

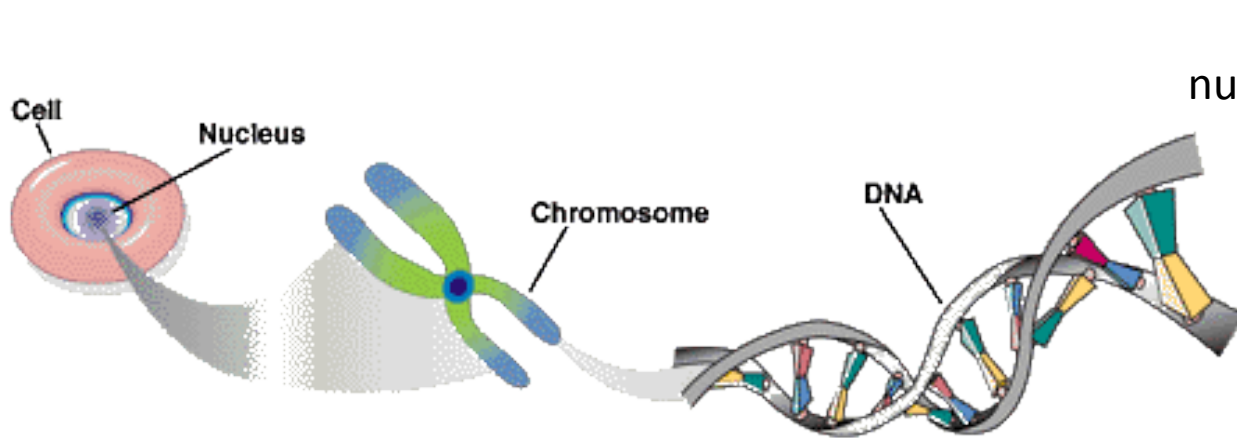
Datenstandards bei Sequenzierungsdaten

A. Herrmann
Arbeitsgruppe J.Hampe, UKSH Kiel

Einleitung

- Datenformate:
 - Sequenzen: FASTA, FASTQ
 - Alignment: SAM/BAM
- Metadaten eines Sequenzierungsexperimentes am Beispiel Sequence Read Archive (SRA)

Nukleotidsequenzen



IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)

FASTA Format

```
>sequence_name_1  
ACGTACGT  
ACGTACGT  
>sequence_name_2  
ACGTACGT  
ACGTACGT
```

R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Alignmentformat SAM

Positionen 12345678901234 5678901234567890123456789012345
Referenz AGCATGTTAGATAA--GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
Reads:
+r001/1 TTAGATAAAGGATA-CTG
+r002 ATAGCT.....TCAGC
-r001/2 CAGCGCCAT

SAM format:

@HD VN:1.3 SO:coordinate

@SQ SN:ref LN:45

r001	163	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*
r002	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*
r001	83	ref	37	30	9M	=	7	-39	CAGCGCCAT	*

BAM: SAM Komprimierung

- BGZF Kompression Format
 - Kleine Blöcke mit gzip komprimiert
 - Zufallszugriff durch Index
 - Schneller Positionszugriff bei vorsortierten SAM File
- Entpacken (BAM → SAM) mit gunzip möglich

Variant Call Format (VCF)

