

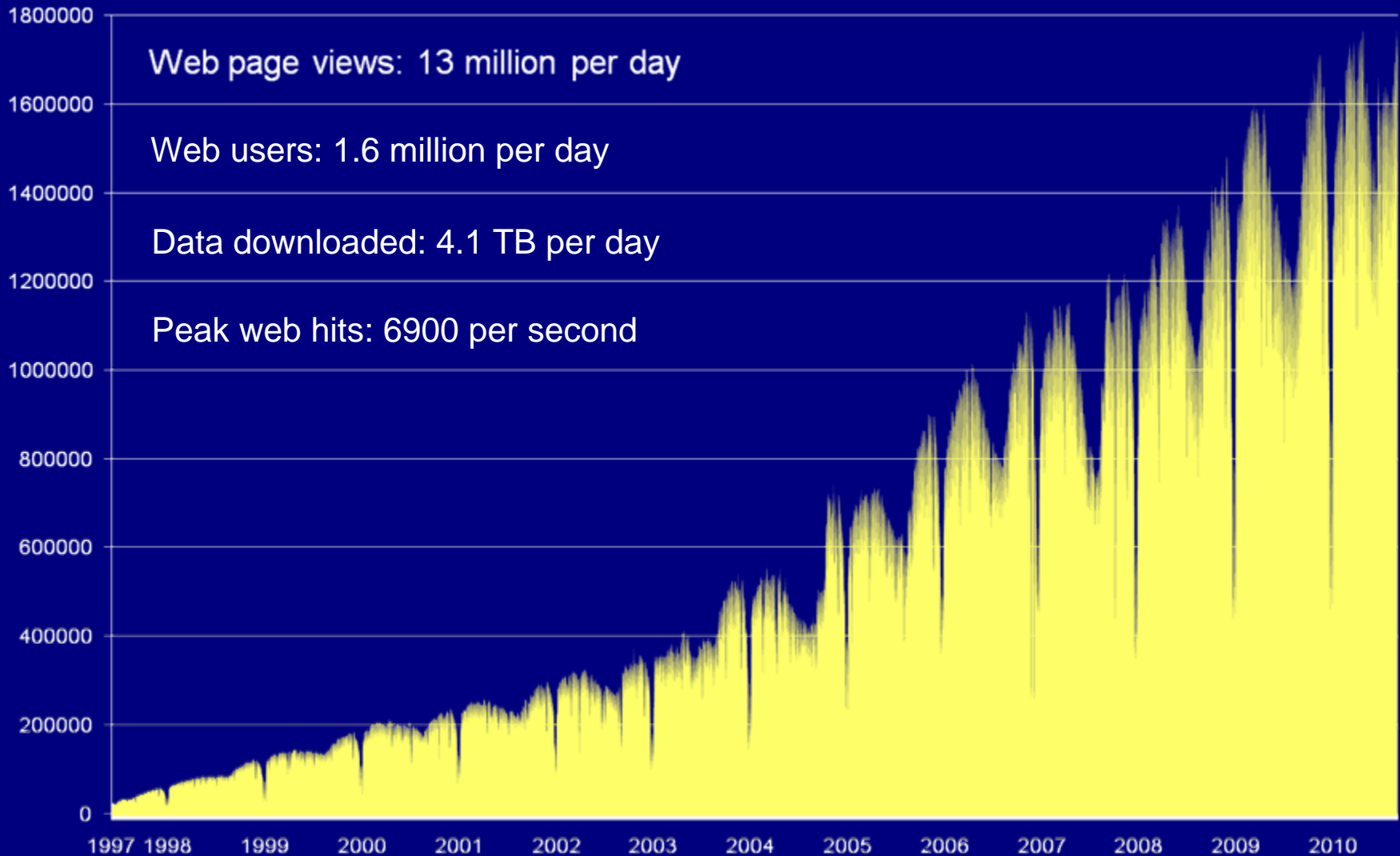
NCBI



National Center for Biotechnology Information

- Created by Public Law 100-607 in 1988 as part of National Library of Medicine at NIH to:
 - Create automated systems for knowledge about molecular biology, biochemistry, and genetics.
 - Perform research into advanced methods of analyzing and interpreting molecular biology data.
 - Enable biotechnology researchers and medical care personnel to use the systems and methods developed.
- Builders and providers of GenBank, Blast, PubMed, dbSNP, dbGaP, RefSeq, and more
- Center for basic research and training in computational biology.

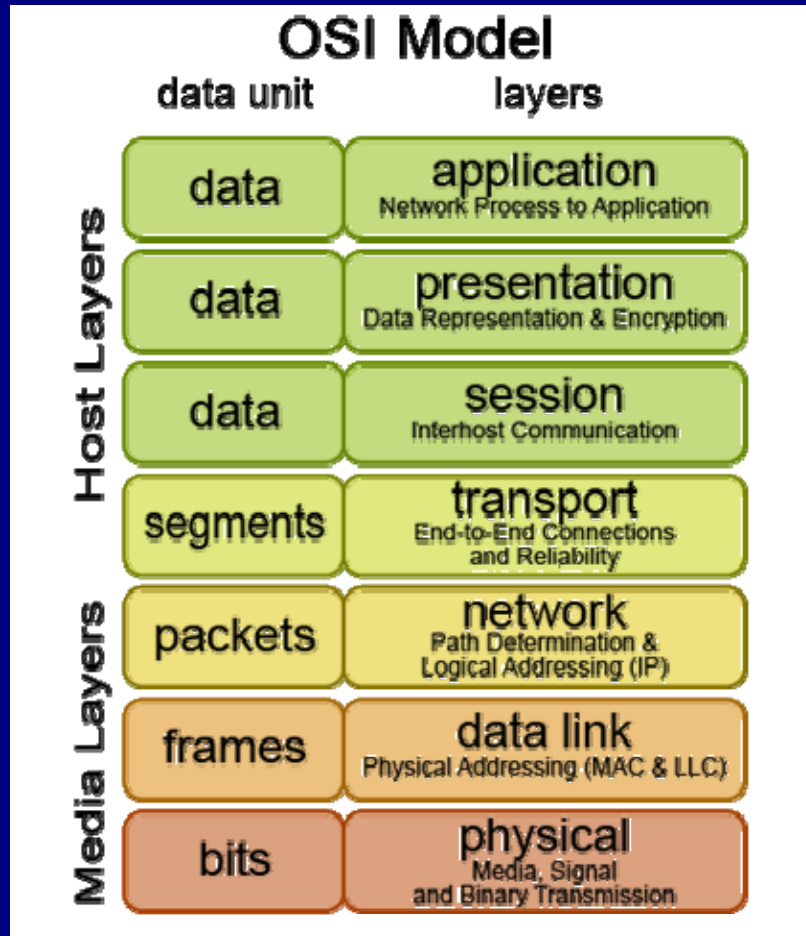
NCBI Daily Users



HL7 and NCBI

- The name HL7 comes from 'Healthcare' and the top level (Level 7) of the Open Systems Interconnection (OSI) model, which carries the meaning of information exchanged between computer applications.

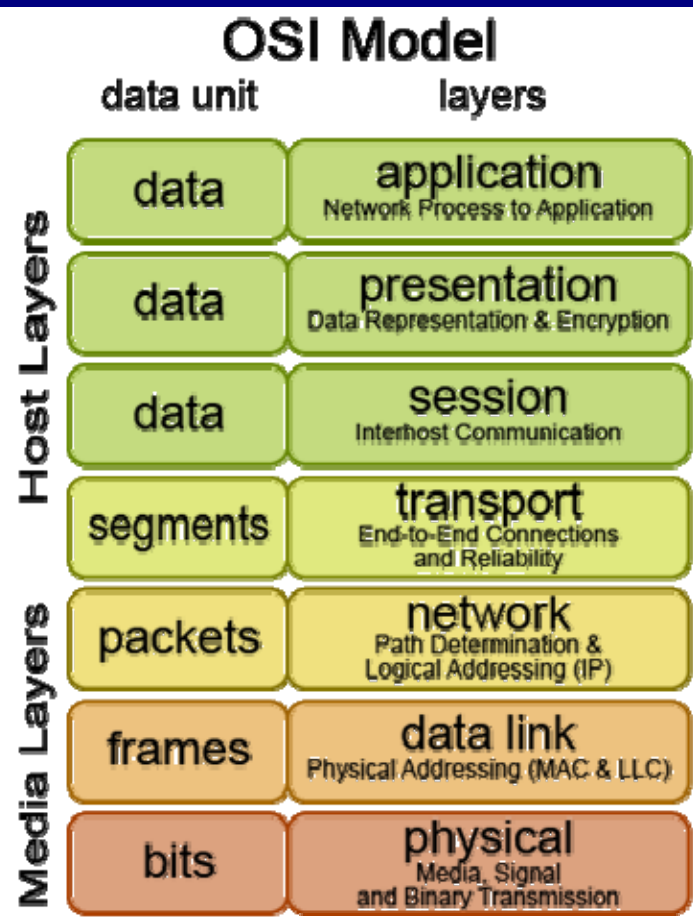
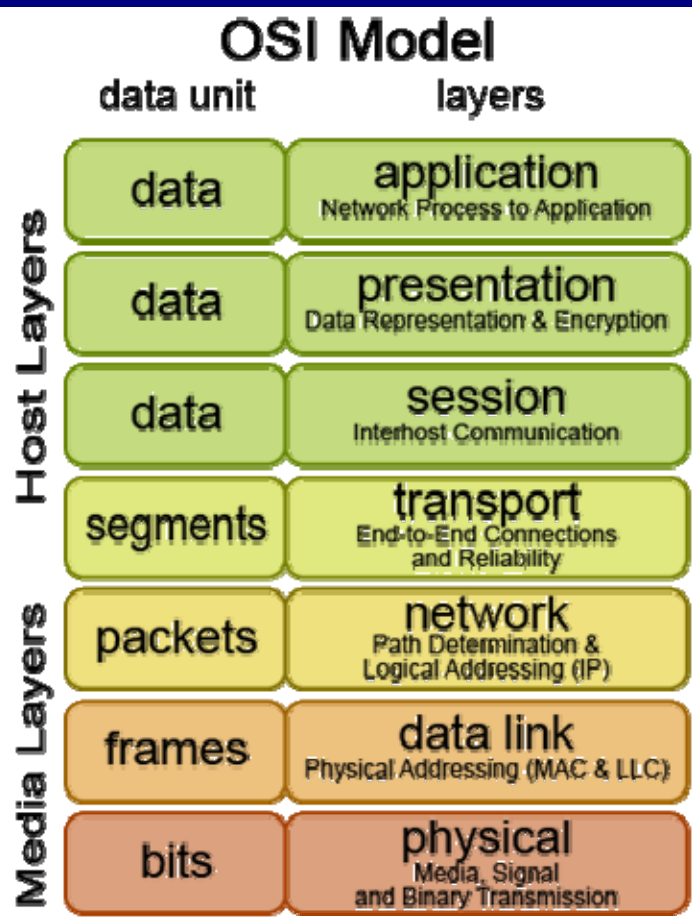
HL7 and NCBI



