LABIMI/F

Deliverable

Alexander Herrmann, Jochen Hampe

Use Case – Genomdaten[1]

	Use Case – Genomdaten
Autor(en)	Alexander Herrmann, Jochen Hampe
Editor(en)	
Datum	29.06.2012
Version des Dokuments	1.0.3

A: Status des Dokuments

Version 1.0.3

[1]Dieses Dokument wurde im Rahmen des Projekts LABIMI/F erstellt. Das Projekt LABIMI/F wird gefördert von der Deutschen Forschungsgemeinschaft (DFG) unter dem Förderkennzeichen RI1000/2-1.

B: Bezug zum Projektplan

Deliverable D2.2: Dokumentierter Use Case, Version 2

C: Abstract

Als konkretes Anwendungsbeispiel für die Rahmenbedingungen der Umsetzung im Projekt „Langzeitarchivierung biomedizinischer Forschungsdaten“ werden hier die für Genomdaten die relevanten Metadatenstandards zusammengefasst. Dazu wurde eine Literaturrecherche sowie eine Nutzungserfassung der aktuell im Rahmen von öffentlichen Datenbanken und Großprojekten genutzten Datenstandards durchgeführt. Es wird dabei deutlich, dass bisherige Metadatenstandards insbesondere auf die Beschreibung der Sequenz selbst und relevanter Motive abzielen. Insbesondere im Bereich der Erfassung von technologieabhängigen Fehlersignaturen und Qualitätsdaten auf der einen Seite und des biologischen Kontext von Sequenzdaten bestehen jedoch noch relevante Lücken.

D: Änderungen

Version	Datum	Name	Kurzbeschreibung
1.0.1	29.06.2012	A. Herrmann	Erste Dokumentversion
1.0.2	27.11.2012	J. Hampe	Erste Überarbeitung
1.0.3	20.02.2013	A.Herrmann	Zweite Überarbeitung
1.0.4	04.03.2013	J. Hampe	Dritte Überarbeitung

E: Inhaltsverzeichnis

1 Einleitung

Biomedizinische Forschung hat in den letzten Jahren eine neue Qualität, insbesondere in Hinblick auf die Menge und die Komplexität der anfallenden Daten gewonnen. Ein wesentlicher Meilenstein ist hier die erste Sequenzierung des menschlichen Genoms im Jahre 2001 (1). Für die Forschungsrealität einer Arbeitsgruppe der krankheits- und patientenorientierten Genomforschung stellen sich hier mehrere Herausforderungen:

- Die vollständige Genomsequenzierung bildet unter anderem die Grundlage für die Möglichkeit, Erbkrankheiten zu erforschen, molekulare Mechanismen der Krebsentstehung besser zu verstehen und Therapien zu individualisieren. Durch neuen Hochdurchsatz-Genotypisierungs- und Sequenzierungstechnologien, stehen heute die technischen Methoden zur Verfügung, komplette menschlichen Genome in einigen Wochen zu sequenzieren und die wesentliche genetische Variabilität eines Individuums in wenigen Stunden zu erfassen.
- Inhaltorientierte Arbeitsgruppen nutzen üblicherweise heterogene Dienstleister für die Datengenerierung (Sequenzier- und Genotypisierungszentren akademischer oder kommerzieller Art). Mit der Übergabe der Daten an den Auftraggeber ist für diese plattformorientierten „Provider“ der Auftrag abgeschlossen – eine langfristige Archivierung ist dort in der Regel aus praktischen und finanziellen Gründen und auch wegen des fehlenden biologisch/medizinischen Kontextes nicht vorgesehen.

Für die inhaltsorientierten Forschungsgruppen stellen sich damit ganz neue Aufgaben: Die Sequenz- und Genomdaten sollten allein schon

- für die Einhaltung guter wissenschaftlicher Praxis 10 Jahre,
- im medizinischen Bereich bis zu 30 Jahre archiviert

werden. Einerseits sollten die Daten sicher gespeichert werden, um den Archivierungspflichten nachzukommen, andererseits stellen sie auch eine wichtige Interaktionsbasis dar, um evtl. später bei neueren größeren genomischen Metaanalysen verwendet zu werden. Archivierung heißt hier also auch technisch effiziente und mit entsprechenden Rechten fein granulierbare Zugriffsmöglichkeiten zu schaffen. Die personelle und technische Infrastruktur inhaltsorientierter Genomforscher ist auf diese Herausforderungen bisher praktisch nicht eingestellt.

Im Rahmen dieses DFG-Vorhabens soll eine modellhafte Implementierung im Sinne einer Machbarkeitsstudie für ein Langzeitarchivierungssystem für komplexe, anwendungsorientierte Genom- und Sequenzdaten entwickelt werden. Hier werden im Folgenden zunächst die relevanten Metadatenstandards evaluiert.