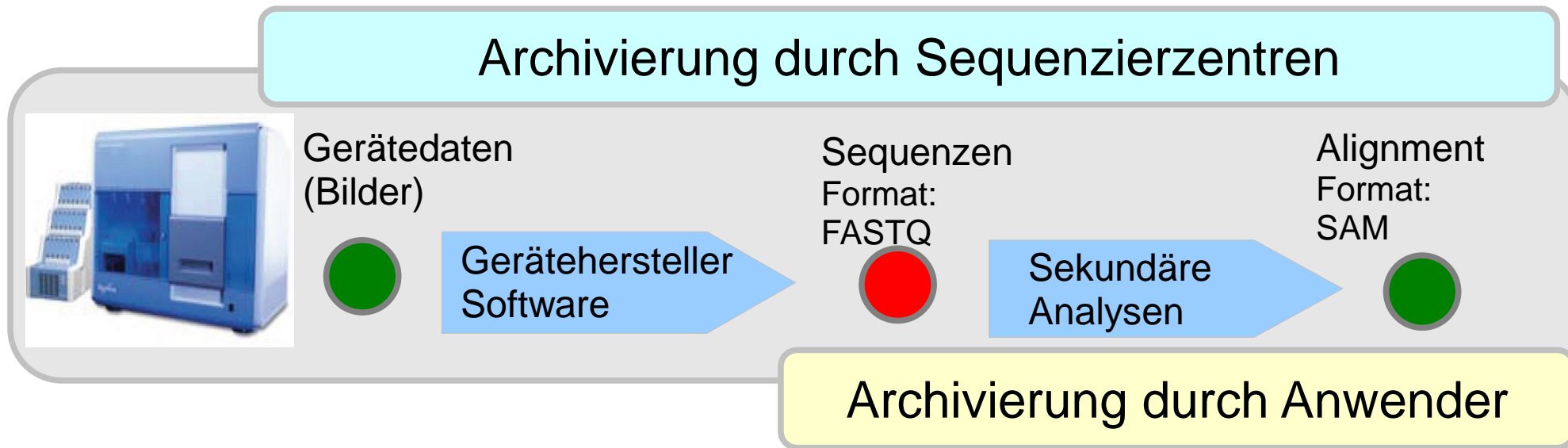
Sequenzvariationen

- Header mit Metadaten
- Jede Zeile - eine Position in Genome

```
#CHR POS ID REF ALT QUAL FILTER INFO
20 14370 rs6054257 G A 29 PASS AF=0.5
20 17330 . T A 3 q10 AF=0.017
20 1110696 rs6040355 A G,T 67 PASS AF=0.333,0.667

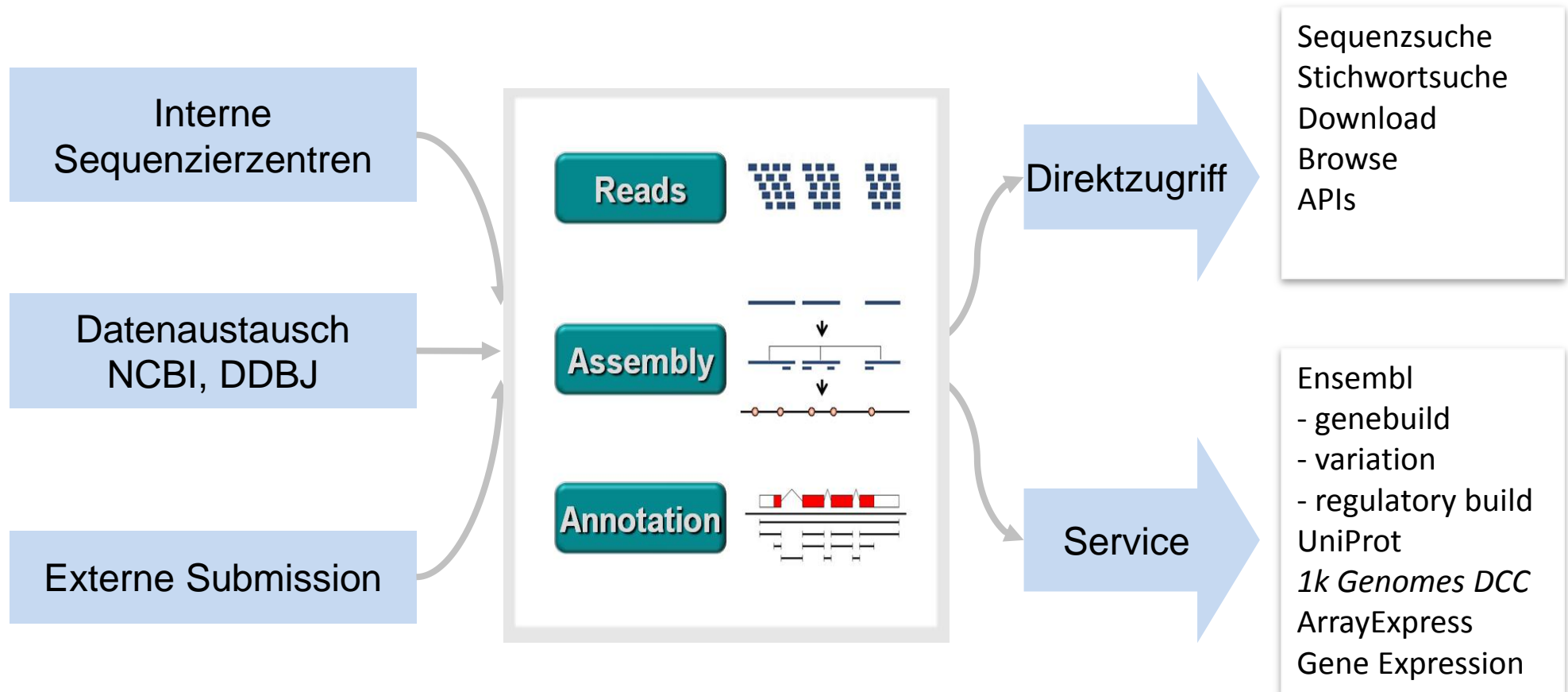
... FORMAT NA00001
... GT:GQ:DP:HQ 0|0:48:1:51,51
... GT:GQ:DP:HQ 0|0:49:3:58,50
... GT:GQ:DP:HQ 1|2:21:6:23,27
```

