

Kapitel 4: Genom-Datenbanken

n Nukleotidsequenz-Datenbanken

- Ausgangsproblematik
- Beispieldatenbanken

n Kartierungs-Datenbanken

- Genomkarten
- Beispieldatenbanken

n Genexpressions-Datenbanken

- Ausgangsproblematik
- Beispieldatenbanken
- Projekt GeWare, Universität Leipzig (E. Rahm et al.): Data warehouse design and implementation to support gene expression analysis



Nukleotidsequenz: Rohdaten

n Daten über den Sequenzierprozess

- Geräterohdaten (Spektren, Sequenzen)
- Benutzte Programme
- Labordaten (Maschinen, Personal, Datum, ...)

n NCBI Trace File Archive

n Viele Sequenzier-Center

- Sanger
- University of Washington
- Celera
- ...



Sequenzdaten

- n Technische Herkunft: Wer, wann, wie, Methode, ...
- n Biologische Herkunft: Clone, Organismus, Linie, ...
- n Literaturreferenzen
- n Fehlerraten
- n Sequenz als Kerninformation
- n Informationen (Features) zu Sequenzteilen
 - Location: Start -Ende, Genau -Ungenau
 - Key: CDS (Coding Sequence(s)) , Repeat, RNA-Strukturen, homologe Sequenzen, Marker, Exon/ Intron Boundaries, Funktion, Motiv, Polymorphismus, ...
 - Qualifier: Ergänzungen, z.B. kodiertes Protein, Regulationsmechanismen, ...



Nukleotidsequenz-Datenbanken: Beispiel-Datenbanken

- n European Molecular Biology Laboratory (EMBL) am European Bioinformatics Institute (EBI)
- n Los Alamos National Laboratory seit 1979; GenBank am NCBI (National Center for Biotech. Information)
- n DNA Data Bank of Japan: 1986; DDBJ am NIG (National Inst. of Genetics)
- n Zusammenschluss in der "International Nucleotide Sequence Database Collaboration" (seit 1988)
 - Täglicher Datenaustausch
 - Lokale Datenbank jeweils verantwortlich für eingebrachte Sequenzen



EMBL-Datenbank

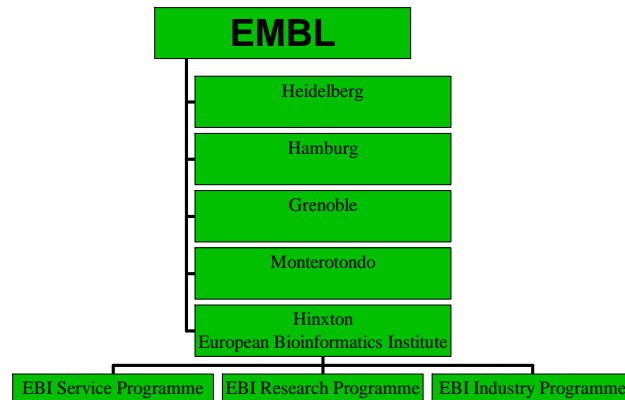
n Erste (seit 1982) und derzeit größte europäische DNA-Sequenzdatenbank (am European Molecular Biology Laboratory in Hinxton, England)*

n Datenquellen

- Lokale Forschergruppen
- Überregionale Sequenzierungsprojekte

n Verfügbarkeit (als vierteljährlich publizierte Releases)

- Flatfile
- SRS (Sequence Retrieval System mit proprietärem EBML-Format)
- XML (BSML = Bioinformatic Sequence Markup Language)
- Oracle Dump Files



* <http://www.ebi.ac.uk/embl/>

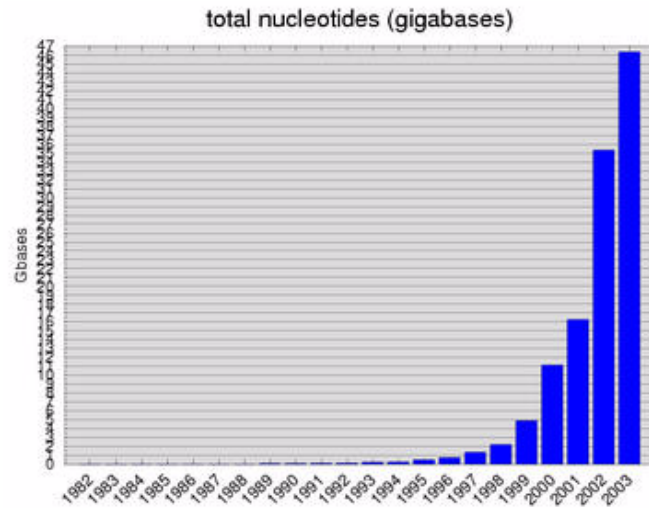


EMBL: Größe

n Release 76 (Sep. 2003)

