

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

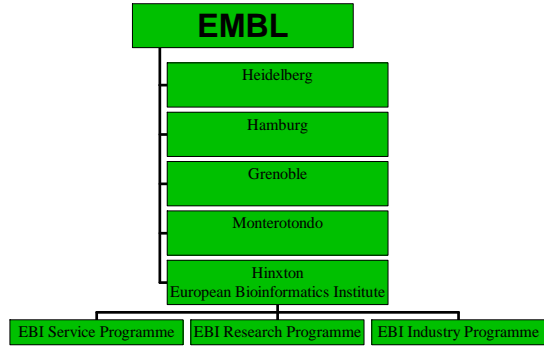
n Erste (seit 1982) und derzeit größte europäische DNA-Sequenzdatenbank (am European Molecular Biology Laboratory in Hinxtun, England)*

n Datenquellen

- Lokale Forschergruppen
- Überregionale Sequenzierungsprojekte

n Verfügbarkeit (als vierteljährlich publizierte Releases)

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



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

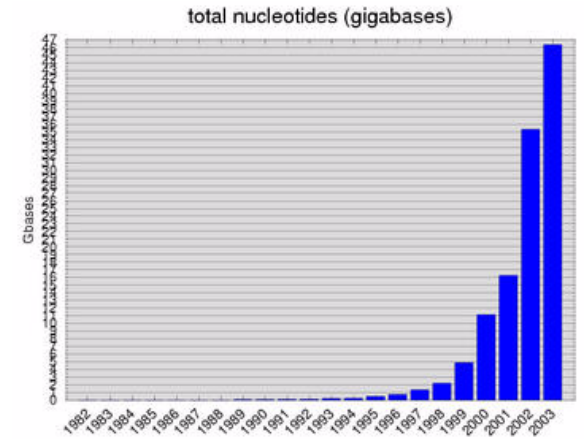


EMBL: Größe

n Release 76 (Sep. 2003)

n Stand Nov. 2003

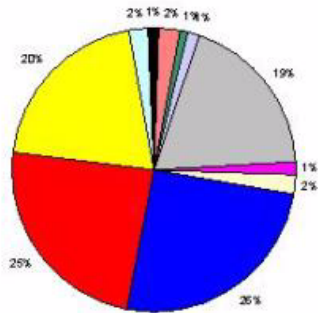
- 46,389,602,205 Basen in 32,049,770 Records
- Über 100 000 Spezies vertreten



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



EMBL: Spezies (Verteilung)



Homo sapiens	A. All Other Organisms	Mus musculus
Drosophila melanogaster	Arabidopsis thaliana	Danio rerio
Ciona intestinalis	Oryza sativa (japonica)	Rattus norvegicus
Gallus gallus	Zea mays	



EMBL: Spezies (Beispiele)

No.	Description	SeqLength (nt)	Genome	Proteins
1	AKV murine leukemia virus	8,374	J01988	SRS, FastA
2	Abelson murine leukemia virus	5,894	J02008	SRS, FastA
3	Abelson murine leukemia virus	5,894	AF030812	SRS, FastA
4a	Abulion mosaic virus subgenome DNA A	2,629	X15983	SRS, FastA
4b	Abulion mosaic virus subgenome DNA B	2,585	X15984	SRS, FastA
5	Acontium latent virus	8,657	AB051848	SRS, FastA
6	Acute bee paralysis virus	9,491	AF150629	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	U08724	SRS, FastA
11	Adeno-associated virus 3B	4,722	AF028705	SRS, FastA
12	Adeno-associated virus 4	4,767	U88790	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	J02007	SRS, FastA
15b	African cassava mosaic virus DNA 2	2,724	J02008	SRS, FastA
16a	African cassava mosaic virus-(Cameroun) component A	2,777	AF112352	SRS, FastA
16b	African cassava mosaic virus-(Cameroun) component B	2,726	AF112353	SRS, FastA



EMBL: Beispieleintrag (Entry-Model)

Global identifier → ID HS185041 standard; RNA; HUM; 1089 BP.

Accession id → AC J00231:

Local identifier & version → NI g185041

Description: free → DT 17-DEC-1994 (Rel. 42, Last updated, Version 6)

Keyword: free → DE Human Ig gamma3 heavy chain disease CHM protein mRNA.