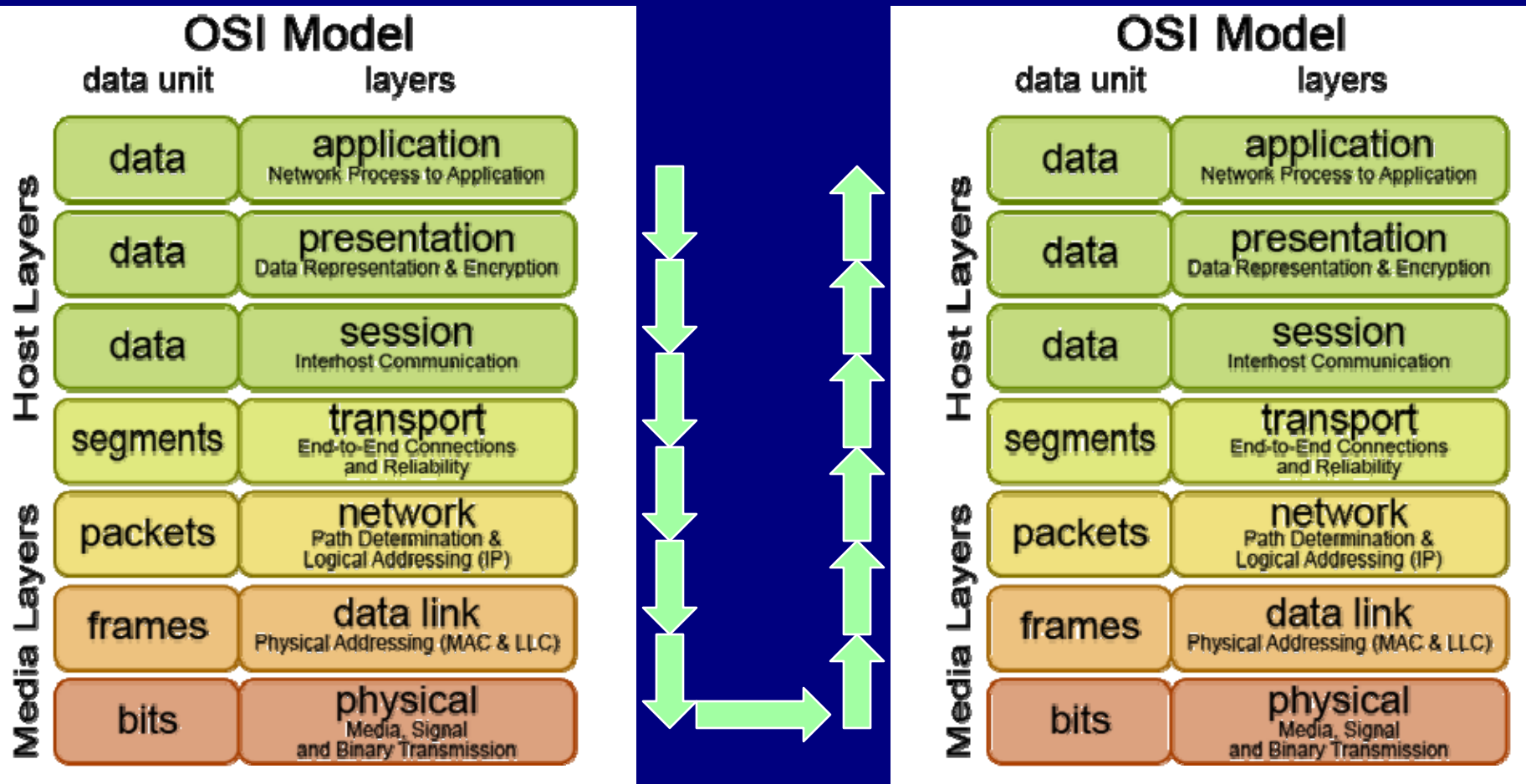
HL7 (1987)

HL7 and NCBI

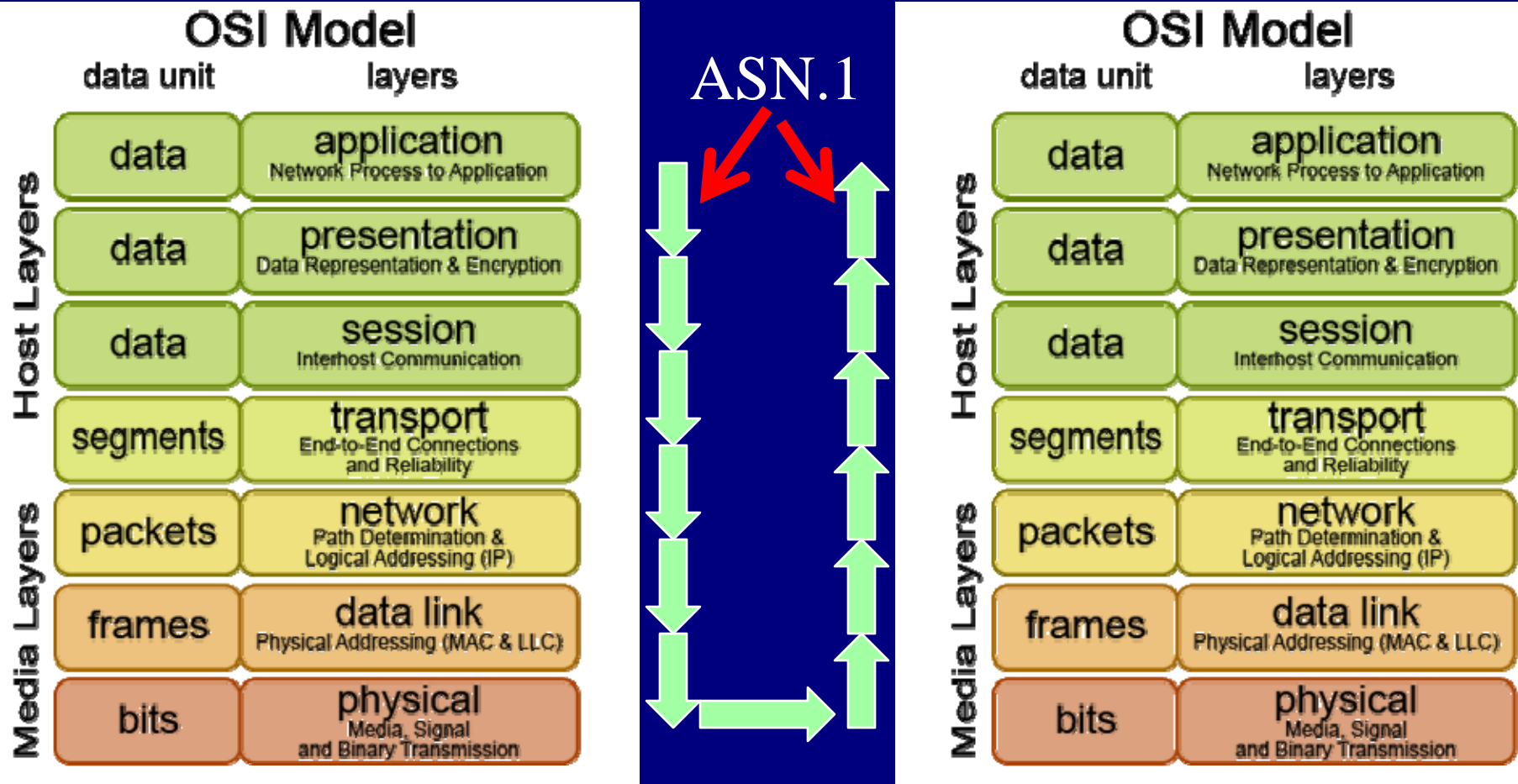
HL7
v.2x



HL7 and NCBI

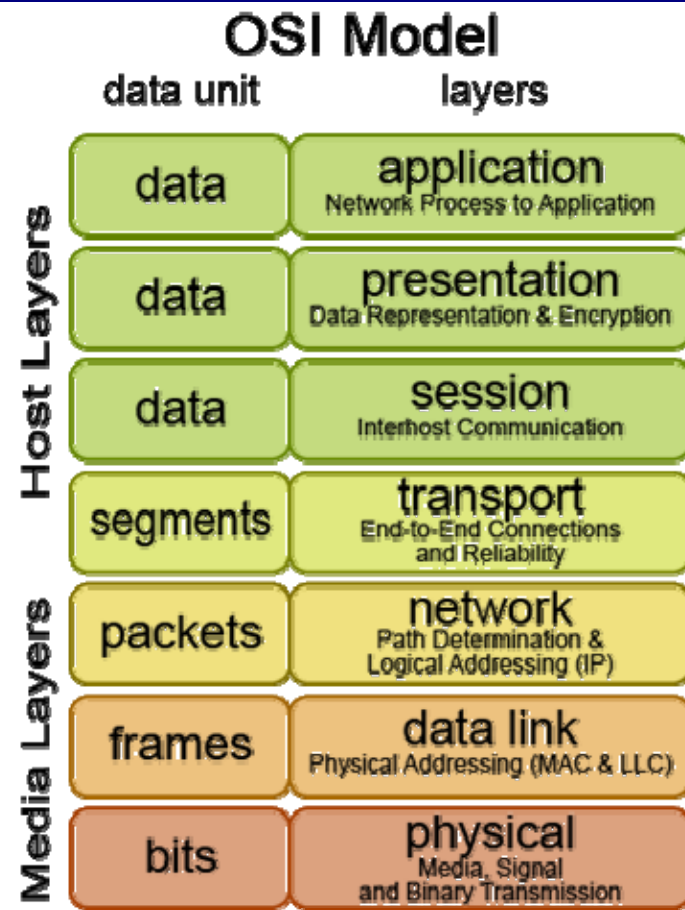
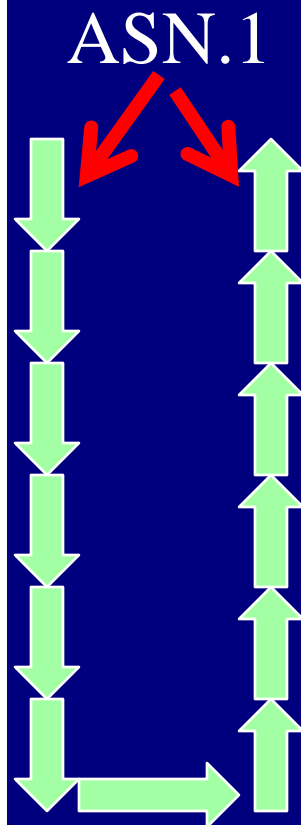
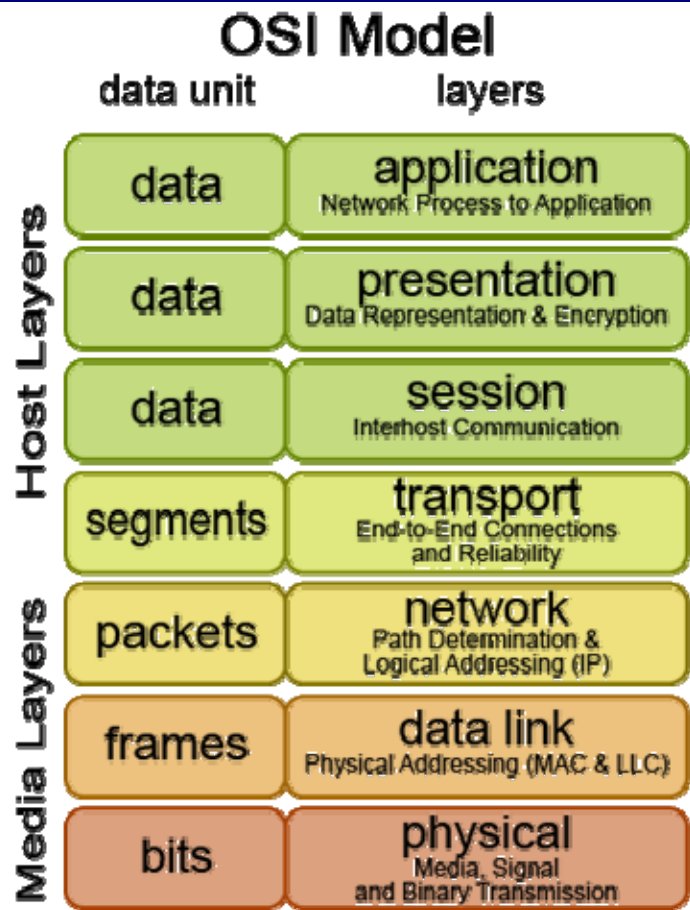