Sicht der Anwender



- Sicherung der Sequenzen:
 - FASTQ
 - Proprietäre Formate der Gerätehersteller
- Aufbewahrung
10 Jahre / 30 Jahre ?

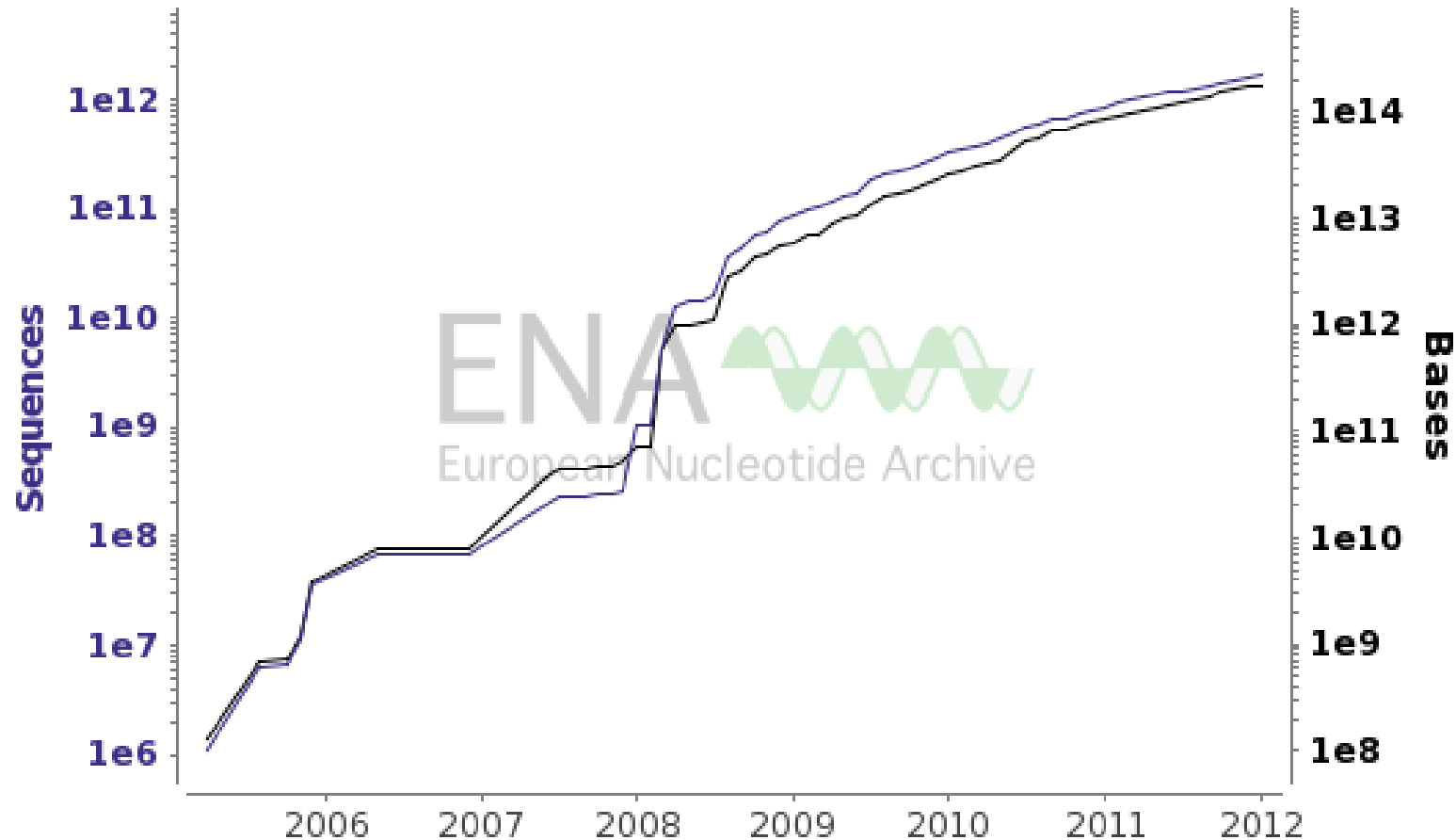
European Nucleotide Archive



EBI SRA

- Primäre Archive für Next-Generation Sequenzierungsdaten und Alignments (BAM)
- Veröffentlichung der Daten mit Publikation
- Sperrfrist möglich
- Freier Zugriff auf die Daten
