

**Schriftenreihe der TMF – Technologie- und Methodenplattform
für die vernetzte medizinische Forschung e. V.**

Band 9



Medizinisch Wissenschaftliche Verlagsgesellschaft

Zur Schriftenreihe der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung

In der TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. haben sich Netzwerke und vernetzt arbeitende Einrichtungen zusammengeschlossen, um gemeinsam die Fragestellungen und Herausforderungen von medizinischer Forschung an verteilten Standorten zu lösen. Durch den Community-Ansatz erfahren die Ergebnisse der TMF eine breite inhaltliche Abstimmung in der medizinischen und medizininformatisch-biometrischen Fachwelt. Mit ihrer Schriftenreihe macht die TMF die Projektergebnisse einer breiteren Leserschaft zugänglich. Zudem bieten Referenzwerke zum Themenfeld der Gesundheits- telematik Orientierungshilfen in der praktischen Umsetzung.

TMF e. V.
Neustädtische Kirchstraße 6
10117 Berlin
www.tmf-ev.de

Bisher in der Schriftenreihe erschienen:

Band 1:

Generische Lösungen zum Datenschutz für die Forschungsnetze in der Medizin

von Carl-Michael Reng | Peter Debold | Christof Specker | Klaus Pommerening

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2006

Band 2:

Biomaterialbanken – Rechtliche Rahmenbedingungen

von Jürgen Simon | Rainer Paslack | Jürgen Robiński | Jürgen W. Goebel | Michael Krawczak

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2006

Band 3:

Checkliste und Leitfaden zur Patienteneinwilligung

Grundlagen und Anleitung für die klinische Forschung

von Urs Harnischmacher | Peter Ihle | Bettina Berger | Jürgen Goebel | Jürgen Scheller

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2006

Band 4:

Datenqualität in der medizinischen Forschung

von Michael Nonnemacher | Dorothea Weiland | Jürgen Stausberg

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2007

Band 5:

Biomaterialbanken – Checkliste zur Qualitätssicherung

von Michael Kiehntopf | Klas Böer

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2008

Band 7:

Verwertungsrechte in der vernetzten medizinischen Forschung

von Jürgen W. Goebel | Jürgen Scheller

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2009

Band 8:

Regulatorische Anforderungen an Medizinprodukte

von Kurt Becker | Sandra Börger | Horst Frankenberger | Dagmar Lühmann | Thomas Norgall

Christian Ohmann | Annika Ranke | Reinhard Vonthein | Andreas Ziegler | Andreas Zimolung

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft, 2011

Schriftenreihe der TMF – Technologie- und Methodenplattform
für die vernetzte medizinische Forschung e. V.

Band 9

M. Krawczak | M. Freudigmann (Hrsg.)

Qualitätsmanagement von Hochdurchsatz- Genotypisierungsdaten

mit Beiträgen von

T. Bettecken | M. Freudigmann | M. Krawczak
T. T. Lu | B. Müller-Myhsok | A. Pfeufer | A. Schillert
M. Steffens | T. F. Wienker | A. Wolf | A. Ziegler



Medizinisch Wissenschaftliche Verlagsgesellschaft

Die Herausgeber

Prof. Dr. Michael Krawczak

Institut f. Med. Informatik u. Statistik
Universitätsklinikum Schleswig-Holstein
Arnold-Heller-Str. 3, Haus 31
24105 Kiel

Mathias Freudigmann

TMF – Technologie- und Methodenplattform
für die vernetzte medizinische Forschung e.V.
Neustädtische Kirchstraße 6
10117 Berlin

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG
Zimmerstr. 11
D- 10969 Berlin
www.mwv-berlin.de

ISBN 978-3-941468-58-0

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Informationen sind im Internet über <http://dnb.d-nb.de> abrufbar.

