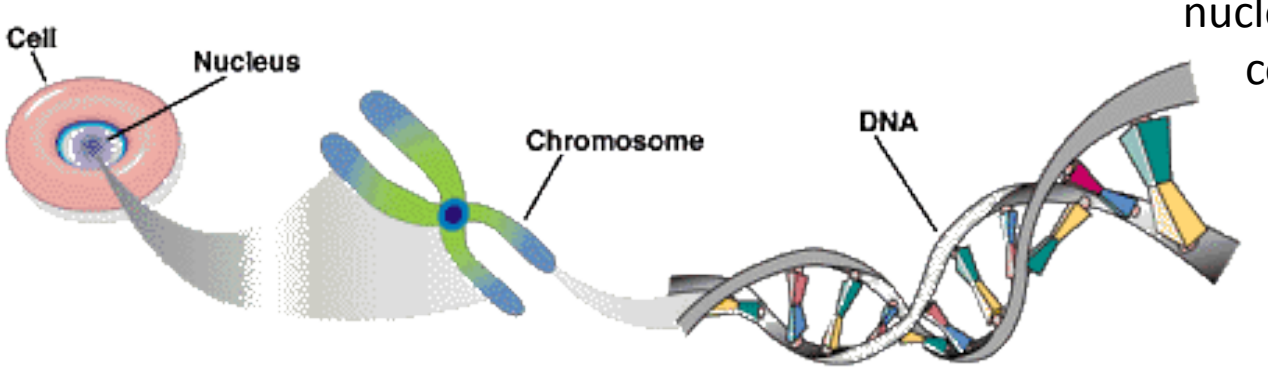
Datenstandards bei Sequenzierungsdaten

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)

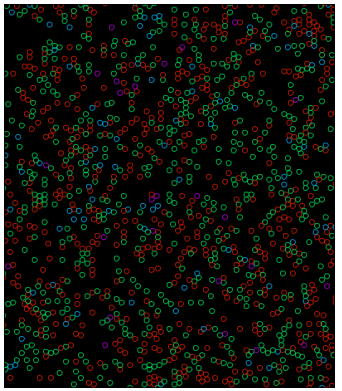
FASTA Format

```
>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

FASTQ

Readformat:



```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;8!
```

Qualitätswerte:

0-40 Stufen: $-10 \log_{10}(\Pr\{\text{Die Base ist falsch}\})$

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)

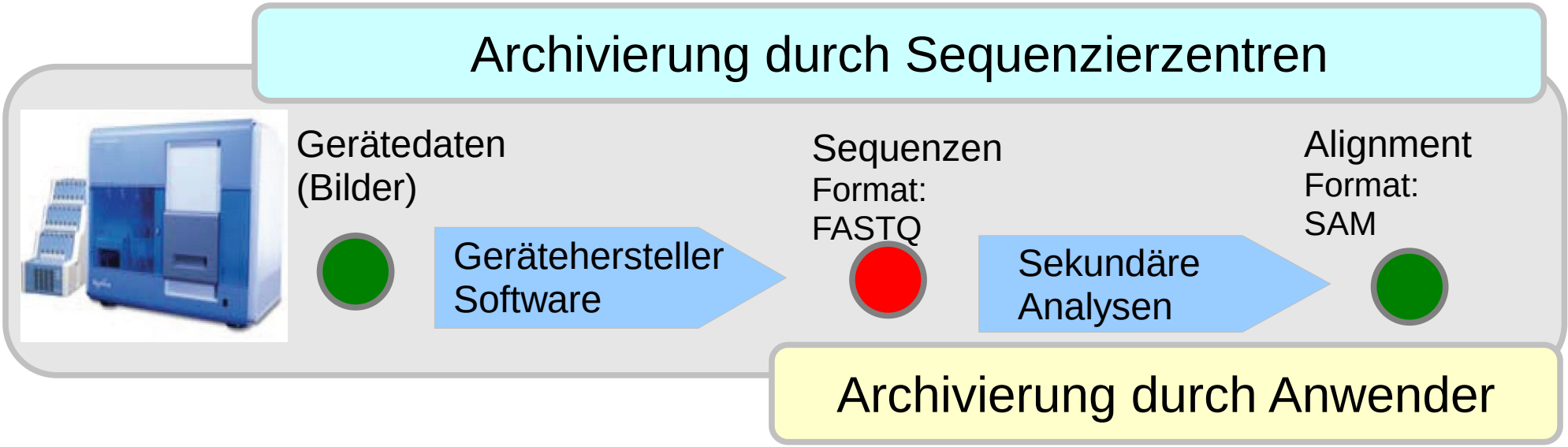
Sequenzvariationen

- 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
```