- Neue Algorithmen für Speicherung

SRA: Wachstum



EMBL-Bank → Index

200M Sequenzen mit 600G Basen

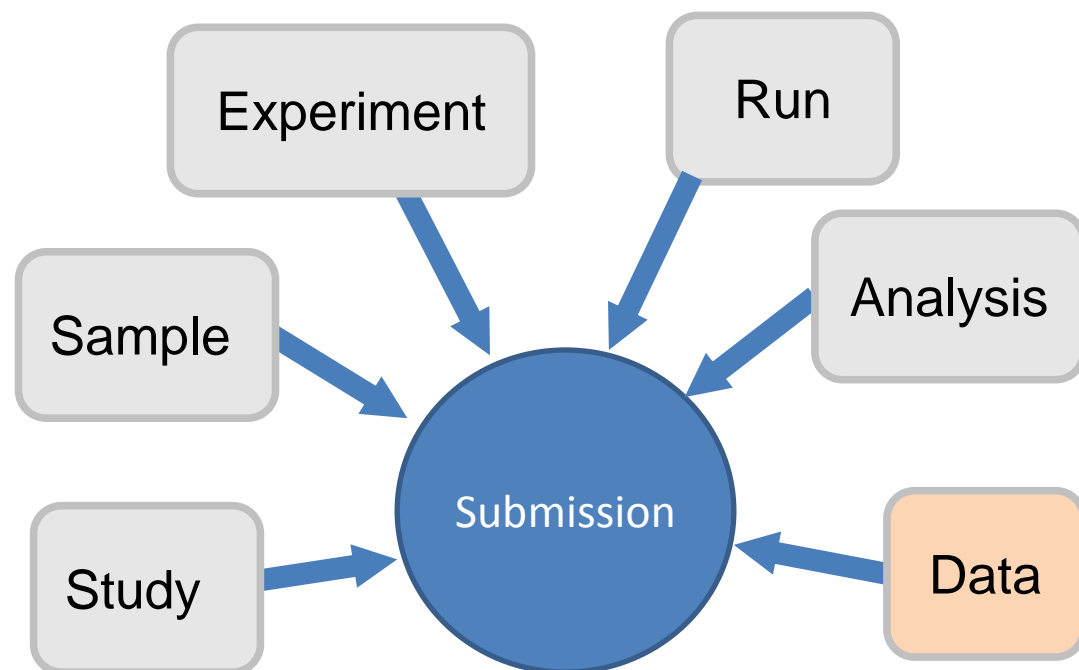
SRA → Kein Sequenzindex

1.3 Billion Sequenzen mit 133 Billion Basen

SRA Submission

http://www.ebi.ac.uk/ena/about/sra_submissions

- Datenformats
- Metadaten Objekte
- Submission Account
- Übertragung
- Anpassen
- Dokumentation