2 Material und Methoden

Für dieses Deliverable wurden, basierend auf eine Literaturrecherche in Pubmed (http://www.ncbi.nlm.nih.gov/pubmed), der englischsprachigen Wikipedia, den Webressourcen der Sequenzierzentren und auf einer Nutzungsanalyse der am meisten genutzten Genomdatenbanken wie der NCBI Genbank, (http://www.ncbi.nlm.nih.gov/genbank), UCSG Genombrowser (http://genome.ucsc.edu/) und der EMBL Nukleotiddatenbank analysiert (http://www.ebi.ac.uk/embl/), die gebräuchlichsten Sequenzdatenstandards analysiert. Dabei wurde besonderes Augenmerk auf die biologische und technische Metadatenabbildung gelegt. Die Metadaten müssen den biologischen Kontext, das technische Format und generische Klassen für die Gestaltung der Zusammenarbeit enthalten. Entsprechend abstrakt gefasste Beschreibungen der technischen Formate sind dann auch die Voraussetzung für die Entwicklung von automatischen Konvertierungsroutinen, die ein Zusammenführen der Daten erlauben.

2.1 Wichtige Formate der Sequenzierungsdaten

2.1.1 FASTA Format

Dieses ist eines der am längsten etablierten Sequenzdatenformate. Das originale FASTA/Pearson Datenformat wurde in der Dokumentation des FASTA-Programmpakets beschrieben (2). Es spiegelt die frühe Fixierung der Datenstandards auf die Sequenz selbst wider, da historisch die Generierung der Sequenz selbst mit dem entscheidenden Aufwand verbunden war. Es ist auch heute noch das verbreitetste Datenformat für Sequenzdaten, selbst im Rahmen der Ausgabeformate von Hochdurchsatzsequenzierern. Das Format unterstützt Metadaten nur rudimentär und wenig strukturiert. Das Format enthält eine einzelne Kopfzeile die den Namen der Sequenz, eine optionale Beschreibung (d.h. Metadaten) in unstrukturierter Form umfasst. Alle weiteren Zeilen dieses textbasierten Formats beinhalten dann die Sequenz selbst. Die Sequenz selbst wird mit einem Größerzeichen („>“) eingeleitet. Die Sequenz ist typischerweise auf 60 Zeichen pro Zeile formatiert. Abhängig von der Anwendung werden Leerzeilen entweder als Ende der Sequenz interpretiert oder auch ignoriert. Ebenfalls anwendungsabhängig werden Leerzeichen oder andere Sequenzsymbole ignoriert oder als Lücken in der Sequenz interpretiert. Die Sequenz selbst wird durch IUB/IUPAC Nukleinsäurencodes als Buchstaben kodiert.

FASTA-Dateinen können multiple Sequenzen enthalten, die häufig aus einem gemeinsamen biologischen oder experimentellen Kontext entstammen. FASTA-Formate werden von einer Vielzahl von Sequenzanalyseprogrammen und Aligmentwerkzeugen aktzeptiert, insbesondere auch BLAST, BLAT und CLUSTAL.

Der folgende Eintrag gibt ein Beispiel eines FASTA-Sequenzeintrages für Proteinsequenz:

```
>FOSB_MOUSE Protein fosB. 338 bp
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
```

2.1.2 FASTQ-Format

Das FASTQ Format ist ein text-basiertes Format zur Speicherung sowohl von Nukleotidsequenzen als auch der korrespondierenden Qualitätsscores (3). Dabei wird sowohl das Nukleotid als auch der Qualtiätsscore in einem einzelnen ASCII Zeichen untergebracht. Das Format wurde ursprünglich am Sanger Center (UK) zur Bünderlung einer FASTA-Sequenz und Ihrer Qualitätsdaten entwickelt. Das Format hat sich in jüngerer Zeit zu einem de-facto Standard zur Speicherung der Ausgabe von Hochdurchsatzsequenzierern der zweiten Generation wie dem Illumina Genome Analyzer entwickelt.

Das FASTQ Datenformat nutzt in der Regel vier Zeilen pro Sequenz. Zeile eins beginnt dabei mit einem '@' Zeichen und wird von einem Sequenzidentifzierungscode und einer optionalen Beschreibung wie im FASTA Format gefolgt. Zeile zwei enthält die eigentliche, alphanumerisch kodierte Sequenz. Zeile drei beginnt mit einem ”+” Zeichen und optional dem gleichen Sequenzidentifizierungscode. Zeile vier kodiert dann für die Qualitätsscores für alle Sequenzzeichen aus Zeile 2. Eine minimale FASTQ Datei könnte etwa wie folgt aussuchen:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

+

!''(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65*

Die ursprünglichen FASTQ Dateien erlaubten auch einen Zeilenumbruch der Sequenz- und Qualitätszeichenketten – für die Robustheit des Parsens wird dies aber aktuell nicht favorisiert, da sowohl "@" als auch "+" in den Qualitätsscores vorkommen können.

