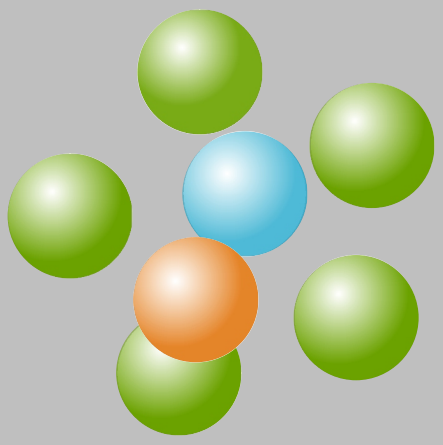

GEM X

Standard Operational Procedures

Datenstrukturen für die Kommunikation genetisch epidemiologischer Projekte im Nationalen Genomforschungsnetz

Kerndaten	SOP 1
erweiterte Kerndaten	SOP 2
Phänotypdaten	SOP 3
Genotypdaten	SOP 4

NGFN
Nationales Genomforschungsnetz

Editoren: Florian Burckhardt (GEM München)
Thomas F. Wienker (GEM Bonn)

Co-Editoren: Andreas Wolf (GEM Kiel)
Tim Strom (GSF München)
Dominik Seelow (MDC Berlin)
Holger Thiele (MDC Berlin)

Diese SOPs werden herausgegeben von:

GEM Plattform (Sprecher: Max P. Baur, Bonn)
Genotypisierungsplattform (Sprecher: Peter Nürnberg, Berlin)

Sie stellen für die beteiligten Einrichtungen des NGFN eine verbindliche Regelung für den Bereich des Datenaustausches dar.

Diese SOPs werden fortgeschrieben und weiterentwickelt. Vorschläge und/oder Korrekturen werden erbeten an den verantwortlichen Editor:

Thomas F. Wienker
Institut für Medizinische Biometrie,
Informatik und Epidemiologie
Universität Bonn
e-mail: wienker@uni-bonn.de
fon: 0228 287 6482
fax: 0228 287 5854

Bonn, den 15. Juni 2003
Version 1.01

Einleitung:

Im NGFN fallen große Datenmengen im Rahmen genetisch epidemiologischer Projekte an. Für diese kooperativen Projekte der Phänotyp-basierten Krankheitsnetze (Kliniker), der Genotypisierungszentren und der Genetisch Epidemiologischen Methodenzentren (GEMs) ist die Kompatibilität dieser Daten für eine qualitätsgesicherte und effiziente Zusammenarbeit von größter Bedeutung. Um in Zukunft einen reibungslosen Transfer und eine anschließende effiziente Verarbeitung dieser Daten zu gewährleisten, wird das im folgenden dargestellte Datenformat für den standardisierten Austausch verbindlich zwischen den GEMs und den Genotypisierungszentren vereinbart. In diesem Zusammenhang ist es wichtig, auf die gestufte Verbindlichkeit der Vereinbarungen hinzuweisen:

- **„muss" oder „darf nicht"** obligate Richtlinie, zwingend notwendig für den reibungslosen Datenaustausch
- **„soll" oder „soll nicht"** empfohlene Richtlinie zur besseren Standardisierung
- **„kann"** optionale Richtlinie

Diese Terminologie folgt den Vorgaben des RCF2119 / BCP14 der internationalen Network Working Group, die die Grundlagen des World Wide Web in kooperativer Weise und auf freiwilliger Basis regelt. Wir möchten Umgangsformen und Kommunikationsstil dieser internationalen Gemeinschaft zur Grundlage dieser SOPs machen.