Taxonomy: ctrl! → OC Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;

References: redundant → RN [1]

X-Ref: free → DR GDB: 119339; IGHG3.

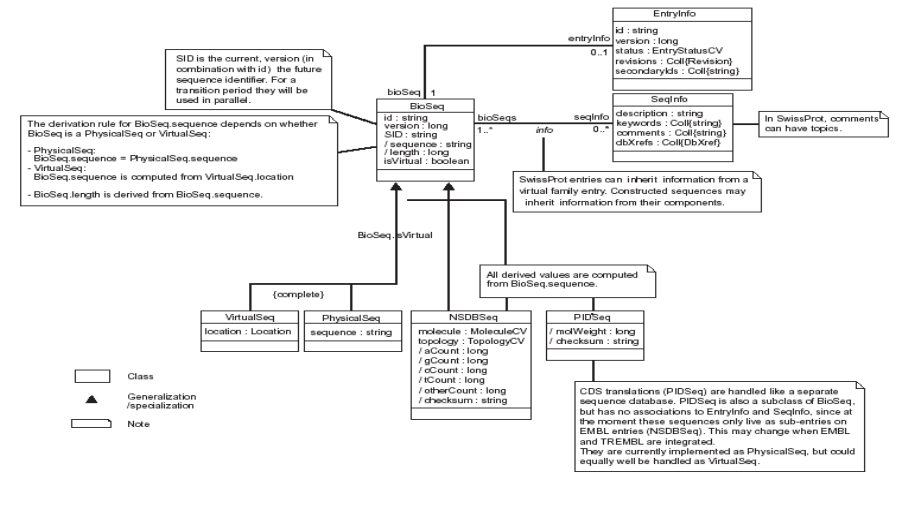
Comment: free → CC The protein isolated from patient CHM is a gamma heavy chain

Feature: partly ctrl → FT CDS 23..964 /codon_start=1

Sequence → SQ
 Sequence 1089 BP: 240 A; 358 C; 271 G; 176 T; 44 other:
 CCGAGCAGTC CTTGTCAGGA AACTGAGAGA NCTGAGTCT TGGCTTTGTC TGGTGGCCG
 TCCAGATGCS GTGCTGTGTC AGGTCAGACT GAGAGAGTCS GCGCCAGAGAC TGCSAGAGCC
 ... 60
 ... 120



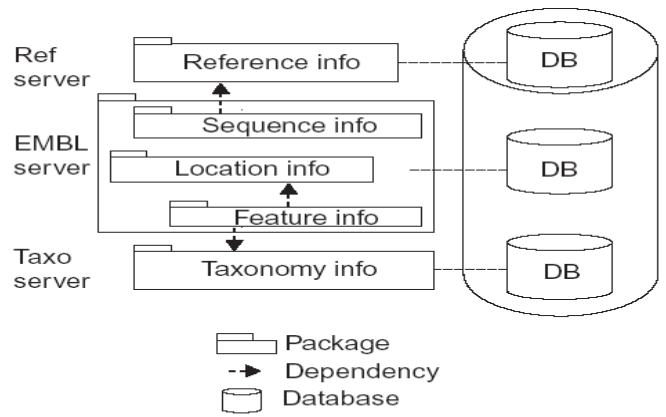
EMBL: UML-Modell (Ausschnitt)



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.