n Stand Nov. 2003

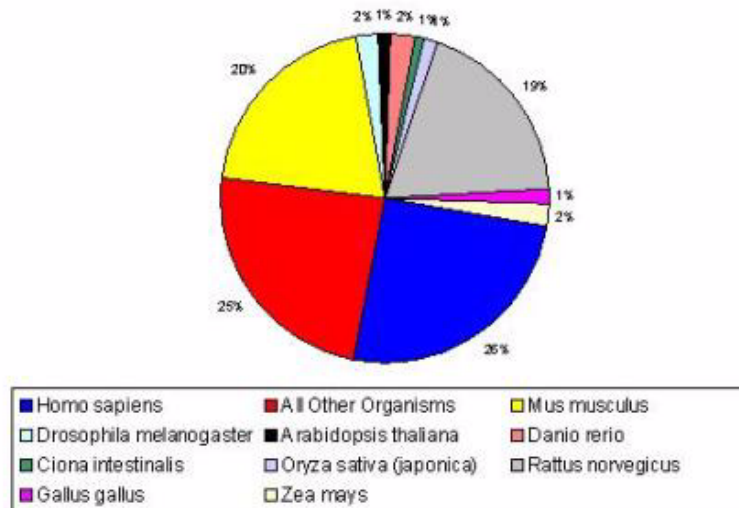
- 46,389,602,205 Basen in
32,049,770 Records
- Über 100 000 Spezies vertreten



<http://www3.ebi.ac.uk/Services/DBStats/> (Gigabase = 10^9 Basen)



EMBL: Spezies (Verteilung)



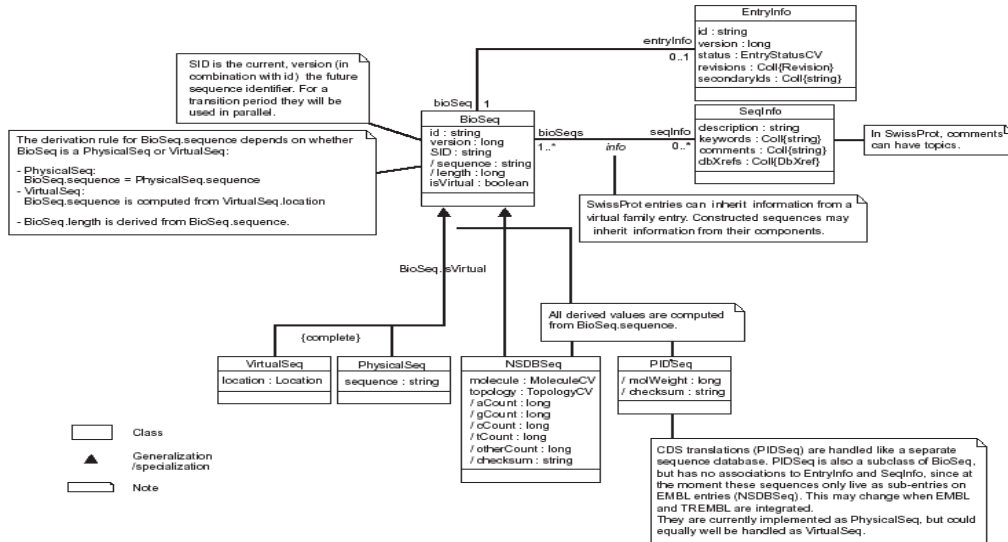
EMBL: Spezies (Beispiele)

The screenshot shows the EMBL Genomes Pages interface. The browser address bar displays the URL: <http://www.ebi.ac.uk/ftp-bin/genomes.org/genomes=viruses>. The page title is "Completed Genomes VIRUSES". A table lists 19 virus genomes with columns for No., Description, Seq length (nt), Genome, and Proteins. A sidebar on the left provides navigation options for different biological categories and links.

No.	Description	Seq length (nt)	Genome	Proteins
1	AKV murine leukemia virus	8,374	J01998	SRS FastA
2	Abelson murine leukemia virus	5,894	J02009	SRS FastA
3	Abelson murine leukemia virus	5,894	AF033812	SRS FastA
4a	Abutilon mosaic virus subgenome DNA A	2,629	X15983	SRS FastA
4b	Abutilon mosaic virus subgenome DNA B	2,585	X15984	SRS FastA
5	Aconitum latent virus	8,657	AB051848	SRS FastA
6	Acute bee paralysis virus	9,491	AF150528	SRS FastA
7	Adeno-associated virus 1	4,718	AF063497	SRS FastA
8	Adeno-associated virus 2	4,679	AF043303	SRS FastA
9	Adeno-associated virus 2	4,675	J01901	SRS FastA
10	Adeno-associated virus 3	4,726	U48704	SRS FastA
11	Adeno-associated virus 3B	4,722	AF028705	SRS FastA
12	Adeno-associated virus 4	4,767	U89790	SRS FastA
13	Adeno-associated virus 6	4,683	AF028704	SRS FastA
14	Aedes albopictus densovirus	4,176	X74945	SRS FastA
15a	African cassava mosaic virus DNA 1	2,779	J02057	SRS FastA
15b	African cassava mosaic virus DNA 2	2,724	J02058	SRS FastA
16a	African cassava mosaic virus-(Cameroon) component A	2,777	AF112352	SRS FastA
16b	African cassava mosaic virus-(Cameroon) component B	2,726	AF112353	SRS FastA