HL7 and NCBI



HL7 and NCBI

NCBI 1989



The NCBI Data Model is defined in ASN.1

- ASN.1 is a data description language similar to a Backus-Naur Form.
- It is a formal language specifically designed to specify complex data structures in a machine, DBMS, and programming language independent manner.
- It is an international standard (ISO 8824, 8825)
- It is used by many data exchange protocols (e.g. X.400, Z39.50, WAIS).

NCBI Described Biological Sequences in ASN.1.

■ ASN.1 definition

```
Bioseq ::= SEQUENCE {  
  id      SET OF Seq-id ,  
  descr   Seq-descr      OPTIONAL,  
  inst    Seq-inst ,  
  annot   SET OF Seq-annot  OPTIONAL}
```

- The minimum required elements are an ID and the instance (e.g. length, topology, residues).

Seq-id

0

1000

Some pieces of an mRNA entry in ASN.1

```

Bioseq ::= {
  id {
    genbank {
      name "HSU03109" ,
      accession "U03109" } ,
    gi 458031 }},
  descr {
    mol-type mRNA ,
    title "Human aspartyl beta-hydroxylase mRNA ..."
    org {
      taxname "Homo sapiens" ,
      common "human" } } ,
  inst {
    repr raw ,
    mol rna ,
    length 2449 ,
    seq-data
      iupacna "AGCTGCCCGCGTTCGCGTGTGTACCCGGTCC..."
  }
  annot {
    ftable { ( ... ) }}}

```

```

Seq-feat ::= {
  data
    cdregion {
      frame one ,
      code {id 1 } } ,
  product
    whole gi 458032 ,
  location
    int {
      from 77 ,
      to 2350 ,
      id gi 458031 } }

```

```

Bioseq ::= {
  id {
    gi 458032 },
  descr {
    title "aspartyl beta-hydroxylase" ,
    method concept-trans } ,
  inst {
    repr raw ,
    mol aa ,
    length 757 ,
    seq-data
      iupacaa "MAQRKNAKSSGNSSSSSGSGSGSTGHKNGRKGG..."
  }
  annot {
    ftable { ( ... ) }}}

```

ASN.1 converted to XML

```
<Bioseq>
- <Bioseq_id>
-   <Seq-id>
-     <Seq-id_genbank>
-       <Textseq-id>
-         <Textseq-id_name>HSU03109</Textseq-id_name>
-         <Textseq-id_accession>U03109</Textseq-id_accession>
-         <Textseq-id_version>1</Textseq-id_version>
-       </Textseq-id>
-     </Seq-id_genbank>
-   </Seq-id>
-   <Seq-id>
-     <Seq-id_gi>458031</Seq-id_gi>
-   </Seq-id>
</Bioseq_id>
- <Bioseq_descr>
-   <Seq-descr>
-     <Seqdesc>
-       <Seqdesc_title>Human aspartyl beta-hydroxylase mRNA, complete
cds.</Seqdesc_title>
-     </Seqdesc>
-   <Seqdesc>
-     <Seqdesc_molinfo>
-       <MolInfo>
-         <MolInfo_biomol value="mRNA">3</MolInfo_biomol>
-       </MolInfo>
-     </Seqdesc_molinfo>
```

Bioinformatics Formats are “Good Enough” (ie. Bad)

```

LOCUS           HSU03109           2449 bp           mRNA           linear           PRI 30-NOV-1995
DEFINITION     Human aspartyl beta-hydroxylase mRNA, complete cds.
ACCESSION     U03109
VERSION       U03109.1   GI:458031
KEYWORDS      .
SOURCE        Homo sapiens (human)
  ORGANISM    Homo sapiens
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
              Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
              Catarrhini; Hominidae; Homo.
REFERENCE     1 (bases 1 to 2249)
  AUTHORS     Korioth,F., Gieffers,C. and Frey,J.
  TITLE       Cloning and characterization of the human gene encoding aspartyl
              beta-hydroxylase
  JOURNAL     Gene 150 (2), 395-399 (1994)
  PUBMED     7821814
REFERENCE     2 (bases 1 to 2449)
  AUTHORS     Korioth,F.
  TITLE       Direct Submission
  JOURNAL     Submitted (03-NOV-1993) Korioth F., Fakultae fuer Chemie-Biochemie
              II, Universitaet Bielefeld, Universitaetsstrasse 25, Bielefeld,
              33615, Germany

FEATURES             Location/Qualifiers
   source             1..2449
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="As-5"

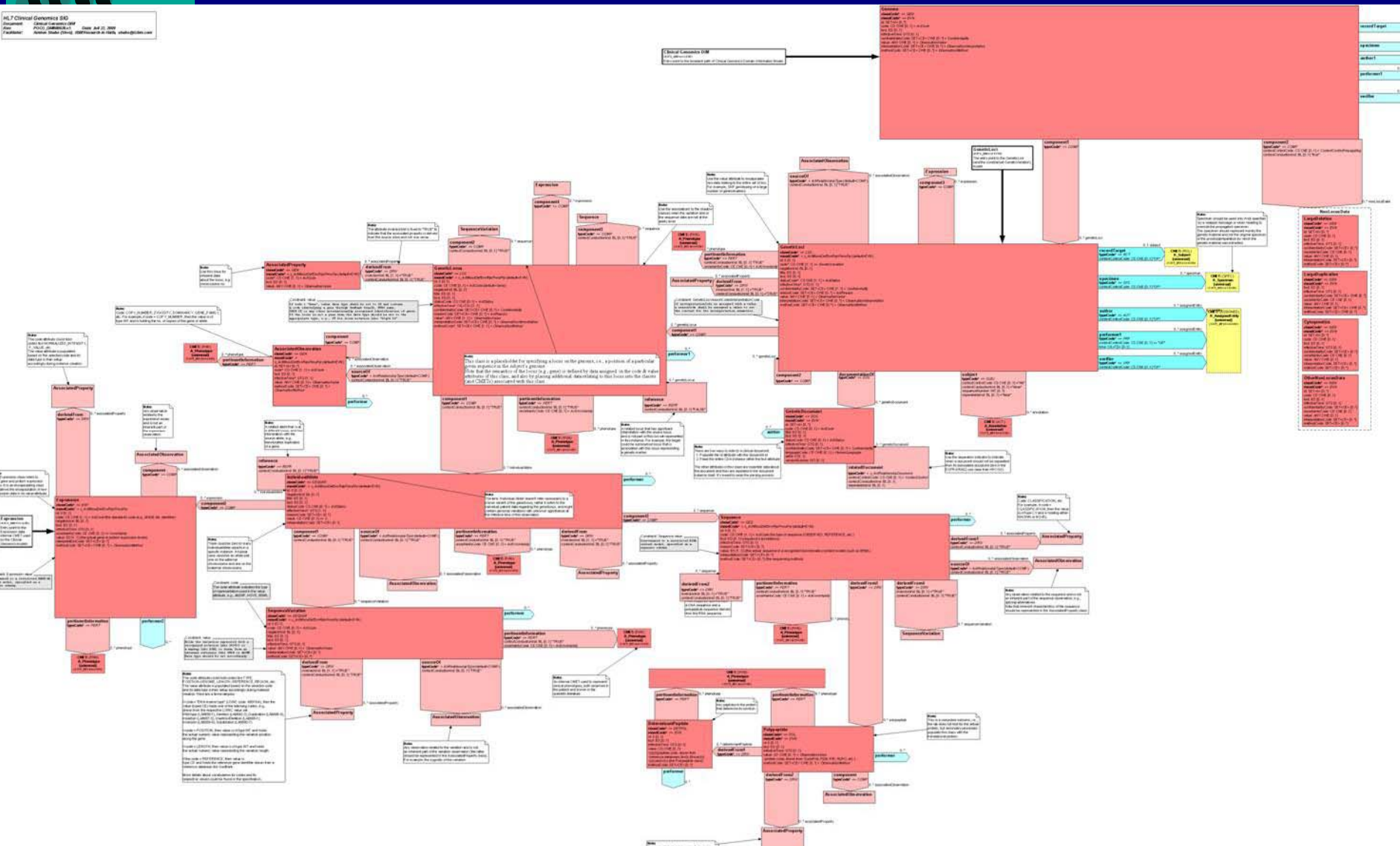
```

