In nicht unerheblichem Maße vollzieht sich die Erforschung der genetischen Grundlagen häufiger Erkrankungen (so genannter „Volkskrankheiten") inzwischen hypothesenfrei. Hierzu werden genetische Einzelbasen-Polymorphismen (SNPs) in großer Zahl im Hinblick auf ihre Krankheitsassoziation untersucht. An dieses Vorgehen knüpft sich die Hoffnung, dass durch das Vorliegen eines so genannten „Kopplungsungleichgewichts" eine Korrelation der analysierten SNPs mit bislang unbekannten, für die Krankheit kausalen, genetischen Varianten sichtbar wird. Zeigt sich im Rahmen einer ersten explorativen Fall-Kontroll-Studie ein Assoziationssignal in einer bestimmten chromosomalen Region, so wird dieses Signal in Folgestudien näher charakterisiert (d.h. gegebenenfalls repliziert und näher eingegrenzt).

Explorative genetische Assoziationsstudien werden heute nahezu durchgängig mit Hochdurchsatzverfahren durchgeführt, wobei bis zu einer Millionen SNPs, die über das gesamte humane Genom verteilt sind, gleichzeitig mit Hilfe so genannter „Microarrays" genotypisiert werden. Im Wesentlichen kommen hierbei zwei Techniken zum Einsatz, die von den Marktführern Affymetrix und Illumina (beide USA) kommerziell vertrieben werden. Beide Verfahren beruhen auf dem Prinzip, das Vorhandensein oder die Abwesenheit von Biomolekülen, die auf einem geeigneten Träger gebunden sind, aus der Stärke eines quantitativen, durch energetische Anregung entstandenen Lichtsignals abzuleiten. Die statistische Verteilung der Signalstärke sollte dabei idealerweise mit drei diskreten Zuständen (0, 1 oder 2) korrespondieren, die den drei möglichen Genotypen für den jeweiligen SNP entsprechen.

Die geschilderten, hoch integrierten Technologien bergen allerdings eine Reihe praktischer Probleme. Die Verlässlichkeit der mit Microarrays erzeugten, qualitativen Genotypisierungsdaten ist in Abhängigkeit von der verwendeten Technologie deutlich geringer als 100 Prozent. Mögliche Fehlerquellen ergeben sich auf der gesamten Prozessierungsschiene, beginnend vom primären Untersuchungsgut, d.h. der aus Blut extrahierten DNA, bis hin zur Übertragung der qualitativen Genotypdaten in eine statistische Analyseumgebung. Soll der Einfluss dieser Fehlerquellen möglichst gering gehalten werden, ergibt sich die Notwendigkeit eines umfangreichen Qualitätsmanagements (QM). Ohne entsprechendes QM besteht die Gefahr, dass wissenschaftliche Studien durch die Nutzung mangelhafter und unkorrigierter Hochdurchsatz-Genotypisierungsdaten systematisch verfälscht und ihre Ergebnisse damit unbrauchbar werden. Diese Problematik betraf in Deutschland insbesondere die umfangreichen Hochdurchsatz-Genotypisierungsdaten, die 2007 am Ende des Nationalen Genomforschungsnetzes II (NGFN2) generiert wurden und als Grundlage für umfangreiche Assoziationsstudien durch Wissenschaftler im NGFN2 und in dessen Folgeprogrammen dienen sollten.

Das QM von Hochdurchsatz-Genotypisierungsdaten erfolgte lange Zeit relativ unstrukturiert in den Daten erzeugenden Laboren oder bei den Institutionen, die mit der abschließenden statistischen Analyse dieser Daten befasst waren. Etablierte Standards gab es zur Zeit des NGFN2 nicht, und über die Gütekriterien der Daten entschieden Wissenschaftler meist nach eigenem Gutdünken. Dies hatte zur Folge, dass die damals bereits existierenden Bestände an