© MWV Medizinisch Wissenschaftliche Verlagsgesellschaft Berlin, 2011

Dieses Werk ist einschließlich aller seiner Teile urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Verfasser haben große Mühe darauf verwandt, die fachlichen Inhalte auf den Stand der Wissenschaft bei Drucklegung zu bringen. Dennoch sind Irrtümer oder Druckfehler nie auszuschließen. Daher kann der Verlag für Angaben zum diagnostischen oder therapeutischen Vorgehen (zum Beispiel Dosierungsanweisungen oder Applikationsformen) keine Gewähr übernehmen. Derartige Angaben müssen vom Leser im Einzelfall an Hand der Produktinformation der jeweiligen Hersteller und anderer Literaturstellen auf ihre Richtigkeit überprüft werden.

Projekt-/Produktmanagement: Silke Hutt, Berlin
Layout & Satz: eScriptum GmbH & Co. KG – Publishing Services, Berlin
Printed in Germany

Zuschriften und Kritik an:

MWV Medizinisch Wissenschaftliche Verlagsgesellschaft mbH & Co. KG, Zimmerstr. 11, D- 10969 Berlin, lektorat@mwv-berlin.de

mit Beiträgen von

Dr. med. Thomas Bettecken

CAGT – Centrum für Angewandte Genotypisierung
Max-Planck-Institut für Psychatrie
Kraepelinstraße 2-10
80804 München

Mathias Freudigmann

TMF – Technologie- und Methodenplattform
für die vernetzte medizinische Forschung
Neustädtische Kirchstraße 6
10117 Berlin

Prof. Dr. Michael Krawczak

Institut f. Med. Informatik u. Statistik
Universitätsklinikum Schleswig-Holstein
Arnold-Heller-Str. 3, Haus 31
24105 Kiel

Dr. Timothy T. Lu

Institut f. Med. Informatik u. Statistik
Universitätsklinikum Schleswig-Holstein
Arnold-Heller-Str. 3, Haus 31
24105 Kiel

Prof. Dr. Bertram Müller-Myhsok

Max-Planck-Institut für Psychatrie
Kraepelinstraße 2-10
80804 München

Dr. Arne Pfeufer

zum Zeitpunkt der Bucherstellung:
Helmholtz Zentrum München, Deutsches
Forschungszentrum für Gesundheit und Umwelt,
München
jetzt:
European Academy of Bozen/Bolzano
Viale Druso, 1 / Drususallee 1
39100 Bolzano/ Italien

Dr. Arne Schillert

Institut für Medizinische Biometrie und Statistik
Universitätsklinikum Schleswig-Holstein, Campus
Lübeck
Ratzeburger Allee 160
23538 Lübeck

Dr. Michael Steffens

Institut für Medizinische Biometrie, Informatik und
Epidemiologie (IMBIE)
Universität Bonn
Sigmund-Freud-Str. 25
53105 Bonn

Prof. Dr. Thomas F. Wienker

Institut für Medizinische Biometrie, Informatik und
Epidemiologie (IMBIE)
Universität Bonn
Sigmund-Freud-Str. 25
53105 Bonn

Dr. Andreas Wolf

Institut für Medizinische Informatik und Statistik
Universitätsklinikum Schleswig-Holstein, Campus
Kiel
Brunswinker Straße 10
24105 Kiel

Prof. Dr. Andreas Ziegler

Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck
Ratzeburger Allee 160, Haus 4
23538 Lübeck

Editorial der TMF

Seit 2007 widmet sich in der TMF eine interdisziplinäre Arbeitsgruppe ‚Molekulare Medizin‘ den vielfältigen methodischen, infrastrukturellen und ethischen Fragestellungen und Herausforderungen, die sich seit Aufkommen der Hochdurchsatzverfahren in der genomischen Forschung und deren Anwendung in der klinischen Patientenversorgung – auf dem Weg zum Erfüllen des Leistungsversprechens einer „personalisierten Medizin“ – ergeben. Die Aktivitäten in der TMF flankieren hierbei die wissenschaftlichen Projekte und Arbeiten im Nationalen Genomforschungsnetz (NGFN₂), aus dessen Reihen sich ganz maßgeblich die Wissenschaftler und Experten der TMF-Arbeitsgruppe rekrutierten.