Datenkomponenten:

Informationen über Probanden (Indexpatient, Kontrollproband, betroffene und unbetroffene Familienmitglieder) genetisch epidemiologischer Studien lassen sich in vier Komponenten gliedern:

1. **Kerndaten zum Projekt, zum Probanden und zur Rekonstruktion der Familie („Linkage"-Format konform)**
2. **erweiterte Kerndaten mit weiteren, studienspezifischen Kovariaten (z.B. Ethnizität, Angaben zu asserviertem Material, Einverständniserklärung inkl. deren Erstreckung)**
3. **Phänotypdaten (Minimalinformation: „affection status", „liability class"), phänotypspezifische Kovariate (z.B. Rauchen und andere Expositionen)**
4. **Genotypdaten**

Die vier Komponenten sollen getrennt gehalten werden und sind über Fremdschlüssel verknüpfbar.

Datenformate:

Für den Datenaustausch zwischen den GEMs und den Genotypisierungszentren werden nur faktisch anonymisierte oder pseudonymisierte Daten verwendet. Das Austauschformat muss textbasiert sein. Der verwendete Zeichensatz muss ISO8859-1 (Latin 1) sein.

Eine **Textdatei** enthält Zeichen mit Codewerten > x001F und ist zeilenorientiert. Die Zeilen entsprechen **Daten-records** unterschiedlicher Länge und werden durch einen **Recordseparator (RS)** getrennt. Hier ist das virtuelle „new line" Zeichen (nl) nach dem POSIX-Standard gefordert (entweder als Doppelzeichen cr = x000D und lf = x000A

z.B. in Windows Umgebungen, oder nur lf = x000A in UNIX oder MAC Umgebungen). Die Datenfelder eines Datensatzes müssen in einer Zeile stehen und müssen durch einen **Feldseparator (FS)** getrennt werden. Als Standard-Feldseparator wird das Semikolon (x003B) vereinbart. **Kerndaten und Genotypdaten** können alternativ als Feldseparator das virtuelle Zeichen „whitespace“ (ws) haben, da hier keine Leerzeichen in den Daten vorkommen. Nach dem POSIX-Standard ist „whitespace“ definiert als ein oder mehrere aufeinander folgende Zeichen „blank“ (BL = x0020 oder „horizontal tab“ (HT = x0009). **Phänotypdaten und erweiterte Kerndaten** sollen das (Semikolon = x003B = „;“); als Feldseparator haben, da es in deren Datenfeldern durchaus sein kann, dass Leerzeichen vorkommen (z.B. Medikamentennamen).

Diese Spezifikationen sind weitverbreitete internationale Standards für den Datenaustausch im Textformat. Darüber hinausgehend werden hier weitere Festlegungen getroffen, indem drei Typen von Records (Zeilen) festgelegt werden.

1. **Datenzeilen**
2. **Kommentarzeilen**
3. **Leerzeilen**

Datenzeilen sind definitionsgemäß alle Zeilen, die nicht Kommentarzeilen oder Leerzeilen sind. **Kommentarzeilen** haben als erstes nicht-whitespace Zeichen einen sogenannten „comment-char“ (cc). Für diese SOP werden als virtuelles cc-Zeichen die Zeichen „asterisk“ = x002A oder „hatchmark/sharp“ = x0023 oder „bang“ = x0021 vereinbart. **Leerzeilen** enthalten entweder kein Zeichen oder ausschließlich „whitespace“-Zeichen.

Direkt aufeinander folgende Kommentarzeilen am Anfang einer Datei werden als **Header** bezeichnet; die erste Nicht-Kommentarzeile beendet den Header, so dass z.B. auf eine Leerzeile oder Datenzeile(n) folgende Kommentarzeilen nicht mehr Teil des Headers sind. Optional können Datenzeilen einen „**trailing comment“** im Stil der Programmiersprache C/C++ haben, der durch einen Doppel-Slash (zweimal aufeinander folgend „slash“ = x002F = “/“) eingeleitet wird und mit dem Zeilenende endet. Ein trailing comment muss sich auf den Inhalt der Datenzeile beziehen. Es wird weiterhin ein spezieller Kommentar für das Dateiende vereinbart: eine Kommentarzeile mit dem Inhalt „EOF“ (drei Buchstaben direkt aufeinanderfolgend e = x0065 / E = x0045 und o = x006F / O = x004F und f = x0066 / F = x0046) soll in der letzten Zeile einer Datei mit Genotypdaten stehen („* EOF“).

