

Aus dem Institut für Medizinische Biometrie und Statistik
der Universität zu Lübeck
Direktor: Univ.-Prof. Dr. rer. nat. Andreas Ziegler

Statistische Qualitätssicherung von Hochdurchsatz-Genotypisierungsdaten

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

– Aus der Medizinischen Fakultät –

vorgelegt von
Arne Schillert, geb. Neumann
aus Leipzig

Lübeck, 2010

1. Berichterstatter: Prof. Dr. rer. nat. Andreas Ziegler

2. Berichterstatterin: Prof. Dr. med. Christine Klein

Tag der mündlichen Prüfung: 30.06.2010

Zum Druck genehmigt. Lübeck, 30.06.2010

gez. Prof. Dr. med. Werner Solbach
- Dekan der Medizinischen Fakultät -

Inhaltsverzeichnis

Abbildungsverzeichnis	iv
Tabellenverzeichnis	v
1 Einleitung	1
2 Affymetrix Microarrays	5
2.1 Molekulargenetische Grundlagen	5
2.2 Datengenerierung mit Affymetrix Microarrays	6
2.2.1 Synthese der Oligonukleotide	8
2.2.2 Ablauf eines Microarray-Experiments	9
2.2.3 Aufbau von Microarrays zur Genotypisierung	10
3 Genotyp-Bestimmung – Genotype Calling	15
3.1 Systematische Literaturrecherche	15
3.2 Vorverarbeitung der Daten	18
3.2.1 Quantilnormalisierung	19
3.2.2 RMA	20
3.2.3 PLIER	23
3.3 Beschreibung ausgewählter Algorithmen	23
3.3.1 Dynamic Model Algorithmus	24
3.3.2 RLMM und BRLMM	25
3.3.3 CRLMM	29
3.3.4 CHIAMO	31
3.3.5 JAPL	32
3.3.6 Birdseed	33
3.4 Systematischer Vergleich der ausgewählten Algorithmen	35
3.4.1 Vergleichsgrundlage: HapMap II	35

3.4.2	Genotype Calling	36
3.4.3	Ergebnisse	40
3.5	Diskussion	44
4	Cluster-Plots	46
4.1	Clustervaliditätsmaße	49
4.1.1	Kennzahlen der Kompaktheit	51
4.1.2	Kennzahlen der Verbundenheit	52
4.1.3	Kennzahlen der Separierbarkeit	52
4.1.4	Stabilität der Cluster	53
4.1.5	Kombination von Kennzahlen	54
4.2	Generierung von Cluster-Plots mit dem R-Paket <i>acpa</i>	60
4.2.1	Grundlagen	60
4.2.2	Aufbau des R-Paketes <i>acpa</i>	62
4.3	Automatische Bewertung	62
4.3.1	Vergleichsgrundlage: Gutenberg-Herz-Studie	63
4.3.2	Methoden	65
4.3.3	Ergebnisse	66
4.4	Diskussion	67
5	Zusammenfassung	69
	Literaturverzeichnis	71
	Lebenslauf	81
	Publikationsverzeichnis	83

Abbildungsverzeichnis

2.1	Aufbau der DNA	7
2.2	Prinzipskizze der Photolithographie	8
2.3	Aufbau eines Affymetrix GeneChip Microarrays	9
2.4	Schematische Darstellung des Hybridisierungsschrittes	11
2.5	Schematische Darstellung des Scannens	12
2.6	Falschfarbenbild der Intensitätswerte eines <i>probe sets</i>	13
3.1	Literaturrecherche – Schlagwörter	16
3.2	Literaturrecherche – Übersicht	17
3.3	Effekt Quantilnormalisierung	20
3.4	CCS-Transformation	28
3.5	Intensitäten getrennt nach Strängen	30
3.6	Flussdiagramm zum Calling mit CRLMM	38
3.7	Umkodierung der Calling-Ergebnisse für CRLMM	39
3.8	Flussdiagramm zum Calling mit JAPL	41
3.9	ADP für alle SNPs	42
3.10	ADP getrennt nach MAF	42
3.11	ADP getrennt nach homozygoten und heterozygoten Genotypen	43
4.1	Schematische Darstellung von Cluster-Plots	47
4.2	Cluster-Plot – Beispiele	49
4.3	Validitätskriterien für Cluster	51
4.4	Prinzip der Silhouette-Statistik	58
4.5	Cluster-Plots mit ACPA-Ellipsen	59
4.6	Struktur von <i>acpa</i>	63
4.7	ROC-Kurven der Clustermaße	66

Tabellenverzeichnis

2.1	Affymetrix Genotyping-Arrays	14
3.1	Datenbanken für die Literaturrecherche	15
3.2	Übersicht über die Populationen in Hapmap I + II	36
4.1	Der Algorithmus von ACPA	60
4.2	Häufigkeitsverteilung der Qualität der SNPs	65

1 Einleitung

Ziel der Arbeit

Genomweite Assoziationsstudien (GWA-Studien) haben sich in den letzten Jahren als Standardwerkzeuge der genetischen Epidemiologie etabliert. In diesen Studien wird der Zusammenhang zwischen einer komplexen genetischen Erkrankung oder einem quantitativen intermediären Phänotyp und Einzelbasen-Polymorphismen (*single nucleotide polymorphism* SNP) untersucht. Auf diesem Weg konnten in den letzten vier Jahren genetische Risikofaktoren für mehr als 60 komplexe genetische Erkrankungen identifiziert werden [49].

Die zugrundeliegende Technologie sind sogenannte Microarray-Experimente. Dabei werden bei der von der Firma Affymetrix verwendeten Methode Millionen von Oligonukleotiden auf einen Chip aufgebracht und fragmentierte und markierte DNA dazu gegeben. Sind die Sequenzen komplementär, kommt es zu einer Bindung der markierten Fragmente. Nachdem die ungebundene DNA abgewaschen wurde und an die Markierung ein Farbstoff gebunden hat, wird mittels Laserlicht der Farbstoff angeregt. Diese Fluoreszenzintensitäten werden durch Einscannen gemessen und in Graustufenbilder konvertiert. Diese numerisch gespeicherten Intensitätswerte werden dann in diskrete Genotypen umgewandelt. Diesen Vorgang der Genotypbestimmung nennt man Genotype Calling. Um die Güte dieser Zuordnung zu verbessern, wurde eine Vielzahl von Algorithmen entwickelt [5, 20, 41, 43, 53, 54, 72, 73].

Ein theoretischer sowie ein praktischer Vergleich dieser Verfahren ist bisher nicht erfolgt. Erstes Ziel dieser Arbeit ist daher eine systematische Betrachtung der Algorithmen zum Genotype Calling. Hierfür wird zunächst eine Literaturrecherche durchgeführt. Anschließend werden die am häufigsten verwendeten Algorithmen miteinander verglichen.

Obwohl in den vergangenen Jahren eine Reihe von Algorithmen zum Genotype Calling entwickelt wurde, sind diese Methoden nicht fehlerfrei. Daher ist eine stringente Qualitätskontrolle der Genotypdaten erforderlich [65, 74, 76]. Dazu gehört der Ausschluss von SNPs, die zu viele fehlende Werte nach der Genotypisierung aufweisen, eine zu geringe Häufigkeit des seltenen Allels besitzen oder bei denen die Wahrscheinlichkeit für eine Abweichung vom Hardy-Weinberg-Gleichgewicht groß ist. Als *Travemünde-Kriterien* haben sich bestimmte Werte für diese Grenzen etabliert [60]. Trotz dieser rigorosen Standard-Qualitätskontrolle gelingt es nicht, alle SNPs zu identifizieren, deren Genotypbestimmung fehlerhaft ist. Diese systematischen Fehler können sowohl zu einem erhöhten Anteil falsch-positiver Befunde als auch zu einer verringerten Wahrscheinlichkeit der Entdeckung wahrer Effekte führen [23].

Des Weiteren werden diese Genotypinformationen nicht nur in GWA-Studien verwendet, sondern sie bilden auch die Grundlage für komplexere statistische Analysen, wie z.B. Haplotypanalysen, der Imputation weiterer Genotypen oder Untersuchung von Gen-Gen- bzw. Gen-Umwelt-Interaktionen.

In der Haplotypanalyse wird der Zusammenhang von Blöcken von SNPs mit einem Phänotyp untersucht. Dieser Zusammenhang von SNPs, die dicht zusammen auf einem Chromosom liegen, wird auch für die Imputation weiterer SNPs verwendet. Dabei wird ausgenutzt, dass eine hohe Korrelation zwischen benachbarten SNPs besteht. Ist ein SNP eines Haplotyps nicht genotypisiert worden, weil dieser z.B. auf einem Chip nicht vorhanden ist, können die Genotypen dennoch durch die Haplotypstruktur "erschlossen", das heißt aufgefüllt bzw. imputiert werden. Die Güte der Imputation hängt dabei maßgeblich von der Qualität der zu Grunde liegenden Genotypen ab [27].

Als ultimative Qualitätskontrolle wird dabei die visuelle Auswertung der Cluster-Plots empfohlen [10, 65, 71, 74]. Bei dieser visuellen Analyse werden die Signalintensitäten für beide Allele gegeneinander aufgetragen und die Punkte entsprechend ihrem Genotyp eingefärbt. Sind auf der Abbildung dann nicht räumlich und farblich klar getrennte Gruppen zu erkennen, spricht das gegen ein fehlerfreies Calling der Genotypen. Da diese Auswertung sehr arbeits- und zeitintensiv ist, wird bisher auf die Bewertung aller Cluster-Plots einer GWA-Studie verzichtet; stattdessen werden nur die Abbildungen für Kandidaten-SNPs betrachtet. Dies ist jedoch, gerade im Hinblick auf die Anforderungen der Imputationen, nur eine Behelfslösung. Es besteht daher dringender Bedarf an der automatischen und objektiven Beurteilung der Cluster-Plots.

Das zweite Ziel dieser Arbeit besteht daher in der Entwicklung und Implementation eines Ansatzes zur automatischen Beurteilung von Cluster-Plots. Hierfür wurden Ideen aus dem Bereich der Clusteranalyse verwendet, aus der eine Vielzahl von Maßen bekannt ist, mit denen die interne Validität eines Clusterings, also auch eines Genotype Callings, beurteilt werden kann [35, 40]. Allerdings wurde bisher noch nicht untersucht, inwieweit die speziellen Eigenschaften von Cluster-Plots, wie das Wissen um die erwartete Lage der Cluster, diese Maße beeinflussen. Eine besondere Schwierigkeit sind bei GWA-Studien die Datenmengen, um zum einen diese Cluster-Plots zu generieren und zum anderen die Cluster-Maße zu berechnen.

Während das erste Ziel dieser Arbeit, die Betrachtung der Validität von Genotypen durch ein vom Bundesministerium für Bildung und Forschung (Förderkennzeichen: 01EZ0874) gefördertes Projekt motiviert wurde, entspricht das zweite Ziel dieser Arbeit einem Teil eines von der Deutschen Forschungsgemeinschaft geförderten Projekts (Förderkennzeichen: ZI 591/17-1).

Aufbau der Arbeit

Zur Bearbeitung der im vorangegangenen Abschnitt formulierten Ziele werden in Kapitel 2 zunächst neben den molekulargenetischen Grundlagen auch die Grundbegriffe zur Datengenerierung mit Affymetrix Microarrays erläutert. Dabei wird der Aufbau der Arrays beschrieben und ein Microarray-Experiment skizziert. Zum Abschluss dieses Kapitels wird gezeigt, wie die Menge eines DNA-Fragmentes in eine digitale Größe überführt wird.

Wie die Signalintensitäten in diskrete Genotypen überführt werden (Genotype Calling), wird in Kapitel 3 thematisiert. Es wird zuerst eine Literaturrecherche durchgeführt, um Algorithmen zum Genotype Calling zu identifizieren und anschließend deren charakteristische Eigenschaften beschrieben. Mittels einer empirischen Evaluation werden die Algorithmen anhand der HapMap-Daten [67, 68] miteinander verglichen. Dabei wird neben der Güte der Genotypbestimmung auch die Nutzbarkeit der einzelnen Programme berücksichtigt. Abschließend werden die Ergebnisse diskutiert und eine Empfehlung

für einen spezifischen Genotype-Calling-Algorithmus ausgesprochen.

Weiterführend wird nach der Genotypbestimmung die Qualitätsbewertung dieses Prozesses durchgeführt. Dazu werden zu Beginn des Kapitels 4 gängige Clustervaliditätsmaße und ein neu entwickeltes Maß zur automatischen Beurteilung der Güte des Genotype Calling beschrieben. Ein wesentlicher Bestandteil dieser Arbeit ist die Entwicklung eines Programms zur automatischen Generierung und Bewertung von Cluster-Plots. Dieses für die Programmiersprache R entwickelte Paket `acpa` wird vorgestellt und seine Nutzung anhand eines Beispiels demonstriert. In dieser Anwendung wird die Bewertung der SNPs anhand von Clustermaßen mit der Bewertung durch zwei erfahrene Beurteiler verglichen. So können Empfehlungen für die Verwendung von Clustermaßen gegeben werden.

2 Affymetrix Microarrays

Die vorliegende Arbeit untersucht die Generierung und Qualitätskontrolle von Genotypdaten für Affymetrix Microarrays. Die molekulargenetischen Grundlagen zum Verständnis der Bedeutung der Daten wird im ersten Abschnitt dieses Kapitels beleuchtet. Der zweite Abschnitt widmet sich den Microarrays von Affymetrix. Dabei wird auf deren Aufbau eingegangen und die typischen Schritte eines Microarray-Experimentes skizziert.

2.1 Molekulargenetische Grundlagen

Die DNA als Träger der Erbinformation

Die Desoxyribonukleinsäure (DNS, engl. DNA) ist Träger der genetischen Erbinformation und ein aus Desoxyribonukleotiden aufgebautes Biopolymer. Ein Nukleotid besteht dabei aus einem ringförmigen Zuckermolekül (der Desoxyribose), einem Phosphatrest und einer der vier möglichen Basen Adenin, Cytosin, Guanin und Thymin (A, C, G und T). Die DNA liegt meist doppelsträngig vor, das heißt, zwei DNA-Moleküle sind spiralförmig ineinander gewunden und bilden so die Struktur einer Doppelhelix. Da die Base Adenin nur mit Guanin und Cytosin nur mit Thymin eine Bindung eingehen kann, sind beide Stränge komplementär und antiparallel zueinander. In Abbildung 2.1 ist der Aufbau der DNA dargestellt. Zu erkennen ist das außen liegende Rückgrat aus Desoxyribose und Phosphatrest, sowie die innen liegenden Basen, welche mit der jeweiligen Base des komplementären Stranges durch Wasserstoffbrückenbindungen miteinander verbunden sind. Die genetische Information ist in der Nukleotidsequenz (der Abfolge

der Basen im DNA-Strang) festgelegt. Stark vereinfacht dargestellt, kodieren bestimmte Nukleotidsequenzen bestimmte Aminosäuren, welche wiederum die Grundbausteine von Proteinen darstellen. Detaillierte Beschreibungen dieser Zusammenhänge finden sich in [22, 75].

Variabilität

Humane Körperzellen besitzen einen doppelten Chromosomensatz. Ein Chromosom ist ein Molekül, welches einen Teil der Gesamt-DNA enthält. Bei der Vererbung erhält jeder Nachfahre so ein Chromosom von der Mutter und eins vom Vater. Unterscheiden sich Bereiche (*Loci*) auf diesen beiden Chromosomen voneinander, spricht man von *Allelen*. Gibt es genau zwei mögliche Varianten, nennt man dies einen *biallelischen Locus*. Für einen biallelischen Locus der DNA, mit den beiden möglichen Varianten *A* und *B* existieren genau drei mögliche Zustände: Homozygot für das *A*-Allel, also *A/A*, homozygot für das *B*-Allel (*B/B*) oder heterozygot *A/B*. Die häufigste Form der Variabilität ist der Austausch einer einzelnen Base. Kommt dieser Austausch häufiger als in einem Prozent der Population vor, so spricht man von einem *Single Nucleotide Polymorphism* (SNP) [69]. Die Erfassung der Variabilität der DNA in Form von SNPs ist Gegenstand der Microarray-Genotypisierungs-Experimente. In großen Projekten, wie dem HapMap-Projekt [67, 68], wurde die Anzahl der SNPs des menschlichen Genoms und ihre Verteilung untersucht. Dabei wurde die Anzahl der SNPs auf etwa 10 Millionen geschätzt [3]. Der aktuelle Microarray von Affymetrix, der *Genome-Wide Human SNP Array 6.0*, enthält 906.600 SNPs. Dabei wurde sich auf biallelische Marker beschränkt. Wie die Genotypinformation für eine Person simultan für diese Anzahl an SNPs anhand von Microarrays ermittelt werden kann, wird in den nächsten Abschnitten dargestellt.

2.2 Datengenerierung mit Affymetrix Microarrays

Wie in Kapitel 1 bereits erwähnt, ist das grundlegende Prinzip der Microarray-Experimente die Hybridisierung fragmentierter, markierter DNA an spezifische

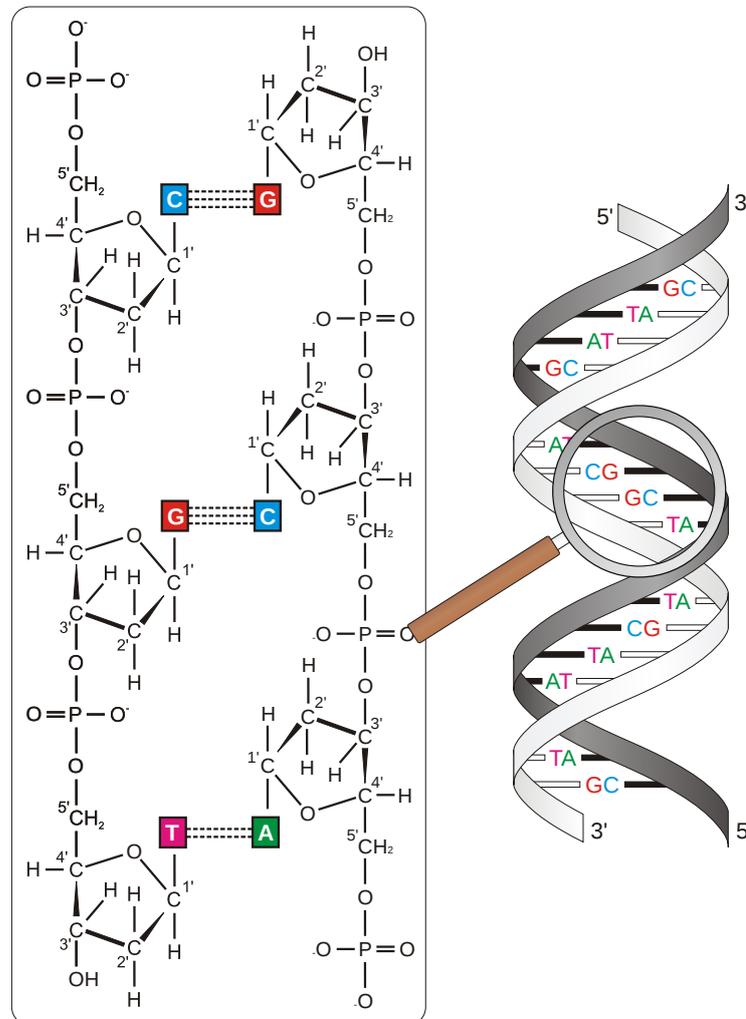


Abbildung 2.1: Schematischer Aufbau der DNA. Dargestellt ist die Strukturformel der DNA. Das Rückgrat eines DNA-Stranges wird aus den Desoxyribosemolekülen gebildet, welche über Phosphatreste miteinander verbunden sind. Die Basen sind farbig hervorgehoben. Die Wasserstoffbrückenbindungen zwischen den Basen (gestrichelte Linie) verbinden die beiden komplementären Stränge miteinander. A = Adenin, T = Thymin, C = Cytosin, G = Guanin. Die Abbildung entspricht Abbildung 1.1 aus [75].

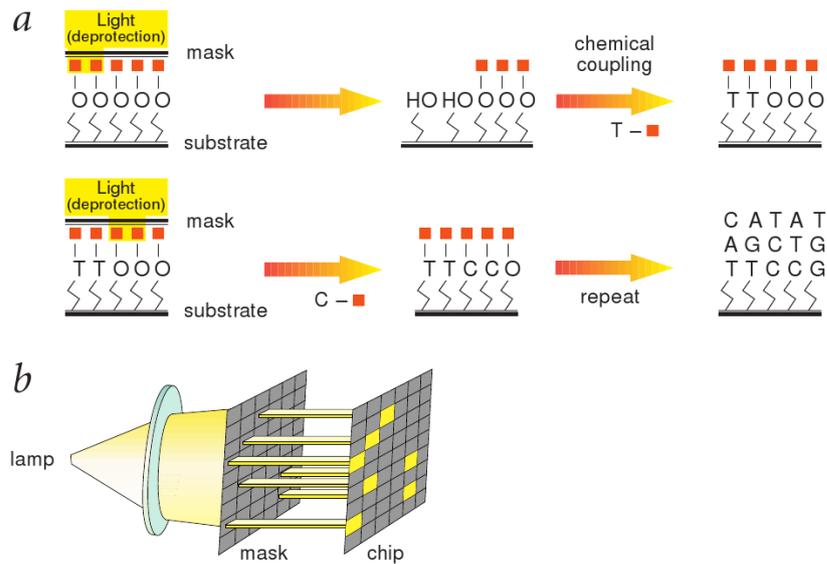


Abbildung 2.2: Prinzipskizze der *Photolithographie*. Die Abfolge der Schritte ist in Leserichtung dargestellt. Zu Beginn werden Bindungsstellen der Linker-Moleküle photochemisch freigelegt. Durch eine Maske wird sicher gestellt, dass nur bestimmte Linker-Moleküle freigelegt werden (Teil *b*). Anschließend kommt der Syntheseschritt, in dem eine der vier Basen hinzugefügt wird (hier Thymin). Die Bindungsstellen sind damit geschützt, so dass durch Wiederholung der beiden Schritte die Oligonukleotide sukzessive auf dem Array gebildet werden. Nachgedruckt mit Erlaubnis von Macmillan Publishers Limited: *Nat Genet*, [46]

Oligonukleotide (Sonden) und die anschließende indirekte Messung der hybridisierten DNA-Menge. Die Oligonukleotide, die auf dem Microarray auf einer Trägerfläche fixiert sind, bestehen aus einer Sequenz von 25 Nukleotiden, die komplementär zu der Sequenz ist, die den SNP umspannt [39]. Alternativ werden die Oligonukleotide auch *probes* genannt. Um jedoch Verwechslungen mit dem deutschen Begriff *Probe*, in diesem Zusammenhang die Daten einer Person, zu vermeiden, wird weitestgehend auf den Originalbegriff verzichtet und stattdessen *Sonde* verwendet.

2.2.1 Synthese der Oligonukleotide auf Affymetrix Microarrays

Die maßgebliche Technologie, die bei der Herstellung der Microarrays Anwendung findet ist die so genannte *Photolithographie* [46] (Abb. 2.2). In der Ausgangssituation befinden sich Linker-Moleküle auf dem Glasträger des Arrays, deren Bindungsstellen

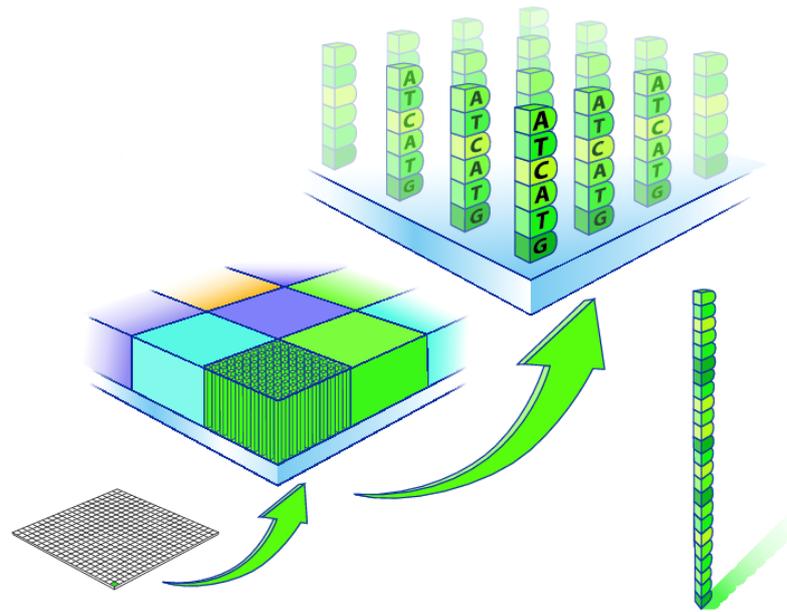


Abbildung 2.3: Schematischer Aufbau eines Affymetrix Microarrays. Die Anordnung der Oligomere auf dem Chip wird skizziert. Aus dem Bildarchiv von Affymetrix: [8].

zunächst durch lichtempfindliche, aber noch unreaktive Gruppen geschützt werden. Durch Auflegen spezieller Masken auf den Glasträger, werden unter Lichteinwirkung an vordefinierten Bereichen diese Bindungsstellen aktiviert. Anschließend wird das gewünschte Nukleotid hinzugefügt und reagiert mit den zuvor aktivierten Bindungsstellen. Das somit neu gebildete Dimer besitzt an seinem freien Ende wiederum eine lichtempfindliche, aber noch unreaktive Bindungsstelle. Durch mehrmaliges Auflegen neuer Masken werden die beschriebenen Schritte so oft wiederholt, bis das fertige Oligomer aus 25 Nukleotiden gebildet wurde [26]. In Abbildung 2.3 ist der Aufbau eines Microarray schematisch dargestellt.