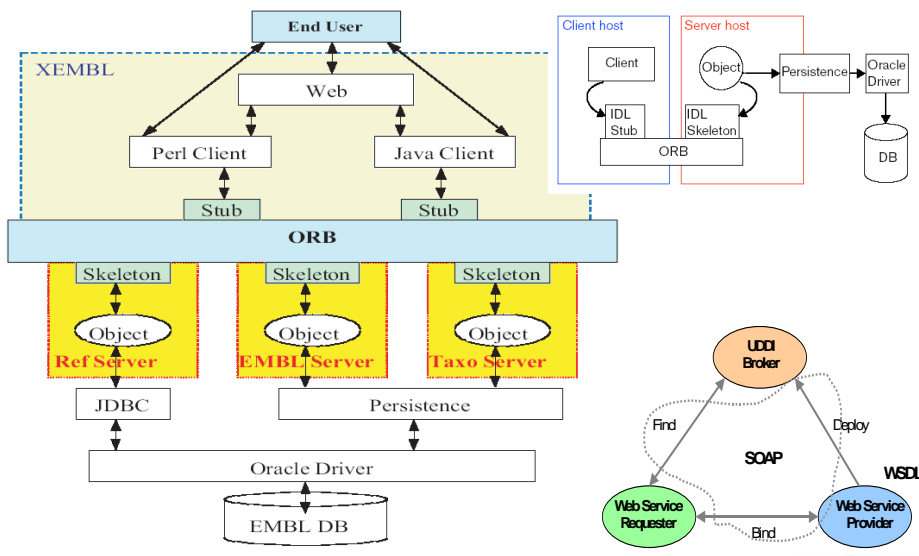
EMBL: Architektur



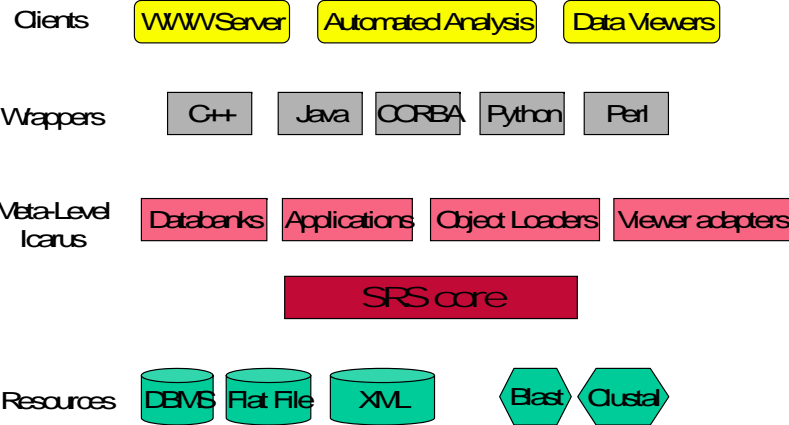
The database partitioning. The database is divided into five main packages: *Sequence Info*, all general information about sequences; *Feature Info*, detailed sequence annotation; *Reference Info*, bibliographic references; *Taxonomy Info*, the taxonomy of the organisms from which the sequences were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