Hochdurchsatz-Genotypisierungsdaten und die aus ihnen abgeleiteten wissenschaftlichen Schlussfolgerungen hinsichtlich ihrer Validität schwer zu beurteilen bzw. zu vergleichen waren. Diesem Umstand sollte durch ein Projekt der TMF abgeholfen werden, im Rahmen dessen die in Deutschland mit dem QM von Hochdurchsatz-Genotypisierungsdaten befassten Fachleute gemeinsame Richtlinien und Standards für den Umgang mit solchen Daten entwickeln sollten. Das QM-Projekt der TMF wurde 2008 geplant und anschließend im Rahmen des Programms Förderung von Instrumenten- und Methodenentwicklungen für die patientenorientierte medizinische Forschung (23.08.2007) durch das BMBF finanziert. Laut Ausschreibungstext war ein erklärtes Ziel dieses Programms, durch die „gemeinsame projektübergreifende Bearbeitung methodischer Fragestellungen [...] Doppelarbeit zu vermeiden, Ressourcen zu sparen und einheitliche Qualitätsstandards zu schaffen." Es sollten „Lösungen für dringende methodische Probleme [...] erarbeitet und der wissenschaftlichen Gemeinschaft breit zur Verfügung gestellt werden." Ein möglicher Beitrag hierzu sollte die Entwicklung „forschungsrelevante[r] Instrumente zur Standardisierung, Harmonisierung und sonstiger Steigerung der Nachhaltigkeit, z.B. von Daten- und Probensammlungen" sein.

Angesichts der erheblichen Investitionen, die im Rahmen des NGFN2 in die Generierung von Hochdurchsatz-Genotypisierungsdaten getätigt wurden, war die Sicherung der nachhaltigen und wissenschaftlich fundierten Nutzung dieser Daten für die patientenbasierte Forschung von höchster Priorität. Da für solche Aktivitäten bei der Förderung der eigentlichen Genotypisierung keine Mittel eingeplant waren, mussten die ersten QM-Maßnahmen von den Projektbeteiligten aus „Bordmitteln" ihrer jeweiligen Institutionen bestritten werden. Insofern erwies es sich als glückliche Fügung, dass das QM der damals bereits in großem Umfang vorliegenden Hochdurchsatz-Genotypisierungsdaten im Rahmen der oben genannten Fördermaßnahme systematisiert werden konnte. Die vielen hochrangigen Publikationen, die aus der anschließenden Nutzung der QM-kontrollierten Daten – auch unter Beteiligung der Projektpartner – her-

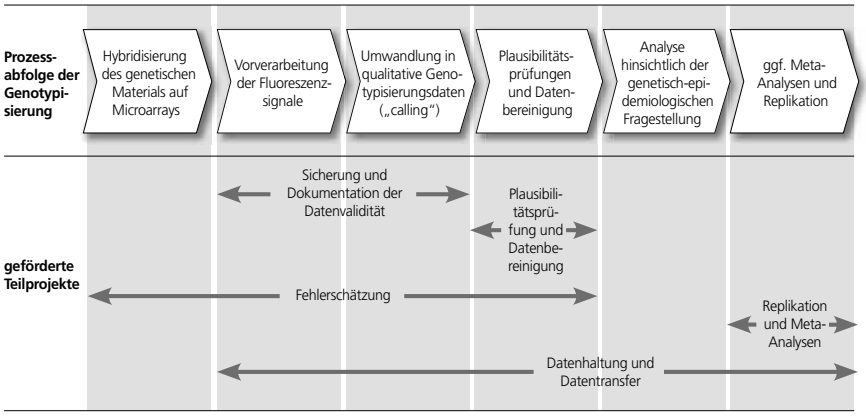

Abbildung 1: Geförderte Teilprojekte und ihr Bezug zu den Prozessschritten der Genotypisierung

vorgegangen sind, belegen den unbestreitbaren Beitrag des Projekts und seiner Ergebnisse zu den förderpolitischen Zielen des Programms.

Das QM-Projekt für Hochdurchsatz-Genotypisierungsdaten wurde zwischen Oktober 2008 und März 2010 durchgeführt. Es war in die unten beschriebenen, inhaltlich abgegrenzten Teilprojekte strukturiert, die wiederum von den beteiligten Institutionen, namentlich den Universitäten Bonn, Kiel und Lübeck, dem MPI für Psychiatrie München, dem DKFZ Heidelberg und dem Helmholtz-Zentrum München, in gemeinsamer Abstimmung standortbezogen bearbeitet wurden.