2.2.2 Ablauf eines Microarray-Experiments

Zusammengefasst lässt sich ein Microarray-Experiment in folgende Arbeitsschritte unterteilen:

1. Restriktion der genomischen DNA mittels spezifischer Enzyme

2. Amplifikation (Vervielfältigung) der DNA mittels Polymerase-Ketten-Reaktion (PCR)
3. Aufreinigung der PCR-Produkte
4. Fragmentierung der PCR-Produkte
5. Markierung der DNA-Fragmente mit Biotin
6. Hybridisierung der Fragmente auf dem Microarray (Pro DNA-Probe ein Microarray)
7. Waschen der Arrays und Anfärbung mittels Fluoreszenzfarbstoff
8. Scannen der Arrays – Konvertierung der Fluoreszenzintensitäten in ein Graustufenbild (Abb. 2.5)
9. Zusammenfassung der Pixelinformationen zu Intensitäten pro Sonde – Generierung der CEL-Datei (Abb. 2.6)

2.2.3 Aufbau von Microarrays zur Genotypisierung

Microarrays zur Hochdurchsatz-Analyse biologischer Daten werden sowohl für Expressions-, als auch für Genotypisierungsexperimente verwendet. Das in Abschnitt 2.2.1 beschriebene Verfahren zur Herstellung der Oligonukleotide, und das in Abschnitt 2.2.2 skizzierte Prinzip der Experimente ist dabei vergleichbar, lediglich die Auswahl der Sonden unterscheidet sich. Während in Expressionsstudien die Menge eines bestimmten RNA-Moleküls mittels einer spezifischen Sonde gemessen werden kann, wird zur Unterscheidung der drei möglichen Genotypen eines biallelischen SNPs pro Allel eine Sonde benötigt. Um Aussagen über die Variabilität der Hybridisierung treffen zu können, werden in der Praxis mehrere Sonden pro Expressionsprodukt oder SNP verwendet [39], dies wird im Folgenden genauer für die Genotypisierungs-Microarrays beschrieben.

Die Sonden, die zur Bestimmung der Menge an DNA-Fragmenten für das A-,

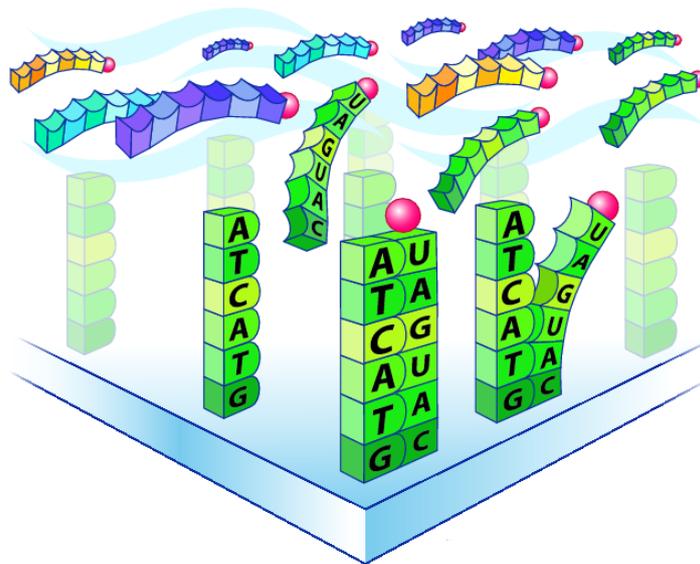


Abbildung 2.4: Schematische Darstellung des Hybridisierungsschrittes anhand eines Microarrays zur Expressionsanalyse. Die Biotin-markierten einzelsträngigen DNA-Fragmente werden auf den Microarray übertragen und dort über einen Zeitraum von 16-18h unter Rotation belassen (Hybridisierung). Während der Zeit der Hybridisierung paaren die einzelsträngigen, markierten DNA-Fragmente mit der entsprechend komplementären Sonde auf dem Microarray und werden somit ebenfalls auf dem Array fixiert. Nicht gebundene Fragmente werden in den folgenden Waschschrritten vom Array entfernt. Im nachfolgenden Färbeschritt wird eine Lösung mit Fluoreszenzfarbstoff (Streptavidin-Phycoerythrin) hinzugefügt, welcher an die Biotin-Moleküle der DNA-Fragmente bindet. Durch Anregung mit Laserenergie während des Scan-Vorgangs können an den Array fixierten DNA-Fragmente nachgewiesen werden. Aus dem Bildarchiv von Affymetrix [8].

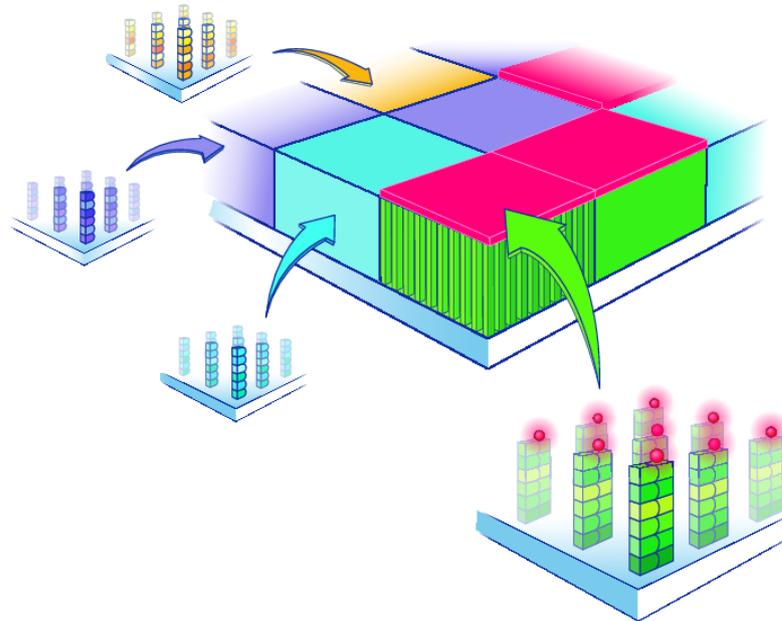


Abbildung 2.5: Schematische Darstellung des Scan-Prozesses. Bereiche, in denen viele DNA-Fragmente gebunden haben, fluoreszieren nach Anregung stärker und sind hier rot dargestellt. Aus dem Bildarchiv von Affymetrix [8].

bzw. das B-Allel eingesetzt werden, sind die so genannten *perfect match probes* (PM-Sonden). Bei diesen wird die 13. Position des Oligomers variiert, so dass die Gesamtsequenz entweder vollständig komplementär zur Sequenz des A-Allels für diesen SNP ist oder komplementär zum B-Allel. Nur Biotin-markierte DNA-Fragmente mit der entsprechenden übereinstimmenden Sequenz (Vorliegen des A- oder des B-Allels) können genügend stark und spezifisch an die entsprechende Sonde binden und geben nach Färbung der Arrays ein Fluoreszenzsignal ab. Liegt keine Übereinstimmung in der Sequenz vor, wird das Fragment mit den folgenden Waschschrritten wieder vom Array entfernt. Im Falle einer Heterozygotie (Vorliegen von Allel A und B) binden die entsprechenden DNA-Fragmente an beide Sonden. Um Aussagen über die Variabilität der Hybridisierung treffen zu können, werden in der Praxis mehrere Sonden pro SNP (oder pro Expressionsprodukt) verwendet [39], welche gleichmäßig auf dem gesamten Array verteilt sind. Im Falle des Affymetrix Genome-Wide Human SNP Array 6.0 sind dies sechs Sonden (3 x A und 3 x B). Auf älteren Microarrays kommen zusätzlich *mismatch probes* (MM-Sonden) zum Einsatz. Diese unterscheiden sich von den PM-Sonden nur in der Wahl der Base an dem SNP, es werden Basen verwendet, die zu keinem der beiden Allele komplementär sind. Damit soll die Korrektur für

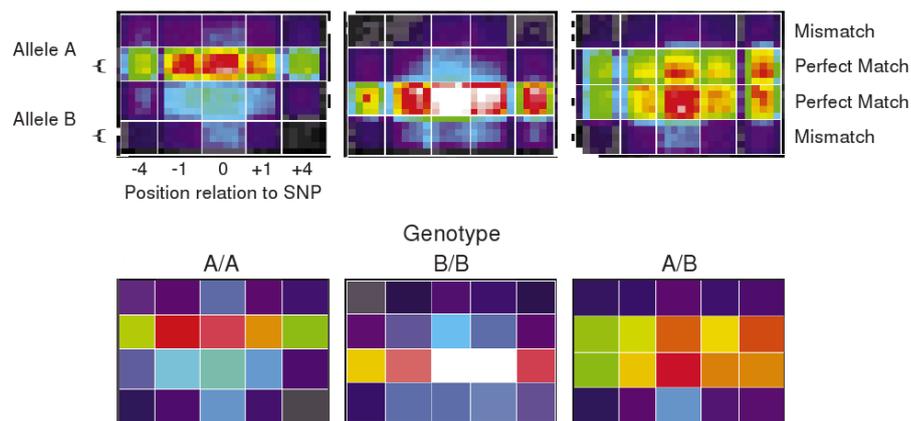


Abbildung 2.6: Falschfarbenbild der Intensitätswerte eines *probe sets*. In der oberen Reihe von links nach rechts ist das Muster für einen SNP mit Genotyp AA, BB bzw. AB gezeigt. Höhere Bindung an der Sonde und damit erhöhte Fluoreszenzintensität ist durch rote Färbung gekennzeichnet, Blautöne repräsentieren geringe Fluoreszenzintensitäten. Wie theoretisch erwartet, findet kaum Hybridisierung an den MM-Sonden statt. Deutlich zu erkennen ist die abnehmende Intensität, wenn sich der SNP außerhalb der Mitte des Oligonukleotides befindet. Für einen SNP mit Genotyp AA zeigen die PM-Sonden für das A-Allel hohe Intensitäten an, während an die PM-Sonden des B-Alles kaum DNA hybridisiert hat. Für den Genotypen BB ist die Situation genau umgekehrt und für heterozygote Genotypen, entsprechen die Intensitäten des A-Allels in etwa dem des B-Allels. In der unteren Reihe wurden die Pixelintensitäten zu einem Wert pro *Probe* zusammengefasst. Dies ist der maßgebliche Schritt bei der Konvertierung der DAT-Dateien in CEL-Dateien. Nachgedruckt mit Erlaubnis von Macmillan Publishers Limited: *Nat Genet*, [46]

Hintergrundrauschen ermöglicht werden [46]. Der Nutzen dieser MM-Sonden wird jedoch angezweifelt [70] und auf Arrays der neueren Generation werden keine MM-Sonden mehr verwendet (Tab. 2.1). Alle Sonden, die zur Bestimmung des Genotyps eines SNPs verwendet werden, werden bei der Analyse als *probe set* zusammengefasst. In Abbildung 2.6 sind in einem Falschfarbenbild die zu erwartenden Signalintensitäten eines *probe sets* für die drei möglichen Genotypen dargestellt. Eine zusätzliche Ebene der Komplexität entsteht beim Einscannen der Arrays, da die Intensitätsinformationen einer Sonde in den Farbwerten mehrerer Pixel abgespeichert sind. Die Zusammenfassung dieser Werte zu einem Intensitätswert pro Oligomer geschieht in der Konvertierung der Affymetrix *DAT*-Dateien in die Affymetrix *CEL*-Dateien. In den letzten Jahren wurde die Anzahl der SNPs pro Array von Affymetrix konstant erhöht (Tab. 2.1). Vom 10K Array zum SNP Array 6.0 auf erfolgte eine Steigerung auf das Einhundertfache. Dies wurde maßgeblich durch Verringerung der Redundanzen und Verzicht auf MM-Sonden erreicht. Algorithmen, die auf die MM-Informationen verzichten, sind

Tabelle 2.1: Übersicht über die Affymetrix Microarray zur Genotypisierung

Name	Kurzbeschreibung
Human Mapping 10K 2.0 Array	10.204 SNPs, PM+MM
Human Mapping 100K Set	116.204 SNPs verteilt auf 2 Arrays, PM+MM
Human Mapping 500K Array Set	500.568 SNPs verteilt auf 2 Arrays, PM+MM
Genome-Wide Human SNP Array 5.0	500.568 SNPs und 420.000 CNV-Sonden, nur PM
Genome-Wide Human SNP Array 6.0	906.600 SNPs und 946.00 CNV-Sonden, nur PM
Axiom Genotyping Solution	567.096 SNPs, komplettes Neudesign der Plattform , Informationen über die genaue Datenstruktur lagen noch nicht vor

PM: *perfect match*-Sonden

MM: *mismatch*-Sonden

CNVs: *copy number variations*

somit eher für eine Anwendung auf verschiedenen Array-Typen geeignet.

3 Genotyp-Bestimmung – Genotype Calling

Zur Identifikation von Calling Algorithmen wurde eine Literaturrecherche durchgeführt. Die Details zu den Suchbegriffen, abgefragten Datenbanken und identifizierten Algorithmen werden im Abschnitt 3.1 dargestellt. Nach einer Beschreibung der Standardverfahren der Datenvorverarbeitung (Abschnitt 3.2) werden in Abschnitt 3.3 die charakteristische Eigenschaften der Algorithmen herausgearbeitet. Mittels eines Realdatensatzes werden in Abschnitt 3.4 die Algorithmen miteinander verglichen und diese Ergebnisse abschließend diskutiert (Abschnitt 3.5).

3.1 Systematische Literaturrecherche

Um einen Überblick über existierende Genotype Calling Algorithmen zu erhalten, wurde am 14. Januar 2009 eine systematische Literaturrecherche durchgeführt. Die Liste der abgefragten Datenbanken ist in Tabelle 3.1 aufgeführt.

Basierend auf den Zusammenfassungen von Artikeln zu bereits bekannten

Tabelle 3.1: Für die Literaturrecherche verwendete Datenbanken.

Datenbank	Internetadresse
Medline	http://www.ncbi.nlm.nih.gov/pubmed
Web of Science	http://isiknowledge.com
Citeseerx	http://citeseerx.ist.psu.edu
Google Scholar	http://scholar.google.de

Calling-Algorithmen und Vorschlägen von Mitgliedern der Arbeitsgruppe „Quality-Management for High-Throughput Genotyping“ der *Telematik-Plattform für medizinische Forschungsnetze* (TMF) wurden Schlagworte ausgewählt, um Publikationen zu finden, in den Calling-Algorithmen für die Affymetrix-Plattform beschrieben werden. Dazu ist die Kombination der Suchbegriffe in Abbildung 3.1

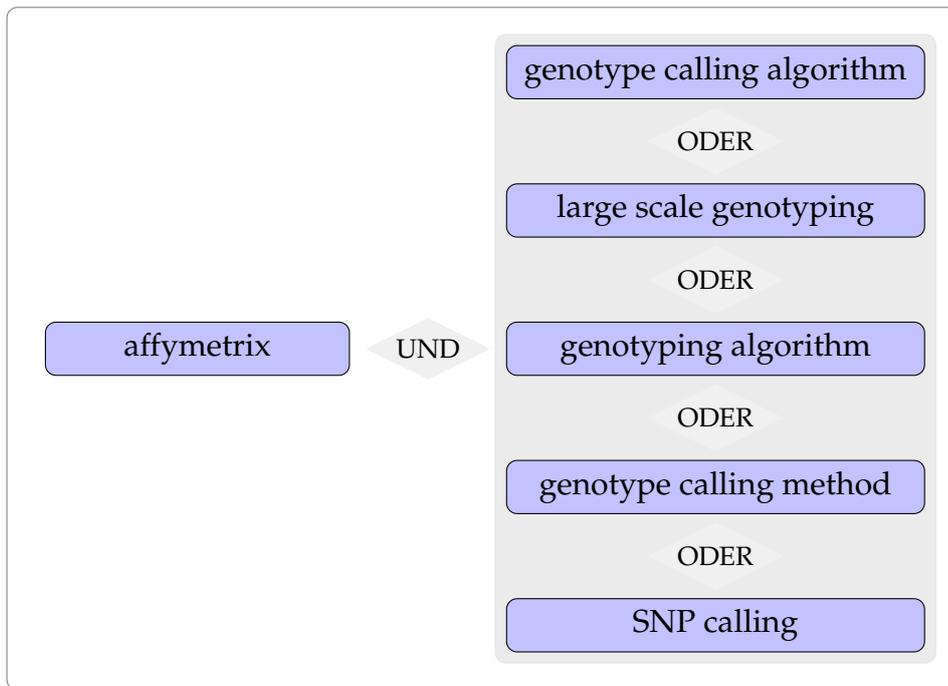


Abbildung 3.1: Kombination der Schlagwörter für die systematische Literaturrecherche.

Suchbegriffe werden durch die logischen Operatoren in den Rauten miteinander verbunden. Insgesamt gibt es also fünf Möglichkeiten die Begriffe zu kombinieren.

aufgelistet. Die Ergebnisse zu den einzelnen Abfragen wurden mittels *Endnote* [1] und *Zotero* [2] gespeichert und dann in *Endnote* zusammengeführt und Duplikate entfernt. Anschließend wurden in einem zweistufigen Auswahlverfahren die relevanten Publikationen identifiziert. Dabei wurden im ersten Schritt 504 Artikel anhand ihrer Zusammenfassungen ausgeschlossen. Im zweiten Schritt wurde die Anzahl der Publikationen von 18 auf 9 reduziert. Eine detaillierte Aufschlüsselung dieses Ablaufs ist in Abbildung 3.2 gegeben. Für die identifizierten neun Algorithmen wurde versucht, die Programme zu erhalten und nach Installation in Betrieb zu nehmen. Für SNIper-HD [36] schlug die Kompilierung fehl und auf eine Anfrage wurde nicht reagiert. Der Code für GEL [53] und MAMS [73] ist nur auf Anfrage erhältlich. Auf gestellte Anfragen für beide Programme wurde

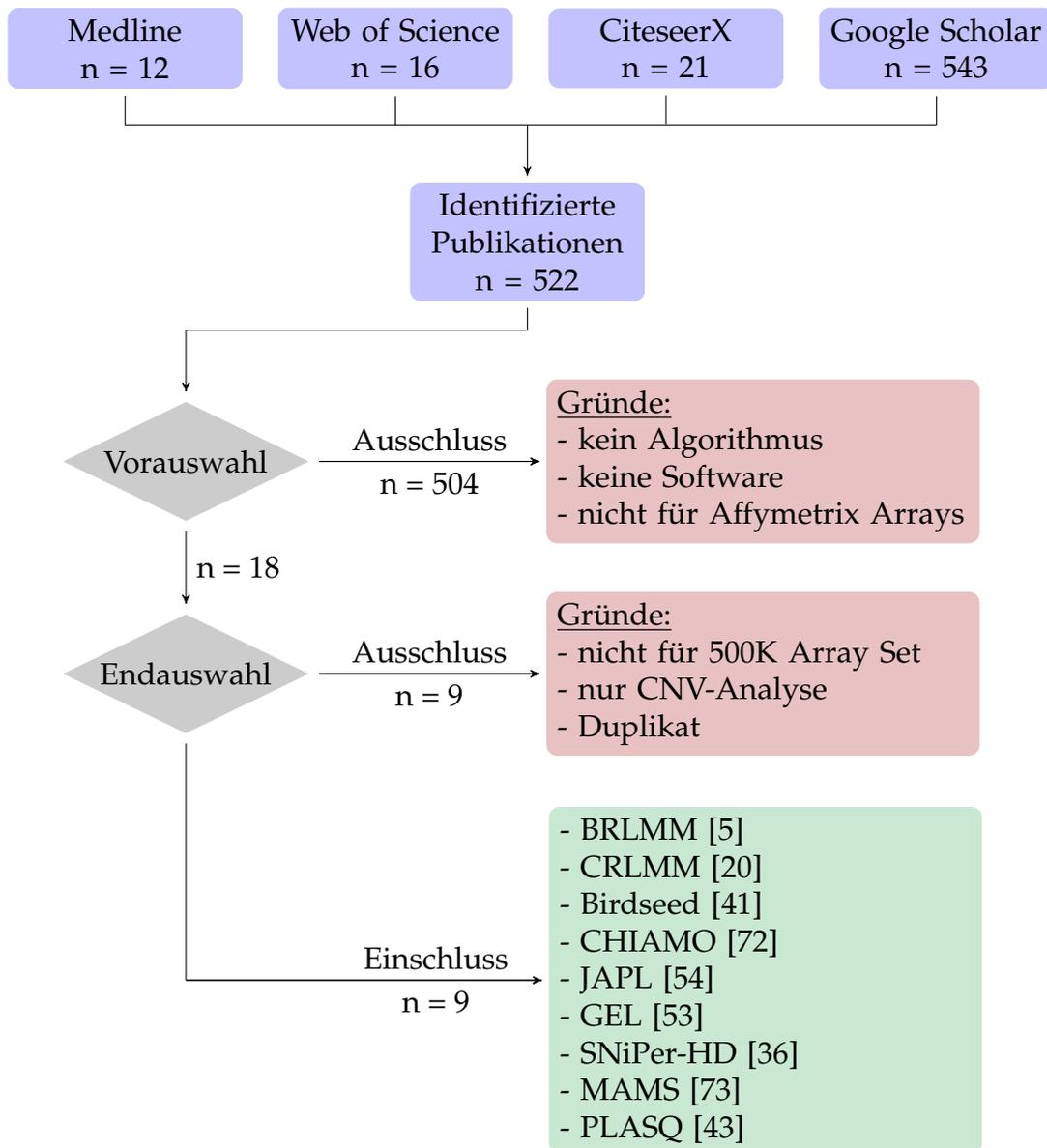


Abbildung 3.2: Schematische Übersicht über den Ablauf der Literaturrecherche. Die Recherche in den vier oben aufgeführten Datenbanken ergab 522 eindeutige Publikationen. In einer Vorauswahl wurden aufgrund der Zusammenfassungen der Publikationen 503 Publikationen ausgeschlossen, da sie entweder keinen neuen Algorithmus beschrieben, keine Software angeboten oder sich nicht für Affymetrix Arrays eigneten. Im zweiten Filterschritt wurden die Artikel ausgeschlossen, deren Algorithmen nicht für das 500k Array Set geeignet waren [29, 37, 44, 57]. Weiterhin wurden zwei Arbeiten als inhaltlich identisch zu bereits aufgenommenen Arbeiten erkannt und eine Arbeit war nicht auffindbar. In zwei weiteren Publikationen ging es um die Analyse von *copy number variations* (CNVs) [58, 61]. Damit blieben neun Algorithmen, die sowohl theoretisch beschrieben als auch praktisch evaluiert werden sollten.

jedoch nicht geantwortet. Der Verweis auf die Internetseite von PLASQ [43] war veraltet, so dass auch dieses Programm nicht mit in die Untersuchung eingehen konnte.

Damit fließen folgende fünf Algorithmen in die Auswertung ein:

1. BRLMM
2. CRLMM
3. Birdseed
4. CHIAMO
5. JAPL

Die charakteristischen Eigenschaften der Algorithmen werden in Abschnitt 3.3 beleuchtet. Zunächst folgt jedoch eine Beschreibung des Schrittes, der dem eigentlichen Genotype Calling vorangeht: Die Vorverarbeitung der Daten.

