INTERSNP

Genomweite Interaktionsanalyse

mit a-priori Information

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Christine Ellen Herold

aus Stuttgart

Bonn Mai 2011

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: PD Dr. rer. nat. Tim Becker

2. Gutachter: Prof. Dr. rer. nat. Jürgen Bajorath

Tag der Promotion: 22.07.2011

Erscheinungsjahr: 2011

Inhaltsverzeichnis

Abkürzungsverzeichnis

CNV	Copy Number Variation (Kopienzahlvariation)
DNA	Deoxyribonucleic Acid (Desoxyribonukleinsäure)
eQTL	Expression Quantitative Trait Loci
FG	Freiheitsgrad
FID	Familien-ID
GWAS	Genome-wide Association Study (genomweite Assoziationsstudie)
GWHA	Genome-wide Haplotype Analysis (genomweite Haplotypanalyse)
GWIA	Genome-wide Interaction Analysis (genomweite Interaktionsanalyse)
HWE	Hardy-Weinberg Equilibrium (Hardy-Weinberg-Gleichgewicht)
LD	Linkage Disequilibrium (Kopplungsungleichgewicht)
LOD	Logarithmic Odds Ratio
MAF	Minor Allele Frequency (Häufigkeit des seltenen Allels)
MC	Monte-Carlo
MB	Megabasen
MPI	Message Passing Interface
MR	Missingrate
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
PAA	Pathway Association Analysis (Pathwayassoziationsanalyse)
PID	Personen-ID
QC	Quality Control (Qualitätskontrolle)
QTDT	Quantitative Transmission Disequilibrium Test
RNA	Ribonucleic acid (Ribonukleinsäure)
RR	Relatives Risiko
OpenMP	Open Multi-Processing
OR	Odds Ratio (Chancenverhältnis)
SNP	Single Nucleotide Polymorphism (Einzelnukleotid-Polymorphismen)

Kapitel 1

Einleitung

1.1 Einführung in das Thema der Arbeit

„Es sind die kleinen Unterschiede, die uns zu unverwechselbaren Individuen machen. Denn abgesehen vom Sonderfall eineiiger Zwillinge gleicht kein Erbgut dem anderen. Jeder Mensch hat eine individuelle Zusammensetzung an Genvarianten. Diese Mischung bestimmt unsere Augenfarbe, die Farbe der Haare und zum Teil auch unsere Persönlichkeit. Diese Genvariationen machen uns aber eventuell auch anfällig für Krankheiten oder beeinflussen die Wirksamkeit von Medikamenten. Sie sind dafür verantwortlich, dass der eine leichter Übergewicht bekommt oder eher zu Asthma neigt als der andere, oder dafür, wie gut depressive Patienten auf Medikamente reagieren. Der Großteil dieser individuellen Gen-Unterschiede beruht auf winzigen Abweichungen im Erbgut: auf der Veränderung nur eines Buchstaben im Alphabet der DNA (Desoxyribonukleinsäure, engl. deoxyribonucleic acid)." [NGFN, 2011]

Vor ca. 20 Jahren war es kaum vorstellbar, mehr als einige wenige DNA-Fragmente gleichzeitig zu untersuchen. Heute können mithilfe der DNA-Chip-Technologie in einem Experiment tausende von Genen parallel analysiert werden. Dazu wurden in den letzten Jahren Hochdurchsatzverfahren entwickelt, welche die Durchführung einer Vielzahl von Analysen in kurzer Zeit ermöglichen und folglich große Datenmengen erzeugen [NGFN, 2011]. Etwa 25 Jahre nachdem 1953 James Watson und Francis Crick die räumliche Struktur der DNA entschlüsselt hatten, wurden parallel zwei Technologien entwickelt, um die Abfolge der Basenpaare auf der DNA zu bestimmen. Fred Sanger entwickelte die Didesoxymethode, auch als Kettenabbruchmethode bekannt, wobei DNA enzymatisch sequenziert wird. Im Gegensatz dazu hatten Maxam und Gilbert die Idee, DNA chemisch abzubauen und so die Sequenz zu bestimmen. Für beide Sequenziermethoden gab es 1980 den Nobelpreis für Chemie.
Über 30 Jahre dominierte Sangers Methode, die auf Sequenzierung durch Synthese beruht, aufgrund der Automatisierbarkeit, der Qualität der Sequenzen und der längeren Leseweiten. In der zweiten Generation der Sequenziermaschinen setzen sich immer mehr die nicht-Sanger-Methoden durch, da diese noch schnelleres Sequenzieren ermöglichen und noch längere Leseweiten erlauben [Schuster, 2008]. Diese neuen Hochdurchsatztechnologien fasst man unter dem Begriff „Next generation sequencing" (NGS) zusammen. Sie können unter anderem die Identifizierung und

Katalogisierung der Häufigkeit von bestimmten Genvarianten (SNPs, engl. Single Nucleotide Polymorphism) beschleunigen. In genomweiten Assoziationsstudien (GWAS) werden über das gesamte Genom verteilte SNPs in Fällen (Patienten) sowie in Kontrollen (gesunden Personen) mit statistischen Verfahren analysiert, um krankheitsassoziierte Gene und ihre natürlich vorkommenden häufigen Varianten zu identifizieren.

1.2 Grundlagen

1.2.1 Genetik

Jede menschliche Zelle enthält 46 Chromosomen, 22·2 homologe Autosomen und zwei Geschlechtschromosomen XX (Frauen) bzw. XY (Männer). Die Chromosomen haben paarweise die gleiche Größe, Gestalt und das gleiche charakteristische Bandmuster. Da jedes Autosom doppelt vorhanden ist, spricht man von einem diploiden Chromosomensatz. Chromosomen sind langkettige Moleküle aus Desoxyribonukleinsäure (DNA), die aus einer linearen Abfolge einzelner Bausteine, den Nukleotiden bestehen [Bickeböller and Fischer, 2007]. Diese setzen sich jeweils aus dem Zucker Desoxyribose, Phosphatresten und einer der vier Basen Adenin (A), Cytosin (C), Guanin (G) oder Thymin (T) zusammen. Ein einfacher Chromosomensatz hat eine Gesamtlänge von ca. 3 Milliarden Basenpaaren. Da immer A und T sowie C und G ein Basenpaar bilden, genügt es für formale Zwecke nur einen DNA-Strang zu betrachten. In der Reihenfolge der Nukleotide ist die genetische Information durch einen „Dreibuchstabencode“ (Basen-Triplett) verschlüsselt. Diese Information wird von den Zellen benötigt um funktionsfähige biologische Produkte (Proteine) herzustellen. Ein Gen ist ein DNA-Abschnitt, der den Code für die Synthese eines Proteins enthält.
Die Vervielfältigung der DNA findet über zwei Arten der Zellteilung statt: Mitose und Meiose. Bei der Mitose wird die genetische Information verdoppelt und auf zwei identische Tochterzellen verteilt. Auf diese Weise werden alle diploiden Körperzellen vermehrt. Die Zellteilung der Keimzellen geschieht durch Meiose. Keimzellen sind Samen- und Eizellen (Gameten), die nur einen haploiden Chromosomensatz besitzen. Wie in Abbildung 1.1 dargestellt, verdoppeln sich die Keimzellen zunächst. Während der 1. Reifeteilung teilen sich die homologen Chromosomen, wobei jeweils zwei Schwesterchromatiden zusammenbleiben. Bei der 2. Reifeteilung werden die Schwesterchromatiden getrennt und es entstehen vier neue Gameten. Durch die Verschmelzung von Ei- und Samenzelle entsteht wieder ein vollständiger diploider Chromosomensatz. Die genetische Variabilität entsteht jedoch schon vor der 1. Reifeteilung, wenn sich die homologen Chromosomen aneinanderlagern. Einerseits werden mütterliche und väterliche Erbanlagen dadurch vermischt, dass sich Chromosomenpaare zufällig zusammensetzen und somit 2^{23} verschiedene Kombinationsmöglichkeiten entstehen, andererseits können bei der Aneinanderlagerung Chromatiden derart auseinander brechen und neu verschmelzen, dass DNA-Bruchstücke der väterlichen und mütterlichen Chromatiden zufällig vermischt werden [Bickeböller and Fischer, 2007]. Dieser nicht seltene Austausch von DNA-Stücken wird Crossover genannt. Nach der 1. Reifeteilung sind somit die Schwesterchromatiden nicht mehr identisch.
Die Trennung von ursprünglich gekoppelten Genen auf einem Chromosom wird Rekombination genannt. Gekoppelt bedeutet, dass die Gene dieses Chromosoms

zusammen vererbt wurden. Rekombination zwischen Genen tritt umso häufiger auf, je weiter die Gene auf dem Chromosom auseinander liegen. Sind die Gene weit voneinander entfernt, kann es zu mehreren Crossovers kommen. Zwei weit auseinander liegende Gene auf einem Chromosom werden durch Rekombination regelmäßig getrennt und können somit als ungekoppelt angesehen werden. Die Rekombinationsrate kann deshalb als Maß für die Entfernung zweier gekoppelter Gene auf einem Chromosom benutzt werden, da die Häufigkeit der Rekombinationen zwischen zwei Genen unter den selben Bedingungen immer gleich ist [Hirsch-Kauffmann and Schweiger, 2000]. Je dichter die Genorte zusammen liegen, desto kleiner die Rekombinationshäufigkeit und je weiter die Gene auseinander liegen, desto größer die Rekombinationshäufigkeit.

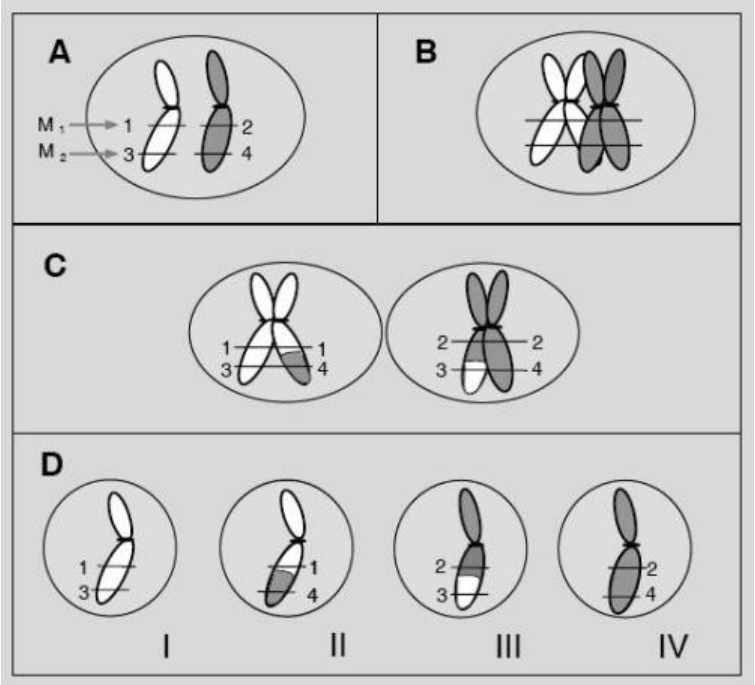

Abbildung 1.1: Quelle: Abb. 1.7 aus Bickeböller & Fischer, 2007 A: Diploider Chromosomensatz, B: Verdopplung und Überlagerung, C: 1. Reifeteilung, D: 2. Reifeteilung, aus der Ausgangszelle entstehen vier Keimzellen. Zur Vereinfachung ist nur ein Chromosomenpaar abgebildet.

Die genetische Information der Zellen ist in der Regel identisch. Jedoch variieren ca. 0,1% der DNA zwischen verschiedenen Individuen. Stellen im Genom, die auf den homologen Chromosomen verschiedene Ausprägungen, sogenannte Allele, haben, nennt man polymorph. Die bei einer Person vorhandenen Kombinationen der beiden Allele heißen Genotyp. Als Phänotyp wird das äußere Erscheinungsbild oder die äußerlich sichtbare Ausprägung bezeichnet. Marker sind Polymorphismen mit einer definierten Lage, deren Allele nach den Mendelschen Regeln vererbt werden. In Analysen werden SNPs als Standard verwendet, während Mikrosatelliten nur vereinzelt und Minisatelliten kaum noch verwendet werden. Ein SNP ist eine Sequenzvariation der DNA, die durch den Austausch einer einzigen Base charakterisiert ist und mit relevanter Häufigkeit ($> 1\%$) in der Population vorkommt. Im menschlichen Genom findet sich im Durchschnitt alle 1.000 Basenpaare ein SNP.

Existieren an einer bestimmten Stelle zwei verschiedene Allele, ist das Individuum für diesen Marker hetero-, ansonsten homozygot. Die Abbildung 1.2 stellt eine DNA-Sequenz zweier homologer Chromosomen dar, also die jeweils von Mutter und Vater geerbten Chromosomenkopien einer Person. Die Sequenzen sind überwiegend gleich, jedoch unterscheiden sie sich an der markierten Position, dem SNP. Es handelt sich hier um einen C/T-SNP und die Person ist heterozygot mit dem Genotypen (C,T). Wie man sieht, ist für die Definition entscheidend welchen DNA-Strang der Doppelhelix man wählt. Würde man den anderen Strang als Referenz ansehen, hätte man hier einen G/A-SNP.

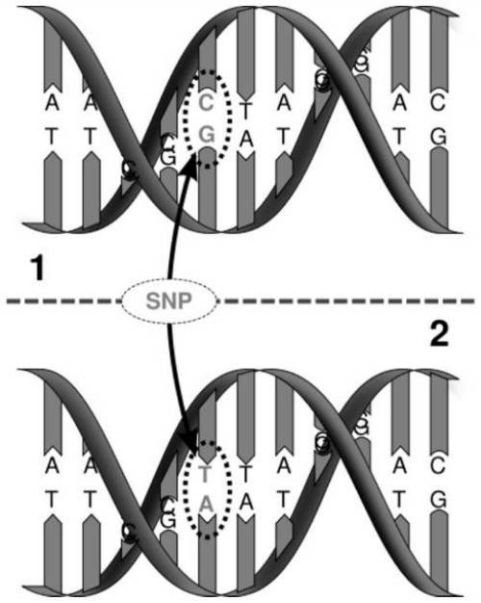

Abbildung 1.2: Die Abbildung stellt zwei Chromosomenkopien dar, die sich in einem einzigen Basenpaar (markierte Stelle) unterscheiden (Quelle: http://www.dnabaser.com/articles/SNP/SNP-single-nucleotide-polymorphism.html).

Es gibt schätzungsweise 7 bis 9 Millionen SNPs im Humangenom [International HapMap Consortium, 2007]. Die NCBI dbSNP Datenbank (Version 129 basierend auf NBCI genome build 36.3) beinhaltet mehr als $14,7 \cdot 10^6$ SNPs, wovon $6,6 \cdot 10^6$ SNPs geprüft und eindeutig im menschlichen Genom lokalisiert wurden.

Im einfachsten Fall können die Allele bzw. Genotypen eines einzelnen SNPs das Krankheitsrisiko bei einer bestimmten Person verändern. Oftmals sind die Zusammenhänge aber komplexer und es ist sinnvoll, Haplotypen zu betrachten. Als Haplotypen bezeichnet man Einheiten von Loci auf demselben Chromosom, die gemeinsam vererbt werden. Es gibt drei wichtige Gründe Haplotypen zu betrachten [Clark, 2004]: Zum einen haben Haplotypen direkte biologische Relevanz, zum anderen ist die genetische Variabilität (Mutation, Selektion, Migration) einer Population auf natürliche Weise in Haplotypen organisiert. Zusätzlich führt die Zusammenfassung mehrerer SNPs in Haplotypen zu einer reduzierten Anzahl der Dimensionen bei statistischen Tests. Da die direkte Bestimmung von Haplotypen im Labor sehr aufwendig ist, werden statistische Verfahren verwendet um Haplotypen zu schätzen [Becker and Knapp, 2004]. Diese Methoden benutzen direkt oder indirekt die Information von Personen (Familien) mit bekannter Phase (es ist bekannt, welche Loci von der Mutter und welche vom Vater vererbt wurden),

um für andere Personen (Familien) mit unbekannter Phase (es ist unbekannt, welche Loci von welchem Elternteil stammen) die Wahrscheinlichkeit des Auftretens der verschiedenen möglichen Haplotyperklärungen zu gewichten. Aufgrund der Entwicklungsgeschichte treten die Allele benachbarter SNPs nicht unabhängig voneinander auf (Kopplungsungleichgewicht). Das hat zur Folge, dass von den vielen theoretisch möglichen Haplotypen nur einige tatsächlich vorkommen.
Während der Populationsgeschichte reichert sich der Pool der vorhandenen Haplotypen durch Mutation und Rekombination an. Dabei unterscheidet man Foundermutation und Hot-Spot-Mutation. Der Unterschied wird in Abbildung 1.3 deutlich. Bei einer Foundermutation sind lange DNA-Abschnitte, also der Haplotyp, identisch, während bei einer Hot-Spot-Mutation die Mutation immer wieder neu entsteht und somit auch das Umfeld nicht einheitlich ist.

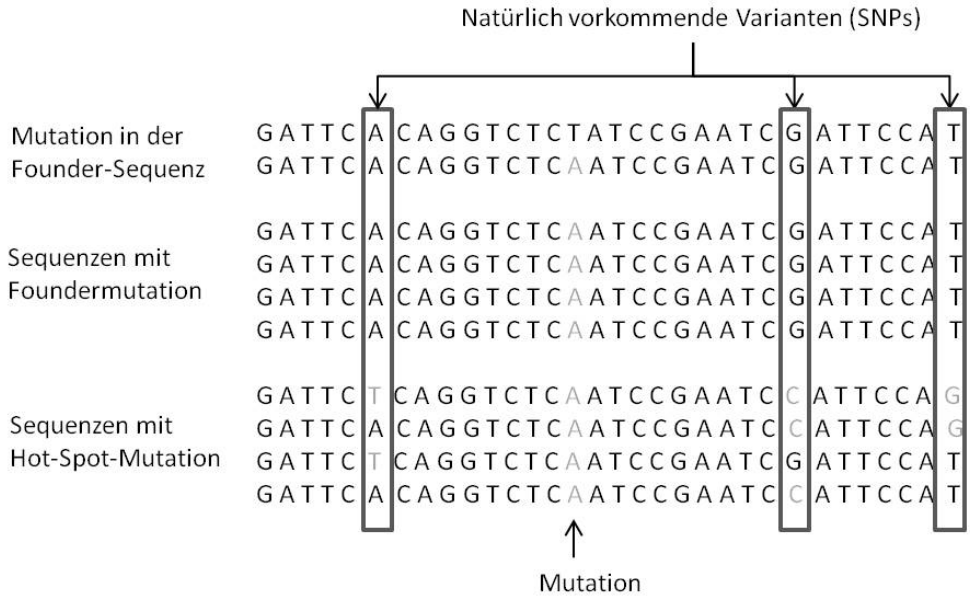

Abbildung 1.3: Unterschied von Foundermutation und Hot-Spot-Mutation (angelehnt an: Spektrum der Wissenschaft 1/2006, von Alison Kendall).

Wie weit die Foundermutation zurückliegt, lässt sich aus der Länge des gemeinsamen Haplotyphintergrunds und der Häufigkeit der Foundermutation bestimmen. Wie in Abbildung 1.4 sichtbar, wird der Haplotyp mit steigender Generationenzahl immer kürzer.

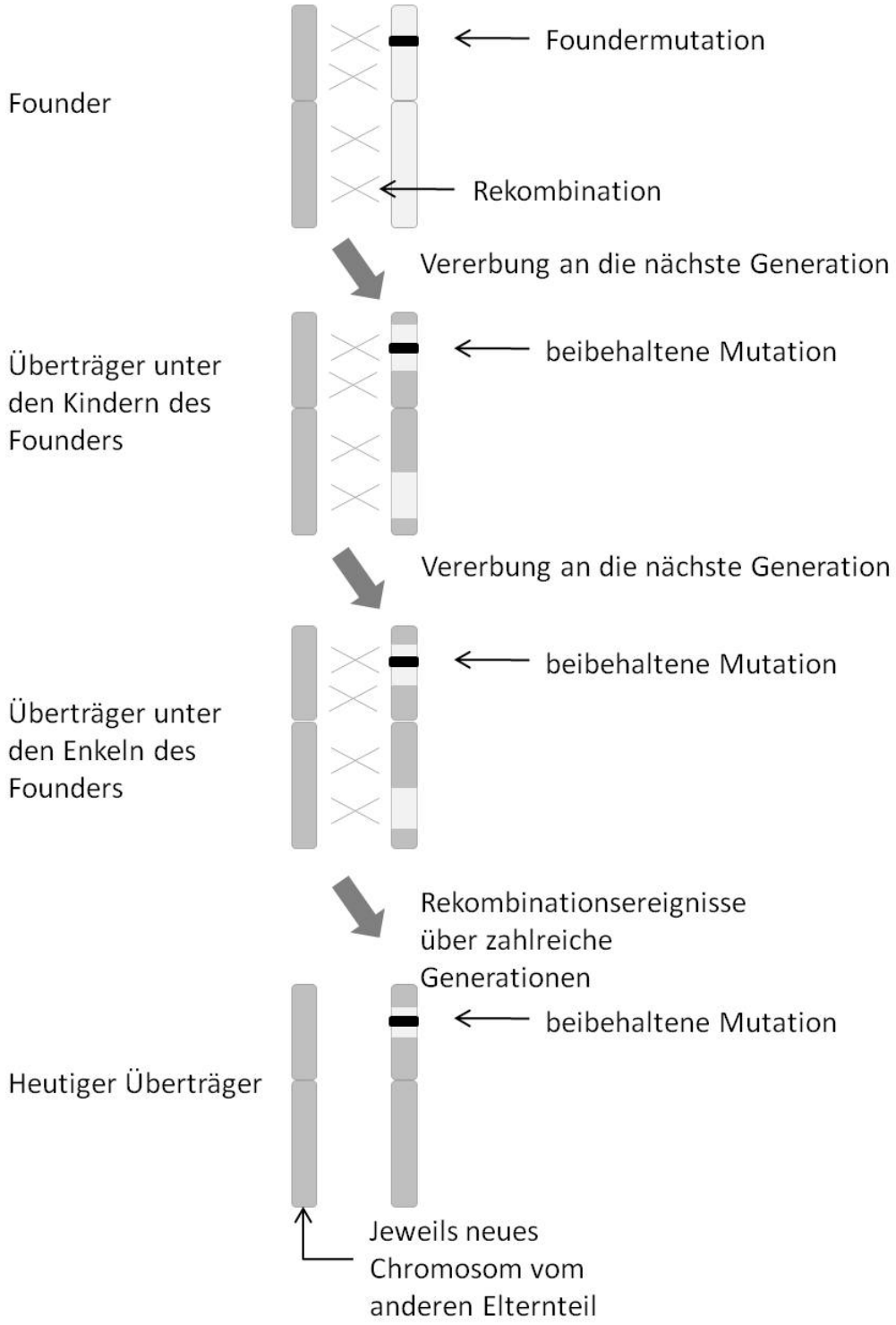

Abbildung 1.4: Diese Abbildung verdeutlicht, dass sich die Haplotyplänge von Generation zu Generation verringert (angelehnt an: Spektrum der Wissenschaft 1/2006, von Alison Kendall).

1.2.1.1 Kopplungsgleichgewicht

Kopplungsgleichgewicht (Linkage Equilibrium) liegt vor, wenn die Allelverteilungen an zwei Genorten unabhängig voneinander sind, anderenfalls handelt es sich

um Kopplungsungleichgewicht (Linkage Disequilibrium, LD). Kopplungsungleichgewicht bedeutet also, dass Allele verschiedener Genorte (Marker/SNPs) häufiger gemeinsam auftreten, als bei zufälliger Verteilung zu erwarten wäre. Zur Erläuterung betrachte man einen Genort 1 mit Allelen A, a und einen Genort 2 mit Allelen B, b. Die zugehörigen Haplotypen sind AB, Ab, aB und ab. Weiter sei $f()$ die Häufigkeit eines Allels oder Haplotyps in der Bevölkerung bzw. in einer Stichprobe. Unter dem Kopplungsgleichgewicht ergibt sich somit [Knapp et al., 2001]:

$$f(AB) = f(A)f(B)$$

Die Abweichungen der Haplotyphäufigkeiten vom Produkt der Allelhäufigkeiten ergeben das Kopplungsungleichgewicht:

$$D = f(AB) - f(A)f(B)$$

$D = 0$ würde somit bedeuten, dass die Loci im Gleichgewicht stehen. LD entsteht meistens durch neue Varianten in einer Gamete. Diese Varianten bleiben über Generationen mit den eng benachbarten Allelen des Ausgangschromosoms im Kopplungsungleichgewicht, da sie auf der Ursprungssequenz weiter vererbt werden. Kopplungsungleichgewicht ist der natürliche Ausgangszustand für die durch Foundermutationen neu entstandenen SNPs und die SNPs, die sich bereits auf dem Haplotyphintergrund befinden. Im Laufe der Generationen reduziert sich das LD durch Rekombination und wird zu einem lokalen Phänomen. Innerhalb von durchschnittlich 20kb bis 100kb großen Bereichen ist starkes LD zwischen SNPs der Normalfall.
Maße für das LD sind D' und r^2. D' ist eine Normierung des Disequilibriumskoeffizient $D = f(AB) - f(A)f(B)$ und ist im Falle eines Kopplungsungleichgewichts von Null verschieden. Dieser Koeffizient gibt also die Stärke einer allelischen Assoziation an und hängt von den Allelhäufigkeiten ab. Der Korrelationskoeffizient r der Vierfeldertafel ist hingegen definiert durch $r^2 = D^2/(f(A)f(a)f(B)f(b))$ und ist das ausschlaggebende Maß für die relative Power zweier SNPs im LD. Im Falle des Kopplungsgleichgewichts sind beide Maße Null. Man spricht vom kompletten LD, wenn $D' = 1$ und der Korrelationskoeffizient $r^2 < 1$ ist, was in unserem Beispiel mit zwei SNPs genau dann der Fall ist, wenn nur drei der vier möglichen Haplotypen existieren. Das perfekte LD zeichnet sich durch $D' = 1$ und $r^2 = 1$ aus, was wiederum bedeutet, dass nur zwei verschiedene Haplotypen existieren. Abbildung 1.5 zeigt eine typische LD-Struktur für benachbarte SNPs.

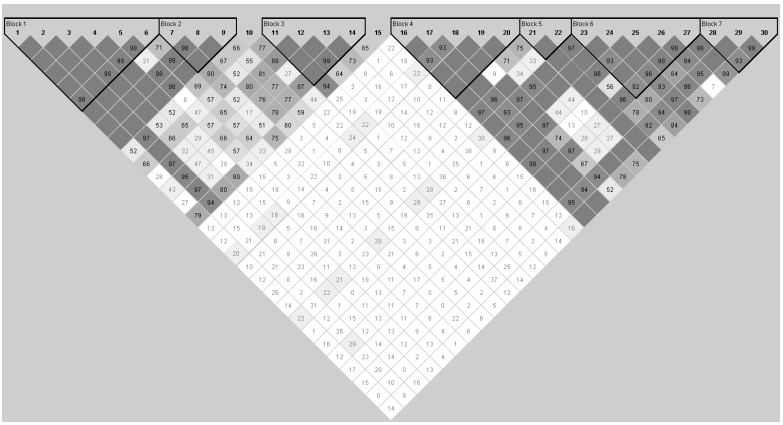

Abbildung 1.5: Diese Abbildung wurde mit der Software Haploview [Barrett et al., 2005] erstellt, die LD-Blöcke visualisiert. Je dunkler die Felder desto größer ist das LD zwischen den SNPs.

1.2.2 Hardy-Weinberg-Gleichgewicht

Eine Grundregel der Populationsgenetik stellt das Hardy-Weinberg-Gleichgewicht (engl. Hardy-Weinberg-Equilibrium, HWE) [Hardy, 1908] dar. Es gilt für große „Standardpopulationen“ bei denen Mutation, Migration und zufälliger Gendrift nicht für die Verteilung der Allelfrequenzen ins Gewicht fallen. Je größer eine Population ist, desto unwahrscheinlicher ist das Auftreten von Zufallsabweichungen und desto eher liegen die Allelfrequenzen im Gleichgewicht. Besteht ein Populationsgleichwicht, ändert sich diese Verteilung von einer Generation zur nächsten nicht. Der Test auf das Hardy-Weinberg-Gleichgewicht deckt allelspezifische Unregelmäßigkeiten in untersuchten Populationen auf. Diese Abweichungen werden in Fall-Kontroll-Studien gewöhnlich durch fehlerhafte Genotypisierung verursacht [Balding et al., 2007].
Gehen wir von einem biallelischen Locus mit Allelen A, a und Allelhäufigkeiten p, q in einer sehr großen Population aus, um das HWE näher zu erläutern. In einer Standardpopulation treten die Allele unabhängig voneinander auf und daher gelten folgende Gleichungen für die Genotyphäufigkeiten:

$$\begin{aligned} P(AA) &= p^2 \\ P(Aa) &= 2pq \\ P(aa) &= q^2 \end{aligned}$$

Es gilt:

$$p^2 + 2pq + q^2 = 1$$

Die Gültigkeit des HWE in einer Stichprobe mit empirisch bestimmten Genotypen lässt sich statistisch prüfen. Beim Test auf das Hardy-Weinberg-Gleichgewicht schätzt man zunächst die Allelhäufigkeiten aus der Stichprobe, indem die Anzahl beobachteter Allele durch die Gesamtzahl der Allele $2N$ geteilt wird, wobei N die Anzahl der Personen ist. Anschließend berechnet man die erwarteten Genotyphäufigkeiten (E_i) unter HWE und vergleicht die beobachtete (O_i) und die erwartete (E_i) Anzahl von Genotypen mit einem χ^2-Test mit einem Freiheitsgrad:

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)^2}{E_i}$$

Die χ^2-Verteilung liefert dann den zugehörigen p-Wert.

1.2.3 Pathways

Das Leben basiert auf einer Folge von biochemischen Prozessen und chemischen Reaktionen, wobei jeder Vorgang dazu führt, dass Moleküle miteinander interagieren und somit eine chemische oder physikalische Veränderung in den lebenden Systemen bewirken [Schreiber, 2001]. Ein Pathway (Stoffwechselweg) beschreibt eine Aneinanderreihung von Reaktionen. Pathways können in metabolische Stoffwechselwege und regulatorische Pfade unterteilt werden. Als metabolischer Stoffwechselweg wird die Gesamtheit aller biochemischen Vorgänge beim Aufbau, Abbau und Umbau eines Organismus sowie dessen Stoffaustausch mit der Umwelt bezeichnet. Die beiden grundlegenden Stoffwechselvorgänge sind Anabolismus (z.B. Photosynthese, Chemosynthese und Verdauungsprozesse) und Katabolismus (Atmung und Gärung). Dabei ermöglichen oder beschleunigen Enzyme, effektive biologische Katalysatoren, die biochemischen Reaktionen in Zellen. Zu den regulatorischen Pfaden gehören Signal- und Transportwege sowie Regulation der Genexpression.
Metabolische Stoffwechselwege und regulatorische Pfade können als interzelluläre Netzwerke beschrieben werden, da sie auf elementaren Bausteinen einer Zelle wie Genen, Transkripten, Proteinen und Metaboliten basieren [Schreiber, 2009]. Metabolite, einfache Moleküle, entstehen als Zwischenstufe oder Abbauprodukt von Stoffwechselvorgängen und werden durch Enzyme ineinander umgewandelt. Proteine können ebenfalls miteinander interagieren und die Aktivität von Genen regulieren, welche auch durch Metabolite beeinflusst werden kann. Daraus entsteht ein komplexes Netzwerk aus Interaktionen und Abhängigkeiten, auf dem wiederum weitere biologische Netzwerke aufbauen können. Dies können interzelluläre Signalnetzwerke, welche Interaktionen zwischen Zellen beschreiben, hormonelle Netzwerke, die die Kommunikation zwischen Geweben und Organen repräsentieren, oder neuronale Netzwerke, welche die Verschaltungen von Neuronen darstellen, sein, um einige Beispiele zu nennen [Schreiber, 2009].
Der deutsche Biochemiker Gerhard Michal hat 1968 erstmals eine graphische Darstellung der in Lebewesen ablaufenden biochemischen Reaktionen und Interaktionen erstellt. Diese Darstellung „Biochemical Pathways“ wurde in Form eines Posters, welches ca. 1.500 Reaktionen und zugehörige Substanzen umfasste (auf der aktualisierten Version ca. 10.000 Reaktionen und Substanzen), veröffentlicht [Michal, 1993]. Moderne Analysemethoden tragen heute zum besseren Verständnis der einzelnen Elemente und Interaktionen in den biologischen Systemen bei und somit zum Verstehen des Gesamtsystems [Schreiber, 2009]. Das Wissen über die Strukturen und die Funktionsweise der Pathways gibt uns neue Möglichkeiten und Ansätze für die Entwicklung von Medikamenten und Therapien, da so gezielt Eingriffe in die Prozesse im menschlichen Organismus vorgenommen werden können.

1.2.4 Statistische Grundbegriffe

Die Aufgabe statistischer Methoden besteht darin, aus Stichproben Aussagen über eine größere Grundgesamtheit abzuleiten [Hilgers et al., 2007]. In dieser Arbeit werden statistische Methoden verwendet um epidemiologische Fragen zu klären. Die Epidemiologie befasst sich einerseits mit der Untersuchung der Verteilung von

Krankheiten, allgemeiner von Phänotypen in Bevölkerungsgruppen, und andererseits mit deren Einflussfaktoren. Die Genetische Epidemiologie spezialisiert sich insbesondere auf die Untersuchung genetischer Einflüsse bei monogenen und komplexen Erkrankungen sowie die Entwicklung statistischer Verfahren hierfür.

1.2.4.1 Testen von Hypothesen

Ein statistischer Test liefert nach bestimmten Regeln die Entscheidung darüber, ob eine vorgegebene Hypothese anhand von Daten unter einem zuvor festgelegten Signifikanzniveau verworfen werden sollte [Heinecke et al., 1992]. Das Festhalten an einer Hypothese bedeutet, dass die Entscheidung offen bleibt, da eine Hypothese, die nicht verworfen werden kann, nicht bewiesen ist. Das logische Prinzip des statistischen Testens gleicht dem des indirekten Beweises. Zum indirekten Beweis einer Hypothese H_1 nimmt man an, dass die Verneinung von H_1 richtig sei. Die Verneinung von H_1 bezeichnet man als Nullhypothese H_0, H_1 heißt auch Alternativhypothese. Wenn es gelingt, aus der Verneinung von H_1, also aus der Nullhypothese, einen Widerspruch abzuleiten, ist der indirekte Beweis gelungen, und an H_1 wird festgehalten.
Beim statistischen Testen führt das Eintreten eines unter H_0 unwahrscheinlichen Ergebnisses in einem entsprechend geplanten Versuch zum Verwerfen von H_0 [Heinecke et al., 1992]. Die Irrtumswahrscheinlichkeit ist dabei sehr wichtig, da nicht ausgeschlossen werden kann, dass eine Fehlentscheidung getroffen wird. Hierbei wird zwischen Fehler 1. Art und Fehler 2. Art unterschieden. Beim Fehler 1. Art wird die in Wirklichkeit richtige Nullhypothese als nicht richtig erkannt und verworfen. Die obere Schranke für die Wahrscheinlichkeit des Fehlers 1. Art wird mit α bezeichnet und hat üblicherweise den Wert $0,05$ oder $0,01$. Beim Fehler 2. Art erkennt man eine in Wirklichkeit richtige Gegenhypothese nicht als richtig und somit wird fälschlicherweise an der Nullhypothese festgehalten. Die Wahrscheinlichkeit für diesen Fehler wird mit β bezeichnet. Bei einem Experiment wird α explizit angegeben, während β nur geschätzt werden kann.
Die Wahrscheinlichkeit, dass eine richtige Gegenhypothese im Test auch tatsächlich als richtig erkannt wird, ist $(1 - \beta)$. Man nennt diese Wahrscheinlichkeit die Power (Mächtigkeit) eines Tests. In der Praxis ist die Power ein entscheidendes Maß für die Verwendbarkeit eines Tests, da man vorhandene Zusammenhänge natürlich immer finden möchte. Die Power hängt von der Art der Daten und der Datenerhebung sowie dem Stichprobenumfang ab. Weitere Einflüsse sind der verwendete Test und natürlich die Irrtumswahrscheinlichkeit α. Zu beachten ist, dass Aussagen der Statistik nie deterministisch sind. Auch eine verworfene Nullhypothese kann in Wirklichkeit richtig sein.