Eine der wesentlichen Herausforderungen auf diesem Feld ist die Vergleichbarkeit und Qualitätsbeurteilung von Genotypisierungsdaten aus Hochdurchsatzverfahren, wie sie heute für genetische Assoziationsstudien verwendet werden. Wie die Herausgeber in ihrem einleitenden Kapitel ausführlich darstellen, fehlten bislang Technologie- und Anbieter-übergreifende Qualitätsstandards in der Arbeit mit diesen Daten. Zugleich ist aber die Vergleichbarkeit und Nutzbarkeit erhobener Genotypisierungsdaten gerade für die wissenschaftliche Ergebnissicherung aus großen und kostenintensiven Projekten wie jenen des NGFN₂ essentiell.

Vor diesem Hintergrund fand sich 2008 die Autorengruppe des vorliegenden Bandes, bestehend aus herausragenden Experten ihres Faches, als Projektteam zusammen, um gemeinsam abgestimmte Verfahrensvorschläge zum Qualitätsmanagement zu erarbeiten. Die Autoren haben die verschiedensten Verfahren der Verarbeitung und Analyse von Genotypisierungsdaten systematisch nach Qualitäts Gesichtspunkten geprüft, verglichen und bewertet.

Mit diesem Buch liegen nun breit abgestimmte Empfehlungen zur Qualitätssicherung von Daten vor, die bei der Hochdurchsatz-Genotypisierung generiert werden. Die TMF freut sich, diesen wichtigen Beitrag in ihrer Schriftenreihe der Öffentlichkeit verfügbar machen zu können. Zusätzlich zum vorliegenden Band stellt die TMF ein Software-Tool, welches die Möglichkeit einer einheitlichen Qualitätskontrolle aller Genotypen bietet, und verschiedene Arbeitsmaterialien zur Verfügung, die auf der TMF-Webseite unter www.tmf-ev.de heruntergeladen werden können.

Für diese grundlegenden Arbeiten gilt allen beteiligten Experten und Autoren großer Dank für ihr Engagement und ihre eingebrachte Expertise. Besonderer Dank gebührt Prof. Dr. Michael Krawczak (Institut für Medizinische Informatik und Statistik, Universitätsklinikum Schleswig-Holstein – Campus Kiel) als Initiator und Leiter des Projekts sowie Mathias Freudigmann, der seitens der TMF-Geschäftsstelle die Projektkoordination übernommen hat. Beide haben zudem die Aufgaben als Herausgeber dieses Bandes wahrgenommen.

Das Projekt wäre in dieser Form nicht möglich gewesen ohne die finanzielle Förderung durch das Bundesministerium für Bildung und Forschung (BMBF), das mit seinem für Methodik und Infrastrukturaufbau äußerst wichtigen Förderprogramm „Förderung von Instrumenten- und Methodenentwicklungen

für die patientenorientierte medizinische Forschung“ (2008–2011) die Voraussetzung für den vorliegenden Beitrag schuf.

Für die TMF – Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V.

Sebastian Claudius Semler
(Wissenschaftlicher Geschäftsführer)

Prof. Dr. Ulrich R. Fölsch
(Vorstandsvorsitzender)