3.2 Vorverarbeitung der Daten

Formal gesehen ist der Prozess des Genotype Calling eine Funktion, die ein Intensitätspaar (I_A, I_B) in einen Genotypen k mit $k \in \{0, 1, 2\}$ konvertiert. Diese Intensitätspaare liegen so jedoch nicht in den CEL-Dateien vor, sondern müssen erst aus den Intensitätsinformationen eines *probe sets* bestimmt werden.

Aufgrund der indirekten Messung der DNA-Menge über Signalintensitäten wird zusätzliche Variabilität eingeführt, welche die Bestimmung der Genotypen erschwert. Werden für einen Calling-Algorithmus die Daten von mehreren Arrays gleichzeitig genutzt, muss man dafür Sorge tragen, dass die Verteilungen der Intensitäten zwischen den Arrays vergleichbar sind. Aus diesen Notwendigkeiten resultieren die folgenden drei Vorverarbeitungsschritte.

1. Hintergrundkorrektur engl. *background correction*
2. Normalisierung engl. *normalisation*

3. Zusammenfassung engl. *summarisation*

Nur die Zusammenfassung der Intensitäten zu Intensitätspaaren bzw. in Ausnahmen Intensitätsquartetts [20] ist zwingend notwendig. Auch die Reihenfolge der einzelnen Schritte kann variieren. Es scheint sich jedoch etabliert zu haben, die Zusammenfassung der Intensitätswerte als finalen Schritt der Datenvorverarbeitung anzuwenden, da dies für alle hier beschriebenen Algorithmen der Fall ist. In den folgenden Abschnitten werden mit der Quantilnormalisierung und dem *Robust Multi Array Averaging* zwei populäre Verfahren zur Normalisierung, bzw. Hintergrundkorrektur und Zusammenfassung vorgestellt.

3.2.1 Quantilnormalisierung

Das Standardverfahren zur Reduzierung der Variabilität zwischen Arrays ist die Quantilnormalisierung von Bolstad et al. [18]. Die Idee hinter der Methode entstammt den Quantil-Quantil-Diagrammen (*QQ-Plots*), bei denen die entsprechenden Quantile von zwei Verteilungen jeweils gegeneinander aufgetragen werden. Sind diese Verteilungen gleich, liegen alle Punkte auf der Hauptdiagonalen. Das Ziel der Quantilnormalisierung ist, alle n Datensätze (Arrays) so zu transformieren, dass sie danach auf der n -dimensionalen Einheitsdiagonalen $d = (\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ liegen. Dazu sei $\mathbf{q}_k = (q_{k1}, \dots, q_{kn})$ der Vektor der k -ten Quantile mit $k = 1, \dots, p$, wobei p die Anzahl der Abstufungen ist. Der Projektion der Quantile auf die Hauptdiagonale entspricht dann gerade

$$proj_d \mathbf{q}_k = \left(\frac{1}{n} \sum_{j=1}^n q_{kj}, \dots, \frac{1}{n} \sum_{j=1}^n q_{kj} \right),$$

das heißt, die Verteilung aller Arrays wäre die gleiche, würde man die ursprünglichen Quantile durch das durchschnittliche Quantil aller Arrays ersetzen.

Die Transformation lautet $x'_i = F^{-1}(G(x_i))$, wobei G die empirische Verteilungsfunktion der Werte eines Arrays ist und F die empirische durchschnittliche Verteilungsfunktion. Der Effekt dieser Transformation wird anhand eines Beispiels verdeutlicht. Dazu werden die empirischen Dichtefunktionen der Intensitäten aller PM-

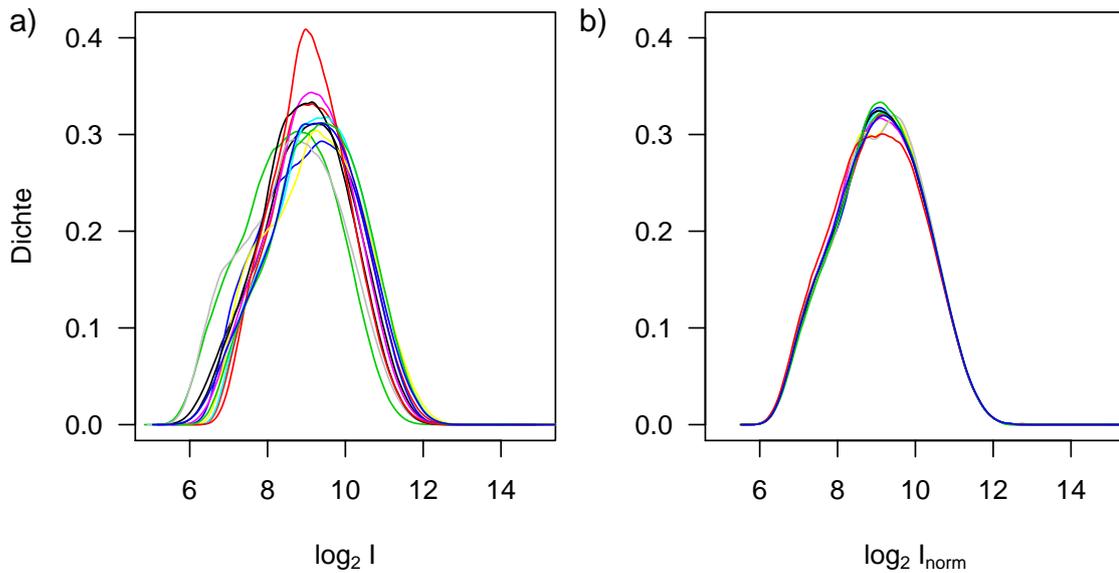


Abbildung 3.3: Beispiel für den Effekt der Quantilnormalisierung. Verwendet wurden 10 Proben des *Nsp*-Chips des Human Mapping 500K Array Sets von Affymetrix für die HapMap-Proben. Die Daten sind in Abschnitt 3.4.1 näher beschrieben. In Abbildung a sind die empirischen Dichtefunktionen über alle PM-Intensitäten einer Probe aufgetragen. Jede Kurve repräsentiert eine Probe. Abbildung b zeigt die empirischen Dichtefunktionen nach der Quantilnormalisierung. Die Proben folgen jetzt einer annähernd gleichen Verteilung. Die Quantilnormalisierung und die Darstellung der Dichten wurden mit Funktionen des Bioconductor-Paketes `aroma.affymetrix` [16] durchgeführt bzw. erstellt.

Sonden einer Probe bestimmt. In Abbildung 3.3 ist dies für die HapMap-Daten (siehe Abschnitt 3.4.1) geschehen. Während links die untransformierten Dichtefunktionen deutliche Unterschiede in Lage und Form zwischen den Proben zeigen, stellt sich rechts, nach der Quantilnormalisierung ein homogenes Bild dar.

3.2.2 Robust Multi Array Averaging (RMA)

Das von Irizarry et al. [38] entwickelte Verfahren *Robust Multi Array Averaging* (RMA) ist eine spezielle Kombination der drei zu Beginn von Abschnitt 3.2 erwähnten Vorverarbeitungsschritte. Dabei wird nur die Information der PM-Sonden verwendet. Da RMA ursprünglich für die Analyse von Expressionsdaten entwickelt wurde, wird sich hier auf Transkripte statt auf SNPs bezogen. Der Aufbau der Arrays ist jedoch vergleichbar (siehe Abschnitt 2.2), die Methoden

also übertragbar. Sei nun PM_{ijn} die Perfect-Match-Intensität der i -ten Probe und

der j -ten Sonde für ein Transkript n . Diese lässt sich laut Irizarry et al. wie folgt zerlegen:

$$PM_{ijn} = bg_{ijn} + sig_{ijn}. \quad \text{mit } bg_{ijn} : \text{Hintergrundrauschen} \\ sig_{ijn} : \text{Signal}$$

Unter dem Hintergrundrauschen bg_{ijn} werden optisches Rauschen beim Scan-Prozess und nicht-spezifische Bindungen von DNA-Fragmenten an die Sonden zusammengefasst. Unter der Annahme, dass dieses Hintergrundrauschen über alle Transkripte eines Arrays konstant ist, das heißt $E(bg_{ijn}) = \beta_i$, soll dieses nun entfernt werden. Dazu wird eine Funktion $B(\cdot)$ mit

$$B(PM_{ijn}) \equiv E(sig_{ijn} | PM_{ijn})$$

gesucht. Für Signalintensitäten sig_{ijn} die positiv sind, ist auch $B(PM_{ijn})$ immer größer 0. Eine Lösung für $B(\cdot)$, die in geschlossener Form darstellbar ist, erhält man unter der Annahme, dass die Signalintensitäten sig exponentialverteilt sind und das Hintergrundrauschen einer Normalverteilung folgt. Irizarry et al. [38] zeigen, dass diese Wahl zur Bestimmung von $B(\cdot)$ für Realdaten gut funktioniert. Die nach dieser Hintergrundkorrektur, Quantilnormalisierung und Transformation mit dem binären Logarithmus erhaltene Intensität Y_{ijn} wird nun verwendet, um daraus die Expressionseffekte zu schätzen. Dazu wird folgendes lineares Modell angenommen:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

Dabei ist μ_{in} der interessierende Expressionswert, α_{jn} der Sondeneffekt und $\varepsilon_{ij} \stackrel{u.i.v.}{\sim} N(0, \sigma^2)$ ein Fehlerterm. Zur eindeutigen Bestimmbarkeit des Modells wird weiterhin angenommen, dass $\sum_j \alpha_j = 0$ für alle *probe sets*. Diese Annahme besagt, dass die gewählten Sonden im Durchschnitt tatsächlich die Expression des Gens messen. Dieses lineare Modell wird nun mittels *Median Polish* [51] gelöst. Dies ist vergleichbar mit einer robusten Variante der ANOVA, die deutlich weniger empfindlich auf Ausreißer reagiert.

3.2.3 Probe Logarithmic Intensity Error (PLIER) estimation

Die *Probe Logarithmic Intensity Error estimation* (PLIER) ist das von Affymetrix eingesetzte Verfahren, um die Intensitäten der *probe sets* zusammenzufassen [6]. Dabei wird besonders Wert darauf gelegt, auch für Proben mit niedrigen Signalintensitäten die Genotypen gut voneinander unterscheiden zu können. In der Nomenklatur von Affymetrix werden die Sonden *features* genannt. Die zur Modellierung der systematischen Fehler benötigten Parameter heißen *feature responses*. Diese sind Skalierungsfaktoren, die das Bindungsverhalten der einzelnen Sonden innerhalb eines *probe sets* beschreiben sollen. Nachdem die *feature responses* bestimmt wurden, kann so zwischen *high performance* und *low performance* Sonden unterschieden werden. Die Intensität eines *probe sets* wird also maßgeblich von den Sonden bestimmt, die konsistentes Bindungsverhalten über alle DNA-Mengen zeigen. Der Beitrag von Sonden mit widersprüchlichen Werten wird mittels der *Geman-McClue*-Funktion verringert. Details dazu finden sich im Anhang von [6]. Die Zuordnungen für jedes *probe set* werden anhand von Trainingsdaten bestimmt und in dem sogenannten *model file* abgespeichert.

Die Wahl der Hintergrundkorrektur steht dem Benutzer frei, je nach Chip können Verfahren gewählt werden, die PM- und MM-Proben verwenden, beziehungsweise Verfahren die analog zu RMA nur die Intensitätsinformationen der PM-Sonden verwenden. Als Normalisierung wird die Quantilnormalisierung empfohlen, wobei jedoch frei gestellt wird, ob diese vor oder nach der Zusammenfassung der Intensitäten eines *probe sets* erfolgen soll.

3.3 Beschreibung ausgewählter Algorithmen

Da die einzelnen Algorithmen in den jeweiligen Referenzarbeiten ausführlich beschrieben wurden, soll hier auf eine Wiedergabe sämtlicher statistisch-technischer Details verzichtet werden. Es werden vielmehr wesentliche Modell Aspekte diskutiert, die einen Vergleich der Algorithmen ermöglichen.

3.3.1 Dynamic Model Algorithmus

Aus historischen Gründen wird zu Beginn der *Dynamic Model* (DM)-Algorithmus von Di et al. [29] beschrieben. Dies ist der erste Algorithmus, der den Anforderungen der Hochdurchsatz-Genotypisierung gerecht wurde und für das Affymetrix GeneChip Human Mapping 100K SNP Array Set konzipiert war. Der für den GeneChip Mapping 10K Array entwickelte Algorithmus MPAM [47] erwies sich als problematisch für den 100K Array, da MPAM hohe Anforderungen an die Allelfrequenz und die Stichprobengröße stellte [57]. Der DM-Algorithmus zeigte diese Probleme nicht mehr und bestimmt die Genotypen generell sehr gut. Allerdings ist die Fehlerhäufigkeit bei der Bestimmung heterozygoter Genotypen höher als bei der Bestimmung von homozygoten Genotypen [57]. Da der DM-Algorithmus jedoch der beste Algorithmus zur Genotypbestimmung ist, der die Arrays einzeln analysiert, wird er bei den neueren Multi-Array-Algorithmen als Qualitätskontrolle vorgeschaltet, um Arrays zu identifizieren und auszuschließen, bei denen die Hybridisierung oder Ähnliches fehlgeschlagen ist. Weiterhin werden die Genotypen aus DM als Startwerte für die Genotypbestimmung mit anderen Algorithmen verwendet. Daher wird jetzt die Methode des DM-Algorithmus beschrieben. In diesem Algorithmus werden die Genotypen jedes SNPs für jeden Array separat bestimmt. Dazu werden die Intensitäten eines *probe quartets*, bestehend aus $\{PM_A, PM_B, MM_A, MM_B\}$ in eine der folgenden vier Klassen eingeteilt:

1. *Null*: Alle vier Sonden zeigen vergleichbare Intensitäten und werden als Hintergrund behandelt.
2. *A*: Nur PM_A ist signifikant größer als die anderen drei und wird deswegen als Signal und $\{PM_B, MM_A, MM_B\}$ als Hintergrund behandelt.
3. *B*: Analog zu 2. Nur mit PM_B anstelle von PM_A .
4. *AB*: $\{PM_A, PM_B\}$ sind signifikant größer als $\{MM_A, MM_B\}$.

Für jedes dieser vier Modelle wird nun die Wahrscheinlichkeit bestimmt. Details dazu finden sich in [25, 29]. Anhand dieser Klassenwahrscheinlichkeiten

$L(1), \dots, L(4)$ wird wie folgt ein Punktwert für jedes dieser Modelle m gebildet:

$$S(m) = L(m) - \max\{L(k), k = 1, 2, 3, 4, k \neq m\}$$

Ist $S(m)$ für ein Quartett größer als Null, dann wird das Modell m von den Daten dieses Quartetts am ehesten unterstützt. Anschließend wird für den Punktwertvektor der n Quartette $V_m = (S_1(m), \dots, S_n(m))$ folgende Hypothese mit Hilfe des *Wilcoxon signed rank test* getestet:

$$H_0 : \text{median}(V_m) = 0 \quad \text{vs.} \quad H_1 : \text{median}(V_m) > 0.$$

Dies wird für jedes Modell durchgeführt, so dass man vier p -Werte $\{p_1, p_2, p_3, p_4\}$ erhält. Der niedrigste p -Wert wird dann als Qualitätsmetrik für die Genotypbestimmung verwendet. Unterschreitet er eine vorher festgelegte Grenze α , so wird der Genotyp des zugehörigen Modells als Genotyp für diesen SNP verwendet.

3.3.2 RLMM und BRLMM

Im Jahr 2006 haben Rabbee und Speed einen Algorithmus zur Genotypbestimmung für das Affymetrix GeneChip Human Mapping 100K SNP Array Set vorgestellt [57]. Dieser sollte zum einen die Problematik der unterschiedlichen Missklassifizierungshäufigkeiten zwischen heterozygoten und homozygoten Genotypen beheben und zum anderen eine höhere Güte dadurch erreichen, dass Informationen für einen SNP über mehrere Arrays hinweg genutzt werden. Das für den Algorithmus verwendete Akronym RLMM steht für ein Robustes Lineares Modell unter Verwendung von Mahalanobis-Abständen. BRLMM ist die um einen Bayes'schen Schritt erweiterte Version, mit der bessere Schätzung der Lage- und Streumaße der Genotypcluster erreicht werden sollen [5]. Da BRLMM auf RLMM aufbaut, wird zuerst die Funktionsweise von RLMM skizziert und anschließend die Modifikationen in BRLMM erläutert. Wie von Rabbee und Speed [57] beschrieben, läuft RLMM in folgenden drei Schritten ab:

Vorverarbeitung der Daten

Nach der Quantilnormalisierung werden die Intensitäten in RLMM per RMA zusammengefasst. Dadurch erhält man für jeden SNPs des i -ten Arrays ein Intensitätspaar $\theta_i = (\theta_{Ai}, \theta_{Bi})$.

Bildung der Entscheidungsregionen

Mit Hilfe eines Trainingsdatensatzes werden nun Regionen berechnet in denen das Intensitätspaar eines neuen Arrays liegen sollte, um zu diesem Genotyp zugeordnet zu werden. Dazu wird der Vektor der Clusterzentren \mathbf{m} und der Vektor der Streuungsparameter \mathbf{S} berechnet, mit

$$\mathbf{m} = \left(m_A^{AA}, m_B^{AA}, m_A^{AB}, m_B^{AB}, m_A^{BB}, m_B^{BB} \right) \quad \text{und}$$

$$\mathbf{S} = \left((s_A^2)^{AA}, (s_B^2)^{AA}, (r)^{AA}, (s_A^2)^{AB}, (s_B^2)^{AB}, (r)^{AB}, (s_A^2)^{BB}, (s_B^2)^{BB}, (r)^{BB} \right).$$

Die Entscheidungsregion eines Genotypclusters wird dabei durch den Schwerpunktvektor \mathbf{m}_g und die 2×2 Kovarianzmatrix \mathbf{S}_g definiert.

Für SNPs mit einer niedrigen Allelfrequenz des selteneren Allels (MAF, *minor allele frequency*) ist diese Konstruktion der Entscheidungsregionen nicht möglich, da nicht alle Genotypcluster besetzt sind, beziehungsweise zu wenige Proben zur Schätzung der Parameter zur Verfügung stehen. Daher wird für 5000 zufällig ausgewählte SNPs, bei denen alle Genotypcluster besetzt sind, \mathbf{m} und \mathbf{S} bestimmt. Stark vereinfacht ausgedrückt, werden nun diese Schätzwerte unter Verwendung der Annahme der multivariaten Normalverteilung für eine Regressionsanalyse verwendet. Daraus lassen sich dann die fehlenden Parameter für den Genotypcluster schätzen und somit die Entscheidungsregionen bilden.

Zuordnung neuer Proben

Soll nun der Genotyp einer neuen Probe bestimmt werden, so wird der Mahalanobis-Abstand $D_g^2(\boldsymbol{\theta})$ des Intensitätsvektors $\boldsymbol{\theta}$ dieser Probe zu den Zentren der Genotypcluster berechnet. Genauer wird

$$D_g^2(\boldsymbol{\theta}) = \sqrt{(\boldsymbol{\theta} - \mathbf{m}_g)^T \mathbf{S}_g^{-1} (\boldsymbol{\theta} - \mathbf{m}_g)}$$

bestimmt. Einer Probe wird dann der Genotyp g^* zugewiesen, zu dessen Clusterzentrum sie den kleinsten Abstand $D_{g^*}^2(\boldsymbol{\theta})$ besitzt. Gleichzeitig fungiert dieser minimale Abstand als Qualitätsmetrik. Je kleiner er ist, umso dichter ist die Probe am entsprechenden Cluster, umso höher das Vertrauen in diese Zuweisung.

BRLMM

Im Vergleich dazu wird bei BRLMM folgendermaßen vorgegangen: Die Vorverarbeitung besteht aus der RMA ohne Hintergrundkorrektur. In den Affymetrix Power Tools, dem von Affymetrix zur Verfügung gestellten Programm, ist jedoch inzwischen die Zusammenfassung der Werte mittels PLIER voreingestellt.

Anschließend erfolgt die *Cluster Center Stretch* (CCS) - Transformation. Diese ergibt sich als

$$\frac{\operatorname{asinh}(K((I'_A - I'_B)) / (I'_A + I'_B))}{\operatorname{asinh}(K)}. \quad (3.1)$$

(I'_A, I'_B) sind dabei die normalisierten Intensitäten und K eine Konstante, deren Erhöhung dafür sorgt, dass Kontraste nahe 0 künstlich vom Nullpunkt entfernt werden, wodurch dieser Cluster der Heterozygoten breiter wird. Der Effekt ist in Abbildung 3.4 dargestellt. Mit wachsendem K verbreitert sich der mittlere Cluster. Durch diesen breiteren Cluster erhöht sich die Wahrscheinlichkeit, dass einer neuen Probe der heterozygote Genotyp zugewiesen wird. Für diese CCS-transformierten Intensitäten werden wieder Entscheidungsregionen berechnet und neue Proben dem Genotypen zugeordnet, zu dem der Mahalanobis-Abstand am kleinsten ist. Aus dieser Transformation resultiert auch eine neue Qualitätsmetrik. Dazu sei d_1 der Abstand der Probe zu dem nächsten Cluster und d_2 der zweitkleinste Abstand. Die Qualitätsmetrik c ist dann d_1/d_2 und ist auf Werte zwischen

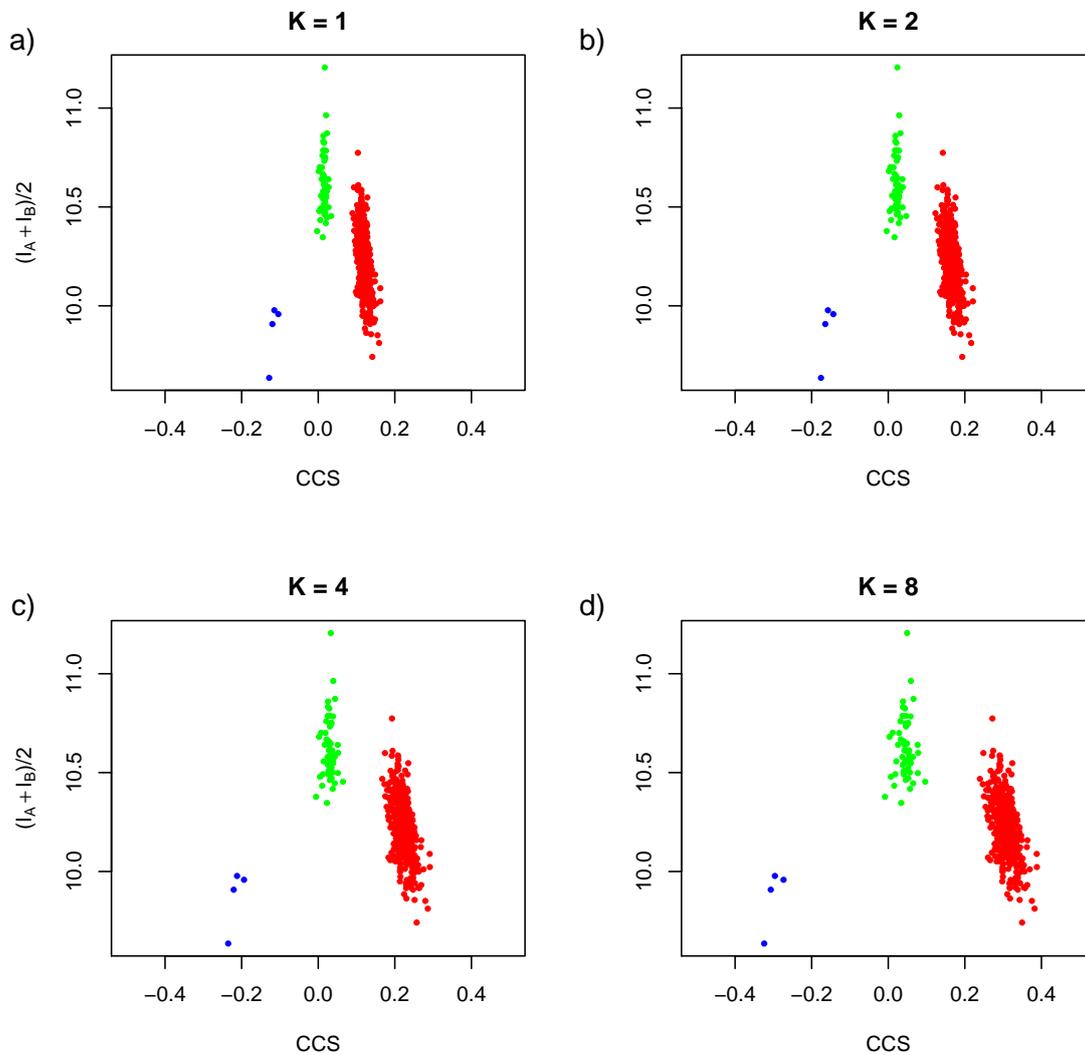


Abbildung 3.4: Effekt der CCS-Transformation. Für verschiedene Werte des Parameter K sind CCS und die durchschnittliche Intensität gegeneinander aufgetragen. Mit steigendem K erhöht sich der Abstand der Punkte auf der x -Achse zum Nullpunkt. Dadurch erhöht sich die Variabilität innerhalb der Cluster, insbesondere für den Cluster der Heterozygoten. Dies soll für Varianzgleichheit zwischen dem Cluster der Heterozygoten und denen der Homozygoten sorgen. In der Software-Implementierung ist $K = 4$ voreingestellt.