1.2.4.2 p-Wert

Um statistisch zu überprüfen, ob ein SNP mit einer Krankheit assoziiert ist, wird zunächst eine Nullhypothese aufgestellt. Die Nullhypothese lautet in diesem Fall: „Der SNP ist nicht mit der Krankheit assoziiert“. Anhand der Daten wird mit Hilfe einer Teststatistik ein p-Wert ausgerechnet, der die Entscheidung beeinflusst, ob die Nullhypothese verworfen wird oder nicht. Der p-Wert quantifiziert also die Wahrscheinlichkeit, dass das gefundene Testergebnis (oder ein noch extremeres Ergebnis) zu beobachten ist, wenn die Nullhypothese richtig ist. Wenn p kleiner ist als das zuvor festgelegte Signifikanzniveau α, wird die Nullhypothese verworfen

und die Alternativhypothese angenommen. Beim p-Wert gilt das Plausibilitätskriterium „Je kleiner, desto besser“. Dabei ist zu beachten, dass der p-Wert lediglich besagt, ob ein signifikanter Unterschied existiert. Er enthält jedoch keine Information über die Größe des gefundenen Effekts [Weiß, 2008].

1.2.4.3 Multiples Testen - Bonferroni-Korrektur

Wenn zur selben bzw. inhaltlich zusammengehörigen Fragestellung mehrere Hypothesen getestet werden, spricht man vom multiplen Testen . Dies führt zu einem Anstieg des Fehlers 1. Art für das Gesamtexperiment, da für jede Hypothese mit der Wahrscheinlichkeit α ein Fehler 1. Art begangen werden kann. Es gibt verschiedene Ansätze das Ansteigen des Fehler 1. Art zu kontrollieren und somit den p-Wert für das multiple Testen zu korrigieren. Eine davon ist die Bonferroni-Korrektur. Sie besteht darin, für k Tests jeweils das Signifikanzniveau α/k zu wählen. Sind die einzelnen Tests voneinander abhängig, ist diese Korrektur sehr streng und es kommt zu einer Reduzierung der Power. In solchen Fällen ist es sinnvoll, Simulationsmethoden zu verwenden [Westfall and Young, 1993]. Dieser Ansatz wird in Abschnitt 3.5.8 ausführlich erläutert.

1.3 Genetische Epidemiologie

Die Genetische Epidemiologie beschäftigt sich mit genetischen Risikofaktoren und deren Zusammenwirken mit Umweltfaktoren bei der Entstehung und dem Verlauf von Krankheiten [Bickeböller and Fischer, 2007]. Das Hauptziel der Genetischen Epidemiologie ist daher die Lokalisation, Identifikation und Bestimmung der Effektstärke von DNA-Sequenzvariationen im menschlichen Erbgut (Genom), die bei der Entstehung einer Krankheit mitverantwortlich sind. Das Gebiet der Genetischen Epidemiologie vereint Forschungsmethoden der Humangenetik, der traditionellen Epidemiologie, der genetischen Statistik und der Bioinformatik. Die Erkenntnisse der Genetischen Epidemiologie sollen Prognose, Präventionsmaßnahmen und neue Therapieformen für die erforschten Krankheiten ermöglichen. Mit der Gründung der Zeitschrift Genetic Epidemiology 1984 und der International Genetic Epidemiology Society (IGES) 1992 hat sich die Genetische Epidemiologie in den 1980er Jahren als eigenständiges Forschungsgebiet etabliert. Klassische Strategien der Genetischen Epidemiologie sind zum einen die Kopplungsanalyse und zum anderen die Assoziationsanalyse.

1.3.1 Studientypen

Die häufigste Studienform in der Genetischen Epidemiologie ist mittlerweile die Fall-Kontroll-Studie, bei welcher die Frequenz von Allelen oder Genotypen zwischen entsprechenden Kollektiven verglichen wird. Es handelt sich hierbei um retrospektive Stichproben mit erkrankten Fällen und gesunden Kontrollen. Als Marker werden in dieser Arbeit ausschließlich SNPs betrachtet. Ein signifikanter Frequenzunterschied eines Allels oder Genotyps zwischen Fällen und Kontrollen kann also ein Hinweis darauf sein, dass das Allel direkt oder indirekt eine Rolle bei der Entstehung der Erkrankung spielt.
Im Gegensatz zu retrospektiven Studien werden bei prospektiven Studien die Daten erst nach Festlegung der Hypothese erhoben. Dadurch kann ein genauerer

kausaler Zusammenhang zwischen Risikofaktor und Krankheit hergestellt werden, jedoch sind diese Studien recht aufwendig und kostspielig.
Familienbasierte Tests vermeiden populationsspezifische Stratifikationseffekte, indem sie die nicht vererbten Allele der Eltern von betroffenen Personen als Kontrollallele nutzen [Spielman et al., 1993; Balding et al., 2007]. Bei Fall-Kontroll-Studien können jedoch Probleme auftreten, wenn Populationsstratifikation (siehe Abschnitt 3.3.3) in der Stichprobe vorliegt, d.h. es können positive Testergebnisse entstehen, ohne dass ein biologischer Zusammenhang zwischen dem untersuchten Marker und der Krankheit existiert, geringere Effekte können verstärkt oder wahre Assoziationen maskiert werden. Die familienbasierten Tests sind gegenüber solchen Effekten robust. Jedoch sind diese familienbasierten Assoziationsstudien im Vergleich zu Fall-Kontroll-Studien auf Grund des hohen Rekrutierungsaufwandes stark zurückgedrängt worden. Im Hinblick auf die in den Fokus rückende Untersuchung von seltenen Varianten könnten sie jedoch wieder an Bedeutung gewinnen. Die Familienstruktur bietet nämlich eine zusätzliche Möglichkeit, die Korrektheit der Bestimmung seltener Allele zu überprüfen. Zur Korrektur für Stratifikation in Fall-Kontroll-Studien siehe auch Abschnitt 3.3.3.

1.3.2 Relatives Risiko und Odds Ratio

Eine Assoziation einer Erkrankung zu einem genetischen Polymorphismus liegt vor, wenn die Häufigkeit eines bestimmten Allels in der Population der Erkrankten sich von der Häufigkeit in einer Kontrollgruppe unterscheidet [Knapp et al., 2001]. Um zu testen, ob ein SNP mit einer Krankheit assoziiert ist, gibt es verschiedene Ansätze der Einzelmarkeranalyse.
Der genotyp- oder allelbasierte χ^2-Test vergleicht die beobachteten Häufigkeiten b_{ij} mit den erwarteten Häufigkeiten e_{ij}, wobei $i = 1, 2, 3$ für die drei verschiedenen Genotypen steht und $j = 1, 2$ für den Fall-Kontroll-Status.
Für den genotypbasierten Test ergibt sich somit für die 2×3-Tafel (siehe Tabelle 1.1) die folgende Teststatistik, welche durch eine χ^2-Verteilung mit zwei Freiheitsgraden (FG) approximiert werden kann:

$$T_G = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(b_{ij} - e_{ij})^2}{e_{ij}}$$

Um die Anzahl der Freiheitsgrade zu berechnen, multipliziert man die (Anzahl der Spalten-1) mit der (Anzahl der Zeilen-1). Diese 2×3-Tafel (siehe Tabelle 1.1) lässt sich in eine 2×2-Tafel (Vierfeldertafel, siehe Tabelle 1.2) vereinfachen, wenn die beiden Allele A und T unabhängig voneinander auftreten, also wenn das HWE gilt.

	Genotyp AA	Genotyp AT	Genotyp TT	
Fälle	F_2	F_1	F_0	N_F
Kontrollen	K_2	K_1	K_0	N_K
	N_2	N_1	N_0	N

Tabelle 1.1: 2×3-Feldertafel für die Genotypverteilung bei einer Fall-Kontroll-Studie.

	Allel A	Allel T	
Fälle	$2 \cdot F_2 + F_1$	$F_1 + 2 \cdot F_0$	$2N_F$
Kontrollen	$2 \cdot K_2 + K_1$	$K_1 + 2 \cdot K_0$	$2N_K$
	N_A	N_T	$2N$

Tabelle 1.2: Vierfeldertafel für die Allelverteilung bei einer Fall-Kontroll-Studie.

Aus der Vierfeldertafel kann nun die Teststatistik T_A für den allelbasierten χ^2-Test erstellt werden:

$$T_A = \frac{2N\left((2F_2 + F_1) \cdot (K_1 + 2K_0) - (F_1 + 2F_0) \cdot (2K_2 + K_1)\right)^2}{2N_F \cdot N_A \cdot 2N_K \cdot N_T}$$

Da es bei dem allelbasierte χ^2-Test eine Merkmalsausprägung weniger gibt, reduziert sich die Anzahl der Freiheitsgrade auf eins. Die Formel für den allelbasierten χ^2-Test lässt sich auch folgendermaßen schreiben [Knapp et al., 2001]:

$$T_A = \frac{(P_A^F - P_A^K)^2}{[P_A(1 - P_A)](\frac{1}{2N_F} + \frac{1}{2N_K})}$$

mit

P_A	$=$	$N_A/2N$	Allelfrequenz von A
P_A^F	$=$	$(2F_2 + F_1)/2N_F$	Allelfrequenz von A bei Fällen
P_A^K	$=$	$(2K_2 + K_1)/2N_K$	Allelfrequenz von A bei Kontrollen
P_{AA}	$=$	P_A^2	

Im Gegensatz zum genotyp- oder allelbasierten χ^2-Test muss beim Armitage-Trendtest [Armitage, 1955] die Population, aus der die Daten stammen, nicht im HWE stehen. Die Teststatistik ist der des χ^2-Tests ähnlich, jedoch wird ein zusätzlicher Korrekturterm ergänzt:

$$T_{trend} = \frac{(P_A^F - P_A^K)^2}{[P_A(1 - P_A) + \underbrace{(P_{AA} - P_A^2)}_{\texttt{Korrekturterm}}](\frac{1}{2N_F} + \frac{1}{2N_K})}$$

Dieser Korrekturterm berücksichtigt die Abweichungen vom HWE.

Wenn nachgewiesen wurde, dass ein SNP mit einer Krankheit signifikant assoziiert ist, möchte man in der Regel die Art des Zusammenhangs beschreiben. Als Maß für die Stärke einer Assoziation wird das relative Risiko (RR) oder das Odds Ratio (OR), auch Chancenverhältnis genannt, verwendet [Knapp et al., 2001]. Das relative Risiko (RR) definiert die Wahrscheinlichkeit, dass eine Krankheit (D) bei Personen, die mindestens ein Allel H am Markergenort aufweisen, auftritt ($P(D|H^+)$), relativ zu der Wahrscheinlichkeit, dass diese Krankheit bei Personen, die das Allel H nicht besitzen, auftritt ($P(D|H^-)$). Hieraus ergibt sich für das relative Risiko (RR):

$$RR_{H^+} := \frac{P(D|H^+)}{P(D|H^-)}$$

Wenn es sich bei H^+ wirklich um einen Risikofaktor handelt, dann ist RR von Null verschieden. Bei gleich großem Risiko ist RR = 1 liegt keine Assoziation vor. Das Odds Ratio (OR) gibt im Gegensatz zum RR das Chancenverhältnis an und wird wie folgt definiert:

$$OR_{H^+} := \frac{\frac{P(D|H^+)}{1-P(D|H^+)}}{\frac{P(D|H^-)}{1-P(D|H^-)}}$$

Anhand der Allel-Vierfeldertafel (siehe Tabelle 1.3) sollen die Unterschiede noch einmal verdeutlicht werden.

	Allel A	Allel B
Fälle	a	b
Kontrollen	c	d

Tabelle 1.3: Allel-Vierfeldertafel

Grundsätzlich ist die Formel

$$\frac{a \cdot d}{b \cdot c}$$

anzuwenden. Jedoch gibt es dabei einen Unterschied je nach Art der Kontrollen. Hat man gesunde Kontrollen, erhält man gemäß obiger Formel einen Schätzer für das OR, bei Bevölkerungskontrollen erhält man einen Schätzer für das RR.

1.3.3 Monogene und komplexe Krankheiten

Monogene Krankheiten entstehen durch genetische Variation (Mutationen) innerhalb eines bestimmten Gens. Diese Veränderungen können einerseits vererbt worden, aber auch spontan entstanden sein. Die Weitergabe der defekten Gene an die Nachfahren erfolgt je nach Krankheit in verschiedenen Vererbungsmodi [Bickeböller and Fischer, 2007]:

- Autosomal dominant: Als autosomal dominant werden Erbkrankheiten bezeichnet, die schon beim Vorhandensein nur *eines* defekten Gens auftreten (z.B. die neurodegenerative Krankheit Chorea Huntington). Bei diesen Krankheiten gibt es männliche und weibliche Erkrankte, die Übertragung kann über beide Geschlechter geschehen. Somit wird bei voll penetranten Erbgang die Krankheit mit einer Wahrscheinlichkeit von 1/2 vererbt, wenn ein Elternteil heterozygot für das Krankheitsgen ist.

- Autosomal rezessiv: Bei einer autosomal rezessiven Krankheit (z.B. Zystische Fibrose, Phenylketonurie) entsteht die Krankheit nur, wenn die Mutation im homozygoten Zustand, also doppelt vorliegt. Gesunde Eltern eines betroffenen Kindes sind heterozygot am Krankheitsgenort. Aus diesem Grund erben weitere Kinder mit einer Wahrscheinlichkeit von 1/4 von beiden Eltern die Mutation.

- Außerdem gibt es noch die X-chromosomal rezessive und die X-chromosomal dominante Vererbung, auf die hier aber nicht näher eingegangen wird.

Monogene Krankheiten sind in der Bevölkerung relativ selten. Sie zeichnen sich durch hohe bzw. vollständige Penetranz am Krankheitslocus aus. Als Penetranz

wird die bedingte Wahrscheinlichkeit bezeichnet, dass eine Person mit einem bestimmten Genotyp einen Phänotyp ausbildet.
Im Gegensatz zu den monogenen Krankheiten sind die komplexen oder multifaktoriellen Krankheiten in der Bevölkerung weit verbreitet. In den letzten Jahren sind vor allem Krankheiten wie Diabetes, Krebs, Herz-Kreislauferkrankungen, psychiatrische und neurodegenerative Erkrankungen in den Fokus der Genetischen Epidemiologie gerückt. Ihre Entstehung lässt sich nicht alleine auf die Variation *eines* Gens zurückführen. Man vermutet vielmehr, dass genetische Variationen an *mehreren* Loci das Krankheitsrisiko erhöhen und weitere Faktoren (z.B. Umwelteinflüsse) zum Ausbruch der Krankheit beitragen. Das resultiert in moderater oder schwacher Penetranz, aber auch in einer großen Uneinheitlichkeit des Phänotyps. So sind die oben dargestellten Vererbungsmodi in allgemeiner, abgeschwächter Form definiert. Im Beispiel der rezessiven Vererbung würde das bedeuten, dass Personen mit zwei Risikoallelen ein erhöhtes Krankheitsrisiko haben, während Personen mit einem Risikoallel nur ein sehr leicht erhöhtes, aber messbares, Krankheitsrisiko haben.

1.3.4 Kopplungsanalyse

Zur Risikoberechnung für bestimmte Krankheiten werden in der Genetischen Epidemiologie die Kopplungs- und die Assoziationsanalyse angewendet. Die Kopplungsanlyse beschäftigt sich mit Stammbäumen und vergleicht das Vererbungsmuster von Krankheiten mit denen von genetischen Markern. Liegen zwei Gene in relativer Nähe, so werden sie häufiger gemeinsam vererbt als Gene auf verschiedenen Chromosomen, bei denen Unabhängigkeit zu erwarten ist [Bickeböller and Fischer, 2007]. Man spricht von Genkopplung. Je weiter die Gene von einander entfernt sind, desto unabhängiger werden sie vererbt. Die Rekombinationshäufigkeit sei mit θ bezeichnet. Bei vollständiger Kopplung gilt $\theta = 0$. Wenn keine Kopplung vorliegt, d.h. wenn freie Rekombination möglich ist, gilt $\theta = 0,5$. Sei weiter $L(\theta)$ die auf die betrachteten Stammbäume bedingte Wahrscheinlichkeit der Transmissionen der Genotypen für die Rekombinationsfrequenz θ und sei $\hat{\theta}$ der Maximum-Likelihood-Schätzer für θ, also der Wert der $L()$ maximiert. Die Bewertung von Kopplungsanalysen erfolgt in der Statistik anhand von LOD-Scores (Logarithm of Odds), indem man das folgende Risikoverhältnis aufstellt:

$$\mathrm{LOD}(\theta) = \log_{10} \frac{L(\theta)}{L(0,5)}$$

Eine Kopplung wird als signifikant betrachtet, wenn der LOD-Score über 3 liegt, also wenn das Verhältnis der Wahrscheinlichkeiten 1000 übersteigt. Ein hoher LOD-Score bei einem genetischen Marker bedeutet also, dass der Marker häufiger als erwartet mit dem Krankheitslocus gekoppelt vererbt wird, was wiederum bedeutet, dass der Krankheitslocus sich in der Nähe des Markers befindet. Auf diese Weise kann man die Lokalisation eines Krankheitsgenorts bis auf 20 MB (= 20 Millionen Basenpaare) genau bestimmen. In dieser Kandidatenregion kann man dann die Lokalisation des Krankheitsgens mit Assoziationsstudien (siehe Abschnitt 1.3.5) verfeinern. Durch die Möglichkeit der Durchführung von genomweiten Assoziationsstudien (siehe Abschnitt 1.3.6) hat dieses zweistufige Kopplungs-Assoziations-Paradigma der Genetischen Epidemiologie allerdings stark an Bedeutung verloren und wird in dieser Arbeit auch nicht weiter ausgeführt.

1.3.5 Assoziationsanalyse

Der zweite Hauptansatz in der Genetischen Epidemiologie ist die Assoziationsanalyse, bei welcher nach statistischen Zusammenhängen zwischen genetischen Polymorphismen und qualitativen oder quantitativen Merkmalen gesucht wird. Der Unterschied zu den Kopplungsanalysen besteht darin, die Annahme zu nutzen, dass das gleiche Allel oder der gleiche Genotyp mit dem Merkmal in der ganzen Bevölkerung in gleicher Weise assoziiert ist. Assoziationsstudien werden unter anderem eingesetzt, um durch Kopplungsanalyse ermittelte Kandidatenregionen im Genom weiter einzugrenzen oder Kandidatengene direkt zu untersuchen und die verantwortlichen Varianten zu finden. Beim Fall-Kontroll-Ansatz werden Allele von Erkrankten (Fällen) und Gesunden (Kontrollen) aus ethnisch ähnlichen Populationen verglichen. Wenn die Allele eines bestimmten Locus in Fällen häufiger auftreten als in den Kontrollen, kann das ein Hinweis sein, dass sie mit dem Phänotyp (z.B. Krankheit) assoziiert sind. Beim genomweiten Ansatz werden viele Marker, die über das ganze Genom verteilt sind, auf den Zusammenhang mit einer Krankheit untersucht. Findet man in einem der verschiedenen Ansätze Assoziation, gibt es vier mögliche Erklärungen:

- Direkte Assoziation: Ein Allel eines SNPs verursacht direkt ein erhöhtes Krankheitsrisiko. Es ist kausal.
- Indirekte Assoziation: Steht ein SNP in hinreichend starkem LD zum kausalen SNP, wird dort ebenfalls Assoziation zur Krankheit beobachtet. In der Tat bedeutet LD-Korrelation der Allelausprägungen, dass die Assoziation zwischen Krankheit und kausalen SNP auch als Assoziation zwischen der Krankheit und dem SNP im LD sichtbar wird.
- Assoziation wird durch Confounder verursacht, sogenannte „spurious association“. Es handelt sich also nicht wie gewünscht um eine Assoziation zwischen Marker und Phänotyp, sondern um eine Assoziation, die auf unbeobachteten Confoundern (z.B. Populationsstratifikation, Selektion, Inzucht) beruht.
- Das Ergebnis ist falsch positiv (Fehler 1. Art).

1.3.6 Genomweite Assoziationsstudien

Genomweite Assoziationsstudien (GWAS) untersuchen eine große Anzahl über das ganze Genom verteilter Marker (SNPs) auf den Zusammenhang mit einer Krankheit. Da SNP-Ausprägungen lokal stark abhängig sind (im LD), reicht es aus eine repräsentative Auswahl von SNPs im Genom zu analysieren. 1999 wurde das SNP-Konsortium [Thorisson and Stein, 2003] aus bedeutenden pharmazeutischen Unternehmen und Instituten gegründet, um eine umfassende SNP-Karte des menschlichen Genoms zu erstellen. Zunächst war das Ziel, 300.000 SNPs in zwei Jahren zu identifizieren und ihre Lage im Genom zu bestimmen. Tatsächlich wurden in dieser Zeitspanne jedoch ca. $1,4$ Millionen SNPs entdeckt. 2002 wurde darauf aufbauend das HapMap-Projekt [International HapMap Consortium, 2007] ins Leben gerufen. Bei diesem Projekt sollte der Schwerpunkt auf der Bestimmung der Haplotypen liegen, also Kombinationen von SNPs, die gemeinsam vererbt werden. Das Projekt war für drei Jahre geplant, es arbeiteten Gruppen aus Industrie

und akademischen Instituten aus Japan, Großbritannien, Kanada, China, Nigeria und USA zusammen. Das gemeinsame Ziel der beiden Projekte, SNP-Konsortium und HapMap-Projekt, war es, Gene zu entdecken, die bei weit verbreiteten Erkrankungen wie Asthma, Diabetes und anderen eine Rolle spielen. Basierend auf den Daten der HapMap und anderen Quellen werden die SNPs auf den SNP-Chips von bekannten Biochip-Herstellern wie Illumina und Affymetrix ausgewählt. Die Auswahl erfolgt entweder zufällig, physikalisch gleichmäßig über das Genom verteilt ohne Berücksichtigung der LD-Strukturen (Affymetrix), oder basierend auf LD-Strukturen anhand einer Selektionsmethode für tagSNPs (Illumina). Ein tagSNP ist ein SNP, der mit einem anderem SNP im fast perfekten LD steht und ihn somit näherungsweise abbildet. In der Regel verlangt man mindestens ein r^2 von $0,8$ für einen guten tagSNP.
Illumina deckt auf ihrem neusten Chip (HumanOmni2.5-Quad mit ca. $2,5$ Millionen Markern) die verschiedenen HapMap-Phasen und die Daten des „1000 Genomes Projects“ ab. Dabei sind neben den SNPs, auch CNVs und funktionelle Regionen zu finden. Im Gegensatz dazu verwendet Affymetrix neben den HapMap-Daten auch SNPs aus biologischen Datenbanken und publizierten Studien. Der aktuelle Chip von Affymetrix ist der Genome-wide Human SNP Array 6.0, welcher $1,8$ Millionen SNPs und CNVs mit einer durchschnittlichen Distanz von 700 Basen umfasst.

Kapitel 2

Fragestellung und Motivation

2.1 Geschichtlicher Hintergrund und Stand der Forschung

Für die Genetische Epidemiologie ist die Suche nach genetischen Faktoren, die komplexe Merkmale und Charakterisierungen von Effekten beeinflussen, Ziel und Herausforderung zugleich. Die klassische Strategie ist zweistufig: Zuerst erfolgt die Kopplungsanalyse und darauf aufbauend die Assoziationsanalyse. Bei der Kopplungsanalyse werden Regionen lokalisiert und identifiziert, die im Zusammenhang mit der Krankheit stehen könnten. Diese Kopplungsregionen werden dann für die Assoziationsanalyse verwendet, bei welcher nur SNPs dieser Region ausgewertet werden.

Aufgrund der vielen und häufigen genetischen Varianten, die mit geringer Effektstärke zum genetischen Gesamteffekt beitragen hat die Kopplungsanalyse oft nur geringe Power. Deshalb kam 1996 die Idee der genomweiten Assoziationsanalysen auf, die von Risch et al. als neues Paradigma der Genetischen Epidemiologie vorgeschlagen wurden [Risch and Merikangas, 1996]. Diese Studien werden mit großen Stichproben und vielen, repräsentativ ausgewählten SNPs durchgeführt und verzichten auf die einleitende Kopplungsanalyse. Bis heute wurden die Ergebnisse aus über 769 GWAS-Studien (Stand 03.02.2011) publiziert [Hindorff et al., 2011]. Für fast alle komplexen Krankheiten wurden genetische Risikovarianten gefunden. Hunderte Gene und Genomregionen konnten identifiziert werden [Maher, 2008; Manolio et al., 2009]. Da die Effektgröße jedoch meist sehr klein ist, bleibt ein großer Teil des genetischen Beitrags zum Phänotypen der komplexen Krankheiten unerklärt [Maher, 2008]. Maher bezeichnet dies als den „Case of the missing heritability“. Er nennt sieben mögliche Erklärungen für die „fehlende“ Heritabilität: unzureichende Abdeckung der SNP-Chips, Krankheitsmodelle mit vielen häufigen Varianten mit jeweils sehr kleinen Effekten, seltene Varianten, strukturelle Variationen (Deletionen, Duplikationen, etc.), Interaktionen, Epigenetik (vererbte Entscheidungsmuster) sowie auch die Möglichkeit einer Überschätzung der Heritabilität. Da die Einzelmarkeranalyse nicht ausreicht um die Lücken der fehlenden Heritabilität zu schließen, ist ein möglicher nächster Schritt die Entwicklung von Multimarkerverfahren. Multimarkerverfahren betrachten mehrere SNPs simultan. Ansätze von diesen Verfahren sind genomweite Haplotypanalyse (Genome-wide Haplotype Analysis, GWHA), Pathwayassoziationsanalysen (Pathway Association Analysis, PAA) und genomweite Interaktionsanalysen (Genome-wide Interaction

Analysis, GWIA).
Die Haplotypanalyse zeichnet sich dadurch aus, dass die Möglichkeit gegeben wird, nicht genotypisierte SNPs über LD besser repräsentieren zu können als Einzel-SNPs. Außerdem kann dies einen Powergewinn gegenüber der Einzelmarkeranalyse ermöglichen [Trégouët et al., 2009; Becker and Herold, 2009].
Auch die Pathwayassoziationsanalyse ist vielversprechend, da komplexe Krankheiten von hunderten oder tausenden SNPs mit sehr kleinen individuellen Effekten verursacht werden können. In solchen Situationen ist es fast unmöglich, alle relevanten SNPs mit der Einzelmarkeranalyse zu finden. Die Systembiologie könnte mit ihren Datenbanken wie KEGG, Biocarta oder Gene Ontology dabei helfen. Die Idee von Pathwayassoziationsanalysen ist es, durch eine überproportionale Häufung unkorrigiert (moderat) signifikanter SNPs in einem Pathway die Assoziation eines Pathways mit einer Krankheit nachzuweisen. Zur Zeit gibt es drei verschiedene Ansätze, die bereits in verschiedenen Programmen implementiert wurden (GenGen [Wang et al., 2007], SNP Ratio [O'Dushlaine et al., 2009], ALIGATOR [Holmans et al., 2009]). Alle diese Ansätze definieren einen Pathway als eine Menge von Genen. Ihre Pathwayassoziationstests lassen sich in fünf Kategorien unterteilen: Einzelmarkertest, Gen-Bewertungsfunktion, Pathway-Bewertungsfunktion, Prozedur für einen Pathway-Signifikanztest und eine Prozedur für die Korrektur des multiplen Testens.
Neben Haplotyp- und Pathwayanalyse könnte auch die mögliche Existenz von Interaktion ein Grund für den geringen Erfolg der GWAS bei komplexen Krankheiten sein. Die Analyse dieses Phänomens bildet den Schwerpunkt der vorliegenden Arbeit. Da viele DNA-Veränderungen eher schädlich als nützlich sind, ist die negative Selektion („purifying selection“) wichtig, um langfristig die Stabilität der biologischen Strukturen aufrechtzuerhalten. Bei der negativen Selektion werden schädliche Mutationen entfernt, um sicherzustellen, dass sich schädliche Mutationen nicht weiter in der Bevölkerung verbreiten und Verbesserungen in der Struktur so lange wie möglich in der Bevölkerung bewahrt werden. Auch kurzzeitige negative Selektion ist weit verbreitet, besonders bei ökologischen Ursachen. Viele genetische Faktoren funktionieren in erster Linie anhand eines komplexen Mechanismus, in welchem viele verschiedene Gene und weitere Faktoren involviert sind. Das bedeutet, dass Varianten mit starkem Krankheitseffekt in der Regel durch die negative Selektion aussortiert werden, welche wiederum die meist schwachen Effekte bei der Einzelmarkeranalyse erklärt. Bei interagierenden Genen kann es jedoch zu stärkeren Effekten kommen. Allele werden nur dann aussortiert, wenn sie in der krankheitsverursachenden Interaktion auftreten, ansonsten werden die Allele „normal“ vererbt. Letzten Endes ist der Selektionseffekt demnach schwächer und das Entstehen sowie die Erhaltung von Krankheitsvarianten mit Interaktionseffekten, die stärker sind als die üblichen marginalen Effekte, ist plausibel.
Ein weiterer Motivationsgrund für die Suche nach interagierenden Genen sind die gewonnenen Erkenntnisse über die biologischen und biochemischen Pathways der Krankheit. Eine wichtige Frage in biologischen Studien ist außerdem, ob es Faktoren gibt, die Interaktionseffekte ohne marginale Effekte zeigen. Wenn es diese gibt, würde man sie bei einer Einzelmarkeranalyse nicht aufdecken, wenn sie nicht schon vorher zu einer marginalen Korrelation zwischen Genotyp und Phänotyp geführt haben, wenn jeder Locus einzeln betrachtet wird.
Weiteren Aufschluss über die fehlende Heritabilität könnten auch bald die Analysen des „Next-Generation-Sequencing“ (NGS) liefern. In diesem Zusammenhang

wurde im Januar 2008 das „1000 Genomes Project“ [1000 Genomes Project Consortium et al., 2010] ins Leben gerufen, welches sich zum Ziel gesetzt hat, 1.000 menschliche Genome zu sequenzieren und die Daten in einer Datenbank zu veröffentlichen. Somit soll ein detaillierter Katalog von Genvarianten im menschlichen Genom aufgebaut werden. Um den Wettbewerb bei der Entwicklung von kostengünstigeren NGS-Methoden voranzutreiben, hat die US-amerikanische X-Prize Stiftung zehn Millionen Dollar dem Team in Aussicht gestellt, welches es schafft, zehn menschliche Genome in zehn Tagen für nicht mehr als 100.000 Dollar zu sequenzieren (`http://genomics.xprize.org`). Auch wenn die Sequenzierung in den kommenden Jahren kostengünstiger werden wird, stellt das NGS die Bioinformatik aufgrund der großen Datenmengen und benötigten Rechenkapazitäten vor neue Herausforderungen.

2.2 Fragestellung und Motivation

In den vergangenen Jahren wurde das Gebiet der Genetischen Epidemiologie durch den Erfolg von GWAS revolutioniert. Die meisten dieser Studien haben eine Einzelmarkeranalyse-Strategie verfolgt, in welcher jede Variante einzeln auf Assoziation mit einem spezifischen Phänotyp getestet wurde. Dies hat zur Identifikation von hunderten Regionen im Genom geführt, die mit einer komplexen Krankheit assoziiert sind. Trotzdem ist ein großer Teil ihrer Heritabilität unerklärt. Deshalb müssen Interaktionen zwischen genetischen Varianten als eine mögliche Erklärung für „Case of missing heribility“ [Maher, 2008] in Betracht gezogen werden.
Das grundsätzliche Problem der genomweiten Interaktionsanalyse ist jedoch die große Anzahl der auszuführenden Tests. Für einen SNP-Chip mit einer Million SNPs sind $0,5 \cdot 10^{12}$ SNP-Paare zu testen. Bei einer mittelgroßen Fall-Kontroll-Studie mit 1.500 Personen wurde die dazu nötige Rechenzeit auf einem leistungsfähigen 3GHz Linux-Rechner auf mehr als sieben Monate hochgerechnet [Herold et al., 2009]. Durch massive Parallelisierung kann die Rechenzeit reduziert werden, was Steffens et al. [2010] in Kooperation mit dem Institut für Numerische Simulation in Bonn gezeigt haben. Sie haben eine genomweite Interaktionsanalyse aller SNP-Paare unter Einsatz von 256 CPUs in sieben Stunden durchgeführt. Besteht jedoch kein Zugang zu Parallelrechnern, ist die Durchführung einer kompletten GWIA praktisch unmöglich. Weiterhin ist die Durchführung einer genomweiten Analyse aller SNP 3er-Kombinationen ($0,16 \cdot 10^{18}$ Tests) selbst bei Parallelisierung utopisch.
Um diese Rechenprobleme zu lösen, wurde im Rahmen meiner Doktorarbeit die Idee entwickelt, die Anzahl der Tests zu reduzieren, indem nur „interessante“ Markerkombinationen berechnet werden. Anhand von a-priori Information werden zunächst nur bestimmte SNPs für die Multimarkeranalyse selektiert. Grundlagen für diese Informationen können statistische (Einzelmarkerergebnisse) oder genetische/biologische (Genlokation, Funktionsklasse oder Pathwayinformation) Kriterien sein. Diese Herangehensweise reduziert gleichzeitig die Anzahl der Tests bei der Korrektur des multiplen Testens und kann zu besserer Power führen.
Im weitesten Sinne soll dieser Ansatz zur Schließung weiterer Lücken der fehlenden Heritabilität beitragen. Das Verstehen von genetischen Variationen könnte dann zur besseren Vorbeugung, Diagnose und Behandlung von Krankheiten führen. Die Umsetzung dieser Idee in meiner Software INTERSNP wird im Folgenden vorgestellt.

