# Kapitel 4: Genom-Datenbanken

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



# Nukleotidsequenz: Rohdaten

- Daten über den Sequenzierprozess
  - Geräterohdaten (Spektren, Sequenzen)
  - Benutzte Programme
  - Labordaten (Maschinen, Personal, Datum, ...)
- NCBI Trace File Archive
- Viele Sequenzier-Center
  - Sanger
  - University of Washington
  - Celera
  - ...



# Sequenzdaten

- Technische Herkunft: Wer, wann, wie, Methode, ...
- Biologische Herkunft: Clone, Organismus, Linie, ...
- Literaturreferenzen
- Fehlerraten
- Sequenz als Kerninformation
- 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, ...

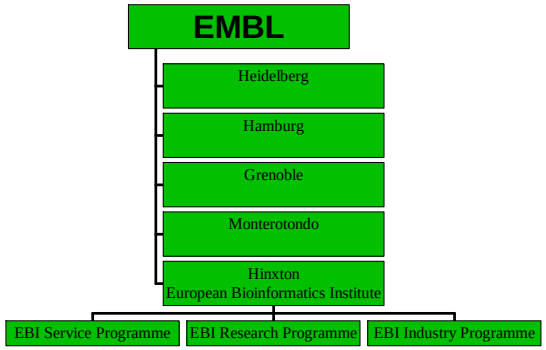


# Nukleotidsequenz-Datenbanken: Beispiel-Datenbanken

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

# EMBL-Datenbank

- Erste (seit 1982) und derzeit größte europäische DNA-Sequenzdatenbank (am European Molecular Biology Laboratory in Hinxton, England)*
- Datenquellen
  - Lokale Forschergruppen
  - Überregionale Sequezierungsprojekte
- Verfügbarkeit (als vierteljährlich publizierte Releases)
  - Flatfile
  - SRS (Sequence Retrieval System mit proprietärem EBML-Format)
  - XML (BSML = Bioinformatic Sequence Markup Language)
  - Oracle Dump Files

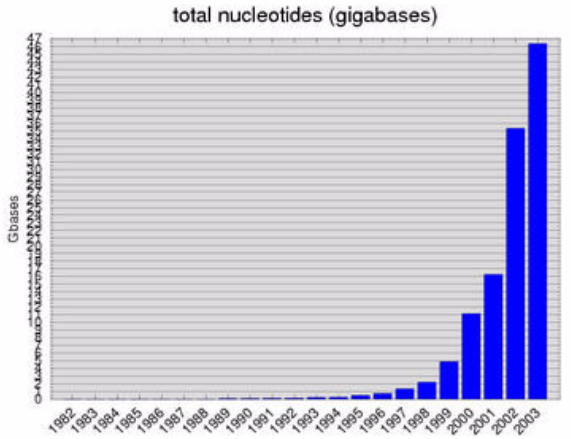



* http://www.ebi.ac.uk/embl/



# EMBL: Größe

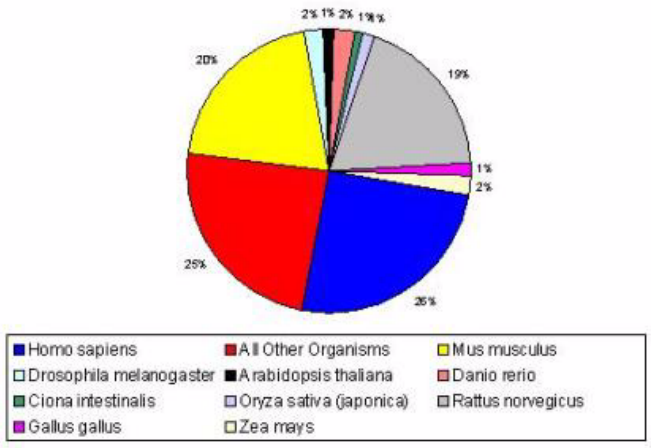
- Release 76 (Sep. 2003)
- Stand Nov. 2003
  - 46,389,602,205 Basen in 32,049,770 Records
  - Über 100 000 Spezies vertreten