Datenaustausch:

Datenaustausch kann über die Medien Diskette, CD-ROM, Email-Attachment im MIME-Format oder FTP erfolgen. Erlaubte Zeichen im **Dateinamen** sind [a...z], [A...Z], [0...9], <.> (Punkt), <-> (Minus), <_> (Unterstrich), die Gesamtlänge darf höchstens 14 Zeichen (POSIX-konform) sein. Als Dateinamenserweiterung sollen im Namen nur **ein** Punkt und maximal 3 Zeichen [a..z], [A..Z], [0..9] zulässig sein. Diese Dateinamenskonvention stellt einen Kompromiss dar zwischen den Erfordernissen unterschiedlicher Betriebssysteme und Medien und vermeidet zuverlässig Probleme im Datenaustausch sowohl zwischen verschiedenen Betriebssystemen wie auch zwischen Shell und Anwendungsprogrammen (sofern diese korrekt programmiert und kompiliert sind).
Die Textdateien zum Datenaustausch können komprimiert werden. Für die Plattform DOS/WINDOWS wird als Kompressionsverfahren „ZIP“ vorgeschlagen, die Dateierweiterung muß dann „.zip“ sein. Für UNIX-Systeme werden TAR und GUNZIP vereinbart; die Dateierweiterungen sind dann „.tar“ bzw. „.gz“, oder, falls **beide** Verfahren angewendet werden „.tgz“..

Lokale Datenspeicherung:

Die Speicherung der Informationen vor Ort wird weiterhin in einem beliebigen Format auf eine beliebige Weise möglich sein (z.B. Datenbankformate). Alle hier beschriebenen Angaben beziehen sich ausschliesslich auf ein gemeinsames und verbindliches Austauschformat.

Kerndaten:

Die Kerndatensätze beschreiben eine Einzelperson innerhalb eines Projektes und erlauben ggf. eine Rekonstruktion des Familienstammbaumes durch „Verzeigerung auf die Eltern“ (sog. Linkage-Format).

Als Trennzeichen für Projekt-, Familien- und Personen-ID sind „Minus“ <-> und „Punkt“ <.> im entsprechenden Datenfeld zugelassen. Sie dürfen nicht am Anfang oder Ende des Datenfeldes stehen. Ein Unterstrich <_> als Trennzeichen soll wg. potentieller SQL-Konflikte vermieden werden.

Feldname/ Eigenschaft (Abkürzung)	Erläuterung	max. Zeichen-zahl	Typ	erlaubte Zeichen	Unterscheidung GROSS/klein	Stärke der Empfehlung
Projekt-ID (PRO)	Bezeichnet das Forschungsprojekt eindeutig	15	varchar	[0...9], [a...z], [A...Z], Trennzeichen	nein, a=A	muß
Familien/ Pedigree ID (PED)	Bezeichnet die Familie innerhalb der Studie eindeutig	15	varchar	[0...9], [a...z], [A...Z] , Trennzeichen	nein, a=A	muß
Personen ID (PID)	Bezeichnet die Person innerhalb der Familie eindeutig	15	varchar	[0...9], [a...z], [A...Z] , Trennzeichen	nein, a=A	muß
Vater-ID (FID)	Ist die Personen-ID des Vaters dieser Person	15	varchar	[0...9], [a...z], [A...Z] 0=kein Vater im Datensatz	nein, a=A	muß
Mutter-ID (MID)	Ist die Personen-ID der Mutter dieser Person	15	varchar	[0...9], [a...z], [A...Z] 0=keine Mutter im Datensatz	nein, a=A	muß
Geschlecht/ Sex		1	Integer	0=unbekannt 1=männlich 2=weiblich		muß
Index Indikator	Kennzeichnung des ersten Familienmit-gliedes, auf welches die Rekrutierer aufmerksam wurden, wichtig für Ascertainment Bias in Assoziationsstudien	1	Integer	0= Proband nicht Erstkontakt innerhalb der Familie 1= Proband Erstkontakt (Index)		kann
Geburtsjahr	Nach ISO8601	yyyy	Integer			soll
Geburtsmonat	Nach ISO8601	mm	Integer			kann
Vitalstatus		1	Integer	0=unbekannt 1=lebt 2=verstorben		soll
DNA vorhanden		1	Integer	0=unbekannt 1=vorhanden 2=nicht vorhanden		soll