Inhaltsverzeichnis

Editorial der TMF _____	vii
1 Das TMF-Projekt zum Qualitätsmanagement von Hochdurchsatz-Genotypisierungsdaten _____	1
2 Affymetrix Genotypisierungs-Chips: Genotypbestimmung und Qualitätsfilter _____	9
2.1 Technischer Hintergrund und Motivation _____	11
2.2 Systematische Literaturrecherche _____	12
2.3 Datenvorverarbeitung _____	14
2.3.1 Quantilnormalisierung _____	15
2.3.2 Robust Multi-Array Averaging (RMA) _____	16
2.3.3 Schätzung des logarithmierten Fehlers der Proben-Intensitäten (PLIER) _____	17
2.4 Beschreibung ausgewählter Genotypisierungs-Algorithmen _____	18
2.4.1 Dynamic Model-Algorithmus _____	18
2.4.2 Robustes lineares Modell mit Mahalanobis Distanz (RLMM) und Bayesianisches robustes lineares Modell mit Mahalanobis Distanz (BRLMM) _____	19
2.4.3 Korrigiertes robustes lineares Modell mit Mahalanobis Distanz (CRLMM) _____	21
2.4.4 CHIAMO _____	22
2.4.5 JAPL _____	23
2.4.6 BIRDSEED _____	23
2.5 Systematischer Vergleich ausgewählter Genotypisierungsalgorithmen _____	24
2.5.1 Vergleichsgrundlage HAPMAP II _____	24
2.5.2 Benutzerfreundlichkeit der Programme zur Genotypisierung _____	25
2.5.3 Ergebnisse der Genotypbestimmung _____	28
2.6 Standard-Qualitätskontrolle _____	30
2.7 Diskussion _____	32
2.7.1 Danksagung _____	36
Literaturverzeichnis _____	36
3 Nicht-kanonische Fluoreszenz-Cluster und Begleitinformation für Genotypisierungsdaten _____	39
3.1 Einführung _____	41
3.2 Motivation _____	42
3.2.1 Nicht-kanonische Fluoreszenzintensitätscluster _____	42
3.2.2 Begleitinformation für Hochdurchsatz-Genotypisierungsdaten _____	45

3.3 Nicht-kanonische Fluoreszenz-Clusterbildung	46
3.3.1 Datenbasis	46
3.3.2 Verteilung der call rates	47
3.3.3 Klassifikation der Clusterbildung	47
3.3.4 Übersicht über die Verteilung der Cluster-Klassen	48
3.3.5 Automatische Erkennung nicht-kanonischer Clusterbildung	50
3.3.6 Bedeutung und Interpretation nicht-kanonischer Fluoreszenzintensitäts-Cluster	50
3.4 Begleitinformation für Genotypisierungsdaten	52
3.5 Zusammenfassung	53
3.6 Anhang: Begleitformular für Genotypisierungsdaten	54
4 Etablierung wissenschaftlich fundierter Plausibilitätskriterien für Genotypisierungsdaten	63
4.1 Einleitung	65
4.2 Üblicherweise verwendete QM-Parameter	65
4.2.1 Minor Allele Frequency (MAF)	66
4.2.2 Per Person Call Rate (ppCR)	66
4.2.3 Per SNP Call Rate (psCR)	66
4.2.4 Hardy-Weinberg Equilibrium (HWE)	67
4.3 Empirische Bewertung von Schwellenwerten	67
4.3.1 Verwendete Stichproben	67
4.3.2 Analyse	67
4.4 Ergebnisse	68
4.5 Diskussion	70
Literaturverzeichnis	72
5 Fehlererkennung und Fehlerkorrektur bei Genotypisierungsdaten	73
5.1 Einleitung	75
5.2 Datengrundlage	75
5.3 Ergebnisse	76
5.3.1 Programmimplementation	76
5.3.2 Abschätzung der Fehlerraten	77
5.3.3 Systematische Untersuchung der high level-Qualitätsparameter	79
5.3.4 Systematische Untersuchung der medium level-Qualitätsparameter	81
5.3.5 Systematische Untersuchung der low level-Qualitätsparameter	82
Literaturverzeichnis	85