Bioinformatics Formats are “Good Enough” (ie. Bad)

```
>gi|458031|gb|U03109.1|HSU03109 Human aspartyl beta-hydroxylase mRNA, complete cds
AGCTGCCCGCGTCGCGTGTGTACCCCCGCGCACTGAAGGAGGTCCACCAGCCCTCACCAGCCCCCGCGGA
CCGTGCAATGGCCCAGCGTAAGAATGCCAAGAGCAGCGGCAACAGCAGCAGCAGCGGCTCCGGCAGCGGT
AGCACGAGTGCGGGCAGCAGCAGCCCCGGGGCCCGGAGAGAGACAAAGCATGGAGGACACAAGAATGGGA
GGAAAGCGGACTCTCAGGAACTTCATTCTTCACGTGGTTTATGGTGATTGCATTGCTGGGCGTCTGGAC
ATCTGTAGCTGTCGTTTGGTTTGATCTTGTTGACTATGAGGAAGTTCTAGGAAAAGTAGGAATCTATGAT
GCTGATGGTGATGGAGATTTTGGATGTGGATGATGCCAAAGTTTTATTAGGACTTAAAGAGAGATCTACTT
CAGAGCCAGCAGTCCCGCCAGAAGAGGCTGAGCCACACACTGAGCCCGAGGAGCAGGTTCCCTGTGGAGGC
AGAACCCAGAAATATCGAAGATGAAGCAAAAGAACAATTTCAGTCCCTTCTCCATGAAATGGTACACGCA
GAACATGTTGAGGGAGAAGACTTGCAACAAGAAGATGGACCCACAGGAGAACCACAACAAGAGGATGATG
AGTTTCTTATGGCGACTGATGTAGATGATAGATTTGAGACCCTGGAACCTGAAGTATCTCATGAAGAAAC
CGAGCATAGTTACCACGTGGAAGAGACAGTTTACAAGACTGTAATCAGGATATGGAAGAGATGATGTCT
GAGCAGGAAAATCCAGATTCCAGTGAACCAGTAGTAGAAGATGAAAGATTGCACCATGATACAGATGATG
TAACATACCAAGTCTATGAGGAACAAGCAGTATATGAACCTCTAGAAAATGAAGGGATAGAAATCACAGA
AGTAACTGCTCCCCCTGAGGATAATCCTGTAGAAGATTACAGGTAATTGTAGAAGAAGTAAGCATTTTT
CCTGTGGAAGAACAGCAGGAAGTACCACCAGAAAACAATAGAAAACAGATGATCCAGAACAAAAAGCAA
AAGTTAAGAAAAAGAAGCCTAAACTTTTTAAATAAATTTGATAAGACTATTAAGCTGAACTTGATGCTGC
AGAAAAACTCCGTAAAAGGGGAAAAATTGAGGAAGCAGTGAATGCATTTAAAGAACTAGTACGCAAATAC
CCTCAGAGTCCACGAGCAAGATATGGGAAGGCGCAGTGTGAGGATGATTTGGCTGAGAAGAGGAGAAGTA
```

Embedding Bioinformatics Formats

HL7 Clinical Genomics SD
Document ID: ClinicalGenomicsSD
Date: 2003-03-03
Version: 1.0.0





BSML

Cover Pages: Bioinformatic Sequence Markup Language (BSML) - Windows Internet Explorer

http://xml.coverpages.org/bsml.html

Favorites | Davis WeatherLink - My ... | AccuWeather.com - Rock... | Metro - Rider Tools - Metr... | StormWatch 7 Weather C...

Cover Pages: Bioinformatic Sequence Markup La...

Cover Pages
Hosted by OASIS

Online resource for markup language technologies

SEARCH | ABOUT | INDEX | NEWS | CORE STANDARDS | TECHNOLOGY REPORTS | EVENTS | LIBRARY

SEARCH
Advanced Search

ABOUT
Site Map
CP RSS Channel
Contact Us
Sponsoring CP
About Our Sponsors

NEWS
Cover Stories
Articles & Papers
Press Releases

CORE STANDARDS
XML
SGML
Schemas
XSL/XSLT/XPath
XLink
XML Query
CSS
SVG

TECHNOLOGY REPORTS
XML Applications
General Apps
Government Apps
Academic Apps

EVENTS

LIBRARY
Introductions
FAQs
Bibliography
Technology and Society

Last modified: November 26, 2001

Technology Reports

Bioinformatic Sequence Markup Language (BSML)

[January 10, 2001] In January 2001, LabBook Inc. announced the availability of BSML (XML DTD) version 2.2. "BSML is an extensible language specification and container for bioinformatic data. BSML was developed under a 1997 grant from the National Human Genome Research Institute (NHGRI) as an evolving public domain standard for the bioinformatics community. The objectives of LabBook are to offer BSML and other XML data formats for effective management, communication, and interactive visualization of bioinformatic data." Background: "Genome research projects typically involve a variety of data (sequences, annotations, analysis results, database links, graphical images, etc.) that may be distributed over multiple storage locations and networks. Creation, management, analysis, and communication of these data often require the use of various computer software applications and databases that utilize non-interchangeable data formats. The lack of standards in bioinformatics is a serious obstacle to productivity. Other obstacles include the loss of information content and state by transmission of data in HTML, and a lack of persistence in bioinformatic analyses and searches because the results are simply pictures in viewers."

[April 12, 2000] The proposed Bioinformatic Sequence Markup Language (BSML) is a public domain protocol for Graphic Genomic Displays. The project goals are in some respects similar to those of the [Chemical Markup Language](#). According to the RFC document of December 1997, which specifies a public domain standard for the encoding and display of DNA, RNA and protein sequence information, this markup language is to be based upon SGML and XML: "BSML is written to conform with the XML standard." Goals in the ending project are to "describe the features of genetic sequences, describe the features of graphic objects used to represent sequence features, determine procedures for assigning graphic objects to sequence features, and determine how to store and transmit encoded sequence and graphic information." BSML is a TopoGEN project, funded by an SBIR [Small Business Innovative Research] from the [National Center for Human Genome Research](#), to develop the public domain protocol. The SBIR with which this project is associated has Joseph Spitzner, Ph. D. (TopoGEN Software Director) as its Principal Investigator.

[July 27, 1999] ["VGI Releases Free BSML Basic Browser for Biotech Research"](#) - ["Visual Genomics, Inc. \(VGI\)](#) announced today the release of BSML Basic Browser. This browser is the first in a family of products to bring visual management, analysis, presentation, and communication of the ever increasing amount of bioinformatics data to genomics researchers. Using [Bioinformatics Sequence Markup Language \(BSML\)](#), an open XML standard developed by VGI and sponsored by National Human Genome Research Institute, the BSML Basic Browser's graphical user interface is a gateway to the visualization, analysis, presentation and communication of genomic data. All information underlying the graphical presentation is contained within the BSML document, allowing the user to drill-down through the data to any level of resolution, from the chromosome to the base pair. Dynamic, interactive datamining is made easy by the intuitive 'point and click'

HOSTED BY
OASIS

SPONSORED BY

IBM

ISIS PAPYRUS

Microsoft

ORACLE

PRIMETON
普元·软件

XML DAILY NEWSLINK
Receive [daily news](#) updates from Managing Editor, Robin Cover.

Subscribe >>

View Archives >>

Internet | Protected Mode: On



BSML

Genomics - Stem Cell - - Windows Internet Explorer

http://www.visualgenomics.com/bsml/index.html

bsml

Genomics - Stem Cell -

Visual Genomics

Search: GO

[Login](#) | [Register](#)

[Custom Microarray Print](#)
Ideal for Custom DNA Protein arrays
High Throughput Microarray Printing

[Cell Lysis Made Easy](#)
Lyse Any Sample In 40 Sec or Less Get A
Free In Lab Demo Now!

Ads by Google

Submit Link | Submit Article | Latest Links | Latest Articles | Top Hits | Contact | Advanced Search

Genomics - Stem Cell

CATEGORIES

- ▶ Genome
- ▶ Internet
- ▶ Legal
- ▶ Medical
- ▶ Stem Cells
- ▶ Tobacco
- ▶ Wine

ARTICLES

Ads by Google

[OEM Sensing Solutions](#)
Miniature spectrometers & sensors for embedding into medical devices
www.oceanoem.com

[Custom Microarray Print](#)
Ideal for Custom DNA Protein arrays High Throughput Microarray Printing
digilabglobal.com/Microarra

[ICL Increases Your Risk](#)

Internet | Protected Mode: On

Cover Pages: Bioinformatic Sequence Marku

http://xml.coverpages.org/bs

Cover Pages: Bioinformatic Sequence Marku

Cover Pages
Hosted by OASIS Online resou

SEARCH
Advanced Search

Technology R

Bioinform

[January 10, 2001] language specific Research Institute (BSML and other XM Background: "Geno graphical images, e communication of t data formats. The la content and state b are simply pictures

[April 12, 2000] The The project goals at 1997, which specifi language is to be b "describe the featur procedures for assit information." BSML [Genome Research](#) (TopoGEN Software

[July 27, 1999] "VG release of BSML B communication of t [Language \(BSML\)](#). Basic Browser's gra information underlyi any level of resoluti

ABOUT
Site Map
CP RSS Channel
Contact Us
Sponsoring CP
About Our Sponsors

NEWS
Cover Stories
Articles & Papers
Press Releases

CORE STANDARDS
XML
SGML
Schemas
XSL/XSLT/Path
XLink
XML Query
CSS
SVG

TECHNOLOGY REPORTS
XML Applications
General Apps
Government Apps
Academic Apps

EVENTS

LIBRARY
Introductions
FAQs
Bibliography
Technology and Society

NCBI

WWW
110101

BSML

The screenshot shows a Windows Internet Explorer browser window with the address bar displaying <http://www.visualgenomics.com/bsml/index.html>. The browser's Favorites bar contains several links, including 'Davis WeatherLink - My...', 'AccuWeather.com - Rock...', 'Metro - Rider Tools - Metr...', and 'StormWatch 7 Weather C...'. The main content area of the browser shows the 'Visual Genomics' website. At the top, there is a search bar with a 'GO' button and links for 'Login' and 'Register'. Below the search bar, there are two advertisements: 'Custom Microarray Print' and 'Cell Lysis Made Easy'. A navigation bar at the bottom of the website includes links for 'Hits', 'Contact', and 'Advanced Search'. A large white error box is overlaid on the page, containing the text: **Genomics - Stem Cell** ...>> **Error: This page it is not active.** The status bar at the bottom of the browser window indicates 'Internet | Protected Mode: On' and '100%' zoom level.

NCBI



BSML

The screenshot shows a Windows Internet Explorer browser window with the address bar displaying <http://www.visualgenomics.com/bsml/index.html>. The browser's Favorites bar includes links for Davis WeatherLink, AccuWeather.com, Metro - Rider Tools, and StormWatch 7 Weather C... The page content features a navigation menu with categories like Genome, Internet, Legal, Medical, Stem Cells, Tobacco, and Wine. A sidebar on the left contains sections for SEARCH, ABOUT, NEWS, CORE STANDARDS, TECHNOLOGY REPORTS, and EVENTS. The main content area displays a "Cover Pages" section and a "Bioinform" section with a date of [January 10, 2001].