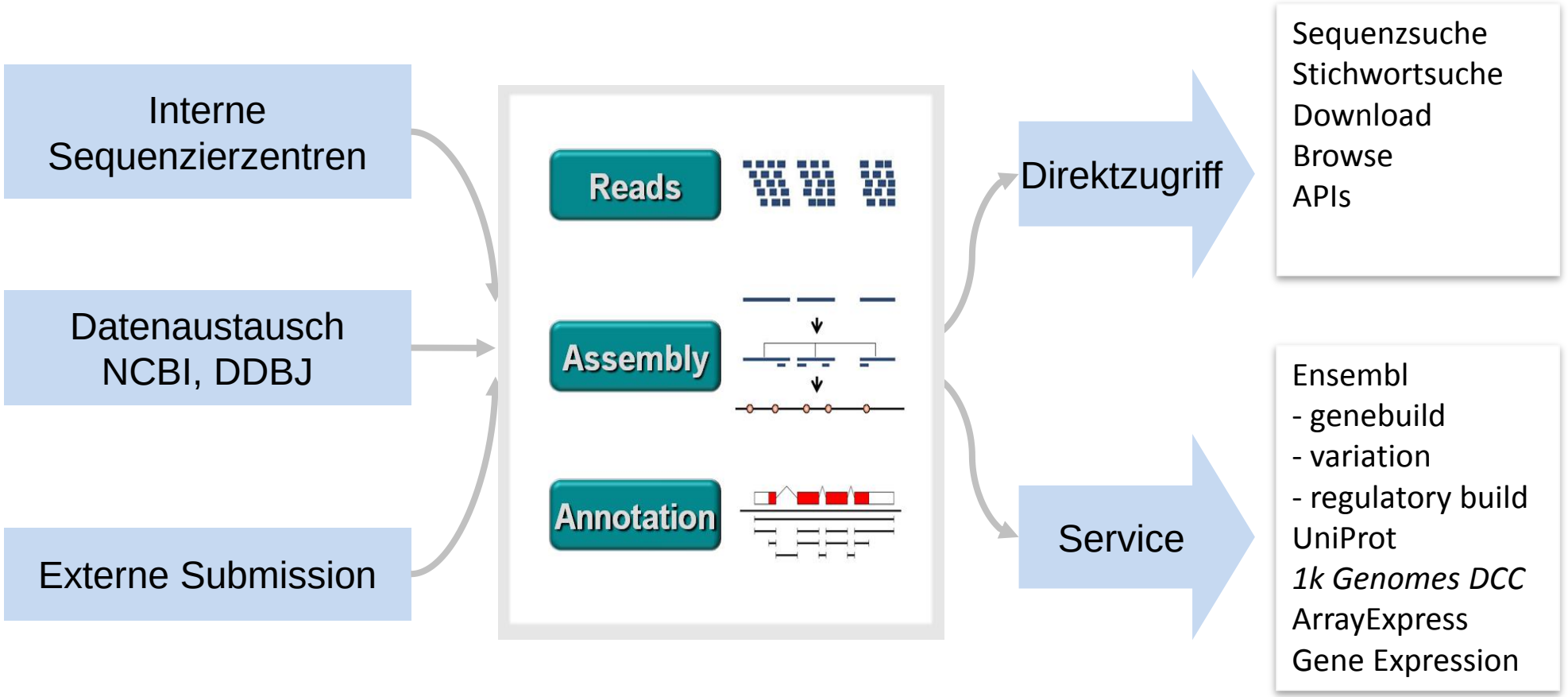
```
…       FORMAT      NA00001
…       GT:GQ:DP:HQ 0|0:48:1:51,51
…       GT:GQ:DP:HQ 0|0:49:3:58,50
…       GT:GQ:DP:HQ 1|2:21:6:23,27
```