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

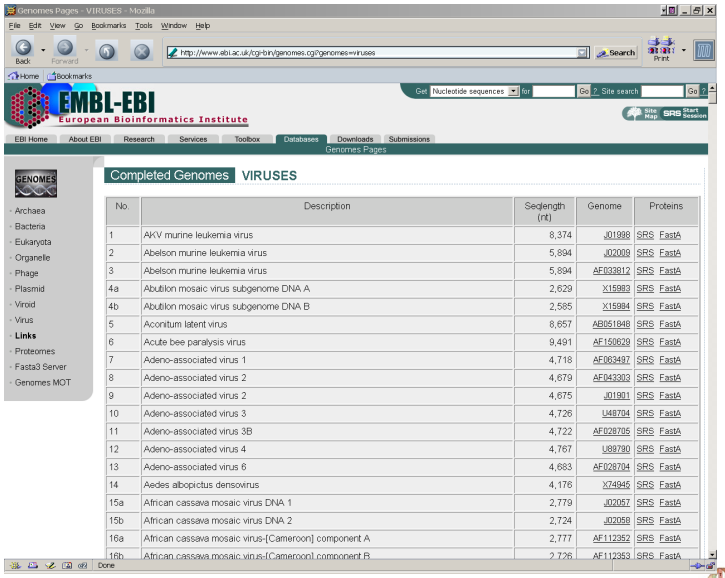


# EMBL: Spezies (Verteilung)






# EMBL: Spezies (Beispiele)



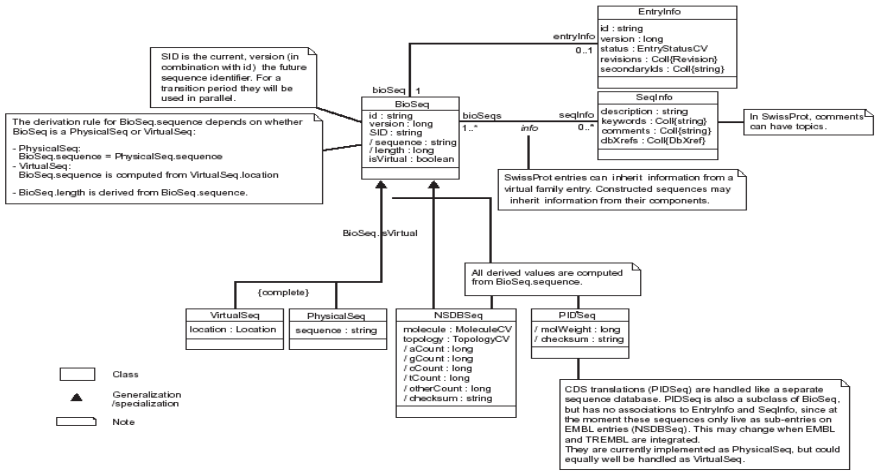
# EMBL: Beispieleintrag (Entry-Model)

Global identifier
Accession id
Local identifier & version
Description: free
Keyword: free
Taxonomy: ctrl !
References: redundant
X-Ref: free
Comment: free
Feature: partly ctrl
Sequence

```
ID   HSIGHAF    standard; RNA; HUM; 1089 BP.
XX
AC   J00231;
XX
NI   g185041
XX
DT   17-DEC-1994 (Rel. 42, Last updated, Version 6)
XX
DE   Human Ig gamma3 heavy chain disease OMM protein mRNA.
XX
KW   C-region; gamma heavy chain disease protein;
XX
OC   Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria; Primates;
XX
RN   [1]
RP   1-1089
RX   MEDLINE; 82247835.
...
DR   GDB; 119339; IGHG3.
DR   GDB; G00-119-339.
...
CC   The protein isolated from patient OMM is a gamma heavy chain
FH
FT   CDS             23. .964
FT                   /codon_start=1
FT                   567112"
XX
SQ   Sequence 1089 BP; 240 A; 358 C; 271 G; 176 T; 44 other;
     CCTGGACCTC CTGTGCAAGA ACATGAAACA NCTGTGGTTC TTCCTTCTCC TGGTGGCAGC        60
     TCCCAGATGG GTCCTGTCCC AGGTGCACCT GCAGGAGTCG GGCCCAGGAC TGGGGAAGCC       120
     ...
```