In das FASTQ Datenformat wird eine Reihe von herstellerspezifischen Metadaten integriert:

Illumina sequence identifiers

Illumina ist aktuell der dominierende Hersteller von Zweitgenerationssequenzierern, ca. 70%-90% der neuen Sequenzen werden über Illumina-Geräte erzeugt. Der Kode ist beispielsweise wie folgt aufgebaut:

@HWUSI-EAS100R:6:73:941:1973#0/1

Dabei entsprechen die Bestandteile des Kodes folgenden Informationen:

HWUSI-EAS100R	Eindeutige Identifizierung des Sequenziers
6	Flowcell Spur
73	Tile innerhalb der Flowcell Spur
941	'x'-Koordinate des Clusters innerhalb der Tile
1973	'y'-Koordinate des Clusters innerhalb der Tile
#0	Indexnummer für Multiplexproben (0: kein Multiplex)
/1	Paarzuordnung für Läufe mit gepaarten Enden

Mit dem Release von Casava 1.8 hat sich das Format der "@" Zeile wie folgt geändert: Dies wird wiederum am Beispiel erläutert:

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

EAS139	Eindeutige Identifizierung des Sequenziers
136	Identifizerungsnummer des Laufes auf dem Gerät
FC706VJ	Identifizerungsnummer der Flowcell
2	Flowcell Spur
2104	Tile innerhalb der Flowcell Spur
15343	'x'-Koordinate des Clusters innerhalb der Tile
197393	'y'-Koordinate des Clusters innerhalb der Tile

1	Paarzuordnung für Läufe mit gepaarten Enden
Y	Y falls der Sequenzread gefiltert ist, sonst N
18	0 wenn keine Kontrollbits eingeschaltet sind, sonst eine gerade Zahl
ATCACG	Indexsequenz

NCBI Sequence Read Archive

Bei FASTQ Dateien des NCBI/EBI Sequenzarchivs wird häufig eine erweiterte Metadatenannotation vorgenommen wie das folgende Beispiel zeigt:

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Hier wird ein zusätzlicher eindeutiger Identifierungskode durch das NCBI erzeugt. In der Beschreibung bleibt die Herstellerinformation erhalten.

2.1.3 Alignmentformate

Die neben den Rohdatenformaten sind insbesondere aggregierte Sequenzdatenformate von hoher Bedeutung, weil diese Variabilität innerhalb von Populationen und den Sequenzzusammenhang über größere genomische Distanzen darstellen.

SAM Format

Das SAM (Sequence Alignment/Map) ist ein text-basiertes, Tabulator-getrenntes Format für die Darstellung von Sequenzalignments (4). Die Kopfzeilen beginnen mit dem „@“ Symbol. Jede Zeile des Aligmentfiles bestehen aus – aufgrund der technischen Termini ist die Beschreibung in Englisch gehalten:

Spalte	Feld	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)

6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Jedes Bit in dem FLAG-Feld ist wie folgt definiert:

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

Ein Beispiel ist hier gegeben:

*091123:8:1:840:1201#0 0 chr1 1911 0 86M * 0 0*
GCAAGCTGAGCACTGGAGTGGAGTTTCCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGC
CATTGTTCCTCTTCTGTCCCCTG
A>?BABB?>>@?;?;,1=35/446?>A:5=@=>===(66(4:5=:??5=4:6>4?9686345;
35@6884/11466767(8(.785 RG:Z:dd0 NM:i:3 NH:i:7
CC:Z:chr15 CP:i:100336560 HI:i:0

2.2 Archivierungssystem ENA

Auf europäischer Ebene gibt es Bemühungen für ein einheitliches Archivierungsformat im „European Nucleotide Archive“ (http://www.ebi.ac.uk/ena).

Das „European Nucleotide Archive (ENA)” erfasst und repräsentiert Informationen bezüglich Informationen über experimentelle Workflows im Bereich der Nukleinsäurensequenzierung. Ein typischer Workflow beinhaltet die Isolierung und Herstellung von Material für die Sequenzierung, ein Lauf auf einem Sequenzierer, die Erzeugung der eigentlichen Sequenzdaten und eine nachfolgende bioinformatische Analyse-Pipeline. ENA erfasst diese Informationen in einem Datenmodell, das die Eingangsinformationen (Probe, Versuchsaufbau, Maschinen-Konfiguration), Ergebnisse wie (Maschinendaten, Sequenzspuren, Qualitäts-Scores) und abgeleitete Informationen (Assembly, Mapping, und funktionelle Annotation) abdeckt.

Daten erreichen die ENA aus einer Vielzahl von Quellen. Dazu gehören Einreichungen von Rohdaten, Sequenzalignments und Metadaten. Es gibt Bemühungen eine strukturierte Einspeisung von Daten aus den wichtigsten europäischen Sequenzierzentren zu erreichen und einen umfassenden Austausch mit den Partnern in der Internationalen Nucleotide Sequence Database Collaboration (INSDC) zu etablieren.

2.2.1 Sequence Read Archive (SRA)

Das European Nucleotide Archive (ENA) kann Sequenzen aus Next Generation Sequencing Projekten und Technologien wie Roche 454, Illumina Genome Analyzer und ABI SOLiD in sein Sequenzarchiv aufnehmen (Sequence Read Archive -SRA). Hier werden nur die öffentlichen Sequenzierungsdaten gespeichert. ENA arbeitet eng mit dem NCBI und der DDBJ zusammen und tauscht die Daten auf eine tägliche Basis aus.

Alle Informationen über Sequenzdaten für Archivierung werden über ein XML-Metadatenformat abgedeckt. Eine typische SRA XML Spezifikationen (siehe Abb. 1) enthält fünf XML Dateien: Submission, Study, Sample, Experiment und Run XML. Die Sequenzdaten selber werden extra übertragen.