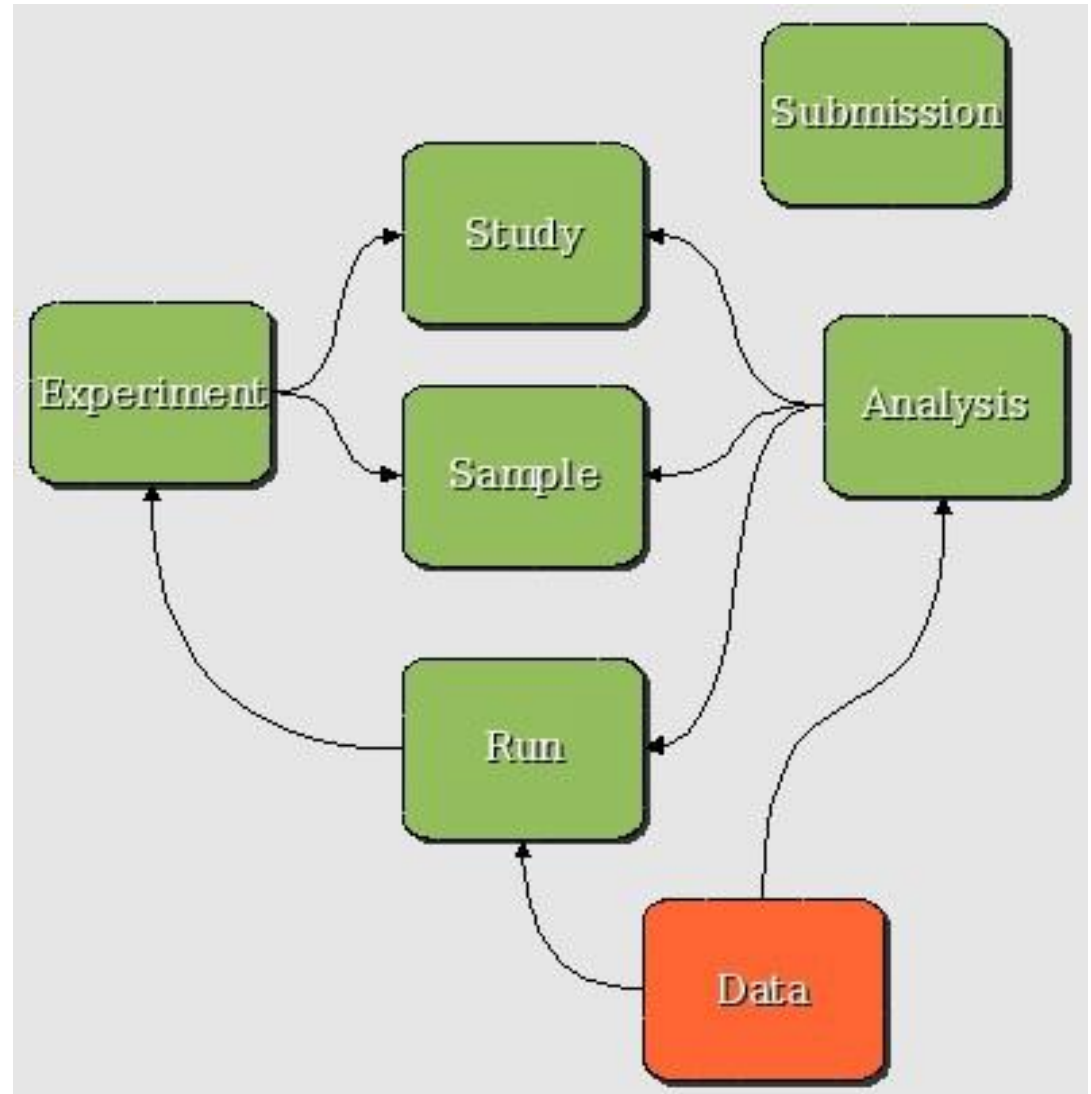
Submission Datenformat

- Bevorzugte Format **BAM**
- Andere:
 - SRF, Fastq, SFF, SOLiD_native,
Illumina_native, PacBio_HDF5,
CompleteGenomics_native
- Minimum Information: Basen und deren Qualität
- Sequenzierungen mit Barcode demultiplexen
- Technische Sequenzen eliminieren

Metadaten Objekte

- Submission
- Study
- Sample
- Experiment
- Run
- Analysis

- Für EGA andere:
 - DAC, Policy, Dataset



Metadaten Objekte (1)