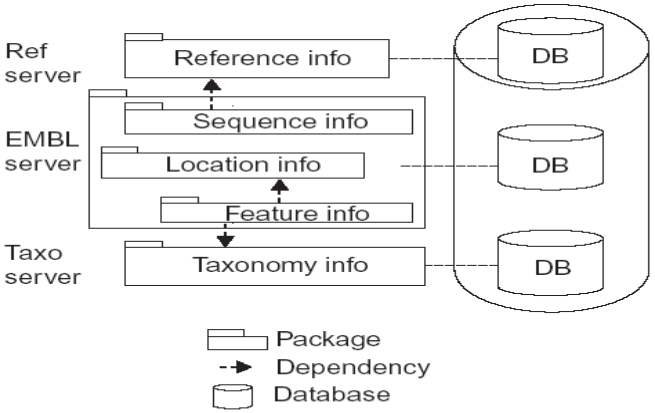
# EMBL: UML-Modell (Ausschnitt)




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.



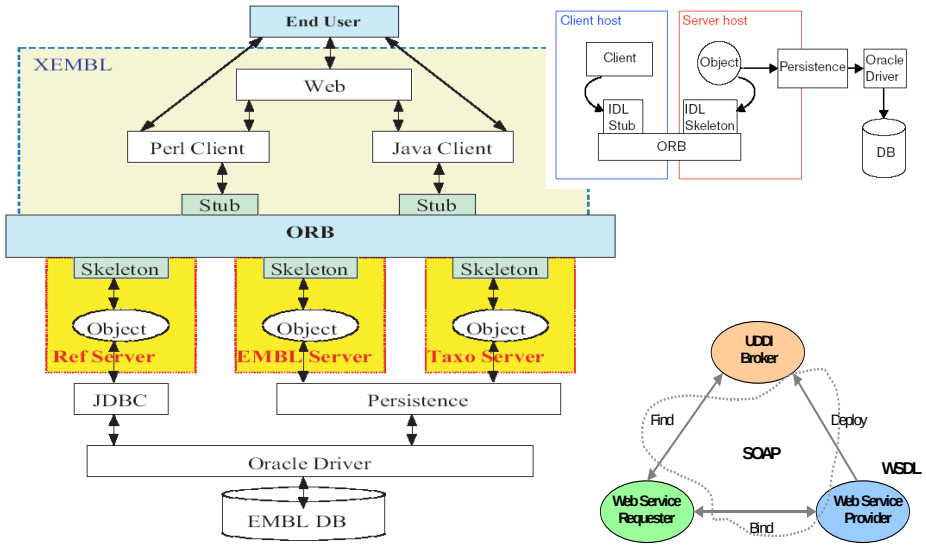
# EMBL: Archi-tektur



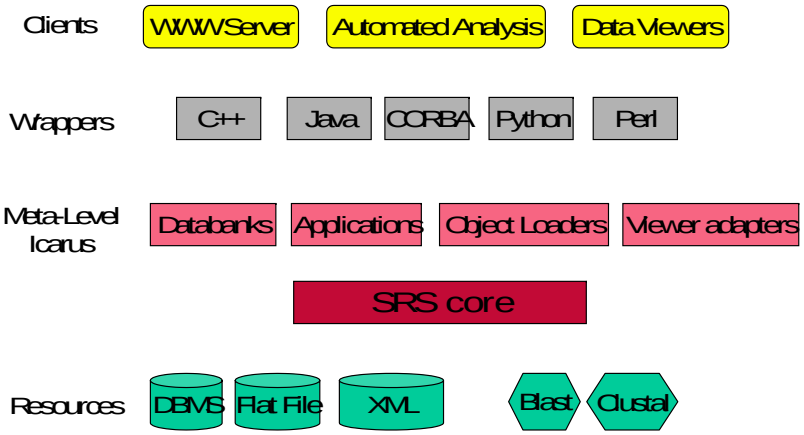

The database partitioning. The database is divided into five main packages: *Sequence Info*, all general information about sequences; *Feature Info*, detailed sequence annotation; *Reference Info*, bibliographic references; *Taxonomy Info*, the taxonomy of the organisms from which the sequences were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



# EMBL: E-Service-Architektur



# EMBL: Sequence Retrieval System






# GenBank

- NCBI-Datenbank
- Derzeit (Nov. 2003) über $20 * 10^9$ Basen
- Modell in ASN.1
- 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(AE010126.1:1..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