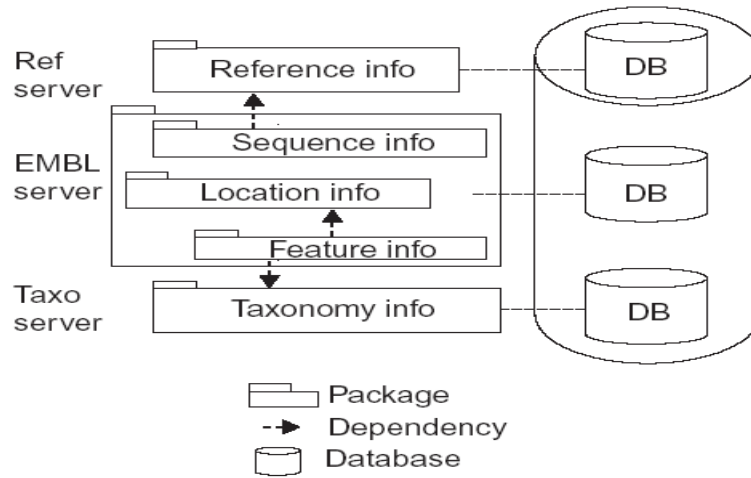
EMBL: UML-Modell (Ausschnitt)



Sequence Info. This package defines class *BioSeq*, which represents biological sequences, and class *SeqInfo*, which describes general information about these sequences. The administrative data associated with database entries are defined in *EntryInfo*. The biological classes of sequence *NSDBSeq*, which is for nucleotide sequences, and *PIDSeq*, which is for protein sequences, are subclasses of *BioSeq*. *VirtualSeq* and *PhysicalSeq* are storage classes of sequence, that is, virtual or literal.