0 und 1 beschränkt. Werte nahe 0 sprechen für eine hohe Wahrscheinlichkeit. Ein Wert von 1 bedeutet, dass die zwei nächsten Cluster gleich weit entfernt liegen. Der Standardwert für BRLMM in den Affymetrix Power Tools ist $c = 0.5$, was bedeutet, dass der Abstand zum nächsten Cluster maximal halb so groß wie der Abstand zum übernächsten Cluster sein darf.

Der andere wesentliche Unterschied zwischen RLMM und BRLMM ist die Art und Weise, wie die Clusterzentren und Varianzen bestimmt werden. BRLMM nutzt dafür einen Bayes'schen Ansatz. Als A-priori-Wahrscheinlichkeiten dienen die Mittelwerte und Varianzen über *alle* SNPs, die mit dem DM-Algorithmus sicher bestimmt werden konnten. Für jeden SNP werden die A-priori-Wahrscheinlichkeiten mit den DM-Schätzungen aller Arrays für diesen SNP aktualisiert und daraus die Entscheidungsregionen konstruiert. Die Zuordnung neuer Genotypen erfolgt dann analog zu RLMM.

3.3.3 CRLMM

Ein weiterer Algorithmus, der auf RLMM aufbaut, ist CRLMM von Carvalho et al. [20]. Dabei steht das **C** für *corrected*. Der Ablauf lässt sich folgendermaßen skizzieren:

1. Die Effekte der Sondensequenz und der Fragmentlängen auf die Intensitäten werden geschätzt und die logarithmierten Intensitäten werden für diese Effekte adjustiert.
2. Die Quantilnormalisierung wird angewendet, um die Variation zwischen den Arrays zu verringern.
3. Die PM-Intensitäten eines SNPs werden allel- und strangspezifisch zusammengefasst.
4. Für die Differenzen der Intensitäten wird eine Mischverteilung modelliert und die Sequenz- und Fragmentlängen-Effekte nochmals berücksichtigt.

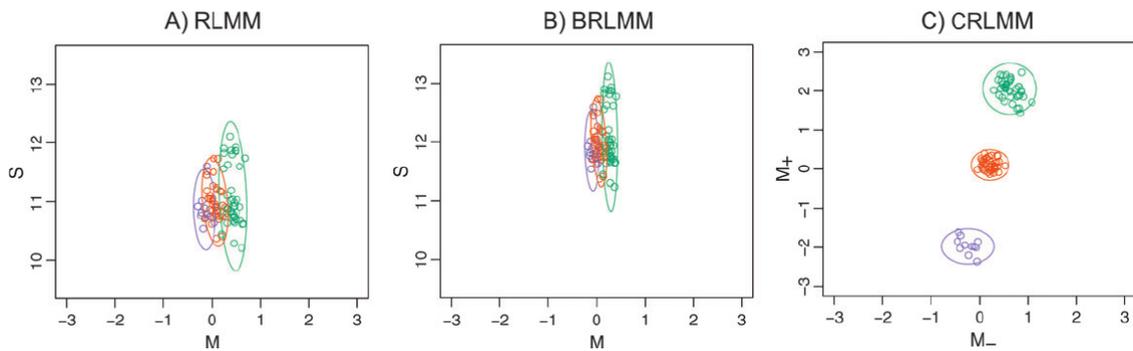


Abbildung 3.5: Effekt der Unterscheidung der Intensitäten nach den Strängen. Die Unterscheidung der Intensitäten nach den Strängen wird hiermit motiviert. M ist die Differenz der logarithmierten Intensitäten $M = \log_2(I_A) - \log_2(I_B)$, M_- und M_+ nach Strängen unterschieden und S ist die mittlere logarithmierte Intensität. Für RLMM (A) und BRLMM (B)) entspricht das in etwa der Kontrastdarstellung aus Abschnitt 3.3.2. Für CRLMM werden die Differenzen der logarithmierten Intensitäten gegeneinander aufgetragen. Es ist aus C) zu erkennen, dass M_- kaum zur Unterscheidung der Genotypcluster beiträgt, während mit M_+ die drei Cluster klar voneinander trennbar sind. Durch das Zusammenfassen der Intensitäten in M geht diese Unterscheidbarkeit verloren, wodurch die Cluster in A) und B) nicht so gut unterschieden werden können. Nachgedruckt mit Erlaubnis von Oxford University Press: *Biostatistics* [20]

5. Entscheidungsregionen werden mit Hilfe der HapMap-Daten gebildet. Dazu wird ein gemischtes Modell aufgestellt, um die Bayes-Schätzer zu erhalten.
6. Für eine neue Probe wird für jedes der Genotyp-Modelle aus dem vorherigen Schritt die Wahrscheinlichkeit berechnet. Der Probe wird der Genotyp zugewiesen, dessen Wahrscheinlichkeit sie maximiert.

Die maßgeblichen Unterschiede zu den bisher beschriebenen Algorithmen sind

1. Die Berücksichtigung der Sequenz- und Fragmentlängen als Kovariablen und
2. Die Trennung der Intensitäten nach den Strängen. In Abbildung 3.5 ist ein motivierendes Beispiel für diese Trennung gezeigt. Während in der klassischen Kontrastdarstellung für RLMM und BRLMM die Cluster nicht klar voneinander trennbar sind, ermöglicht die Darstellung der Intensitätsdifferenzen getrennt nach Strang eine klare Trennung der Genotypcluster.

3.3.4 CHIAMO

CHIAMO (ital. ich rufe) wurde vom Wellcome Trust Case Control Consortium im Rahmen der GWA-Studien zu sieben häufigen Erkrankungen entwickelt [72].

Vorverarbeitung der Daten

Die Datenvorverarbeitung unterscheidet sich teilweise von RMA und PLIER und wird deswegen kurz skizziert. Für den s -ten SNP der i -ten Person wird das k -te Quartett der Intensitäten wie folgt definiert:

$$I_{isk} = (I_{isk}^{PA}, I_{isk}^{PB}, I_{isk}^{MA}, I_{isk}^{MB}).$$

Durch PA sind beispielsweise die Perfect-Match-Intensitäten für das A-Allel gekennzeichnet, MB markiert entsprechend die Mismatch- Intensitäten des B-Allels. Die Anzahl K der Quartette variiert zwischen den Chips, für das Affymetrix 500k Array Set ist $K \in \{6, 10\}$. Durch Quantilnormalisierung erhält man zunächst I' . Anschließend erfolgt eine log-Transformation, wodurch man folgende log-normalisierte Intensitäten erhält

$$Y_{isk} = (Y_{isk}^{PA}, Y_{isk}^{PB}, Y_{isk}^{MA}, Y_{isk}^{MB}); \quad Y = \log(I'). \quad (3.2)$$

Anschließend wird eine Hintergrundkorrektur durchgeführt:

$$Y_{isk}^A = \begin{cases} Y_{isk}^{PA} - \frac{1}{2}(Y_{isk}^{MA} + Y_{isk}^{MB}) & \text{falls } Y_{iks}^{PA} \geq Y_{iks}^{MA} \\ 0 & \text{sonst} \end{cases}$$

Die Korrektur für die Intensitäten des B-Allels erfolgt analog. Die einzelnen Intensitäten pro Allel eines SNPs werden dann durch Mittelwertbildung zusammengefasst. Dadurch erhält man pro SNP ein Intensitätspaar (X_{is}^A, X_{is}^B) , wobei X den Mittelwert über entsprechende Y bezeichnet.

Gegenüber der *Median Polish*-Methode, die bei RMA verwendet wird, resultiert diese Transformation in kompakteren Clustern mit weniger Ausreißern [72].

Calling

Das Calling wird über ein komplexes Bayes'sches Hierarchisches Gemischtes Modell modelliert. Für Details wird auf den Anhang von Wellcome Trust Case Control Consortium [72] verwiesen. Hervorhebenswert erscheinen folgende Punkte:

- Der Algorithmus kann mit gruppierten Daten umgehen, das heißt die Individuen können aus verschiedenen Studien bzw. Regionen stammen.
- Es wird das Vorliegen des Hardy-Weinberg-Gleichgewichtes angenommen, wobei diese Annahme durch geeignete Auswahl eines Wertes für einen Parameter abgeschwächt werden kann.
- Zur Klassifikation werden die Allelhäufigkeiten pro SNP geschätzt. Als zusätzliche A-priori-Wahrscheinlichkeiten können die Allelhäufigkeiten der CEU-HapMap-Population verwendet werden.
- Pro SNP erhält man ein Tripel von Posterior-Wahrscheinlichkeiten $p = \{p_{AA}, p_{AB}, p_{BB}\}$, das die Wahrscheinlichkeiten für die verschiedenen Genotypen des SNPs für diese Probe angibt.

3.3.5 JAPL

JAPL (nach J'appelle, frz. ich rufe) wurde von Plagnol et al. [54] mit der Maßgabe entwickelt, systematische Fehler in der Genotypbestimmungen zweier Gruppen, wie Fällen und Kontrollen zu verringern. Dies geschieht, um den Anstieg des Anteils an falsch-positiven Assoziationen in einer GWA-Studie zu verhindern [23]. Zu diesem Zweck wurde der von Moorhead et al. [50] verwendete Algorithmus weiterentwickelt.

In der praktischen Anwendung ist die Vorverarbeitung der Daten bei CHIAMO und JAPL identisch, die Daten werden bei beiden Verfahren mittels einer eigenen Implementierung der Quantilnormalisierung transformiert [9] und anschließend, wie in Abschnitt 3.3.4 zusammengefasst. Der Algorithmus verwendet zur Bestim-

mung der Genotypen lediglich die Kontraste der Intensitäten, welche hier als

$$contr = \frac{\sinh(2 \cdot I_A \cdot I_B)}{\sinh(2)}$$

berechnet werden. Die Genotypen werden dann mittels des EM-Algorithmus [11] bestimmt. Zum Ansatz von Moorhead et al. [50] unterscheidet sich der Algorithmus in folgenden Punkten:

- Die Annahme, dass die Verteilung der SNPs dem Hardy-Weinberg-Gleichgewicht folgt kann genutzt werden, um die A-priori-Verteilung der Genotyphäufigkeiten zu schätzen.
- Die zu Grunde liegende Mischverteilung basiert auf t -Verteilungen anstelle von Normalverteilungen, um eine höhere Robustheit gegenüber Ausreißern zu erzielen.
- Die Verteilung der Allelhäufigkeiten zwischen Fällen und Kontrollen wird unter der Nullhypothese keiner Assoziation als gleich angenommen.
- Die Lage der Genotypcluster kann jedoch für Fälle und Kontrollen unterschiedlich sein.
- Analog zu CHIAMO ist das Ergebnis für eine Probe und einen SNP ein Tripel von Wahrscheinlichkeiten.

3.3.6 Birdseed

Obwohl eigentlich für den Genome-Wide Human SNP Array 6.0 entwickelt, ist Birdseed [41] auch auf Daten des Human Mapping 500K Array Set anwendbar. Der Algorithmus wurde am Broad-Institut aus Boston, Massachusetts in enger Zusammenarbeit mit Affymetrix entwickelt.

Nach der Quantilnormalisierung werden die Intensitäten entweder mittels *Median Polish* oder PLIER zusammengefasst. Analog zu CRLMM, werden mittels der HapMap-Daten die SNP-spezifischen Lage- und Streumaße sowie Genotyp-

häufigkeiten geschätzt. Diese Informationen werden im *model file* gespeichert und sind integraler Bestandteil des Callings. Sie dienen als Startwerte für den Algorithmus. Dieser nutzt ein bivariates Mischverteilungsmodell. Der Ablauf lässt sich wie folgt skizzieren:

1. Initialisierung: Startwerte für die Lage und Größe der Cluster werden aus dem *model file* importiert und auf die Intensitäten des aktuellen Datensatzes skaliert.
2. Mittels des EM-Algorithmus werden die Parameter für die Cluster mit den Daten aus der Stichprobe aktualisiert. Dabei werden vier Modelle untersucht, die die Anzahl der Genotypen festlegen. Anhand von Kriterien, wie dem *Bayesian Information Criterion* (BIC), der Entfernung der geschätzten Clusterzentren zu den a priori festgelegten Schwerpunkten und des Abstandes der geschätzten Clusterzentren zueinander, wird entschieden, welche Genotypgruppen dieser SNP in dieser Stichprobe besitzt. Erst dann beginnt die eigentliche Genotypzuweisung. Dabei wird einer Probe der Genotyp zugewiesen zu dessen Cluster sie den minimalen Abstand besitzt.

Die maßgeblichen Unterschiede von Birdseed zu den anderen Algorithmen sind folgende:

1. Die Anzahl der Genotypgruppen wird vor der eigentlichen Bestimmung der Genotypen festgelegt.
2. in den EM-Algorithmus ist ein Kriterium implementiert, welches verhindert, dass sich während der Maximierung die Zentren der Cluster zu sehr annähern.

3.4 Systematischer Vergleich der ausgewählten Algorithmen

Nach der theoretischen Beschreibung der Algorithmen, sollen sie nun praktisch miteinander verglichen werden. Ein großes Problem dabei ist die Auswahl eines geeigneten Datensatzes. Um die Ergebnisse der Genotypbestimmung bewerten zu können, muss man die wahren Genotypen kennen. Daher wurden in der Vergangenheit häufig die HapMap-Daten verwendet [41, 45]. Die Struktur dieses Datensatzes wird in Abschnitt 3.4.1 beschrieben. Anschließend wird in Abschnitt 3.4.2 ein Überblick über die technische Umsetzung der Genotypbestimmung gegeben, bevor in Abschnitt 3.4.3 die Ergebnisse des Vergleichs gezeigt werden.

3.4.1 Vergleichsgrundlage: HapMap II

Im Rahmen des HapMap-Projektes [67, 68] wurden Personen verschiedenster Ethnizitäten typisiert. Damit sollte eine genaue Bestimmung der Muster von Variationen im menschlichen Genom [67] ermöglicht werden. In Tabelle 3.2 sind die Herkunftsregionen und Anzahl der untersuchten DNA-Proben aufgelistet. Neben einer afrikanischen Subpopulation wurden noch zwei asiatische und eine Population europäischer Herkunft untersucht. In der initialen Studie wurde die Bestimmung der Genotypen von ca. 600.000 SNPs angestrebt. Durch eine Weiterführung des Projekts wurde diese Anzahl im Jahr 2007 noch um 3,1 Millionen SNPs erhöht. Diese Daten wurden auf verschiedenen Plattformen in verschiedene Zentren generiert. Um die Qualität der Daten sicher zu stellen, wurde eine ausführliche Qualitätskontrolle durchgeführt. Dazu gehörten folgende Überprüfungen der Konkordanz (Übereinstimmung) der Genotypen:

- Alle Zentren haben zu Beginn des Projekts die selben 1500 SNPs typisiert. Die durchschnittliche Übereinstimmung mit zwei anderen konkordanten Zentren lag bei 99,5%.

Tabelle 3.2: Übersicht über die Populationen in Hapmap I + II

Kürzel	Ursprung	Anzahl
YRI	Yoruba in Ibadan, Nigeria	90 (30 Trios)
JPT	Tokyo, Japan	45
CHB	Han Chinesen aus Peking, China	45
CEU	Einwohner Utahs mit nord- oder westeuropäischem Ursprung	90 (30 Trios)

Die Probanden aus JPT und CHB sind jeweils nicht verwandte Individuen.

Ein Trio, auch Kernfamilie genannt, besteht aus Mutter, Vater und einem Kind.

- Immer 96 Proben wurden zusammen analysiert (eine Mikrotiterplatte), davon waren fünf Duplikate und eine Leerprobe.
- Die Genotypisierung von Trios erlaubt die Überprüfung auf Mendel-Fehler, also Genotypen die nicht den Gesetzen der Vererbungslehre folgen.
- Eine Zufallstichprobe von Genotypen, die ein Zentrum generiert hat, wurde in einem anderen Zentrum wiederholt.

Die Genotypdaten sind frei verfügbar und wurden aus Gründen des praktikableren Formats von der PLINK-Homepage [4] heruntergeladen. Um einen im Hinblick auf die Allelfrequenzen homogenen Datensatz zu erhalten, wurde sich in dieser Arbeit auf die 90 Proben europäischer Herkunft (CEU-Proben) beschränkt.

3.4.2 Genotype Calling

Für die HapMap-Daten stehen neben den Genotypen auch die CEL-Dateien für das Affymetrix Human Mapping 500K Array Set zur Verfügung [7]. Diese wurden genutzt, um mit den in Abschnitt 3.3 beschriebenen Algorithmen die Genotypen zu bestimmen.

Der Ablauf der Datengenerierung wird anhand von Flussdiagrammen verdeutlicht. Generell sind neben den Intensitätsinformationen aus den CEL-Dateien

noch weitere Informationen zur Generierung der Genotypen nötig:

- *CDF (chip definition file)*: Datei, die der Intensitätsmatrix in den CEL-Dateien die entsprechenden SNPs und Sonden zuordnet.
- *annotation file*: Datei, die zu allen SNPs eines Arrays die Informationen über genomische Position, rs-Nummer, Strang, Allele und Allelfrequenzen innerhalb der einzelnen HapMap-Populationen bereit stellt.

Die CDF-Dateien sind zwingend notwendig, um aus den CEL-Dateien die Intensitäten zu extrahieren. Mit Hilfe der Annotationsinformationen erfolgen folgende Schritte:

- Zuordnung der rs-Nummern zu den Affymetrix-SNP-Nummern
- Konvertierung $\{0, 1, 2\}$ -kodierter Genotypen in basenkodierte Genotypen, z.B. $0 \rightarrow A/A$, $1 \rightarrow A/C$ und $2 \rightarrow C/C$
- Orientierung der Genotypen auf den 5'-Strang
- Zuweisung eines SNPs zu einem Chromosomen und einer Basenpaarposition

In Abbildung 3.6 ist der Arbeitsablauf exemplarisch für CRLMM dargestellt. Die Implementierung dieses Algorithmus zeichnet sich durch eine hohe Anwenderfreundlichkeit aus. So werden angegebene Dateinamen und Pfade auf Existenz überprüft und die CDF-Dateien werden automatisch dem Array-Typ zugeordnet, der in den Daten erkannt wurde. Die Aufteilung von Genotypen und Qualitätsmaßen in zwei Dateien wird auch in der Implementierung von BRLMM und Birdseed verwendet. Dadurch konnte das eingesetzte Perl-Skript `cr1mm2tped.pl` nach leichten Modifikationen auch für diese Algorithmen eingesetzt werden. In Abbildung 3.7 ist die Zusammenführung der numerischen Genotypen und der Gütemaße anhand der CRLMM-Struktur illustriert.

Der Aufwand für das Calling mit JAPL ist ungleich höher (Abb. 3.8). So wird die Normalisierung der Daten dem Nutzer selbst überlassen. Gerade die Aktu-

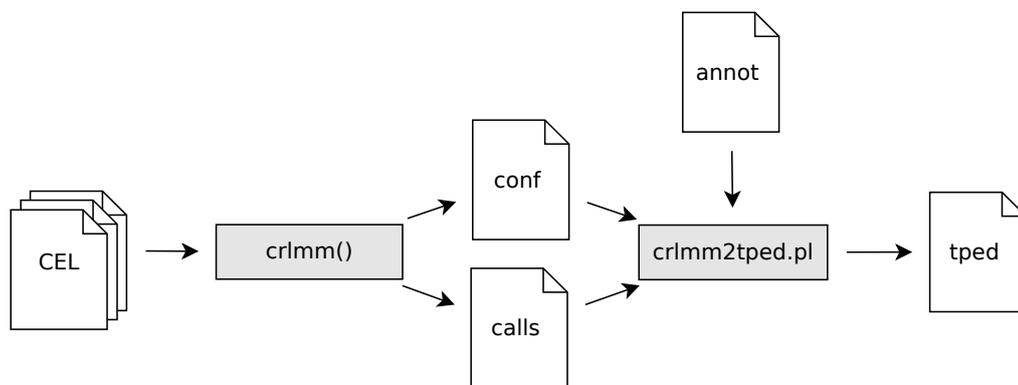


Abbildung 3.6: Flussdiagramm zum Calling mit CRLMM. Die Schritte bei der Bestimmung der Genotypen aus den CEL-Dateien sind skizziert. Dateien werden durch die abgeknickte Ecke symbolisiert, Skripte sind grau hinterlegt. Die Funktion `crlmm()` aus dem Bioconductor-Paket `oligo` lädt die benötigten CDF-Dateien selbständig nach. Die relevanten Ergebnisdateien `calls` und `conf` enthalten die numerisch kodierten Genotypen und die entsprechenden Qualitätswerte. Mittels `crlmm2tped.pl`, einem eigens geschriebenen Perl-Skript, werden die Informationen aus diesen beiden Dateien kombiniert und zusammen mit den Annotationsinformationen aus `annot` in ein `tped` konvertiert. Details zu diesem Schritt werden in Abbildung 3.7 gezeigt. Die weitere Verarbeitung/Analyse kann dann mit PLINK erfolgen.

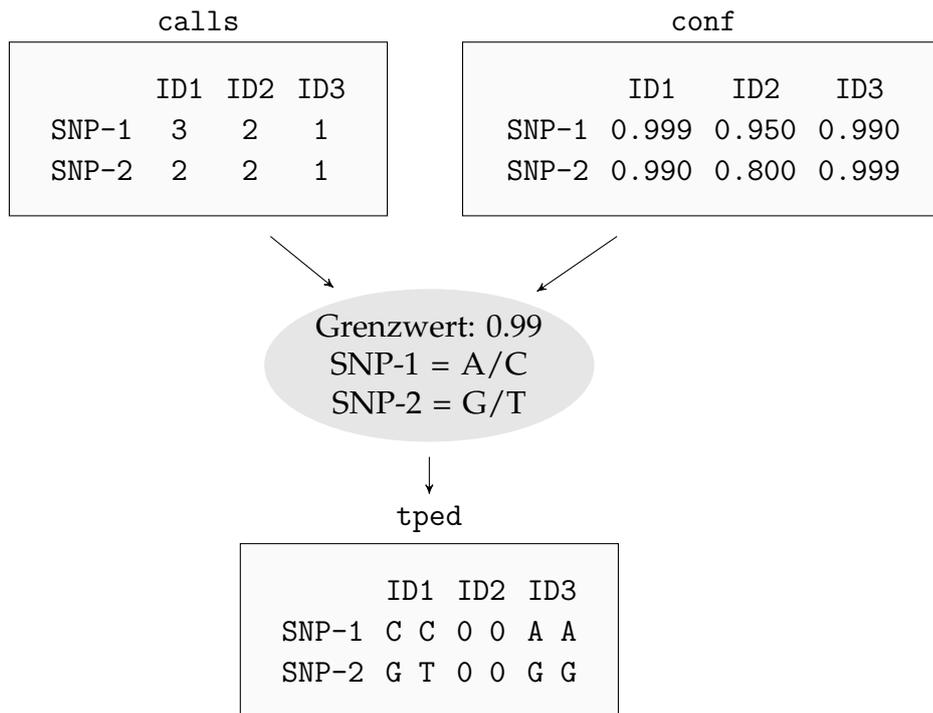


Abbildung 3.7: Illustration der Umkodierung der Calling-Ergebnisse für CRLMM. In diesem fiktiven Beispiel liegen Informationen für zwei SNPs und drei Personen vor. In der `calls`-Datei sind die numerischen Genotypen abgespeichert, die `conf`-Datei enthält die entsprechenden Qualitätswerte. Aus der Annotationsdatei erhält man die Information, dass SNP-1 ein A/C SNP und SNP-2 ein G/T SNP ist. In dem Programm wird nun für jeden Genotyp der entsprechende Qualitätswert extrahiert und mit einem vorher festgelegten Grenzwert (hier 0.99) verglichen. Überschreitet der Qualitätswert diesen Grenzwert, werden die Genotypen anhand der Annotationsdaten gebildet, ansonsten wird dieser Wert als fehlender Wert kodiert (hier 0 0).

alisierung der Annotationsinformationen gestaltete sich schwierig, da innerhalb der verwendeten Hilfsdateien verschiedene Symbole zur Trennung der Spalten genutzt wurden. Diese Anforderung sind nicht dokumentiert. Weiterhin werden keine Plausibilitätskontrollen hinsichtlich nutzerdefiniert Dateinamen durchgeführt. Darf das Programm aufgrund fehlender Schreibrechte die Ausgabedatei nicht anlegen, wird trotzdem die Berechnung bis zum Ende durchgeführt. JAPL legt pro Chromosom ein Verzeichnis an und schreibt pro SNP eine GEN-Datei in das entsprechende Verzeichnis. Diese Art der Ausgabe ist sehr ungewöhnlich und die Konvertierung in ein gängiges Format erforderte erheblichen Aufwand. Auf die Angabe der Flussdiagramme für BRLMM, Birdseed und CHIAMO wird verzichtet, da diese dem Ablauf CRLMMs, bzw. bei CHIAMO dem von JAPL, sehr stark ähneln.