- **Sicherstellung der Datenvalidität**
  Alle von den Microarrays abgelesenen, primären Fluoreszenzsignale müssen zunächst vorverarbeitet und dann in qualitative Genotypen umgewandelt werden (sogenanntes *allele calling* oder *genotype calling*). Erst die daraus hervorgehenden Sekundär- bzw. Tertiärdaten werden statistisch ausgewertet und bilden damit die Grundlage für die Beantwortung der ursprünglichen genetisch-epidmiologischen Fragestellung. Wegen ihres entscheidenden Einflusses auf die Validität der erzeugten Genotypen, d.h. auf die Korrespondenz zwischen abgeleitetem und tatsächlichem Genotyp, wurden daher verschiedene Verfahren zur Datenvorverarbeitung und zum *genotype calling* auf ihre Vor- und Nachteile hin untersucht und bewertet (Kapitel 2). Ein besonderes Augenmerk lag dabei auch auf dem *genotype calling* aus nicht-kanonischen Intensitätsclustern, d.h. aus Signalmustern, die zunächst keine eindeutige Zuordnung zu den drei möglichen SNP-Genotypen erlauben (Kapitel 3).
- **Plausibilitätsprüfung und Datenbereinigung**
  Die aus den Fluoreszenzintensitäten mittels *genotype calling* gewonnenen qualitativen Genotypisierungsdaten können auch ohne Rückgriff auf die Primärdaten mit Hilfe einer Reihe von Plausibilitätskriterien überprüft werden. Dafür bieten sich mehrere Parameter an, unter anderem die Vereinbarkeit der Genotypverteilung eines SNP mit dem Hardy-Weinberg-Gleichgewicht sowie die Ausfallrate per SNP oder per genotypisiertem Individuum. Fehlerhafte Datensätze („Ausreißer“) können so identifiziert und eliminiert werden, was wiederum die Validität des gesamten Datenbestandes erhöht. Aus diesem Grund wurden die derzeit gängigen Plausibilitätskriterien für Hochdurchsatz-Genotypisierungsdaten auf ihre wissenschaftliche Fundiertheit und ihren praktischen Nutzen hin untersucht (Kapitel 2.6, Kapitel 4).
- **Fehlermodelle und Fehlerschätzung**
  Die Rate an fehlerhaft bestimmten Genotypen ist natürlicherweise eines der zentralen Qualitätsmerkmale von Genotypisierungsdatensätzen. Ihre Schätzung ist daher ein essentieller Bestandteil der Qualitätskontrolle. Im TMF-Projekt wurden unterschiedliche Fehlermodelle anhand empirischer Daten miteinander verglichen. Die dabei geschätzten Fehlerraten wurden außerdem genutzt, um den tatsächlichen Einfluss verschiedener

Qualitätsmaßnahmen im Rahmen der Datenvorverarbeitung zu quantifizieren (Kapitel 5).

- **Replikation bzw. Meta-Analysen von Genotypisierungsdaten**
  Die geringe statistische Power für den Nachweis realistischer Genotyp-Phänotyp-Beziehungen ist ein maßgeblicher Grund dafür, dass genomweite Daten oftmals mit den Ergebnissen zusätzlicher Genotypisierungen entweder im Sinne einer „Replikation" oder im Rahmen von Metaanalysen kombiniert werden. Für dieses Vorgehen wurden im TMF-Projekt Qualitätsmerkmale erarbeitet, um die beträchtlichen Fehlerrisiken bei der Zusammenführung heterogener Datensätze möglichst weitgehend zu minimieren. Unterstützung erhält das QM von Replikationsstudien und Metaanalysen zusätzlich durch eine im TMF-Projekt entwickelte Software, die den Betreibern derartiger Experimente die Möglichkeit einer einheitlichen Qualitätskontrolle aller Genotypen bietet (Kapitel 6).
- **Datenhaltung und Datentransfer**
  Neben der Einhaltung der Datenschutzbestimmungen bilden die Optimierung des Speicherbedarfs und der Zugriffszeiten sowie die Minimierung der Risiken von Fehlübertragungen und Datenverlusten die wesentlichen IT-seitigen Anforderungen an das QM von Hochdurchsatz-Genotypisierungsdaten. Hierfür existierende Konzepte und Tools wurden recherchiert, vergleichend bewertet und auf ihr Optimierungspotenzial hin überprüft (Kapitel 7). Beim QM-kontrollierten Transfer von Genotypisierungsdaten an Dritte stellt sich zudem ein erheblicher Dokumentierungsbedarf, der sich über die Art der Vorverarbeitung und des *genotype callings* hinaus auch auf administrative Hintergrundinformationen erstreckt. Für Art und Umfang dieser Dokumentation wurde im TMF-Projekt ein abgestimmter, generischer Vorschlag erarbeitet und von den Projektteilnehmern konsentiert (Kapitel 3.4).