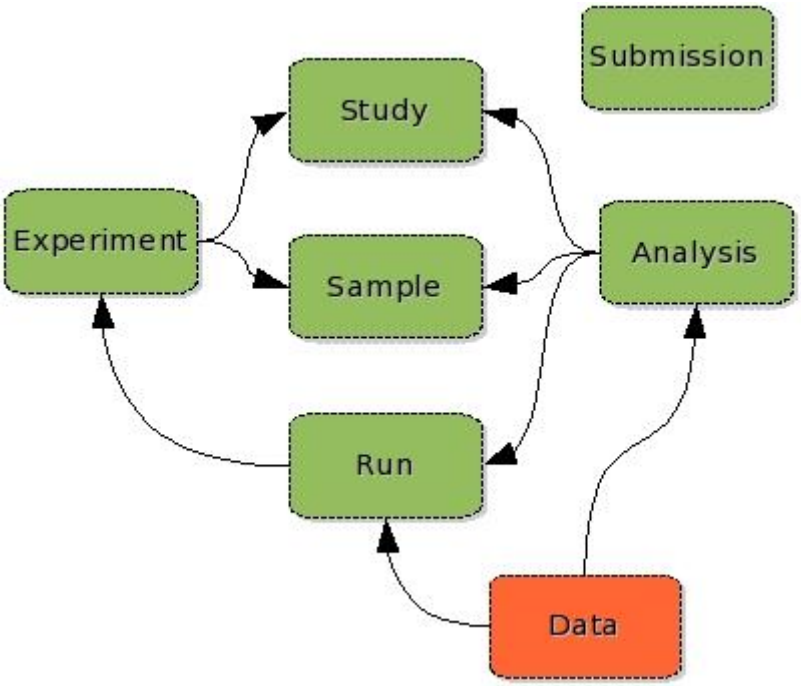

Abbildung 1: SRA Spezifikation eines Sequenzierungsprojektes

2.2.2 Beispiel der XML Spezifikation für ENA

Unten sind Beispiel XML-Dateien für die Übertragung der FASTQ und BAM Files an das ENA System dargestellt. Die Daten stammen aus einem Transkriptom-Sequenzierungsexperiment.

Submission.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<SUBMISSION_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_3/SRA
.submission.xsd">
<SUBMISSION alias="tai2010" center_name="ikmb_kiel">
        <ACTIONS>
            <ACTION>
                <ADD source="tai2010_study.xml" schema="study"/>
            </ACTION>
            <ACTION>
                <ADD source="tai2010_sample.xml" schema="sample"/>
            </ACTION>
            <ACTION>
```

```
            <ADD source="tai2010_experiment.xml" schema="experiment"/>
        </ACTION>
        <ACTION>
            <ADD source="tai2010_run.xml" schema="run"/>
        </ACTION>
        <ACTION>
            <RELEASE/>
        </ACTION>
    </ACTIONS>
    <FILES>
        <FILE filename="tai2010_study.xml" checksum_method="MD5"
checksum="a2d30c4c15d3f655b168a5fe6bbfbf29"/>
        <FILE filename="tai2010_sample.xml" checksum_method="MD5"
checksum="e72031ee1855e055643fb532625605c7"/>
        <FILE filename="tai2010_experiment.xml" checksum_method="MD5"
checksum="3641b3b2fed81339d9d8bcdffd4eb41c"/>
        <FILE filename="tai2010_run.xml" checksum_method="MD5"
checksum="2f5241bac46dcd0b492889bef972f605"/>
        <FILE filename="GM10847_GCK9J_s76.bam" checksum_method="MD5"
checksum="1c796e4cf2720ce35b14f0c259a943bf"/>
        <FILE filename="GM10847_s76.fq.gz" checksum_method="MD5"
checksum="126ec08e0505776f7eb753adf9f599f3"/>
    </FILES>
</SUBMISSION>
</SUBMISSION_SET>
```

Study.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<STUDY_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:noNamespaceSchemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_3/SRA
.study.xsd">
    <STUDY alias="tai2010_study" center_name="ikmb_kiel">
        <DESCRIPTOR>
            <STUDY_TITLE>Statistical Inference of Allelic Imbalance from
Transcriptome Data </STUDY_TITLE>
            <STUDY_TYPE existing_study_type="Transcriptome Analysis"/>
            <STUDY_ABSTRACT>
```