6	Daten- und Qualitätsmanagement bei der inhaltlichen Replikation von Genotypisierungsexperimenten	87
6.1	Einleitung	89
6.2	Motivation	89
6.3	Primäre Datentypen und Datenformate	90
6.4	Qualitätskontrolle mit Hilfe von REPLICHECKSNP	93
6.4.1	CHECKPOS	94
6.4.2	CHECKBLAT	95
6.4.3	CHECKHWE	96
6.5	Zusammenfassung	98
	Literaturverzeichnis	98
7	Datenstrukturen, Speicherung und Transfer von Genotypisierungsdaten	99
7.1	Einleitung	101
7.2	Motivation	101
7.3	Primäre Datentypen und Datenformate	102
7.3.1	Vorsätze für Maßeinheiten	103
7.3.2	Primäre Datentypen	104
7.3.3	Datenformate	105
7.4	Datentransfer	108
7.4.1	Datentransport über feste Medien	108
7.4.2	Datentransfer über ein Netzwerk	109
7.4.3	Datenkomprimierung	111
7.5	Datenspeicherung	113
7.5.1	Relationale Datenbanken und binary large objects (BLOBs)	113
7.5.2	Datensicherung und Archivierung	119
7.6	Datenschutz	121
7.6.1	Mehrstufige Berechtigungskonzepte	122
7.6.2	Datentreuhändermodell	124
7.7	Testen von Software und Softwareentwicklung	127
7.7.1	Testen von Programmen und Dateiformaten	127
7.7.2	Programmentwicklung	130
7.8	Zusammenfassung	131
	Literaturverzeichnis	132

Abbildungsverzeichnis

Abbildung 1:	Geförderte Teilprojekte und ihr Bezug zu den Prozessschritten der Genotypisierung _____	4
Abbildung 2:	Schematischer Ablauf der Literaturrecherche _____	14
Abbildung 3:	Quantilnormalisierung _____	16
Abbildung 4:	Effekt einer Veränderung der Konstanten K der CCS-Transformation _____	21
Abbildung 5:	Flussdiagramm des genotype callings mit CRLMM _____	26
Abbildung 6:	Illustration der Umkodierung der calling-Ergebnisse für CRLMM _____	26
Abbildung 7:	Flussdiagramm zur Genotyp-Bestimmung mit JAPL _____	27
Abbildung 8:	Genauigkeits- vs. Fehlend-Diagramme (ADP) _____	28
Abbildung 9:	Genauigkeits- vs. Fehlend-Diagramme (ADP) bei unterschiedlichen MAF _____	29
Abbildung 10:	Genauigkeits- vs. Fehlend-Diagramme (ADP) für (a) homozygote und (b) heterozygote Genotypen _____	29
Abbildung 11:	Typische Anordnung der Fluoreszenz-Intensitäten beider SNP-Allele _____	42
Abbildung 12:	Fluoreszenzintensitäts-Cluster eines X-chromosomalen SNPs bei weiblichen und männlichen Individuen _____	43
Abbildung 13:	Beispiel für Fluoreszenz-Intensitäten eines SNPs, die in mehr als drei Cluster zerfallen _____	44
Abbildung 14:	Beispiel für Fluoreszenz-Intensitäten eines SNPs mit schlechter Trennung der drei kanonischen Cluster _____	44
Abbildung 15:	Häufigkeit von <i>call rates</i> ausgewählter SNPs _____	49
Abbildung 16:	Verteilung der Cluster-Typen ausgewählten SNPs ($n=1.000$) _____	49
Abbildung 17:	Schema des INFINIUM II ASSAYS von ILLUMINA _____	50
Abbildung 18:	Beispiel eines SNPs in Cluster-Klasse 6 _____	51
Abbildung 19:	Genomischer Inflationsfaktor λ ($rls\$lambda$) in Abhängigkeit vom Schwellenwert der MAF ($rls\$minor$) _____	69
Abbildung 20:	Genomischer Inflationsfaktor λ ($rls\$lambda$) in Abhängigkeit vom Schwellenwert für den p -Wert im HWE Test ($rls\$hardy$) _____	69
Abbildung 21:	Genomischer Inflationsfaktor λ ($lambda$) in Abhängigkeit der Schwellenwerte für den p -Wert im HWE-Test ($hardy$) und für die MAF ($minor$) _____	70
Abbildung 22:	Change point-Analyse von λ und $psCR$ _____	71
Abbildung 23:	Fehlerraten im einfachen Fehlermodell, getrennt nach Variationstypen _____	78
Abbildung 24a:	Abhängigkeit der Fehlerrate des einfachen Fehlermodells von unterschiedlichen Qualitätsparametern _____	80
Abbildung 24b:	Abhängigkeit der Fehlerrate des einfachen Fehlermodells von unterschiedlichen Qualitätsparametern _____	81
Abbildung 25:	Abhängigkeit der residualen Fehlerrate (einfaches Fehlermodell) vom <i>silhouette score</i> bei der Qualitätskontrolle des AGWH SNP ARRAY 6.0 _____	82
Abbildung 26:	Abhängigkeit der Fehlerrate (einfaches Fehlermodell) von <i>silhouette score</i> und MAF bei der Qualitätskontrolle des AGWH SNP ARRAY 6.0 _____	82
Abbildung 27:	Relative Fluoreszenzintensitäten vor Normalisierung _____	84
Abbildung 28:	Beispiel einer REPLICHECKSNP Eingabedatei _____	92

