LABIMI/F

Deliverable

Alexander Herrmann, Jochen Hampe

Use Case – Genomdaten[1]

	Use Case – Genomdaten
Autor(en)	Alexander Herrmann, Jochen Hampe
Editor(en)	Frank Dickmann, Jochen Hampe
Datum	27.09.2011
Version des Dokuments	1.0.2

A: Status des Dokuments

Version 1.0.1

[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 D2.1: Dokumentierter Use Case, Version 1

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. Vor allem 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
0.0.1	27.09.2011	A. Herrmann	Erste Dokumentversion
0.0.2	10.10.2011	A. Herrmann	Erste Überarbeitung
1.0.1	29.11.2011	J. Hampe	Zweite Überarbeitung

E: Inhaltsverzeichnis

1 Einleitung 5
2 Material und Methoden 7
2.1 Wichtige Formate der Sequenzierungsdaten 7
2.1.1 FASTA Format 7
2.1.2 FASTQ-Format 8
2.1.2.1 Illumina sequence identifiers 9
2.1.2.2 NCBI Sequence Read Archive 10
2.1.3 Alignmentformate 10
2.1.3.1 SAM Format 10
2.2 Archivierungssystem ENA 12
2.2.1 Prozess der Dateneinreichung zur ENA 12
2.2.2 Beispiel der XML Spezifikation für ENA 13
3 Ergebnisse 19
3.1 Use Case: Ablauf der Analyse der Sequenzierungsdaten 19
4 Diskussion 21
5 Literaturverzeichnis 22

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 (Venter, Adams et al. 2001). 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 neue Hochdurchsatz-Genotypisierungs- und Sequenzierungs-technologien, 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 granulierbare Zugriffsmöglichkeiten zu schaffen. Die personelle und technische Infrastruktur inhaltsorientierter Genom-forscher ist auf diese Herausforderungen bisher praktisch nicht eingestellt.

Im Rahmen DFG-Vorhabens LABIMI/F 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/intellektuelle/inhaltliche 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 (Pearson and Lipman 1988). 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 verbreiteteste Datenformat für Sequenzdaten, selbst im Rahmen der Ausgabeformate von Hochdurchsatzsequenzierern. Das Format unterstützt Metadaten nur rudimentär und wenig strukturiert. Es 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 jeweilige Sequenz selbst. Jede 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-Dateien können multiple Sequenzen enthalten, die häufig aus einem gemeinsamen biologischen oder experimentellen Kontext entstammen. FASTA-Formate werden von einer Vielzahl von Sequenzanalyseprogrammen und Alignment-Werkzeugen akzeptiert, 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. Dabei wird sowohl das Nukleotid als auch der Qualitäts-Score in einem einzelnen ASCII Zeichen untergebracht. Das Format wurde ursprünglich am Sanger Center (UK) zur Bündelung 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 Qualitäts-Scores für alle Sequenzzeichen aus Zeile 2. Eine minimale FASTQ Datei könnte etwa wie folgt aussehen:

```
@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 Qualitäts-Scores vorkommen können.

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

2.1.2.1 Illumina sequence identifiers

Illumina ist aktuell der dominierende Hersteller von Zweitgenerationssequenzierern, ca. 70%-90% der neuen Sequenzen werden über Illumina-Geräte erzeugt. Der Code 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

2.1.2.2 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
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

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

2.1.3 Alignment-Formate

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.

2.1.3.1 SAM Format

Das SAM (Sequence Alignment/Map) ist ein text-basiertes, Tabulator-getrenntes Format für die Darstellung von Sequenzalignments. 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 um 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.

Das European Nucleotide Archive (ENA) kann Sequenzen aus Next Generation Sequencing-Projekten und -Technologien wie Roche 454, Illumina Genome Analyzer and ABI SOLiD in sein Sequenzarchiv aufnehmen (Sequence Read Archive – SRA). ENA arbeitet eng mit dem NCBI und der DDBJ zusammen und tauscht die öffentlichen Daten täglich aus. Das Europäische „Genome-phenome Archive” (EGA) nimmt Daten entgegen, die nur unter kontrollierten Bedingungen zugänglich gemacht werden sollen. So sollen sensitive Information der Einreicher/Forscher geschützt werden.

2.2.1 Prozess der Dateneinreichung zur ENA

Die XML-Spezifikationen für Archivierungsdaten sehen wir folgt aus: Alle Teilbereiche der Sequenzdaten werden über ein XML-Metadatenformat abgedeckt. Eine typische SRA Einreichung enthält fünf SRA XML-Dateien: Submission, Study, Sample, Experiment und Run XML.

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>
```

3 Ergebnisse

3.1 Use Case: Ablauf der Analyse der Sequenzierungsdaten

Im Folgenden wird der an der Universität Kiel im Rahmen des Projektes LABIMI/F der beispielhaft implementierte Use Case (Abb. 1) 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 Rechenzenrum 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 USB-Festplatten per Post verschickt. Der Austausch der großen Datenmengen mit externen Auftraggebern stellt für die Sequenzierzentren eine organisatorische und technische Herausforderung dar.