Next-generation sequencing and the availability of high-density genotyping arrays have facilitated an analysis of somatic and meiotic mutations at unprecedented level, but drawing sensible conclusions about the functional relevance of the detected variants still remains a formidable challenge. In this context, the study of allelic imbalance in intermediate RNA phenotypes may prove a useful means to elucidate the likely effects of DNA variants of unknown significance. We developed a statistical framework for the assessment of allelic imbalance in next-generation transcriptome sequencing (RNA-seq) data that requires neither an expression reference nor the underlying nuclear genotype(s), and that allows for allele miscalls. Using extensive simulation as well as publicly available whole-transcriptome data from European-descent individuals in HapMap, we explored the power of our approach in terms of both genotype inference and allelic imbalance assessment under a wide range of practically relevant scenarios. In so doing, we verified a superior performance of our methodology, particularly at low sequencing coverage, compared to the more simplistic approach of completely ignoring allele miscalls. Because the proposed framework can be used to assess somatic mutations and allelic imbalance in one and the same set of RNA-seq data, it will be particularly useful for the analysis of somatic genetic variation in cancer studies.

```
            </STUDY_ABSTRACT>
        </DESCRIPTOR>
        <STUDY_ATTRIBUTES>
            <STUDY_ATTRIBUTE>
                <TAG>Publication</TAG>
                <VALUE>Human Mutation, 2010</VALUE>
            </STUDY_ATTRIBUTE>
        </STUDY_ATTRIBUTES>
    </STUDY>
</STUDY_SET>
```

Sample.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<SAMPLE_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:noNamespaceSchemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_3/SRA
.sample.xsd">
    <SAMPLE alias="GM10847" center_name="ikmb_kiel">
        <TITLE>coriell cell line GM10847</TITLE>
        <SAMPLE_NAME>
```

```
            <TAXON_ID>9606</TAXON_ID>
            <SCIENTIFIC_NAME>homo sapiens</SCIENTIFIC_NAME>
            <COMMON_NAME>human</COMMON_NAME>
        </SAMPLE_NAME>
        <DESCRIPTION>Homo sapiens lymphoblastoid cell lines (30 CEPH trios
and 30 Yoruban trios) were purchased from Coriell Institute for Medical
Reseach (Camden, NJ) </DESCRIPTION>
        <SAMPLE_ATTRIBUTES>
            <SAMPLE_ATTRIBUTE>
                <TAG>Sample type</TAG>
                <VALUE>RNA</VALUE>
            </SAMPLE_ATTRIBUTE>
            <SAMPLE_ATTRIBUTE>
                <TAG>Gender</TAG>
                <VALUE>Female</VALUE>
                <UNITS>Associated family: 1334-2</UNITS>
                <UNITS>Family relationship: mother</UNITS>
            </SAMPLE_ATTRIBUTE>
        </SAMPLE_ATTRIBUTES>
    </SAMPLE>
</SAMPLE_SET>
```

Experiment.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<EXPERIMENT_SET xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:noNamespaceSchemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_3/SRA
.experiment.xsd">
    <EXPERIMENT alias="tai2010_GM10847_SE76" center_name="ikmb_kiel">
        <TITLE>tai2010 GM10847 RNA sequencing</TITLE>
        <STUDY_REF refname="tai2010_study"/>
        <DESIGN>
            <DESIGN_DESCRIPTION>RNA was extracted from 10^8 cells using the
RNEasy kit (Qiagen). The mRNA-Seq libraries for Illumina/Solexa GAII
sequencing were prepared according to the manufacturer’s instructions,
starting with 5 mg RNA. </DESIGN_DESCRIPTION>
            <SAMPLE_DESCRIPTOR refname="GM10847"/>
            <LIBRARY_DESCRIPTOR>
```

```
                <LIBRARY_NAME>GM10847_SE76</LIBRARY_NAME>
                <LIBRARY_STRATEGY>RNA-SEQ</LIBRARY_STRATEGY>
                <LIBRARY_SOURCE>TRANSCRIPTOMIC</LIBRARY_SOURCE>
                <LIBRARY_SELECTION>RT-PCR</LIBRARY_SELECTION>
                <LIBRARY_LAYOUT>
                    <SINGLE/>
                </LIBRARY_LAYOUT>
                <LIBRARY_CONSTRUCTION_PROTOCOL>RNEasy kit
(Qiagen)</LIBRARY_CONSTRUCTION_PROTOCOL>
            </LIBRARY_DESCRIPTOR>
            <SPOT_DESCRIPTOR>
                <SPOT_DECODE_SPEC>
                    <SPOT_LENGTH>76</SPOT_LENGTH>
                    <READ_SPEC>
                        <READ_INDEX>0</READ_INDEX>
                        <READ_CLASS>Application Read</READ_CLASS>
                        <READ_TYPE>Forward</READ_TYPE>
                        <BASE_COORD>1</BASE_COORD>
                    </READ_SPEC>
                </SPOT_DECODE_SPEC>
            </SPOT_DESCRIPTOR>
        </DESIGN>
        <PLATFORM>
            <ILLUMINA>
                <INSTRUMENT_MODEL>Illumina Genome Analyzer
II</INSTRUMENT_MODEL>
            </ILLUMINA>
        </PLATFORM>
        <PROCESSING/>
    </EXPERIMENT>
</EXPERIMENT_SET>
```

Run.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<RUN_SET  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:noNamespaceSchemaLocation="ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_3/SRA
.run.xsd">
```