EMBL: E-Service-Architektur



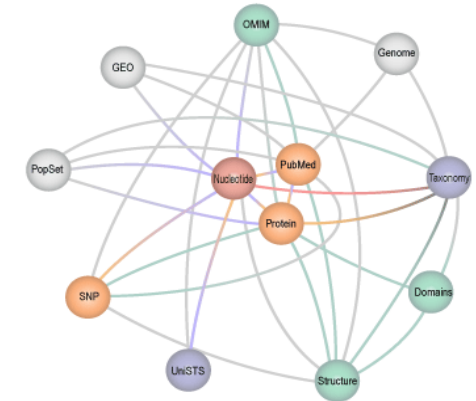
EMBL: Sequence Retrieval System



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: Beispielintrag

```

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)



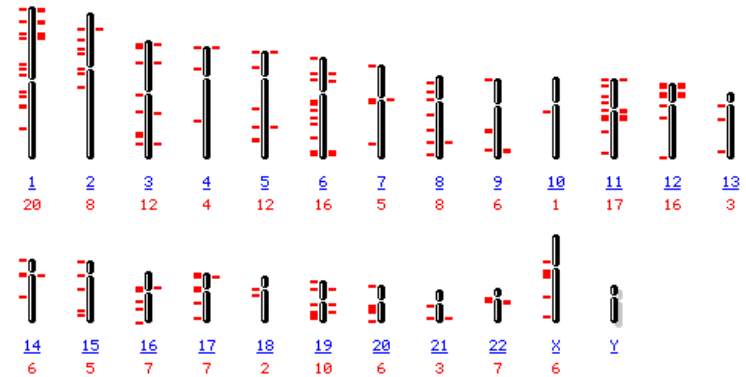
Kartierungs-Datenbanken

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



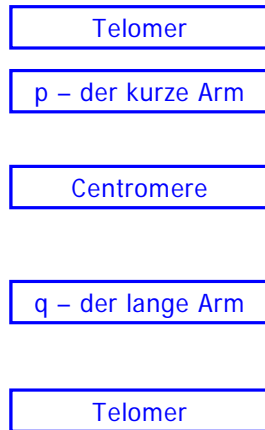
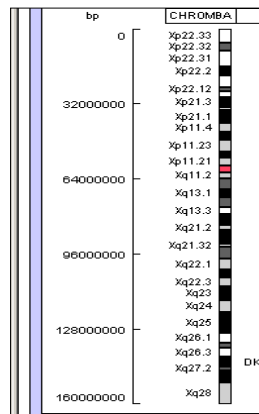
Motivation

- 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

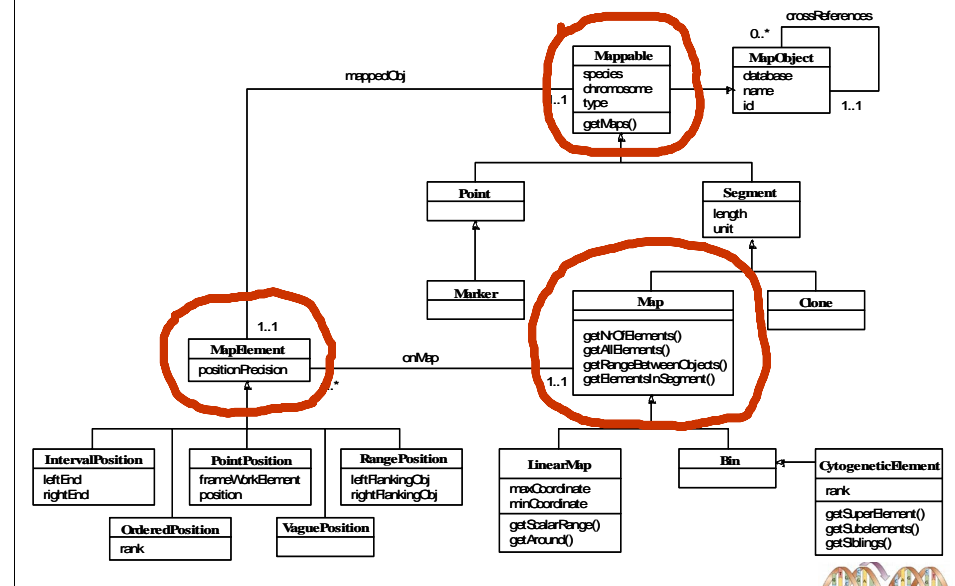


Gen-Loci

- 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)



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

The screenshot shows the GDB web interface with several search categories:

- 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...

On the right, there are "Browsing Options" including "Genetic Diseases by Chromosome", "Lists of Genes by Chromosome" (a grid of chromosome numbers 1-22, X, Y), and "Lists of Genes by Symbol Name" (a grid of letters A-Z).

Below these is a search form with fields for "Name", "Library Address" (Library, Plate location, Plate Row position, Plate Column position, Location Type), "Cytogenetic Localization" (Chromosome, Left Marker, Right Marker), and "All Localizations". There are also "Related Segments" and "Marker" sections.



GDB: Map Viewer

The screenshot shows the GDB Map Viewer interface. On the left, there are navigation controls like "Map Viewer: Help", "Human Maps: Help", "FTP", "Chr. X Resource", "Data As Table View", "Maps & Options", "Region Shown" (131735501 to 148027498), "Go", "out", "zoom", "in", "idcogram", and "master".

The main area displays a "Master Map: STS" for Chromosome 8, with a region of 131M-148M bp. It shows a chromosome map with various markers and a list of STS markers. The list includes:

marker	Kbp	STS	Map	Polymorph
DXS1192	132940			Y
GDB:192503	133228			
DXS52	146535			
DXS7074	146854			
RH12600	147189			
GDB:618098	147279			
DXS7153	147534			
DXS7072	147720			