Kapitel 3

Genomweite Interaktionsanalyse mit INTERSNP

3.1 INTERSNP - Was ist das?

INTERSNP [Herold et al., 2009] ist eine in C/C++ geschriebene Software für genomweite Interaktionsanalyse (GWIA) von Fall-Kontroll-Studien. Die Idee von INTERSNP ist es, SNPs anhand von a-priori Information vor der Multimarkeranalyse zu selektieren, um die Anzahl der auszuführenden Tests zu reduzieren. Statistische (Einzelmarkerergebnisse) und/oder genetische/biologische (Genlokation, Funktionsklasse oder Pathwayinformation) Kriterien werden genutzt um die „interessanten“ Multimarkerkombinationen auszuwählen. Für Multimarkeranalysen mit mehreren SNPs wurden verschiedene statistische Verfahren in INTERSNP implementiert. Einerseits handelt es sich um ein log-lineares Modell und die logistische Regression für Fall-Kontroll-Datensätze, anderseits um die lineare Regression für quantitative Zielgrößen (Traits). Die beiden Regressionsmodelle ermöglichen die Verwendung von Kovariaten und die Formulierung einer Vielzahl von Interaktionstests. Um die Signifikanz der Analyseergebnisse nach Korrektur für multiples Testen überprüfen zu können, sind in INTERSNP zusätzlich Monte-Carlo-Simulationen implementiert, die auf genomweiter Permutation des Fall-Kontroll-Status basieren. Auch existiert eine parallelisierte Version der Software, um die Rechenzeit bei genomweiten Analysen zusätzlich zu reduzieren oder eine genomweite Analyse ohne Einschränkungen zu ermöglichen. Im Folgenden werden die Qualitätskontrolle, die statistischen Methoden, die Ein- und Ausgabedateien sowie die Verwendung von INTERSNP detailliert beschrieben.

3.2 Qualitätskontrolle

Allgemein werden unter dem Begriff Qualitätskontrolle (QC, engl. Quality Control) unterschiedliche Ansätze und Maßnahmen zusammengefasst, mit denen festgelegte Qualitätsanforderungen gewährleistet werden sollen. In unserem Fall bezieht sich die Qualitätskontrolle auf SNP-Daten und dient dazu, die Anzahl der durch Artefakte erzeugten falsch-positiven Ergebnisse möglichst gering zu halten. Ein häufiger Grund für scheinbare Assoziationen sind Genotypisierungsfehler, die dann

entstehen, wenn der in der molekulargenetischen Analyse ermittelte Genotyp einer Person nicht dem tatsächlichen Genotyp entspricht [Kiewert, 2006]. Genotypisierungsfehler können in jedem Schritt der DNA-Analyse entweder durch menschliche oder technische Fehler oder durch die Qualität der DNA verursacht werden. Dazu gehören beispielsweise Fehler bei der Probenentnahme, Vertauschen von Proben, Pipettierungsfehler, kontaminiertes und unvollständige DNA-Material sowie Fehler bei der Datenübertragung [Bonin et al., 2004; Miller et al., 2002].

Treten fehlende Werte und Genotypisierungsfehler zufällig und unabhängig vom Fall-Kontroll-Status auf, so erhält man zwar reduzierte Power, aber eine Erhöhung der falsch positiven Assoziationen ist nicht zu erwarten. Werden Fälle und Kontrollen jedoch nicht unter gleichen Bedingungen genotypisiert, so dass etwa bei den Fällen gehäuft falsche Genotypen bestimmt werden, können auch falsch positive Ereignisse auftreten. Aus diesem Grund ist es wichtig, vor der Analyse der Daten eine gründliche Qualitätskontrolle durchzuführen, sowohl bezüglich der Arbeitsschritte im Labor wie auch statistisch bzw. bei der statistischen Auswertung. Die statistische Qualitätskontrolle überprüft die Kriterien „Missingrate“ pro Person und SNP, Abweichungen vom Hardy-Weinberg Equilibrium (HWE) bei Fällen und Kontrollen (siehe Abschnitt 1.2.2) und je nach Studie und Genotypisierungsplattform auch die Frequenz des selteneren Allels (Minor Allele Frequency, MAF). Bei der MAF ist zu beachten, dass bei seltenen Allelen die Zuverlässigkeit der Genotypisierung geringer ist und deshalb Allele mit einer MAF kleiner als typischerweise $0,001$ bei der Analyse ausgeschlossen werden. Diese statischen Qualitätskontrollen wurden in INTERSNP implementiert und werden im Abschnitt 3.5.4 näher erläutert.

3.3 Statistische Methoden

Im Folgenden werden die in INTERSNP implementierten statistischen Modelle erläutert, die eine simultane Analyse mehrerer Variablen ermöglichen und somit anhand von Interaktionstermen Abhängigkeitsstrukturen aufdecken können. Der Einfachheit halber werden hier in der Regel zwei SNPs betrachtet. Methoden für drei SNPs sind ebenfalls implementiert und werden an einigen Stellen erwähnt. Die im weiteren Verlauf verwendete $2 \times 3 \times 3$-Kontingenztafel (siehe Tabelle 3.1) setzt sich aus jeweils drei Genotypen (AA, AB, BB) für SNP1 und SNP2 zusammen sowie dem Fall-Kontroll-Status. Dabei sind A, B allgemeine Platzhalter für die beiden möglichen Allele eines SNPs.

(a) Fälle

SNP2 \ SNP1	AA	AB	BB
AA	0,003	0,016	0,003
AB	0,152	0,032	0,045
BB	0,423	0,281	0,045

(b) Kontrollen

SNP2 \ SNP1	AA	AB	BB
AA	0,017	0,003	0,003
AB	0,116	0,162	0,023
BB	0,367	0,263	0,046

Tabelle 3.1: $2 \times 3 \times 3$-Feldertafel für die Häufigkeiten der Fälle und Kontrollen.

3.3.1 Log-lineares Modell

Ein Weg, um Beziehungen zwischen statistischen Variablen zu erforschen ist die Verwendung eines log-linearen Modells (Abbildung 3.1). Nach Bishop et al. [2007] werden die beobachteten Daten x_{ijk} der 2×3×3-Kontingenztafel (siehe Tabelle 3.1) durch ein log-lineares Modell der erwarteten Zellhäufigkeiten m_{ijk} zu den Zelleinträgen angepasst, wobei $i = 1, 2, 3$ den Genotypen von SNP1 entspricht, $j = 1, 2, 3$ den Genotypen von SNP2 und $k = 1, 2$ Auskunft über den Fall-Kontroll-Status gibt. Die Gleichung für dieses Modell lautet wie folgt [Steffens et al., 2010]:

$$
\begin{aligned}
\log(m_{ijk}) = \; & \underbrace{u}_{\texttt{Gesamtmittelwert}} + \underbrace{u_{1(i)}}_{\texttt{Genotypfrequenz SNP1}} + \underbrace{u_{2(j)}}_{\texttt{Genotypfrequenz SNP2}} \\
& + \underbrace{u_{3(k)}}_{\texttt{Fall-Kontroll-Status}} + \underbrace{u_{12(ij)}}_{\texttt{genotypische Assoziation SNP1 - SNP2 (LD)}} \\
& + \underbrace{u_{13(ik)}}_{\texttt{marginale Effekte SNP1}} + \underbrace{u_{23(jk)}}_{\texttt{marginale Effekte SNP2}} \\
& + \underbrace{u_{123(ijk)}}_{\texttt{Interaktionseffekt SNP1-SNP2-Status}}
\end{aligned}
$$

Möchte man dieses log-lineare Modell als Test auf Interaktion verwenden, also die 3-Wege-Interaktion aus SNP1, SNP2 und Fall-Kontroll-Status, ergeben sich die beiden Hypothesen H_0 ($u_{123(ijk)} = 0$) und H_1 ($u_{123(ijk)} \neq 0$), die gegeneinander getestet werden.
Mit den Maximum-Likelihood-Schätzern x_{ijk} der Zellhäufigkeiten für die Zelleinträge, erhält man die folgende Teststatistik:

$$T = -2\left(\sum_{i,j,k} x_{ijk} \log \frac{\hat{m}_{ijk}}{x_{ijk}}\right),$$

welche χ^2-verteilt ist mit $(I\check{}1)(J\check{}1)(K\check{}1) = 4$ Freiheitsgraden. Basierend auf dem Startwert $\hat{m}_{ijk}^{(0)} = 1$ können die Maximum-Likelihood Schätzer $\hat{m}_{ijk}$ iterativ berechnet werden:

$$
\begin{aligned}
\hat{m}_{ijk}^{(1)} &= \hat{m}_{ijk}^{(0)} \cdot \frac{X_{ij+}}{\hat{m}_{ij+}^{(0)}}, \\
\hat{m}_{ijk}^{(2)} &= \hat{m}_{ijk}^{(1)} \cdot \frac{X_{i+k}}{\hat{m}_{i+k}^{(1)}}, \\
\hat{m}_{ijk}^{(3)} &= \hat{m}_{ijk}^{(2)} \cdot \frac{X_{+jk}}{\hat{m}_{+jk}^{(2)}}, \\
\hat{m}_{ijk}^{(4)} &= \hat{m}_{ijk}^{(3)} \cdot \frac{X_{ij+}}{\hat{m}_{ij+}^{(3)}} \\
&\dots
\end{aligned}
$$

Die Iteration operiert direkt auf der Genotyptafel und konvergiert in der Regel sehr schnell, meistens nach weniger als zehn Iterationen. Für flexiblere Modellierung,

Testen auf bzw. unter Interaktion, allelische und genotypische Tests sowie Berücksichtigung von Kovariaten muss jedoch, wie im nächsten Abschnitt beschrieben, die lineare oder logistische Regression verwendet werden.

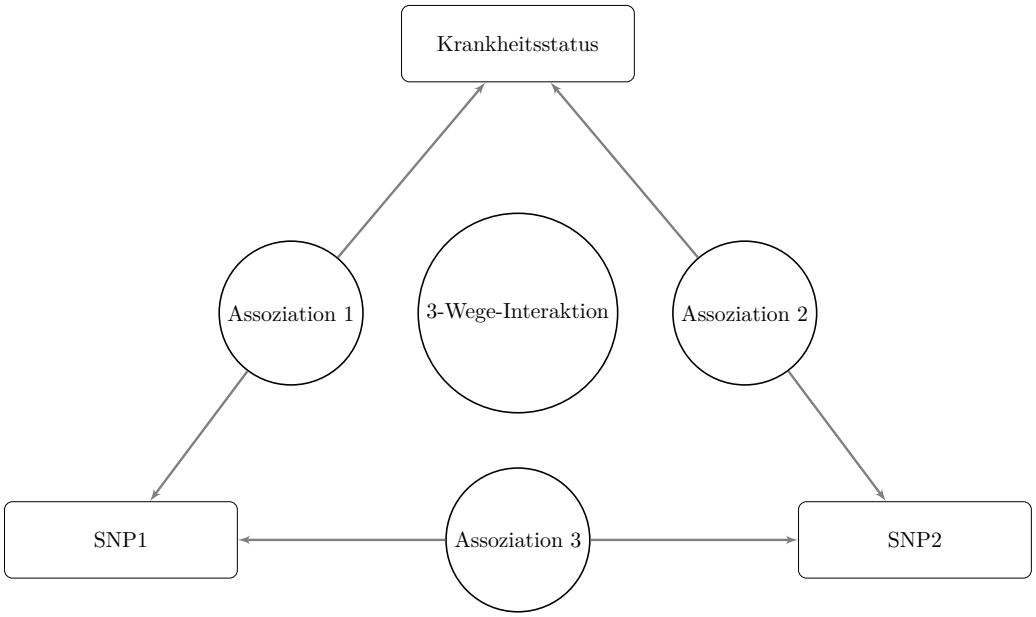

Abbildung 3.1: Log-lineares Modell: Assoziation 1 ist der marginale Effekt von SNP1 bezüglich des Krankheitsstatus. Assoziation 2 ist der marginale Effekt von SNP2 bezüglich des Krankheitsstatus. Assoziation 3 ist die allelische oder genotypische, Fall-Kontroll-Status unabhängige Assoziation zwischen SNP1 und SNP2 die z.B. durch LD verursacht wird [Steffens et al., 2010].

3.3.2 Regresssionsmodell

Regressionsmodelle beschreiben den Zusammenhang zweier oder mehrerer statistischer Variablen durch eine Gleichung. Diese wird durch Parameterschätzung den beobachteten Daten so angepasst, dass die Datenpunkte möglichst wenig von den unter dem Regressionsmodell erwarteten Daten abweichen. Der Phänotyp ist die Zielvariable y, alle weiteren Variablen bezeichnet man als Einflussvariablen x_i. Die Wahrscheinlichkeit, dass eine Person ein Fall und keine Kontrolle ist, sei mit p bezeichnet [Cordell and Clayton, 2002].
Bei der logistischen Regression betrachtet man

$$\text{logit}(p) := \ln\left(\frac{p}{1-p}\right) = \beta^T x,$$

wobei β der Vektor der geschätzten Koeffizienten ist und x der Vektor, in dem die Genotypen kodiert sind. Die logistische Regression, welche für Fall-Kontroll-Daten verwendet wird, wurde nach dem Modell von Cordell und Clayton implementiert. Bevor näher auf die Tests der logistischen Regression eingegangen wird, soll kurz die Verwendung der Funktion logit motiviert werden [Sachs and Hedderich, 2009]. Die Wahrscheinlichkeit p_i, dass ein bestimmter Phänotyp die Ausprägung y_i aufweist, liegt im Bereich von 0 bis 1. Das Ziel ist es, dieses Intervall durch Transformation in den Bereich $-\infty$ bis ∞ abzubilden. Berechnet man das Verhältnis der

Wahrscheinlichkeiten p_i und $1 - p_i$, also die Odds Ratios, erweitert sich der Zielbereich von 0 bis ∞. Logarithmiert man diese Odds Ratios, wird der gewünschte Bereich von $-\infty$ bis ∞ ermöglicht:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \ln\left(\frac{p(y=1)}{p(y=0)}\right) := \text{logit}(p)$$

Wird die Gleichung noch nach p_i aufgelöst, ergibt sich die logistische Regressionsgleichung:

$$p_i = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$

Im Kontext der Genetik ermöglicht das Modell der logistischen Regression marginale Effekte ein- und auszuschließen, allelische und genotypische Test zu unterscheiden sowie für Kovariaten zu adjustieren. Nehmen wir beispielhaft zwei SNPs. Für jeden SNP i mit $i = 1, 2$ stellen wir den allelischen Effekt x_i mit der Kodierung der Genotypen (1,1), (1,2) und (2,2) als $x_i = -1, 0, 1$ dar. Für Modelle mit einem Dominanzeffekt gilt somit, dass der Dominanzterm $x_{i,D}$ die Werte $-0,5$ und $0,5$ sowie $-0,5$ für die Genotypen (1,1), (1,2) und (2,2) annimmt. Der Dominanzparameter modelliert Abweichungen von multiplikativen Effekten des Allels und erkennt damit insbesondere auch rezessive Effekte. Wir erhalten beispielsweise x_1x_2 als Interaktionsterm, der die allelische Interaktion zwischen SNP1 und SNP2 repräsentiert, während $x_{1,D}x_{2,D}$ die Interaktion zwischen dem Dominanzterm von SNP1 und SNP2 darstellt.

Sei β_0 der Achsenabschnitt, der die Grundlinie der Likelihood $L_0 := \text{logit}(p) = \beta_0$ definiert. Somit stellt die Likelihood $L_1^A := \beta_0 + \beta_1 x_1$ den allelischen Effekt von SNP1 dar und der Vergleich mit L_0 führt zu einem Likelihood-Ratio-Test mit einem Freiheitsgrad. Analog entsteht aus dem Vergleich von $L_1^G = \beta_0 + \beta_1 x_1 + \beta_{1,D} x_{1,D}$ und L_0 ein Genotyptest für SNP1 mit zwei Freiheitsgraden. Im Allgemeinen bezeichnen wir mit L^A und L^G Likelihoods, die allelische Terme oder aber allelische und genotypische Terme für SNP1 und SNP2 enthalten. Zusätzlich verwenden wir Likelihoods wie $L_{1,2}^{A,I}$ und $L_{1,2}^{G,I}$, welche auch Interaktionsterme beinhalten, beispielsweise $\beta_0+\beta_1x_1+\beta_2x_2+\beta_{1,2}x_1x_2$ und $\beta_0+\beta_1x_1+\beta_{1,D}x_{1,D}+\beta_2x_2+\beta_{2,D}x_{2,D}+\beta_{1,2}x_1x_2 + \beta_{1,2,D}x_1x_{2,D} + \beta_{1,D,2}x_{1,D}x_2 + \beta_{1,D,2,D}x_{1,D}x_{2,D}$. Testen auf Interaktion entspricht also dem Testen, ob der Regressionskoeffizient zum Interaktionsterm Null ist. Bei einem allelischen Test mit einem Freiheitsgrad wäre die Nullhypothese somit $H_0 : \beta_{1,2} = 0$ und bei einem genotypischen Test mit vier Freiheitsgraden $H_0 : \beta_{1,2} = \beta_{1,2D} = \beta_{1D,2} = \beta_{1D,2D} = 0$. Die verschiedenen Likelihoods sind in der Tabelle 3.2 zusammengefasst.

Möchte man anstatt von Fall-Kontroll-Daten quantitative Traits (Zielgrößen) für die Analyse verwenden, bietet sich die lineare Regression an. Bezüglich der Kodierung der Genotypen sind logistische Regression und lineare Regression analog. Bei der linearen Regression betrachtet man jedoch die Gleichung $y = \beta^T x$, wobei y die Ausprägung des quantitativen Merkmals ist, und berechnet die Signifikanz über die Fehlerquadratsummen (SSE = Sum of Squared Errors). Die Fehler messen für jede Person i die Abweichung zwischen dem beobachteten Wert y_i und dem von einem Modell geschätzten $\hat{y}_i$. Details zur Beschreibung der Implementierung der logistischen/linearen Regression befinden sich im Anhang A.

	Likelihood
L_0	β_0
L_1^A	$\beta_0 + \beta_1 x_1$
L_1^G	$\beta_0 + \beta_1 x_1 + \beta_{1,D} x_{1,D}$
$L_{1,2}^A$	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
$L_{1,2}^G$	$\beta_0 + \beta_1 x_1 + \beta_{1,D} x_{1,D} + \beta_2 x_2 + \beta_{2,D} x_{2,D}$
$L_{1,2}^{A,I}$	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2$
$L_{1,2}^{G,I}$	$\beta_0 + \beta_1 x_1 + \beta_{1,D} x_{1,D} + \beta_2 x_2 + \beta_{2,D} x_{2,D} +$ $\beta_{1,2} x_1 x_2 + \beta_{1,2,D} x_1 x_{2,D} + \beta_{1,D,2} x_{1,D} x_2 + \beta_{1,D,2,D} x_{1,D} x_{2,D}$

Tabelle 3.2: Likelihoods, die in INTERSNP verwendet werden.

3.3.3 Adjustierung für Stratifikation

Im Fall-Kontroll-Design kann es durch Populationsstratifikation zu Verzerrung der Ergebnisse der statistischen Auswertung und somit zu falschen Assoziationsbefunden kommen. Populationsstratifikation entsteht, wenn die Stichprobe aus verschiedenen Subpopulationen („Schichten“) besteht, die erstens unterschiedliche Krankheitshäufigkeiten und zweitens am untersuchten Merkmal (SNP) unterschiedliche Merkmalshäufigkeiten (Allelfrequenzen) in den Schichten aufweisen. Letzteres kann aufgrund unterschiedlichen ethnischen Hintergrunds auch systematisch auftreten. Schichtung kann aber ebenso durch die Genotypisierung von Personen in verschiedenen Batches, also durch unterschiedliche Laborbedingungen für verschiedene Samples, entstehen. Wenn die Schichtung einer Stichprobe bekannt ist, kann man die Populationszugehörigkeit einer Person als Kovariate im Regressionsmodell verwenden, d.h. die Populationszugehörigkeit wird sowohl bei der Likelihood L1 als auch bei der Alternativ-Likelihood L2 mitmodelliert, z.B.: $L_{1,2}^{A,I}$ vs $L_{1,2}^{A}$: $\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_{1,2} x_1 x_2+\gamma_1 c_1+\gamma_2 c_2$ vs $\beta_0+\beta_1 x_1+\beta_2 x_2+\gamma_1 c_1+\gamma_2 c_2$, wobei c_1 und c_2 die Werte der Kovariaten pro Person beinhalten. Dadurch können Assoziationsergebnisse, die ausschließlich auf Stratifikation zurückzuführen sind, vermieden werden. In der Regel ist es allerdings so, dass die Subpopulationen unbekannt sind. In diesem Fall können mithilfe der Principal Component Analysis [Price et al., 2006] aus GWAS-Daten Populationsschichten bestimmt und die Wahrscheinlichkeit der Zugehörigkeit der Personen zu den Schichten geschätzt werden. So erhält man für jede Subpopulation i und jede Person j die Wahrscheinlichkeit $c_i(j)$, dass Person j zur Schicht i gehört. Die $c_i = c_i(j)$ können dann als zusätzliche Parameter ins Regressionsmodell aufgenommen werden, wodurch eine Adjustierung für die errechnete Stratifikation erzielt wird.

3.4 Implementierung von INTERSNP

3.4.1 Programmaufbau

Wie sich mithilfe von INTERSNP die Anzahl der Tests der genomweiten Interaktionsanalyse (GWIA) anhand von Prioritäten reduzieren lässt, wird schematisch in Abbildung 3.2 darstellt und im Folgenden genauer beschrieben. Angenommen, es steht ein genomweiter Datensatz mit 1 Million SNPs zur Verfügung, so wä-

ren das $5 \cdot 10^{11}$ Paare, die auf Interaktion getestet werden müssten. Auf einem Standard-Desktopcomputer würde die Berechnung Monate dauern. Aus diesem Grund werden in INTERSNP zwei Möglichkeiten angeboten, die Analyse zu beschleunigen. Zum einen steht eine parallelisierte Version zur Verfügung, die in Abschnitt 3.4.3 näher beschrieben wird, und zum anderen kann die Anzahl der Tests reduziert werden, was die Grundidee von INTERSNP ist. Um die zweite Möglichkeit zu realisieren, müssen Auswahlkriterien definiert werden, nach denen die Daten selektiert werden sollen. In INTERSNP dienen die Ergebnisse der Einzelmarkeranalyse, genetische Kriterien oder Pathwayinformation als Prioritäten. Genetische Informationen werden aus einer Datei mit Annotationen (Annotationfile) entnommen, Informationen zu den Pathways findet man in einer Datei mit Pathways (Pathwayfile). Mit der reduzierten Datenmenge kann die Multimarkeranalyse durchgeführt werden. Im Folgenden werden die Dateien nach dem Schlüsselwort in der Parameterauswahldatei (Selectionfile) von INTERSNP bezeichnet d.h. statt „Datei mit Pathways“ wird „Pathwayfile“ usw. verwendet.

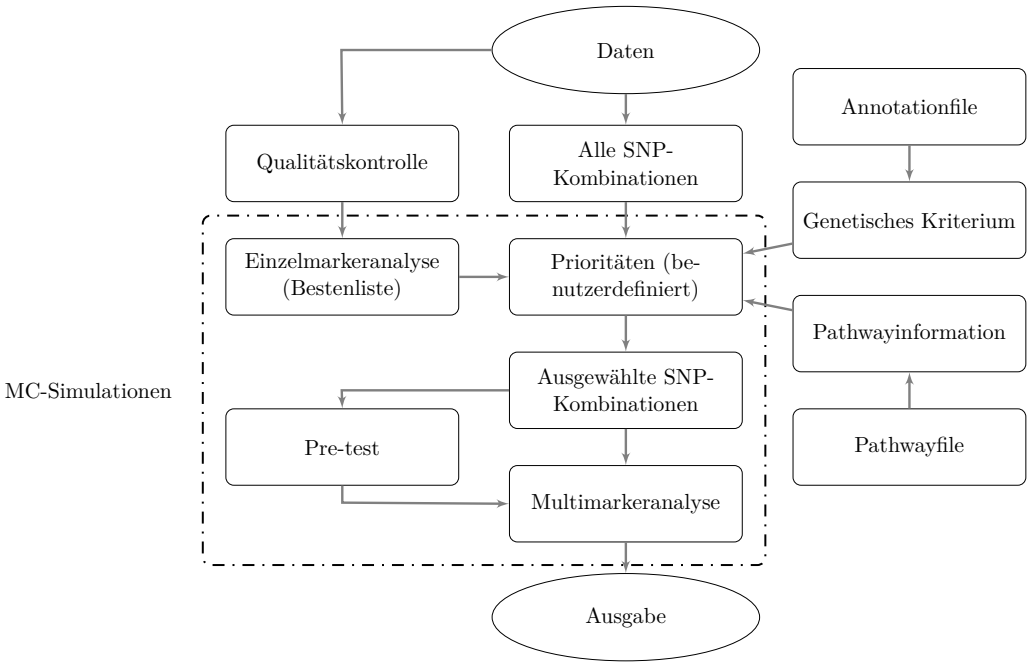

Abbildung 3.2: Programmablauf von INTERSNP.

Nach der Beschreibung der Grundidee von INTERSNP soll der Ablauf des Programms genauer beschrieben werden. Zunächst wird das Selectionfile, eine Auswahldatei für Programmparameter mit Schlüsselworten eingelesen, die unter anderem Auskunft darüber geben, welche Eingabedateien benötigt werden. So werden nur die relevanten Dateien eingelesen. Das ist deshalb sinnvoll, weil das Einlesen der Dateien relativ viel Rechenzeit kostet (jedoch meist < 20 min, auch bei großen Datensätzen). Die erste Reduzierung der Daten kann durch die Schlüsselwörter `POSCHOICE` und `NEGCHOICE` erfolgen. Hierbei können bestimmte Regionen (z.B. ein Chromosom) aus- oder eingeschlossen werden (siehe Abschnitt 3.5.6.4). Danach werden alle erforderlichen Daten eingelesen und die Informationen in Ma-

trizen gespeichert. Nach dem Einlesen folgt die Qualitätskontrolle, bei welcher sowohl Personen als auch SNPs, die den definierten Qualitätskriterien nicht genügen, gelöscht werden. Die Qualitätskontrolle (siehe Abschnitt 3.5.4) basiert auf einem iterativen Algorithmus, welcher SNPs und Personen löscht, deren Missingrate zu hoch ist. Zusätzlich wird auf Abweichungen vom HWE überprüft. Bevor es schließlich zur Auswahl der Kriterien kommt, wird für alle Daten, die die Qualitätskontrolle überstanden haben, eine Einzelmarkeranalyse durchgeführt. Danach werden die verschiedenen Kriterien zum Reduzieren der Daten eingelesen und es werden Schnittmengen der Daten gebildet, so dass nur noch die SNPs analysiert werden, die alle ausgewählten Kriterien erfüllen. Wird beispielsweise verlangt, dass die SNPs unter den besten 1.000 Einzelmarkerergebnissen und gleichzeitig in einer kodierenden Region sein sollen, so werden von der Menge der SNPs, die unter den besten 1.000 Einzelmarkerergebnissen liegen, nur diejenigen ausgewählt, die sich in einer kodierenden Region befinden. Folglich kann die Anzahl der Tests deutlich reduziert werden. Näheres zu den Kriterien und Strategien findet man unter Abschnitt 3.5.6. Der nächste Schritt ist die Analyse des qualitätskontrollierten und reduzierten Datensatzes. Bei der Multimarkeranalyse kann der Benutzer über das Selectionfile entscheiden, ob eine 2-Markeranalyse oder eine 3-Markeranalyse durchgeführt werden soll und welcher Test ausgewählt werden soll. Die Beschreibung der Tests ist in Abschnitt 3.5.5 zu finden. Werden für die Korrektur des multiplen Testens Monte-Carlo- (MC-) Simulationen verwendet, folgt nach der Analyse eine N-fache Wiederholung des kompletten Prozesses ab der Qualitätskontrolle, wobei N die vom Benutzer festgelegte Anzahl von Simulationen ist. Die Ergebnisse werden in verschiedene Ausgabedateien geschrieben, welche im Abschnitt 3.5.9 genauer erläutert werden. Namen und Anzahl der Ergebnisse können über das Selectionfile (siehe Abschnitt 3.5.2) benutzerspezifisch angepasst werden.

3.4.2 Hardware

Das in C/C++ geschriebene Programm INTERSNP kann sowohl auf Windows- als auch Unix-basierten Rechnern ausgeführt werden. Die entsprechenden vorkompilierten Versionen stehen auf der Homepage (`http://intersnp.meb.uni-bonn.de`) zur Verfügung. Um sehr große Datensätze zu analysieren ist ein großer Arbeitsspeicher erforderlich, da alle eingelesenen Dateien im Arbeitsspeicher bleiben, so dass Cluster-Systeme für die GWIA zu empfehlen sind.

3.4.3 Parallelisierung

Um die Rechenzeit der GWIA so gering wie möglich zu halten, wurde bereits die Möglichkeit vorgestellt die Anzahl der Tests anhand von a-priori Information zu reduzieren. Ein weiterer Ansatz, die Analyse zu beschleunigen ist die Parallelisierung des Programms, um die Ressourcen eines modernen Hochleistungsrechners besser zu nutzen. Für die Parallelisierung haben wir OpenMP (Open Multi-Processing, [OpenMP, 2008]) verwendet, da es mit diesem Standard relativ einfach ist aus einem bestehendem seriellen Programm eine parallelisierte Version zuerstellen. Der OpenMP Standard dient zur „Shared-Memory-Programmierung“ (gemeinsamer Hauptspeicher) in C/C++ /Fortran auf Multiprozessor-Computern. Die Parallelisierung findet hierbei auf Prozess-, bzw. Schleifenebene statt. Der OpenMP-Standard definiert dazu spezielle Compiler-Direktiven, die diesen anweisen, beispielsweise eine for-Schleife auf mehrere Prozesse und/oder Prozessoren zu

verteilen. Daneben existieren Bibliotheksfunktionen sowie Umgebungsvariablen, die zur OpenMP-Programmierung dienen. Ein großer Vorteil von OpenMP ist, dass im Allgemeinen so programmiert werden kann, dass die Programme auch korrekt laufen, wenn der Compiler die OpenMP-Anweisungen nicht kennt. Die Anweisungen werden in diesem Fall als Kommentare aufgefasst und übergangen. Eine mit OpenMP für mehrere Prozesse aufgeteilte *for*-Schleife kann auch mit einem einzelnen Prozess sequentiell abgearbeitet werden.
Da der vorhandene Programmcode nicht völlig neu geschrieben und umstrukturiert werden sollte, haben wir uns zunächst auf OpenMP beschränkt. In Abbildung 3.1 werden die Hauptschleifen n und a dargestellt, die im Programm parallelisiert wurden. In INTERSNP wurde die Parallelisierung folgendermaßen umgesetzt:

```
for (n=0; n<=nsim; n++) // Iteration über die Simulationen
{
    for (a=startA; a<=stopA; a++) // Iteration über SNP1
    {
        for (b=startB; b<=stopB; b++) // Iteration über SNP2
        {
            test(a,b); // Interaktionsfunktion für SNP1 und SNP2

            for (c=startC; b<=stopC; c++) // Iteration über SNP3
            {
                test(a,b,c); // Interaktionsfunktion für SNP1,
                             // SNP2 und SNP3
            }
        }
    }
}
```

Listing 3.1: Hauptschleifen in INTERSNP, wobei die beiden äußeren Schleifen parallelisiert sind.

In der n-Schleife werden die MC-Simulationen ausgeführt. Sie ist deshalb sinnvoll zu parallelisieren, da in jedem Schleifendurchlauf die komplette Analyse aller QC-SNPs mit randomisiertem Fall-Kontroll-Status (Affektionstatus) neu berechnet wird. Bei der Parallelisierung werden also die verschiedenen Simulationsdurchgänge auf die zur Verfügung stehenden Prozessoren verteilt. Angenommen, man hat 20 Prozessoren und 1.000 Simulationsdurchläufe, so werden 20 Durchläufe gleichzeitig gestartet und jeder Prozessor rechnet 50 Analysen. Die a-Schleife läuft durch die Liste der SNPs und setzt SNP1 fest. Mit der b-Schleife wird dann SNP2 bestimmt. Wird die a-Schleife parallelisiert, werden die einzelnen Tests auf die Prozessoren verteilt. Damit der Sourcecode nicht vom Benutzer geändert werden muss, gibt es zwei parallelisierte Versionen. Die Version intersnpN, bei welcher die n-Schleife für MC-Simulationen parallelisiert wurde und die Version intersnpA für die a-Schleife. Der Einsatz von intersnpA kann beispielsweise für eine komplette GWIA sinnvoll sein.

3.4.4 Datenbanken

Allgemein handelt es sich bei einer Datenbank um eine strukturierte Ansammlung inhaltlich zusammengehörender und miteinander in Verbindung stehender

Daten. In dieser Arbeit interessieren uns die biologischen Datenbanken, die Informationen über Nukleotid-Sequenzen von Genen, Aminosäuresequenzen von Proteinen, Funktion, Struktur und Lokalisation auf dem Chromosom, klinische Auswirkung von Veränderung sowie Ähnlichkeiten von biologischen Sequenzen beinhalten. Die meisten bieten einen einfachen Zugang zu Veröffentlichungen, Sammlungen von genetischen Krankheiten, DNA-, Protein- und Strukturdatenbanken sowie Bioinformatik-Tools über das Internet. Das Institut NCBI (National Center for Biotechnology Information) bietet beispielsweise mit über 40 verlinkten Datenbanken einen Zugang zu wichtigen DNA-, RNA- und Protein-Datenbanken. Über PubMed wird wissenschaftliche Literatur zur Verfügung gestellt.

Für die Pathway-Analyse in INTERSNP werden nur Pathway-Datenbanken bzw. daraus extrahierte Informationen benötigt. Pathway-Datenbanken sind meist webbasierte Dienste, die Zugang zu Informationen über Stoffwechselwege, biochemische Reaktionswege und an deren Einzelreaktionen beteiligte Enzyme und Substrate bieten. Sie sind mit Suchmaschinen ausgestattet und mit weiteren Quellen verknüpft. KEEG- [Kanehisa et al., 2006] und GO-Datenbank [Harris et al., 2004] bieten aber auch die Möglichkeit, Dateien herunterzuladen. In meiner Arbeit wird ausschließlich die KEEG-Datenbank verwendet, auch wenn die Möglichkeit besteht, andere Pathwayinformationen zu benutzen solange das Format eingehalten wird. Das KEGG-Projekt wurde im Mai 1995 im Rahmen des Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan gegründet. Alle Daten in der KEGG-Datenbank, die manuell aus Publikationen erstellt wurden und die dazugehörigen Softwaretools sind als Teil des Japanese GenomeNet verfügbar. KEGG besteht aus vier Hauptkomponenten:

- PATHWAY: Repräsentation von Funktionen höherer Ordnung im Netzwerk der interagierenden Moleküle
- GENES: Zusammenstellung von Genkatalogen für komplett- und teilsequenzierte Genome
- LIGAND: Zusammenstellung der chemischen Komponenten der Zelle, Enzymmolekülen und enzymatischen Reaktionen
- BRITE: Funktionelle Hierarchien und Ontologien

Für die Analyse mit INTERSNP ist nur die PATHWAY-Komponente von Relevanz. Die Pathwayinformationen unterteilen sich in verschiedene Kategorien wie Metabolismus (z.B. Glykolyse), Prozesse der genetischen Information (z.B. Transkription), Prozesse von Umweltinformationen (z.B. Signaltransduktion), verschiedene zelluläre Prozesse (z.B. Zellwachstum), menschliche Krankheiten (z.B. Krebs) und Medikamentenwirkung (z.B. Penicillin). Abbildung 3.3 ist eine Beispieldarstellung eines Pathways. Das Schaubild stellt relevante Gene und ihren Zusammenhang mit Morbus Parkinson dar. In der Menge der Gene befinden sich die für uns interessanten SNPs. Ein Pathway stellt also eine Menge von Genen dar, die wiederum eine Menge von SNPs beinhaltet.