Sicht der Anwender



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

European Nucleotide Archive

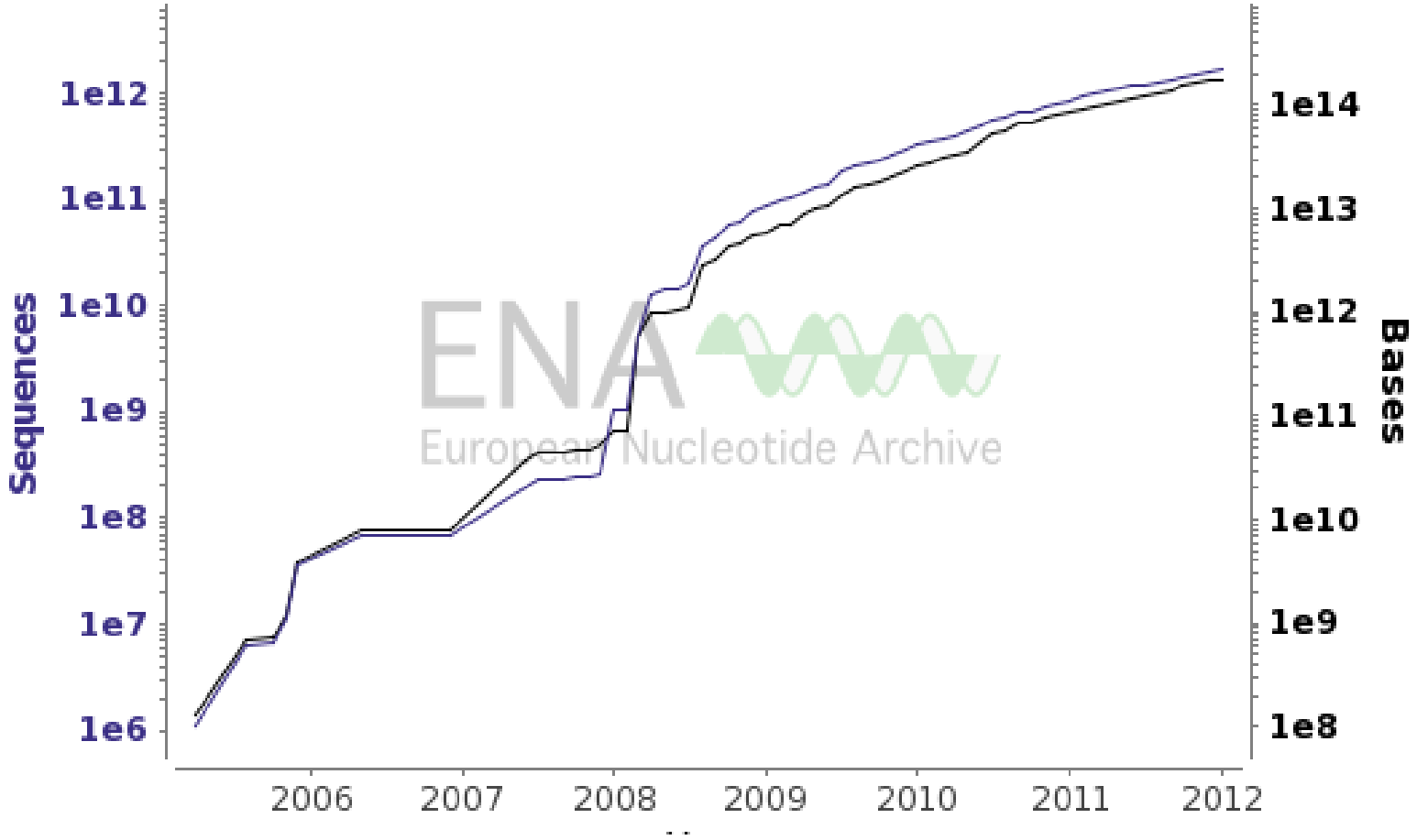