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

Name	Bundle	Status	Sequence position	RH position	Cytolocation
0052577		Unknown	0.052 Mb		1p36.3
0052578		Unknown	0.053 Mb		1p36.3
054113		Unknown	0.054 Mb		1p36.3
A0711		Unknown	0.076 Mb		1p36.3
0DB_229298		Unknown	0.09 Mb		1p36.3
L31440		Unknown	0.123 Mb		1p36.3
VM-4202		Unknown	0.13 Mb		1p36.3
SV		Unknown	0.194 Mb		1p36.3
GDB_1318434		Unknown	0.272 Mb		1p36.3
VM-4202		Unknown	0.276 Mb		1p36.3
L31440		Unknown	0.491 Mb		1p36.3
0DB_229295		Unknown	0.506 Mb		1p36.3
L28245		Unknown	0.514 Mb		1p36.3
VM-4202		Unknown	0.553 Mb		1p36.3
GDB_1318434		Unknown	0.558 Mb		1p36.3
G01853		Unknown	0.641 Mb		1p36.3
L28277		Unknown	0.649 Mb		1p36.3
L28245		Unknown	0.651 Mb		1p36.3
RH98513		Transcribed	0.650 Mb	1pter to 1qter	1p36.3
RH98513		Unknown	0.650 Mb		1p36.3
RH98514		Unknown	0.658 Mb		1p36.3
0DB_229295		Unknown	0.664 Mb		1p36.3
L31440		Unknown	0.679 Mb		1p36.3
RH37473		Transcribed	0.716 Mb		1p36.3
SV		Unknown	0.78 Mb		1p36.3
AL033601		Unknown	0.811 Mb		1p36.3



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			
Genebridge 1200202010 2210001010 1011111110 0000000000 0100100000 1110000201 1001100012 0000100002 0100011101 111			
RHdb entry Help			
RH98513			
Primer sequences Help			
AAAAAATCATGGAGGCCATG 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	
RH98514	509 bp	qterminal to RH98513	
0DB_229295	6,472 bp	qterminal to RH98513	
L31440	20,780 bp	qterminal to RH98513	



LocusLink [\[http://www.ncbi.nlm.nih.gov/LocusLink\]](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

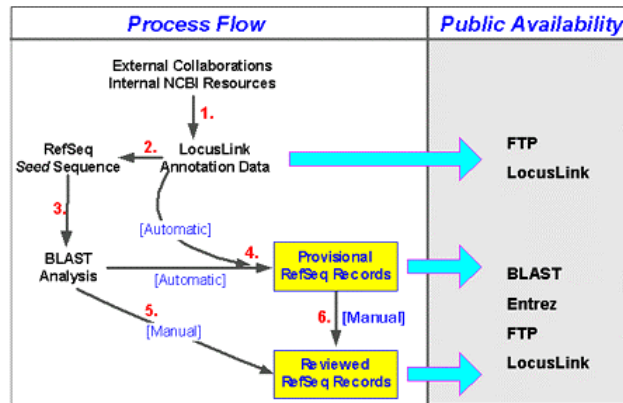
n Technische Implementierung

- NCBI: Entrez Search Interface
- Tab-delimited Files



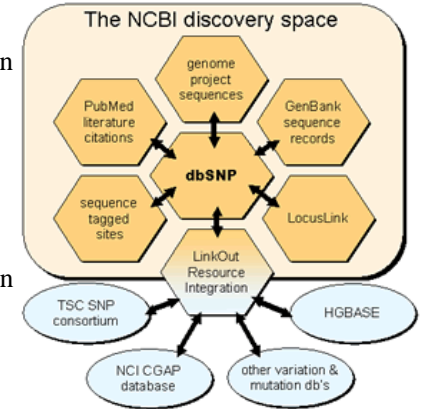
LocusLink: Integrationsworkflow

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

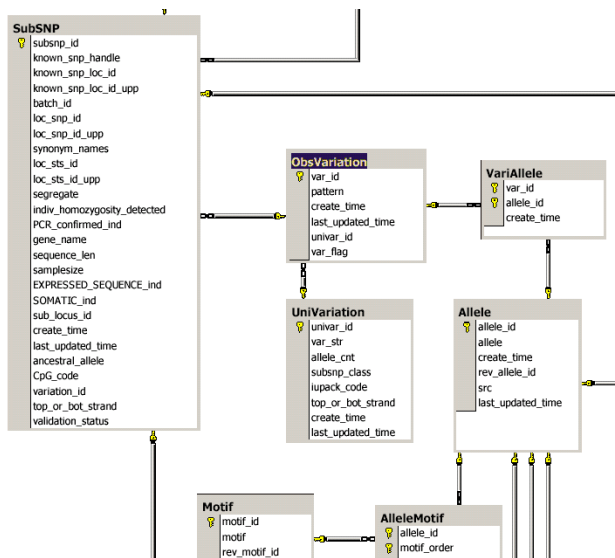


dbSNP

- 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 Genlokalisati-onen und ggf. medizinischen Relevanz



dbSNP: E/R-Modell (Auszug)



dbSNP: Beispiel