3.4.3 Ergebnisse

Als Maß für die Güte eines Calling Algorithmus soll die Konkordanz (Übereinstimmung) der Genotypen mit den von HapMap bestimmten Genotypen verwendet werden. Diese Konkordanz wird nur zwischen Genotypen berechnet, die in beiden Datensätzen vorliegen, das heißt ein fehlender Wert wird nicht als falscher Genotyp gewertet. Allerdings soll dieses Fehlen berücksichtigt werden. Daher haben Lin et al. [45] in ihrem Vergleich von CRLMM mit BRLMM *accuracy vs. drop rate plots* (ADP) eingeführt. Dabei wird für verschiedene Grenzwerte der Qualitätsmaße die Konkordanz und der Anteil fehlender Genotypen bestimmt. In Abbildung 3.9 ist ein ADP für alle 482.203 SNPs des 500K Array Sets dargestellt, die auch in den HapMap-Genotypen vorlagen. Die Algorithmen BRLMM, Birdseed und CRLMM zeigen vergleichbare Kurvenverläufe mit einer maximalen Konkordanz knapp unter 99,5% bei BRLMM. CHIAMO erreicht eine maximale Konkordanz von 99,2% und JAPL 98,7%. Da SNPs mit einer kleinen MAF eine besondere Herausforderung darstellen, da ein Cluster nur spärlich bzw. gar nicht besetzt ist, wurden ADPs für seltene und häufige SNPs getrennt berechnet. Als Grenzwert wurde eine MAF von zehn Prozent festgelegt, das heißt bei

90

Personen erwartet man für diese Grenze ungefähr eine Person im Cluster der Homozygoten des selteneren Allels. In Abbildung 3.10 sind die entsprechen-

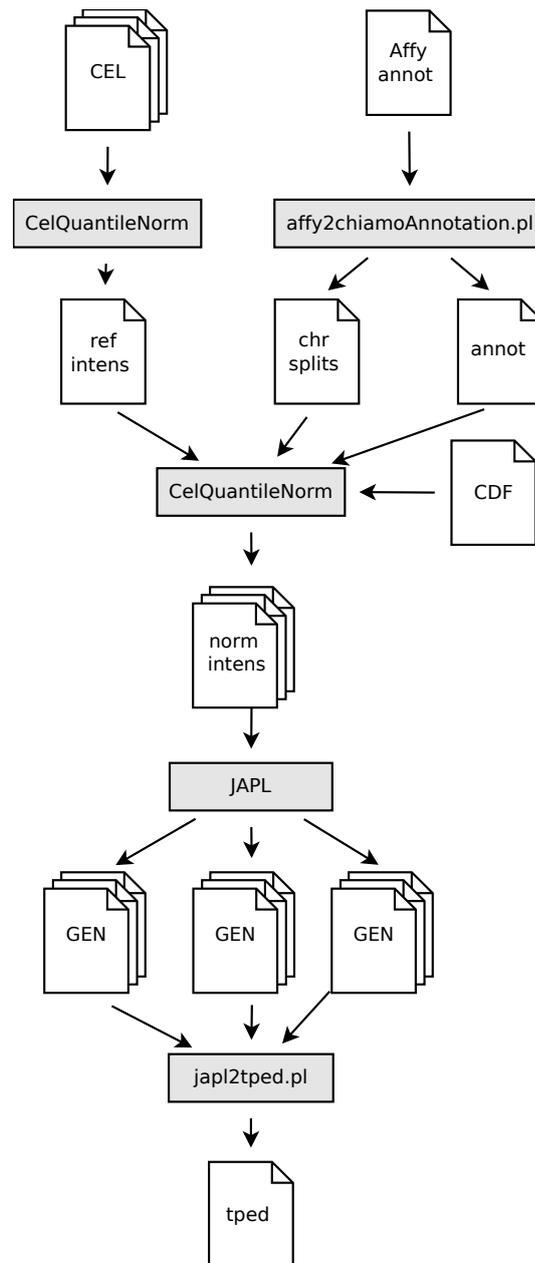


Abbildung 3.8: Flussdiagramm zum Calling mit JAPL. Die Schritte bei der Bestimmung der Genotypen aus den CEL-Dateien sind skizziert. Dateien werden durch die abgeknickte Ecke symbolisiert, Skripte sind grau hinterlegt. Die Pfade bis zu den normalisierten Intensitäten (*norm intens*) symbolisieren die Datenvorverarbeitung. Dies geschah mit dem Programm *CelQuantileNorm* [9], jedoch erforderte die Integration aktueller Annotationsinformationen erheblichen Arbeitsaufwand. Diese Schritte sind für JAPL und CHIAMO identisch. Anschließend erfolgt das eigentliche Calling mit JAPL. Für jeden SNP wird eine GEN-Datei in einem chromosomen-spezifischen Verzeichnis angelegt. Die Umwandlung der Tripel von A-posteriori-Wahrscheinlichkeiten in Genotypen und Zusammenfassung in einer *tped*-Datei erfolgt mittels eines eigens geschriebenen Perl-Skripts.

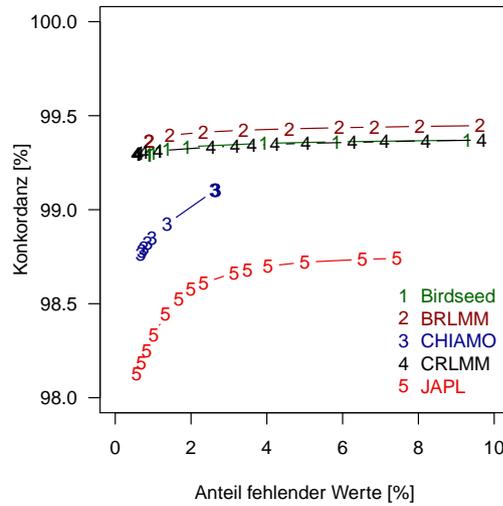


Abbildung 3.9: Accuracy vs. Drop-rate Plot (ADP). für alle SNPs. Die Grenzwerte wurden so gewählt, dass der Anteil fehlender Genotypen zwischen 0 und 10% lag. Für CHIAMO war es jedoch nicht möglich, den Anteil auf mehr als 3% zu erhöhen. Während die Konkordanz von BRLMM, Birdseed und CRLMM nahezu vergleichbar ist und auch kaum durch den Anteil fehlender Werte verändert wird, ist die Konkordanz für JAPL und CHIAMO deutlich niedriger.

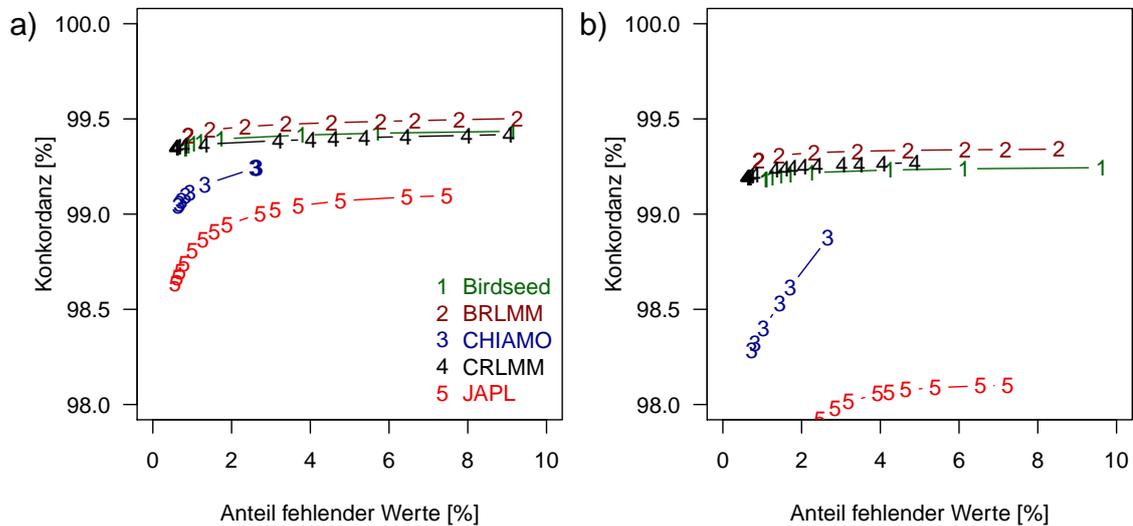


Abbildung 3.10: ADP getrennt nach MAF. In a) ist der ADP für SNPs mit einer MAF größer als zehn Prozent dargestellt (312.883 SNPs). b) zeigt den ADP für die SNPs mit niedriger MAF (160.320). Bei Birdseed, BRLMM und CRLMM beträgt der Unterschied in der Konkordanz zwischen häufigen und seltenen SNPs ungefähr 0,2%, während es für CHIAMO ca. 0.5% sind. Für JAPL ist die Konkordanz bei seltenen SNPs 1% niedriger als für häufige SNPs.

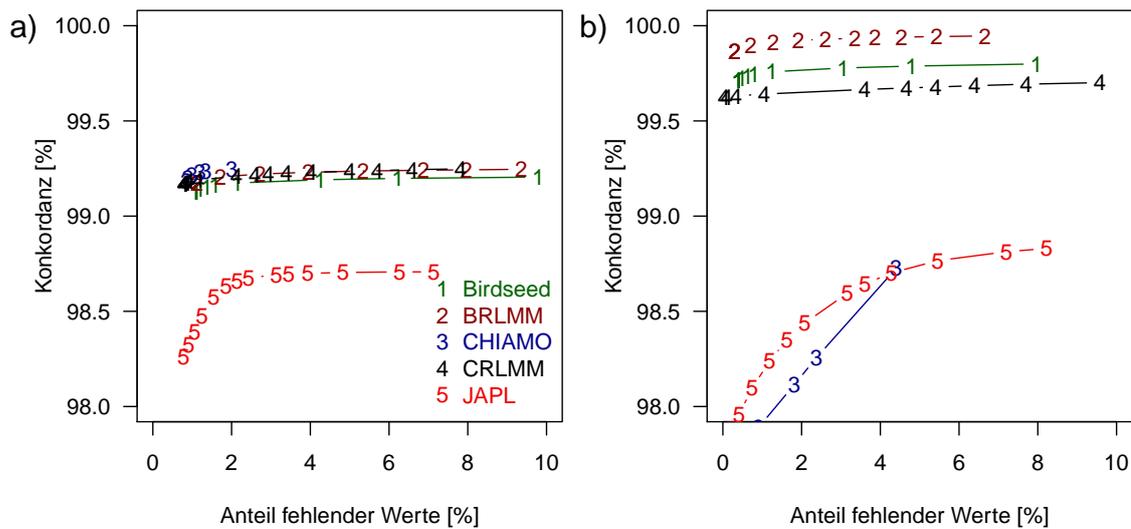


Abbildung 3.11: ADP getrennt nach homozygoten und heterozygoten Genotypen. In a) ist der ADP der homozygoten Genotypen dargestellt, b) zeigt den der Heterozygoten. Bis auf JAPL zeigen alle Algorithmen eine ähnliche Konkordanz (ca. 99,3%) bei der Bestimmung homozygoter Genotypen. Für heterozygote Genotypen ist die Konkordanz von BRLMM nahezu perfekt, während hier CHIAMO mit maximal 98,7% die niedrigste Übereinstimmung zeigt.

den ADPs abgebildet. Während die Ergebnisse für SNPs mit einer höheren MAF weitestgehend mit den Ergebnissen für alle SNPs übereinstimmen, ist der Unterschied zwischen den Konkordanzen für die häufigen SNPs und den Konkordanzen für die seltenen SNPs beträchtlich. Insbesondere die Konkordanzen für CHIAMO und JAPL verringern sich deutlich.

Da bekannt ist, dass sich Missklassifizierungshäufigkeiten zwischen heterozygoten und homozygoten Genotypen bei einigen Algorithmen unterscheiden [57], wurde dies ebenfalls betrachtet. Dabei zeigt sich (Abb. 3.11), dass bis auf CHIAMO alle Algorithmen für die homozygoten Genotypen (links) eine niedrigere Konkordanz haben als für heterozygote Genotypen. Während die Konkordanz für homozygote Genotypen bei CHIAMO mit am höchsten ist, zeigt es die niedrigste Übereinstimmung für heterozygote Genotypen.

3.5 Diskussion

Mit der Literaturrecherche konnten insgesamt neun Algorithmen identifiziert werden, mit denen das Calling von Affymetrix Chips möglich ist. Allerdings ist es dem Autor nur bei fünf der neun Verfahren möglich gewesen, ein funktionsfähiges Computerprogramm zu erhalten. Dieses mag weitestgehend der geringen Nachfrage nach alternativen Algorithmen geschuldet sein. Mit BRLMM und Birdseed bietet Affymetrix selbst zwei Algorithmen an, die in einer Software mit grafisch ansprechender Benutzeroberfläche verfügbar sind. Daher ist es nachvollziehbar, dass für viele Anwender die Motivation, nach Alternativen zu suchen, gering ist; insbesondere dann, wenn die Anwender nicht im größeren Stil viele GWA-Studien analysieren. Allerdings ist die Verfügbarkeit von Software in einer benutzerfreundlichen Umgebung kein Indikator für die inhaltliche Qualität der Software. Aus diesem Grund wurden hier die existierenden Algorithmen miteinander verglichen. Dabei fiel zunächst auf, dass bereits die Vorverarbeitung der Daten sehr divers gestaltet ist. So werden in einigen Algorithmen die MM-Informationen gar nicht benutzt während sie in anderen integraler Bestandteil des Verfahrens sind. Insbesondere im Hinblick auf die Weiterentwicklung der Arrays hat sich letzteres als Nachteil erwiesen. Während Algorithmen wie CRLMM problemlos auch auf Daten des 5.0 Array oder des 6.0 Array, die beide keine MM-Sonden enthalten, angewendet werden können, ist die direkte Weiterverwendung von CHIAMO nicht möglich.

Die empirische Evaluation der Daten ist sehr stark von der Güte des Goldstandards abhängig. Zwei Punkte sind daher bei der Verwendung der HapMap-Daten zu beachten:

- Die Daten wurden experimentell gewonnen, sind also ebenfalls fehlerbehaftet.
- BRLMM, Birdseed, CRLMM und CHIAMO nutzen die HapMap-Daten zum Training ihrer Modelle. Aus dem Bereich des maschinellen Lernens ist jedoch bekannt, dass das Training und die Evaluation auf dem selben Datensatz zu einer Unterschätzung des Fehlers führt [19].

Wenn man die HapMap-Daten jedoch als Goldstandard verwendet, lassen sich folgende Schlüsse daraus ableiten: Die drei Algorithmen Birdseed, BRLMM und CRLMM liefern vergleichbare Ergebnisse, mit leichten Vorteilen für BRLMM. CRLMM ist jedoch deutlich komfortabler in der Anwendung. Insbesondere die einfache Installation als R-Paket und die umfassenden Plausibilitätsprüfungen der Nutzereingaben machen es für Einsteiger empfehlenswert. Unabhängig vom Algorithmus muss allerdings festgestellt werden, dass insbesondere für SNPs mit einer niedrigen MAF der Anteil falsch bestimmter Genotypen sehr hoch sein kann. Gerade für kleine Studien ist dieser kritische Bereich schon viel eher erreicht als bei großen Studien, da ein Cluster zur Genotypisierung nur als solches erkannt werden kann, wenn dieses genügend Beobachtungen aufweist. Um Fehlklassifikationen von SNPs nach dem Calling automatisch identifizieren zu können, wurde im Rahmen der vorliegenden Arbeit das R-Paket *acpa* entwickelt. Dieses wird im nächsten Kapitel diskutiert.

4 Bewertung von Cluster-Plots

Wie in Kapitel 3 ausführlich beschrieben wurde, werden im Genotype Calling die normalisierten Signalintensitäten in Genotypen konvertiert. Dieser Prozess lässt sich, für jeden analysierten Marker getrennt, anschaulich grafisch darstellen. Dazu werden die normalisierten Signalintensitäten des A-Allels gegen die normalisierten Intensitäten des B-Allels aufgetragen. Der zugewiesene Genotyp wird farblich kodiert. Abbildung 4.1 zeigt schematische Darstellungen dieser *Cluster-Plots*. Dabei wird neben der theoretischen Lage der Punkte auch die praktisch beobachtete Situation gezeigt, in der statt Punkten ellipsenförmige Cluster für jede Genotypgruppe zu erkennen sind. Weiterhin wird als alternative Darstellungsform eine Kontrastdarstellung gezeigt, die die optische und auch statistische Beurteilung der Cluster-Plots vereinfachen kann. Die visuelle Auswertung von Cluster-Plots ist nicht nur laut Affymetrix das Mittel der Wahl, um Fehler bei der Genotyp-Bestimmung aufzudecken [10] sondern wird auch in mehreren Publikationen zur Qualitätssicherung von Hochdurchsatz-Genotypisierungsdaten empfohlen [65, 71, 76]. Dabei geht die Auswertung von Cluster-Plots über den bloßen Ausschluss von SNPs mit schlechtem Clustering hinaus, wie folgendes Beispiel zeigt:

Barrett et al. [14] haben in einer GWA-Studie die genetischen Ursachen des Typ-1-Diabetes untersucht. Dazu wurde eine Fall-Kontroll-Studie durchgeführt und die Ergebnisse dieser Studie mit zwei bereits publizierten Studien per Metaanalyse kombiniert [15, 24, 52, 72]. Dabei zeigte sich eine Inflation der Teststatistik, das heißt, man erhielt mehr positive Assoziationen, als man unter der Nullhypothese erwarten würde. Zwei Ursachen dafür wurden diskutiert:

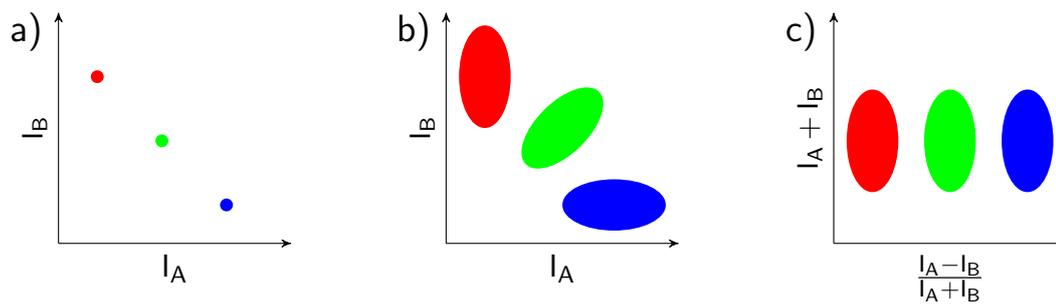


Abbildung 4.1: Schematische Darstellung von Cluster-Plots. In a) wird der Idealfall gezeigt, in dem die Punkte für die Genotypgruppen AA, AB und BB an den Koordinaten $(1,0)$, $(0.5, 0.5)$ und $(0, 1)$ liegen. Durch Unterschiede in den DNA-Mengen der einzelnen Proben und Messfehler beim Scannen entstehen aber tatsächlich die in b) gezeigten Punktwolken. c) zeigt eine alternative Darstellung der Daten aus b). Statt die Intensitäten gegeneinander aufzutragen, werden der Kontrast gegen die Summe beider Intensitäten dargestellt. Abbildung nach Abb. 4.9 aus [75].

1. Populationsstratifikation und
2. systematische Unterschiede in den Genotypisierungsfehlern zwischen Fällen und Kontrollen

Durch Adjustierung für die geographische Region, wurde versucht, einer potentielle Populationsstratifikation Rechnung zu tragen. Da der Einfluss dieser Korrektur gering war und der Effekt der Inflation in einer Analyse nicht auftrat, in der nur Fälle verwendet wurden, wurde Populationstratifikation als Ursache der Inflation verworfen und Genotypisierungsfehler als wahrscheinlichste Ursache angenommen. Warum ist dies so bedeutsam? Ein üblicher Weg, um einer solchen Inflation zu begegnen, ist die genomische Kontrolle [28]. Dabei wird die Teststatistik auf geeignete Weise korrigiert. Für den Trend-Test wird häufig der Quotient aus dem beobachteten Median der Teststatistiken und dem Median der χ^2_1 -Verteilung verwendet [13]. Anschließend werden die Teststatistiken durch diesen Inflationsfaktor λ dividiert. Durch diese Korrektur werden die p -Werte für die Assoziation der Marker mit der Krankheit größer, das heißt der pauschale Einsatz der genomischen Kontrolle führt zu konservativen Ergebnissen, so dass die Wahrscheinlichkeit, neue Assoziationen zu finden, sinkt. Daher haben Barrett et al. versucht, dieses konservative Vorgehen zu vermeiden. Unter Berufung auf die untersuchten Cluster-Plots der SNPs mit den kleinsten p -Werten wurde auf eine λ -Korrektur verzichtet.

Neben der direkten Verwendung in GWA-Studien bilden die Genotypinformationen die Grundlage für komplexere statistische Analysen, wie z.B. Haplotypanalysen, der Imputation weiterer Genotypen oder Untersuchung von Gen-Gen bzw. Gen-Umwelt-Interaktionen.

In der Haplotypanalyse wird der Zusammenhang von Blöcken von SNPs mit einem Phänotyp untersucht. Dieser Zusammenhang von SNPs, die dicht zusammen auf einem Chromosom liegen, wird auch für die Imputation weiterer SNPs verwendet. Dabei wird ausgenutzt, dass eine hohe Korrelation zwischen benachbarten SNPs besteht. Ist ein SNP eines Haplotyp nicht genotypisiert worden, weil dieser z.B. auf einem Chip nicht vorhanden ist, können die Genotypen dennoch durch die Haplotypstruktur „erschlossen“ werden, d.h. aufgefüllt bzw. imputiert werden. Die Güte der Imputation hängt dabei maßgeblich von der Qualität der zu Grunde liegenden Genotypen ab [27].

Eine Evaluation aller Cluster-Plots ist also notwendig. Die visuelle Beurteilung ist jedoch aufgrund des zu hohen personelle und zeitlichen Aufwandes nicht praktikabel. Daher wäre es wünschenswert, wenn es eine automatische Bewertung von Cluster-Plots gäbe.