ENA

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

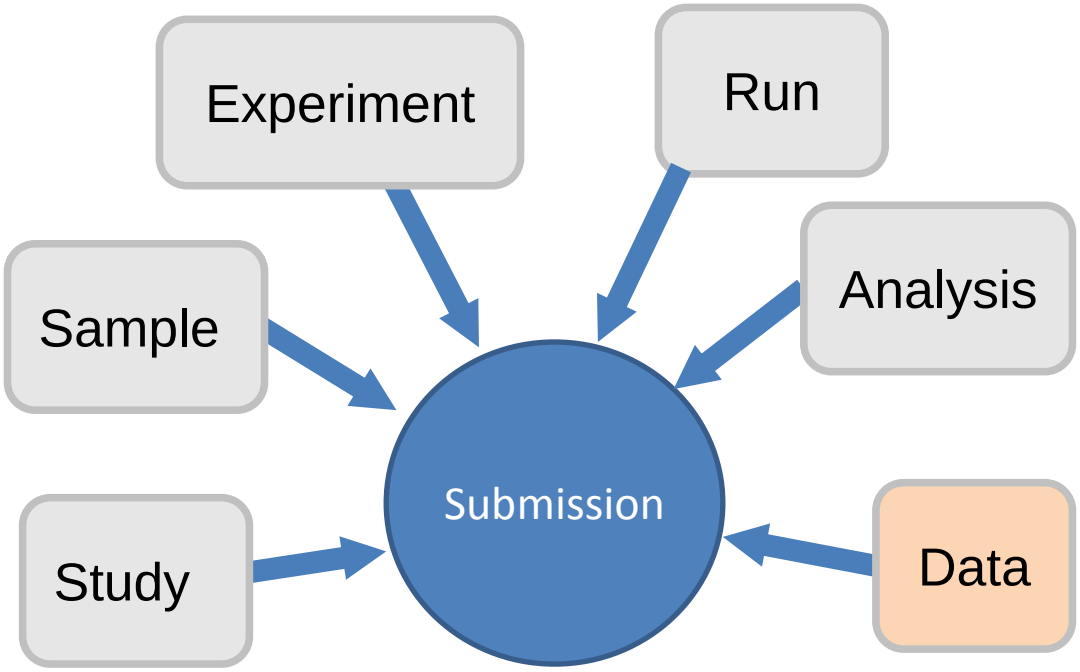
1.3 Billion Sequenzen mit 133 Billion Basen