Tabellenverzeichnis

Abbildung 29:	Beispiel einer Ausgabedatei des REPLICHECKSNP-Moduls CHECKPOS _____	94
Abbildung 30:	Beispiel einer Ausgabedatei des REPLICHECKSNP-Moduls CHECKBLAT _____	96
Abbildung 31:	Beispiel einer Ausgabedatei des REPLICHECKSNP-Moduls CHECKHWE _____	97
Abbildung 32:	Konverterbedarf mit und ohne Standardformat _____	105
Abbildung 33:	Normalisierte Speicherung von Genotypen in einer relationalen Datenbank _____	114
Abbildung 34:	Binär codierte Speicherung von Genotypen _____	114
Abbildung 35:	Speicherung von Genotypen in Form von BLOBs _____	115
Abbildung 36:	Vergleich der Zugriffszeit _____	116
Abbildung 37:	Anhängigkeit der Zugriffszeit vom Datenvolumen bei de-normalisierter Speicherung (BLOB-Größe) _____	117
Abbildung 38:	Abhängigkeit der upload-Zeit eines BLOBs in Abhängigkeit von der BLOB-Größe _____	118
Abbildung 39:	Datentreuhändermodell für das Einpflegen neuer Daten _____	125
Abbildung 40:	Datentreuhändermodell für die Zusammenführung von Phänotypen und Genotypen _____	126

Tabellenverzeichnis

Tabelle 1:	Zieldatenbanken der Literaturrecherche von Genotypisierungsalgorithmen _____	12
Tabelle 2:	Schlagwortkombinationen für die Literaturrecherche von Genotypisierungsalgorithmen _____	13
Tabelle 3:	Filter für die Standard-Qualitätskontrolle von Hochdurchsatz-Genotypisierungsdaten _____	30
Tabelle 4:	Anzahl der SNPs auf dem ILLUMINA 550K-CHIP mit reduzierter SNP-bezogener <i>call rate</i> _____	47
Tabelle 5:	Klassifizierung der Fluoreszenzintensitätscluster _____	48
Tabelle 6:	Qualitätskontrollexperiment des NGFN II – GWAS ADD-ON-Projekts _____	76
Tabelle 7:	Fehlermodelle zur Bestimmung der Fehlerrate ϵ _____	77
Tabelle 8:	Fehlerraten der untersuchten Genotypisierungschips _____	78
Tabelle 9:	Fehlerraten im vollständigen Fehlermodell für den ILLUMINA HUMANHAP 550K _____	79
Tabelle 10:	Layout des AFFYMETRIX GENOME-WIDE HUMAN SNP ARRAY 6.0 _____	83
Tabelle 11:	Inhalt der REPLICHECKSNP Eingabedatei _____	92
Tabelle 12:	Einheitenvorsätze in SI- und IEC-Notation _____	103
Tabelle 13:	IEC-Nomenklatur _____	104
Tabelle 14:	Zugriffszeiten in Abhängigkeit der BLOB-Größe _____	117