```
    <RUN alias="tai2010_GM10847_SE76_GCK9J" center_name="ikmb_kiel"
run_center="center_jena" run_date="2010-08-02T10:00:00">
        <EXPERIMENT_REF refname="tai2010_GM10847_SE76"/>
        <DATA_BLOCK>
            <FILES>
                <FILE filename="GM10847_s76.fq.gz" filetype="fastq"/>
            </FILES>
            <FILES>
                <FILE filename="GM10847_GCK9J_s76.bam" filetype="bam"/>
            </FILES>
        </DATA_BLOCK>
    </RUN>
</RUN_SET>
```

2.3 Europäische „Genome-phenome Archive" (EGA)

Das Europäische „Genome-phenome Archive" (EGA) (http://www.ebi.ac.uk/ega) stellt eine Plattform für Archivierung und Verteilen der personalisierten genetischen und phänotypischen Daten. EGA verwaltet und bietet sicheren Zugriff auf die gespeicherten Daten für die eingetragenen Benutzer. Jeder Benutzer wird durch evtl. sicherheitstechnisch schwache Email-Identifizierungsverfahren überprüft. Eine unabhängige Ethik-Kommission überwacht die EGA Archivierungsmethoden und die Infrastruktur.

Die Personen oder Organisationen, die die Daten erzeugt und in EGA hochladen, stellen für EGA eine Datenzugriff-Organisation (consortium data access committee (CDAC)). CDAC verwaltet die Zugriffrechte und kann bestimmte oder komplette Projektdaten für Benutzer freigeben. EGA stellt dann für Benutzer die Möglichkeit auf die freigegebenen Daten zuzugreifen.

2.4 Sequenzierungsmetadaten

Mit massiver Ausweitung der Anwendung der Next-Gene-Sequenzierungsdaten wird die Verwendung standardisierter Begriffe für Annotation der Sequenzierungsdaten notwendig. Unterschiedliche Konsortien wie GSC (The Genomic Standards Consortium: http://gensc.org) versuchen die minimalen Beschreibungsdaten für ein Sequenzierungsexperiment zu definieren und zu standardisieren. Z.B. befasst sich ein Unterprojekt MIGS/MIMS (Minimum Information About a (Meta)Genome

Sequence) (6) mit der Annotation der Sequenzierung des unter anderen menschlichen Genoms. Solche Informationen sind in der SRA XML Spezifikationsformat für Sequenzierungsexperimente.

3 Ergebnisse

3.1 Use Case: Ablauf der Analyse der Sequenzierungsdaten

Im Folgenden wird der an der Universität Kiel im Rahmen des Projektes beispielhaft implementierte Use-Case (Abb. 2) dargestellt. Das Experiment wird von Wissenschaftlern erarbeitet und die notwendigen aufbereiteten Proben die Sequenzierungszentren bereitgestellt. In großen Sequenzierzentren werden die Daten in ein lokales LIMS System integriert. Die Sicherung und die Archivierung der Daten geschieht bei großen Zentren zentral über das Rechenzentrum über Archivierungssysteme (wie Tape) oder über lokale USB Festplatten. Bei externen Sequenzieraufträgen werden die Daten meistens über einen FTP-Server des Sequenzierzentrums ausgetauscht oder die Daten werden mit externen USB-Platten per Post verschickt. Der Austausch der großen Datenmengen bei kleineren Datennetzbandbreiten oder möglichen Beschädigungen von Datenträgern bei Postversand mit externen Auftraggebern stellt für die Sequenzierungszentren eine Herausforderung dar.

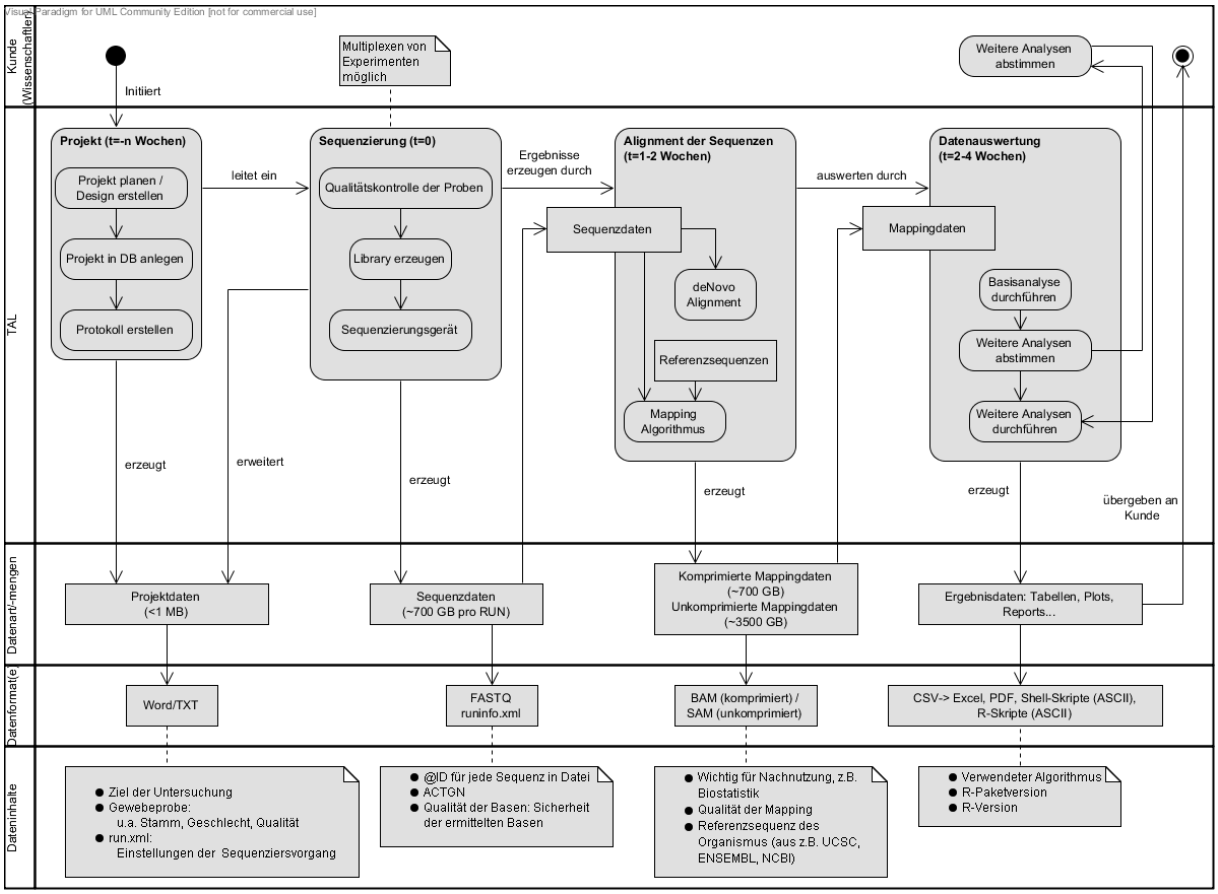

Abbildung 2: Workflow eines Sequenzierungsprojekts als UML-Diagramm