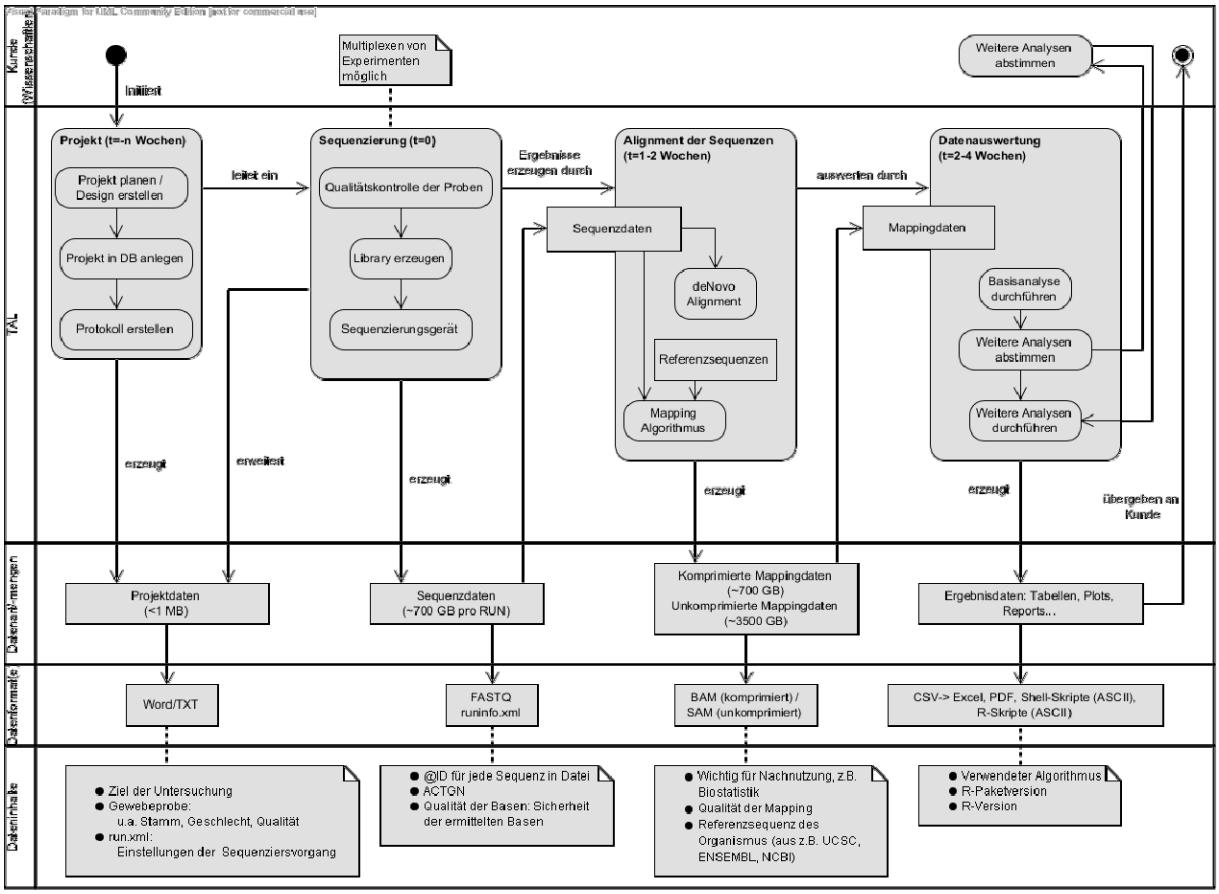

Abb 1: Workflow eines Sequenzierungsprojekts als UML-Diagramm.

Für die notwendige Annotation der Sequenzierdaten trägt der jeweilige Wissenschaftler die Verantwortung. Unterschiedliche Sequenzierzentren bilden dies teils durch interne LIMS[2] Systeme ab. Die Projektwissenschaftler müssen die erforderlichen Informationen entsprechend der lokalen Spezifikationen selber zusammenstellen. Häufig erfolgt keine strukturierte Annotation der Metadaten – das heißt die Sequenzdaten sind im LIMS System und für die biologische Information werden traditionelle Dokumentationssysteme (papierbasierte Laborbücher) genutzt. Der Medienbruch und das Fehlen von Metadaten stellen natürlich ein Problem für die Langzeitarchivierung der digitalen Daten dar. Ein Leitfaden für die Mindestanforderungen an die Metadaten für Archivierungsdaten könnte hier die internationale Spezifikation (ENA) bieten - beschreibt aber nur die frei zugänglichen Daten. Da im Rahmen der Langzeitarchiverung primär abgeschlossene Projekte und Datensätze relevant sind, ist dies keine relevante Einschränkung. Dieses Format (siehe vorigen Abschnitt) soll aus Grundgerüst für den hier geplanten Use Case genutzt werden.

[2] Laboratory Information Management System.

4 Diskussion

Die Vielfalt, und insbesondere die formalen und strukturellen Schwächen der aktuellen Formate, die sich als inoffizielle de-facto Standards durchgesetzt haben sind im Methodenteil klar herausgestellt worden.

Im Rahmen dieses Projektes soll der Workflow über ein ENA-kompatibles XML-Format abgebildet werden und wird dann im Deliverable 2.2 dokumentiert werden.

5 Literaturverzeichnis

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.

Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.