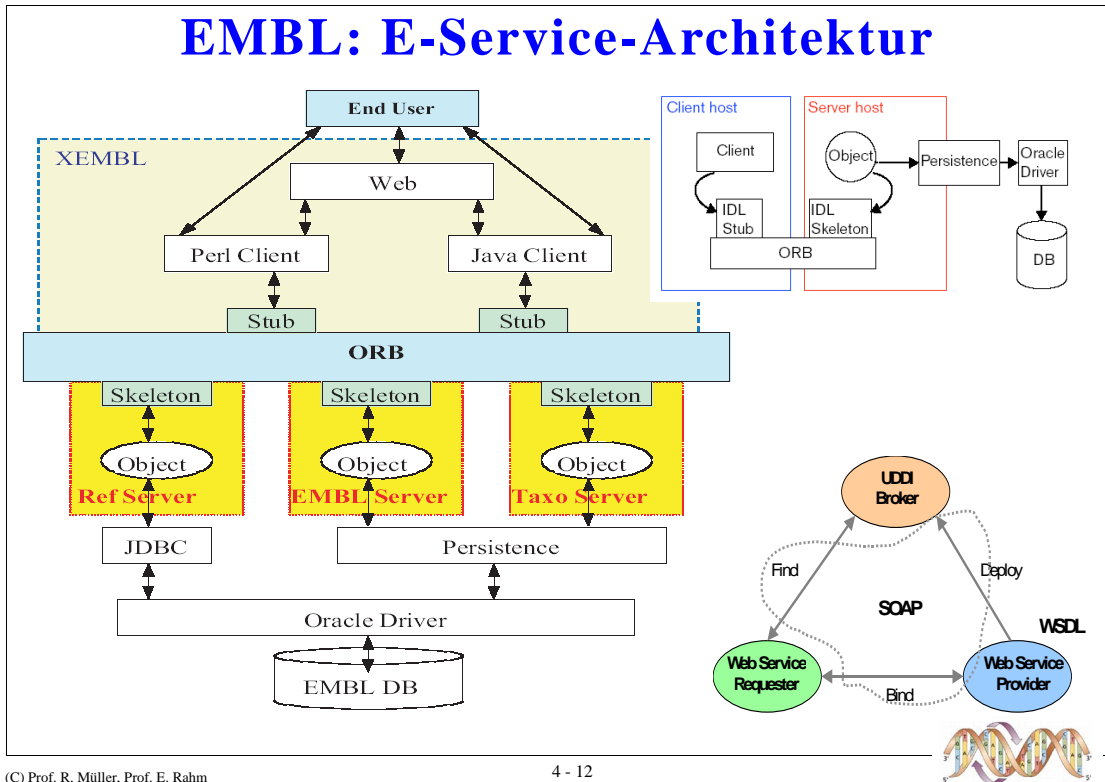
EMBL: Archi- tektur



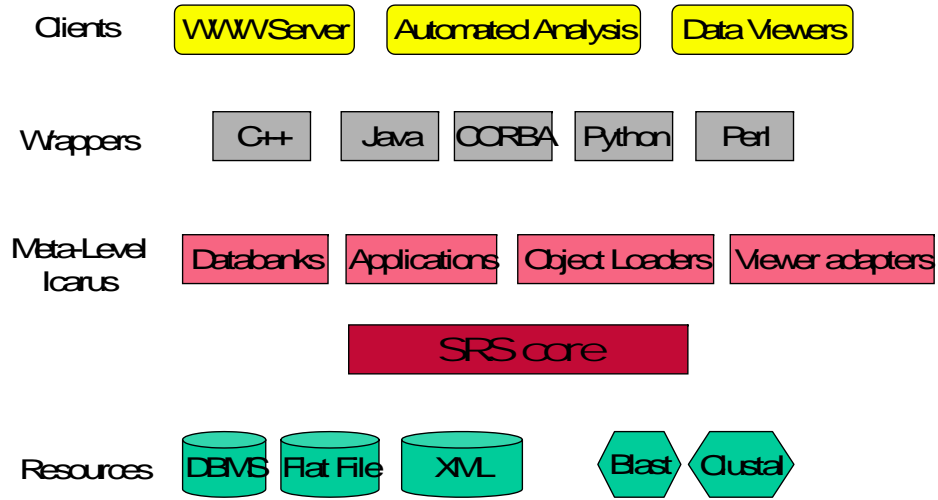
The database partitioning. The database is divided into five main packages: *Sequence Info*, all general information about sequences; *Feature Info*, detailed sequence annotation; *Reference Info*, bibliographic references; *Taxonomy Info*, the taxonomy of the organisms from which the sequences were obtained; *Location Info*, representing locations on sequences.



EMBL: E-Service-Architektur



EMBL: Sequence Retrieval System



GenBank

n NCBI-Datenbank

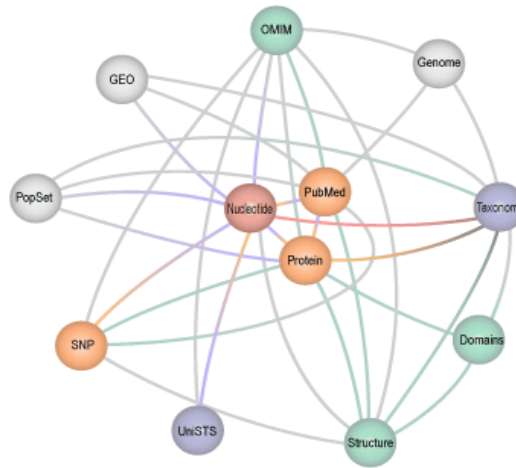
n Derzeit (Nov. 2003) über $20 * 10^9$ Basen

n Modell in ASN.1

n Zugriff über "Entrez"

- Ähnlich SRS bei EBML
- Keine Joins
- "Neighbours" - "Related Documents"
- Click-And-Browse

Entrez is the text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others.



GenBank: Beispieleintrag

```
LOCUS       AE009950                1908256 bp    DNA     circular CON 27-FEB-2002
DEFINITION  Pyrococcus furiosus DSM 3638, complete genome.
ACCESSION   AE009950
VERSION     AE009950.1  GI:18980902
KEYWORDS    .
SOURCE      Pyrococcus furiosus DSM 3638
  ORGANISM  Pyrococcus furiosus DSM 3638
            Archaea; Euryarchaeota; Thermococci; Thermococcales;
            Thermococcaceae; Pyrococcus.

<<<<< deleted for brevity >>

REFERENCE   4 (bases 1 to 1908256)
AUTHORS     Weiss,R.B.
TITLE       Direct Submission
JOURNAL     Submitted (12-FEB-2002) Human Genetics, University of Utah, 20
            South 2030 East, Salt Lake City, UT 84112, USA
FEATURES    Location/Qualifiers
     source   1..1908256
             /organism="Pyrococcus furiosus DSM 3638"
             /strain="DSM 3638"
             /db_xref="taxon:186497"
CONTIG      join(AE010125.1:1..14559,AE010127.1:61..8666,AE010128.1:21..11327,
AE010129.1:61..8659,AE010130.1:61..8716,AE010131.1:61..11112,
AE010132.1:61..11093,AE010133.1:61..11664,AE010134.1:61..3717,
AE010135.1:61..13488,AE010136.1:61..6244,AE010137.1:61..11952,
AE010138.1:61..10516,AE010139.1:61..10851,AE010140.1:61..14818,

<<<<< deleted for brevity >>

AE010288.1:61..12641,AE010289.1:61..11338,AE010290.1:61..11204,
AE010291.1:61..11397,AE010292.1:61..13064,AE010293.1:61..9294,
AE010294.1:61..12888,AE010295.1:61..10029,AE010296.1:61..11091,
AE010297.1:61..13483,AE010298.1:61..2120)
//
```

Figure 2: A GenBank CON entry for a complete bacterial genome. The information toward the bottom of the record describes how to generate the complete genome from the pieces.



Weitere Nukleotid-Datenbanken

- n UniGene, dbEST, RZPD, ...
- n Vielzahl von Datenbanken für spezifische Aspekte
 - Organismen (Hefe, Fliege, Maus, HIV, ...)
 - Ribosomen, Immunsystem
 - Motifs: Transkriptionsfaktoren, Promotoren, ...
- n Terminologie-Datenbanken
 - GeneOntology (> 7000 Begriffe: Funktion, Prozess, Zelllokation)
 - NCBI TaxonomyDatabase (119000 Organismen)