Overlaid on the browser is a smaller window titled "The XEMBL service has been discontinued and replaced with two supported | EBI - Windows Internet Explorer" with the address bar showing <http://www.ebi.ac.uk/xembl/>. This window displays a search bar for EMBL-EBI and a navigation menu. The main content area shows a "XEMBL - Discontinued" message with a "Please Note" section stating: "The XEMBL service has been discontinued and replaced with two supported XML formats (EMBLxml, INSDseq). For further information, please refer to <http://www.ebi.ac.uk/embl/xml/>."

MAGE-ML

Cover Pages: MicroArray and Gene Expression Markup Language (MAGE-ML) - Windows Internet Explorer

http://xml.coverpages.org/mageML.html

Cover Pages: MicroArray and Gene Expression M...

Cover Pages
Hosted by OASIS

Online resource for markup language technologies

SEARCH | ABOUT | INDEX | NEWS | CORE STANDARDS | TECHNOLOGY REPORTS | EVENTS | LIBRARY

Last modified: February 08, 2002

Technology Reports

MicroArray and Gene Expression Markup Language (MAGE-ML)

[February 08, 2002] Microarray Gene Expression Markup Language (MAGE-ML) "is a language designed to describe and communicate information about microarray based experiments. MAGE-ML is based on XML and can describe microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results. MAGE-ML has been automatically derived from Microarray Gene Expression Object Model (MAGE-OM), which is developed and described using the Unified Modelling Language (UML) -- a standard language for describing object models. Descriptions using UML have an advantage over direct XML document type definitions (DTDs), in many respects. First they use graphical representation depicting the relationships between different entities in a way which is much easier to follow than DTDs. Second, the UML diagrams are primarily meant for humans, while DTDs are meant for computers. Therefore MAGE-OM should be considered as the primary model, and we will explain MAGE-ML by providing simplified fragments of MAGE-OM, rather than XML DTD or XML Schema." [from the description by Ugis Sarkans]

MicroArray and GeneExpression Markup Language (MAGE-ML). Description from the Revised Gene Expression RFP:

The MAGE-ML model defines the elements for supporting gene expression data. Because the exchange of gene expression data can be abstracted from the source from which it was obtained, it can be represented by XML files, which are both human readable and machine readable. This facilitates an independence between the export and the import of the gene expression data as illustrated below. Ad hoc queries, when the XML files are directly accessible, can take advantage of the suite of W3C recommendations, including XSLT or XMLQuery. Queries against repositories could be specified a number of ways, including through an IDL interface that had as its query language either of the above choices or OQL based on MAGEOM. The DTD file, MAGE-ML.dtd, is generated from MAGE-OM from a fixed set of rules. In one area, BioAssayData, further modifications were made to offer alternatives and efficiency to the parsing... The vocabulary of MAGE-ML is organized into sub-vocabularies in such a way that the sub-vocabularies are independent of each other. These sub-vocabularies are driven by the packages and Identifiable classes of the MAGE-OM, which correspond to discreet groupings of events and results of Gene expression experiments. This will allow a valid XML document to contain the data from an individual sub-vocabulary, such as BioMaterial or ArrayDesign, or to contain any combination of these sub-vocabularies, such as all the BioAssay and BioAssayData for an experiment. Implementations may impose additional ordering, such as ArrayDesigns before their Arrays, or they may require that they be exported to separate files.

HOSTED BY
OASIS

SPONSORED BY

IBM

ISIS PAPYRUS

Microsoft

ORACLE

PRIMETON
普元·软件

XML DAILY NEWSLINK
Receive daily news updates from Managing Editor, Robin Cover.

Subscribe >>

View Archives >>

Internet | Protected Mode: On

NCBI

WWW 110101

MAGE-ML

Cover Pages: MicroArray and Gene Expression Markup Language (MAGE-ML) - Windows Internet Explorer

http://xml.coverpages.org/mageML.html

Cover Pages: MicroA... MAGE - Workgroups - FGED - Windows Internet Explorer

http://www.mged.org/Workgroups/MAGE/mage.html

MAGE - Workgroups - FGED

FGED SOCIETY

Google Custom Search Search

HOME MEETINGS **WORKGROUPS** MISSION FGED BOARD SITE MAP

>> Quick Links

Home > Workgroups > MAGE

MicroArray and Gene Expression - MAGE

FGED Sponsors
illumina

This is the homepage for the MAGE group. The group aims to provide a standard for the representation of microarray expression data that would facilitate the exchange of microarray information between different data systems.

MAGE-TAB

MAGE-TAB is the currently recommended best practice approach. More details available from: <http://www.mged.org/mage-tab/>

Other Detailed Information

Through the **OMG** (Object Management Group) the establishment of a data exchange model (**MAGE-OM: Microarray Gene Expression - Object Model**) and data exchange format (**MAGE-ML: Microarray Gene Expression - Markup Language**) for microarray expression experiments has been done. MAGE-OM has been modelled using the Unified Modelling Language (UML) and MAGE-ML has been implemented using XML (eXtensible Markup Language). MAGEstk (or MAGE Software Toolkit) is a collection of packages that act as converters between MAGE-OM and MAGE-ML under various programming platforms.

There are [guidelines](#) on how to encode MIAME in MAGE-ML.

Please subscribe to the MAGE mailing lists from [here](#).

MAGE Links:

- [Introduction](#)
- [Websites associated with MAGE](#)
- [MAGE-OM](#)
- [MAGE-ML](#)

Implementations may impose additional ordering, such as ArrayDesigns before their Arrays, or they may require that they be exported to separate files.

NCBI

WWW
110101

MAGE-ML

Representation of microarray expression data that would facilitate the exchange of microarray information between different data systems.

MAGE-TAB
MAGE-TAB is the currently recommended best practice approach. More details available from:
<http://www.mged.org/mage-tab/>

Other Detailed Information

TECHNOLOGY REPORTS
XML Applications
General Apps
Government Apps
Academic Apps

EVENTS

LIBRARY
Introductions
FAQs
Bibliography
Technology and Society
Semantics
Tech Topics
Software
Related Standards
Historic

OM: Microarray Gene Expression - Object Model) and data exchange format (MAGE-ML: Microarray Gene Expression - Markup Language) for microarray expression experiments has been done. MAGE-OM has been modelled using the Unified Modelling Language (UML) and MAGE-ML has been implemented using XML (eXtensible Markup Language). MAGEstk (or MAGE Software Toolkit) is a collection of packages that act as converters between MAGE-OM and MAGE-ML under various programming platforms.

There are [guidelines](#) on how to encode MIAME in MAGE-ML.

Please subscribe to the MAGE mailing lists from [here](#).

MAGE Links:

- [Introduction](#)
- [Websites associated with MAGE](#)
- [MAGE-OM](#)
- [MAGE-ML](#)

Implementations may impose additional ordering, such as ArrayDesigns before their Arrays, or they may require that they be exported to separate files.

The Importance of Live Archives

- NCBI converts external formats de jour into internal ASN.1 (or XML or other)
- NCBI converts ASN.1 to external formats de jour
- NCBI ASN.1 model has been very stable, but we still must update occasionally.
- May be backward compatible but must support old form.
- Occasionally must convert archive.

Forget Format – Let's Look at Content

```

<subjectOf2>
  <geneticLocus>
    <component1>
      <individualAllele moodCode="EVN">
        <text>breast cancer 1, early onset</text>
        <value code="83990" displayName="BRCA1" codeSystemName="NCBI Entrez">
          <translation code="20473" displayName="BRCA1" codeSystem="HGNC"/>
        </value>
      <component2>
        <sequence moodCode="EVN">
          <code code="BSMLcon3"/>
          <value mediaType="text/xml">
            <bsml:Bsml xmlns:bsml="urn:bsml.org">
              <bsml:Definitions>
                <bsml:Sequences>
                  <bsml:Sequence id="seq1" molecule="dna" ic-acckey="U14680 REGION: 101..199" db-source="GenBank" title="BRCA1, exon 2" representation="raw" local-acckey="this could be used by the genetic lab">
                    <bsml:Seq-data>
                      GCTCCA CTCCATGAGG TATTCTTCA
                      CATCCGTGTC CCGGCCCGGC CGCGGGGAGC CCCGCTTCAT CGCCGTGGGC
                      TACGTGGACG ACACGCAGTT CGTGCGGTTT GACAGCGACG CCGCGAGCCA
                      GAGGATGGAG CCGCGGGCGC CGTGGATAGA GCAGGAGGGG CCGGAGTATT
                      GGGACCAGGA GACACGGAAT GTGAAGGCC AGTCACAGAC TGACCGAGTG
                      GACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG CCG
                    </bsml:Seq-data>
                  </bsml:Sequence>
                  <bsml:Sequence id="seq2" molecule="dna" ic-acckey="U14680 REGION: 200..253" db-source="GenBank" title="BRCA1, exon 3" representation="raw" local-acckey="this could be used by the genetic lab">
                    <bsml:Seq-data>
                      GTTCTCA
                      CACCATCCAG ATAATGTATG GCTGCGACGT GGGGTCGGAC GGGCGCTTCC
                      TCCGCGGGTA CCGGCAGGAC GCCTACGACG GCAAGGATTA CATCGCCCTG
                      AACGAGGACC TGCCTCTTG GACCGCGGCG GACATGGCGG CTCAGATCAC
                      CAAGCGCAAG TGGGAGGCGG CCCATGTGGC GGAGCAGCAG AGAGCCTACC
                      TGGATGGCAC GTGCGTGGAG TGGCTCCGCA GATACCTGGA GAACGGGAAG
                      GAGACGCTGC AGCGCACGG
                    </bsml:Seq-data>
                  </bsml:Sequence>
                </bsml:Sequences>
              </bsml:Definitions>
            </bsml:Bsml>
          </value>
        </sequence>
      </component2>
    </individualAllele>
  </geneticLocus>
</subjectOf2>