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

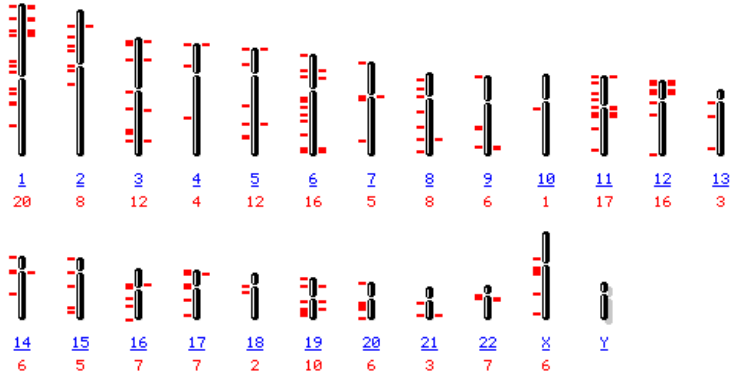
# Kartierungs-Datenbanken

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



# Motivation

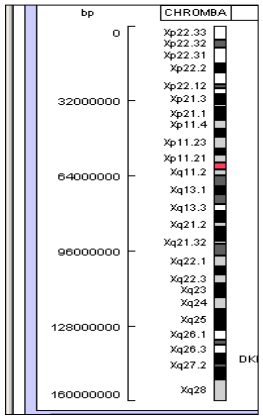
- Bestimmung der Gen-Loci: Welches Gen liegt an welcher Position (in welchen Modifikationen) auf welchem Chromosom?
- Medizinische Relevanz: Numerische und strukturelle Chromosomen-Abberationen, Lokalisation von medizinisch relevanten Punktmutationen