Der Projektablauf wurde durch die TMF kontrolliert und koordiniert. Diese organisatorische Unterstützung schloss neben der Erstellung und Pflege eines Projektplans auch die Unterstützung der internen Kommunikation, der Außendarstellung und der Qualitätssicherung von Deliverables der anderen Teilprojekte mit ein.

Alle Teilprojekt-Leiter waren parallel in anderen Forschungsprojekten und -verbünden engagiert, die patientenbasierte genetische Forschung auf der Grundlage von Hochdurchsatz-Genotypisierungsdaten betrieben bzw. betreiben. Hierzu zählten neben dem NGFN auch einige der Kompetenznetze in der Medizin sowie Verbünde aus der Exzellenzinitiative von BMBF und DFG. Dadurch konnte sichergestellt werden, dass alle aktuellen technischen und methodischen Entwicklungen auf dem Gebiet der Hochdurchsatz-Genotypisierung unmittelbaren Eingang in die Arbeit im TMF-Projekt finden konnten. Außerdem war dem Projekt ein Advisory Board mit drei internationalen Experten (Françoise Clerget-Darpoux, Paris/Frankreich; Cornelia van Duijn, Rotterdam/Niederlande; David Clayton, Cambridge/Großbritannien) auf dem Gebiet der genetischen Epidemiologie zugeordnet, deren Aufgabe in der Evaluation der

methodischen Qualität, der Bewertung der Validität von Methoden und Ergebnissen des Projekts und in der fortlaufenden Information über konkurrierende Entwicklungen bestand. Wir möchten den drei Kollegen auf diesem Weg ganz herzlich für die freundschaftliche und konstruktive Zusammenarbeit danken!

Das Projekt war fachlich überaus einträglich und – trotz des eng bemessenen Zeitrahmens – ausnehmend kollegial. Dabei hat sich insbesondere die administrative Projektbegleitung durch die TMF als hilfreich erwiesen. Die für das Projekt charakteristische zügige und effiziente Abarbeitung der Aufgaben und die rasche Integration der Erfahrungen und Resultate der Teilprojekte wurden maßgeblich durch die engmaschige Durchführung von Projekttreffen, mehrfach in der Form von Web-Konferenzen, befördert.

Im Rahmen des TMF-Projekts zum QM von Hochdurchsatz-Genotypisierungsdaten ist eine Vielzahl praktisch relevanter Erkenntnisse erzielt worden, die auch zukünftig einen wertvollen Beitrag zur genetisch-epidemiologischen Forschung darstellen dürften. Die detaillierte Darstellung und Verbreitung der Ergebnisse in einer größeren Fach-Community erfolgt nun in Form des vorliegenden Bandes der TMF-Schriftenreihe.