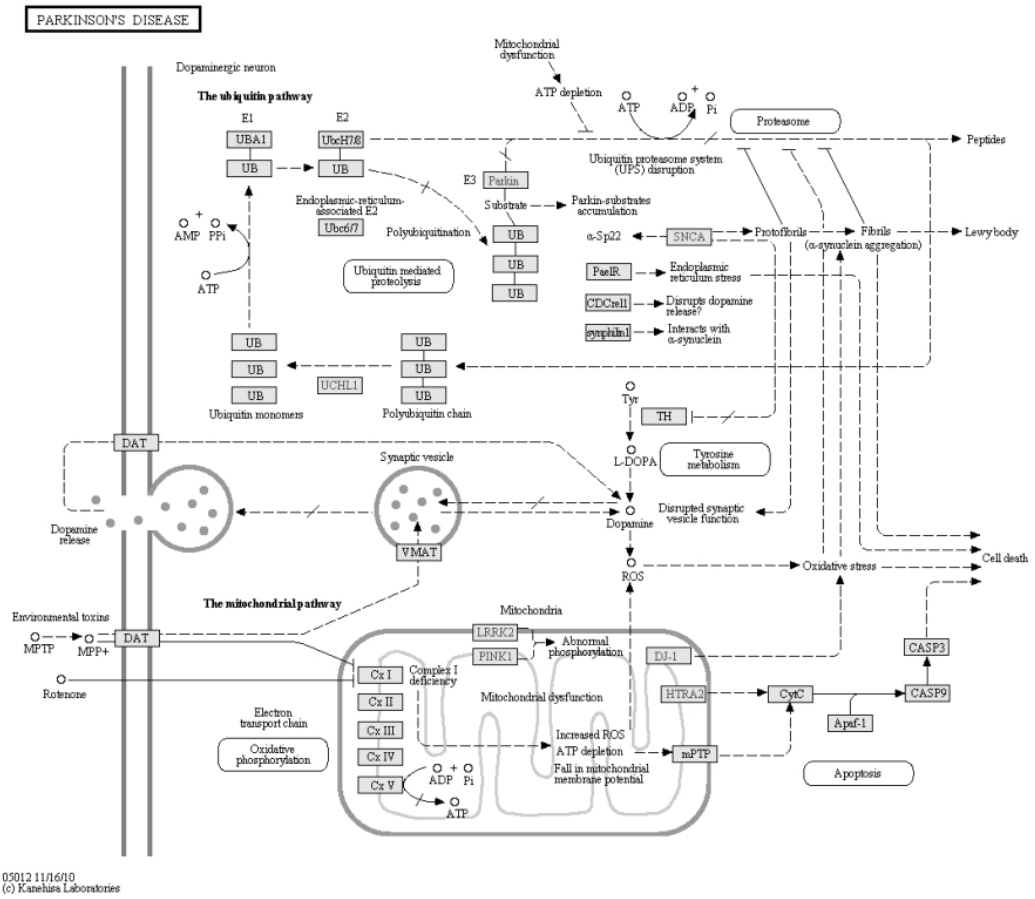

Abbildung 3.3: Quelle: KEGG-Datenbank: ein an Morbus Parkinson beteiligter Pathway.

3.5 Arbeiten mit INTERSNP

3.5.1 INTERSNP starten

INTERSNP kann sowohl auf Unix- als auch Windows-Rechnern von der Kommandozeile gestartet werden. Mit dem GNU-Compiler würde das Kompilieren und der Aufruf folgendermaßen aussehen:

Kompilieren: g++ intersnp.cpp -o intersnp -O3
Aufruf des Programm: ./intersnp selectionfile.txt

Natürlich kann auch jeder andere C++ -Compiler verwendet werden. Neben dem Aufruf des Programms muss zusätzlich nur noch der Pfad des Selectionfiles angegeben werden, in welchem alle wichtigen Parameter und Optionen für die Analyse mit INTERSNP gesetzt werden. Es müssen also keine weiteren Parameter auf der Kommandozeile, außer dem Selectionfile, übergeben werden. Somit ist der Umgang und Programmstart sehr einfach und übersichtlich.
Für die Software INTERSNP wurde unter `http://intersnp.meb.uni-bonn.de` eine Homepage eingerichtet auf welcher unter anderem die aktuellste Version von INTERSNP zu finden ist. Neben der Software, die als zip-Paket heruntergeladen werden kann, gibt es dort auch eine Dokumentation und Tipps zur Anwendung, insbesondere auch zur Kompilierung der parallelisierten Version.

3.5.2 Selectionfile

Mit dem Selectionfile werden die für den Programmaufruf notwendigen Parameter und Optionen konfiguriert. Die Eingabedateien, Analysen und zusätzliche Optionen können über das Selectionfile ausgewählt und spezifiziert werden. Die Datei besteht aus drei Spalten: Schlüsselwort (Keyword), Parameter und Kommentar (optional). Die Spalten können entweder durch Leerzeichen oder Tabs getrennt werden. Es ist sinnvoll, das Selectionfile, welches auf der Homepage zur Verfügung gestellt wird, zu verwenden und nur die Parameter entsprechend zu modifizieren. Die folgende Datei enthält alle verwendeten Schlüsselworte.

```
Keyword Parameter   Comment
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
ANNOTATIONFILE ./annotation.txt // Pfad zum Annotation-File.
PATHWAYFILE ./KEGG_2_snp_b129.txt.ann // Pfad zum Pathway-File.
COVARIATEFILE // Pfad zum Covariate-File.
MODELFILE  // Pfad zum Model-File.
SNPLIST 0 // Analyse der SNPs der SNP-Liste.
SNPFILE  // Pfad zum SNP-File.
COMBILIST 0 // Analyse der Paare/Tripel der Combi-Liste.
COMBIFILE  // Pfad zum Combi-File.
ONLY_MALE 0 // 1=nur Männer werden analysiert.
ONLY_FEMALE 0 // 1=nur Frauen werden analysiert.
POSCHOICE chr4; // Auswahl bestimmter Regionen, hier Chromosom 4.
NEGCHOICE  // Ausschluss bestimmter Regionen.
HWE_P_CASE 0.001 // QC-Grenzwert für HWE in Fällen.
HWE_P_CONTROL 0.01 // QC-Grenzwert für HWE in Kontrollen.
MRDIFF 1 // QC-Grenzwert für die Missingrate (MR): Personen und SNPs
            schlechter als die durchschnittliche MR + MRDIFF werden
            ausgeschlossen.
MAF 0 // QC-Grenzwert für MAF.
QT 0 // Für quantitative Analysen muss QT auf 1 gesetzt werden.
MISSING_PHENO -99 // Festlegen eines Wertes für fehlende Phänotypen
                     in den quantitativen Datensätzen.
SINGLE_MARKER 1 // 1=Armitage-Trendtest, 2=genotypischer Test
                mit 2 FG, 3=logistische Regression mit 1 FG,
                4=logistische Regression mit 2 FG.
PRETEST 0 // Pre-test 1=ja, 0=nein.
PRETEST_CUTOFF  // Pre-test Cutoff.
TWO_MARKER 1 // 1=2-Markeranalyse ist ausgewählt.
THREE_MARKER 0 // 1=3-Markeranalyse ist ausgewählt.
TEST 1 // 1=Chi^2-Test, 2=log-lineares Modell, 3-12=logistische
            Regression, M=benutzerdefiniertes logistisches
            Regressionsmodell.
COVARIATES 2-4; // Auswahl bestimmter Kovariaten, hier die 2te-4te.
SEXCOV 0 // 1=Geschlecht als Kovariate.
SINGLETOP 500 // Länge der Bestenliste der Einzelmarker-p-Werte.
M_WITH_SINGLETOP 2 // Anzahl der SNPs (0,1,2,3), die aus der
                     Einzelmarker-Bestenliste stammen sollen.
GENETIC_IMPACT 1 // Genetisches Kriterium: 0=Genwüste,
                     1=Gen-LD-Bereich, 2=Exon,3=in einer kodierenden
                     Region, 4=nicht-synonym.
M_WITH_GENETIC_IMPACT 1 // Anzahl der SNPs (0,1,2,3), die das
                          genetische Kriterium enthalten sollen.
SNP1  // rs-Nummer für den ersten festen SNP.
SNP2  // rs-Nummer für den zweiten festen SNP.
SNP3  // rs-Nummer für den dritten festen SNP.
```

```
PATHWAY 1 // 1=Pathway-Information wird berücksichtigt,
            0=ohne Pathway-Information.
SIMULATION 0 // Anzahl der MC-Simulationen.
MC_WITH_SM      0 // 1=MC-Simulationen bei der Multimarkeranalyse
                     UND der Einzelmarkeranalyse. 0=MC-Simulationen
                     nur bei der Multimarkeranalyse.
PRINTTOP 100 //Die n besten Multimarker-p-Werte werden in das
                Ausgabedatei geschrieben.
ANNOTATE  // Ausgabedatei mit Annotationsinformation.
GENECOL  // Angabe der Spalte mit der Annotationsinformation.
OUTPUTNAME ./meinProjekt/test // Name der Ausgabedateien mit Pfad.
END
```

Listing 3.2: Diese Datei stellt ein komplettes Selectionfile dar. In der ersten Spalte das Schlüsselwort, gefolgt von dem Parameter und einem Kommentar, der das Schlüsselwort näher erläutert.

Um den Aufbau des Selectionfiles darzustellen, wird ein kurzes Beispiel (siehe Listing 3.3) im Detail erläutert. Diese Datei enthält nicht alle möglichen Optionen. `TPED` ist der Pfad zu der tped-Eingabedatei (SNP-Daten, siehe Abschnitt 3.5.3) und `TFAM` der Pfad zu der tfam-Eingabedatei (Personeninformation, siehe Abschnitt 3.5.3). In diesem Beispiel wird zusätzlich ein Annotationfile (`ANNOTATIONFILE`) benutzt um das genetische Kriterium anwenden zu können. Nachdem die Eingabedateien definiert sind, folgen Qualitätskriterien, um Individuen und SNPs, die schlechter als ein angegebener Grenzwert sind, auszuschließen (`HWE_P_CASE`, `HWE_P_CONTROL`, `MRDIFF`, `MAF`). Jede Analyse startet mit einer Einzelmarkeranalyse. Der Test kann vom Benutzer ausgewählt werden (`SINGLEMARKER`). In unserem Beispiel wird der Armitage-Trendtest (`SINGLEMARKER = 1`, siehe Abschnitt 3.5.5.1) verwendet. Darüber hinaus soll eine 2-Marker-Analyse durchgeführt werden. Aus diesem Grund wird `TWO_MARKER` auf 1 gesetzt und `TEST 5` gewählt, also einen Test auf additive Interaktion (siehe Tabelle 3.3). Um die Anzahl der Tests zu reduzieren, wird das statistische und das genetische Kriterium verwendet. `SINGLETOP 50000` und `M_WITH_SINGLETOP 2` bedeutet, dass nur diejenigen SNP-Paare analysiert werden, die unter den 50.000 besten Einzelmarkerergebnissen liegen. Zusätzlich sollen sich diese Paare in einer kodierenden Region befinden (`GENETIC_IMPACT 3, M_WITH_GENETIC_IMPACT 2`). Zum Schluss wird der Pfad und der Name (`OUTPUTNAME`) für die Ausgabedateien angegeben.

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
ANNOTATIONFILE ./annotation.txt // Pfad zum Annotation-File.
HWE_P_CASE 0.001 // QC-Grenzwert für das HWE in Fällen
HWE_P_CONTROL 0.01 // QC-Grenzwert für das HWE in Kontrollen
MRDIFF 1 // QC-Grenzwert für die Missingrate (MR): Personen und
            SNPs schlechter als die durchschnittliche MR + MRDIFF
            werden gelöscht.
MAF  0 //   QC-Grenzwert für MAF
SINGLE_MARKER 1 //  Armitage-Trendtest
TWO_MARKER 1 // 2-Marker-Analyse
TEST 5 // Logistisches Regressionsmodell: Test auf additive
        Interaktion
SINGLETOP 50000 // Die besten 50.000 Einzelmarkerergebnisse
M_WITH_SINGLETOP 2 // Alle Paare befinden sich unter den
                      50.000 Einzelmarkerergebnissen.
GENETIC_IMPACT 3 // Genetisches Kriterium: in einer kodierenden
                    Region
M_WITH_GENETIC_IMPACT 2 // Alle Paare liegen in einer kodierenden
```

```
                          Region von einem Gen.
OUTPUTNAME ./meinProjekt/test // Name der Ausgabedateien mit Pfad
END
```

Listing 3.3: Dieses Selectionfile dient als Beispiel und stellt nur die für Strategie III notwendigen Optionen dar, die im Text ausführlich beschrieben werden.

Im Folgenden werden zunächst die Formate der Eingabedateien beschrieben und dann auf die Optionen und Parameter eingegangen. Dabei wird jeweils das Schlüsselwort des Selectionfiles näher erläutert.

3.5.3 Eingabedateien

In diesem Abschnitt sollen die verschiedenen Eingabedateien vorgestellt und wichtige Informationen zum Format erläutert werden. Beispiele sind auf unserer Homepage (`http://intersnp.meb.uni-bonn.de`) zu finden. Alle Dateien werden über das Selectionfile (siehe Abschnitt 3.5.2) ausgewählt.

3.5.3.1 tped/tfam

Das Datenformat der Eingabedateien ist das transponierte PLINK-Format mit tped- und tfam-Files [Purcell et al., 2007]. Hierbei handelt es sich um das Standardformat von PLINK, welches vorteilhaft ist, wenn viel mehr SNPs als Personen in einem Datensatz vorhanden sind, was bei GWAS fast immer der Fall ist. Die Dateien werden nämlich dann nur länger, aber gehen nicht in die Breite. Insbesondere ist der Zugriff auf bestimmte SNPs als einfacher Zeilenzugriff realisierbar. Das tped-File (`TPED`) enthält SNP- und Genotypinformationen. Die Genotypen werden in zwei Spalten, eine pro Allel, dargestellt. Die ersten vier Spalten sind für Chromosom, rs-Nummer, genetische Distanz und Basenpaarposition reserviert. In jeder Zeile wird ein SNP repräsentiert. Um die Laufzeit zu verbessern ist es sinnvoll, die Datei vor der Analyse nach Chromosom und Position zu sortieren.

```
16  rs8466895   0   37354 T T T T ... C C C T
16  rs216590    0   41263 G A A A ... A A G A
16  rs216596    0   45320 G G G G ... G G A G
16  rs2541594   0   45444 T T C T ... C C C C
16  rs8466998   0   49427 T T T T ... C C C T
16  rs216590    0   52259 G A A A ... A A G A
...
```

Listing 3.4: Ausschnitt aus einem tped-File: In der ersten Spalte steht das Chromosom gefolgt von rs-Nummer, genetischer Distanz, Basenpaarposition und jeweils zwei Spalten für die Genotypen der Personen.

Das tfam-File (`TFAM`) beinhaltet die Familieninformationen. Hier stehen in jeder Zeile Informationen zu einer Person. Die Spalten definieren Familien-ID, Personen-ID, Vater, Mutter, Geschlecht und Affektionstatus. Sind Mutter und Vater vorhanden steht in den jeweiligen Spalten ihre Personen-ID, ansonsten steht eine 0. Das Geschlecht unterteilt sich in männlich (1), weiblich (2) oder unbekannt (jede andere Zahl). Der Affektionstatus spiegelt den Fall-Kontroll-Status wieder, wobei „1“ für eine Kontrolle steht und „2“ für einen Fall. Dies sind im Prinzip die ersten sechs Spalten aus einem ped-File des ped/map-Formats [Purcell et al., 2007].

```
co1 co1 0 0 1 1
co2 co2 0 0 2 1
...
ca1 ca1 0 0 2 2
ca2 ca2 0 0 1 2
...
```

Listing 3.5: Ausschnitt aus einem tfam-File: In der ersten Spalte befindet sich die Familien-ID gefolgt von Personen-ID, Vater-ID, Mutter-ID, Geschlecht und Affektionstatus.

Person co1 ist somit eine männliche Kontrolle (siehe Listing 3.5) und hat den Genotyp (T,T) für SNP rs8466895 auf Chromosom 16 (siehe Listing 3.4). Person ca1 wäre im Gegensatz dazu ein weiblicher Fall (siehe Listing 3.5) der für SNP rs8466895 den Genotyp (C,C) hat (siehe Listing 3.4).

3.5.3.2 Annotationfile

Um das genetische Kriterium anwenden zu können und die Annotationsinformationen in der Einzelmarkerdatei einzubinden, wird ein Annotationfile (`ANNOTATIONFILE`) benötigt. Hierbei wird das Semikolon-getrennte Format des Illumina Human-610-chip Annotationfile verwendet. Die ersten Spalten geben Auskunft über rs-Nummer, Chromosom, Basenpaarposition und Genome-Build-Nummer. Weitere Spalten geben detaillierte Informationen über die jeweiligen SNPs. Es können auch hier eigene Annotationfiles verwendet werden, wobei es sinnvoll ist, die Datei nach Chromosom und Position zu sortieren, um die Laufzeit zu reduzieren.

```
name;chr;coordinate;genome_build;gene_symbol;gene_id;accession;
location;location_relative_to_gene;coding_status;
amino_acid_change;id_with_mouse;phast_conservation
rs12354060;1;10004;36.2;LOC653635;653635;XR_017611.1;intron;
-1762;NULL;NULL;NULL;NULL
rs2691310;1;46844;36.2;LOC642894;642894;XR_016145.1;
flanking_5UTR;-672;NULL;NULL;NULL;NULL
rs2531266;1;59415;36.2;OR4F5;79501;NM_001005484.1;coding;
[461/456];SYNON;A154A (NP_001005484.1);0.64;0.979
rs4124251;1;97215;36.2;LOC727901;727901;XR_015157.1;3UTR;
[3300/491];NULL;NULL;NULL;NULL
rs8179466;1;224176;36.2;LOC728481;728481;XR_015292.1;intron;
-42;NULL;NULL;NULL;NULL
...
```

Listing 3.6: Auszug aus einem Annotationfile.

Für die Zuordnung der SNPs zu den verschiedenen Kategorien (`GENETIC_IMPACT`) sind die folgenden Spalten aus dem Annotationfile (Ausschnitt aus der Originaldatei) wichtig:

- 1. Spalte: rs_Number (rs-Nummer)
- 2. Spalte: chromosome (Chromosom)
- 5. Spalte: gene_id (Genname)
- 8. Spalte: gene_location: coding, intron, 3UTR, 5UTR, UTR, flanking_3UTR, flanking_5UTR (Genlokalisationen)

- 9. Spalte: location_relative_to_gene: numbers below zero: distance to nearest gene, location within gene (detailliertere Genlokalisationsinformation)
- 10. Spalte: SNP coding_status: -1, NULL, SYNON, COMPLEX, NONSYNON (Kodierungsstatus des SNPs)

3.5.3.3 Pathwayfile

Damit INTERSNP die Pathwayinformationen bei den Analysen berücksichtigen kann, muss ein Pathwayfile (`PATHWAYFILE`) eingelesen werden. Das Pathwayfile beinhaltet in der ersten Spalte den Pathwaynamen, in der zweiten Spalte die rs-Nummer und in der dritten den Gennamen. Ein KEGG-Pathwayfile kann von unserer Homepage heruntergeladen werden. Natürlich ist es auch hier möglich, eigene Dateien zu erstellen, jedoch muss darauf geachtet werden, dass das Format eingehalten wird. Um die Laufzeit zu verbessern, ist es sinnvoll, die Datei nach Pathwayname zu sortieren.

```
hsa00010    rs61487361  HK2
hsa00010    rs2286168   ALDH3B1
hsa00010    rs41275697  ADH1B
hsa00010    rs191970    PGM1
hsa00010    rs7583259   GALM
...
```

Listing 3.7: Ausschnitt aus dem Pathwayfile: Bei dieser Datei kann eine Überschriftenzeile verwendet werden, die Reihenfolge der Spalten muss beibehalten werden. In der ersten Spalte steht der Pathwayname, gefolgt von rs-Nummer und Genname.

3.5.3.4 Covariatefile

Um Kovariaten in die Analyse zu integrieren, muss ein Covariatefile (`COVARIATEFILE`) erstellt werden. In dieser Datei steht in jeder Zeile ein Individuum mit den zugehörigen Kovariaten. In den ersten beiden Spalten müssen Familien-ID und Personen-ID stehen, danach können bis zu 10 Kovariaten (`COVARIATES`) folgen. Die Personen werden über PID und FID mit dem tfam-File abgeglichen. Das Geschlecht kann separat ohne Covariatefile als Kovariate ausgewählt werden (`SEXCOV`). Der Wert für fehlende Daten ist „-“ oder „x“.

```
FID PID COV1 COV2
co1 co1 1.31 5.75
co2 co2 7.24 5.97
...
ca1 ca1 1.85 5.12
ca2 ca2 2.36 6.42
...
```

Listing 3.8: Ausschnitt aus dem Covariatefile: Auch hier wird eine Überschriftenzeile verwendet. Nach Familien- und Personen-ID folgen die Kovariaten.

3.5.3.5 Modelfile

Das Modelfile (`MODELFILE`) wird benötigt, um benutzerdefinierte Modelle für die logistische/lineare Regression zu erstellen. Die erste Spalte präsentiert den Namen

des Parameters, gefolgt von je einer Spalte für L1 und L2 mit einem 0/1 Indikator, der angibt, ob der Parameter in den jeweils zu vergleichenden Likelihoods L1 und L2 verwendet wird. Um das benutzerdefinierte Modell zu verwenden, ist es außerdem notwendig, `TEST` auf `M` zu setzen.

```
PARAMETER   L1  L2
x1          1   1
x1D         0   0
x2          1   1
x2D         0   0
x1x2        1   0
...
```

Listing 3.9: Dies ist ein Beispiel für ein Modelfile. Es beschreibt den Test auf allelische Interaktion.

Dieses Modelfile beschreibt den Test auf allelische Interaktion. Dieser kann jedoch auch direkt mit `TEST 5` und ohne Modelfile aufgerufen werden.

3.5.3.6 SNPfile

Das SNPfile (`SNPFILE`) bietet dem Benutzer die Möglichkeit, nur bestimmte SNPs aus einem tped-File für die Analyse zu verwenden. Es werden genau die ausgewählten SNPs aus der Liste verwendet. Damit die Datei eingelesen wird muss `SNPLIST` auf `1` gesetzt werden.

```
rs11248850
rs7190878
rs7404049
rs4984707
rs11649498
...
```

Listing 3.10: SNPlistfile: Es werden genau die SNPs aus der Liste verwendet.

3.5.3.7 Combifile

Falls nur ganz bestimmte Paare oder Tripel analysiert werden sollen, ist die Verwendung des Combilistfile (`COMBIFILE`) sinnvoll. In dieser Datei stehen die SNP-Nummern der Paare oder Tripel in einer Zeile. Eine nahliegende Anwendung wäre die Replikation von Vorbefunden. SNP-Paare von einer ersten Studie A, deren Ergebnisse repliziert werden sollen, schreibt man in das Combifile und analysiert dann nur diese mit dem Datensatz der Studie B. Für diese Anwendung muss `COMBILIST 1` ausgewählt sein.