# Gen-Loci

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

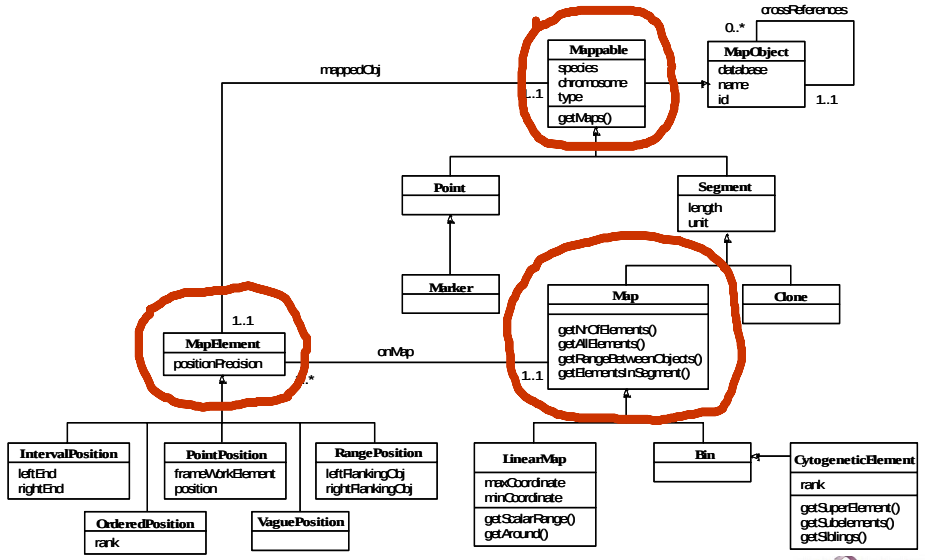
Centromere

q – der lange Arm

Telomer



# OMG Standard für Genome Maps

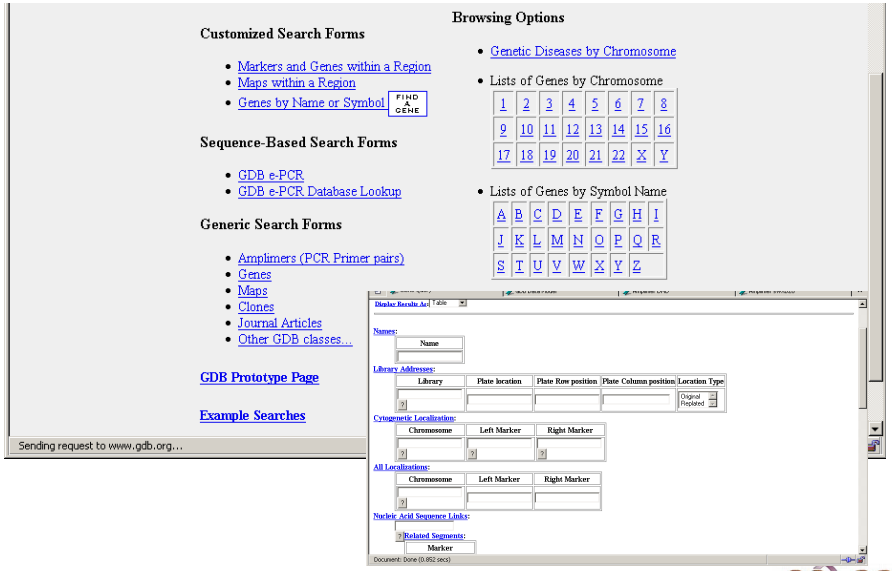


# Genome Database: GDB

- Jahrelang Standarddatenbank für Kartierungs-Daten des Humane Genome Projects
- Anzahl Objekte
  - 14.000 Gene mit Position
  - 150.000 DNA-Marker
- Verfahren der Integration
  - Submission-based
  - Idee der "Community Curation"
  - Chromosome Editors
- Implementierung
  - OPM, Sybase
  - OPM-Datenschema mit ca. 75 Klassen
  - Sybase-Implementierung mit ca. 140 Tabellen

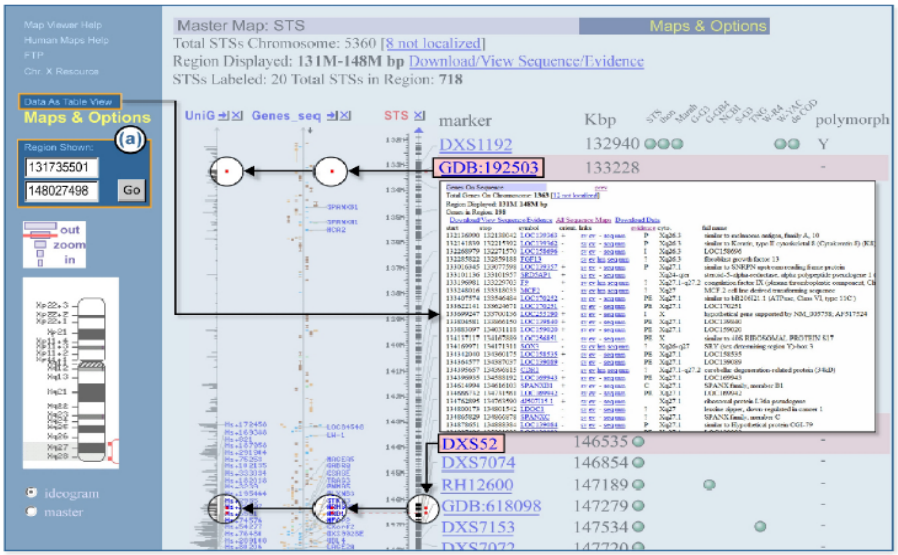


# GDB: Interface






# GDB: MapViewer






# GDB: Bewertung

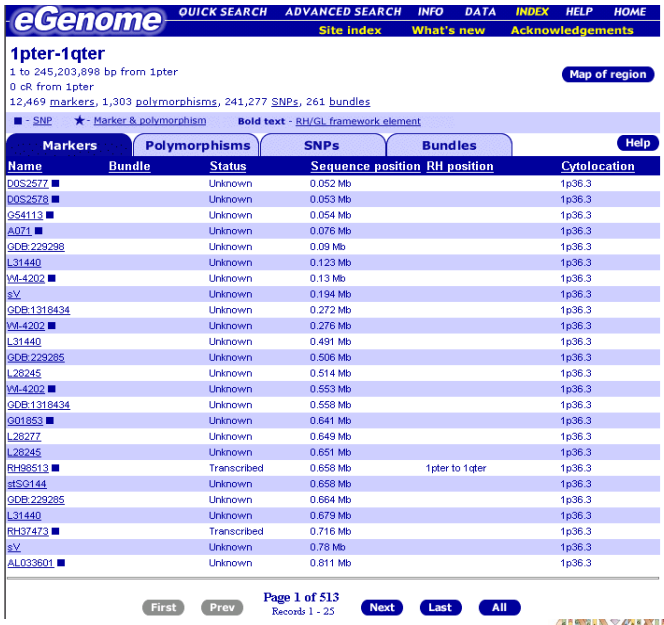
- Sehr technisch orientiert
- Modell ähnlich zu OMG-Standard (OPM)
- Komplizierte Search-Forms kaum benutzt
- Community Curation kaum benutzt
- Relativ langsam