2 Affymetrix Genotypisierungs-Chips: Genotypbestimmung und Qualitätsfilter

Dr. Arne Schillert, Univ.-Prof. Dr. Andreas Ziegler
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck

2.1 Technischer Hintergrund und Motivation

Grundlage genomweiter Assoziationsstudien sind Microarray-Experimente. Im Falle des von der Firma Affymetrix entwickelten Verfahrens wird die jeweils präparationsspezifische DNA-Konzentration der Allele eines diallelischen Einzelnukleotidpolymorphismus (*single nucleotide polymorphism*, SNP) durch eine Hybridisierungsreaktion ermittelt. Jede Zielsequenz, d.h. jedes Allel eines SNPs, korrespondiert mit einem wohldefinierten Satz von kurzen, meist 25 Nukleotide umfassenden DNA-Fragmenten („Sonden", englisch: *probes*), die auf ein geeignetes Trägermedium (den „SNP-Chip") aufgebracht sind. Zur Bestimmung des SNP-Genotyps einer DNA-Präparation, d.h. der relativen Konzentration der beiden Allele, wird zunächst die DNA durch eine PCR (Polymerase-Kettenreaktion, englisch: *polymerase chain reaction*) amplifiziert, dann fragmentiert und mit einem Fluoreszenzfarbstoff markiert. Ist eines der markierten DNA-Fragmente komplementär zu einer der auf dem Chip gebundenen Sonden, so bindet es an diese Sonde. Nach dem Abwaschen der ungebundenen DNA wird mittels Laserlicht der Farbstoff angeregt. Die Intensitäten der daraus resultierenden Fluoreszenzreaktionen werden gemessen und in Graustufenbilder konvertiert. Die nunmehr numerisch vorliegenden Intensitäten werden normalisiert und zur Umwandlung in diskrete Genotypen in SNP-spezifischen *probe sets* zusammengefasst [44, 45].

Für jedes Allel eines SNPs enthält der SNP-Chip verschiedene *perfect match probes* (PM-Sonden). Diese Sonden umfassen jeweils 25 Nukleotide und sind vollständig komplementär zur Zielsequenz. Die verschiedenen Genotypisierungschips der Firma Affymetrix unterscheiden sich allerdings erheblich hinsichtlich ihres technischen Designs, wie im Folgenden kurz skizziert wird. Da die Hybridisierungsreaktion von verschiedenen Faktoren (z.B. Positionseffekten) beeinflusst wird, kommen in der Praxis pro Zielsequenz mehrere, um einige Basenpositionen gegeneinander verschobene Sonden zum Einsatz. Im Falle des Affymetrix Mapping 500K Array Set enthält der zugehörige Chip zudem Probenquartette, die für jedes Allel PM-Sonden und *mismatch probes* (MM-Sonden) umfassen. Die MM-Sonden sind so konstruiert, dass im Vergleich zur PM-Sonde an Position 13 des Oligonukleotides eine homomere Base ausgetauscht ist. Die PM- und MM-Sonden eines Allels sind auf dem Chip benachbart; die PM-Sonden der beiden Allele aber nicht notwendigerweise. Mit den MM-Sonden sollte die Korrektur der Fluoreszenzintensitäten für Hintergrundrauschen ermöglicht werden [26]. Der diesbezügliche Nutzen der MM-Sonden wurde jedoch angezweifelt [41], und weder auf dem Affymetrix Genome-Wide Human SNP Array 5.0 noch auf dem Affymetrix Genome-Wide Human SNP Array 6.0 werden MM-Sonden verwendet. Durch den Verzicht auf die MM-Sonden und durch eine Verringerung der Redundanzen wurde die Anzahl der pro Chip analysierbaren SNPs erheblich erhöht. Beim Affymetrix Genome-Wide Human SNP Array 5.0 werden im Regelfall nur noch acht Sonden, vier je Allel, eingesetzt. Beim Affymetrix Genome-Wide Human SNP Array 6.0 sind es je drei, insgesamt also sechs; die Sonden beider Alllele sind benachbart angeordnet, um lokale Effekte ausgleichen zu können.

Um bei der Genotypisierung eine möglichst hohe Güte zu erreichen, wurde eine Vielzahl von Algorithmen zur Überführung der primären Fluoreszenzsignale in qualitative Genotypen (*genotype calling*) entwickelt. Gegenstand des ersten Teils dieses Kapitels ist eine systematische Betrachtung dieser Algorithmen. Zunächst wird das Ergebnis einer systematischen Literaturrecherche zur Identifikation von Algorithmen zur Genotypbestimmung beschrieben. Anschließend werden die am häufigsten verwendeten Verfahren zur Datenvorverarbeitung skizziert. Danach werden die wesentlichen Genotypisierungsalgorithmen explizit vorgestellt. Den Abschluss des ersten Teils dieses Kapitels bildet ein Vergleich der Algorithmen im Hinblick auf ihre Benutzerfreundlichkeit und die Güte der Genotypbestimmung.