- Submission
 - Neu oder Änderungen
 - Veröffentlichungstermin
- Study
 - Projektziel mit Beschreibung
- Sample
 - Probenbeschreibung
 - Organismus

Metadaten Objekte (2)

- Experiment
 - Library Information
 - Platform Information
 - Verbindet Study mit Sample und Run
- Run
 - Datenfiles mit primären Sequenzdaten
 - md5sum
- Analysis
 - BAM files mit Referenzsequenzen

Metadaten Objekte (3)

- Existierenden Study und Sample Objekte
- Eindeutiger Name für jeden Objekt
- Jeder Objekt bekommt Zugriffsnummer
 - Submission: ERANNNNNNN
 - Study: ERPNNNNNNN
 - Sample: ERSNNNNNNN
 - Experiment: ERXNNNNNNN
 - Run: ERRNNNNNNN
 - Analysis: ERZNNNNNNN

SRA Webin: metadaten submission

Home **New Submission** Studies Samples Experiments Runs Projects

Start >> Study >> Sample >> Run >> Finish

You are about to make a new submission into ENA's Sequence Read Archive (SRA).

Please select the type of submission you would like to make:

I wish to do a complete submission (study, samples and sequence reads)

When submitting sequence reads you will be asked to provide information about sequenced samples, libraries, instruments and data files. Each submission will be associated with a single study and data for different studies must be submitted in separate submissions. Please quote assigned study accession numbers (ERPNNNNNN) when citing submitted data.

I wish to register studies

I wish to register samples


I wish to register a project

The first step in the sequence read submission is to upload data files. If you have already uploaded them into your dropbox please proceed directly to the next step. To use our data upload tool please follow the instructions indicated in the following link.

[SRA File Upload instructions](#)

Or directly access to the tool [SRA-FileUploader](#).

SRA FileUploader

ENA 
European Nucleotide Archive

Username: Password: Upload Protocol:

Local Upload Directory:

Upload	Name	Size	Date	MDS Checksum	Progress
<input type="checkbox"/>					

Overwrite Existing Files Upload Directory Tree Selection:

FASTQ gezippt

ENA Browser

Text search **Sequence search**

Enter or paste text or ENA accession number:

Search

Upload file of accessions:

Choose File

No file chosen

Search

SRA Study: ERP000013 : Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment -

View: [XML](#)

Download: [XML](#)

[Send Feedback](#)

Submitting Centre

Center for Microbial Ecology at Michigan State University

Study Type

Metagenomics

Read Count

2,632

Base Count

475Kb

Abstract

Understanding the relationship between gene diversity and function for important environmental activities is a major ecological research goal. We applied gene-targeted-metagenomics and pyrosequencing to aromatic dioxygenase genes to obtain greater sequence depth than possible by other methods.

[Bulk download Fastq/Submitted files](#)

Navigation

Fastq Files

Submitted Files

Attributes

Please note that submitted files are available in the Submitted Files tab.


View: [TEXT](#)

Download: [TEXT](#)

[Select columns](#)

Study	Sample	Run	Organism	Instrument Model	Library Layout	Run Read Count	Run Base Count	ftp	Aspera	Galaxy
ERP000013	ERS000029	ERR003029	soil metagenome	454 GS FLX	SINGLE	2,024	354Kb	Fastq file#1	Fastq file#1	Fastq file#1
ERP000013	ERS000030	ERR003030	soil metagenome	454 GS FLX	SINGLE	608	121Kb	Fastq file#1	Fastq file#1	Fastq file#1

SRA FileDownloader

ENA 
European Nucleotide Archive

[Contact the ENA Helpdesk](#)

Accession Number:

Local Download Directory:

Remote Files

Download	Name	Size	Progress
<input type="checkbox"/>	ERR003029.fastq.gz	136Kb	<div style="width: 0%;"></div> 0%
<input type="checkbox"/>	ERR003030.fastq.gz	46Kb	<div style="width: 0%;"></div> 0%

Selection:

Zusammenfassung

- SRA – öffentliche Archivierungsstelle für Next-Generation Sequenzierungsdaten
 - Sequenzarchivierung mit Metadaten
 - Sequenzdateisuche über Metadatenindex
- Verdoppelung der SRA Datenmenge pro Jahr
- Wichtige Sequenzierungsformate:
FASTQ, BAM