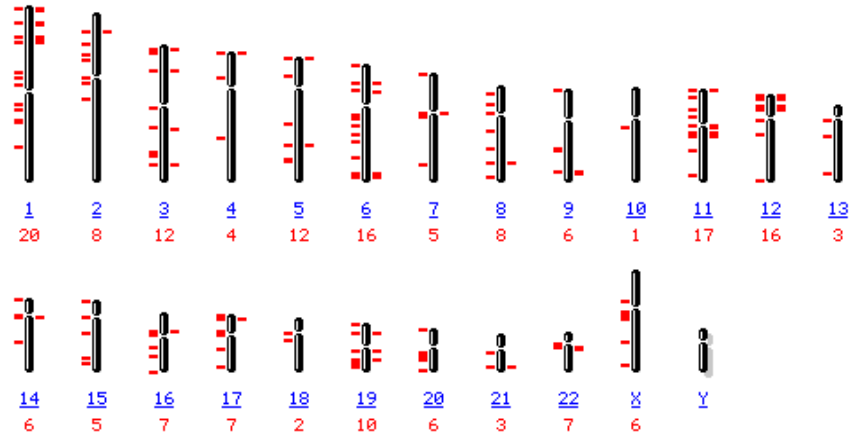
Kartierungs-Datenbanken

- n Motivation
- n The Genome Database (GDB)
- n eGenome
- n LocusLink
- n dbSNP



Motivation

- n Bestimmung der Gen-Loci: Welches Gen liegt an welcher Position (in welchen Modifikationen) auf welchem Chromosom?
- n Medizinische Relevanz: Numerische und strukturelle Chromosomen-Abberationen, Lokalisation von medizinisch relevanten Punktmutationen

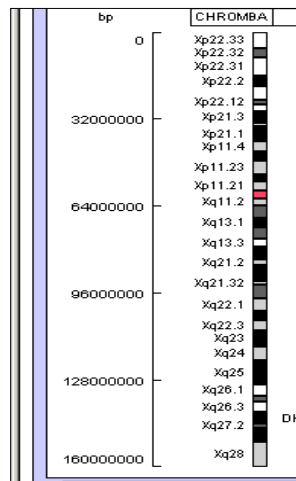


Gen-Loci

n Gen-Locus: Ein "Ort" auf einem Chromosom

n Enthält z.B.

- Gene oder Genfragmente
- DNA-Marker (Eindeutige Gen-identifizierende Sequenzen mit durchschnittlicher Länge von 300-500 Basenpaare)
- Polymorphe Strukturen (Unterschiedliche Allele vorhanden)



Telomer

p – der kurze Arm

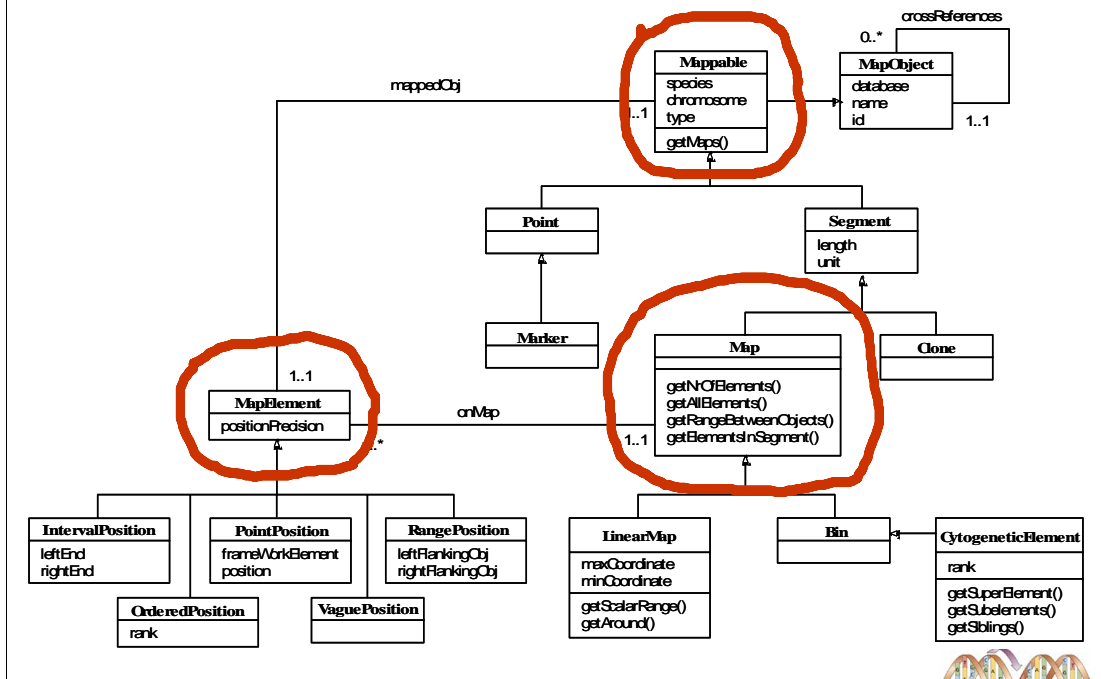
Centromere

q – der lange Arm

Telomer



OMG Standard für Genome Maps



Genome Database: GDB

n Jahrelang Standarddatenbank für Kartierungs-Daten des Humane Genome Projects

n Anzahl Objekte

- 14.000 Gene mit Position
- 150.000 DNA-Marker

n Verfahren der Integration

- Submission-based
- Idee der "Community Curation"
- Chromosome Editors

n Implementierung

- OPM, Sybase
- OPM-Datenschema mit ca. 75 Klassen
- Sybase-Implementierung mit ca. 140 Tabellen



GDB: Interface

Customized Search Forms

- [Markers and Genes within a Region](#)
- [Maps within a Region](#)
- [Genes by Name or Symbol](#)

Sequence-Based Search Forms

- [GDB e-PCR](#)
- [GDB e-PCR Database Lookup](#)

Generic Search Forms

- [Amplimers \(PCR Primer pairs\)](#)
- [Genes](#)
- [Maps](#)
- [Clones](#)
- [Journal Articles](#)
- [Other GDB classes...](#)

[GDB Prototype Page](#)

[Example Searches](#)

Sending request to www.gdb.org...