Mit der systematischen Analyse von Cluster-Plots haben sich bis jetzt lediglich Lovmar et al. [48] beschäftigt, dies jedoch nicht im Kontext von Hochdurchsatz-Verfahren. In diesem Kapitel wird daher ein Ansatz zur automatischen Beurteilung von Cluster-Plots beschrieben. Dabei wird auf folgende wichtige Teilprobleme eingegangen:

1. Wie kann man Cluster-Plots von erfolgreichen Genotypbestimmungen von Cluster-Plots für fehlgeschlagene Genotypbestimmungen unterscheiden? Welche statistischen Verfahren existieren für diese Art von Problemen? Anders formuliert: Wie könnte ein neues Verfahren zur automatischen Analyse von Cluster-Plots aussehen? Diesen Fragen wird in Abschnitt 4.1 nachgegangen.
2. Wie lassen sich die Intensitätsdaten und die Genotypdaten effizient auswerten? Da für eine GWA-Studie mit ca. 3500 Personen für den Affymetrix 6.0 Array die Datei der Intensitätswerte ca. 66 GB groß ist, verbieten sich Ansätze, die alle Daten gleichzeitig einlesen würden. Ein praktikabler Ansatz wird in Abschnitt 4.2 vorgestellt und dessen Leistungsfähigkeit in Abschnitt 4.3

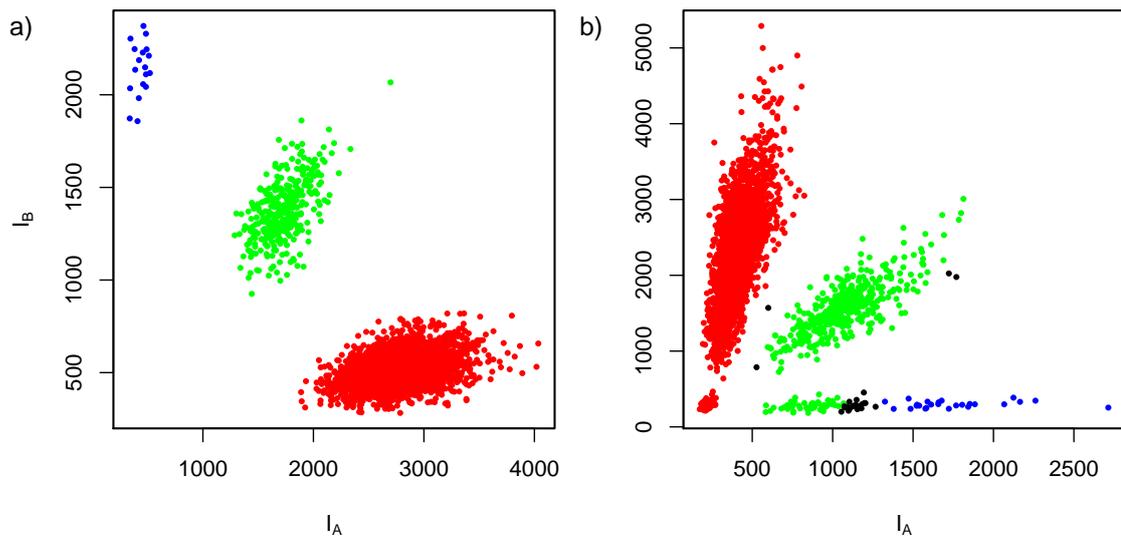


Abbildung 4.2: Beispiele für Cluster-Plots. Aufgetragen sind die normalisierten Intensitäten des A-Allels gegen die normalisierten Intensitäten des B-Allels. Der Genotyp ist farblich kodiert, rot ist der Cluster der Homozygoten des häufigeren Allels, grün der Cluster der Heterozygoten, blau der Cluster der Homozygoten des selteneren Allels und schwarz kennzeichnet Proben, für die der Algorithmus keinen Genotyp zugeordnet hat. Die Daten stammen aus der Gutenberg-Herz-Studie (siehe Abschnitt 4.3.1). In Abbildung a) ist der Cluster-Plot für den SNP rs10492932 dargestellt, Abbildung b) zeigt den Cluster-Plot für rs12407460. Während links die Genotypbestimmung erfolgreich verlief, sieht man rechts deutlich Unstimmigkeiten zwischen der Lage der Cluster und der Farbe. Dieser SNP weist eine geringe Datenqualität in der Stichprobe auf.

beurteilt.

4.1 Clustervaliditätsmaße

Wie von Ziegler und König [75] beschrieben wurde, ist die Bewertung von Cluster-Plots ein Prozess, der der Bewertung der internen Validität des Clusterings in einer Cluster-Analyse entspricht. Dieser Bereich wurde bereits detailliert untersucht [31–34] und verschiedene Kenngrößen vorgeschlagen. Ziel ist es, anhand solcher Statistiken SNPs zu detektieren, bei denen das Calling fehlgeschlagen ist. In Abbildung 4.2 wird je ein Beispiel für ein erfolgreiches und für ein fehlgeschlagenes Calling gezeigt. Die Unterschiede zwischen den beiden Cluster-Plots sind deutlich: Während in a) die Gruppen klar voneinander getrennt und farblich homogen sind, ist in b) einer Vielzahl von Proben, die laut ihren Inten-

sitätswerten homozygot für das B-Allel sein sollten, ein heterozygoter Genotyp zugewiesen worden. Wie dieser optische Eindruck auch anhand von geeigneten Kenngrößen erfassbar gemacht werden kann, ist in Handl et al. [35] und in Kim und Ramakrishna [40] ausgezeichnet zusammengefasst. Formal lässt sich die Güte des Clustering anhand folgender Kriterien bestimmen:

- *Kompaktheit* beschreibt die Homogenität von Clustern. Ein typisches Maß ist die Varianz. Je größer die Varianz, desto geringer die Kompaktheit.
- *Verbundenheit* untersucht die Nähe von Punkten einer Gruppe zueinander. Es sind zumeist Maße, die untersuchen, welcher Gruppe die nächsten Nachbarn angehören.
- *Separierbarkeit* bestimmt den Abstand zwischen zwei Gruppen. Dieser kann beispielsweise als mittlerer Inter-Cluster-Abstand definiert werden, also beispielsweise der Abstand der Zentren zueinander.
- *Kombinationen* der vorangegangenen Kriterien werden von verschiedenen Verfahren genutzt. Dabei wird sowohl die Homogenität der einzelnen Cluster, als auch der Abstand der Cluster zueinander berücksichtigt. Als Beispiele seien der Dunn Index und die Silhouette-Statistik genannt.
- *Stabilität* ist eine spezielle Form der internen Validierung. Es wird dabei der Frage nachgegangen, wie empfindlich die Genotypbestimmung auf kleine Schwankungen der Signalintensitäten reagiert. Die Konsistenz der Ergebnisse wird dann als Schätzung für die Qualität des Clustering verwendet.

In Abbildung 4.3 sind Cluster dargestellt, die nach diesen aufgezählten Gesichtspunkten gruppierbar sind. Es wird deutlich, dass mit den Kriterien Verbundenheit und Separierbarkeit, Gruppierungen in b) und c) ähnlich bewertet werden, wohingegen mit Maßen der Kompaktheit langgestreckte Cluster schlecht bewertet werden würden. Im folgenden Abschnitt werden nun einige Kennzahlen aus den jeweiligen Bereichen vorgestellt. Zur Vereinfachung wird von nur einer Gruppe von Individuen ausgegangen, z.B. einer Kohortenstudie. Im folgenden werden die Cluster mit k , $k = 1, 2, 3$ und die Stichprobengröße pro Cluster mit n_k

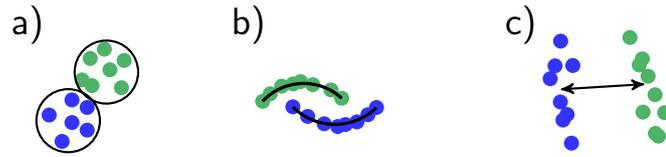


Abbildung 4.3: Schematische Abbildung der Kriterien für die interne Validität von Clusterergebnissen. Gezeigt werden Beispiele für Cluster, die kompakt sind (a), Verbundenheit zeigen (b) oder anhand ihrer Schwerpunkte separierbar sind (c). Nach Abb. 2 aus [35].

und $n = n_1 + n_2 + n_3$ bezeichnet. Als Abstandsmaß zwischen zwei Proben i und i' wird der euklidische Abstand verwendet. Damit erhält man für den euklidischen Abstand zweier Proben

$$d(i, i') = \sqrt{(c(i) - c(i'))^2 + (s(i) - s(i'))^2}. \quad (4.1)$$

Falls notwendig werden Indizes zur Clusterzugehörigkeit verwendet. Dabei kennzeichnet $c_k(i)$ den Kontrast eine Probe aus dem Cluster i und $d_{k,k'}(i, i')$ den Abstand zwischen i und i' , wobei i aus dem Cluster k und i' aus dem Cluster k' stammt. Das Zentrum des Clusters k wird wie folgt bezeichnet

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \begin{pmatrix} c_k(i) \\ s_k(i) \end{pmatrix}.$$

Dadurch lässt sich der Abstand der Probe i aus dem Cluster k und dem Clusterzentrum von k als $d(i, \boldsymbol{\mu}_k)$ darstellen.

4.1.1 Kennzahlen der Kompaktheit

Typische Kennzahlen für die Kompaktheit von Clustern sind die *Cluster-spezifische Intra-Cluster-Varianz* $\text{Var}(\text{IC}_k)$ und die *Gesamt-Intra-Cluster-Varianz* $\text{Var}(\text{IC})$. Sie basieren auf der quadratischen euklidischen Distanz d^2 und sind wie folgt definiert:

$$\sigma_k^2 = \text{Var}(\text{IC}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} d_k^2(i, \boldsymbol{\mu}_k) \quad \text{und} \quad \sigma_w^2 = \text{Var}(\text{IC}) = \frac{1}{n} \sum_{k=1}^3 \sum_{i=1}^{n_k} d_k^2(i, \boldsymbol{\mu}_k).$$

In der Praxis wird jedoch die *root mean square distance (RMSD)* verwendet, für die $RMSD = \sqrt{\text{Var}(\text{IC})}$ gilt. Kleine Werte von RMSD sollten also für ein gutes Clustering, also ein gutes Calling sprechen. Da jedoch die Variation der Gesamtintensitäten über die Genotypgruppen vergleichbar sind, entstehen typischerweise langgezogene Cluster.

4.1.2 Kennzahlen der Verbundenheit

Die Verbundenheit von Clustern wird typischerweise mit sogenannten „Nächster Nachbar“-Methoden bestimmt. Dazu wird aus den Clusterzugehörigkeiten der J nächsten Proben eine Kennzahl *Conn (Connectivity)* bestimmt. Für eine Probe i des Clusters k wird dazu mittels Gleichung (4.1) der j -te nächste Nachbar $nn_{i(j)}$ bestimmt. Hat dieser denselben Genotyp wie Probe i , gelten diese beiden als verbunden, und die Verbundenheit ist gleich 0. Falls sie zu verschiedenen Clustern gehören, ist die Nicht-Verbundenheit gleich $1/j$. Für mehrere nächste Nachbarn ergibt sich somit

$$C_{i,nn_{i(j)}} = \begin{cases} 0 & \text{falls } i \text{ und } nn_{i(j)} \text{ denselben Genotyp haben,} \\ \frac{1}{j} & \text{falls } i \text{ und } nn_{i(j)} \text{ verschiedene Genotypen haben.} \end{cases}$$

Daraus ergibt sich dann für die J nächsten Nachbarn eine Kennzahl für die Verbundenheit mit

$$Conn = \sum_{i=1}^n \sum_{j=1}^J C_{i,nn_{i(j)}}.$$

Ist *Conn* groß, liegen die Cluster zweier Genotypgruppen dicht zusammen, oder sie überlappen sich sogar.

4.1.3 Kennzahlen der Separierbarkeit

Zur Messung der Separierbarkeit von Clustern wurden verschiedene Maßzahlen vorgeschlagen. Vor dem Hintergrund des Genotyp Calling erscheint der minimale Abstand zwischen zwei Clustern als sinnvoll. Alternativ kann man den

durchschnittlichen Inter-Cluster-Abstand zwischen den Homozygoten-Clustern und dem Cluster der Heterozygoten betrachten. Der Abstand zwischen zwei Clustern ist dabei der Abstand der Clusterzentren, das heißt $d(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'})$. Ohne Beschränkung der Allgemeinheit wird nun angenommen, dass die Cluster für $k = 1, 2, 3$ von links nach rechts geordnet vorliegen.

Damit ist der minimale Inter-Cluster-Abstand

$$\min D = \min_{k \neq k'} d(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'}),$$

und der durchschnittliche Inter-Cluster-Abstand (*meanD*)

$$\text{meanD} = \frac{d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) + d(\boldsymbol{\mu}_2, \boldsymbol{\mu}_3)}{2}.$$

4.1.4 Stabilität der Cluster

Einen anderen Ansatz verfolgen Teo et al. [66]. Sie untersuchen den Einfluss von Zufallsschwankungen ausgesetzten Signalintensitäten auf die Clusterergebnisse. Genauer gesagt, werden normalverteilte Fehlerterme $\varepsilon_a, \varepsilon_B \stackrel{\text{u.i.v.}}{\sim} N(0, \sigma^2)$ zu den normalisierten Intensitäten I_A und I_B addiert. Die variierten Intensitäten $\tilde{I}_A = I_A + \varepsilon_A$ und $\tilde{I}_B = I_B + \varepsilon_B$ werden dann zur Genotypbestimmung verwendet. Die Autoren schlagen als Varianzen Werte von $\epsilon = 2\%$, 10% und 20% der mittleren Intensität vor. SNPs sollten ausgeschlossen werden, wenn sich damit 1% , 2% , 5% oder 10% der Genotypen ändern. Obwohl dieses Verfahren intuitiv erfassbar ist, mangelt es an der Möglichkeit der praktischen Umsetzung. Das Stören der Signalintensitäten durch Hinzufügen eines normalverteilten Fehlers müsste sehr häufig erfolgen, um die Variabilität der Veränderungen beurteilen zu können. Jedes Mal müssten die Genotypen aus den veränderten Signalintensitäten neu bestimmt werden. Dies wäre ein immenser Aufwand im Datenmanagement und an Rechenzeit.

4.1.5 Kombination von Kennzahlen

Neben diesen Statistiken, die immer nur ein Kriterium der Clustergüte berücksichtigen, existieren eine Vielzahl von Indizes, die sowohl Kompaktheit als auch Separierbarkeit bzw. Verbundenheit der Cluster berücksichtigen. Übersichten zu diesem Thema finden sich in [32, 33, 35, 40]. Eine Auswahl dieser Kriterien soll jetzt näher beschrieben werden.

Cluster Separation Criterion

Speziell auf die Problematik des Genotyp Calling zugeschnitten ist der Ansatz von Plagnol et al. [54]. Als Maß für die Separierbarkeit wird der Abstand der Zentren benachbarter Cluster betrachtet. Um die Kompaktheit der beiden Gruppen zu berücksichtigen, wird diese Differenz anschließend durch die Summe der Standardabweichungen dieser beiden Gruppen dividiert. Besonders ansprechend an diesem Maß ist, dass es lediglich die Abstände und Variabilität der Kontraste (siehe Gleichung (3.1)) berücksichtigt. Somit wird die typische langgezogene Form der Cluster ignoriert, die ansonsten zu Verzerrungen des Maßes führen könnte. Im Detail werden die Cluster-spezifischen Kontrast-Mittelwerte und -Streuungen mit \bar{c}_k und σ_k bezeichnet. Ohne Beschränkung der Allgemeinheit wird nun angenommen, dass die Cluster für $k = 1, 2, 3$ von links nach rechts geordnet sind. Daraus ergibt sich das *Cluster Separation Criterion* als

$$\text{CSC} = \min \left\{ \frac{\bar{c}_2 - \bar{c}_1}{\sigma_1 + \sigma_2}, \frac{\bar{c}_3 - \bar{c}_2}{\sigma_2 + \sigma_3} \right\}. \quad (4.2)$$

Für den Fall, dass pro SNP mehr als eine Gruppe untersucht wird, wie beispielsweise in Fall-Kontroll-Studien, sollte das Minimum der gruppenspezifischen CSCs verwendet werden. Das CSC gehört zu den *Dunn Indizes* (DI), bei denen ein Maß für die Separierbarkeit durch ein Maß für die Kompaktheit dividiert wird [33]. Üblicherweise wird dabei als Maß für die Kompaktheit das Maximum der Standardabweichungen σ_k verwendet und nicht, wie in Gleichung (4.2), die Summe. Ein guter SNP zeichnet sich durch große Inter-Cluster Abstände und kleine Variabilität innerhalb der Cluster aus. Daher sollten gute SNPs mit hohen Werten für das CSC oder andere DIs einhergehen. Der DI ist zwar intuitiv ver-

ständig und einfach zu berechnen, reagiert jedoch empfindlich auf Ausreißer.

Davies-Bouldin Index

Analog zum CSC, werden beim *Davies-Bouldin (DB) Index* die Inter-Cluster-Abstände mit den entsprechenden Intra-Cluster-Varianzen verglichen. Dazu wird das Verhältnis aus Summe der Intra-Cluster-Varianzen zu den Abständen zwischen den Clustern gebildet:

$$DB = \frac{1}{3} \sum_{k=1}^3 \max_{k \neq k'} \left\{ \frac{\sigma_k^2 + \sigma_{k'}^2}{d^2(\boldsymbol{\mu}_k, \boldsymbol{\mu}_{k'})} \right\}. \quad (4.3)$$

Der DB Index ähnelt einer F -Statistik, bei der die Quadratsummen zwischen den Gruppen durch die Quadratsummen innerhalb der Gruppen dividiert werden. Da die Cluster kompakt und gut getrennt sein sollten, spricht also ein niedriger DB Index für gut getrennte Genotypcluster. Die genaue Definition des DB Index kann auf verschiedene Weisen erfolgen, so haben Dixon et al. [30] beispielsweise den Euklidischen Abstand zur Bestimmung der Intra-Cluster-Variabilität und der Abstände zwischen den Clustern verwendet. Eine allgemeine Definition des DB Index ist in [40, 42] zu finden.

Caliński Harabasz Index

Ein Maß ähnlich zu DI, DB und CSC ist der *Caliński Harabasz (CH) Index*. Dies ist eine Pseudo- F -Statistik, die die Quadratsummen zwischen den Clustern mit denen innerhalb der Cluster vergleicht. Sie ist definiert als

$$CH = \frac{\sum_{k=1}^3 d^2(\boldsymbol{\mu}_k, \boldsymbol{\mu}) / 2}{\sum_{k=1}^3 \sum_{i=1}^{n_k} d_k^2(i, \boldsymbol{\mu}_k) / (n - 2)},$$

wobei $\boldsymbol{\mu}$ das Zentrum über alle drei Cluster ist.

Modifizierter Hubert Index

Mit dem modifizierten Hubert Index Γ wird die Zuordnung der Genotypen zu den Intensitäten bewertet. Dazu sei $d(i, i')$ der euklidische Abstand zweier Proben und $q(i, i')$ ist der Konkordanzindikator. Genauer ist

$$q(i, i') = \begin{cases} 1 & \text{falls } i \text{ und } i' \text{ derselbe Genotyp zugewiesen wurde,} \\ 0 & \text{falls } i \text{ und } i' \text{ verschiedene Genotypen zugewiesen wurden.} \end{cases}$$

Damit ist der *Modifizierte Hubert Index (MHI)* durch

$$\text{MHI} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n d(i, i') q(i, i').$$

gegeben. Alternativ kann der MHI auch in einer standardisierten Form verwendet werden. Dabei wird der Korrelationskoeffizient nach Pearson zwischen d und q anstelle des Produkts verwendet [17]. Dazu bezeichnen \bar{d} und s_d den durchschnittlichen Abstand und die durchschnittliche Standardabweichung über alle $n(n-1)/2$ Probenpaare. Analog dazu wird das arithmetische Mittel und die Standardabweichung des Konkordanzindikators mit \bar{q} und s_q bezeichnet. Dann ist der *standardisierte MHI (MHIS)* gegeben durch

$$\text{MHIS} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n \frac{(d(i, i') - \bar{d})(q(i, i') - \bar{q})}{s_d s_q}.$$

Silhouette-Statistik

Die Silhouette-Statistik von Rousseeuw [59], auch Silhouette-Index bzw. Silhouette-Score genannt, ist eine populäre Methode, um die Güte eines Clusterings zu beurteilen. Lovmar et al. [48] benutzten es, um die Qualität von TaqMan Daten zu beurteilen. Zur Berechnung der Silhouette-Statistik wird der durchschnittliche Abstand jeder Probe zu allen Proben desselben Clusters berechnet. Dieser Abstand wird dann mit dem durchschnittlichen Abstand der Probe zu allen Proben des nächstgelegenen Clusters verglichen. Im Detail sei $a_k(i)$ der durchschnittliche Abstand der Probe i des Clusters k zu allen Proben desselben Genotypclusters,

das heißt

$$a_k(i) = \frac{1}{n_k - 1} \sum_{i \neq i', i'=1}^{n_k} d_{k,k}(i, i').$$

Analog dazu wird mit $b_{k'}(i)$ der durchschnittliche Abstand von i zu Proben des Clusters k' , $k \neq k'$ definiert:

$$b_{k'}(i) = \frac{1}{n_{k'}} \sum_{i'=1}^{n_{k'}} d_{k,k'}(i, i').$$

$b_k(i)$ ist dann das Minimum der beiden durchschnittlichen Abstände:

$$b_k(i) = \min_{k' \neq k} \{b_{k'}(i)\}.$$

Die Silhouette $SI_k(i)$ misst dann, wie gut die Zuweisung des Genotype Calling war, indem die Unterschiede zwischen $b_k(i)$ und $a_k(i)$ betrachtet werden:

$$SI_k(i) = \frac{b_k(i) - a_k(i)}{\max\{a_k(i), b_k(i)\}} = \begin{cases} 1 - a_k(i)/b_k(i) & \text{falls } a_k(i) < b_k(i), \\ 0 & \text{falls } a_k(i) = b_k(i), \\ b_k(i)/a_k(i) - 1 & \text{falls } a_k(i) > b_k(i). \end{cases}$$

Für jeden Cluster wird nun der *cluster-spezifische Silhouette-Index* (SI_k) als arithmetisches Mittel aller Silhouetten eines Clusters bestimmt. Daraus wiederum ergibt sich der *Silhouette Index* (SI) als arithmetisches Mittel der SI_k :

$$SI_k = \frac{1}{n_k} \sum_{i=1}^{n_k} SI_k(i) \quad \text{und} \quad SI = \frac{1}{n} \sum_{k=1}^3 SI_k.$$

Falls ein Genotypcluster nur eine Probe enthält, wird die Silhouette dieser Probe gleich 0 gesetzt.

In Abbildung 4.4 ist das Konzept des Silhouette-Verfahrens dargestellt. Der SI ist auf das Intervall $[-1, 1]$ beschränkt. Klar trennbare Cluster haben Werte nahe 1, während Werte nahe -1 auf ein völliges Versagen des Calling Algorithmus hindeuten. Falls $SI_k(i)$ nahe 0 ist, könnte diese Probe genausogut dem nächstgelegenen Cluster zugeordnet werden. Ist $SI_k(i)$ negativ, würde sie besser in den nächstgelegenen Cluster passen.

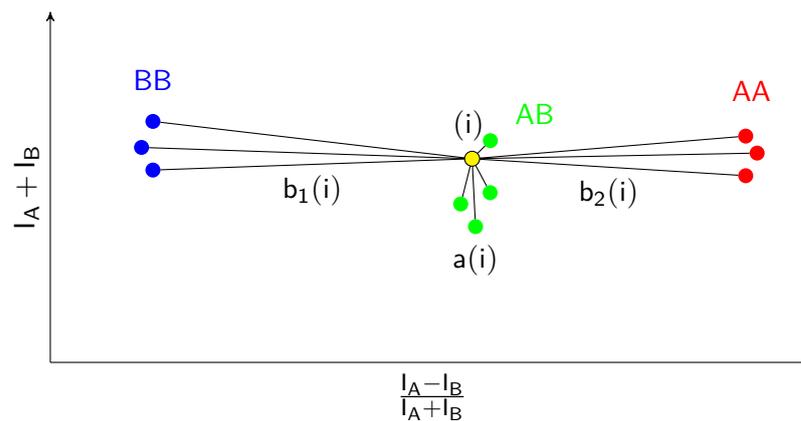


Abbildung 4.4: Prinzip der Silhouette-Statistik. Die Basisidee der Silhouette-Statistik zur Bewertung der Qualität eines Clusterings wird hier anhand einer Probe i aus dem Cluster $k = 2$ illustriert. Die Summe der Signalintensitäten $I_A + I_B$ ist gegen den Kontrast der Signalintensitäten $(I_A - I_B) / (I_A + I_B)$ aufgetragen. Für jede Probe i des Clusters k wird die Silhouette $s_k(i)$ berechnet. $a_k(i)$ ist der durchschnittliche euklidische Abstand von i zu allen Proben aus dem selben Genotypcluster (grün). $b_k(i)$ ist der durchschnittliche euklidische Abstand von i zu allen Proben des nächstgelegenen Clusters, also entweder $b_1(i)$ (blau) oder $b_3(i)$ (rot) [48, 59]. Die durchschnittliche Silhouette Breite $s(k)$ eines Clusters $k = 1, 2, 3$ ist das arithmetische Mittel aus allen Silhouettes $s(i)$ für jeden Genotype Cluster. Die *Overall Average Silhouette Width* wird als arithmetische Mittel der drei Cluster-Silhouette Breiten bestimmt. Abbildung nach Abbildung I von [48].