Rasko Leinonen, EMBL-EBI

SRA Submission

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

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

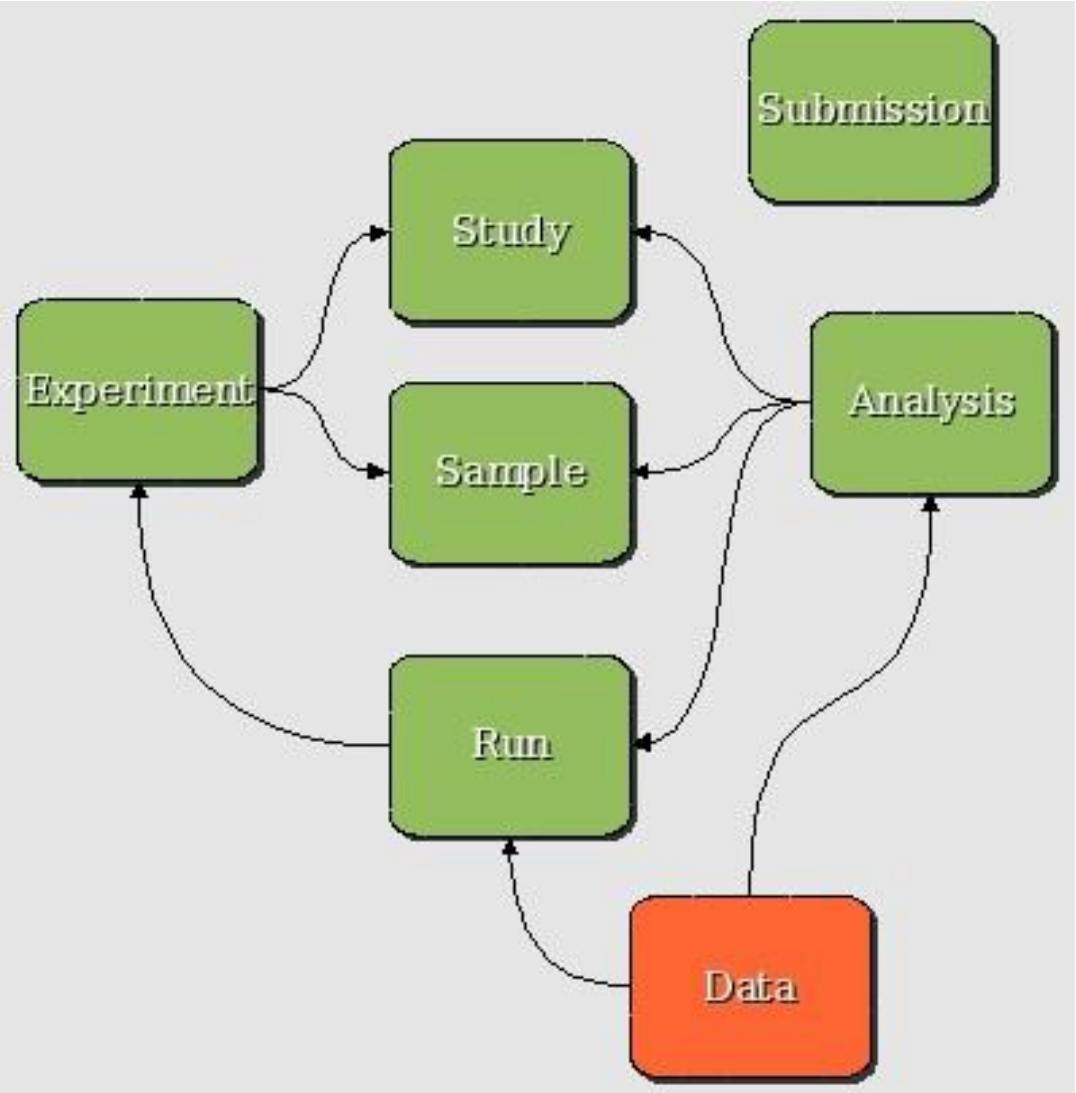


Submission Datenformat

- 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

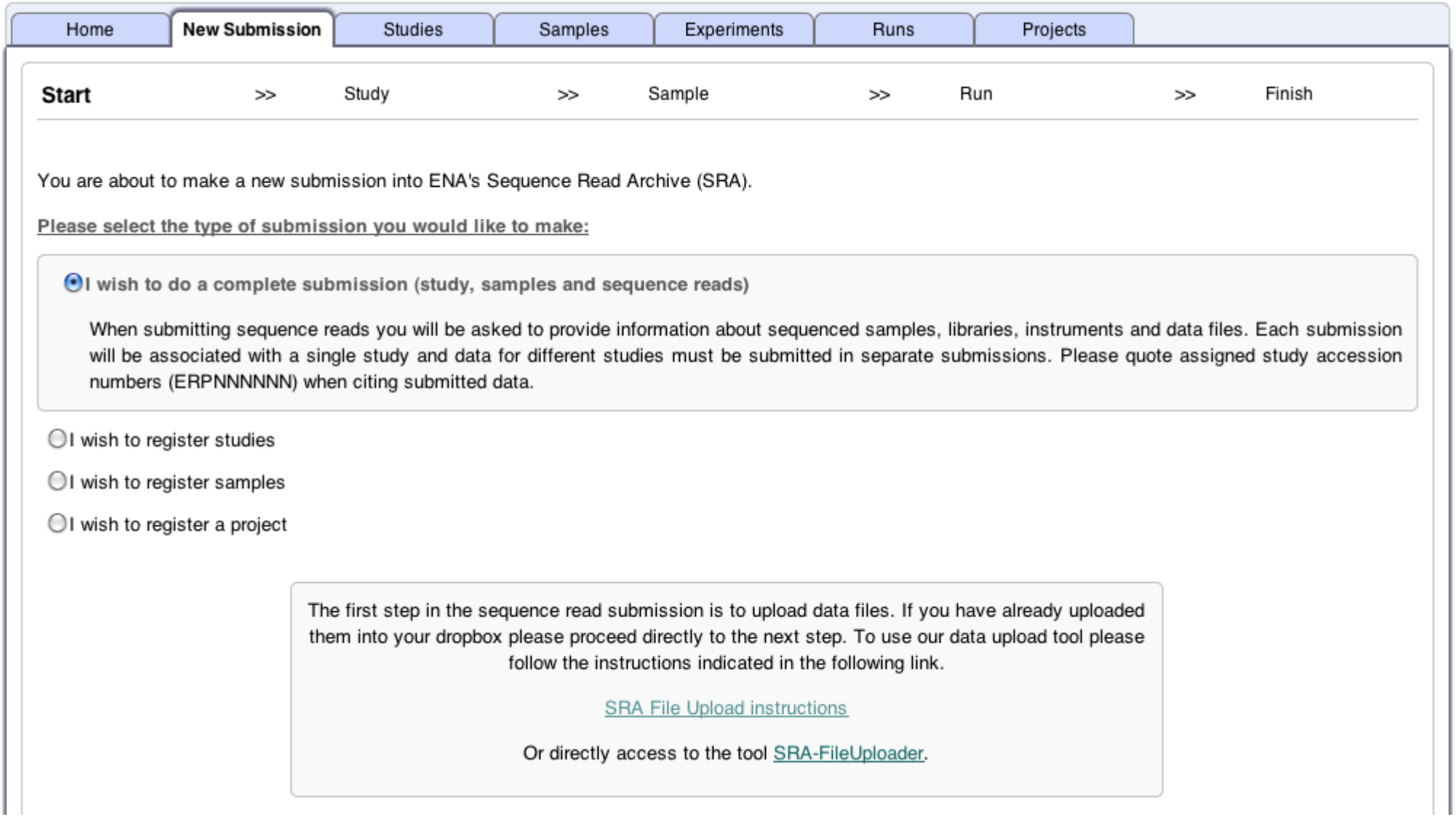
Metadaten Objekte (2)