```

Forget Format – Let's Look at Content

```

<subjectOf2>
<geneticLocus>
  <component1>
    <individualAllele moodCode="EVN">
      <text>breast cancer 1, early onset</text>
      <value code="83990" displayName="BRCA1" codeSystemName="NCBI Entrez">
        <translation code="20473" displayName="BRCA1" codeSystem="HGNC"/>
      </value>
    <component2>
      <sequence moodCode="EVN">
        <code code="BSMLcon3" />
        <value mediaType="text">
          <bsml:Bsm1 xmlns:bsml="http://www.ncbi.nlm.nih.gov/Sequence/BSML" />
            <bsml:Definitions>
              <bsml:Sequence id="seq1" molecule="dna" local-acckey="U14680 REGION: 101..199" title="BRCA1, exon 3" representation="raw" local-acckey="this could be used by the genetic lab">
                <bsml:Seq-data>
                  GCTCCCACTCCATGAGG TATTTCCTCA
                  CATCCGTGTC CCGGCCCGGC CGCGGGGAGC CCCGCTTCAT CGCCGTGGGC
                  TACGTGGACG ACACGCAGTT CGTGCGGTTT GACAGCGACG CCGCGAGCCA
                  GAGGATGGAG CCGCGGGCGC CGTGGATAGA GCAGGAGGGG CCGGAGTATT
                  GGGACCAGGA GACACGGAAT GTGAAGGCC AGTCACAGAC TGACCGAGTG
                  GACCTGGGGA CCTGCGCGG CTACTACAAC CAGAGCGAGG CCG
                </bsml:Seq-data>
              </bsml:Sequence>
              <bsml:Sequence id="seq2" molecule="dna" local-acckey="U14680 REGION: 200..253" db-source="GenBank" title="BRCA1, exon 3" representation="raw" local-acckey="this could be used by the genetic lab">
                <bsml:Seq-data>
                  GTTCTCA
                  CACCATCCAG ATAATGTATG GCTGCGACGT GGGGTCGGAC GGGCGCTCC
                  TCCGCGGGTA CCGGCAGGAC GCCTACGACG GCAAGGATTA CATCGCCCTG
                  AACGAGGACC TGCCTCTTG GACCGCGCGG GACATGGCGG CTCAGATCAC
                  CAAGCGCAAG TGGGAGGCGG CCCATGTGGC GGAGCAGCAG AGAGCCTACC
                  TGGATGGCAC GTGCGTGGAG TGGCTCCGCA GATACCTGGA GAACGGGAAG
                  GAGACGCTGC AGCGCACGG
                </bsml:Seq-data>
              </bsml:Sequence>
            </bsml:Definitions>
          </value>
        </sequence>
      </component2>
    </individualAllele>
  </component1>
</geneticLocus>
</subjectOf2>

```

Forget Format – Let's Look at Content

```

LOCUS           HSU03109           2449 bp           mRNA           linear           PRI 30-NOV-1995
DEFINITION     Human aspartyl beta-hydroxylase mRNA, complete cds.
ACCESSION      U03109
VERSION        U03109.1   GI:458031
KEYWORDS       .
SOURCE         Homo sapiens (human)
   ORGANISM    Homo sapiens
               Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
               Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
               Catarrhini; Hominidae; Homo.
REFERENCE      1 (bases 1 to 2249)
   AUTHORS     Korioth,F., Gieffers,C. and Frey,J.
   TITLE       Cloning and characterization of the human gene encoding aspartyl
               beta-hydroxylase
   JOURNAL     Gene 150 (2), 395-399 (1994)
   PUBMED     7821814
REFERENCE      2 (bases 1 to 2449)
   AUTHORS     Korioth,F.
   TITLE       Direct Submission
   JOURNAL     Submitted (03-NOV-1993) Korioth F., Fakultae fuer Chemie-Biochemie
               II, Universitaet Bielefeld, Universitaetsstrasse 25, Bielefeld,
               33615, Germany

FEATURES             Location/Qualifiers
   source             1..2449
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="As-5"

```

Forget Format – Let's Look at Content

LOCUS HSU03109 2449 bp mRNA linear PRI 30-NOV-1995

ACCESSION U03109
 VERSION U03109.1 GI:458031

SOURCE Homo sapiens (human)
 ORGANISM Homo sapiens
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
 Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 2249)
 AUTHORS Koriath,F., Gieffers,C. and Frey,J.
 TITLE Cloning and characterization of the human gene encoding aspartyl
 beta-hydroxylase
 JOURNAL Gene 150 (2), 395-399 (1994)
 PUBMED 7821814

REFERENCE 2 (bases 1 to 2449)
 AUTHORS Koriath,F.
 TITLE Direct Submission
 JOURNAL Submitted (03-NOV-1993) Koriath F., Fakultae fuer Chemie-Biochemie
 II, Universitaet Bielefeld, Universitaetsstrasse 25, Bielefeld,
 33615, Germany

FEATURES Location/Qualifiers
 source 1..2449
 /organism="Homo sapiens"
 /mol_type="mRNA"
 /db_xref="taxon:9606"
 /clone="As-5"

Why Versions Matter

U12345: position 10 is “C”

U12345.1: position 10 is “C”



U12345

U12345.1

AGCTGCCCGCGTCGCGTGTGTACCCCGCGCACTGA

Why Versions Matter

U12345: position 10 is “C”

U12345.1: position 10 is “C”



U12345
U12345.1

AGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTGA

U12345
U12345.2

TAGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTG

Why Versions Matter

U12345: position 10 is “C”

U12345.1: position 10 is “C”



U12345
U12345.1

AGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTGA

U12345: position 10 is “G”



U12345
U12345.2

TAGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTG

Why Versions Matter

U12345: position 10 is “C”

U12345.1: position 10 is “C”



U12345

U12345.1

AGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTGA

U12345: position 10 is “G”

U12345.1: position 10 is still “C”



U12345

U12345.2

TAGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTG

Why Versions Matter

U12345: position 10 is “C”

U12345.1: position 10 is “C”



U12345

U12345.1

AGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTGA

U12345: position 10 is “G”

U12345.1: position 10 is still “C”



U12345

U12345.2

TAGCTGCCCCGCGTCGCGTGTGTACCCCCGCGCACTG

Another Reason Why Live Archives Are Important



HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model, Release 1

OBX|3|CWE|51958-7^Transcript reference sequence
identifier^LN||NM_170707.1^^2.16.840.1.113883.6.280|||||F|200
80702100909|||||||Laboratory for Molecular
Medicine^L^22D1005307^^^CLIA&2.16.840.1.113883.4.7&IS
O|1000 Laboratory Lane^Ste.
123^Cambridge^MA^99999^USA^B



HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model, Release 1

OBX|3|CWE|51958-7^Transcript reference sequence
identifier^LN|NM_170707.1^2.16.840.1.113883.6.280|||||F|200
80702100909|||||||Laboratory for Molecular
Medicine^L^22D1005307^^^CLIA&2.16.840.1.113883.4.7&IS
O|1000 Laboratory Lane^Ste.
123^Cambridge^MA^99999^USA^B