Alternativ zum Standard SI, bei dem die euklidischen Abstände über beide Achsen bestimmt werden, könnte man lediglich die euklidischen Abstände der Kontraste, wie für das CSC (siehe Abschnitt 4.1.5), verwenden. Ein weiterer einfacher Indikator für die Qualität eines SNPs ließe sich aus der Anzahl der positiven Silhouettes erzeugen. Dieser *modifizierte Silhouette Index (mSI)* ist dann der Anteil der Silhouettes $SI_k(i)$ mit einem positiven Vorzeichen [30].

ACPA-Ellipsen

Ein alternativer Ansatz wurde von Schillert et al. [62] verfolgt. Sie haben die Anzahl der Proben bestimmt, die zu dicht an einem benachbarten Cluster liegen. Falls diese Anzahl eine festgelegte Grenze überschreitet, wird dieser SNP ausgeschlossen. Dieser, *ACPA* genannte, Algorithmus wird in Tabelle 4.1.5 detailliert beschrieben. Zusammengefasst werden dabei die euklidischen Abstände der Proben eines Clusters zum Schwerpunkt bestimmt und anhand dieser Abstände

eine Ellipse um die Datenpunkte eines Clusters konstruiert. Im nächsten Schritt wird gezählt, wieviele Proben aus fremden Clustern innerhalb dieser Ellipse liegen. Diese Schritte werden für jeden Cluster wiederholt. Als Entscheidungskriterium dient die Anzahl aller Proben, die in einer falschen Ellipse liegen. Die Form der Ellipse wird anhand der ersten beiden Hauptkomponenten einer Hauptkomponentenanalyse des jeweiligen Clusters bestimmt. Durch Variation der Größe der

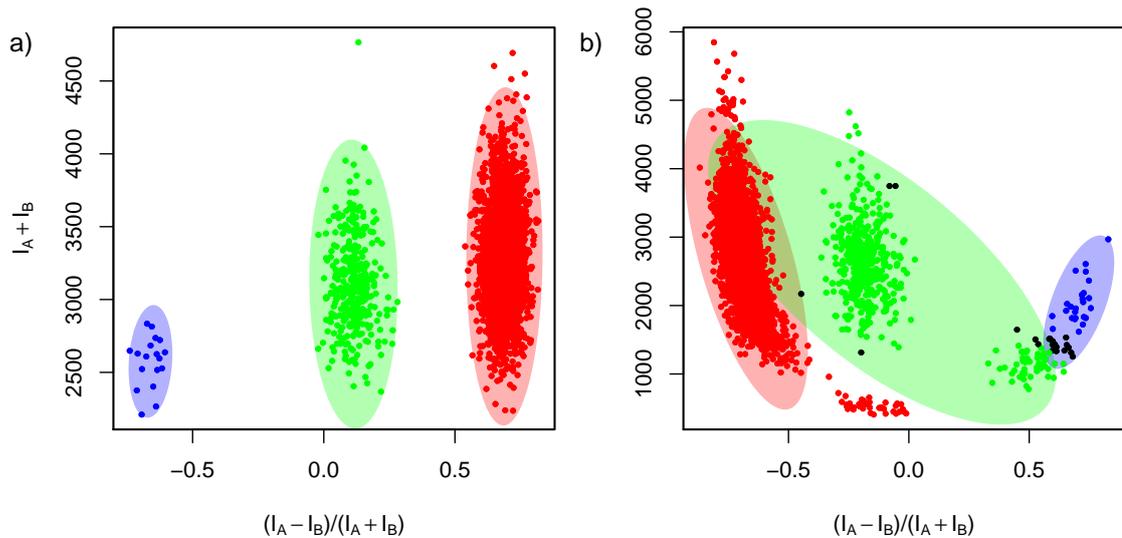


Abbildung 4.5: Beispiele für Cluster-Plots mit den ACPA-Ellipsen. Aufgetragen sind die Kontraste der normalisierten Intensitäten gegen die Summe der normalisierten Intensitäten. Es sind die selben SNPs wie in Abbildung 4.2. Während in a) die Ellipsen den Trend der Daten exakt wiedergeben ist in b) die grüne Ellipse durch die Fehler in der Genotypisierung aus der Senkrechten verrückt worden. Es entsteht eine große Überlappung mit Punkten des roten Clusters, weshalb dieser SNP von der weiteren Analyse ausgeschlossen werden sollte.

Ellipse lässt sich die Empfindlichkeit des Verfahrens verändern. Abbildung 4.5 zeigt die zwei SNPs aus Abbildung 4.2 erneut, jedoch in der Kontrast-Darstellung und mit den ACPA-Ellipsen. In der rechten Abbildung ist deutlich zu erkennen, dass die Ellipse für den Heterozygoten-Cluster, grün dargestellt, die Punktwolke des linken homozygoten-Clusters, rot dargestellt, teilweise überdeckt. Die Anzahl von Proben in einer falschen Ellipse ist hier sehr hoch, was für ein fehlgeschlagenes Calling spricht.

Tabelle 4.1: Der Algorithmus von ACPA

1. Für Cluster $k = 1$ bis 3:
 - a) Führe Hauptkomponentenanalyse mit Proben des Clusters k durch.
 - b) Transformiere alle Daten anhand der ersten zwei Hauptkomponenten aus Schritt a).
 - c) Berechne den euklidischen Abstand vom Zentrum des Clusters k zu allen Proben des Clusters
 - d) Definiere die Clustergrenze b als $b = q_3 + f \text{IQR}$, wobei q_3 das obere Quartil und IQR der Interquartilsabstand der Abstände aus c) ist. Standardwert: $f = 3$.
 - e) Berechne die euklidischen Abstände vom Zentrum des Clusters k zu alle Proben, die nicht aus k sind.
 - f) Zähle die Anzahl c_k der Proben, die nicht zu k gehören, deren Abstand zum Zentrum von k aber kleiner als f ist.
2. Berechne die Summer alle Proben, die die Grenzen eines anderen Clusters überschreiten, das heißt $c = \sum c_k$.
3. Die Genotypbestimmung eines SNPs wird als unzureichend eingestuft, falls c eine Grenze t überschreitet. Standardwert: $t = 25$.

4.2 Generierung von Cluster-Plots mit dem R-Paket *acpa*

Im vorangegangenen Abschnitt wurden Maße zur Beurteilung der Validität des Clusterings definiert. In diesem Abschnitt wird deren praktische Umsetzung beschrieben. Die Herausforderung ist hier, dass die Maßzahlen für jeden beliebigen SNP einer GWA-Studie bestimmt werden müssen. Dies ist sehr rechenintensiv.

4.2.1 Grundlagen

Im Detail sind dabei zwei Probleme maßgeblich:

1. Wie bereits in Abschnitt 3.3 angedeutet, werden innerhalb eines Callings die Genotypen getrennt von den Intensitäten abgelegt. Diese müssen also wieder SNP-spezifisch zusammengeführt werden. Erschwerend kommt hinzu, dass für Fall-Kontroll-Studien die Daten der Fälle und Kontrollen getrennt aufbereitet wurden, die Cluster-Plots für alle Teile einer Studie aber simultan generiert werden sollen.
2. Aufgrund der Menge an individuellen Genotyp- und Intensitätsinformationen stößt man an die Grenzen heutiger Rechnersysteme. Für eine, inzwischen übliche Studie mit 3.000 Probanden und dem Genome-Wide Human SNP Array 6.0 von Affymetrix müssen $906.600 \cdot 3.000 \cdot 2 = 5,4 \cdot 10^9$ Daten jeweils für die Genotypen und die Intensitäten gespeichert werden. Während für die Speicherung der Genotypdaten die Möglichkeit besteht, diese binär kodiert abzulegen und damit massiv Speicherplatz eingespart wird [12, 55], besteht diese Möglichkeit für die Intensitätsdaten nicht. Für die hier beschriebene Konstellation beträgt die physikalische Größe der Genotypdatei ca. 1 GB, die der Intensitätsdatei jedoch 60 GB. Somit sind Methoden notwendig, die die SNPs nacheinander abarbeiten.

Aufgrund dieser Anforderungen, wurde die Generierung der Cluster-Plots als R-Paket implementiert. R ist eine weit verbreitete Open-Source Software, die ihre Stärken in der Implementierung statistischer Methoden und der visuellen Darstellung von Daten hat [56]. R ist beliebig erweiterbar, wobei eine Vielzahl von Erweiterungen, Pakete genannt, bereits existiert. Eines dieser R-Pakete ist GenABEL von Aulchenko et al. [12], welches für die Verwaltung und Auswertung von Hochdurchsatz-Genotypdaten entwickelt wurde und jetzt hier zur Verwaltung der Genotypdaten für das Paket *acpa* (Automated Cluster Plot Analysis) eingesetzt wird. In der jetzigen Entwicklungsstufe können Daten der Calling-Algorithmen BRLMM, CRLMM, Birdseed, Chiamo und JAPL ausgewertet werden.

4.2.2 Aufbau des R-Paketes *acpa*

Bei der Zusammenführung der Intensitätsdaten mit den Genotypdaten sind eine Reihe von Punkten zu beachten:

- Die SNPs werden in den Intensitätsdateien mit einer so genannten Affymetrix-ID bezeichnet, die Genotypdaten jedoch mit rs-Nummern.
- Die ID der Proben in der Genotypdatei muss nicht identisch mit dem Namen der CEL-Datei sein, die für diesen Probanden verwendet wurde.
- Die Intensitätsdaten liegen für einen SNP in mehreren Dateien vor, zum Beispiel bei Fall-Kontroll-Studien.

In Abbildung 4.6 ist die Verarbeitung und Zusammenführung der notwendigen Informationen schematisch dargestellt. Aufgrund der Größe der Intensitätsdatei erfolgt die Auswertung in Stücken, das heißt, immer nur eine bestimmte Anzahl an Zeilen der Intensitätsdatei wird eingelesen und bearbeitet. Um die Zugriffszahlen auf den Massenspeicher gering zu halten, sollten mindestens 50 Zeilen simultan bearbeitet werden. Die Größe dieser Stücke ist über einen Parameter wählbar. Des Weiteren besteht die Möglichkeit, anhand einer Liste nur bestimmte SNPs für die Analyse auszuwählen. Dies entspricht der üblichen Vorgehensweise bei einer GWA-Studie, bei der man die SNPs, die eine starke Assoziation zeigten, auf korrekte Cluster-Plots überprüft. Weiterhin besteht dazu die Möglichkeit, die SNPs interaktiv zu bewerten, das heißt die Cluster-Plots für eine Reihe von SNPs werden nacheinander gezeigt und die Eingabe zur Qualität des SNPs wird protokolliert.

4.3 Automatische Bewertung

In diesem Abschnitt wird die Güte der automatischen Evaluation der Cluster-Plots empirisch untersucht.

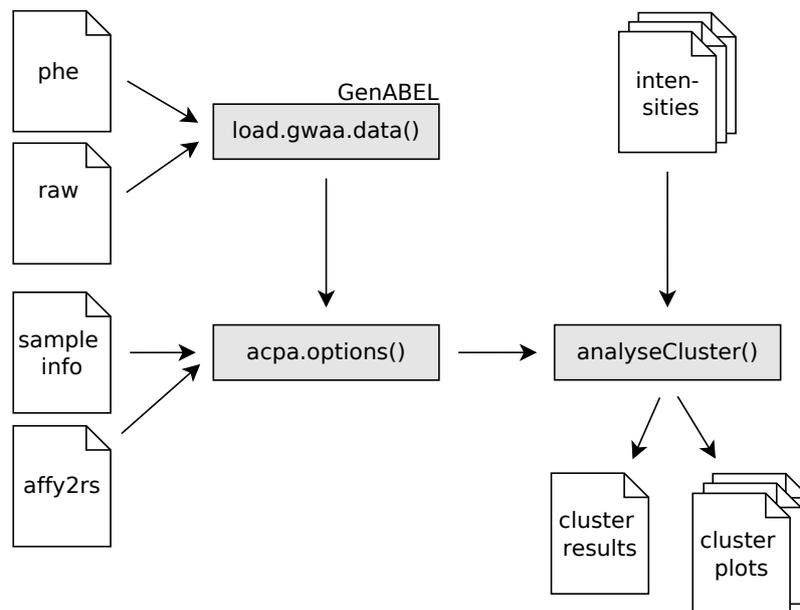


Abbildung 4.6: Struktur von `acpa`. Es wird erwartet, dass die Genotypdaten binär kodiert in einer `raw`-Datei und die zugehörigen Phänotypinformation in einer `phe`-Datei abgespeichert sind. Diese werden dann mit der Funktion `load.gwaa.data()` aus dem GenABEL-Paket eingelesen. Zusammen mit Informationen zu der Zuordnung zwischen Proben-ID und CEL-Namen (`sample info`) und der Zuordnung von Affymetrix-ID und rs-ID wird dies an die Funktion `acpa.options()` übergeben, welche die Korrektheit der Argumente überprüft wird. Die eigentliche Auswertung erfolgt mittels `analyseCluster()`, wofür noch die Dateinamen der Intensitätsdaten benötigt werden. Je nach Konfiguration werden Clustermaße berechnet und/oder Cluster-Plots produziert. Diese werden entweder in einer pdf-Datei zusammengefasst, oder pro SNP wird eine png-Datei produziert.

4.3.1 Vergleichsgrundlage: Gutenberg-Herz-Studie

Im Rahmen der Gutenberg-Herz-Studie, einer prospektiven Kohortenstudie aus dem Raum Mainz, werden neben einer Vielzahl von klinischen Variablen auch genetische Marker untersucht. Für die ersten 3500 Probanden, das so genannte A1-Kollektiv, lag die Genotypinformation in Form von CEL-Dateien für den Affymetrix Genome-Wide Human SNP Array 6.0 vor. Daraus wurden mit Birdseed [41] die Genotypen bestimmt. Zuerst erfolgte die Qualitätskontrolle auf Proben-Ebene Es wurden Personen ausgeschlossen, falls

- der Anteil fehlender Genotypen größer als 3% war, oder

- der Anteil heterozygoter Genotypen mehr als drei Standardabweichungen von der durchschnittlichen Heterozygotie entfernt war.

Anschließend fand die Qualitätskontrolle auf Marker-Ebene statt. Dabei wurden die Travemünde-Kriterien angewandt [60]. SNPs wurden ausgeschlossen, falls

- sie bei weniger als 98% der Personen bestimmt werden konnten (für Fälle und Kontrollen separat),
- der Anteil des seltenen Allels kleiner als 1% ist oder
- der p-Wert für die Abweichung vom Hardy-Weinberg-Gleichgewicht kleiner als 0,0001 ist.

Für die Daten des A1-Kollektivs standen damit für 3.194 Personen die Genotypinformation für 649.461 Marker zur Verfügung.

Tabelle 4.2: Häufigkeitsverteilung der Qualität der SNPs

		sicher			
		beide ja	nur einer	beide unsicher	
gut	beide ja	4563	222	44	4829
	uneinig	28	58	39	125
	beide nein	26	12	8	46
		4617	292	91	5000

Zeilenweise ist die Anzahl der SNPs aufgetragen, die beide, nur einer oder keiner der Beurteiler für gut befunden hat. Spaltenweise erfolgt die Abstufung nach dem Grad der Sicherheit. Die fett markierten Zahlen kennzeichnen die Anzahl von SNPs, die tatsächlich verwendet wurden, nämlich 4.617, von denen 26 als schlecht gekennzeichnet wurden.

4.3.2 Methoden

Die Clustermaße wurden auf ihre Eignung zur Bewertung der SNP-Qualität überprüft. Um die Ergebnisse der automatischen Bewertung einschätzen zu können, wurden diese mit dem Goldstandard verglichen. Dieser Goldstandard besteht in der unabhängigen und verblindeten Beurteilung der Cluster-Plots durch zwei erfahrene Beurteiler. Dabei wurde neben der eigentlichen gut/schlecht Entscheidung auch die Sicherheit der Antwort abgefragt. Dafür wurden 5000 SNPs *zufällig* aus den *qualitätskontrollierten* SNPs des A1-Kollektivs ausgewählt. In Tabelle 4.2 sind die Ergebnisse dieser Bewertung aufgelistet. Für 4875 SNPs stimmten die Bewertungen der Beurteiler überein, für 4617 dieser SNPs waren sich beide Beurteiler auch sicher. Um eine realistische Einschätzung der Güte der Clustermaße zu erhalten, werden im Folgenden lediglich diese 4617 SNPs für die Auswertung verwendet. Von diesen wurden 26 als schlecht eingestuft. Dies bedeutet, dass in einer GWA-Studie für den Human SNP Array 6.0 ca. 5085 SNPs mit „schlechten“ Cluster-Plots zu erwarten sind. Das zugehörige 95%-Konfidenzintervall beträgt [4947, 5227].

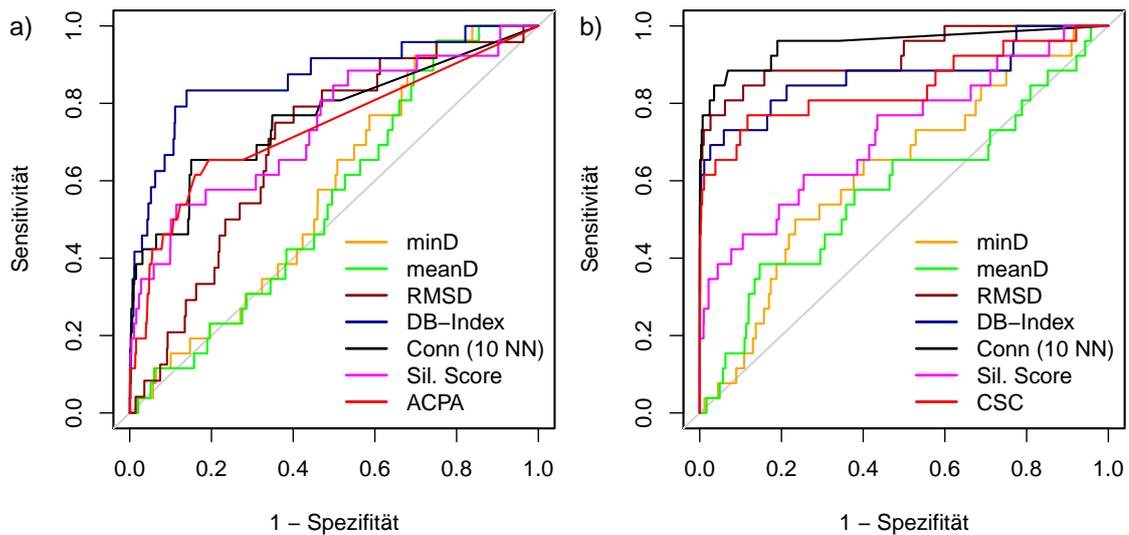


Abbildung 4.7: ROC-Kurven der Clustermaße. In a) wurden die normalisierten Intensitäten (I_A, I_B) zur Berechnung der Clustermaße verwendet, in b) nur der Kontrast $((I_A - I_B) / (I_A + I_B))$. Die Verbundenheit, beruhend auf den 10 nächsten Nachbarn – Conn (10 NN), erreicht bei einer Spezifität nahe 100% eine Sensitivität von ca. 65%. Eine vergleichbare Güte zeigen RMSD, DB Index und CSC. Keines der Maße schneidet unter Verwendung der normalisierten Intensitäten besser ab, als unter Verwendung des Kontrastes.

4.3.3 Ergebnisse

Für eine Auswahl der in Abschnitt 4.1 vorgestellten Clustermaße wurden die typischen Kennzahlen des diagnostischen Testens berechnet und in Abbildung 4.7 in zwei *Receiver Operating Characteristic* (ROC)-Kurven grafisch dargestellt. Da das CSC lediglich die Kontraste zur Berechnung der Lage- und Streumaße berücksichtigt, wurde dieses Prinzip auf alle Maße angewandt und der herkömmlichen Variante unter Verwendung der Intensitäten I_A und I_B gegenüber gestellt. Da dieser zusätzliche Screening-Schritt nicht dazu führen sollte, dass SNPs mit einer problemlosen Genotypzuweisung als problematisch klassifiziert werden, ist darauf zu achten, dass die Spezifität sehr hoch ist. Die höchste Sensitivität für die Spezifität von 99,9% beträgt 65,4% und wird unter Verwendung der Kennzahl für die Verbundenheit *Conn* erreicht. In die Berechnungen floss dabei die Information über die Gruppenzugehörigkeit der 10 nächsten Nachbarn ein. Zur Berechnung der Abstände zwischen den Proben wurden die Kontraste verwendet.

4.4 Diskussion

Mit dem neu entwickelten R-Paket *acpa* besteht die Möglichkeit, alle Cluster-Plots einer GWA-Studie automatisch zu beurteilen. Die Zeit für die Berechnung der implementierten Clustermaße beträgt für eine GWA-Studie mit 3500 Proben für den Human SNP Array 6.0 nach Parallelisierung auf acht Prozesse derzeit etwa eine Woche. Diese Zeit lässt sich auf zwei Weisen verkürzen. Zum einen können mehr als acht Prozesse zeitgleich laufen. Wird *acpa* z.B. auf einem Computer mit 40 Kernen gestartet, reduziert sich die Rechenzeit auf weniger als drei Tage. Zum anderen müssen auch nicht alle Maße gleichzeitig berechnet werden. So hat sich in dieser Arbeit gezeigt, dass ein Teil der Maßzahlen, wie z.B. *minD* und *meanD* nicht gut zur automatischen Beurteilung der Cluster-Plots geeignet sind. Eine Beschränkung auf die zentralen drei Maße würde allerdings nur eine geringe Reduktion der Rechenzeit bewirken, da der größte CPU-Zeitaufwand im Datenmanagement liegt.

In Fall-Kontroll-Studien wird das Calling von Fällen und Kontrollen üblicherweise getrennt durchgeführt. Um Unterschiede in den Genotypisierungsfehlern auszuschließen, ist es daher erforderlich, die Cluster-Plots beider Gruppen gemeinsam zu betrachten. Dies ist besonders relevant, da durch diese Unterschiede in den Genotypisierungsfehlern der Anteil der falsch-positiven Werte stark ansteigt [23]. Entsprechend müssen die Clustermaße auch für beide Gruppen berechnet werden. Diese Option wurde in *acpa* realisiert: Es können mehrere Intensitätsdateien angegeben und die Cluster-Plots dieser Gruppen gemeinsam angezeigt werden. In dieser Arbeit wurde das Calling als Problem der Clusteranalyse verstanden, so dass die im Rahmen der Clusteranalyse entwickelten Clustermaße zur Beurteilung der internen Validität zur automatischen Beurteilung von Cluster-Plots verwendet werden können. Eine wesentliche Erkenntnis dieser Arbeit ist dabei, dass die Wahl der Transformation für die Intensitäten die Güte der Vorhersage maßgeblich beeinflusst. Betrachtet man beispielsweise das Maß *Connectivity*, so sieht man einen deutlich Anstieg der Sensitivität durch die Transformation. Für eine Spezifität von 99,9% ergibt sich in der Darstellung (I_A, I_B) eine Sensitivität von 28%, während die Transformation auf die Kontraste eine Sensitivität von 65% liefert. Durch diese Transformation der Intensitätswerte können mehr als doppelt so viele SNPs mit fehlgeschlagenem Calling identifiziert werden. Auch wenn *acpa* schon jetzt ein viel versprechender Ansatz zur automatischen

Beurteilung von Cluster-Plots ist und unter Beibehaltung fast aller SNPs von hoher Qualität etwa zwei Drittel der SNPs geringer Qualität ausschließt, lassen sich derzeit nicht alle Formen schlechter Cluster-Plots mit *acpa* identifizieren. Dieses gilt insbesondere für SNPs, die die kanonische Clusterzahl von drei überschreiten. Daher sollen in einem weiterführenden DFG-Projekt Alternativen zur Beurteilung mittels Clustervaliditätsmaßen untersucht werden. Ist die Clusterzahl größer als drei, liegt eine endliche Mischverteilung vor, die es zu identifizieren gilt. In dem Projektteil werden dabei zwei Ansätze verfolgt werden. Zum einen werden Tests auf Mischung zweier Normalverteilungen betrachtet werden. Diese haben allerdings den Nachteil, dass bestimmte Modelle nicht identifizierbar und daher nicht testbar sind [64]. Allerdings haben Chen und Li [21] vor kurzem einen Ansatz vorgestellt, der das Problem der Nicht-Identifizierbarkeit der Modelle umgehen kann. Zum anderen soll untersucht werden, ob durch die Schätzung der Anzahl der Modi der bivariaten Verteilung der Intensitäten [63] die problematischen Cluster-Plots identifiziert werden können.