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

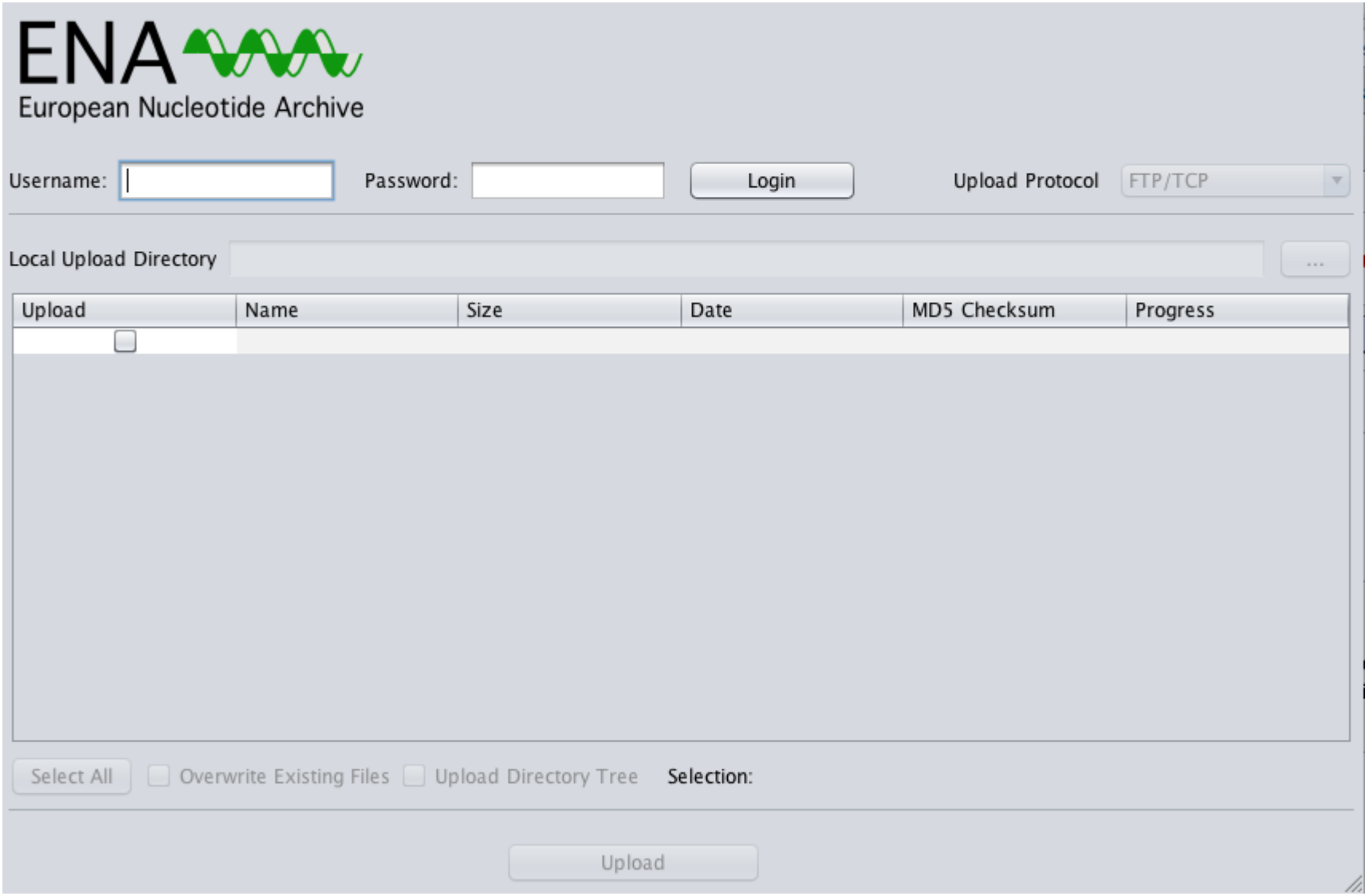
Metadaten Objekte (3)

- Existierenden Study und Sample Objekte
- Eindeutiger Name für jeden Objekt
- Jeder Objekt bekommt Zugriffnummer
  - Submission: ERANNNNNN
  - Study: ERPNNNNNN
  - Sample: ERSNNNNNN
  - Experiment: ERXNNNNNN
  - Run: ERRNNNNNN
  - Analysis: ERZNNNNNN

SRA Webin: metadaten submission



SRA FileUploader



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	Study Type	Read Count	Base Count
Center for Microbial Ecology at Michigan State University	Metagenomics	2,632	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.