NCBI

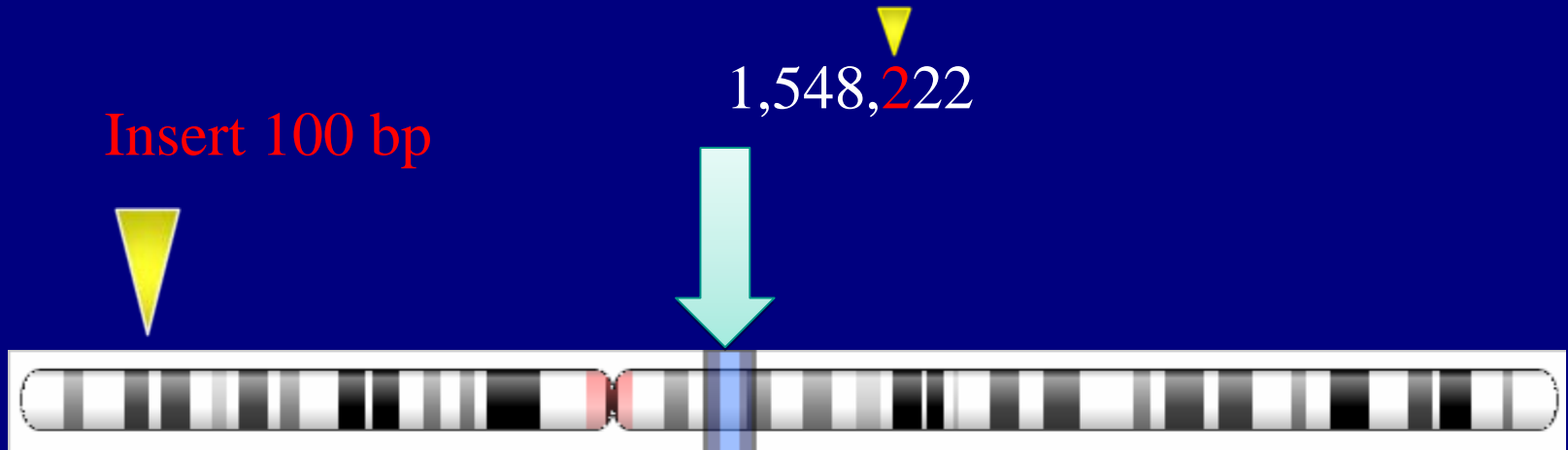
WWW
110101

The Changing Genome

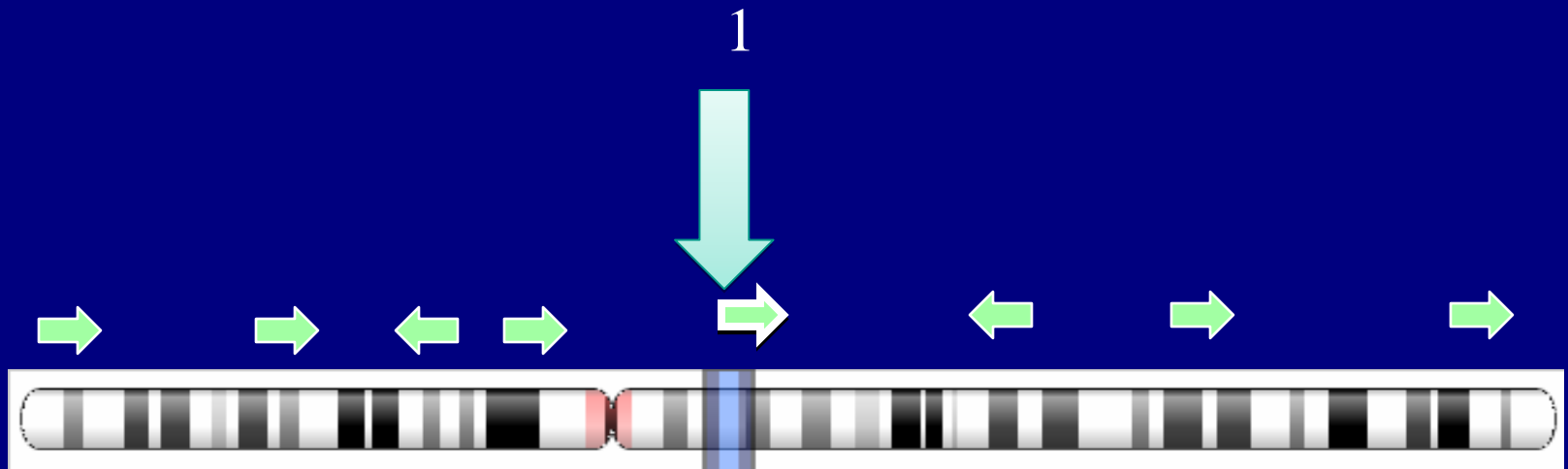
1,548,122



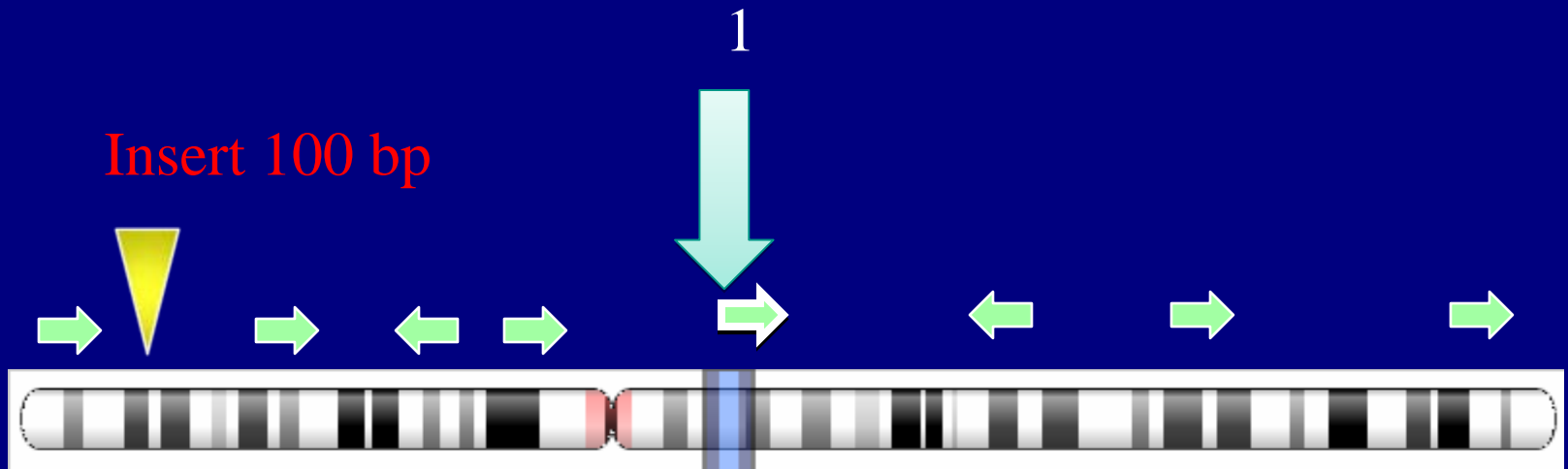
The Changing Genome



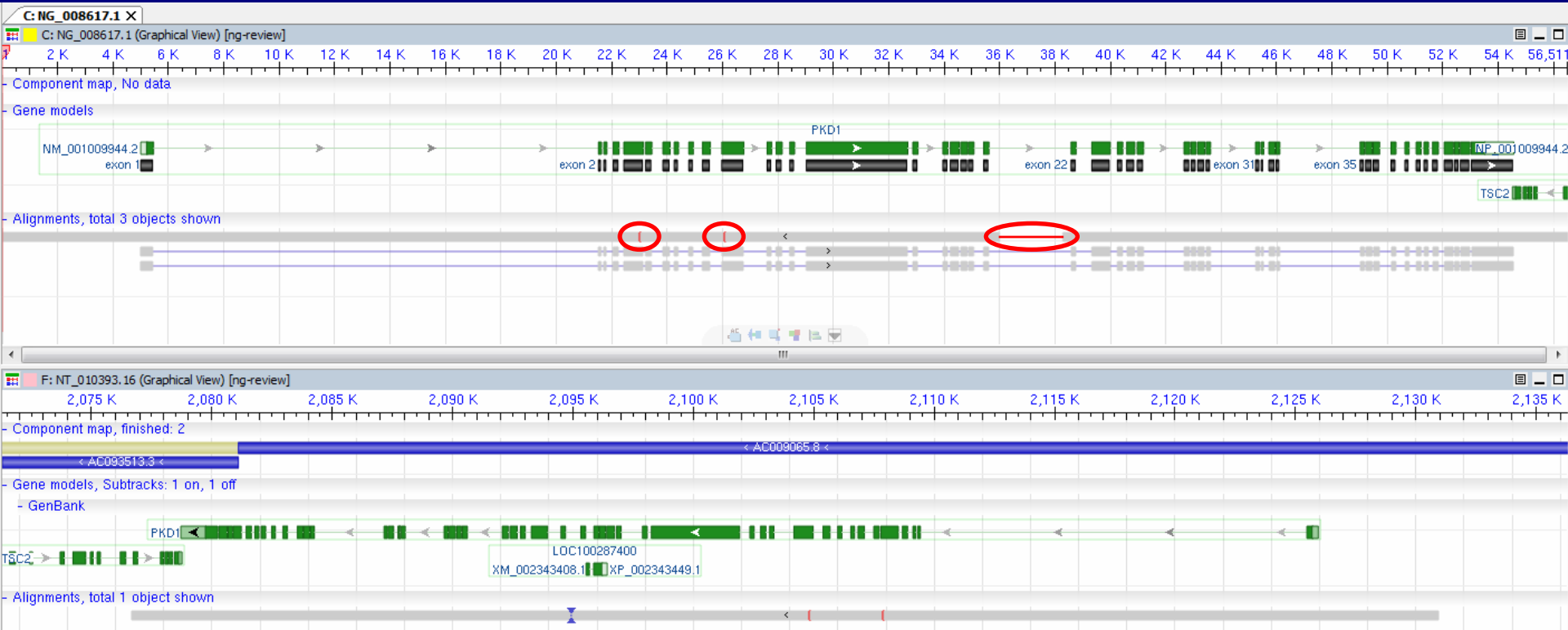
RefSeqGene/LRG



RefSeqGene/LRG



RefSeqGene / LRG



The Genome is Versioned

Genome Reference Consortium - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Genome Reference Consortium

Genome Reference Consortium





GRC Home Human Mouse Help Report an Issue Contact Us Curators Only

The Genome Reference Consortium

At the time the human reference was initially described, it was clear that some regions were recalcitrant to closure with existing technology. What was not as clear was the degree to which structural variation affected our ability to produce a truly representative genome sequence at some loci. It is now apparent that some regions of the genome are sufficiently variable that they are best represented by multiple sequences in order to capture all of the sequence potentially available at these loci.

In order to improve the representation of the reference human genome we have formed the Genome Reference Consortium (GRC). The goal of this group is to correct the small number of regions in the reference that are currently misrepresented, to close as many remaining gaps as possible and to produce alternative assemblies of structurally variant loci when necessary. We will provide mechanisms by which the scientific community can report loci in need of further review. In addition, information about loci currently under review and genome assembly production cycles will be made readily available. The human reference assembly is the cornerstone upon which all whole genome studies are based. It is critical to ensure that we have the best possible view of the genome to facilitate continued progress in understanding and improving human health.

The Genome Reference Consortium consists of:

-  wellcome trust **sanger** institute
The Wellcome Trust Sanger Institute
-  THE **Genome** CENTER
AT WASHINGTON UNIVERSITY The Genome Center at Washington University
-  EMBL-EBI
The European Bioinformatics Institute
-  **NCBI**
The National Center for Biotechnology Information

GRC News and Updates

GRC in the News

Tue, 29 Jan 2009
The GRC is highlighted in a Nature news feature.

GRCh37 is now available in Map Viewer

Fri, 14 Aug 2009
NCBI has annotated and released the latest version of the public human genome assembly (GRCh37).
[see all](#)

Resolved Issues

Mouse (MG-3729) *Sep23, 2010*

AC113082.6 has been updated to AC113082.7 (vector sequence removed), fixing the half-dovetail join between AC113082 and CT025561.

Human (HG-858) *Sep23, 2010*

Certificate has been submitted for the overlap between AC026369.21 and AC215219.3
[see all](#)

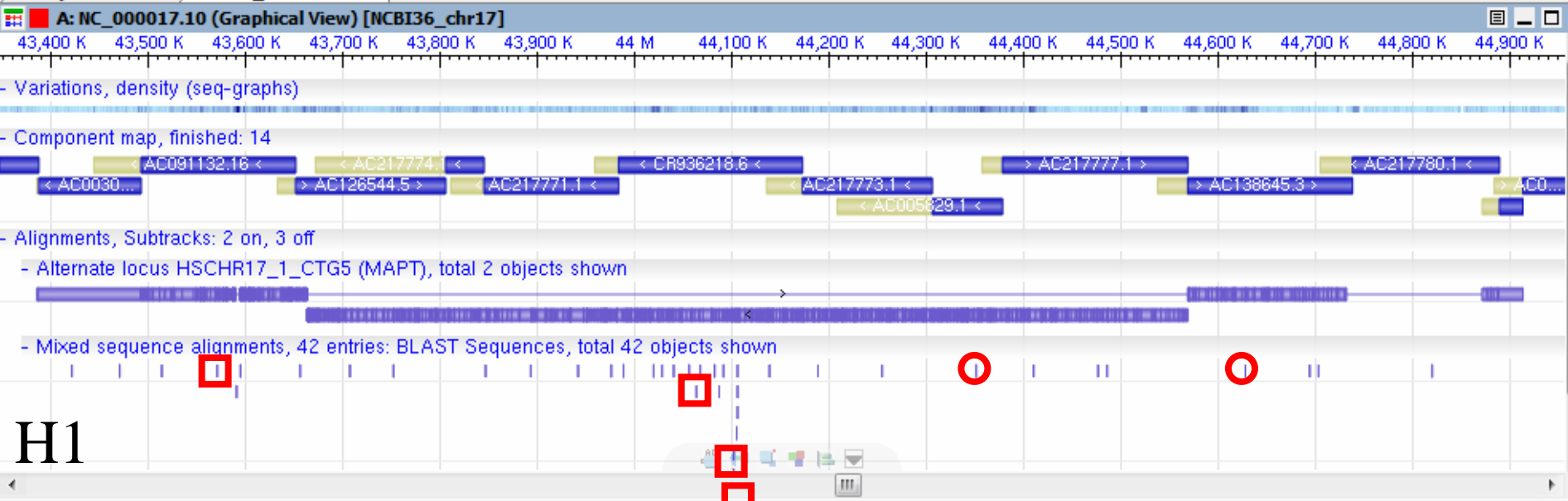
FTP | NHGRI | The Wellcome Trust | HHS | NIH | Accessibility |