Browsing Options

- [Genetic Diseases by Chromosome](#)
- [Lists of Genes by Chromosome](#)

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	X	Y

- [Lists of Genes by Symbol Name](#)

A	B	C	D	E	F	G	H	I
J	K	L	M	N	O	P	Q	R
S	T	U	V	W	X	Y	Z	

The screenshot shows a web browser window with the following search forms:

- Name:** A text input field.
- Library Address:** A table with columns: Library, Plate location, Plate Row position, Plate Column position, Location Type. A "Digital Profiles" button is to the right.
- Cytogenetic Localization:** A table with columns: Chromosome, Left Marker, Right Marker.
- All Localizations:** A table with columns: Chromosome, Left Marker, Right Marker.
- Nucleic Acid Sequence Links:** A "Related Segments" section with a "Marker" input field.

At the bottom right of the browser window, there is a small graphic of a DNA double helix.

GDB: Bewertung

- n Sehr technisch orientiert
- n Modell ähnlich zu OMG-Standard (OPM)
- n Komplizierte Search-Forms kaum benutzt
- n Community Curation kaum benutzt
- n Relativ langsam



eGenome

- n Kartierungsdatenbank
- n Z.Z. mehr als 135.000 DNA-Marker (Nov. 2003)
- n Technische Realisierung
 - Abspeicherung der Daten in CompDB, einer Oracle-Datenbank
 - Export als Flatfiles verfügbar



eGenome: Beispiel

eGenome QUICK SEARCH ADVANCED SEARCH INFO DATA INDEX HELP HOME
 Site Index What's new Acknowledgements

1pter-1qter
 1 to 245,203,898 bp from 1pter
 0 cR from 1pter
 12,469 markers, 1,303 polymorphisms, 241,277 SNPs, 261 bundles [Map of region](#)

■ - SNP ★ Marker & polymorphism **Bold text** - RH/GL framework element

Markers		Polymorphisms	SNPs	Bundles	Help
Name	Bundle	Status	Sequence position	RH position	Cytolocation
D0S2577	■	Unknown	0.052 Mb		1p36.3
D0S2578	■	Unknown	0.053 Mb		1p36.3
G54113	■	Unknown	0.054 Mb		1p36.3
A071	■	Unknown	0.076 Mb		1p36.3
GDB.229298		Unknown	0.09 Mb		1p36.3
L31440		Unknown	0.123 Mb		1p36.3
WI-4202	■	Unknown	0.13 Mb		1p36.3
sY		Unknown	0.194 Mb		1p36.3
GDB.1318434		Unknown	0.272 Mb		1p36.3
WI-4202	■	Unknown	0.276 Mb		1p36.3
L31440		Unknown	0.491 Mb		1p36.3
GDB.229285		Unknown	0.506 Mb		1p36.3
L28245		Unknown	0.514 Mb		1p36.3
WI-4202	■	Unknown	0.553 Mb		1p36.3
GDB.1318434		Unknown	0.558 Mb		1p36.3
G01859	■	Unknown	0.641 Mb		1p36.3
L28277		Unknown	0.649 Mb		1p36.3
L28245		Unknown	0.651 Mb		1p36.3
RH98513	■	Transcribed	0.658 Mb	1pter to 1qter	1p36.3
stSG144		Unknown	0.658 Mb		1p36.3
GDB.229285		Unknown	0.664 Mb		1p36.3
L31440		Unknown	0.679 Mb		1p36.3
RH37473	■	Transcribed	0.716 Mb		1p36.3
sY		Unknown	0.78 Mb		1p36.3
AL033801	■	Unknown	0.811 Mb		1p36.3

Page 1 of 513
 Records 1 - 25 [Next](#) [Last](#) [All](#)



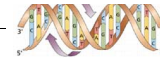
eGenome: Beispiel (2)

RH98513
1p36.3

Map of region

List of region

Position	Description	Clones & Sequences	Help
Sequence position Help			
Base pairs 657,800 to 657,927 from 1pter (UCSC)			
RH map position(s) Help			
1pter to 1qter 0 to 3613.7 cR from 1pter RH positions 1pter to 1qter			
Cytogenetic position(s) Help			
1p36.3 (for sequence 657,800 to 657,927 bp from 1pter) 1pter-1qter (for RH position 0 to 3613.7 cR from 1pter)			
RH score Help			
Genebridge4 1200202010 2210001010 1011111110 0000000000 0100100000 1110000201 1001100012 0000100002 0100011101 111			
RHdb entry Help			
RH98513			
Primer sequences Help			
AAAAAGTCATGGAGGCCATG CTATATGGATGCCCCAC			
Neighboring elements Help			
Elements within <input type="text" value="50"/> kb GO			
Element	Distance	Orientation	
G01853	16,225 bp	pterminal to RH98513	
L28277	8,326 bp	pterminal to RH98513	
L28245	7,034 bp	pterminal to RH98513	
stSG144	509 bp	qterminal to RH98513	
GDB:229285	6,472 bp	qterminal to RH98513	
L31440	20,780 bp	qterminal to RH98513	