Die Frage, ob das Calling für einen SNP erfolgreich ist oder nicht, ist oft nicht studienspezifisch, sondern hängt von den Eigenschaften der Sonde des SNPs, bzw. des zugehörigen DNA-Fragmentes ab. Insbesondere Länge und Sequenz spielen dabei eine große Rolle. Gibt es bezüglich des Hybridisierungsverhaltens starke Unterschiede zwischen den Sonden für das A-Allel und für das B-Allel lassen sich die Genotypen nicht sauber unterscheiden. Dies ist anhand der Cluster-Plots erkennbar. Über mehrere Studien hinweg kann man so eine Liste von SNPs erstellen, die grundsätzlich von den weiteren Analysen ausgeschlossen werden sollten.

5 Zusammenfassung

Einleitung

Genomweite Assoziationsstudien sind zur Zeit der Standardansatz in der Genetischen Epidemiologie zur Identifikation neuer Gene für komplexe genetische Erkrankungen. Der technologische Fortschritt im Bereich der Chip-Technologie ermöglicht die simultane Untersuchung von bis zu einer Million Einzelnukleotid-Polymorphismen (SNPs). Dies stellt große Anforderungen an die Qualitätssicherung zur Vermeidung falscher Schlussfolgerungen. Im so genannten Calling werden Signalintensitäten in Genotypen umgewandelt. Hierfür wurden eine Reihe verschiedener Algorithmen entwickelt, deren Eigenschaften und Güte bisher nicht umfassend miteinander verglichen wurden. Nach dem Calling muss eine wirksame Qualitätskontrolle erfolgen. Hier kommt der Beurteilung von Cluster-Plots eine große Bedeutung zu, da dieses das Mittel der Wahl ist, um SNPs zu identifizieren, bei denen das Calling fehlgeschlagen ist. Entsprechend dieser beiden fundamentalen Aspekte der Qualitätssicherung werden in dieser Dissertationsschrift verschiedene Calling-Algorithmen verglichen und ein automatisches Verfahren zur Beurteilung von Cluster-Plots vorgestellt.

Material und Methoden

Die Charakteristika der mittels einer Literaturrecherche identifizierten Calling Algorithmen wurden beschrieben. Anschließend wurden die Algorithmen anhand der Übereinstimmung mit den HapMap-Daten miteinander verglichen. Zur Beurteilung der Güte der Genotypbestimmung werden Clustervaliditätsmaße genutzt. Dazu wurde das R-Paket `acpa` geschrieben, das neben der Generierung von Cluster-Plots auch die Berechnung der Clustervaliditätsmaße gestattet. Die Güte dieser Maße wurde mittels ROC-Kurven anhand eines Datensatzes aus der Gutenberg-Herz-Studie bestimmt. Als Goldstandard diente dabei die Bewertung durch zwei erfahrene Beurteiler.

Ergebnisse

Für fünf der neun identifizierten Algorithmen war funktionsfähige Software erhältlich. Von diesen zeigte BRLMM die größte Übereinstimmung mit den HapMap-Daten. Die Unterschiede zu Birdseed und CRLMM waren jedoch gering. Die verbleibenden zwei Algorithmen CHIAMO und JAPL schnitten deutlich schlechter ab. Alle Algorithmen zeigten geringere Konkordanzen für SNPs mit niedriger Allelfrequenz.

Das R-Paket *acpa* ermöglicht die Verwendung nahezu beliebig großer Intensitätsdateien. Die Auswertung der Clustermaße zeigte für das Maß *Connectivity* bei einer Spezifität von 99,9% eine Sensitivität von 65,4%. Das heißt, dass zwei Drittel der SNPs, deren Calling fehlgeschlagen ist, mit *acpa* identifiziert wurden, während gleichzeitig nahezu alle SNPs guter Qualität auch als solche erkannt wurden.

Diskussion

In dieser Arbeit wurde gezeigt, dass die Wahl des Calling-Algorithmus die Qualität der Genotypdaten maßgeblich beeinflusst. Unter Berücksichtigung der Übereinstimmung mit den Daten und der benutzerfreundlichen Implementierung wird CRLMM empfohlen. Für die anschließende Qualitätskontrolle der Daten mittels der Beurteilung der Cluster-Plots wurde das R-Paket *acpa* entwickelt. Die dort implementierten Clustervaliditätsmaße sind gut geeignet, um SNPs mit fehlerhaftem Calling zu entdecken. Dabei erhöht sich die Vorhersagekraft deutlich, wenn statt der Intensitäten der Kontrast der Intensitäten zur Berechnung der Lage- und Streumaße verwendet wird.

Literaturverzeichnis

- [1] EndNote - bibliographies made easy. (Tag des Zugriffs: 07.01.2009)
<http://www.endnote.com/>

- [2] Zotero. (Tag des Zugriffs: 07.01.2009)
<http://www.zotero.org/>

- [3] What is the HapMap. (Tag des Zugriffs: 09.04.2010)
<http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html>

- [4] PLINK. (Tag des Zugriffs: 10.04.2010)
<http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml>

- [5] BRLMM algorithm for Affymetrix mapping arrays. (Tag des Zugriffs: 11.03.2010)
http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf

- [6] Guide to probe logarithmic intensity error (PLIER estimation. (Tag des Zugriffs: 11.03.2010)
http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf

- [7] Affymetrix HapMap 500K data. (Tag des Zugriffs: 22.09.2009)
http://hapmap.ncbi.nlm.nih.gov/downloads/raw_data/affy500k/

- [8] Affymetrix - image library. (Tag des Zugriffs: 29.01.2010)

- http://www.affymetrix.com/about_affymetrix/media/image-library.affx
- [9] CelQuantileNorm. (Tag des Zugriffs: 30.03.2010)
<http://www.wtccc.org.uk/info/software.shtml>
- [10] Affymetrix: Affymetrix Genotyping Console 3.0 User Manual. Affymetrix, Santa Clara, CA (2008)
- [11] Aitkin M, Rubin DB: Estimation and hypothesis testing in finite mixture models. *J Roy Stat Soc B Met* 47, 67–75 (1985)
- [12] Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294–6 (2007)
- [13] Bacanu SA, Devlin B, Roeder K: The power of genomic control. *Am J Hum Genet* 66, 1933–1944 (2000)
- [14] Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS: Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* (2009)
- [15] Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, Schulze TG, Cichon S, Rietschel M, Nöthen MM, Georgi A, Schumacher J, Schwarz M, Jamra RA, Höfels S, Propping P, Satagopan J, Detera-Wadleigh SD, Hardy J, McMahon FJ: A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 13, 197–207 (2008)
- [16] Bengtsson H, Simpson K, Bullard J, Hansen K: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Technical Report 745, Department of Statistics, University of California, Berkeley (2008)

-
- [17] Bezdek JC, Pal NR: Some new indexes of cluster validity. *IEEE T Syst Man Cy B* 28, 301–315 (1998)
- [18] Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193 (2003)
- [19] Boulesteix A: Over-optimism in bioinformatics research. *Bioinformatics* 26, 437–439 (2010)
- [20] Carvalho B, Bengtsson H, Speed TP, Irizarry RA: Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* 8, 485–99 (2007)
- [21] Chen J, Li P: Hypothesis test for normal mixture models: The EM approach. *Ann Stat* 37, 2523–2542 (2009)
- [22] Clark DP: *Molecular Biology: Das Original mit Übersetzungshilfen: Understanding the Genetic Revolution*. Spektrum Akademischer Verlag, München, 1. Aufl. (2006)
- [23] Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JMM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA: Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 37, 1243–6 (2005)
- [24] Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC, Warram JH, Todd JA: Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 40, 1399–1401 (2008)
- [25] Cutler DJ, Zwick ME, Carrasquillo MM, Yohn CT, Tobin KP, Kashuk C, Mathews DJ, Shah NA, Eichler EE, Warrington JA, Chakravarti A: High-

- throughput variation detection and genotyping using microarrays. *Genome Res* 11, 1913–1925 (2001)
- [26] Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG: The Affymetrix GeneChip platform: an overview. *Methods Enzymol* 410, 3–28 (2006)
- [27] de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF: Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 17, R122–128 (2008)
- [28] Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 55, 997–1004 (1999)
- [29] Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, mei Shen M, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S: Dynamic model based algorithms for screening and genotyping over 100 k SNPs on oligonucleotide microarrays. *Bioinformatics* 21, 1958–1963 (2005)
- [30] Dixon SJ, Heinrich N, Holmboe M, Schaefer ML, Reed RR, Trevejo J, Brereton RG: Use of cluster separation indices and the influence of outliers: application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles. *J Chemometr* 23, 19–31 (2009)
- [31] Halkidi M, Batistakis Y, Vazirgiannis M: Clustering algorithms and validity measures. In: *Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on*, 3–22 (2001)
- [32] Halkidi M, Batistakis Y, Vazirgiannis M: Cluster validity methods: part I. *SIGMOD Rec* 31, 40–45 (2002)
- [33] Halkidi M, Batistakis Y, Vazirgiannis M: Clustering validity checking methods: part II. *SIGMOD Rec* 31, 19–27 (2002)

- [34] Halkidi M, Vazirgiannis M, Batistakis Y: Quality scheme assessment in the clustering process. In: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, 265–276: Springer-Verlag (2000)
- [35] Handl J, Knowles J, Kell DB: Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 3201–3212 (2005)
- [36] Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA: SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics* 23, 57–63 (2007)
- [37] Huentelman MJ, Craig DW, Shieh AD, Corneveaux JJ, Hu-Lince D, Pearson JV, Stephan DA: SNiPer: improved SNP genotype calling for affymetrix 10K GeneChip microarray data. *BMC Genomics* 6, 149 (2005)
- [38] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264 (2003)
- [39] Kennedy GC, Matsuzaki H, Dong S, min Liu W, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SPA, Jones KW: Large-scale genotyping of complex DNA. *Nat Biotechnol* 21, 1233–7 (2003)
- [40] Kim M, Ramakrishna RS: New indices for cluster validity assessment. *Pattern Recogn Lett* 26, 2353–2363 (2005)
- [41] Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 40, 1253–1260 (2008)

- [42] Kovacs F, Legany C, Babos A: Cluster validity measurement techniques. In: Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence: Budapest Tech: Budapest (2005)
- [43] Laframboise T, Harrington D, Weir BA: PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. *Biostatistics* 8, 323–336 (2007)
- [44] Lamy P, Andersen CL, Wikman FP, Wiuf C: Genotyping and annotation of affymetrix SNP arrays. *Nucleic Acids Res* 34, e100 (2006)
- [45] Lin S, Carvalho B, Cutler D, Arking D, Chakravarti A, Irizarry R: Validation and extension of an empirical bayes method for SNP calling on affymetrix microarrays. *Genome Biol* 9, R63 (2008)
- [46] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: High density synthetic oligonucleotide arrays. *Nat Genet* 21, 20–4 (1999)
- [47] Liu W, Di X, Yang G, Matsuzaki H, Huang J, Mei R, Ryder TB, Webster TA, Dong S, Liu G, Jones KW, Kennedy GC, Kulp D: Algorithms for large-scale genotyping microarrays. *Bioinformatics* 19, 2397–2403 (2003)
- [48] Lovmar L, Ahlford A, Jonsson M, Syvanen A: Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics* 6, 35 (2005)
- [49] Manolio TA, Collins FS: The HapMap and genome-wide association studies in diagnosis and therapy. *Annu Rev Med* 60, 443–456 (2009)
- [50] Moorhead M, Hardenbol P, Siddiqui F, Falkowski M, Bruckner C, Ireland J, Jones HB, Jain M, Willis TD, Faham M: Optimal genotype determination in highly multiplexed SNP data. *Eur J Hum Genet* 14, 207–15 (2006)
- [51] Mosteller F, Tukey JW: Data Analysis and Regression: A Second Course in Statistics. Addison Wesley, Reading MA, 1. Aufl. (1977)

-
- [52] Mueller PW, Rogus JJ, Cleary PA, Zhao Y, Smiles AM, Steffes MW, Bucksa J, Gibson TB, Cordovado SK, Krolewski AS, Nierras CR, Warram JH: Genetics of kidneys in diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *J Am Soc Nephrol* 17, 1782–1790 (2006)
- [53] Nicolae DL, Wu X, Miyake K, Cox NJ: GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics* 22, 1942–7 (2006)
- [54] Plagnol V, Cooper JD, Todd JA, Clayton DG: A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet* 3, e74 (2007)
- [55] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81, 559–575 (2007)
- [56] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009). ISBN 3-900051-07-0
URL <http://www.R-project.org>
- [57] Rabbee N, Speed TP: A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 22, 7–12 (2006)
- [58] Rigai G, Hupe P, Almeida A, Rosa PL, Meyniel J, Decraene C, Barillot E: ITALICS: an algorithm for normalization and DNA copy number calling for affymetrix SNP arrays. *Bioinformatics* 24, 768–774 (2008)
- [59] Rousseeuw P: Silhouettes- a graphical aid to the interpretation and validation of cluster-analysis. *J Comput Appl Math* 20, 53–65 (1987)
- [60] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann H, Barrett JH, König IR, Stevens

- SE, Szymczak S, Tregouet D, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H: Genomewide association analysis of coronary artery disease. *N Engl J Med* 357, 443–53 (2007)
- [61] Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I: A hidden markov model for joint estimation of genotype and copy number in high-throughput SNP chips. *Johns Hopkins University, Dept. of Biostatistics Working Papers* 136 (2007)
- [62] Schillert A, Schwarz DF, Vens M, Szymczak S, König IR, Ziegler A: ACPA: automated cluster plot analysis of genotype data. *BMC Proc* 3 Suppl 7, S58 (2009)
- [63] Schlattmann P: Medical Applications of Finite Mixture Models. Springer, Berlin (2009)
- [64] Tarpey T, Yun D, Petkova E: Model misspecification: finite mixture or homogeneous? *Statistical Modeling* 8, 199–218 (2008)
- [65] Teo YY: Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol* 19, 133 (2008)
- [66] Teo YY, Small KS, Clark TG, Kwiatkowski DP: Perturbation analysis: a simple method for filtering SNPs with erroneous genotyping in genome-wide association studies. *Ann Hum Genet* 72, 368–374 (2008)
- [67] The International HapMap Consortium: The international HapMap project. *Nature* 426, 789–796 (2003)
- [68] The International HapMap Consortium: A second generation human haplotype map of over 3,1 million SNPs. *Nature* 449, 851–861 (2007)
- [69] Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G,

- Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082 (1998)
- [70] Wang Y, Miao Z, Pommier Y, Kawasaki ES, Player A: Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips. *Bioinformatics* 23, 2088–2095 (2007)
- [71] Weale ME: Quality control for genome-wide association studies. *Methods Mol Biol* 628, 341–372 (2010)
- [72] Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–78 (2007)
- [73] Xiao Y, Segal MR, Yang YH, Yeh R: A multi-array multi-SNP genotyping algorithm for affymetrix SNP microarrays. *Bioinformatics* 23, 1459–67 (2007)
- [74] Ziegler A: Genome-wide association studies: quality control and population-based measures. *Genet Epidemiol* 33, S45–S50 (2009)
- [75] Ziegler A, König IR: A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform. Wiley-VCH Verlag, Weinheim, 2. Aufl. (2010)
- [76] Ziegler A, König IR, Thompson JR: Biostatistical aspects of genome-wide association studies. *Biom J* 50, 8–28 (2008)

Danksagung

Zum Gelingen dieser Arbeit haben viele Personen beigetragen, dafür gebührt Ihnen mein Dank.

An erster Stelle sei Herr Univ.-Prof. Dr. A. Ziegler genannt. Ihm danke ich für die Vergabe des Themas und die umfassende wissenschaftliche Betreuung, sowie für die Möglichkeit in zwei sehr spannenden Projekten mitarbeiten zu dürfen.

Herrn Univ.-Prof. Dr. S. Blankenberg danke ich für den Zugang zu den Daten der Gutenberg-Herz-Studie und die angenehme Zusammenarbeit. Der tiefere Einblick in eine genetisch-epidemiologische Studie dieser Größe war sehr beeindruckend.

Bei Frau Dr. T. Zeller bedanke ich mich für die Bereitstellung der CEL-Dateien und die umfassende Einführung in die Laborseite der Microarray-Technologie.

Für die technische Unterstützung bei der Implementierung von `acpa bin` bin ich Herrn O.-J. Frahm zu Dank verpflichtet.

Bei Frau M. Vens, M.Sc. bedanke ich mich für die Offenlegung der Fallstricke, die in der Quantilnormalisierung für CHIAMO enthalten sind.

Frau Dipl.-Inform. S. Szymczak und Herr Dipl.-Inf. D.F. Schwarz haben die ermüdende Aufgabe der visuellen Beurteilung von 5000 Cluster-Plots übernommen. Dafür und für das Korrekturlesen gebührt Ihnen mein Dank.

Lebenslauf

Studium

- ab 11/2007 Promotionsstudium der Humanbiologie,
Universität zu Lübeck
- 04/2006 – 10/2007 Promotionsstudiengang *Angewandte Statistik und Empirische Methoden*, Georg-August-Universität Göttingen
- 19/01/2004 Diplom in Biomathematik, Note 1,3
Thema der Diplomarbeit: „Methoden zur statistischen Auswertung von Proteomdaten“
- 10/1998 – 02/2004 Studium der Biomathematik, Ernst-Moritz-Arndt Universität Greifswald
- 07/2002 – 12/2002 Auslandssemester, Massey University Palmerston North, Neuseeland

Wehrdienst

- 11/1997–08/1998 Stabsdienstsoldat Panzergrenadierbrigade 41, Eggesin

Schulbildung

- 20/06/1997 Abitur, Note 2,0
- 09/1990 – 07/1997 Friedrich-Ludwig-Jahn-Gymnasium, Greifswald

09/1985 – 07/1990 Polytechnische Oberschule Friedrich Engels, Greifswald

Berufliche Tätigkeiten

seit 11/2007 wissenschaftlicher Mitarbeiter, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck

10/2005 – 10/2007 wissenschaftlicher Mitarbeiter, Abteilung für genetische Epidemiologie, Georg-August-Universität Göttingen

10/2004 – 09/2005 wissenschaftlicher Mitarbeiter, Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie, Charité Berlin

07/2004 – 09/2004 wissenschaftlicher Mitarbeiter, Institut für Psychologie, Humboldt-Universität zu Berlin

02/2004 – 04/2004 Programmierer, Decodon GmbH, Greifswald

Weiterbildungen

06/2007 Sommerschule „Biometrie in regulatorischen Guidelines zur Arzneimittelzulassung – methodische und praktische Aspekte“, St. Andreasberg

08/2006 Sommerschule „Logical reasoning in human genetics“, Helsinki

Publikationsverzeichnis

Zeitschriftenartikel

- [1] Dressel R, Schindehütte J, Kuhlmann T, Elsner L, Novota P, Baier PC, **Schillert A**, Bickeböller H, Herrmann T, Trenkwalder C, Paulus W, Mansouri A: The tumorigenicity of mouse embryonic stem cells and in vitro differentiated neuronal cells is controlled by the recipients' immune response. *PloS One* 3, e2622 (2008)
- [2] Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, Hall AS, Linsel-Nitschke P, Kathiresan S, Wright B, Trégouët D, Cambien F, Bruse P, Aherrahrou Z, Wagner AK, Stark K, Schwartz SM, Salomaa V, Elosua R, Melander O, Voight BF, O'Donnell CJ, Peltonen L, Siscovick DS, Altshuler D, Merlini PA, Peyvandi F, Bernardinelli L, Ardissino D, **Schillert A**, Blankenberg S, Zeller T, Wild P, Schwarz DF, Tiret L, Perret C, Schreiber S, Mokhtari NEE, Schäfer A, März W, Renner W, Bugert P, Klüter H, Schrezenmeier J, Rubin D, Ball SG, Balmforth AJ, Wichmann H, Meitinger T, Fischer M, Meisinger C, Baumert J, Peters A, Ouwehand WH, Deloukas P, Thompson JR, Ziegler A, Samani NJ, Schunkert H: New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet* 41, 280–282 (2009)
- [3] Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AVA, Ketkar S, Hwang S, Johnson AD, Dehghan A, Teumer A, Pare G, Atkinson EJ, Zeller T, Hayward C, Aspelund T, Mitchell BD, Boerwinkle E, Struchalin M, Cavalieri M, Singleton A, Giallauria F, Metter J, de Boer I, Haritunians T, Lumley T, Siscovick D, Psaty BM, Zillikens MC, Oostra BA, Feitosa M, Province M, de Andrade PhD M, Turner ST, **Schillert A**, Ziegler A, Wild PS, Schnabel RB, Wilde S, Muenzel TF, Leak TS,

- Illig T, Klopp N, Meisinger C, Wichmann HE, Koenig W, Zgaga L, Zemunik T, Kolcic I, Minelli C, Hu FB, Åsa Johansson, Igl W, Zaboli G, Wild SH, Wright AF, Campbell H, Ellinghaus D, Schreiber S, Aulchenko YS, Felix JF, Rivadeneira F, Uitterlinden AG, Hofman A, Imboden M, Nitsch D, Brandstätter A, Kollerits B, Kedenko L, Mägi R, Stumvoll M, Kovacs P, Boban M, Campbell S, Endlich K, Völzke H, Kroemer HK, Nauck M, Völker U, Polasek O, Vitart V, Badola S: Multiple new loci associated with kidney function and chronic kidney disease. *Nat Genet* advance online publication (2010)
- [4] Malzahn D, **Schillert A**, Mueller M, Bickeboller H: The longitudinal non-parametric test as a new tool to explore gene-gene and gene-time effects in cohorts. *Genet Epidemiol* accepted (2010)
- [5] Paterson AD, Waggott D, **Schillert A**, Infante-Rivard C, Bull SB, Yoo YJ, Pinnaduwa D: Transmission-ratio distortion in the framingham heart study. *BMC Proc* 3 Suppl 7, S51 (2009)
- [6] **Schillert A**, Schwarz DF, Vens M, Szymczak S, König IR, Ziegler A: ACPA: automated cluster plot analysis of genotype data. *BMC Proc* 3 Suppl 7, S58 (2009)
- [7] Vasan RS, Glazer NL, Felix JF, Lieb W, Wild PS, Felix SB, Watzinger N, Larson MG, Smith NL, Dehghan A, Grosshennig A, **Schillert A**, Teumer A, Schmidt R, Kathiresan S, Lumley T, Aulchenko YS, König IR, Zeller T, Homuth G, Struchalin M, Aragam J, Bis JC, Rivadeneira F, Erdmann J, Schnabel RB, Dörr M, Zweiker R, Lind L, Rodeheffer RJ, Greiser KH, Levy D, Haritunians T, Deckers JW, Stritzke J, Lackner KJ, Völker U, Ingelsson E, Kullo I, Haerting J, O'Donnell CJ, Heckbert SR, Stricker BH, Ziegler A, Reffelmann T, Redfield MM, Werdan K, Mitchell GF, Rice K, Arnett DK, Hofman A, Gottdiener JS, Uitterlinden AG, Meitinger T, Blettner M, Friedrich N, Wang TJ, Psaty BM, van Duijn CM, Wichmann H, Munzel TF, Kroemer HK, Benjamin EJ, Rotter JI, Witteman JC, Schunkert H, Schmidt H, Völzke H, Blankenberg S: Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. *JAMA* 302, 168–178 (2009)

- [8] Vens M, **Schillert A**, König IR, Ziegler A: Look who is calling: a comparison of genotype calling algorithms. *BMC Proc* 3 Suppl 7, S59 (2009)

Kongressbeiträge

- [9] **Schillert A**, Frahm O-J, Zeller T, Schwarz DF, Blankenberg S, Ziegler A: Automated Evaluation of Signal Intensity Plots – Cluster Validity Measures are Great (2009). (Poster) 18th Annual Meeting of the International Genetic Epidemiology Society, Honolulu, HI, USA; *Genet Epidemiol*, 33:752–835
- [10] Rotival M, Zeller T, Wild P, Szymczak S, **Schillert A**, Cambien F, Ziegler A, Tired L, Blankenberg S: Integrating large-scale genetic and monocyte expression data reveals major regulators of biological processes (2009). (Poster) 18th Annual Meeting of the International Genetic Epidemiology Society, Honolulu, HI, USA; *Genet Epidemiol*, 33:752–835
- [11] Zeller T, **Schillert A**, Szymczak S, Wild P, Rotival M, Cambien F, Tired L, Ziegler A, Blankenberg S: Identification of novel genetic variants associated with coronary artery disease by combined analyses of genome-wide variability and gene expression (2009). (Vortrag) 59th Annual Meeting of The American Society of Human Genetics, Honolulu, HI, USA
- [12] Ziegler A, **Schillert A**, Vens M, Zeller T, Blankenberg S: On the quality of genotype calling algorithms (2010). (Vortrag) 56. Biometrisches Kolloquium, Dortmund