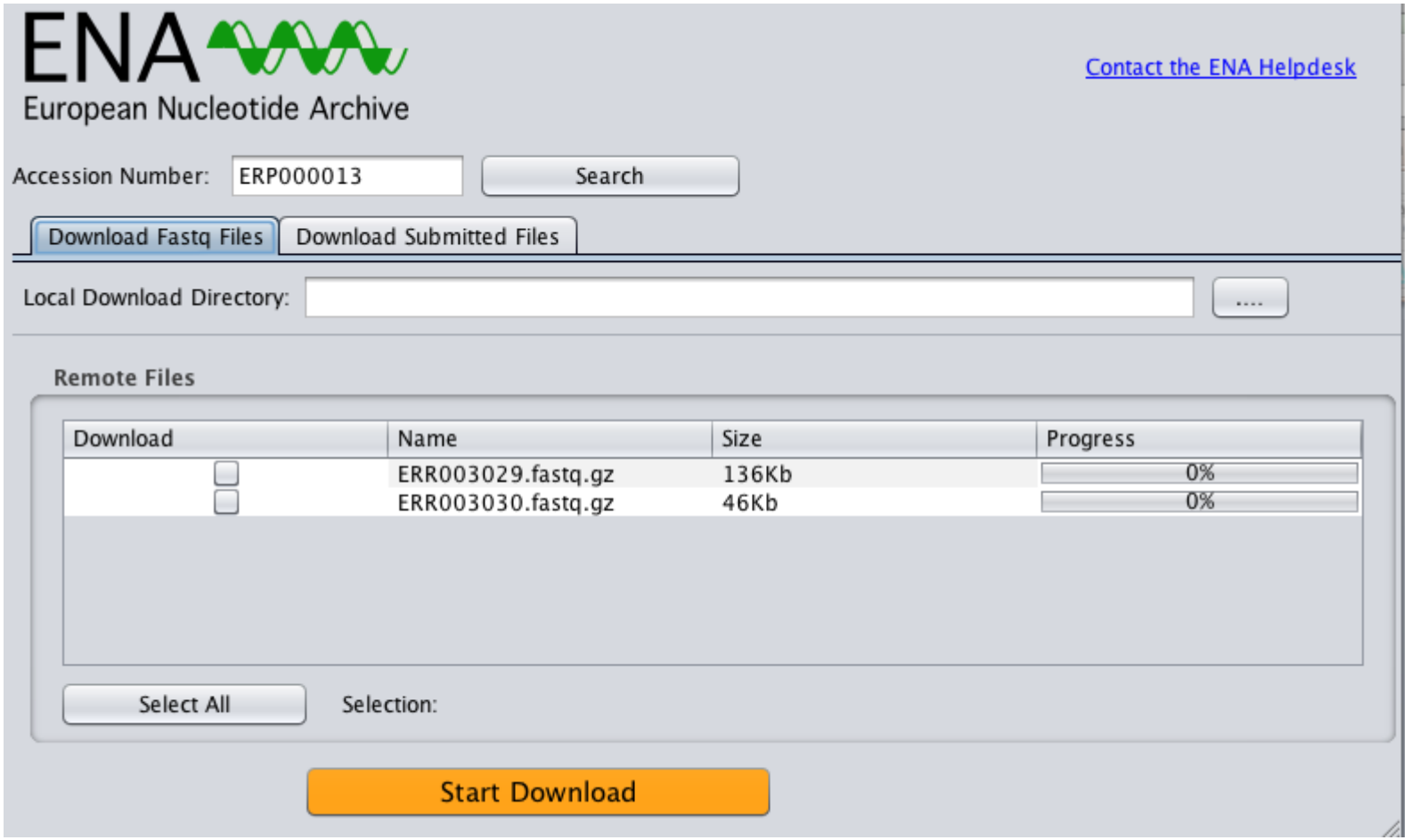
View: TEXT

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



Zusammenfassung

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