# eGenome

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

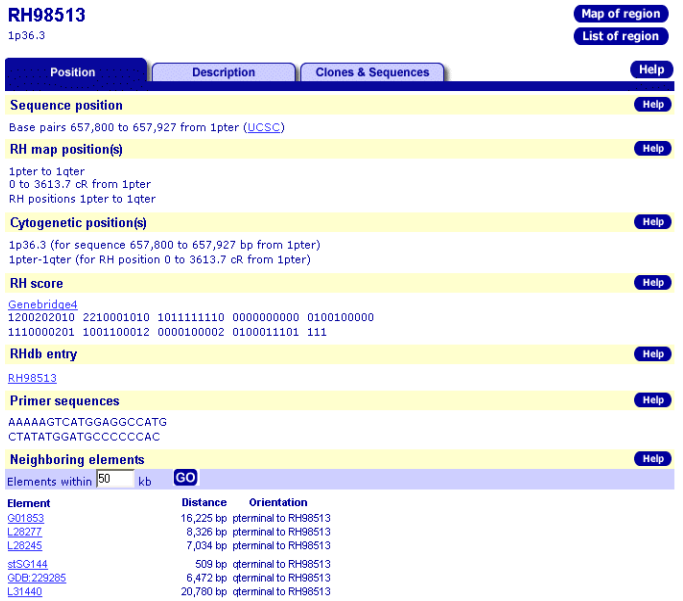



# eGenome: Beispel






# eGenome: Beispiel (2)



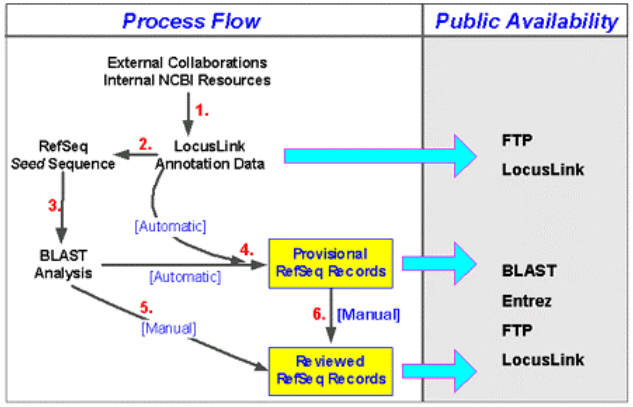



# LocusLink [http://www.ncbi.nlm.nih.gov/LocusLink]

- Repository von Genen und "some non Genes"
  - Vielfältige Informationen über Position hinaus
  - Proteine, Funktionen, RNA, Phänotypen, ....
  - 32.000 Gene
- Technische Implementierung
  - NCBI: Entrez Search Interface
  - Tab-delimited Files

# LocusLink: Integrationsworkflow

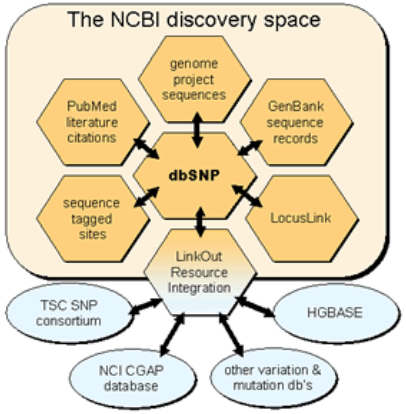
- Mischung aus manueller und automatischer Bearbeitung
- Objektstatus: Provisional - Reviewed
- Kein Releasekonzept, keine Versionierung






# dbSNP

- Single Nucleotide Polymorphism Database*
- 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
- Bestimmte SNPs führen aber zu veränderten Stoffwechselkompetenz ihrer Träger
  - Allergische Reaktionen
  - Langsamerer Abbau von Medikamenten
  - Prädisposition für bestimmte Krankheiten