Erweiterte Kerndaten:

Der Umfang der erweiterten Kerndaten (Ethnizität, Einverständniserklärung, etc) hängt vom konkreten Projekt ab und soll hier nicht behandelt werden. In der ersten Phase genügt eine Implementierung der Austauschschnittstelle auf reiner ASCII-Basis. Letztlich soll nach einem Erfahrungsaustausch unter allen Beteiligten eine Umsetzung des Austauschformats in XML angestrebt werden.

Phänotypdaten:

Der Umfang der Phänotypinformationen ist Projekt-spezifisch und muss in Absprache mit den Klinikern operationalisiert werden. Der Minimalphänotyp für die Analyse muss den „affection status“ enthalten. Ihm soll eine „liability class“ (s. Terwilliger/Ott (1994), S. 65 ff.) zugeordnet werden.

Feldname/ Eigenschaft (Abkürzung)	Erläuterung	max. Zeichen-zahl	Typ	erlaubte Zeichen	Unterscheidung GROSS/klein	Stärke der Empfehlung
Probanden-ID	Hintereinanderreihung Familien- und Personen-ID als Fremdschlüssel für die Verknüpfung mit Kern- und Genotypdaten	30	varchar	[0...9], [a...z], [A...Z] , Trennzeichen <.>, <->	nein, a=A	muß
„affection status“	Operationalisierter Phänotyp	1	Integer	0=unbekannt 1=unbetroffen 2=betroffen		muß
„liability class“	Sicherungsgrad, Abstufung des „affection status“	1	Integer	0=undefiniert 1...maxliab		soll

Genotypdaten:

Die Genotypinformationen werden aufgeteilt in:

1. Informationen über die verwendeten Marker und Details zur Datenherkunft (Genotypisierungszentrum, Zeitstempel, usw.)
2. Allelische Ausprägungen des Markers für jede Person

Informationen zu 1.) stehen am Kopf der Datei in aufeinanderfolgenden Kommentarzeilen (im Dateiheader); vereinbarungsgemäß werden diese Kommentarzeilen eingeleitet mit <*>, <#> oder <!>.
Informationen zu 2.) sind in jeweils einer Datenzeile für jede Person aufgeführt.

Zu 1.) **Header**

- Projekt-ID
- Name des Markers (Mikrosatelliten: D-numbers, CHLC, GATA, oder andere Bezeichner, s. GDB oder /HUGO/HGN) (SNPs: rs-number, ss-number, oder andere). Für den Fall, dass ein Marker noch keinen offiziellen Namen hat, kann der offizielle Name zu einem späteren Zeitpunkt vom Verantwortlichen eingeholt werden.
- Markertyp: „STR“ oder „SNP“; bei SPNs wird mittels IUPAC-Code die Ausprägung der Allelität angegeben.
- Genotypisierungszentrum (Berlin, GMC@MDC = „B“, München GAC@GSF = „M“ und Kiel, Med. Univ. Klinik = „K“)
- Verwendete Primer bzw. interne Testversion des Genotypisierungszentrums, z.B. für Berlin „.Bxx“, wobei „xx“ eine zweistellige fortlaufende Integerzahl ist. Diese Testversion soll mit einem geeigneten Separator (Unterstrich = x005F = „_“) an den Markernamen angehängt werden. Eine Angabe wie „D12S329_B05“ wird empfiohlen.
- Name des Verantwortlichen für das Projekt auf Seiten des Genotypisierungszentrums (Signatur) und dessen Email-Adresse, ggf. auch Telefon und FAX.
- Datumsstempel im Format nach ISO8601 (yyyy-mm-dd) und Signatur des Verantwortlichen, der das Ergebnis gegenüber den GEMs und den PIs des Projektes vertritt.
- Änderungen an den Datensätzen müssen mit Datum und Personensignatur versehen kommentiert werden