```
rs11248850 rs7190878
rs7404049 rs4984707
rs11649498 rs1040499
...
```

Listing 3.11: Combilistfile: Es werden nur die SNP-Paare/Tripel aus der Liste berechnet.

3.5.4 Qualitätskriterien

Hohe Ausfallraten bei bestimmten Personen oder SNPs sind ein Hinweis auf schlechte DNA-Qualität bzw. Probleme bei der Genotypisierung. Entsprechend sind auch die tatsächlich vorhandenen Genotypen solcher Personen oder SNPs stärker fehlerbehaftet und sollten von der Analyse ausgeschlossen werden. Um die Daten der einzelnen Personen und SNPs auf ein einheitliches Qualitätsniveau zu bringen, wird in INTERSNP ein iterativer QC-Algorithmus angewendet. Häufig werden bei der Qualitätskontrolle alle SNPs und Personen, die unter einer definierten Callrate liegen, entfernt, was dazu führen kann, dass mehr Personen oder SNPs „gelöscht“ werden als nötig. Vorteil des iterativen QC-Algorithmus in INTERSNP ist, dass durch SNPs, die bei allen Personen eine schlechte Callrate haben, nicht zusätzlich auch Personen ausgefiltert werden, deren Callrate aufgrund dieser SNPs sinkt. Statt dessen werden Personen und SNPs ausgeschlossen, die im Sinne der Callrate schlechtere Qualität haben als der Großteil der Daten. Der iterative Algorithmus entfernt alle SNPs und Personen, deren Missingrate größer ist als die durchschnittliche Missingrate + eine benutzerdefinierte Missingrate-Differenz (`MRDIFF`). Startpunkt des Algorithmus ist die durchschnittliche Missingrate des Gesamtdatensatzes, also über alle SNPs und Personen. In jeder Iteration werden entweder die SNPs oder die Personen mit einer Missingrate größer als die durchschnittliche Missingrate + `MRDIFF` selektiert und als „qc-failed“ markiert. Danach wird eine neue durchschnittliche Missingrate berechnet und weitere SNPs und Personen, die über der Missingrate + `MRDIFF` liegen, ausgefiltert. Der Algorithmus endet, wenn keine Veränderungen mehr auftreten, d.h. die durchschnittliche Missingrate + `MRDIFF` immer größer als die Missingrate eines einzelnen SNPs oder einer einzelnen Person ist. Der Algorithmus terminiert schnell, in der Regel spätestens nach drei kompletten SNP/Personen-Durchgängen.
Neben dem QC-Algorithmus wurden noch weitere Qualitätskontrollen für SNPs implementiert. Beim HWE-Test werden SNPs, die zu einem bestimmtem Signifikanzniveau (z.B.: $\alpha = 10^{-6}$ für Fälle, $\alpha = 10^{-4}$ für Kontrollen) vom HWE abweichen, gelöscht. Dabei erwartet man Gültigkeit des HWE in der Kontrollgruppe (`HWE_P_CONTROL`), jedoch nicht zwingend in der Gruppe der Fälle (`HWE_P_CASE`), wo echte Krankheitsassoziation zu HWE-Abweichungen führen kann. Besonders starke Abweichungen vom HWE werden allerdings in der Regel auch bei Fällen durch Genotypisierungsfehler verursacht und sollten deshalb von der Assoziationsanalyse ausgeschlossen werden. Des Weiteren kann der Grenzwert für die `MAF` vorgegeben werden, also die Häufigkeit des seltenen Allels eines SNPs in der Gesamtstichprobe.

3.5.5 Tests

3.5.5.1 Einzelmarkeranalyse

Die Einzelmarkeranalyse (`SINGLEMARKER`) wird bei jedem Aufruf von INTERSNP für alle QC-SNPs durchgeführt, bevor weitere Analyseansätze gestartet werden. Der Benutzer kann über das Selectionfile wählen, ob der Armitage-Trendtest (`TEST 1`) [Armitage, 1955], der Genotyptest mit zwei Freiheitsgraden (`TEST 2`), die logistische/lineare Regression mit einem Freiheitsgrad (`TEST 3`) oder die logistische/lineare Regression mit zwei Freiheitsgraden (`TEST 4`) zur Analyse verwendet werden soll. Alle Chromosomen, auch X und Y, können mit der logistischen/linearen Re-

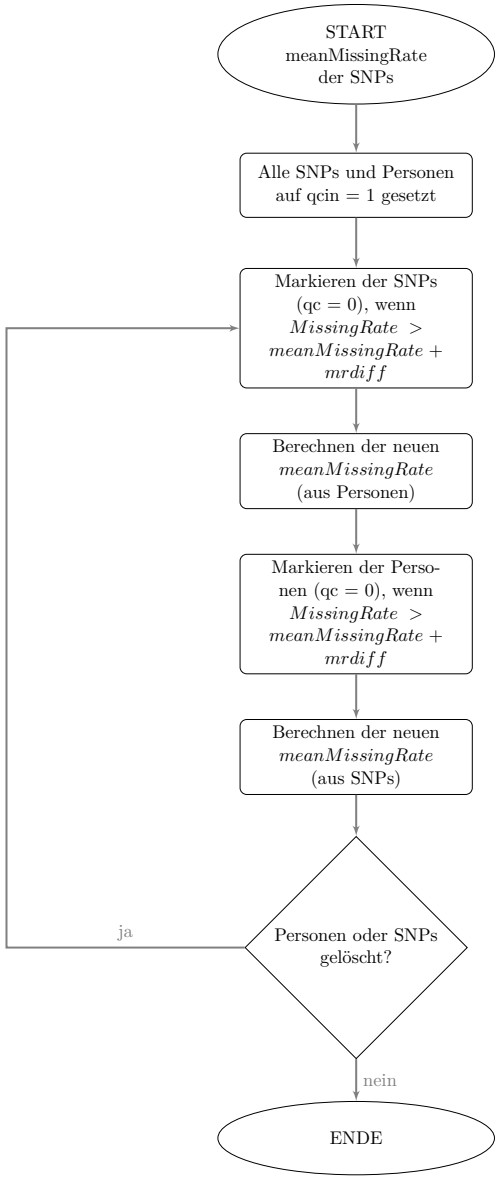

Abbildung 3.4: Iterativer QC-Algorithmus in INTERSNP für Personen und SNPs.

gression einbezogen werden, die zusätzlich den Vorteil hat, dass Kovariaten berücksichtigt werden können. Bei der logistischen/linearen Regression kann zwischen allelischem (`TEST 3`) oder genotypischem Test (`TEST 4`) unterschieden werden. Für die X-chromosomalen SNPs, die nicht mit logistischer/linearer Regression berechnet werden, wird der allel-basierte Test von Clayton [2008] mit einem Freiheitsgrad verwendet. Claytons Ansatz garantiert, dass heterozygote Männer und homozygote Frauen auf die gleiche Weise zu der Teststatistik beitragen, unter Berücksichtigung der Tatsache, dass in der Regel nur eines der X-Chromosomen in Frauen aktiv sein kann.

3.5.5.2 Multimarkeranalyse

Bei gleichzeitiger Analyse mehrerer SNPs (`TWO_MARKER` oder `THREE_MARKER`) ist eine wichtige Frage, ob man „auf“ oder „unter“ Interaktion, also unter Einbeziehung

(„voller“ Test) oder ohne Einbeziehung der marginalen Effekte testet. Im ersten Fall wird nicht explizit auf Interaktion getestet, sondern ein Test durchgeführt, der Power hat, wenn Interaktion vorliegt. Da beide Strategien je nach Krankheitsmodell sinnvoll sein können, bietet unsere Software beide an. Wenn es das Ziel ist, Gene zu entdecken, die an der Ätiologie der Krankheit beteiligt sind, ist der volle Test nützlich. Es scheint plausibel, dass interagierende Gene zumindest auch schwache marginale Effekte zeigen. Auch wenn solche marginalen Effekte klein sind, verbessert ihre Einbeziehung in die statistische Analyse die Chance, betreffende Paare zu entdecken. Dies wurde von Marchini et al. [2005], die den vollen Genotyptest untersucht haben, gezeigt. Ein Nachteil der vollen Teststrategie ist, dass (starke) Assoziationen von bestimmten SNPs dazu führen können, dass Paare, die diesen SNP enthalten, signifikant werden, auch wenn der andere SNP weder marginal assoziiert noch ein Interaktionspartner ist. Dieses Problem kann gelöst werden, indem die Ausgabe gefiltert wird: Für jeden SNP schreibt INTERSNP nur die besten 50 SNP-Kombinationen, welche den SNP enthalten, in die Ausgabedatei.

Ein expliziter Test auf Interaktion kann sinnvoll sein, wenn entweder die Interaktion an sich das Forschungsziel ist oder wenn eine bestimmte Region oder ein SNP als assoziiert bekannt ist. In diesem Fall kann die Suche nach Interaktionspartnern hilfreich sein um weitere Gene zu finden, die in der Ursache der Krankheit involviert sind.

In INTERSNP wurden verschiedene Multimarkertests für zwei und drei SNPs implementiert, welche in Tabelle 3.3 (Tests 1-12) zusammengefasst sind. Die Nummerierung der Tests wird im Selectionfile beibehalten, d.h. `TEST 1` entspricht dem χ^2-Test mit acht Freiheitsgraden. Die Tabelle bezieht sich auf die logistische Regression für Fall-Kontroll-Daten. Jedoch gibt es für quantitative Datensätze analoge Tests der linearen Regression. Dabei ist zu beachten, dass statt der Likelihoods die Fehlerquadratsummen (SSE) verwendet werden. Um quantitative Datensätze zu verwenden, muss im Selectionfile `QT` auf `1` gesetzt und der Wert für den fehlenden Phänotyp (`MISSING_PHENO`) festgelegt werden.

3.5.6 Prioritäten

Die Grundidee von INTERSNP ist es, die Anzahl der zu analysierenden Paare anhand von Kriterien zu beschränken. Neben statischen und genetischen Kriterien können auch Pathwayinformation zur Reduzierung der Tests beitragen. Außerdem ist die Möglichkeit gegeben, unabhängig von diesen Kriterien, nur eine gezielte Gruppe von SNPs auszuwählen, was im Abschnitt 3.5.6.4 beschrieben wird.

3.5.6.1 Statistisches Kriterium

Für jeden SNP wird ein Einzelmarker-p-Wert mit dem Armitage-Trendtest oder der logistischen/linearen Regression von den eigenen Daten berechnet. Eine Liste der besten n SNPs wird basierend auf diesen p-Werten erstellt. Die Länge n der Liste (`SINGLETOP`) wird vom Benutzer festgelegt. Zudem kann vom Benutzer ausgewählt werden, wie viele SNPs von jeder Kombination in der Bestenliste liegen sollen (`M_WITH_SINGLETOP`).

Nr.	Art	Formel	FG	Kommentar
1	χ^2-Test		8	Voller genotypischer Test
2	Log-lineares Modell	$l_{1,2}^{G,I}$ vs $l_{1,2}^{G}$	4	Test auf genotypische Interaktion
3	Logistische Regression	$L_{1,2}^{A,I}$ vs L_0	3	Voller allelischer Test
4	Logistische Regression	$L_{1,2}^{G,I}$ vs L_0	8	Voller genotypischer Test
5	Logistische Regression	$L_{1,2}^{A,I}$ vs $L_{1,2}^{A}$	1	Test auf additive Interaktion
6	Logistische Regression	$L_{1,2}^{G,I}$ vs $L_{1,2}^{G}$	4	Test auf genotypische Interaktion
7	Logistische Regression	$L_{1,2}^{A,I}$ vs L_1^{A}	2	Zusätzlicher allelischer Effekt von SNP 2
8	Logistische Regression	$L_{1,2}^{G,I}$ vs L_1^{G}	6	Zusätzlicher genotypischer Effekt SNP 2
9	Logistische Regression	$L_{1,2,3}^{A,I}$ vs L_0	7	Voller additiver Test (3 SNPs)
10	Logistische Regression	$L_{1,2,3}^{A,I}$ vs $L_{1,2,3}^{A}$	4	Test auf allelische Interaktion (3 SNPs)
11	Logistische Regression	$L_{1,2,3}^{A,I}$ vs $L_{1,2,3}^{A,I_2}$	1	Test auf 3-facher allelischer Interaktion
12	Logistische Regression	$L_{1,2,3}^{A,I}$ vs $L_{1,2}^{A,I}$	4	Zusätzlicher allelischer Effekt von SNP3

Tabelle 3.3: Likelihoods der Multimarkertests in INTERSNP. Test 1-8 wurden für die Analyse von zwei SNPs modelliert und die Test 9-12 für drei SNPs.

3.5.6.2 Genetisches Kriterium

Entsprechend den Kriterien *genetische Lokalisation* und *Funktionsklasse*, werden die SNPs in fünf ineinander geschachtelte Gruppen mit aufsteigendem genetischen Einfluss aufgeteilt:

- 0: „Genwüste“: Abstand zum nächsten Exon des nächsten Gens ist größer als 100 kb.
- 1: Gen-LD-Bereich: Abstand zum nächsten Exon des nächsten Gens ist weniger als 100 kb oder der SNP liegt in einem Intron von einem Gen.
- 2: Exon: Lage in einem Exon eines Gens.
- 3: Kodierend: SNP liegt in einer kodierenden Region eines Gens.
- 4: Nicht-synonym: SNP verursacht einen nicht-synonymen Aminosäureaustausch.

Der Benutzer legt den gewünschten genetischen Einfluss (`GENETIC_IMPACT`) fest und wählt zusätzlich die Anzahl der SNPs (`M_WITH_GENETIC_IMPACT`) des Paares (Tripels), die mindestens das jeweilige Kriterium erfüllen soll. Die SNP-Annotation stammt aus einem entsprechenden Annotationfile, welches in das Programm geladen wird. Eine Einschränkung aufgrund von genetischen Merkmalen kann sinnvoll

sein, da mehr GWAS-Befunde in kodierenden Regionen liegen als aus der Verteilung der SNPs auf kommerziellen SNP-Chips zu erwarten wäre [Manolio et al., 2009].

3.5.6.3 Pathwayinformationen

Pathwayinformationen (`PATHWAY`) können bei allen Analysestrategien dazu beitragen, die Anzahl der Tests einzugrenzen, indem nur SNPs analysiert werden die in Genen liegen, welche einem bestimmten Pathway angehören. Die Pathwayinformationen werden über eine Eingabedatei, die eine Liste von Pathways mit den zugehörigen rs-Nummern der SNPs beinhaltet, vom Benutzer bereitgestellt. So können Experten auf diesem Gebiet Pathway-basierte Interaktionsanalysen auf die Pathways, die eine eventuelle Relevanz für den Phänotypen haben, eingrenzen. Wenn kein Expertenwissen vorhanden ist oder nicht existiert, wird die Nutzung der KEEG Datenbank [Kanehisa et al., 2006] empfohlen.

3.5.6.4 Gezielte Auswahl

Neben den Prioritäten gibt es noch weitere Möglichkeiten, die Anzahl der Tests zu reduzieren. Man kann beispielsweise nur eine bestimmte Region auswählen (`POSCHOICE`) oder eine bestimmte Region ausschließen (`NEGCHOICE`). Dabei muss zuerst das Chromosom angegeben und nachfolgend kann die Region durch die Angabe der Position noch verfeinert werden (z.B.: chr4;chr12,123000-160000;chr24;). Hier wird also Chromosom 4, Chromosom 12 von Basenpaar 123000 bis 160000 und Chromosom 24 (Y-Chromosom) ausgewählt.
Es ist auch möglich, bestimmte SNPs festzusetzen (`SNP1, SNP2, SNP3`). Das bedeutet, dass alle Kombinationen mit diesem einen SNP (z.B. `SNP1 rs11248850`) berechnet werden. Es besteht die Möglichkeit, bis zu drei SNPs festzusetzen. Diese Option kann mit den oben erwähnten Prioritäten kombiniert werden, beispielsweise SNP1, der zuvor in der Literatur gefunden wurde, wird festgesetzt und der SNP2, ein potentieller Interaktionspartner, soll im gemeinsamen Pathway liegen.
Sollen nur die Kombinationen von ganz bestimmten SNPs berechnet werden, empfiehlt es sich `SNPLIST` im Selectionfile auszuwählen. Bei dieser Option werden nur die SNPs kombiniert, die im SNPfile (`SNPFILE`) enthalten sind. Eine weitere Einschränkung bietet die Wahl der `COMBILIST` im Selectionfile, bei welcher nur die Kombinationen berechnet werden, die im Combifile (`COMBIFILE`) angegeben werden. Eine typische Anwendung der `COMBILIST` ist die Replikation von Ergebnissen aus Studien anderer Gruppen. Die Paare mit den besten Interaktionsergebnissen können so explizit überprüft werden.
Auf der Personenseite ist es auch möglich, wenn erwünscht, nur Männer (`ONLY_MALE 1`) oder nur Frauen (`ONLY_FEMALE 1`) zu analysieren.

3.5.7 Pre-test

Lineare und logistische Regression erlauben flexible Modellierung und Verwendung von Kovariaten, jedoch kann ihre Rechenzeit aufgrund der Operationen an den relativ großen Design-Matrizen (Anhang A) sehr zeitintensiv sein. Soll eine genomweite Interaktionsanalyse durchgeführt werden, macht es Sinn, erst einmal einen einfachen, leicht zu berechnenden Pre-test auf alle SNP-Paare anzuwenden.

Danach kann dann ein verfeinerter Test nur auf die SNP-Paare, die beim Pre-test einen p-Wert kleiner als die Obergrenze (`PRETEST_CUTOFF`) besitzen, angewendet werden. Eine Cutoff-Grenze von $0,01$ ist dabei eine vernünftige Wahl, da dies die Anzahl der Berechnungen des „wirklichen“ Tests im Durchschnitt um mindestens einen Faktor von 100 reduziert, aber gleichzeitig garantiert, dass keine „echten“ p-Werte kleiner als 10^{-5} verloren gehen. Da p-Werte bei einer kompletten GWIA im Bereich von 10^{-12} liegen sollten [Becker et al., 2011], wird der Verlust der Paare mit weniger signifikanten p-Werten als 10^{-5} in den meisten Fällen irrelevant sein. Auch eine Obergrenze von $0,001$ oder $0,0001$ für den Pre-test wäre akzeptabel. Die einzige Situation, in der relevante Paare verloren gehen könnten, ergibt sich, wenn Kovariaten einen extrem großen Einfluss auf die p-Werte haben, da der Pre-test die Kovariaten ignoriert.

3.5.7.1 Pre-test allelischer Interaktion (logistische Regression)

Wenn `PRETEST 1` in Kombination mit `TEST 3` (voller allelischer Test, 3 FG) oder `TEST 5` (Test auf additive Interaktion, 1 FG) ausgewählt ist, wird ein log-lineares Modell mit 1 FG (Test auf allelische Interaktion) als Pre-test angewendet. Das log-lineare Modell wird analog zu dem log-linearen Modell im genotypischen Fall (`TEST 2`, siehe Abschnitt 3.3.1) definiert. Dort ist die $2 \times 3 \times 3$-Genotyp-Kontingenztafel (siehe Tabelle 3.1) festgelegt. Im allelischen Fall betrachtet man die $2 \times 2 \times 2$-SNP-Allel Tafel, wobei die erste Dimension die beiden Allele von SNP1 widerspiegelt, die zweite die beiden Allele von SNP2 und die dritte den Fall-Kontroll-Status. Doppelt heterozygote Fälle/Kontrollen tragen 0,5 Einheiten zu jeder Zelle der 2×2-Tafel der Fälle/Kontrollen bei. In der Praxis ist dies eine einfache Umschreibung der Doppeltheterozygoten, jedoch wird der log-lineare Test dadurch konservativ wiedergeben, wenn seine Teststatistik mit einer χ^2-Verteilung mit 1 FG bewertet wird. In diesem Sinne hat der log-lineare Test mit einem Freiheitsgrad nicht alle Eigenschaften, die einen klassischen statistischen Test auszeichnen. Jedoch sind seine p-Werte stark mit den p-Werten für den Test auf allelische Interaktion vom logistischen Modell korreliert. Aus diesem Grund kann dieser als Pre-test verwendet werden, wenn die Cutoff-Grenze wie oben erläutert gewählt wird.

3.5.7.2 Pre-test genotypischer Interaktion (logistische Regression)

Im genotypischen Fall ist der log-lineare Test (`TEST 2`) der offensichtliche Pre-test. Da dieser Test ein „wirklicher“ Test ist, wäre eine mögliche Strategie, eine komplette GWIA mit `TEST 2` durchzuführen, eine „Combilist“ (`COMBILIST`, siehe Abschnitt 3.5.6.4) mit den besten Ergebnissen zu erstellen und dann einen anderen Test mit der „Comiblist“ zu starten.

3.5.7.3 Pre-test allelischer Interaktion (lineare Regression)

Wenn `PRETEST 1` in mit `TEST 3` (voller allelischer Test, 3 FG) oder `TEST 5` (Test auf allelische Interaktion, 1 FG) ausgewählt ist, wird eine ANOVA mit einem Freiheitsgrad als Pre-test verwendet. Doppelheterozygote werden genau wie bei der logistischen Regression behandelt. Dabei sollte beachtet werden, dass bei der Auswahl von `TEST 3` und dem Pre-test Paare mit marginalen Effekten in beiden SNPs, aber moderaten Interaktionsergebnissen, verloren gehen können.

3.5.7.4 Pre-test genotypischer Interaktion (lineare Regression)

Wenn `PRETEST 1` in Kombination mit `TEST 4` (voller genotypischer Test, 8 FG) oder `TEST 6` (Test auf genotypische Interaktion, 4 FG) ausgewählt wurde, wird eine ANOVA mit 4 FG (Test auf genotypische Interaktion) als Pre-test verwendet. Zu beachten ist auch hier, dass Paare mit marginalen Effekten in beiden SNPs, aber moderaten Interaktionsergebnissen verloren gehen können.

3.5.8 Multiples Testen

3.5.8.1 Monte-Carlo-Simulation

Multiples Testen bedeutet grundsätzlich, dass zur selben bzw. zu inhaltlich zusammengehörenden Fragestellungen verschiedene Hypothesen getestet werden. Das Problem des multiplen Testens besteht darin, dass man bei jedem einzelnen Test mit einer Wahrscheinlichkeit α einen Fehler 1. Art begehen kann und sich dieser Fehler aufsummiert. Somit wird die Nullhypothese fälschlicherweise zu oft verworfen. Das Ansteigen des Fehlers kann durch Bonferroni-Korrektur, welche in Abschnitt 1.2.4.3 beschrieben ist, kontrolliert werden. Jedoch ist diese für SNPs im LD zu konservativ. Hier bietet die Korrektur mittels MC-Simulation eine Alternative. Dabei wird der p-Wert anhand von Permutationen bestimmt. Dazu werden die Genotypdaten beibehalten, aber die Phänotyplabel zufällig auf die Individuen verteilt, um „Vergleichsdatensets“ zu erstellen, die die LD-Struktur unverändert lassen, aber unter der Nullhypothese „keine Assoziation“ erzeugt werden. Nur wenn viele solcher Datensets analysiert werden, kann der korrigierte p-Wert berechnet werden. Geprüft wird folgende Nullhypothese: „Keine der beobachteten Kombinationen (SNP Paare) zeigt Assoziation/Interaktion“. Das Konzept ist einfach, kann aber zu einer Herausforderung an die vorhandenen Rechenkapazitäten werden, da die Prozedur oft wiederholt werden muss, um eine zuverlässige p-Wert-Schätzung zu erhalten.
Als Beispiel wird hier der Ablauf unter Strategie III (alle Paare der besten 50.000 Einzelmarkerergebnisse und beide SNPs aus einer kodierenden Region) vorgestellt:

1. Berechne für jeden SNP den Einzelmarker-p-Wert.

2. Erstelle eine Liste der besten Einzelmarkerergebnisse (50.000).

3. Berechne Multimarker-p-Wert mit Test 3 (allelischer Test unter Einbeziehung von marginalen Effekten und Interaktion).

4. Erstelle geordnete Liste der besten Multimarker-p-Werte.

5. Führe m Simulationen durch: Permutiere den Fall-Kontroll-Status zufällig unter Beibehaltung des Fall-Kontroll-Zahlenverhältnisses. Wiederhole die Schritte 1-4. Die Einzelmarker-Bestenlisten werden sich dabei für jedes Replikat unterscheiden. Dies ist notwendig, um den Selektionsprozess zu imitieren und um den Fall zu berücksichtigen, dass die marginalen Effekte die Teststatistik von Test 3 beeinflussen.

6. Jetzt ist es möglich den korrigierten p-Wert für den besten Multimarker-p-Wert in den realen Daten zu berechnen, aber auch für die k besten p-Werte p_k. Der korrigierte p-Wert für den k–ten p-Wert p_k berechnet sich als s/m,

wobei s die Anzahl der simulierten Datensätze ist, für die der beste p-Wert kleiner oder gleich p_k ist und m die Anzahl der Replikationen. Zu beachten ist, dass man p_k auch für $k > 1$ mit dem besten p-Wert p_1 der simulierten Daten vergleichen muss. Der Vergleich mit p_k aus der Simulation könnte zu der sinnlosen Situation führen, bei der der korrigierte p-Wert für ein p_k mit k>1 besser ist, als der korrigierte p-Wert für p_1.

Um die MC-Simulation für die Korrektur des multiplen Testens zu verwenden, muss mit dem Schlüsselwort `SIMULATION` die Anzahl der Simulation festgelegt werden. Soll zusätzlich auch die Einzelmarkeranalyse korrigiert werden, muss `MC_WITH_SM` ausgewählt werden. Um zu korrigiertem Niveau $\alpha = 0,05$ zu testen, sind 1.000 Simulationen in der Regel ausreichend.

3.5.9 Ausgabedateien

Der Name der Ausgabedatei wird durch das Schlüsselwort `OUTPUTNAME` festgelegt und im Folgenden mit „*“ gekennzeichnet.

3.5.9.1 Einzelmarkeranalyse

Alle Einzelmarkerergebnisse, also korrigierter und unkorrigierter p-Wert und einige zusätzliche Informationen werden in die Datei `*Singlemarker.txt` geschrieben. Neben rs-Nummer, Chromosom und Position findet man in dieser Datei auch die Allele, p-Wert für HWE der Fälle und Kontrollen, Allelhäufigkeiten, Allelfrequenzen und Konfidenzintervalle. Wird die logistische oder lineare Regression verwendet, dann stehen zusätzlich die vom Modell geschätzten „Betas“ und deren Standardfehler in der Datei (siehe Listing 3.12). Im `*SinglemarkerTop.txt` stehen nur die n besten Ergebnisse (Auswahl über `SINGLETOP`), um sich schnell einen Überblick verschaffen zu können. Es besteht darüber hinaus die Möglichkeit, Annotationsinformation in die Einzelmarkerergebnisdatei zu schreiben. Hierzu muss im Selectionfile `ANNOTATE` auf `1` gesetzt und mit `GENECOL` die Spalte mit der notwendigen Information im Annotationfile angegeben werden.

```
No Chr rs_No Position Gene minor major MAF SNP_MR HWE_Ca
HWE_Co P_Single-marker P_Corr A_Ca_N B_Ca_N A_Co_N B_Co_N
A_Ca B_Ca A_Co B_Co OR_A LCL_A RCL_A OR_B LCL_B RCL_B beta1
se1 OR1 LCL1 RLC1
1 16 rs1541449 2815484 PRSS21 G A 0.31268 0 0.609333
0.20498 0.000385391 0.319866 243 349 217 477 0.410473 0.589527
0.31268 0.68732 1.53052 1.21692 1.92495 0.653372 0.519495
0.821749 0.405504 0.11393 1.50006 1.19986 1.87537
2 16 rs17136872 4193035 SRL C T 0.252161 0.233281 0.657819
0.159986 0.00050054 0.393873 44 248 175 519 0.150685 0.849315
0.252161 0.747839 0.526175 0.365807 0.756849 1.90051 1.32127
2.73368 -0.605159 0.164863 0.545988 0.395229 0.754253
3 16 rs3810818 4372030 VASN T G 0.169617 0.0217729 0.660532
0.924103 0.000949132 0.613098 142 438 115 563 0.244828 0.755172
0.169617 0.830383 1.58717 1.20464 2.09119 0.63005 0.478197
0.830125 0.467145 0.141255 1.59543 1.20959 2.10435
```

Listing 3.12: Die Abbildung zeigt einen Ausschnitt aus einer Einzelmarkerergebnisdatei (`*Singlemarker.txt`). In diesem Beispiel wurde die logistische Regression für die Einzelmarkeranalyse verwendet. Zusätzlich findet man die Genzuordnung der SNPs.

3.5.9.2 Multimarkeranalyse

Bei der Multimarkeranalyse werden neben korrigiertem (Bonferroni-Korrektur) und unkorrigiertem p-Wert auch die rs-Nummer, Chromosom, Position und Einzelmarker-p-Wert für jeden beteiligten SNP ausgegeben (`*BestMarkerCombi2.txt` bzw. `*BestMarkerCombi3.txt`). In der zweiten Datei der Multimarkeranalyse (`*BestMarkerCombi2Details.txt` bzw. `*BestMarkerCombi3Details.txt`) befinden sich weitere Details wie die Kontingenztafeln für Fälle und Kontrollen und die geschätzten Regressionsparameter. Die Anzahl der Ergebnisse, die in die `*BestMarkerCombi`-Datei geschrieben werden sollen, kann mit `PRINTTOP` festgelegt werden.

```
No Chr_1 rs_No_1 Pos_No_1 p-Single-marker_1 Chr_2 rs_No_2
Pos_No_2 p-Single-marker_2 p-value p-value_corr
1 16 rs550713 1066680 0.99159 16 rs887250 5604128 0.324954
1.31293e-06 0.480979
2 16 rs761068 1391520 0.102839 16 rs2283479 4254530 0.205624
8.41023e-06 0.985018
3 16 rs3848375 4617191 0.0788968 16 rs1860307 5660562 0.744311
8.5981e-06 0.98636
```

Listing 3.13: Ausschnitt aus der Multimarkerergebnisdatei (`*BestMarkerCombi2.txt`). Bei dieser Analyse wurden SNP-Paare mit der logistischen Regression getestet. Neben den Informationen zu den beiden SNPs (Chromosom, rsNummer, Position und Einzelmarker-p-Wert) werden p-Wert und korrigierter p-Wert der Multimarkeranalyse ausgegeben.

```
1: SNP1: rs550713 SNP2: rs887250

cases:    controls:
1 6 11   0 1 16
3 18 63   1 15 93
1 18 169   3 51 166

L1:
OR0: 1.33304 OR_lcl0: 0.817128 ORrcl0: 2.1747
OR1: 3.57134 OR_lcl1: 1.89087 ORrcl1: 6.74528
OR2: 1.8229 OR_lcl2: 1.06852 ORrcl2: 3.10987
OR12: 4.85244 OR_lcl12: 2.43543 ORrcl12: 9.66816
```

Listing 3.14: Für weitere Details steht das zugehörige `*BestMarkerCombi2Details.txt` zur Verfügung. Neben der Kontingenztafel für Fälle und Kontrollen befinden sich die Parameter der logistischen Regression in dieser Datei.

3.5.9.3 Monte-Carlo-Simulationen

Wurden MC-Simulationen zum Korrigieren des p-Wertes verwendet, werden diese MC-korrigierten p-Werte in die Datei `*ToplistMC.txt` geschrieben.

```
No Chr_No_1 rs_No_1 Chr_No_2 rs_No_2 Chr_No_3 rs_No_3 Pos_No_1
Pos_No_2 Pos_No_3 p-value MC-p-value
1 16 rs550713 16 rs887250 - - 1066680 5604128 - 1.31293e-06 0.3
2 16 rs761068 16 rs2283479 - - 1391520 4254530 - 8.41023e-06 1
3 16 rs3848375 16 rs1860307 - - 4617191 5660562 - 8.5981e-06 1
```

Listing 3.15: In der ToplistMC.txt Datei findet man den unkorrigierten Interaktions-p-Wert und den MC-korrigierten p-Wert.

3.5.9.4 LOG-File

Die Logdatei (`*logfile.txt`) ist das automatisch geführte Protokoll über die wichtigsten Schritte der Analyse. Sie fasst zunächst die verwendeten Optionen und Parameter des Selectionfiles zusammen und gibt dann an, welche Dateien eingelesen und die Anzahl der SNPs und Personen, die nach der Qualitätskontrolle für die Analyse verwendet wurden. Anschließend werden die zehn besten Ergebnisse der Einzelmarker- und Multimarkeranalyse aufgelistet. Außerdem werden Warnungen und Fehlermeldungen protokolliert. Wichtige Informationen werden zusätzlich auch auf dem Bildschirm ausgegeben.

3.5.9.5 Qualitätskontrolle

Alle SNPs und Personen, die die Qualitätskriterien nicht erfüllen, werden nach der Qualitätskontrolle von der Analyse ausgeschlossen. Zur besseren Übersicht werden diese SNPs in die Datei `*deletedSnps.txt` bzw. diese Personen in die Datei `*deletedPerson.txt` geschrieben.

```
Row_No rs_No
30      rs11248914
58      rs13226
60      rs1802752
```

```
FID PID
co1 co1
ca2 ca2
```

3.5.9.6 Fehlermeldungen und Warnungen

Treten Fehler während des Programmablaufs auf, werden diese Fehler in die Datei `*errorMessage.txt` geschrieben. Dies könnten auch Warnungen sein, die nicht zum Abbruch des Programms führen.

3.5.10 Beispiel-Strategien

Im Folgenden werden mögliche Strategien (I-VI) vorgestellt, die zeigen, wie die in INTERSNP implementierten Kriterien (statistische, genetische und Pathwayinformationen) zum Reduzieren der Tests angewendet werden können. Außerdem wird ein Beispiel für die Durchführung einer genomweiten Interaktionsanalyse mit dem Pre-Test vorgestellt. Bei den Beispielen werden nur die obligatorischen Optionen angegeben.

I. Alle SNP-Paare mit mindestens einem SNP aus den 10 besten Einzelmarkerergebnissen:

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
SINGLE_MARKER 1 // Armitage-Trendtest.
TWO_MARKER 1 // 2-Markeranalyse.
TEST 2 // Log-lineares Modell, Test auf genetische Interaktion.
SINGLETOP 10 // 10 besten Einzelmarkerergebnisse.
M_WITH_SINGLETOP 1 // Es muss mindestens einer der beiden SNPs
                      unter den besten 10 Einzelmarkerergebnissen
                      sein.
```

```
OUTPUTNAME ./meinProjekt/test // Name der Ausgabedateien mit Pfad
END
```

Mit einer kleinen Anzahl von besten SNPs hofft man echte Einzelmarkerassoziationen widerspiegeln zu können, die möglicherweise durch andere Studien bestätigt sind. Mit Strategie I könnte dann genomweit nach Interaktionspartnern zu den SNPs mit den stärksten marginalen Effekten gesucht werden.

II. Alle SNP-Paare aus der Liste der 1.000 besten Einzelmarkerergebnisse:

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
SINGLE_MARKER 1 // Armitage-Trendtest.
TWO_MARKER 1 // 2-Markeranalyse.
TEST 5 // Logistisches Regressionsmodell: Test auf
        additive Interaktion
SINGLETOP 1000 // 1.000 besten Einzelmarkerergebnisse.
M_WITH_SINGLETOP 2 // Alle SNP-Paare liegen unter den 1.000
                    besten Einzelmarkerergebnissen.
OUTPUTNAME ./meinProjekt/test // Name der Ausgabedateien mit Pfad
END
```

Hier wird auf Interaktion zwischen SNPs getestet, die einen Hinweis auf Einzelmarkerassoziation zeigen, aber nicht notwendigerweise nach Korrektur für multiples Testen signifikant sind. Die Betrachtung solcher Paare ist besonders interessant, wenn man davon ausgeht, dass beide interagierende SNPs marginale Effekte zeigen.

III. Alle SNP-Paare aus Liste der 50.000 besten Einzelmarkerergebnisse, die zusätzlich aus kodierenden Genregionen stammen:

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
ANNOTATIONFILE ./annotation.txt // Pfad zum Annotation-File.
SINGLE_MARKER 1 // Armitage-Trendtest.
TWO_MARKER 1 // 2-Markeranalyse.
TEST 5 // Logistisches Regressionsmodell:
        Test auf additive Interaktion.
SINGLETOP 50000 // 50.000 besten Einzelmarkerergebnisse.
M_WITH_SINGLETOP 2 // Alle Paare befinden sich unter den
                    50.000 Einzelmarkerergebnissen.
GENETIC_IMPACT 3 // Genetisches Kriterium: in einer
                    kodierenden Region.
M_WITH_GENETIC_IMPACT 2 //  Alle Paare liegen in einer
                           kodierenden Region von einem Gen.
OUTPUTNAME ./meinProjekt/test // Name der Ausgabedateien mit Pfad
END
```

In Ergänzung zu Strategie II wird verlangt, dass die SNPs eines Paares zusätzlich in kodierenden Regionen eines Gens liegen, da solche Regionen eine höhere a-priori Wahrscheinlichkeit haben, mit einer Krankheit assoziiert zu sein. Dieses Kriterium erlaubt es, das statistische Kriterium weniger regide zu handhaben und nur schwache Assoziation auf Einzelmarkerebene (SINGLETOP 50000) zu verlangen.

IV. Alle SNP-Paare nicht-synonymer SNPs:

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
ANNOTATIONFILE ./annotation.txt // Pfad zum Annotation-File.
SINGLE_MARKER 1 // Armitage-Trendtest.
TWO_MARKER 1 // 2-Markeranalyse.
TEST 5 // Logistisches Regressionsmodell:
        Test auf additive Interaktion.
GENETIC_IMPACT 4 // Genetisches Kriterium: nicht-synonym.
M_WITH_GENETIC_IMPACT 2 // Alle nicht-synonymen SNP-Paare.
OUTPUTNAME ./meinProjekt/test // Name der Ausgabedateien mit Pfad
END
```

SNPs, die zu einem nicht-synonymen Basenpaaraustausch des Genprodukts führen, haben eine besonders hohe a-priori Wahrscheinlichkeit in Krankheiten involviert zu sein. Deshalb ist es eine sinnvolle Strategie, alle solche SNP-Paare auf Interaktion zu testen.

V. Alle SNP-Paare aus der Liste der 50.000 besten Einzelmarkerergebnisse, die zusätzlich in einem gemeinsamen Pathway liegen:

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
PATHWAYFILE ./pathway.txt // Pfad zum Pathway-File.
SINGLE_MARKER 1 // Armitage-Trendtest.
TWO_MARKER 1 // 2-Markeranalyse.
TEST 2 // Log-lineares Modell, Test auf genetische
        Interaktion.
SINGLETOP 50000 // 50.000 besten Einzelmarkerergebnisse.
M_WITH_SINGLETOP 2 // Alle SNP-Paare, die unter den 50.000
                    besten Einzelmarkerergebnissen liegen.
PATHWAY 1 // Alle SNP-Paare, die in einem gemeinsamen
            Pathway liegen.
OUTPUTNAME ./meinProjekt/test // Name des Ausgabedateien mit Pfad
END
```

In Ergänzung zu Strategie II wird verlangt, dass die SNPs eines Paares in Genen eines gemeinsamen Pathways liegen. Da Pathways zusammenwirkende biologische Einheiten definieren, ist es naheliegend, Interaktionseffekte zwischen den entsprechenden SNPs zu vermuten. Das statistische Kriterium wird deshalb weniger strikt angewandt, da moderate Assoziation auf Einzelmarkerebene (SINGLETOP 5000) vorausgesetzt wird.

VI. Genomweite Interaktionsanalyse mit dem Pre-test:

```
TPED ./test.tped // Pfad zum tped-File.
TFAM ./test.tfam // Pfad zum tfam-File.
SINGLE_MARKER 1 // Armitage-Trendtest.
PRETEST 1 // Pre-Test ist ausgewählt.
PRETEST_CUTOFF 0.01 // Pre-test Cutoff ist bei 0,01.
TWO_MARKER 1 // 2-Marker-Analyse.
TEST 5 // Nur Tests, deren p-Wert < 0,01 sind, werden tatsächlich
        mit der logistischen Regression berechnet.
OUTPUTNAME ./meinProjekt/test // Name des Ausgabedateien mit Pfad
END
```

Eine komplette genomweite Interaktionsanalyse ohne Prioritäten ist mit der parallelisierten Version `intersnpA` (siehe Abschnitt 3.4.3) von INTERSNP

möglich. Die Rechenzeit wird durch den in Abschnitt 3.5.7 beschriebenen Pre-test reduziert.

Kapitel 4

Datenanalyse mit INTERSNP

4.1 Anwendung

4.1.1 Androgenetische Alopezie

Als erstes soll ein GWAS-Datensatz als Beispiel verwendet werden, der von Hillmer et al. [2008] bereits veröffentlicht wurde und erstmals die Region 20p11 als Suszeptibilitätsregion für Androgenetische Alopezie belegt. Androgenetische Alopezie ist die häufigste Ursache für Haarausfall beim Menschen, besonders bei Männern [Rexbye et al., 2005]. Es handelt sich hierbei um erblich bedingter Haarausfall, dessen Ursache das Steroidhormon Dihydrotestosteron (DHT) ist.
In die Analyse gingen 296 Fälle und 347 Kontrollen ein. Das stärkste Assoziationssignal (rs4548330) wurde in einem unabhängigen Sample bestätigt ($p = 2,7 \cdot 10^{-15}$ kombiniert). In diesem Anwendungsbeispiel analysieren wir die ursprüngliche GWAS (643 Individuen) noch einmal und führen mit INTERSNP eine genomweite Interaktionsanalyse durch. Es werden nur die 300.454 QC-SNPs für die Analyse verwendet, die mindestens eine Callrate von 95% in Fällen und Kontrollen aufweisen. Die niedrigere Callrate bei den anderen SNPs entstand nicht durch Genotypisierungsfehler, sondern dadurch, dass das SNP-Sample in Batches mit unterschiedlicher SNP-Dichte genotypisiert worden ist.

Nr	Chr	rs-Nr	Pos	p-Wert	p_{Bonf}	p_{MC}	Kommentar
1	23	rs4548330	496582	6,85e-10	2,06e-04	0	repliziert
2	23	rs5919235	496583	1,04e-09	3,11e-04	0,0001	repliziert
3	23	rs2497938	496629	2,93e-09	8,81e-04	0,0005	repliziert
4	23	rs1041668	496581	3,03e-09	9,09e-04	0,0006	repliziert
5	23	rs5919200	496579	5,92e-09	1,78e-03	0,0012	repliziert
6	23	rs775358	496577	8,36e-09	2,51e-03	0,0016	repliziert
7	23	rs12396249	496648	5,47e-08	0,016	0,0095	repliziert
8	23	rs5919393	496649	6,11e-08	0,018	0,0108	repliziert
9	20	rs1998076	468034	1,30e-07	0,039	0,0248	repliziert
10	13	rs4976846	388824	2,92e-07	0,088	0,0549	nicht repliziert

Tabelle 4.1: Einzelmarkerergebnisse mit Bonferroni-Korrektur und MC-korrigierten p-Werten

Als Hintergrundinformation sind zunächst die Einzelmarkerergebnisse in Tabelle 4.1 dargestellt. Die MC-korrigierten p-Werte basieren auf 10.000 Simulationen. Die ersten acht Ergebnisse beinhalten Marker vom bestätigten X-chromosomalen Locus [Hillmer et al., 2008]. Diese SNPs erreichten bereits in der anfänglichen GWAS nach Bonferroni-Korrektur genomweite Signifikanz. Die mit MC-Simulation korrigierten p-Werte sind jedoch genauer, da die Abhängigkeit des Tests hinsichtlich des LDs berücksichtigt wird. So ist SNP rs1998076 in Zeile 9 bereits nach Bonferroni-Korrektur genomweit signifikant ($p = 0,039$) und der p-Wert verbessert sich durch die MC-Simulationen auf $p = 0,0248$. Dieser SNP gehört zum Locus auf Chromosom 20, welcher in einem unabhängigen Sample repliziert wurde [Hillmer et al., 2008]. Der letzte SNP in Tabelle 4.1 war schließlich nicht signifikant, weder nach Bonferroni-Korrektur noch nach MC-Simulationen. Dieser SNP wurde auch in dem unabhängigen Sample nicht repliziert.
Bei den Interaktionsergebnissen werden wir uns nur auf die Ergebnisse der Strategie V (alle Paare aus der Liste der 50.000 besten Einzelmarkerergebnisse, die zusätzlich in einem gemeinsamen Pathway liegen) beschränken. Ergebnisse dieser Strategie sind in Tabelle 4.2 dargestellt. Hierbei wurde der Test auf genotypische Interaktion verwendet, welcher mit dem log-linearen Modell (`Test 2`) gerechnet werden kann. Das beste Interaktionspaar besteht aus den SNPs rs608139 (Chromosom 2) und rs4678398 (Chromosom 3). Beide SNPs zeigen ein moderates Einzelmarkerergebnis ($p = 0,0080$ und $p = 0,0091$) und liegen in Genen, die mit dem Pathway hsa04530 aus der KEGG Datenbank verknüpft sind.

rs-Nr1	**p1**[a]	**rs-Nr2**	**p2**[b]	**p-Wert**[c]	$\mathbf{p_{Bonf}}$	$\mathbf{p_{MC}}$	**Pathway**
rs608139	0,008	rs4678398	0,009	1,25e-06	0,023	0,0091	hsa04530
rs9436297	0,011	rs17863168	0,012	7,87e-05	1,000	0,4788	hsa04080
rs2892805	0,009	rs348458	0,013	0,00015	1,000	0,7204	hsa00830
rs2892805	0,009	rs610529	0,014	0,00025	1,000	0,8667	hsa00830
rs1199333	0,009	rs5750854	0,009	0,00038	1,000	0,9533	hsa04010
rs9816982	0,001	rs2186598	0,015	0,00038	1,000	0,9535	hsa04080
rs1464443	0,003	rs10487888	0,008	0,00050	1,000	0,9810	hsa04012
rs2575357	0,007	rs3741049	0,012	0,00052	1,000	0,9828	hsa00620
rs918938	0,011	rs7789059	0,014	0,00057	1,000	0,9886	hsa04514
rs11851957	0,002	rs1938958	0,009	0,00062	1,000	0,9917	hsa04080

Tabelle 4.2: Ergebnisse der Strategie V: Alle Paare sind aus der Liste der 50.000 besten Einzelmarkerergebnisse und liegen in einem gemeinsamen Pathway.
[a] p-Wert Armitage-Trendtest SNP 1
[b] p-Wert Armitage-Trendtest SNP 2
[c] p-Wert Test auf Interaktion (4 FG)

Das Interaktionsergebnis dieser beiden SNPs lässt auf genotypische Interaktion schließen ($p = 1,249 \cdot 10^{-6}$). Der Effekt wird durch den Überschuss von doppelt Heterozygoten in den Kontrollen beeinflusst (16,2% versus 3,2% in Fällen, siehe Tabelle 4.3). Der Interaktions-p-Wert hält der Bonferroni-Korrektur mit der Anzahl der 2-Markertests (18.458 Tests, korrigiert $p = 0,023$) stand. Zu beachten ist, dass diese beiden SNPs aus der Liste der 50.000 besten Einzelmarkerergebnisse gewählt wurden. Die Bonferroni-Korrektur mit der tatsächlich ausgeführten Anzahl von Tests ist trotzdem ausreichend, da Test 2 (log-lineares Modell) mar-

(a) Fälle

SNP2 \ SNP1	AA	AC	CC
AA	0,003	0,016	0,003
AC	0,152	0,032	0,045
CC	0,423	0,281	0,045

(b) Kontrollen

SNP2 \ SNP1	AA	AC	CC
AA	0,017	0,003	0,003
AC	0,116	0,162	0,023
CC	0,367	0,263	0,046

Tabelle 4.3: Genotypfrequenzen der 2-Markeranalyse für das beste Interaktionspaar (rs608139 und rs4678398) des Datensatzes zur Androgenetischen Alopezie.

ginale Effekte nicht berücksichtigt. Für die MC-Korrektur wurden die Daten mit 10.000 Permutationen simuliert. Für unser bestes Ergebnis wurde der korrigierter p-Wert von 0.0091 erhalten. Somit ist die Signifikanz durch die Korrektur für Strategie V bestätigt. Dieses Ergebnis muss natürlich noch in einer unabhängigen Replikation nachgewiesen werden, da alle vorgeschlagenen Strategien aus Abschnitt 3.5.10 durchgeführt wurden. Außerdem hat sich der gefundene Pathway nicht als unmittelbar plausibel für den betrachteten Phänotypen erwiesen.
Laut KEEG-Datenbank ist der Pathway für die Funktionalität der Tight Junction verantwortlich. Tight Junctions befinden sich in allen Zellverbänden und dienen zur Kommunikation mit den Zellen untereinander. Sie verhindern entweder die laterale Diffusion von anderen Membranen oder bilden eine mehr oder weniger durchlässige Barriere für den transepithelialen Transport [Shen et al., 2011]. Somit ist ein Zusammenhang von Haarausfall mit dem Pathway nicht direkt ersichtlich. Nach dem aktuellen Kenntnisstand wird Haarausfall nicht durch fehlende interzelluläre Kommunikation verursacht, sondern durch die Verschiebung des Wachstumszyklus, was eher auf hormonelle Störungen zurückgeführt werden kann. Trotzdem zeigt dieses Beispiel, dass es nützlich sein kann a-priori Information bei genomweiten Interaktionsanalysen zu verwenden. Bei einer kompletten GWIA von allen SNPs ($4,5 \cdot 10^{10}$ Tests) hätte man 56.250 SNP-Paare mit einem p-Wert kleiner als den von uns beobachteten besten Paar von Strategie V erwartet. Somit wäre dieses Interaktionspaar in einer konventionellen GWIA wahrscheinlich nie aufgefallen.