Für die notwendige Annotation der Sequenzierdaten trägt der Wissenschaftler die Verantwortung. Unterschiedliche Sequenzierzentren bilden dies teils durch interne LIMS Systeme ab. Die Projektwissenschaftler müssen die erforderlichen Informationen über selbst entsprechend der lokalen Spezifikationen zusammenstellen. Häufig erfolgt auch keine strukturierte Annotation der Metadaten – das heißt die Sequenzdaten sind im LIMS System und für die biologische Information werden traditionelle Dokumentationssysteme (Stichwort Laborbuch) genutzt. Z. B. sind für Labore der Sicherheitsstufe 1 die papierbasierten Laborbücher zwingend vorgeschrieben. Die parallele elektronische Buchführung ist eher ein Ausnahmefall. Dies stellt natürlich ein Problem für die digitale Langzeitarchivierung dar. Es gibt unterschiedliche Anstrengungen für Minimaldatensätze zur Dokumentation von Forschungsdaten die je nach Bereich zur Anwendung kommen könnten und für viele biologische Felder im MIBBI-Projekt zusammengefasst sind (7). Ein möglicher Leitfaden für die Mindestanforderungen an die Metadaten für Archivierungsdaten bietet die SRA XML Spezifikation. Dieses Format (siehe vorigen Abschnitt) soll als Grundgerüst für den geplanten Use-Case genutzt werden.

3.2 Datenaufbewahrung

Die Datenarchivierung ist gemäß Gendiagnostikgesetz § 12 für mindestens 10 Jahre und länger ausgelegt (5). Die Daten werden im Beispiel des Standorts Kiel zweifach auf einem RAID-Festplattenverbund und Tape-Roboter des Rechenzentrums Kiel archiviert. Die Daten dienen ausschließlich der wissenschaftlichen Forschung und werden unbegrenzt lange abhängig von jeweiliger Patientenverfügung aufbewahrt. Am Standort Kiel wird nach 20 Jahren durch eine Kommission aus öffentlichen Trägern entschieden, ob eine weitere Speicherung die Patientendaten für wissenschaftliche Forschung sinnvoll ist oder die Daten sofort gelöscht werden sollten. Die Löschung auf Grund eines Patientenantrages bzw. der Zurücknahme der Einwilligungserklärung eines Patienten ist jederzeit möglich. Der Daten- und Probenerheber, im Fall Kiel das Biobankprojekt POPGEN (http://www.popgen.de), stellt die vollständige Durchführung sicher. Bei strittigen Fällen wäre das Landesamt für Datenschutz zuständig.

3.3 Ethik

Die Speicherung genetischer Daten zu reinen Archivierungszwecken im Sinne des ursprünglichen Projektziels ist ethisch unbedenklich, solange die entsprechenden technischen und organisatorischen Voraussetzungen vor einem unbefugten Datenzugriff schützen.

Ethische Probleme treten bei der Archivierung zum Zwecke der Nachnutzung von Forschungsdaten auf: aus ethischer Sicht liegt hier ein Zielkonflikt vor:

Einerseits besteht von Seiten der Spender und Betroffenen ein Interesse an möglichst effektiver Datennutzung mit dem Ziel der Aufklärung der jeweils zugrundeliegenden Erkrankungen. Dies würde für einen möglichst einfachen Datenzugang sprechen mit unkompliziertem Zugriff aller interessierten Forscher. Dieses Vorgehen ist vom National Institutes of Health (NIH) im Rahmen von dieser Agentur finanzierten GWAS-Studien gewählt worden (8). Hier ist eine weitgehende öffentliche Zugänglichkeit der Daten in dbGAP als zwingende Fördervorraussetzung festgeschrieben worden.

Im Sinne des Schutzes der Privatsphäre der Spender bestehen bei diesem Vorgehen etliche Probleme, die in Krawczak et al. 2010 (Investig Genet. 2010; 1: 3) dargestellt sind. Bereits mit 30-80 SNPs (zum Bsp. aus einem Mundschleimhautabstrich typisierbar) wäre hier eine Re-identifizierung von Individuen gegen eine genomweite SNP-Datenbank möglich. Derartige Betrachtungen müssen daher klar in der Einverständniserklärung für die Probanden verankert sein.

Die praktische Relevanz dieser Überlegungen im Lichte der weit verbreiteten sozialen Netzwerke für den Einzelnen bedarf einer allgemeinen gesellschaftlichen Diskussion.

4 Diskussion

Die aktuellen *de-facto* Standards für Sequenzierungsdaten wurden im Methodenteil vorgestellt. Die Vielfalt, und insbesondere die formalen und strukturellen Schwächen der aktuellen Metadatenformate für Sequenzierungsdaten, zeigen den schwierigen Weg der Standardisierung. Die ersten Initiativen, wie MIGS, könnten die notwendigen minimalen Informationen für die Anforderungen der Archivierung liefern.