LocusLink [<http://www.ncbi.nlm.nih.gov/LocusLink>]

n Repository von Genen und "some non Genes"

- Vielfältige Informationen über Position hinaus
- Proteine, Funktionen, RNA, Phänotypen,
- 32.000 Gene

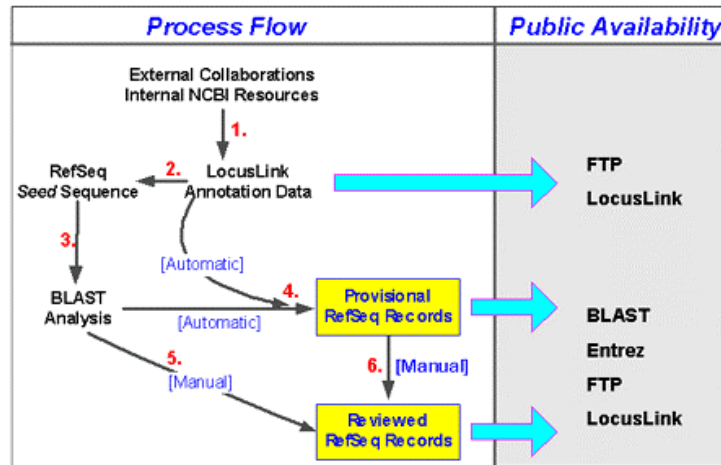
n Technische Implementierung

- NCBI: Entrez Search Interface
- Tab-delimited Files



LocusLink: Integrationsworkflow

- n Mischung aus manueller und automatischer Bearbeitung
- n Objektstatus: Provisional - Reviewed
- n Kein Releasekonzept, keine Versionierung



dbSNP

n Single Nucleotide Polymorphism Database*

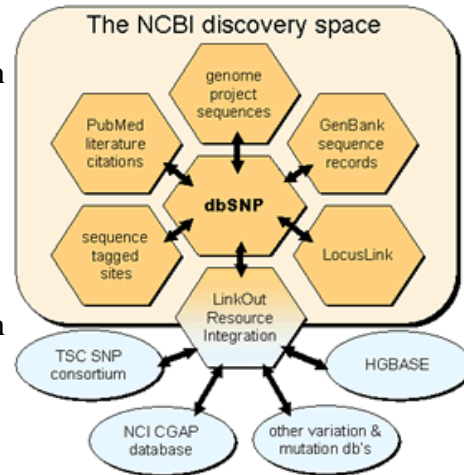
n SNP ("snip"): Kleinste genetische Variation auf dem Level einer einzelnen Base

- Beispiel: Variation des DNA-Segments von AAGGTTA zu ATGGTTA
- Ca. 1.000.000 SNP's im menschlichen Genom
- Viele SNP's ohne phänotypische Auswirkung

n Bestimmte SNPs führen aber zu veränderten Stoffwechselkompetenz ihrer Träger

- Allergische Reaktionen
- Langsamere Abbau von Medikamenten
- Prädisposition für bestimmte Krankheiten

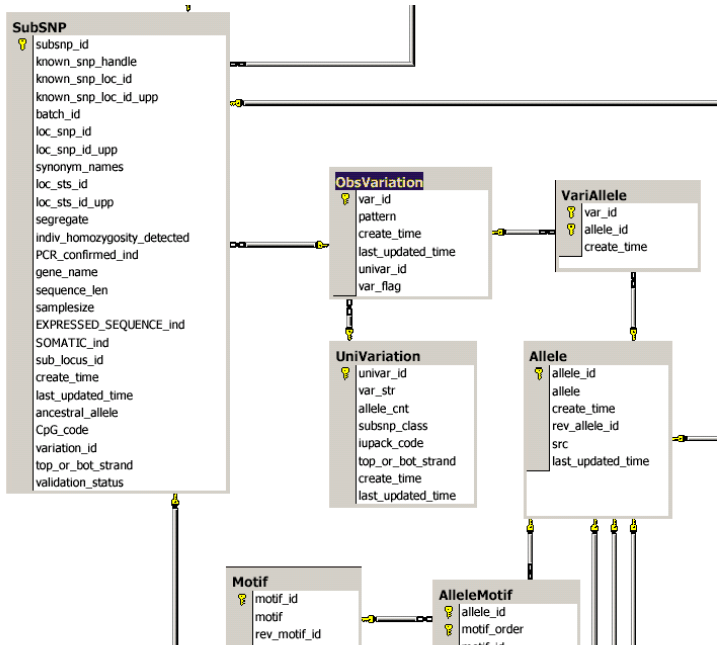
n Zielsetzung von dbSNP: Speicherung aller bekannten "snips", ihrer Genlokalisationen und ggf. medizinischen Relevanz