4.1.2 Interaktionsanalyse mit eQTLs

Im zweiten Anwendungsbeispiel wird ein Projekt beschrieben, welches in Zusammenarbeit mit der Arbeitsgruppe von Nancy Cox (Section of Genetic Medicine, University Chicago) im Rahmen eines Forschungsaufenthalts an der Universität Chicago bearbeitet wurde. Studien mit der SCAN Datenbank (SNP and Copy number ANnotation) [Gamazon et al., 2010] haben gezeigt, dass SNPs, die mit komplexen Krankheiten assoziiert sind, mit höherer Wahrscheinlichkeit eQTLs (expression Quantitative Trait Loci) sind als es ihre Repräsentation auf den Hochdurchsatzplattformen erwarten lässt. Bei eQTLs handelt es sich um Marker, die

Transkription oder Expression eines bestimmten Gens regulieren. Nach ihrer Lage werden sie in cis (lokal) und trans (global) unterteilt. In dieser Analyse werden cis-regulatorische eQTLs als SNPs definiert, die innerhalb von 4 MB vom Anfang und Ende des Gens liegen, dessen Expression gemessen wird. Alle übrigen SNPs sind trans-regulatorische eQTLs.
Um ein Verständnis für Interaktionen auf Basis des Transkriptoms zu entwickeln, könnte dieser Ansatz helfen, herauszufinden, ob Interaktion bei komplexen Krankheiten eine Rolle spielt. Somit ergibt sich die Hypothese, dass die eQTL-Interaktionsanalyse nützlich sein kann, signifikante, den Fall-Kontroll-Status betreffende Interaktionen zu finden, indem sie gute Kandidaten liefert. Bei den verwendeten HapMap-Daten [International HapMap Consortium, 2007] handelt es sich um Daten aus dem International HapMap-Projekt (siehe Abschnitt 1.3.5), welches sich zum Ziel gesetzt hat die Variationen der Haplotypen des menschlichen Genoms zu beschreiben. Für unsere Interaktionsanalyse wurden nur unabhängige Individuen von dem CEU (Caucasians of northern and western European ancestry from UT, USA d.h. Personen europäischer Herkunft) HapMap-Datensatz verwendet, da weder INTERSNP noch eine andere Software die Möglichkeit bietet, eine Interaktionsanalyse der quantitativen Traits in Eltern-Kind-Trios auf genomweiter Ebene durchzuführen. Somit gingen 30 Männer und 30 Frauen in die Analyse ein. Für die Analyse mit INTERSNP wurden mehr als 2 Millionen SNPs des Affymetrix GeneChips Human Exon 1.0 ST Array (mit einer MAF $> 5\%$, ohne Mendelfehler) eines original HapMap-Trio-Datensatzes von Lymphoblastzelllinien in CEU verwendet.

4.1.2.1 Analysestrategie

Bevor die Analysestrategie beschrieben wird, werden noch weitere Fakten zu der Studie genannt. Es handelt sich um eine Analyse von 10.104 Transkripten, über 2 Millionen SNPs und 60 Personen. Für jedes Transkript können theoretisch alle SNPs auf Interaktion getestet werden. Es geht also um 10.104 genomweite Interaktionsanalysen. Folglich wäre die Anzahl von Interaktionstests sehr groß. Um die rechnerischen Hürden zu überwinden, wurde die Anzahl der Tests mit Hilfe von a-priori Information reduziert. Es gingen nur Paare von cis und trans eQTLs in die Analyse ein, die einen marginalen Effekt ($p < 0,0001$) mit dem Transkript zeigten. Die Einzelmarkeranalyse wurde basierend auf den Familiendaten, also 90 Personen, mit dem QTDT (quantitative trait disequilibrium test) [Abecasis et al., 2000] durchgeführt und die Ergebnisse in der SCAN Datenbank gespeichert. Diese Einzelmarker-Informationen aus der SCAN Datenbank wurden verwendet, da INTERSNP nicht mit familienbasierten Daten rechnen kann. Der Expressionsdatensatz wurde mit der linearen Regression (siehe Abschnitt 3.3.2) analysiert (`TEST 5`).
Um valide Ergebnisse aus dieser eQTL-Interaktionsanalyse mit 10.104 Transkripten zu bekommen, war es wichtig eine sinnvolle Strategie für die Korrektur der p-Werte zu entwickeln. Im Folgenden wird von transkriptweiter und experimentweiter Signifikanz die Rede sein. Um den transkriptweiten korrigierten p-Wert zu ermitteln, multipliziert man den p-Wert mit der Anzahl der Interaktionstests, die für das jeweilige Transkript durchgeführt wurden. Für die experimentweite Signifikanz wird der p-Wert zusätzlich mit der Anzahl aller Transkripte korrigiert. Die Beurteilung der Signifikanz der Interaktions-p-Werte ist aufgrund zweier Phänomene schwierig: multiples Testen und Ungenauigkeit der asymptotischen Verteilung

der Interaktionsteststatistik aufgrund der kleinen Samplegröße. Beide Probleme können durch die p-Wert-Berechnung anhand von MC-Simulationen gelöst werden. Zu diesem Zweck wurde in INTERSNP eine MC-Prozedur implementiert, welche die bei den Personen beobachteten Ausprägungen pro Replikat über die Personen permutiert.
Es ist bekannt, dass Bonferroni-Korrektur mit der Anzahl der durchgeführten Tests zu einer konservativen Prozedur führt, wenn die SNPs im LD liegen. Studien haben gezeigt, dass n Tests mit SNPs im LD ungefähr $n/2$ unabhängigen Tests entsprechen, sowohl für Einzelmarker- [Gao et al., 2010] als auch Interaktionsanalysen [Becker et al., 2011]. Folglich werden p-Werte, die größer als $0,10$ nach Bonferroni-Korrektur sind, in der Regel über einem α-Level von $0,05$ nach der MC-basierten Korrektur für das Multiple Testen liegen. Aus diesem Grund wurde folgende Strategie beschlossen: Für jedes Transkript T wird der minimale Interaktions-p-Wert $minP(T)$ berechnet und die Bonferroni-Korrektur $bonfP(T)$ bestimmt, indem $minP(T)$ mit Anzahl der durchgeführten Interaktionsanalysen für T multipliziert wird. Wenn $bonfP(T)$ kleiner als $0,10$ ist, berechnen wir den „echten“ korrigierten p-Wert $p_{adj} = mcP(T)$ zu T, indem die Signifikanz von $minP(T)$ mit 999 Permutationsreplikaten berechnet wird. Auf diesem Weg berücksichtigt man das LD zwischen den SNPs, man entfernt aber auch gleichzeitig falsche Befunde, die durch die kleine Samplegröße verursacht sind.
Interaktionen sollen als transkriptweit signifikant angesehen werden, wenn $mcP(T)$ kleiner als $\alpha = 0,05$ ist. Da 10.104 Transkripte betrachtet werden, wird ein Signifikanzniveau von $\alpha_{exp} = 0,05/10.104 = 4,95 \cdot 10^{-6}$ für den transkriptweiten p-Wert $mcP(T)$ erwünscht, um zusätzlich Signifikanz über das ganze Experiment zu rechtfertigen. Um dieses Signifikanzniveau zu erreichen, wurden 10^7 Permutationsreplikate verwendet um $mcP(T)$ neu zu berechnen. Die 10^7 Replikate wurden nur auf solche Transkripte T angewendet, für welche $mcP(T)$ bei 999 Simulationen Null war. So kann davon ausgegangen werden, dass die Studie ohne den Verlust von wahren Funden analysiert wurde. Für einen $mcP(T)$ von $0,001$, der auf 999 Replikaten basiert, ist es extrem unwahrscheinlich ($p = 1,6 \cdot 10^{-44}$ gemäß Binomialtest), dass dieser kleiner ist als $4,95 \cdot 10^{-6}$, basierend auf 10^7 Replikaten. Weiterhin wurde festgelegt, dass alle Interaktionen mit einem unkorrigierten p-Wert kleiner als 10^{-4} potentiell interessante Befunde sind. Dies war auch das Kriterium für die Aufnahme in die Datenbank. Überrepräsentation von cis-trans oder trans-trans Interaktionen, also auch Überrepräsentation autoimmuner Gene (siehe Abschnitt 4.1.2.2), wurde auf der Basis dieser Befunde berechnet.

4.1.2.2 Ergebnisse der eQTL-Interaktionsanalyse

Insgesamt wurden $8,57 \cdot 10^8$ eQTL-Paare auf Interaktion getestet. Von diesen waren $1,4 \cdot 10^6$ cis-cis, $1,2 \cdot 10^7$ cis-trans und $8,43 \cdot 10^8$ trans-trans Paare. Zunächst wurde untersucht, ob es einen Überschuss an signifikanten Interaktionen in einer dieser Typklassen gibt. Auf dem nominalen α-Niveau von $1 \cdot 10^{-4}$ haben wir 225.779 signifikante trans-trans Interaktionen und 2.731 signifikante cis-trans Interaktionen identifiziert, was einem empirischen Niveau von $2,7 \cdot 10^{-4}$ bzw. $2,2 \cdot 10^{-4}$ entspricht. Beide Darstellungen reflektieren hohe Signifikanzabweichungen vom nominalen Niveau ($p < 10^{-300}$). Jedoch wäre es voreilig daraus zu schließen, dass wir einen Hinweis haben, dass es mehr signifikante Interaktionen gibt als per Zufall erwartet wird. Da die Samplegröße eher klein ist und da abhängig von den Allel-

frequenzen die Anzahl der Individuen, die zum Messen der Interaktion informativ sind, sehr klein sein können, kann der Überschuss an signifikanten Befunden teilweise durch fehlende Anpassung an die asymptotische Verteilung der Teststatistik bedingt sein.
Um die Relevanz unserer Befunde zu überprüfen, wurde der Monte-Carlo Simulationsansatz gewählt. Von den 10.104 Transkripten in dieser Studie wurden 363 Transkripte identifiziert, welche einheitlich mit autosomalen Genen mit mindestens einem SNP-Paar mit signifikanter eQTL-Interaktion ($P_{adj} <= 0,05$) korrespondieren. Zu beachten ist, dass diese Anzahl kleiner ist als die erwartete Anzahl von 505 nominal signifikanten Transkripten. Das hängt damit zusammen, dass wegen der kleinen Samplegröße die Monte-Carlo-Simulationsprozedur, welche einem exakten Test sehr nahe kommt, noch immer zu konservativ ist.

Die besten 25 Interaktionsergebnisse sind im Detail in Tabelle 4.1.2.2 dargestellt. Es handelt sich bei den Interatkionspaaren um trans-trans Paare, also Paare von SNPs, die weiter als 4MB vom Anfang oder Ende des Gens liegen, dessen Expression gemessen wurde oder auf einem anderen Chromosom. Es ist zunächst zu beachten, dass für kein Transkript die experimentweite Signifikanz ($p_{adj} <= 4,95 \cdot 10^{-6}$) erreicht werden konnte.

Nr	Transkript	Gen	Chr1	rs-Nr1	Pos1	Chr2	rs-Nr2	Pos2	p^a	$p_{MC}{}^b$
1	3535922	LOC730432, STYX	1	rs7575439	13820492	2	rs11588651	5597990	8,79E-08	5,00E-05
2	3829020	PDCD5	3	rs2243380	173636035	6	rs507759	117831072	7,30E-25	5,00E-05
3	3250373	TSPAN15	9	rs10988411	73111699	9	rs1927938	131326648	1,77E-17	1,00E-04
4	2563785	HLA-C	2	rs734504	173603531	14	rs2305462	91968561	2,69E-09	2,00E-04
5	3296046	KCNMA1	1	rs10871652	8146523	18	rs7521313	65514804	8,18E-11	3,00E-04
6	3127775	TNFRSF10A	3	rs912521	69678007	13	rs6778947	28059336	1,87E-12	3,00E-04
7	2900059	HIST1H2BM	12	rs4077639	8274582	16	rs2965689	85051436	9,01E-08	3,00E-04
8	2654394	FXR1	4	rs2243380	130523962	6	rs7679511	117831072	7,47E-12	4,00E-04
9	2764054	SEPSECS	6	rs11013291	55927036	10	rs4342427	23440197	9,22E-09	4,00E-04
10	3842315	ZNF580	7	rs11631328	43446724	15	rs2240984	34314192	4,01E-09	4,00E-04
11	3340697	UVRAG	14	rs2267361	56324057	22	rs1483113	35430206	1,81E-08	4,00E-04
12	3696016	CTRL, PSMB10	1	rs4953469	222402409	2	rs11579031	47232898	2,82E-10	1,00E-03
13	2705706	TNFSF10	1	rs17218936	212499946	3	rs12131852	41994396	3,96E-08	1,00E-03
14	3456732	ITGA5	2	rs4731689	173570788	7	rs13000260	129797075	3,65E-09	1,00E-03
15	3908358	SULF2	2	rs2235076	59450732	6	rs2192574	102622953	1,20E-08	1,00E-03
16	3329247	DGKZ	3	rs10501885	161011181	11	rs13085339	97757171	1,49E-09	1,00E-03
17	3289235	SGMS1	3	rs944689	173676326	9	rs6445063	100406359	4,15E-09	1,00E-03
18	3937943	BCR, FLJ42953	4	rs10514419	189229434	16	rs10003920	76107378	1,39E-08	1,00E-03
19	3512769	ZC3H13	4	rs13216409	130523962	6	rs7679511	117800529	4,99E-10	1,00E-03
20	3398241	ZBTB44	5	rs2011869	45975431	22	rs7724350	37810643	1,67E-06	1,00E-03
21	2732611	MRPL1	6	rs11013291	55944009	10	rs2745739	23440197	4,98E-07	1,00E-03
22	3724858	TBX21	6	rs1881730	99926277	10	rs4574651	80177785	1,57E-07	1,00E-03
23	3645338	PRSS21	7	rs13245386	43451773	7	rs6960889	43479536	3,16E-12	1,00E-03
24	2884301	IL12B	9	rs7138427	23120554	12	rs10965654	6707113	6,82E-14	1,00E-03
25	3631498	LARP6	10	rs11637113	59033936	15	rs7901219	63411765	1,94E-09	1,00E-03

Tabelle 4.4: Die besten 25 Ergebnisse der eQTL-Interaktionsanalyse. Diese eQTL-Paare waren alle vom Typ trans-trans.
[a] unkorrigierter p-Wert
[b] MC-korrigierter p-Wert

Um mehr über die gefundenen Gene zu erfahren, wurden die Interaktionsergebnisse mit den bereits veröffentlichten Ergebnissen der GWAS-Datenbank [Hindorff et al., 2011] verglichen. Für über die Hälfte der 25 besten Ergebnisse, deren Transkript einem oder mehreren Genen zugewiesen wurde, findet man in der Datenbank Informationen zu GWAS-Publikationen. Tabelle 4.1.2.2 zeigt neben den Interaktions-p-Werten unserer Analyse und den gefundenen Genen auch Publikationen und Krankheiten, in welchen die gefunden Gene involviert sind. Da Lymphozyten als Proxygewebe für Autoimmunerkrankungen gelten, ist es sehr interessant, dass einige unserer gefundenen Gene mit Krankheiten assoziiert sind, bei denen das Immunsystem eine Rolle spielt.
Autoimmunkrankheiten führen im Allgemeinen zu einer Abwehrreaktion des Immunsystems gegen körpereigenes Gewebe. Von daher sind die gefundenen Interaktionen bezüglich der Lymphozytenexpression in der Tat sehr vielversprechende Kandidaten für die Interaktionsanalyse auf Krankheitsebene. Die gefundenen Gene HLA-C und IL12B, die jeweils für humane Leukozytenantigene und Interleukin kodieren, sind nachweislich für die Regulierung von Immunreaktionen zuständig, was durch die zahlreichen hochrangigen Publikationen in der GWAS-Datenbank belegt wird. Das Antigen HLA-C ist ein Molekül des Haupthistokompatibilitätskomplexes, dessen Aufgabe die Präsentation von Antigenen ist. Interleukin 12 gehört zur Gruppe der Cytokine, welche sich wie Wachstumsfaktoren verhalten und T-Zellen und natürliche Killerzellen (NK-Zellen) aktivieren können [Safran et al., 2010]. Beide Gene sind hauptsächlich an Autoimmunkrankheiten wie Psoriasis, Vitiligo und Morbus Crohn beteiligt. Bei der Psoriasis (Schuppenflechte) handelt es sich um eine autoimmune Hautkrankheit. Hüffmeier et al. [2010] haben bei ihrer GWAS, bei welcher unsere Arbeitsgruppe mitgewirkt hat, für den Phänotypen Psoriasis Arthritis (Schuppenflechte und Gelenkrheuma) ebenfalls HLA-C und IF12B als auf Krankheitsebene assoziierte Gene beschrieben. Der nächste Schritt ist nun, unsere gefunden Interaktions-eQTL-SNPs für die Gene HLA-C und IF12B in dem Fall-Kontroll-Datensatz von Hüffmeier et al. auf Interaktion zu testen.
Neben Psoriasis wurden die beiden Gene mit einer weiteren Hautkrankheit, Vitiligo, in Verbindung gebracht. Die Ursache von Vitiligo (Weißfleckenkrankheit) ist eine Pigmentstörung. Weitere Krankheiten im Zusammenhang mit der Immunabwehr sind HIV1 (Humanes Immundefizienz-Virus), welches zum Ausbruch von AIDS (Acquired Immunodeficiency Syndrome) führen kann und Morbus Crohn, eine chronisch entzündliche Darmerkrankung. Neben den genannten Genen wurden Interaktionen für weitere Gene gefunden, die auf den ersten Blick nicht mit dem Immunsystem oder den Lymphozyten in Verbindung gebracht werden können, jedoch können auch diese Interaktionen grundsätzlich wertvolle Hinweise auf bis jetzt unbekannte Zusammenhänge liefern.

Nr	Transkript	p-Wert	p_{MC}	Gene	Erstautor	Publikation	Krankheit/ Trait	Region
1	3535922	8,79E-08	5,00E-05	LOC730432, STYX	-	-	-	-
2	3829020	7,30E-25	5,00E-05	PDCD5	-	-	-	-
3	3250373	1,77E-17	1,00E-04	TSPAN15	-	-	-	-
4	2563785	2,69E-09	2,00E-04	HLA-C	The International HIV Controllers Study	Science. 2010 Dec 10;330(6010):1551-7	HIV-1 control	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	E. Ellinghaus	Nat Genet. 2010 Nov;42(11):991-5	Psoriasis	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	Genetic Analysis of Psoriasis Consortium; the Wellcome Trust Case Control Consortium	Nat Genet. 2010 Nov;42(11):985-90	Psoriasis	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	C. Quan	Nat Genet. 2010 Jul;42(7):614-8	Vitiligo	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	J. Bei	Nat Genet. 2010 Jul;42(7):599-603	Nasopharyngeal carcinoma	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	J. Fellay	PLoS Genet. 2009 Dec;5(12):e1000791	HIV-1 control	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	R. Nair	Nat Genet. 2009 Feb;41(2):199-204	Psoriasis	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	S. Limou	J Infect Dis. 2009 Feb 1;199(3):419-26	AIDS progression	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	D. Melzer	PLoS Genet. 2008 May 9;4(5)	Protein quantitative trait loci	6p21.33
4	2563785	2,69E-09	2,00E-04	HLA-C	Y. Liu	PLoS Genet. 2008 Mar 28;4(3)	Psoriasis	6p21.33

4	2563785	2,69E-09	2,00E-04	HLA-C	F. Capon	Hum Mol Genet. 2008 Jul 1;17(13):1938-45	Psoriasis	6p21.33
5	3296046	8,18E-11	3,00E-04	KCNMA1	A. Morrison	Circ Cardiovasc Genet. 2010 Jun 1;3(3):248-55	Mortality among heart failure patients	10q22.3
6	3127775	1,87E-12	3,00E-04	TNFRSF10A	-	-	-	-
7	2900059	9,01E-08	3,00E-04	HIST1H2BM	-	-	-	-
8	2654394	7,47E-12	4,00E-04	FXR1	International Schizophrenia Consortium	Nature. 2009 Aug 6;460(7256):748-52	Schizophrenia	3q26.33
9	2764054	9,22E-09	4,00E-04	SEPSECS	-	-	-	-
10	3842315	4,01E-09	4,00E-04	ZNF580	-	-	-	-
11	3340697	1,81E-08	4,00E-04	UVRAG	-	-	-	-
12	3696016	2,82E-10	1,00E-03	CTRL, PSMB10	-	-	-	-
13	2705706	3,96E-08	1,00E-03	TNFSF10	-	-	-	-
14	3456732	3,65E-09	1,00E-03	ITGA5	-	-	-	-
15	3908358	1,20E-08	1,00E-03	SULF2	J. Lasky-Su	Am J Med Genet B Neuropsychiatr Genet. 2008 Dec 5;147B(8)	Attention deficit hyperactivity disorder	20q13.13
16	3329247	1,49E-09	1,00E-03	DGKZ	R. Ferreira	Nat Genet. 2010 Sep;42(9):777-80	Immunoglobulin A	11p11.2
17	3289235	4,15E-09	1,00E-03	SGMS1	-	-	-	-
18	3937943	1,39E-08	1,00E-03	BCR, FLJ42953, LOC728468	D. Gudbjartsson	Nat Genet. 2008 May;40(5):609-15	Height	22q11.23
19	3512769	4,99E-10	1,00E-03	ZC3H13	-	-	-	-
20	3398241	1,67E-06	1,00E-03	ZBTB44	-	-	-	-
21	2732611	4,98E-07	1,00E-03	MRPL1	-	-	-	-
22	3724858	1,57E-07	1,00E-03	TBX21	-	-	-	-

23	3645338	3,16E-12	1,00E-03	PRSS21	-	-	-	-
24	2884301	6,82E-14	1,00E-03	IL12B	E. Ellinghaus	Nat Genet. 2010 Nov;42(11):991-5	Psoriasis	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	Genetic Analysis of Psoriasis Consortium; the Wellcome Trust Case Control Consortium	Nat Genet. 2010 Nov;42(11):985-90	Psoriasis	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	U. Hüffmeier	Nat Genet. 2010 Nov;42(11):996-9	Psoriatic arthritis	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	D. McGovern	Hum Mol Genet. 2010 Sep 1;19(17):3468-76	Crohn's disease	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	R. Nair	Nat Genet. 2009 Feb;41(2):199-204	Psoriasis	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	X. Zhang	Nat Genet. 2009 Feb;41(2):205-10	Psoriasis	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	J. Barrett	Nat Genet. 2008 Aug;40(8):955-62	Crohn's disease	5q33.3
24	2884301	6,82E-14	1,00E-03	IL12B	M. Parkes	Nat Genet. 2007 Jul;39(7):830-2	Crohn's disease	5q33.3
25	3631498	1,94E-09	1,00E-03	LARP6	-	-	-	-

Tabelle 4.5: Veröffentlichungen aus der GWAS-Datenbank [Hindorff et al., 2011] zu den besten 25 Ergebnissen der eQTL-Interaktionsanalyse.

Die Ergebnisse unserer eQTL-Interaktionsanalyse werden in der Datenbank SCAN zugänglich gemacht.

4.1.3 Bipolare Störungen

Als letztes Anwendungsbeispiel wird ein GWAS-Datensatz vorgestellt, welcher sehr interessante GWAS-Ergebnisse bei bipolaren Störungen geliefert hat [Cichon et al., 2011]. Bipolare Störungen sind schwere psychische Störungen, die sich durch eine stark erhöhte Suizidrate der Betroffenen auszeichnet. Patienten leiden unter extremen, in Phasen verlaufenden Stimmungsschwankungen, die zwischen übersteigertem Glücksgefühl und Depression wechseln. Auch in der vorliegenden Fall-Kontroll-Studie ist aufgrund des großen Datensatzes die genomweite Interaktionsanalyse eine rechnerische Herausforderung. In die Analyse gingen 1.158 Fälle (DSM-IV Diagnose von bipolarer Störung nach Diagnosekriterien für Geistesstörungen) und 2.172 Kontrollen ein. Sowohl die Fälle auch als auch die Kontrollen sind deutscher Abstammung. Die ca. 500.000 SNPs wurden auf einem Illumina BeadArray (Illumina, San Diego, USA) genotypisiert.
Nach der Qualitätskontrolle mit PLINK reduzierte sich die Anzahl der SNPs auf 473.227. Um sowohl bei der Einzelmarkeranalyse als auch bei der Multimarkeranalyse für Stratifikation zu korrigieren, wurden Kovariaten unter Verwendung der Multidimensionalen Skalierung (MDS) berechnet. Die SNP-Paare mit Randeffekten wurden dann mit einem logistischen Regressionsmodell unter genotypischer Interaktion (Test 4, volles Modell mit 8 FG) getestet. Das Signifikanzniveau für genomweite Signifikanz wurde auf $0,05$ festgelegt. Somit ergibt sich für unsere Analyse (500K Chip) eine unkorrigierte Signifikanzschwelle von $1 \cdot 10^{-12}$. Das Interaktionspaar rs912607 und rs6590281 erreichte einen p-Wert von $p = 9,49 \cdot 10^{-13}$ und kann somit als genomweit signifikant betrachtet werden. Beide SNPs zeigten in der Einzelmarkeranalyse moderate Ergebnisse ($p1 = 4,22 \cdot 10^{-4}$, $p2 = 1,19 \cdot 10^{-3}$). Die stärkste Effektgröße mit einer Odds Ratio von 8,2 [3,59-18,74] wurde für den doppelt homozygoten Genotyp (AA,CC) (siehe Tabelle 4.7) mit einer Frequenz von 0,26% in den Fällen und 0,003% in den Kontrollen gemessen.

Nr	Chr1	rs-Nr1	Gen1	p1[a]	Chr2	rs-Nr2	Gen2	p2[b]	p-Wert[c]
1	13	rs912607	B3GALTL (Intron)	4,22E-04	11	rs6590281	-	1,19E-03	9,49E-13
2	13	rs912607	B3GALTL (Intron)	4,22E-04	11	rs4937269	-	1,32E-03	7,68E-12
3	9	rs6479458	-	1,88E-03	10	rs944963	EBF3 (Intron)	1,81E-02	4,33E-11
4	4	rs1435442	-	2,10E-04	2	rs10184538	-	1,17E-01	5,46E-11
5	19	rs1064395	NCAN (3'UTR)	3,02E-06	17	rs10853029	BCAS3 (Intron)	1,30E-03	6,14E-11
6	14	rs12888576	-	1,45E-04	9	rs2027963	SARDH (Intron)	1,62E-04	7,39E-11
7	9	rs2027963	SARDH (Intron)	1,62E-04	3	rs17057445	-	4,22E-04	1,06E-10
8	17	rs2074404	WNT3 (Intron)	4,84E-04	7	rs12537171	-	1,55E-02	1,92E-10
9	5	rs17826588	ADCY2 (Intron)	2,79E-06	4	rs1435442	-	2,10E-04	1,98E-10
10	4	rs7698262	-	3,77E-04	8	rs2599676	ZMAT4 (Intron)	1,66E-01	1,98E-10

Tabelle 4.6: Die besten Ergebnisse der Interaktionsanalyse in einer deutschen Bevölkerungsgruppe mit bipolarer Störung.
[a] p-Wert logistische Regression SNP 1.
[b] p-Wert logistische Regression SNP 2.
[c] p-Wert Interaktion (unkorrigiert).

(a) Fälle

SNP2 \ SNP1	AA	AC	CC
AA	0,004	0,056	0,074
AC	0,016	0,134	0,312
CC	0,026	0,144	0,233

(b) Kontrollen

SNP2 \ SNP1	AA	AC	CC
AA	0,011	0,051	0,103
AC	0,021	0,152	0,318
CC	0,003	0,086	0,255

Tabelle 4.7: $2 \times 3 \times 3$-Feldertafel für Fälle und Kontrollen des Datensatzes zur bipolaren Störung.

Der erste SNP rs912607 aus dem besten Paar liegt im Gen B3GALTL (Chromosom 13q12) während der zweite SNP rs6590281 in einer Inter-Genregion (Chromosom 11q24) liegt. Das Produkt von B3GALTL ist eine Beta-1,3-Glukosyltransferase, die in der Chemie der komplexen Zuckerverbindungen (Glykane) eine Rolle spielt [Safran et al., 2010]. Glykane sind Bestandteile vieler Proteine und anderer Bausteine von Organismen. Defekte in diesem Gen verursachen unter anderem das Peters-Plus-Syndrom sowie autosomal rezessive Syndrome mit verschiedenen Symptomen, welche auch psychomotorísche Retardation beinhalten. Der SNP rs912607 in Gen B3GALTL ist auch im zweitbesten signifikanten Paar involviert und interagiert dabei mit SNP rs4937269 (Chromosom 11q24), der im LD mit SNP rs6590281 des besten Interaktionspaars steht. Auch hier sollte das beste GWIA Ergebnis in einer Replikation bestätigt werden. Aus diesem Grund werden zur Zeit die besten Ergebnisse in großen unabhängigen Datensätzen zur bipolaren Störung aus Europa, USA und Australien weiter verfolgt.

4.2 Laufzeittabellen

Für die Laufzeittabelle wurde der Datensatz des dritten Anwendungsbeispiels zu bipolaren Störungen verwendet. Der Datensatz besteht aus 1.158 Fällen und 2.172 Kontrollen und 473.227 SNPs nach der Qualitätskontrolle. Die Analyse wurde auf einem IBM-Hochleistungsrechner (High Performance Computer Cluster, HPC-Cluster) durchgeführt, der aus 34 Knoten (8 × Blade LS42 und 26 × Blade HS22) besteht. Jedem Knoten stehen durch Hyper-Threading 24 logische Prozessoren zur Verfügung, wobei ein Prozessor zur Sicherheit für interne Prozesse freigelassen wird. Die Laufzeiten wurden auf einem Blade HS22 Knoten mit 53GB RAM (Arbeitsspeicher, engl. Random-Access-Memory), 146 GB HDD (Festplattenlaufwerk, engl. hard disk drive = HDD) und 2 × SixCore Intel(R) Xeon(R) CPU X5650 mit 2.67GHz ermittelt. Zunächst wurde die serielle INTERSNP-Version getestet, also nur ein Prozessor pro Analyse, anschließend die parallelisierte Version von INTERSNP, bei der 12 Prozessoren für die Analyse verwendet wurden (es wurden nur die 12 physikalischen Prozessoren benutzt). In Tabelle 4.8 wird `Test 2`, das log-lineare Modell, mit der logistischen Regression (`Test 5`) mit und ohne

Pre-Test verglichen.

Strategie	Test	Modell	FG	Anzahl Tests	Laufzeit[a]	Laufzeit[b]
I	2	log-linear	4	4,73E+06	14m44s	8m57s
	5	Pre-Test1/ lineare Regression	1	4,73E+06	20m41s	14m40s
	5	lineare Regression	1	4,73E+06	46m16s	39m59s
II	2	log-linear	4	5,00E+05	0m46s	0m42s
	5	Pre-Test1/ lineare Regression	1	5,00E+05	2m40s	2m56s
	5	lineare Regression	1	5,00E+05	5m29s	3m29s
III	2	log-linear	4	3,58E+05	2m11s	1m1s
	5	Pre-Test1/ lineare Regression	1	3,58E+05	3m44s	3m47s
	5	lineare Regression	1	3,58E+05	6m40s	3m25s
IV	2	log-linear	4	1,40E+07	41m33s	2m38s
	5	Pre-Test1/ lineare Regression	1	1,40E+07	58m30s	6m40s
	5	lineare Regression	1	1,40E+07	139m27s	15m20s
V	2	log-linear	4	1,29E+06	7m14s	6m12s
	5	Pre-Test1/ lineare Regression	1	1,29E+06	10m6s	9m20s
	5	lineare Regression	1	1,29E+06	17m15s	13m23s
Genomweite Analyse	2	log-linear	4	1,12E+11	144582m (100t1h36m)	12500m (8t16h20m)
	5	Pre-Test1/ lineare Regression	1	1,12E+11	213605m (148t8h5m)	18487m (12t20h7m)

Tabelle 4.8: Laufzeittabelle: Parallelisierung der Multimarker-Analyse mit OpenMP. Das Einlesen der Daten und die Einzelmarkeranalyse dauerten zusätzlich ca. 14min.
[a] Unter Verwendung der seriellen Version von INTERNP.
[b] Unter Verwendung der parallelisierten Version von INTERNSP mit 12 Prozessoren.