Bei der Archivierung der medizinischen Genotyp- oder Phänotyp-Daten sind viele Rahmenbedingungen der nationalen Gesetzgebungen zu erfüllen. Einerseits muss Datenschutz für den Einzelnen sehr hoch sein, andererseits muss der Mechanismus des Datenschutzes ausreichend transparent sein, damit es den legitimen und genehmigten Einsatz der Daten nicht behindert.

Aus ethischer Sicht besteht bei der Archivierung von genetischen und phänotypischen Daten ein Zielkonflikt. Einerseits besteht von Seiten der Spender und Betroffenen ein Interesse an möglichst effektiver Datennutzung mit dem Ziel der Aufklärung der jeweils zugrundeliegenden Erkrankungen. Dies würde für einen möglichst einfachen Datenzugang sprechen. Dem steht auf der anderen Seite der Schutz der Privatsphäre der Spender entgegen, da insbesondere genomweite Daten bei Verfügbarkeit der genetischen Information eines Individuums eine leichte Re-identifizierung erlauben. Die praktische Relevanz dieser Überlegungen im Lichte der weit verbreiteten sozialen Netzwerke für den Einzelnen bedarf einer gesellschaftlichen Diskussion und muss Teil des „informed consent“ sein.

Die europäische ENA und EGA Implementierung der Verwaltung von Sequenzierungsdaten könnte den Leitfaden für nationalen Archivierungsdienst dienen. Die Implementierung fein granularer Zugriffrechte könnte durch erweiterte XML-Annotation geschehen, hierzu gibt es mit XACML (eXtensible Access Control Markup Language) bereits eine mögliche Lösung. Dabei können die komplexen nationalen Datenschutzanforderungen der medizinischen Versorgung leichter erfüllt werden.

Die Nutzung der etablierten Rechenkapazitäten und Infrastruktur des D-Grid könnte die Implementierung des Archivierungssystems für Sequenzierungsdaten erleichtern.

Dabei steigert die dezentrale Speicherung bei D-Grid die Ausfallsicherheit und die Verfügbarkeit der archivierten Daten. Die etablierten Identifizierungsverfahren der Benutzer durch D-Grid-Zertifikate erlauben größere Sicherheit im Vergleich zur Email-Identifikation.

Die Implementierung der Empfehlungen von GSC Konsortien für minimale Annotation eines Sequenzierungsexperiments könnte eine standardisierte Metadatenstruktur liefern. Eine Verpflichtung zu deren Einhaltung könnte ein erster Schritt zur einheitlichen Dokumentation von Sequenzierungsdaten sein.

5 Literaturverzeichnis

1. Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science 291(5507): 1304-1351.
2. Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proc Natl Acad Sci U S A 85(8): 2444-2448.
3. Cock PJA, Fields CJ, Goto N et al (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research 38(6):1767-1771
4. The SAM Format Specification Working Group (2011). The SAM Format Specification. Letzter Zugriff: 2012.06.28, URL: http://samtools.sourceforge.net/SAM1.pdf
5. Scholz C (2012). Aufbewahrungsfristen für ärztliche Unterlagen. Deutsche Gesellschaft für Humangenetik 1, Letzter Zugriff: 2012.05.09, URL: http://www.gfhev.de/de/qualitaetsmanagement/aufwahrungsfristen.pdf
6. Field D, Garrity G, Gray T et al (2008). The minimum information about a genome sequence (MIGS) specification. Nature Biotechnology 26(5):541-547
7. Taylor, C. F. et al. (2008) "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." Nature Biotechnology 26(8):889-896
8. National Institutes of Health (NIH) (2007). Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). National Institutes of Health (NIH), Letzter Zugriff: 2011.09.28, URL: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html