Zu 2.) **Datenrecords**

- Es gibt **eine** Datei pro Genlocus pro Projekt und pro Marker, d.h. eine Untersuchung von 50 SNPs bei 300 Probanden resultiert in 50 Dateien mit je 300 Einträgen.
- Wird derselbe Locus in einem anderen Projekt bei demselben oder anderen Probanden wiederholt untersucht, resultiert das in separaten Dateien. Die Daten werden nicht zusammengeführt, auch wenn es sich um denselben Locus handelt. Wird derselbe Locus mit einem neuen Primerpaar (geänderte Testversion) untersucht, resultiert das ebenfalls in unterschiedlichen Dateien.
- Wie zu Beginn erwähnt, kann die „lokale Datenbank“ natürlich für jeden Probanden einen Multilocus-Genotyp speichern. Z.B. können für Herrn Meier (nach faktischer Anonymisierung) in einem dieser Person zugeordneten Datensatz alle 10 untersuchten SNPs gespeichert werden. Für den Datenaustausch jedoch gilt:

  EINE Datei pro Projekt und pro Marker

  in der dann alle Probanden zeilenweise aufgeführt sind.
- Im Falle von X-chromosomalen oder Y-chromosomalen Markern wird der hemizygote Genotyp bei Männern so wie ein homozygoter Genotyp kodiert. Pseudoautosomale Genotypen werden wie autosomale Genotypen behandelt; die Identifizierung erfolgt implizit über den Markernamen (z.B. im Dateiheader). Mitochondriale Marker sind außerhalb der vorliegenden Spezifikation. Somatische Mosaike sind ggf. über Varianten der Probanden-ID zu kennzeichnen.

Genetische Daten sollen folgende Mindestinformationen enthalten:

Feldname/ Eigenschaft (Abkürzung)	Erläuterung	max. Zeichen-zahl	Typ	erlaubte Zeichen	Unterscheidung GROSS/klein	Stärke der Empfehlung
Probanden-ID	Hintereinanderreihung Familien- und Personen-ID als Fremdschlüssel für die Verknüpfung mit Kern- und Genotypdaten	30	varchar	[0...9], [a...z], [A...Z] , Trennzeichen <.>, <->	nein, a=A	muß
Mikrosatelliten						
Allel 1	Länge des Repeats in Basenpaaren	3	**Integer**	[0...9]		muß
Allel 2	Länge des Repeats in Basenpaaren	3	**Integer**	[0...9]		muß
SNPs						
Allel 1	Einzelbase	1	**char**	[G,A,T,C]	nein, a=A	muß
Allel 2	Einzelbase	1	**char**	[G,A,T,C]	nein, a=A	muß
Qualitätsindikator						
Qualität	Sicherheit des Ergebnisses, entsprechend der verwendeten Methoden (z.B. call rate bei MALDI-TOF)	1	Integer	[0..9] 0 = not defined, 1 = schlechteste Qualität 9 = beste Qualität		soll

Bearbeitungsstand:

Version: 1.01, Revisionsdatum: 2003-06-15

Diese SOPs werden laufend überarbeitet und ggf. den sich wandelnden Bedürfnissen und Erfahrungen angepasst. Die Herausgabe einer überarbeiteten und damit als verbindlich erklärten Version/Revision erfolgt über die Sprecher der GEM-Zentren und der Genotypisierungsplattform.

Die redaktionelle Sichtung und Abstimmung der Vorschläge zu Änderungen und Ergänzungen erfolgt durch die Editoren und Ko-Editoren dieses Dokuments im Sinne der RFC2119/BCP14.