Die Strategien I,II, III, und V lassen sich in weniger als einer Stunde ohne Parallelisierung durchführen. Bei Strategie IV ist die Anzahl der Tests deutlich höher und deshalb dauert diese Analyse entsprechend länger. Bei allen Strategien ist das log-lineare Modell am schnellsten, gefolgt von der linearen Regression mit Pre-test. Die Verwendung des Pre-test ist also sinnvoll, um die Rechenzeit bei der aufwendigeren linearen Regression zu reduzieren. Besonders deutlich wird das bei Strategie IV. Die lineare Regression dauert ca. 150 min, die lineare Regression mit Pre-Test dagegen nur ca. 72 min. Somit ist die Analyse mit Pre-Test doppelt so schnell, was sich bei noch größeren Datensätzen noch stärker bemerkbar macht. Benutzt man die parallelisierte Version mit 12-facher Parallelisierung, so bleiben die Laufzeiten für alle Analysen unter 30 min. Je größer die Anzahl der Tests, desto sinnvoller ist die Parallelisierung und desto deutlicher die Unterschiede zwischen dem seriellen und parallelisierten Ergebnis. Zu beachten ist, dass bereits die Einzelmarkeranalyse zusätzlich ca. 14 min benötigt, da zuerst alle Eingabedateien seriell eingelesen werden müssen.
Am deutlichsten lässt sich der Vorteil der Parallelisierung bei der kompletten

GWIA mit $1,12^{11}$ Tests erkennen. Die serielle Version benötigt mit dem log-linearen Modell ca. 100 Tage, die parallelisierte hingegen nur etwas über eine Woche (8 Tage 16 Stunden) für die Analyse mit 12-facher Parallelisierung. Somit ergibt sich eine Verbesserung um den Faktor 11,5. Dieser Faktor wird durch die Analyse der linearen Regression mit Pre-test bestätigt. Die Laufzeit bei dieser GWIA beträgt beim seriellen Programm 148 Tage und 8 Stunden und mit der parallelisierten Version nur 12 Tage und 20 Stunden. Dies zeigt wiederum, dass die Parallelisierung relativ effizient ist und durch die Verteilung der einzelnen Prozesse nur wenig Zeit verloren geht.

Kapitel 5

Diskussion

5.1 Die Rolle von INTERSNP in der aktuellen Forschung

In den letzten Jahren wurden mit Hilfe der GWAS Hunderte von Loci gefunden, die mit komplexen Krankheiten assoziiert sind. Jedoch bleibt weiterhin ein großer Teil der Heritabilität ungeklärt. Mögliche Schritte zur Schließung dieser Lücke sind unter anderem Multimarkeranalysen, die nicht nur einen SNP, sondern mehrere SNPs gleichzeitig betrachten. Dazu gehören Haplotypanalysen, Pathwayassoziationsanalysen und Interaktionsanalysen. Viele bedeutende Forscher, beispielsweise Sarah Tiskoff (University of Pennsylvania) nennen unter anderem die Untersuchung der Gen-Gen Interaktion für eine aussichtsreiche, wichtige Strategie für die kommenden Jahre [Heard et al., 2010].

In der vorliegenden Arbeit liegt der Schwerpunkt ebenfalls auf den genomweiten Interaktionsanalysen. Auch Cordell [2009] weist in ihrem Nature Review darauf hin, dass die Interaktionsanalyse ein sinnvoller Ansatz sein könnte, um neue Informationen über biologische und biochemische Pathways zu bekommen und die Power von GWAS-Studien zu verbessern. Alle von ihr beschriebenen Herangehensweisen wie logistische/lineare Regression, Bayes-Modelle oder rekursive Partitionierung stoßen jedoch an ihre Grenzen, wenn große GWAS-Datensätze ($>$ 1000 Fälle, $>$ 1000 Kontrollen mit $>$ 300.000 SNPs) analysiert werden sollen. Als Lösung schlägt Cordell vor, die Software zu parallelisieren oder eine Vorauswahl der Daten zu treffen. Die Selektion der Daten könnte beispielsweise anhand von marginalen Effekten oder mit Hilfe von biologischen Informationen erfolgen. Die Interaktionsidee wurde in den letzten Jahren von einigen Forschergruppen weiterverfolgt ([Schüpbach et al., 2010],[Kam-Thong et al., 2010], Wan et al. [2010]), die neue Ansätze zur Datenselektion entwickelten. Das Hauptproblem bei all diesen Methoden bleibt jedoch die riesige Anzahl von SNP-Kombinationen, wenn man eine genomweite Analyse durchführen möchte. Diese Vielzahl von Tests hat zur Folge, dass die Analyse mit seriellen Programmen auf einem normalen Desktopcomputer Monate dauern würde, aber auch Hochleistungsrechner Wochen brauchen. Utopisch werden die Rechenzeiten, wenn man mehr als zwei Marker gleichzeitig betrachten will. Aus diesem Grund haben die verschiedenen Software-Pakete, die es neben INTERSNP gibt, neue Ansätze gefunden, um die Anzahl der Tests zu reduzieren oder die Daten so aufzuarbeiten, dass schnellere Rechenoperationen möglich sind. Das Ziel von INTERSNP ist die Anzahl der SNP-Kombinationen mit Hilfe von a-priori

Information zu reduzieren, was zeitgleich Cordell [2009] als mögliche Lösung vorgeschlagen hat. Dazu können statistische und/oder genetische Kriterien verwendet werden. Auch Pathwayinformationen können als Filter benutzt werden. Zusätzlich wurden in INTERSNP „Pre-Tests“ implementiert und eine parallelisierte Version erstellt. Auch in der Literatur lassen sich einige Ansätze zur Reduzierung der Anzahl der Tests finden. Die wichtigsten Filterkriterien stellt Ritchie [2011] in ihrer Veröffentlichung dar. Sie unterteilt die Filter in drei Kategorien: Statistischer Nachweis von Einzelmarkereffekten, intrinsisches Wissen und extrinsisches biologisches Wissen. INTERSNP wird in dieser Veröffentlichung als Beispiel für einen entsprechend umfassenden Ansatz vorgestellt. Ansätzen wie INTERSNP wird ein großes Potential zugesprochen Ritchie [2011], da sie Informationen aus verschiedenen Quellen vereinen. Jedoch weist Ritchie auch darauf hin, dass bei solchen Informationen Vorsicht geboten werden muss, da das biologische Wissen noch nicht vollständig sei und somit auch Fehler beinhalten könnte.
Neben der Verwendung von Filterkriterien ist es sinnvoll, die Software zu parallelisieren und/oder neue Hardwareansätze, wie beispielsweise das Rechnen auf Grafikkarten [Kam-Thong et al., 2010], einzusetzen.

Um zu zeigen, dass INTERSNP durchaus im Forschungsalltag von Bedeutung ist, führen wir ein kleines Experiment durch. Gibt man bei Google „INTERSNP“ als Suchkriterium ein, erhält man 2.930 Ergebnisse (Stand: 03.05.2011). Natürlich beziehen sich nicht alle Beiträge auf die Software INTERSNP, aber die Anzahl ist doch beachtlich. Ergänzt man das Suchkriterium um „Herold“, erhält man 574 Einträge, die nun ziemlich sicher etwas mit der Software INTERSNP zu tun haben. Natürlich sind viele Seiten redundant, aber man bekommt durch diese Suche einen Eindruck, wie schnell sich die Software im Internet verbreitet hat. Verwendet man Google Scholar und sucht nach der Veröffentlichung zu INTERSNP [Herold et al., 2009], so findet man 16 Zitate. Allerdings wurde die Publikation nur 13-mal in anderen Veröffentlichungen wirklich zitiert, da auch hier wieder redundante Veröffentlichungen aufgelistet werden. Beachtlich ist jedoch, dass 11 Veröffentlichungen aus von uns unabhängigen Arbeitsgruppen stammen. In drei davon wird INTERSNP mit anderer Software verglichen und vier nennen INTERSNP in Verbindung mit genomweiter Interaktionsanalyse. Auch aus diesen Zahlen lässt sich erkennen, dass es sich bei INTERSNP um eine bekannte und etablierte Software handelt, die sich in der Forschungsgemeinschaft bereits kurze Zeit nach der Veröffentlichung verbreitet hat.

Trotz der rechnerischen Hürden, scheint die genomweite Interaktionsanalyse immer mehr an Bedeutung zu gewinnen. Immer mehr Forschergruppen suchen weiterhin nach Ideen die GWIA möglichst effizient zu realisieren. Die zunehmende Popularität der GWIA wird auch dadurch ersichtlich, dass im Januar 2011 beispielsweise eine ganze Ausgabe des Annals of human Genetics diesem Thema gewidmet wurde. Damit INTERSNP für die Zukunft gerüstet ist, werden die im folgenden Abschnitt möglichen Verbesserungen unserer Software vorgestellt. Es ist davon auszugehen, dass die Interaktionsanalyse noch interessante Erkenntnisse und neue Loci, die in Zusammenhang mit Krankheiten stehen, liefern wird. Unsere Anwendungsbeispiele haben bereits gezeigt, dass Loci mithilfe der genomweiten Interaktionsanalyse mit a-priori Information, gefunden werden können, welche bei der Einzelmarkeranalyse nie aufgefallen wären.

5.2 Geplante Verbesserungen und Erweiterungen

5.2.1 Parallelisierung mit MPI

Wie in Abschnitt 3.4.3 beschrieben, wurden Teile von INTERSNP mit OpenMP [OpenMP, 2008] parallelisiert. Bei der genomweiten Analyse konnte eine deutliche Verbesserung in der Laufzeit beobachtet werden. Durch die Parallelisierung mit OpenMP ist es möglich, eine komplette GWIA in ca. einer Woche durchzuführen. Eine noch bessere Performance könnte dadurch erreicht werden, wenn neben OpenMP mit Message Passing Interface (MPI) [MPI, 2009] parallelisiert wird, was jedoch mit großem Aufwand verbunden wäre. MPI ist ein Standard, der den Nachrichtenaustausch bei parallelen Berechnungen auf Multicomputern (verteilter Speicher) beschreibt. Eine MPI-Anwendung erleichtert die Kommunikation zwischen Prozessen, die auf die zur Verfügung stehende Anzahl von Knoten aufgeteilt werden, um eine gemeinsame Aufgabe zu lösen. Die einzelnen Prozesse verständigen sich via Message Passing, d.h. beim Datenaustausch werden Nachrichten von einem zum anderen Prozess geschickt. Ein Vorteil dieser Technologie ist, dass der Nachrichtenaustausch auch über Rechnergrenzen hinweg funktioniert. Der Unterschied zu OpenMP besteht darin, dass die Prozesse nicht nur auf einen gemeinsamen Speicher zugreifen und darüber Daten austauschen, sondern auch auf einem System oder mehreren Systemen mit verteilten Speichern laufen können. Die gemeinsame Verwendung von OpenMP und MPI wird „hybride Parallelisierung“ genannt.
Um zusätzlich MPI in INTERSNP zu verwenden, müssten größere Teile des Programmcodes umgeschrieben und der Aufbau des Programms müsste angepasst werden. Da in Zukunft die Datensätze aufgrund der Next-Generation-Sequencing-Technologie immer größer werden, ist dieser Aufwand sicher sinnvoll und auch notwendig. Außerdem können durch eine „hybride Parallelisierung“ die zu Verfügung stehenden Hochleistungsrechner optimal ausgenutzt werden. Als Beispiel ist die Erweiterung der Epistasis-Option in PLINK (FastEpistasis) [Schüpbach et al., 2010] zu nennen. FastEpistasis wurde gegenüber der ursprünglichen Version dadurch verbessert, dass neben einer prozessorbasierten Parallelisierung in diesem Fall SMP (Symmetric Multiprocessing) auch clusterbasierte Parallelisierung mit MPI erfolgt ist.

5.2.2 Dosage data

Des Weiteren wäre es sinnvoll, „dosage data“ als Eingabe zu erlauben. Bei „dosage data“ liegen keine kompletten Genotypen vor, sondern Wahrscheinlichkeitsgewichte für die drei möglichen Genotypen für jede Person. Solche Daten entstehen unter anderem beim Imputing [Becker et al., 2009]. Imputation bedeutet im Allgemeinen, dass mit statistischen Verfahren fehlende Daten vervollständigt werden. Bei der Beschreibung von SNP-Chip-Datensätzen bedeutet dies, dass SNPs, die man analysieren möchte, nicht genotypisiert, sondern mit Referenzdaten rekonstruiert werden. Für die Imputation gibt es verschiedene Arten von Software, welche sich in zwei Gruppen einteilen lassen. Die eine Gruppe arbeitet mit Hidden Markov Modellen (HMM) wie z.B. IMPUTE [Marchini et al., 2007], die andere nutzt Haplotypfrequenzschätzung mit dem EM-Algorithmus wie z.B. FAMHAP [Becker and Knapp, 2004] oder PLINK [Purcell et al., 2007]. Zu Beginn der Imputation wird eine Referenzdatei festgelegt, beispielsweise HapMap-Daten und/oder

1000 Genomes-Daten, also ein großes Set von genotypisierten SNPs. Dabei sollten die Fälle und Kontrollen des Referenzsamples aus der gleichen Bevölkerungsgruppe stammen. Imputing kann sinnvoll sein, wenn beispielsweise die Daten zweier Gruppen auf verschiedenen SNP-Panels genotypisiert worden sind. Durch die Imputation könnten dann fehlende SNPs geschätzt werden. Imputing ist ein sehr zeitintensives Verfahren, welches sich als durchaus zuverlässig herausgestellt hat und dadurch in letzter Zeit tendenziell an Bedeutung gewinnt [Becker et al., 2009]. Aus diesem Grund ist es sicher sinnvoll, INTERSNP so anzupassen, dass die Ausgabedateien der Imputation eingelesen und analysiert werden können. Da nicht mehr mit eindeutigen Genotypen gerechnet werden kann, müssen die statistischen Tests so modifiziert werden, dass die Unsicherheit, mit der „dosage data“ behaftet ist, adäquat adressiert wird. Dieser Aufwand wird sich aller Voraussicht nach lohnen, da bis jetzt nur einige wenige Programme mit Wahrscheinlichkeitsgewichten arbeiten können und davon keines mit „dosage data“ genomweite Interaktionen berechnen kann.

5.2.3 Bitoperatoren

Ein weiterer Ansatz wäre die Genotypen in binärer (0/1)-Kodierung abzuspeichern und das Auszählen dieser Genotypen über Bit-Operationen durchzuführen. Diese Idee wurde bereits von Wan et al. [2010] in ihrer Software BOOST umgesetzt und hat dadurch die Analyse mit dem log-linearen Modell erheblich beschleunigt, da das Auszählen der Genotyptabellen der zeitkritische Faktor ist. Für Regressionsmodelle ist dieser Teil der Analyse jedoch nicht entscheidend, d.h. der Vorteil der Bitoperatoren ist nur für das log-lineare Modell von Bedeutung. Die Änderung der Datenstruktur in dieser Weise würde sicher auch die Laufzeiten von INTERSNP für das log-lineare Modell um einen beachtlichen Faktor verbessern. Das Prinzip der (0/1)-Kodierung der Genotypen wird anhand des Beispiels 5.1 gezeigt. Die Spalten entsprechen fünf Personen und in die Zeilen stehen für die Genotypen von zwei SNPs.

	Genotyp	Person1	Person2	Person3	Person4	Person5
SNP1	AA	0	0	0	0	1
	AB	1	0	1	1	0
	BB	0	1	0	0	0
SNP2	AA	0	0	0	1	0
	AB	1	1	1	0	0
	BB	0	0	0	0	1

Tabelle 5.1: Beispiel für die 0/1-Kodierung der Genotypen von SNP1 und SNP2 für fünf Personen.

Mit Hilfe dieser Bitschreibweise wird das Auszählen der Genotypen vereinfacht und wesentlich beschleunigt. Betrachtet man z. B. die Genotypkombination (AB, AB) von SNP_i und SNP_j für die Personen k. Für SNP_i und den Genotyp AB beschreibt eine 1/0 Folge der Länge N, wobei N die Anzahl der Personen ist, ob bei der Person k der Genotyp AB vorliegt (Bit wird auf 1 gesetzt) oder nicht (Bit wird auf 0 gesetzt). Eine analoge Folge existiert für den Genotyp AB von SNP_j.

Möchte man nun die Anzahl der Personen mit der Genotypkombination (AB, AB) für SNP_i und SNP_j zählen, verbindet man die 0/1 Folgen des Genotypen AB von SNP_i und SNP_j mit einem bitweisen UND zu einer neuen Bitfolge. Die Anzahl der „Einsen“ in dieser Bitfolge entspricht dann der Anzahl der Personen, die bei beiden SNPs den Genotyp AB besitzen, also doppelt heterozygot sind. Die weiteren Genotypkombinationen behandelt man analog. Anhand des Beispiels 5.2 soll der Gedankengang für die Genotypkombination AB von SNP1 und SNP2 für fünf Personen verdeutlicht werden.

		Person1	Person2	Person3	Person4	Person5	Anzahl (AB, AB)
SNP1	AB	1	0	1	1	0	
SNP2	AB	1	1	1	0	0	
Bitweises UND	AB	1	0	1	0	0	= 2 (Anzahl der „Einsen“ in der Reihe)

Tabelle 5.2: Beispiel für das Auszählen der Genotypkombination (AB, AB) zweier SNPs mit den Personen aus dem obigen Beispiel. In diesem Fall haben zwei Personen die Genotypkombination (AB, AB) an SNP1 und SNP2.

Bitweise Operationen haben den Vorteil, dass sie bei Additions-und Subtraktionsoperationen schneller als floating-point- oder integer-Operationen sind. Bits haben bekannterweise nur zwei Zustände: 1 oder 0. So kann man ein Bit nur setzen oder löschen. Diese Möglichkeit ist allerdings nur durchführbar, wenn vollständige Genotypen vorliegen und keine Wahrscheinlichkeitsgewichte („dosage data“).

5.2.4 Familienbasierte Daten - Trios

Auch sehr interessant können Interaktionsanalysen bei familienbasierten Datensätzen sein beispielsweise bei Trios, also Familien aus Vater, Mutter und Kind. Der Vorteil von familienbasierten Daten ist, dass sie robust gegen Stratifikation sind. Bis jetzt ist es in INTERSNP nur möglich, Fall-Kontroll-Datensätze und quantitative Traits zu analysieren. Für die Analyse von familienbasierten Daten müssen die verschiedenen Assoziations- und Interaktionsanalysen in INTERSNP angepasst werden. Dieser Ansatz wäre aber sicher sinnvoll, da man dann beispielsweise die kompletten HapMap-Daten für Analysen verwenden könnte (siehe Abschnitt 4.1.2). Bis jetzt können nur die unabhängigen Personen, also Vater und Mutter, in die Analyse aufgenommen werden. Für zukünftige Projekte wäre es aber interessant, die vollständigen HapMap-Trios und auch die Daten des 1000-Genome-Projekts verwenden zu können.

Kapitel 6

Zusammenfassung

Die Genetische Epidemiologie hat sich zum Ziel gesetzt, DNA-Sequenzvarianten im menschlichen Genom zu finden, die in der Entwicklung von Krankheiten involviert sind, um so zur Verbesserung von Prognose, Präventionsmaßnahmen und neuen Therapieformen der Krankheiten beizutragen. Im Laufe der letzten Jahre gab es große Fortschritte hinsichtlich der Kosten und des Arbeitsaufwandes der Genotypisierung, was zu neuen Analysestrategien in der Genetischen Epidemiologie geführt hat. Hat man vor zehn Jahren nur einige SNPs pro Person untersucht, ist es heute möglich ca. 1 Million SNPs auf einmal zu analysieren. Früher wurden bestimmte Regionen erst durch die Kopplungsanalyse eingegrenzt, SNPs dieser interessanten Region anschließend genotypisiert und auf Assoziation untersucht. Heute gilt die GWAS, die genomweite Assoziationsanalyse, als Standard und Ausgangspunkt bei der Datenanalyse. Trotz der 769 publizierten GWAS-Studien (Stand 03.02.2011) ist weiterhin ein großer Teil der Heritabilität ungeklärt. Die Einzelmarkeranalyse alleine kann also die Lücke der fehlenden Heritabilität nicht schließen. Aus diesem Grund sind neue Strategien wie die Multimarkeranalyse, die mehrere SNPs simultan betrachtet, erforderlich. Dazu gehören die genomweite Haplotypanalyse, Pathwayassoziationsanalyse und genomweite Interaktionsanalyse. In der vorliegenden Arbeit wurde der Schwerpunkt auf die genomweite Interaktionsanalyse gelegt und eine Software für diese Art der Analyse entwickelt.
Genomweite Interaktionsanalyse (GWIA) aller SNP-Paare von einem Standard-SNP-Chip (ca. 1 Million SNPs) ist rechnerisch ohne massive Parallelisierung auf einem Hochleistungsrecher unmöglich. Darüber hinaus wäre eine GWIA mit allen SNP-Tripeln utopisch, auch wenn die Hochleistungsrechner immer leistungsfähiger werden. Ziel der Software INTERSNP ist es, trotzdem eine genomweite Interaktionsanalyse zu ermöglichen. Um die rechnerischen Hindernisse zu überwinden, werden nur bestimmte Kombinationen von SNPs anhand von a-priori Information für die Interaktionsanalyse auswählt. Somit wird die Anzahl der Interaktionstests reduziert und eine genomweite Interaktionsanalyse ermöglicht. Grundlage dieser a-priori Information können statistische Kriterien (Einzelmarkerassoziationen auf moderater Basis), genetische Relevanz (Lokalisation im Genom) und/oder biologische Relevanz (SNP-Funktionsklassen und Pathwayinformation) sein. INTERSNP bietet für die Multimarkeranalyse der SNPs Tests der logistischen/linearen Regression sowie eines log-linearen Modells an. Für die Korrektur des multiplen Testens steht eine Umgebung für Monte-Carlo Simulationen zur Verfügung. Eine weitere Option, die Anzahl der zu analysierenden Test zu reduzieren, ist die Auswahl des Pre-tests. Beim Pre-test werden die zu berechnenden Paare zuerst mit einer

vereinfachten Teststatistik analysiert und schließlich nur Paare, die beim Pre-Test ein bestimmtes Signifikanzniveau erreichen, mit der komplizierteren Teststatistik berechnet. Zusätzlich steht eine parallelisierte Version zur Verfügung, die eine genomweite Interaktionsanalyse (ca. 500.000 SNPs) in etwa einer Woche ermöglicht. Bei der genomweiten Interaktionsanalyse eines GWAS-Datensatzes bei bipolaren Störungen konnte ein genomweit signifikantes Interaktionspaar gefunden werden, dessen Befund zur Zeit in Replikationsstudien weiter verfolgt wird. Die genomweite Interaktionsanalyse auf Expressionsebene (Leukozyten) in gesunden Personen erbrachte ebenfalls sehr vielversprechende Ergebnisse. Viele der implizierten Gene sind laut GWAS-Datenbank an Autoimmunerkrankungen beteiligt und stellen somit exzellente Kandidaten für Interaktion auf Krankheitsebene dar.
Die Ergebnisse der in der Arbeit dargestellten Anwendungsbeispiele zeigen, dass Interaktion und insbesondere die Verwendung von a-priori Information sinnvolle Ansätze sein können, um weitere Loci zu finden, die in Krankheiten involviert sind und mit herkömmlichen Methoden vielleicht nie gefunden worden wären. Auch die Untersuchung von Expressionsdaten könnte weiteren Aufschluss für das Auffinden der „Missing Heritability“ geben. Zusammenfassend lässt sich sagen, dass die genomweite Interaktionsanalyse neben der GWAS durchaus eine vielversprechende Analysestrategie ist, die weiter verfolgt werden sollte.

Kapitel 7

Ausblick

Durch den medizinischen Fortschritt und die Verbesserung des Lebensstandards wird unsere Gesellschaft in Zukunft immer älter werden und somit werden auch Krankheiten wie Parkinson und Alzheimer mit einer erhöhten Prävalenz auftreten. Diese und andere Krankheiten gehören zur Gruppe der neurodegenerativen Erkrankungen, welche sich durch den fortschreitenden Verlust von Nervenzellen (Neurodegeneration) in Gehirnregionen auszeichnen. Weitere pathologische Kriterien sind die Anlagerung von Proteinen in den Neuronen und anderen Zellen oder extrazellulär, was zu Demenz und Bewegungsstörung führt [Ross and Poirier, 2004]. In Deutschland leiden mehr als eine Million Menschen an Demenzerkrankungen. Aufgrund des demografischen Wandels ist davon auszugehen, dass diese Zahl in den nächsten Jahren noch steigen wird [BMG, 2010]. Abgesehen von den vielen einzelnen persönlichen Tragödien bringt diese Entwicklung eine hohe volkswirtschaftliche Belastung mit sich. Aus diesem Grund ist es wichtig, sich in Zukunft auf die Erforschung der Ursachen von neurodegenerativen Krankheiten zu konzentrieren. Bei Studien mit Alzheimer- und Parkinson-Patienten, welche in Zukunft auch am Deutschen Zentrum für Neurodegenerative Erkrankungen (DZNE) durchgeführt werden, wird neben den herkömmlichen Analyseverfahren sicher auch die Interaktionsanalyse von Bedeutung sein. Die Interaktion der Gene im menschlichen Genom ist längst nicht vollständig aufgeklärt. Auch planen wir ein weiteres Projekt mit der Arbeitsgruppe von Nancy Cox, University of Chicago, wobei es sich diesmal um Expressionsdaten vom Gehirn handelt, welche für die Untersuchung der Ursachen der Demenz sehr interessant sein könnte.
Auch wenn die Hochleistungsrechner immer schneller werden und durch unsere Analysen schon einige hochinteressante Ergebnisse erzielt wurden, steckt noch sehr viel Arbeit in der Erforschung und dem Verstehen der genetischen Ursachen von Krankheiten. Neben den genetischen Faktoren, welche in dieser Arbeit betrachtet wurden, sollten auch Umwelteinflüsse wie Schadstoffe, Lebensstil und Stress nicht unterschätzt werden. Es ist zu erwarten, dass in Zukunft die Erkenntnisse der Genetischen Epidemiologie immer schneller in der Medizin praktische Anwendung finden. Trotzdem wird es noch ein langer Weg sein, die Ursachen von häufigen Krankheiten zufriedenstellend aufzuklären. Große Hoffnung liegt in den Daten des Next-Generation-Sequencing und somit auf den Rare-Variant-Analysen, die im Gegensatz zu GWAS die seltenen Ursachen von Krankheiten untersuchen. Wie lange es im einzelnen Fall dauern wird, bis wir Nutzen aus diesen Daten gewinnen können ist jedoch noch ungewiss. „Die kleinen Unterschiede, die uns zu unverwechselbaren Individuen machen“ [NGFN, 2011] scheinen sich nicht so ein-

fach entschlüsseln zu lassen und werden wahrscheinlich zu einem gewissen Grad auch unbekannt bleiben. Folglich wird die Arbeit auf dem Gebiet der Genetischen Epidemiologie noch lange spannend bleiben und die Forscher immer wieder vor neue Herausforderungen stellen und zu überraschenden Erkenntnissen führen.

Literaturverzeichnis

1000 Genomes Project Consortium, Durbin, R., Abecasis, G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R., Hurles, M., and McVean, G. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

Abecasis, G., Cardon, L., and Cookson, W. (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, 66:279–292.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386.

Balding, D., Bishop, M., and Cannings, C. (2007). *Handbook of statistical genetics*. Wiley.

Barrett, J., Fry, B., Maller, J., and Daly, M. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–5.

Becker, T., Flaquer, A., Brockschmidt, F., Herold, C., and Steffens, M. (2009). Evaluation of potential power gain with imputed genotypes in genome-wide association studies. *Hum Hered.*, 68(1):23–34.

Becker, T. and Herold, C. (2009). Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. *Eur J Hum Genet*, 17(8):1043–9.

Becker, T., Herold, C., Meesters, C., Mattheisen, M., and Baur, M. (2011). Significance Levels in Genome-Wide Interaction Analysis (GWIA). *Ann Hum Genet*, 75(1):29–35.

Becker, T. and Knapp, M. (2004). Maximum-Likelihood Estimation of Haplotype Frequencies in Nuclear Families. *Genet Epidemiol*, 27:21–32.

Bickeböller, H. and Fischer, C. (2007). *Einführung in die Genetische Epidemiologie*. Springer Verlag Berlin Heidelberg.

Bishop, Y., Fienberg, S., and Holland, P. (2007). *Discrete Multivariate Analysis - Theory and Application*. Springer.

BMG (2010). Broschüre des Bundesministerium für Gesundheit: Wenn das Gedächtnis nachlässt.

Bonin, A., Bellemain, E., Eidesen, P. B., Pompanon, F., Brochmann, C., and Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Mol Ecol*, 13(11):3261–73.

Cichon, S., Mühleisen, T. W., Degenhardt, F. A., Mattheisen, M., Miró, X., Strohmaier, J., Steffens, M., Meesters, C., Herms, S., Weingarten, M., Priebe, L., Haenisch, B., Alexander, M., Vollmer, J., Breuer, R., C.Schmäl, Tessmann, P., Moebus, S., Wichmann, H., Schreiber, S., Müller-Myhsok, B., Lucae, S., Jamain, S., Leboyer, M., Bellivier, F., Etain, B., Henry, C., Kahn, J., Heath, S., Consortium, B. D. G. S. B., Hamshere, M., O'Donovan, M., Owen, M., Craddock, N., Schwarz, M., Vedder, H., Kammerer-Ciernioch, J., Reif, A., Sasse, J., Bauer, M., Hautzinger, M., Wright, A., Mitchell, P., Schofield, P., Montgomery, G., Medland, S., Gordon, S. D., Martin, N. G., Gustafsson, O., Andreassen, O., Djurovic, S., Sigurdsson, E., Steinberg, S., Stefansson, H., Stefansson, K., Kapur-Pojskic, L., Oruc, L., Rivas, F., Mayoral, F., Chuchalin, A., Babadjanova, G., Tiganov, A. S., Pantelejeva, G., Abramova, L. I., Grigoroiu-Serbanescu, M., Diaconu, C. C., Czerski, P., Hauser, J., Zimmer, A., Lathrop, M., Schulze, T., Wienker, T., Schumacher, J., Maier, W., Propping, P., Rietschel, M., and Nöthen, M. M. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am J Hum Genet*, 88(3):372–81.

Clark, A. (2004). The role of haplotypes in candidate gene studies. *Genet Epidemiol*, 27(4):321–33.

Clayton, D. (2008). Testing for association on the X chromosome. *Biostatistics*, 9(4):593–600.

Cordell, H. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, 10(6):392–404.

Cordell, H. and Clayton, D. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet*, 70(1):124–41.

Gamazon, E., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E., Nicolae, D., Dolan, M., and Cox, N. (2010). SCAN: SNP and copy number annotation. *Bioinformatics*, 26(2):259–62.

Gao, X., Becker, L., Becker, D., Starmer, J., and Province, M. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet Epidemiol*, 34(1)::100–5.

Hardy, G. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., P. Gaudet, W. K., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la

Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32:258–61.

Heard, E., Tishkoff, S., Todd, J. A., Vidal, M., Wagner, G. P., Wang, J., Weigel, D., and Young, R. (2010). Ten years of genetics and genomics: what have we achieved and where are we heading? *Nat Rev Genet*, 11(10):723–33.

Heinecke, A., Hultsch, E., and Repges, R. (1992). *Medizinische Biometrie. Biomathematik und Statistik*. Springer-Lehrbuch.

Herold, C., Steffens, M., Brockschmidt, F., Baur, M., and Becker, T. (2009). INTERSNP: Genome-wide Interaction Analysis Guided by a priori Information. *Bioinformatics*, 15;25(24):3275–81.

Hüffmeier, U., Uebe, S., Ekici, A., Bowes, J., Giardina, E., Korendowych, E., Juneblad, K., Apel, M., McManus, R., Ho, P., Bruce, I., Ryan, A., Behrens, F., Lascorz, J., Böhm, B., Traupe, H., Lohmann, J., Gieger, C., Wichmann, H., Herold, C., Steffens, M., Klareskog, L., Fitzgerald, T. W. O., Alenius, G., McHugh, N., Novelli, G., Burkhardt, H., Barton, A., and Reis., A. (2010). Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. *Nat Genet*, 42(11):996–9.

Hilgers, R., Bauer, P., and Scheiber, V. (2007). *Einführung in die Medizinische Statistik*. Springer Verlag.

Hillmer, A., Brockschmidt, F., Hanneken, S., Eigelshoven, S., Steffens, M., Flaquer, A., Herms, S., Becker, T., Kortüm, A., Nyholt, D., Zhao, Z., Montgomery, G., Martin, N., Mühleisen, T., Alblas, M., Moebus, S., Jöckel, K., Bröcker-Preuss, M., Erbel, R., Reinartz, R., Betz, R., Cichon, S., Propping, P., Baur, M., Wienker, T., Kruse, R., and Nöthen, M. (2008). Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nat Genet*, 40(11):1279–81.

Hindorff, L., Junkins, H., Hall, P., Mehta, J., and Manolio, T. (2011). Catalog of Published Genome-Wide Association Studies.

Hirsch-Kauffmann, M. and Schweiger, M. (2000). *Biologie für Mediziner und Naturwissenschaftler*. Thieme, Stuttgart.

Holmans, P., Green, E., Pahwa, J., Ferreira, M., Purcell, S., Sklar, P., Consortium, W. T. C.-C., Owen, M., O'Donovan, M., and Craddock, N. (2009). Gene Ontology Analysis of GWAS Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *Am J Hum Genet*, 85:13–24.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61.

Kam-Thong, T., Czamara, D., Tsuda, K., Borgwardt, K., Lewis, C., Erhardt-Lehmann, A., Hemmer, B., Rieckmann, P., Daake, M., Weber, F., Wolf, C. ., Ziegler, A., Pütz, B., Holsboer, F., Schölkopf, B., and Müller-Myhsok, B. (2010). EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *Eur J Hum Genet*, 19(4):465–71.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34:354–357.

Kiewert, A. (2006). *Empfehlungen zur Qualitätssicherung von Genotypisierungsdaten bei familienbasierten Studien mit Mikrosatelliten.* PhD thesis, Universität zu Lübeck.

Knapp, M., Strauch, K., Baur, M. P., and Wienker, T. F. (2001). *Quantitative Methoden in der genetischen Epidemiologie.* Institut für Medizinische Biometrie, Informatik und Epidemiologie, Universität Bonn.