Trotz der Vielzahl der in den vergangenen Jahren entwickelten Algorithmen ist keines der etablierten Verfahren zu 100 Prozent fehlerfrei. Daher ist eine stringente Qualitätskontrolle der quantitativen Genotypdaten erforderlich. Im zweiten Teil des vorliegenden Kapitels werden daher Standardkriterien betrachtet, die bislang für die Qualitätskontrolle von Genotypen empfohlen wurden (für eine englischsprachige Fassung dieses Teils siehe [44])

2.2 Systematische Literaturrecherche

Zur Identifikation von Genotypisierungsalgorithmen wurde im Januar 2009 eine systematische Literaturrecherche in ausgewählten Datenbanken durchgeführt (Tabelle 1).

Tabelle 1: Zieldatenbanken der Literaturrecherche von Genotypisierungsalgorithmen.

Datenbank	URL
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

Basierend auf den Zusammenfassungen von Artikeln über bereits bekannte Genotypisierungsalgorithmen und auf den Vorschlägen der Projektgruppe „Qualitätsmanagement für Hochdurchsatz-Genotypisierungsstudien“ der TMF wurden Schlagworte für die gezielte Suche nach Publikationen zur Affymetrix-Plattform ausgewählt (Tabelle 2).

Tabelle 2: Schlagwortkombinationen für die Literaturrecherche von Genotypisierungsalgorithmen.

Schlagwortkombination
affymetrix AND genotype calling algorithm
affymetrix AND large scale genotyping
affymetrix AND genotyping algorithm
affymetrix AND genotype calling method
affymetrix AND snp calling

Die Ergebnisse der Abfragen wurden zunächst mittels EndNote und Zotero gespeichert, dann in EndNote zusammengeführt. Nach Entfernung der Duplikate verblieben 522 potenziell interessante Publikationen. Unter diesen wurden dann in einem zweistufigen Auswahlverfahren die tatsächlich relevanten Publikationen identifiziert. Zunächst ließen sich 504 Artikel anhand ihrer Zusammenfassungen ausschließen, da sie entweder keinen neuen Algorithmus beschrieben, keine Software anboten oder die beschriebenen Verfahren sich nicht für Affymetrix Arrays eigneten. Im zweiten Schritt wurden jene Artikel ausgeschlossen, deren Algorithmen nicht für das Mapping 500k Array Set geeignet waren; zwei Arbeiten wurden zudem als inhaltlich identisch zu bereits identifizierten Arbeiten erkannt; eine Arbeit war nicht auffindbar. Zwei weitere Publikationen bezogen sich auf Algorithmen für *copy number variations* (CNVs). Insgesamt wurde im zweiten Schritt die Anzahl der Publikationen von 18 auf 9 reduziert (Abbildung 2). Die darin beschriebenen Algorithmen sollten sowohl theoretisch beleuchtet als auch praktisch evaluiert werden.

Für die neun identifizierten Algorithmen wurde versucht, entsprechende Software-Implementierungen zu erhalten und nach Installation in Betrieb zu nehmen. Bei SNiPer-HD [17] schlug die Kompilierung fehl; auf eine Anfrage reagierte der korrespondierende Autor nicht. Der Code für GEL (kurz für: *genotype calling using empirical likelihood*) [31] und MAMS (kurz für: *multi-array multi-SNP genotype calling*) [43] war nur auf Anfrage erhältlich. Auf entsprechende Bitten um Zusendung antworteten die korrespondieren Autoren jedoch nicht. Der Verweis auf die Internetseite von PLASQ (kurz für: *probe-level allele-specific quantification procedure*) [22] war veraltet, so dass auch dieses Programm nicht mit in die Untersuchung einbezogen werden konnte.