Page last updated: May 1, 2009

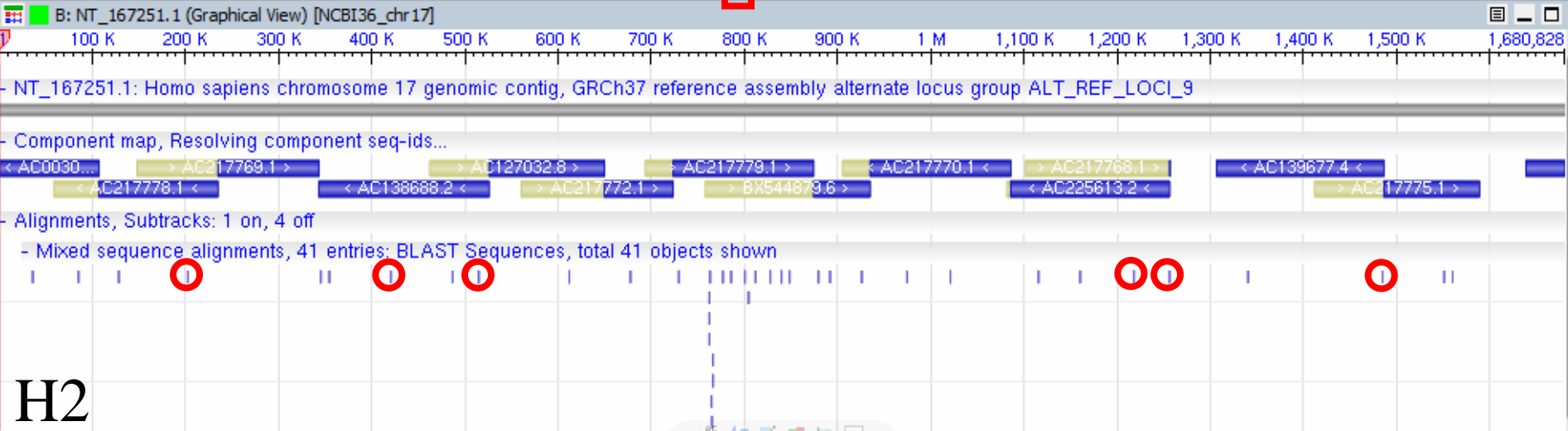
Done Local intranet | Protected Mode: Off 100%

The Genome Changes

17q21.31 microdeletion syndrome



H1



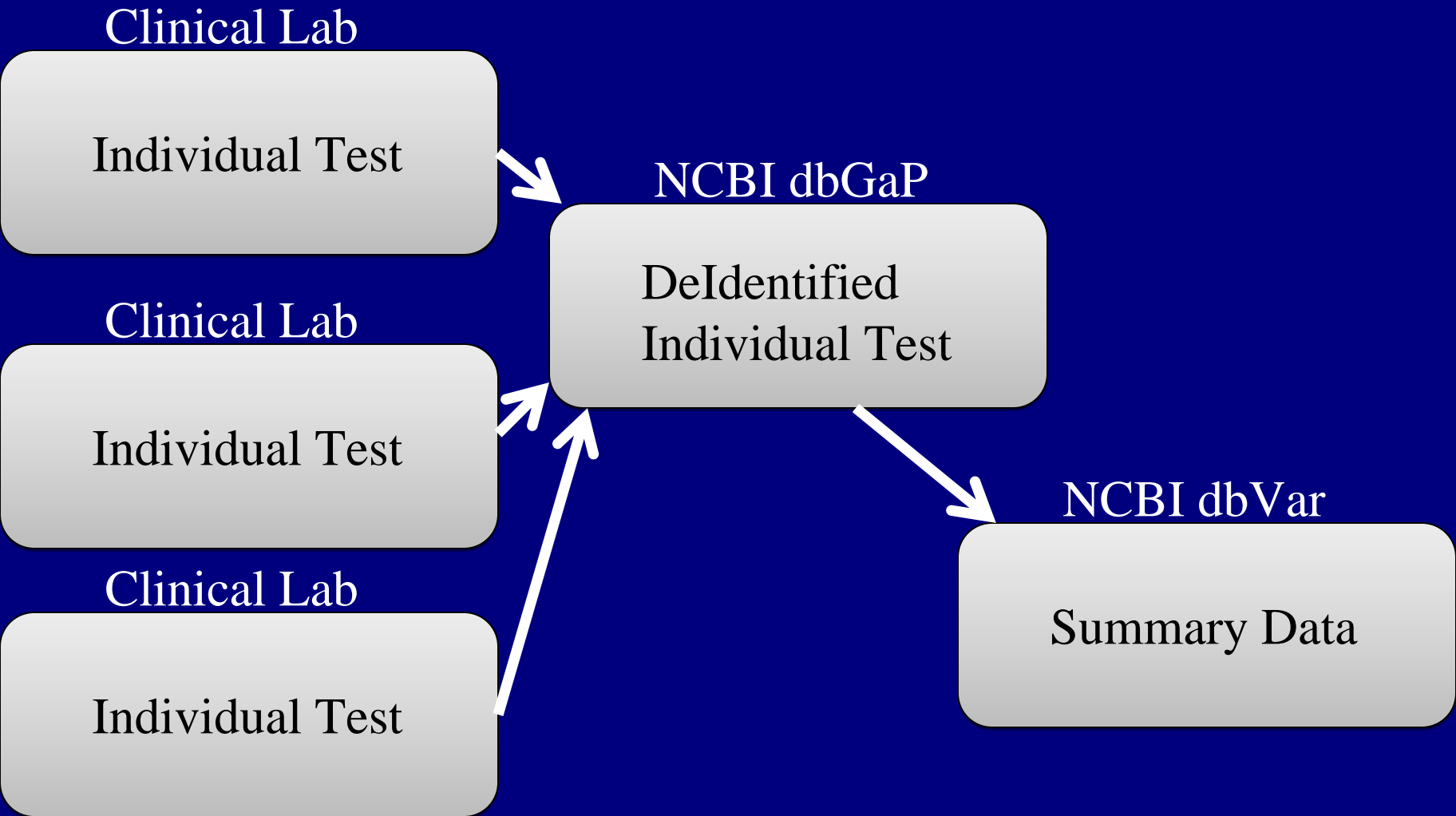
H2

Summary Thoughts

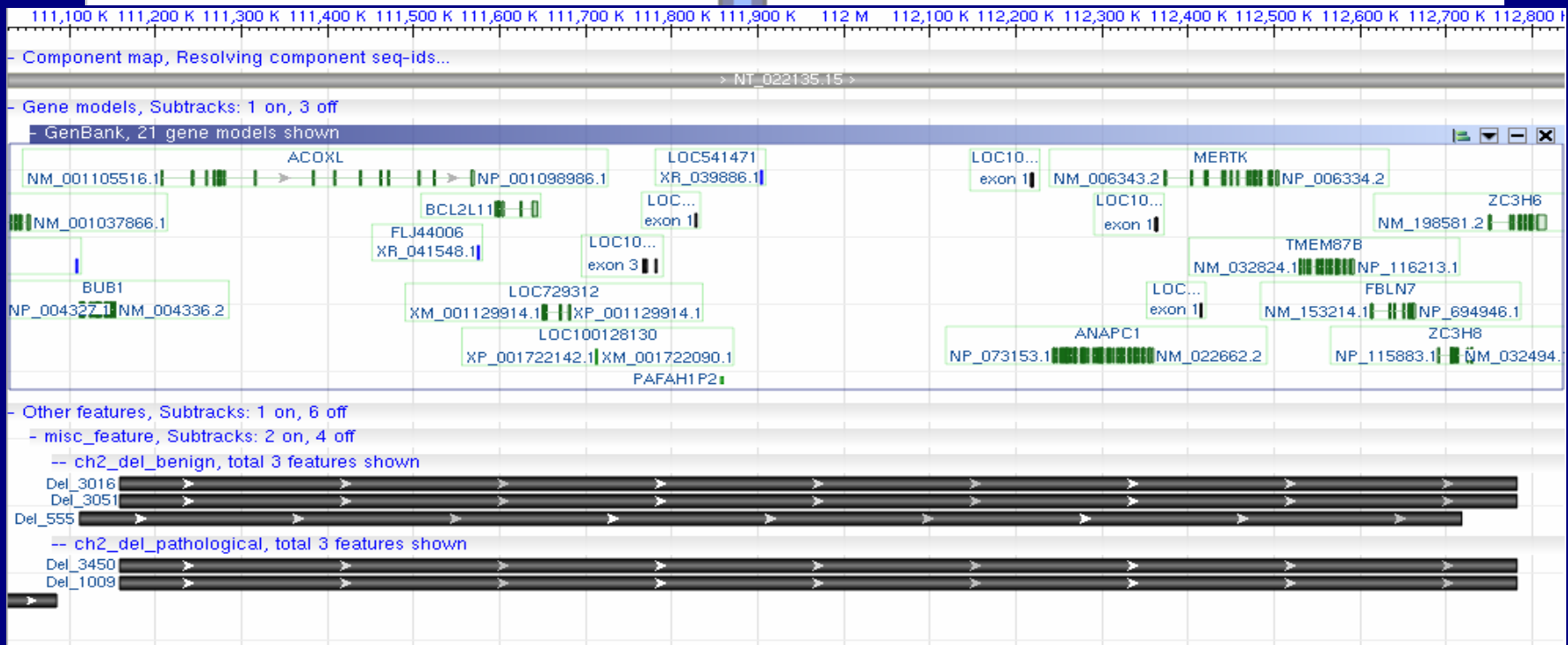
- Bioinformatics Formats come and go
 - Plan for change
- Versions Matter
- Archives Matter
- Focus on the Data
 - Define the Platform
 - Define the Observation
 - Then Add the Interpretation.. This may change.



From Clinical Testing



ISCA Review and Curation



Region 4: 111,000,000-113,000,000

bcNC->unce	555	2	111,112,291	112,718,099	del	bcNC	pCNC->unce	1009	del	pCNC	111,158,601	112,782,250
bcNC->unce	555	2	111,112,291	112,718,099	del	bcNC	pCNC->unce	3450	del	pCNC	111,158,601	112,782,250
bcNC->unce	3016	2	111,158,601	112,782,250	del	bcNC	pCNC->unce	1009	del	pCNC	111,158,601	112,782,250
bcNC->unce	3016	2	111,158,601	112,782,250	del	bcNC	pCNC->unce	3450	del	pCNC	111,158,601	112,782,250
bcNC->unce	3051	2	111,158,601	112,782,250	del	bcNC	pCNC->unce	1009	del	pCNC	111,158,601	112,782,250
bcNC->unce	3051	2	111,158,601	112,782,250	del	bcNC	pCNC->unce	3450	del	pCNC	111,158,601	112,782,250



More Steps Along the Way