Li, J. (2010). *Logistic Regression.* Department of Statistics, The Pennsylvania State University.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21.

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A., Cho, J., Guttmacher, A., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C., Slatkin, M., Valle, D., Whittemore, A., Boehnke, M., Clark, A., Eichler, E., Gibson, G., Haines, J., Mackay, T., McCarroll, S., and Visscher, P. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.

Marchini, J., Donnelly, P., and Cardon, L. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37(4):413–7.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet*, 39:906–913.

Michal, G. (1993). Biochemical Pathways (Poster). Technical report, Boehringer Mannheim, Penzberg.

Miller, C., Joyce, P., and Waits, L. (2002). Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, 160:357–66.

MPI (2009). MPI: A Message-Passing Interface Standard.

NGFN, N. (2011). NGFN Homepage: Genomforschung.

O'Dushlaine, C., Kenny, E., Heron, E., Segurado, R., Gill, M., Morris, D., and Corvin, A. (2009). The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, 25:2762–2763.

OpenMP (2008). The OpenMP API specification for parallel programming.

Press, W., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing.* Cambridge University Press.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.*, 38(8):904–9.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., and Sham, P. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses linkage analyses. *Am J Hum Genet*, 81(3):559–75.

Rexbye, H., Petersen, I., Iachina, M., Mortensen, J., McGue, M., Vaupel, J., and Christensen, K. (2005). Hair loss among elderly men: etiology and impact on perceived age. *J Gerontol A Biol Sci Med Sci*, 60(8):1077–82.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7.

Ritchie, M. (2011). Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Ann Hum Genet*, 75(1):172–82.

Ross, C. and Poirier, M. (2004). Protein aggregation and neurodegenerative disease. *Nat Med*, 10:10–7.

Sachs, L. and Hedderich, J. (2009). *Angewandte Statistik: Methodensammlung mit R.* Springer, Berlin; Auflage: 13. Aufl.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Stein, T. I., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford)*, baq020.

Schüpbach, T., Xenarios, I., Bergmann, S., and Kapur, K. (2010). FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26(11):1468–9.

Schreiber, F. (2001). *Visualisierung biochemischer Reaktionsnetze.* PhD thesis, Universität Passau.

Schreiber, F. (2009). Analyse und Visualisierung biologischer Netzwerke. *Informatik Spektrum*, 32:301–309.

Schuster, S. (2008). Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1):16–8.

Shen, L., Weber, C., Raleigh, D., Yu, D., and Turner., J. (2011). Tight junction pore and leak pathways: a dynamic duo. *Annu Rev Physiol*, 17;73:283–309.

Spielman, R., McGinnis, R., and Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52(3):506–16.

Steffens, M., Becker, T., Sander, T., Fimmers, R., Herold, C., Holler, D., Leu, C., Herms, S., Cichon, S., Bohn, B., Gerstner, T., Griebel, M., Nöthen, M., Wienker, T., and Baur, M. (2010). Feasible and successful: genome-wide interaction analysis involving all 1.9 x 10(11) pair-wise interaction tests. *Hum Hered*, 162(4):899–903.

Thorisson, G. and Stein, L. (2003). The SNP Consortium website: past, present and future. *Nucleic Acids Res*, 31(1):124–7.

Trégouët, D., König, I., Erdmann, J., Munteanu, A., Braund, P., Hall, A., Grosshennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M., Meitinger, T., Wright, B., Preuss, M., Balmforth, A., Ball, S., Meisinger, C., Germain, C., Evans, A., Arveiler, D., Luc, G., Ruidavets, J., Morrison, C., van der Harst, P., Schreiber, S., Neureuther, K., Schäfer, A., Bugert, P., Mokhtari, N. E., Schrezenmeir, J., Stark, K., Rubin, D., Wichmann, H., Hengstenberg, C., Ouwehand, W., Consortium, W. T. C. C., Consortium, C., Ziegler, A., Tiret, L., Thompson, J., Cambien, F., Schunkert, H., and Samani, N. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet*, 41(3):283–5.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., and Yu1, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet*, 87(3):325–40.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet*, 81(6):1278–1283.

Weiß, C. (2008). *Basiswissen Medizinische Statistik*. Springer Medizin Verlag Heidelberg.

Westfall, P. and Young, S. (1993). *Resampling-Based Multiple Testing*. Wiley-Interscience.

Winer, B. J. (1962). *Statistical Principles in Experimental Design*. McGraw-Hill.

Anhang A

Algorithmen

A.1 Logistische Regression

Mit der logistischen Regression kann die Wahrscheinlichkeit der Zugehörigkeit zu einer Gruppe in Abhängigkeit von einer oder mehreren unabhängigen Variablen bestimmt werden [Sachs and Hedderich, 2009]. Die Herleitung der logistischen Regression wurde bereits im Kapitel 3.3.2 beschrieben. Im Folgenden wird noch auf einige Details zur Umsetzung der logistischen Regression [Cordell and Clayton, 2002] in INERSNP eingegangen.
Die Anzahl der Person ist:

$$N = n_1 + n_2$$

wobei n_1 = Anzahl der Fälle und n_2 = Anzahl der Kontrollen ist. Die Likelihood berechnet sich aus

$$L = \prod_{i=1}^{N} p_i^{I_i} (1 - p_i)^{1-I_i}$$

mit I_i als Indikatorfunktion für den Fall-Kontroll-Status (1= Fall, 0=Kontrolle) und p_i als die Wahrscheinlichkeit einer Person i ein Fall zu sein. Die Regressionsgleichung lautet:

$$p_i = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$

oder äquivalent

$$\text{logit}(p) := \ln\left(\frac{p}{1-p}\right) = \beta^T x,$$

wobei β der Vektor der geschätzten Koeffizienten ist und x der Vektor in dem die Genotypen kodiert sind.

In INTERSNP wurde für die Schätzung der β-Gewichte das Newton-Raphson-Verfahren [Press et al., 2007] verwendet, welches in der Regel für Parameteroptimierung benutzt wird, insbesondere für die Maximum-Likelihood-Schätzung. Die Grundidee ist die iterative Bestimmung der Nullstellen einer reellen Funktion. Zuerst wird das Verfahren im allgemeinen Fall erklärt, danach die Anwendung auf die logistische Regression und schließlich der in INTERSNP implementierte Algorithmus erläutert. Im eindimensionalen Fall haben wir:

$$f(\beta) = 0 \text{ mit } f\colon \mathbb{R} \to \mathbb{R} \text{ und } \beta \in \mathbb{R}$$

Im $t+1$ Iterationsschritt berechnet sich β_{t+1} wie folgt:

$$\beta_{t+1} = \beta_t - \frac{f(\beta_t)}{f'(\beta_t)},$$

wobei β_t der t-te Iterationsschritt ist.

Im mehrdimensional Fall funktioniert der Algorithmus analog:

$$f\colon \mathbb{R}^p \to \mathbb{R}^p, \beta \in \mathbb{R}^p$$
$$\beta_{t+1} = \beta_t \underbrace{- \left(J(\beta_t)^{-1} f(\beta_t) \right)}_{:\, \Delta\beta_t}$$

mit der Jacobi-Matrix der partiellen Ableitungen $J(\beta_t) = \frac{\partial f_i}{\partial \beta_j}$, wobei $1 \leq i \leq p$ und $1 \leq j \leq p$. Da die numerische Invertierung von J sehr rechenintensiv ist, wird statt dessen das lineare Gleichungssystem

$$J(\beta_t)\Delta\beta_t = -f(\beta_t)$$

gelöst, d.h. $\Delta\beta_t$ wird bestimmt. Somit ergibt sich β_{t+1} als $\beta_{t+1} = \beta_t + \Delta\beta_t$.
Im Folgenden wird das iterative Verfahren auf unser Problem, die Maximierung der Likelihood $L = L(\beta)$ angewendet [Li, 2010]. Für die Maximierung sucht man die Nullstellen der 1. Ableitung von L, also $L'(\beta) = 0$. Somit ist $L'(\beta) = f(\beta)$ in der „Newton-Raphson-Notation“. Es kann gezeigt werden, dass

$$f(\beta) = L'(\beta) = \frac{\delta L(\beta)}{\delta \beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{p})$$

$$f'(\beta) = L''(\beta) = \frac{\delta^2 L(\beta)}{\delta\beta\delta\beta^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

mit der Diagonalmatrix W:

$$\mathbf{W} = \text{diag} \begin{pmatrix} p(x_1, \beta_t)(1 - p(x_1, \beta_t)) \\ p(x_2, \beta_t)(1 - p(x_2, \beta_t)) \\ \vdots \\ p(x_N, \beta_t)(1 - p(x_N, \beta_t)) \end{pmatrix}$$

$\mathbf{y}$ als Spaltenvektor der y_i mit $i = 1 \ldots N$, also der Vektor mit dem Krankheitsstatus der Personen, $\mathbf{X}$ als $N \times (p+1)$ Eingangsmatrix, wobei p die Anzahl der Parameter ist und $\mathbf{p}_t$ als N-Vektor der gemäß β berechneten Wahrscheinlichkeiten des i-ten Elements $p(x_i, \beta_t)$ der Iteration t.

Der Newton-Raphson-Schritt ist dann also

$$\beta_{t+1} = \beta_t + (\mathbf{X}^T\mathbf{W}_t\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{p}_t)$$

Im Folgenden wird das iterative Verfahren, welches kompakt in Matrixform ausgedrückt werden kann und somit übersichtlicher ist, erklärt [Li, 2010]. Da $\mathbf{W}$ eine $N \times N$ Diagonalmatrix ist, können direkte Matrixoperationen mit ihr sehr ineffizient sein. Deshalb betrachtet man direkt $\tilde{\mathbf{X}} := \mathbf{W}\mathbf{X}$. Die einzelnen Schritte lauten:

1. Setze $\beta = 0$.
2. Bestimme $\mathbf{y}$ indem die Elemente wie folgt gesetzt werden:

$$y_i = \begin{cases} 1, & \text{wenn Person } i \text{ ein Fall ist,} \\ 0, & \text{wenn Person } i \text{ eine Kontrolle ist.} \end{cases}$$

3. Berechne $\mathbf{p}_t$ indem die Elemente wie folgt gesetzt werden:

$$p_t(x_i; \beta_t) = \frac{e^{\beta_t^T x_i}}{1 + e^{\beta_t^T x_i}}$$

   mit i = 1,2,. . . ,N.
4. Berechne die $N \times (p-1)$ Matrix $\tilde{\mathbf{X}} := \mathbf{W}\mathbf{X}$ gemäß:

$$\tilde{\mathbf{X}} = \begin{pmatrix} p(x_1, \beta_t)(1 - p(x_1, \beta_t))x_1^T \\ p(x_2, \beta_t)(1 - p(x_2, \beta_t))x_2^T \\ \vdots \\ p(x_N, \beta_t)(1 - p(x_N, \beta_t))x_N^T \end{pmatrix}.$$

5. Bestimme einen neuen Schätzer für β mittels $\beta = (\mathbf{X}^T\tilde{\mathbf{X}})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{p})$
6. Wenn das Stoppkriterium erreicht wird, halte an, sonst gehe zurück zu Schritt 3.

Das Stoppkriterium ist erreicht, wenn $L(\beta_{t+1}) - L(\beta_t) < \epsilon$ mit $\epsilon < 10^{-6}$.

A.2 Lineare Regression

Das lineare Regressionsmodell beschreibt den Zusammenhang von p Einflussgrößen $x_1, x_2, \ldots, x_p$ und einer Zielvariable y (quantitativer Wert). Das vollständige Modell für insgesamt n Beobachtungen kann wie folgt beschrieben werden [Sachs and Hedderich, 2009]:

$$\begin{array}{ccccccccc} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} & = & \begin{pmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \ddots & \vdots & \\ 1 & X_{n1} & \cdots & X_{pn} \end{pmatrix} & \cdot & \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} & + & \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon \\ \epsilon_p \end{pmatrix} \\ Y & = & X & \cdot & \beta & + & \epsilon \end{array}$$

In Indexnotation lautet die Modellgleichung:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \cdots + \beta_p x_{p1} + \epsilon_i$$

Das Regressionsproblem besteht darin, die Koeffizienten mit Hilfe von Schätzern zu bestimmen. Das Ziel ist dabei die Daten möglichst gut an die lineare Gleichung anzupassen. Eine Lösung ist das Verfahren der kleinsten Abweichungsquadrate, die im Folgenden dargestellt wird [Rolf Fimmers, persönliche Kommunikation].

Mit Minimierung der Abweichungsquadratsumme verstehen wir die Lösung der Gleichung

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 = \min$$

Die Lösungen erhalten wir zunächst durch die Ableitung nach β_0 und Gleichsetzung mit Null:

$$\begin{aligned}
&\frac{\delta}{\delta \beta_0} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \\
&= -2 \cdot \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right) \\
&= -2n \cdot (\bar{Y} - \beta_0) + 2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{p} \beta_j X_{ij} \\
&= 0 \\
&\Longrightarrow \beta_0 = \bar{Y} - \sum_{j=1}^{p} \beta_j \bar{X}_j
\end{aligned}$$

Anschließend leiten wir analog nach β_k ab:

$$\begin{aligned}
&\frac{\delta}{\delta \beta_k} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \\
&= -2 \cdot \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right) \cdot X_{ik} = 0 \\
&\Longrightarrow \sum_{i=1}^{n} \left(Y_i - \bar{Y} + \sum_{j=1}^{p} \beta_j \bar{X}_j - \sum_{j=1}^{p} \beta_j X_{ij} \right) \cdot X_{ik} = 0
\end{aligned}$$

Im Folgenden skizzieren wir die Lösung dieser Gleichungssysteme mit Hilfe der Matrixinvertierung. In der Matrixschreibweise lautet die zu minimierende Gleichung:

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta)$$

Nach einigen Umformungen und Ableitung nach β ergibt sich:

$$(X^T X)\beta = X^T Y \ (*)$$

oder explizit:

$$\begin{pmatrix} n & \sum X_{1i} & \cdots & \sum X_{pi} \\ \sum X_{1i} & \sum X_{1i}^2 & \cdots & \sum X_{1i}X_{pi} \\ \vdots & \ddots & \vdots & \\ \sum X_{pi} & \sum X_{1i}X_{pi} & \cdots & \sum X_{pi}^2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ \vdots & \ddots & \vdots & \\ 1 & X_{n1} & \cdots & X_{pn} \end{pmatrix}^T \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

mit $X^TX = \begin{pmatrix} n & \sum X_{1i} & \cdots & \sum X_{pi} \\ \sum X_{1i} & \sum X_{1i}^2 & \cdots & \sum X_{1i}X_{pi} \\ \vdots & \vdots & \ddots & \vdots \\ \sum X_{pi} & \sum X_{1i}X_{pi} & \cdots & \sum X_{pi}^2 \end{pmatrix}$.

Erste Zeile ist

$$n\beta_0 + \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} X_{ji} = \sum_{i=1}^{n} Y_i \Leftrightarrow \sum_{j=1}^{p} \beta_j + \bar{X}_j = \bar{Y}$$

Die k-te Zeile ist

$$\beta_0 + \sum_{i=1}^{n} X_{ki} + \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} X_{ki}X_{ji} = \sum_{i=1}^{n} X_{ki}Y_i \Leftrightarrow \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right) \cdot X_{ij} = 0$$

Aus $(*)$ ergibt sich die explizite Lösung für β.

$$\beta = (X^TX)^{-1}X^TY$$

d.h. Das Gleichungssystem kann durch die Inversion der Matrix X^TX gelöst werden ($\rightarrow$ Dwyer-Algorithmus).

Einsetzen der Lösung $\beta = (X^TX)^{-1}X^TY$ in die Abweichungsquadratsumme ergibt ($\beta^T = Y^TX(X^TX)^{-1^T} = Y^TX(X^TX)^{-1}$, da X^TX symmetrisch):

$$\begin{aligned} \epsilon^T\epsilon &= Y^TY + Y^TX(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^TY - 2Y^TX(X^TX)^{-1}X^TY \\ &= Y^TY - Y^TX(X^TX)^{-1}X^TY \end{aligned}$$

Aus den Abweichungsquadratsummen für L_1, das uneingeschränkte Modell, und L_2, das eingeschränkte Modell, bildet man schließlich die F-Statistik (vgl. Abschnitt 3.3).

A.3 Matrixinvertierung mit dem Dwyer-Algorithmus

Der Dwyer-Algorithmus wird benutzt um die Inverse einer symmetrischen Matrix zu finden. Da die Matrix X^TX aus A.2 symmetrisch ist, können wir den Dwyer-Algorithmus für die lineare Regression verwenden. Sei M eine symmetrische nicht-singuläre Matrix. Diese Matrix kann durch zwei Dreiecksmatrizen dargestellt werden

$$M = TT^T$$

wobei T eine untere und T^T eine obere Dreiecksmatrix ist (alle Einträge über bzw. unter der Diagonalen sind Null). Die Inverse der Matrix M kann somit in folgender Form ausgedrückt werden:

$$\begin{aligned} M^{-1} = (TT^T)^{-1} &= (T^T)^{-1}T^{-1} \\ &= U^T U, \text{ wobei } U = T^{-1}, \text{ also} \qquad U \cdot T = I \end{aligned}$$

wobei I die Einheitsmatrix ist. Die Inverse der Matrix T ist relativ einfach zu erhalten. Der Dwyer-Algorithmus berechnet die T und U Matrix simultan. Im folgenden wird der Algorithmus skizziert [Winer, 1962]:

Für den Fall $n = 3$ ergibt sich für die Elemente der gesuchten Matrizen:

$$\begin{array}{rr} m_{11}m_{12}m_{13} & d_{11} \\ m_{22}m_{23} & d_{21}d_{22} \\ m_{33} & d_{31}d_{32}d_{33} \end{array}$$

$$\begin{array}{rr} t_{11}t_{12}t_{13} & u_{11} \\ t_{22}t_{23} & u_{21}u_{22} \\ t_{33} & d_{31}u_{32}u_{33} \end{array}$$

$$\begin{array}{llll} t_{11} = \sqrt{m_{11}}; & t_{1j} = m_{1j}/t_{11}; & u_{11} = d_{11}/t_{11} & j = 2,3 \\ t_{22} = \sqrt{m_{22} - t_{12}^2}; & t_{2j} = (m_{2j} - t_{12}t_{1j})/t_{22} & & j = 3 \\ & u_{2k} = (d_{2k} - t_{12}u_{1k})/t_{22}; & & k = 1,2 \\ t_{33} = \sqrt{m_{33} - t_{13}^2 - t_{23}^2}; & u_{3k} = (d_{3k} - t_{13}u_{1k} - t_{23}u_{2k})/t_{33} & & k = 1,2,3 \end{array}$$

Aus den Elementen von U lassen sich dann die Elemente von M^{-1} bestimmen. Damit wäre die Inverse im dreidimensionalen Fall gefunden. Es lässt sich zeigen, dass sich die Lösung im allgemeinen Fall mit p Parametern gemäß

$$\begin{aligned} t_{pp} &= \sqrt{m_{pp} - t_{1p}^2 - t_{2p}^2 - \cdots - t_{(p-1)p}^2} \\ t_{pj} &= (m_{pj} - t_{1p}t_{1j} - t_{2p}t_{2j} - \cdots - t_{(1-p)p}t_{(p-1)j})/t_{pp}, \qquad j > p \\ u_{pk} &= (d_{pk} - t_{1p}u_{1k} - t_{2p}u_{2k} - \cdots - t_{(p-1)p}u_{(p-1)k})/t_{pp}, \qquad k \leq p \end{aligned}$$

berechnen lässt, was uns wiederum sofort die Elemente von M^{-1} liefert.

Anhang B

Optionen in INTERSNP

Schlüsselwort	Beschreibung
TPED	Pfad zum tped-File.
TFAM	Pfad zum tfam-File.
ANNOTATIONFILE	Pfad zum Annotation-File. Hinweis: Wird für das genetische Kriterium und die erweiterte Einzelmarker-Ausgabe benötigt. Hinweis: ANNOTATE oder GENETIC_IMPACT und M_WITH_GENETIC_IMPACT muss ausgewählt sein.
PATHWAYFILE	Pfad zum Pathway-File. Hinweis: PATHWAY muss ausgewählt sein (1).
COVARIATEFILE	Pfad zum Covariate-File. Hinweis: Kovariaten werden mit dem Schlüsselwort COVARIATES ausgewählt. Hinweis: Fehlende Datenwerte sind als „-“ oder „x“ definiert.
MODELFILE	Pfad zum Model-File. Hinweis: Bei TEST muss ein M stehen.
COMBILIST	Um ausgewählte Paare oder Triple zu analysieren.
COMBIFILE	Pfad zum Combi-File. Hinweis: COMBILIST muss auf 1 gesetzt sein.
ONLY_MALE	Es werden nur Männer analysiert (1).
ONLY_FEMALE	Es werden nur Frauen analysiert (1).
POSCHOICE	Auswahl einer bestimmten Region (nur die ausgewählte Region wird analysiert). Beispiel: chr4;chr12,123000-160000;chr24;
NEGCHOICE	Ausschluss einer bestimmten Region (diese Region wird NICHT analysiert).

HWE_P_CASE	QC-Grenzwert für das HWE in Fällen (z.B. p-Wert 0,001).
HWE_P_CONTROL	QC-Grenzwert für das HWE in Kontrollen (z.B. p-Wert 0,01).
MRDIFF	QC-Grenzwert für die Missingrate (MR): Personen und SNPs schlechter als die durchschnittliche $MR+MRDIFF$ werden gelöscht.
MAF	QC-Grenzwert für MAF (Häufigkeit des seltenen Allels).
SINGLE_MARKER	Auswahl des Einzelmarkertests. 1: Armitage's Trend Test (default), 2: Genotypischer Test mit 2 F.G., 3: logistische Regression mit 1 F.G., 4: logistische Regression mit 2 F.G.
TWO_MARKER	Wähle diesen Parameter aus (1) um eine 2-Marker-Analyse durchzuführen.
THREE_MARKER	Wähle diesen Parameter aus (1) um eine 3-Marker-Analyse durchzuführen. Hinweis: Es ist nicht möglich die 2-Marker- und 3-Marker-Analyse gleichzeitig auszuwählen.
TEST	Auswahl des Multimarkertests. 1=χ^2-Test, 2=log-lineares Modell, 3-12=Logistische Regression, M=benutzerdefiniertes Regressionsmodell (MODELFILE auswählen).
COVARIATES	Wähle welche Kovariaten für die Analyse benutzt werden sollen (z.B. `2-4;` → Kovariate 2,3 und 4 werden ausgewählt). Sollen alle Kovariaten mit einbezogen werden, kann man „`1-10;`“ schreiben. Hinweis: Diese Option kann nur verwendet werden, wenn ein Covariate File (Schlüsselwort: COVARIATEFILE) angegeben wurde.
SEXCOV	Geschlecht wird als Kovariate benutzt, 0=Geschlecht wird nicht als Kovariate benutzt.
SINGLETOP	Basierend auf den Einzelmarker-P-Werten wird eine Liste von n top SNPs berechnet. Die Länge dieser Liste kann mit dieser Option festgelegt werden.
M_WITH_SINGLETOP	Anzahl der SNPs (0,1,2,3), die aus der Einzelmarker-Bestenliste stammen müssen. Hinweis: SINGLETOP muss ausgewählt sein. Hinweis: M_WITH_SINGLETOP=3 ist nur möglich wenn THREE_MARKER=1.
GENETIC_IMPACT	Genetisches Kriterium: 0=Genwüste, 1=Gen-LD-Bereich, 2=Exon, 3=in einer kodierenden Region, 4=nicht-synonym.

	Hinweis: Für diese Option wird das Annotation File benötigt (Schlüsselwort: ANNOTATIONFILE).
M_WITH_GENETIC_IMPACT	Anzahl der SNPs (0,1,2,3), die dem genetischen Kriterium entsprechen. Hinweis: GENETIC_IMPACT muss ausgewählt sein. Hinweis: M_WITH_GENETIC_IMPACT=3 ist nur bei THREE_MARKER=1 möglich.
SNP1	rs-Nummer für den ersten festen SNP der Analyse.
SNP2	rs-Nummer für den zweiten festen SNP der Analyse.
SNP3	rs-Nummer für den dritten festen SNP der Analyse (nur möglich wenn THREE_MARKER=1).
PATHWAY	1=Pathway Informationen werden benutzt, 0=pathway Informationen werden ignoriert. Hinweis: Um diese Option zu nutzen, muss ein Pathway File angegeben werden (Schlüsselwort: PATHWAYFILE).
SIMULATION	Anzahl der MC-Simulationen (0=es werden keine Simulationen ausgeführt)
MC_WITH_SM	1=MC-Simulationen korrigieren für die Multimarkeranalyse UND die Einzelmarkeranalyse. 0=MC-Simulationen nur bei der Multimarkeranalyse (default).
PRINTTOP	Die n besten Multimarker-p-Werte werden in die Ausgabedatei geschrieben.
QT	Um quantitative Traits zu analysieren, muss QT auf 1 gesetzt werden (default ist `QT 0`).
MISSING_PHENO	Festlegen eines Wertes für fehlende Phänotypen in den quantitativen Datensätzen.
ANNOTATE	Einzelmarkerausgabedatei soll Annotationsinformationen enthalten.
GENECOL	Angabe der Spalte der Annotationsinformation.
OUTPUTNAME	Name und Pfad der Ausgabedateien.

Eigene Publikationen

1. **Herold, C.** and Becker, T. (2009). Genetic association analysis with FAMHAP: a major program update. Bioinformatics, 25:134–136.

2. Becker, T., Flaquer, A., Brockschmidt, F., **Herold, C.**, and Steffens, M. (2009). Evaluation of potential power gain with imputed genotypes in genome-wide association studies. Hum Hered., 68(1):23–34.

3. Schumacher, J., Laje, G., Jamra, R. A., Becker, T., Mühleisen, T., Vasilescu, C., Mattheisen, M., Herms, S., Hoffmann, P., Hillmer, A., Georgi, A., **Herold, C.**, Schulze, T., Propping, P., Rietschel, M., McMahon, F., Nöthen, M., and Cichon, S. (2009). The DISC locus and schizophrenia - Evidence from an association study in a central European sample and from a meta-analysis across different European populations. Hum Mol Genet., 18(14):2719–27.

4. Becker, T. and **Herold, C.** (2009). Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. Eur J Hum Genet, 17(8):1043–9.

5. Scholl, H., Fleckenstein, M., Fritsche, L., Schmitz-Valckenberg, S., Göbel, A., Adrion, C., **Herold, C.**, Keilhauer, C., Mackensen, F., Mössner, A., Pauleikhoff, D., Weinberger, A., Mansmann, U., Holz, F., Becker, T., and Weber, B. (2009). CFH, C3 and ARMS2 are significant risk loci for susceptibility but not for disease progression of geographic atrophy due to AMD. PLoS One, 4(10):e7418.

6. **Herold, C.**, Steffens, M., Brockschmidt, F., Baur, M., and Becker, T. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics, 15;25(24):3275–81.

7. Redler, S., Brockschmidt, F., Forstbauer, L., Giehl, K., **Herold, C.**, Eigelshoven, S., Hanneken, S., Weert, J. D., Lutz, G., Wolff, H., Kruse, R., Blaumeiser, B., Böhm, M., Becker, T., Nöthen, M., and Betz, R. (2010). The TRAF1/C5 locus confers risk for familial and severe alopecia areata. Br J Dermatol., 162(4):866–9.

8. Brockschmidt, F., Hillmer, A., Eigelshoven, S., Hanneken, S., Heilmann, S., Barth, S., **Herold, C.**, Becker, T., Kruse, R., and Nöthen, M. (2010). Fine mapping of the human AR/EDA2R locus in androgenetic alopecia. Br J Dermatol., 162(4):899– 903.

9. Steffens, M., Becker, T., Sander, T., Fimmers, R., **Herold, C.**, Holler, D., Leu, C., Herms, S., Cichon, S., Bohn, B., Gerstner, T., Griebel, M., Nöthen,

M., Wienker, T., and Baur, M. (2010). Feasible and successful: genome-wide interaction analysis involving all 1.9 x 10(11) pair-wise interaction tests. Hum Hered, 162(4):899–903.

10. Becker, T., **Herold, C.**, Meesters, C., Mattheisen, M., and Baur, M. (2011). Significance Levels in Genome-Wide Interaction Analysis (GWIA). Ann Hum Genet, 75(1):29– 35.

11. Hüffmeier, U., Uebe, S., Ekici, A., Bowes, J., Giardina, E., Korendowych, E., Juneblad, K., Apel, M., McManus, R., Ho, P., Bruce, I., Ryan, A., Behrens, F., Lascorz, J., Böhm, B., Traupe, H., Lohmann, J., Gieger, C., Wichmann, H., **Herold, C.**, Steffens, M., Klareskog, L., Fitzgerald, T. W. O., Alenius, G., McHugh, N., Novelli, G., Burkhardt, H., Barton, A., and Reis., A. (2010). Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. Nat Genet., 42(11):996–9.

12. Leon, C. A., Schumacher, J., Kluck, N., **Herold, C.**, Schulze, T., Propping, P., Rietschel, M., Cichon, S., Nöthen, M., and Jamra, R. A. (2011). Association study of the GRIA1 and CLINT1 (Epsin 4) genes in a German schizophrenia sample. Psychiatr Genet., 21(2):114.

13. John, K., Brockschmidt, F., Redler, S., **Herold, C.**, Hanneken, S., Eigelshoven, S., Giehl, K., Weert, J. D., Lutz, G., Kruse, R., Wolff, H., Blaumeiser, B., Böhm, M., Becker, T., Nöthen, M., and RC, R. B. (2011). Genetic Variants in CTLA4 Are Strongly Associated with Alopecia Areata. J Invest Dermatol., 131(5):1169–72.

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei allen bedanken, die zum Gelingen dieser Arbeit beigetragen und mich während meiner Promotion unterstützt haben.

Ich danke sehr herzlich PD Dr. Tim Becker, der mich während meiner Promotion exzellent betreut und mir bei Fragen und neuen Ideen immer mit Rat und Tat zur Seite stand. Einen besseren Einstieg in die Wissenschaft hätte ich nicht haben können. Vielen, vielen Dank auch für die Möglichkeit eines Auslandsaufenthalts in den USA und die große Unterstützung bei der Organisation. Damit ging ein großer Wunsch in Erfüllung und ich konnte damit einzigartige Erfahrungen sammeln.

Prof. Jürgen Bajorath danke ich für das Interesse an meiner Arbeit und für die freundliche Übernahme des Zweitgutachters. Ebenso danke ich den Prof. Joachim Schultze und Prof. Max Baur für die Teilnahme an der Promotionskommission.

Desweiteren danke ich Prof. Max Baur, dass er mir die Gelegenheit gegeben hat, diese Arbeit an seinem Institut durchzuführen, und dass er mir den großartigen Auslandsaufenthalt in Chicago ermöglichte.

Vielen Dank auch an Prof. Nancy Cox und ihre Arbeitsgruppe für die unvergesslichen Monate an der University of Chicago und die daraus resultierende Gelegenheit an sehr interessanten Forschungsprojekten mitzuarbeiten. Nicht zu vergessen ist auch die buntgemischte Gruppe aus dem International House, die mir ein zweites Zuhause gegeben hat.

Den Kollegen der Humangenetik möchte ich für die freundschaftliche und gute Zusammenarbeit an immer wieder interessanten Projekten danken, auch dafür, dass sie mir die Datensätze für meine Anwendungsbeispiele zur Verfügung gestellt haben.

Darüber hinaus gebührt mein Dank Prof. Thomas Wienker und den Kollegen der Genetischen Epidemiologie, aber auch allen derzeitigen und ehemaligen Kollegen des IMBIE und DZNE für die freundliche Arbeitsatmosphäre und stete Hilfsbereitschaft, besonders meiner Arbeitsgruppe, dem „Team Becker“.

Herrn Waldemar Spitz gilt besonderer Dank für die kompetente Unterstützung in technischen Fragen und zuverlässige Computeradministration.

Vielen Dank auch an Manuel Mattheisen für die immer neuen Ideen und Anre-

gungen zur Verbesserung von INTERSNP.

Aller besten Dank an meine Korrekturhelfer Annette Simon, Stephan Gade, Markus Leber, Michael Knapp und ganz besonders Dmitriy Drichel, die in den letzten Wochen großartige Arbeit geleistet haben. Christian und Stephan danke ich zusätzlich für die schnelle Hilfe bei Latex-Fragen. Auch meine Geschwister Ulrike und Simon sowie meine Eltern wurden nicht müde mein Geschriebenes zu korrigieren. Vielen Dank!

Neben der Arbeit sorgten die IMBIE Runners mit Tim Becker, Ute Klarmann, Kim Schmidt et al. immer wieder für Highlights im Alltag. Euch lieben Dank. Aber auch meine jahrelangen Freunde Annette Simon, Melanie Streusel und die restlichen „HGler“, sowie Christina Wassermann und die „Schwaben“ haben mich unterstützt und für die notwendige Ablenkung im positiven Sinne gesorgt.

Ein abschließender Dank gilt meiner „Großfamilie“, besonders meinen Eltern, die auf ihre Weise zum Erfolg der Arbeit beigetragen und mich immer wieder motiviert haben. Danke für eure Geduld und Unterstützung.

Erklärung

An Eides statt versichere ich, dass ich die Dissertation „INTERSNP Genomweite Interaktionsanalyse mit a-priori Information“ selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist.

- Becker, T., Flaquer, A., Brockschmidt, F., **Herold, C.**, and Steffens, M. (2009). Evaluation of potential power gain with imputed genotypes in genome-wide association studies. Hum Hered., 68(1):23–34.

- Becker, T. and **Herold, C.** (2009). Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. Eur J Hum Genet, 17(8):1043–9.

- **Herold, C.**, Steffens, M., Brockschmidt, F., Baur, M., and Becker, T. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics, 15;25(24):3275–81.

- Steffens, M., Becker, T., Sander, T., Fimmers, R., **Herold, C.**, Holler, D., Leu, C., Herms, S., Cichon, S., Bohn, B., Gerstner, T., Griebel, M., Nöthen, M., Wienker, T., and Baur, M. (2010). Feasible and successful: genome-wide interaction analysis involving all 1.9 x 10(11) pair-wise interaction tests. Hum Hered, 162(4):899–903.

- Hüffmeier, U., Uebe, S., Ekici, A., Bowes, J., Giardina, E., Korendowych, E., Juneblad, K., Apel, M., McManus, R., Ho, P., Bruce, I., Ryan, A., Behrens, F., Lascorz, J., Böhm, B., Traupe, H., Lohmann, J., Gieger, C., Wichmann, H., **Herold, C.**, Steffens, M., Klareskog, L., Fitzgerald, T. W. O., Alenius, G., McHugh, N., Novelli, G., Burkhardt, H., Barton, A., and Reis., A. (2010). Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. Nat Genet., 42(11):996–9.

- Becker, T., **Herold, C.**, Meesters, C., Mattheisen, M., and Baur, M. (2011). Significance levels in genome-wide interaction analysis (GWIA). Ann Hum Genet, 75(1):29–35.