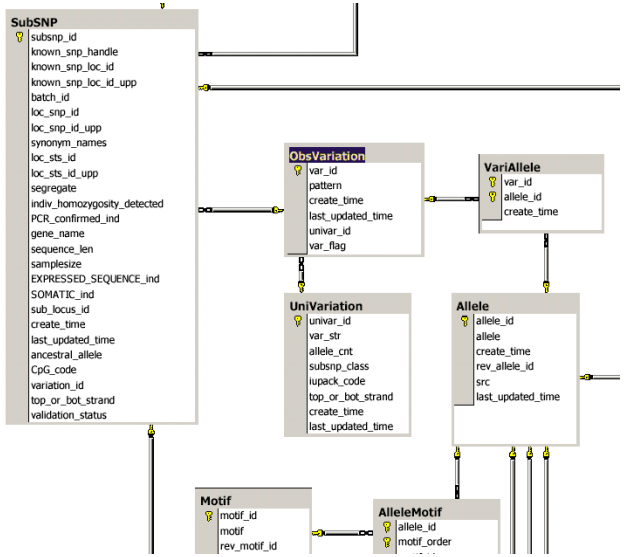



- Zielsetzung von dbSNP: Speicherung aller bekannten "snips", ihrer Genlokalisationen und ggf. medizinischen Relevanz

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

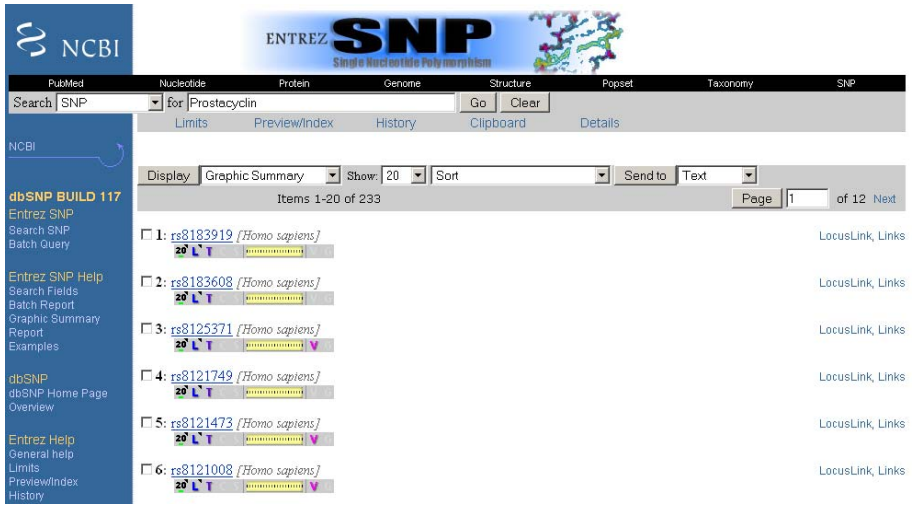


# dbSNP: E/R-Modell (Auszug)

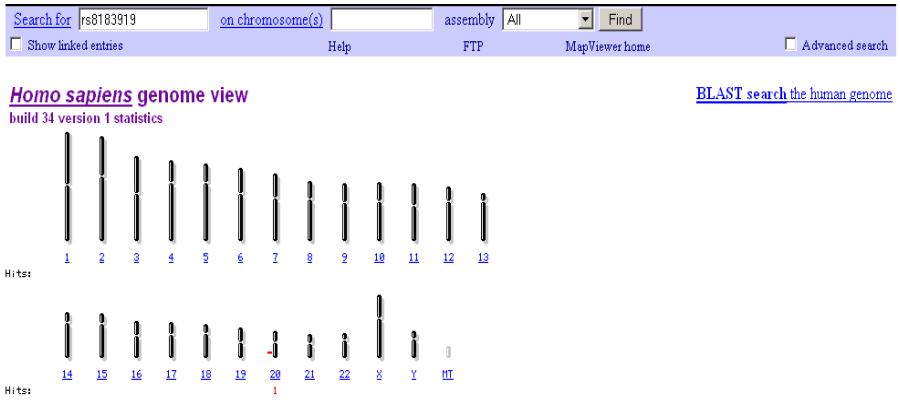





# dbSNP: Beispiel



# dbSNP: Beispiel (2)






# dbSNP: Beispiel (3)