* <http://www.ncbi.nlm.nih.gov/About/primer/snps.html>



dbSNP: E/R-Modell (Auszug)



dbSNP: Beispiel

The screenshot displays the NCBI dbSNP interface. At the top, the NCBI logo is on the left, and the 'ENTREZ SNP Single Nucleotide Polymorphism' logo is in the center. Below the logo is a navigation bar with tabs for PubMed, Nucleotide, Protein, Genome, Structure, Popset, Taxonomy, and SNP. A search bar contains 'Prostacyclin' and has 'Go' and 'Clear' buttons. Below the search bar are links for Limits, Preview/Index, History, Clipboard, and Details. A control bar shows 'Display: Graphic Summary', 'Show: 20', and 'Sort' options. A pagination bar indicates 'Items 1-20 of 233' and 'Page 1 of 12'. The main content area lists six SNPs, each with a checkbox, ID, species, and a graphic representation of the nucleotide sequence. To the right of each SNP are links for 'LocusLink' and 'Links'. A sidebar on the left contains navigation links for dbSNP BUILD 117, Entrez SNP, Entrez SNP Help, dbSNP, and Entrez Help. A DNA double helix graphic is located in the bottom right corner of the page.

SNP ID	Species	Graphic	Links
<input type="checkbox"/> 1: rs8183919	[Homo sapiens]	20 G A T	LocusLink, Links
<input type="checkbox"/> 2: rs8183608	[Homo sapiens]	20 G A T	LocusLink, Links
<input type="checkbox"/> 3: rs8125371	[Homo sapiens]	20 G A T	LocusLink, Links
<input type="checkbox"/> 4: rs8121749	[Homo sapiens]	20 G A T	LocusLink, Links
<input type="checkbox"/> 5: rs8121473	[Homo sapiens]	20 G A T	LocusLink, Links
<input type="checkbox"/> 6: rs8121008	[Homo sapiens]	20 G A T	LocusLink, Links

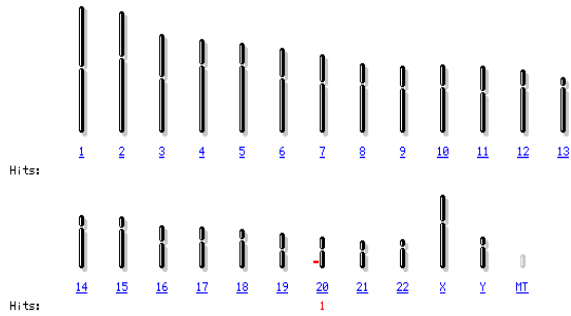
dbSNP: Beispiel (2)

Search for on chromosome(s) assembly
 Show linked entries Help FTP MapViewer home Advanced search

Homo sapiens genome view

[BLAST search the human genome](#)

build 34 version 1 statistics



Search results for query "rs8183919": 1 hit

Chr	Match	Map element	Type	Maps
20	rs8183919	rs8183919	SNP	Variation



dbSNP: Beispiel (3)

Region Shown:

48,860,726
48,861,313

default
 master

[Download/View Sequence/Evidence](#)

Variations Labeled: 6 Total Variations in Region: 6

Variation	Map	Gene	Het	Validation	Genotypes Avail	Linkout Avail
rs6019897		LTC		100%	>80->90->95%	
rs6125670		LTC		100%	>80->90->95%	
rs8183919		LTC		100%	>80->90->95%	
rs6019898		LTC		100%	>80->90->95%	
rs1066894		LTC		100%	>80->90->95%	

Gene Meaning

- L** LOCUS: Any part of the marker position on sequence map is within a 2kb interval 5' of the most 5' feature of gene (CDS, mRNA, gene), OR the marker position is within a 500 base interval 3' of the most 3' feature of the gene. Both strands of sequence are examined for gene features, so a marker can potentially be a variation on multiple genes at a single location.
- T** TRANSCRIPT: Any part of marker position overlaps with mRNA location (or overlaps with UTR/intron and mRNA feature is missing), BUT marker position is not within the coding region of the transcript.
- C** CODING: Any part of the marker position overlaps with a coding sequence (CDS) region (or overlaps with exon region in the unlikely case an exon is annotated but CDS is missing).
- L** The marker is not within the gene locus (as defined above) for any annotated gene.
- T** The marker is not within a transcript region for any annotated gene.
- C** The marker is not within a coding region for any annotated gene.

Mouseover Text
Mouseover on any letter will show the set of gene symbols for each respective functional category

Marker heterozygosity
This 0 – 100% scale indicates the average heterozygosity as the range $avg_hets2[SE(avg_het)]$, where $SE(avg_het)$ is the standard error of the estimator. Thus the graph shows an approximate 95% confidence interval for the marker.

Het Meaning

- No allele frequencies or measures of observed heterozygosity were submitted for this marker.
- Average heterozygosity shown as a 95% confidence interval (0.26 – 0.30). The estimate has a small standard error.
- Average heterozygosity shown as a 95% confidence interval (0.00 – 0.40) The estimate has a large standard error.

