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

Web page views: 13 million per day

Web users: 1.6 million per day

Data downloaded: 4.1 TB per day

Peak web hits: 6900 per second
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 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**

  \[
  \text{Bioseq ::= SEQUENCE} \{
  \text{id \hspace{1em} SET OF Seq-id ,}
  \text{descr \hspace{1em} Seq-descr \hspace{1em} OPTIONAL,}
  \text{inst \hspace{1em} Seq-inst ,}
  \text{annot \hspace{1em} SET OF Seq-annot \hspace{1em} OPTIONAL}\}
  \]

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

  \[
  \text{Seq-id} \hspace{1em} 0 \hspace{1em} 1000
  \]
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 "AGCTGCCGCCTCAGGTCGATGTAACCCGATCC ..."
  }
  annot {
    ftable { ( ... ) }}
}

Bioseq ::= {
  id { gi 458032 },
  descr {
    title "aspartyl beta-hydroxylase",
    method concept-trans },
  inst {
    repr raw,
    mol aa,
    length 757,
    seq-data
    iupacaa "MAQRKNKSSGSSSSGSGSTGKHNGRRKGG ..."
  }
  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>
  <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>
    </Seq-descr>
  </Bioseq_descr>
  <Seqdesc>
    <Seqdesc_molinfo>
      <MolInfo>
        <MolInfo_biomol value="mRNA">3</MolInfo_biomol>
      </MolInfo>
    </Seqdesc_molinfo>
  </Seqdesc>
</Bioseq>
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., Fakultaet 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
AGCTGCCCGCTCGCTGTGTAACCCCCGCGCACTGAAGGAGGTCCACCAGCCCTCACCAGCCCCCGCGGA
CCGTGCAATGGCCCAGCGTAAGAATGCCAAGAGCAGCGGCAACAGCAGCAGCAGCGGCTCCGGCAGCGGT
AGCACGAGTGCGCGGACGACGACGCCCCGAGCAGAGAAGACATGGAGGACACAAGAATGGGAGGAAAGGCG
ACTCTCAGAATTTTGATGCTGAGATGGAGACAAAGGTTCAGCTGCACACCAGCGACGAGGAGGAGGACGAG
AGGTTCCTGTGGAGGCAAGCCCAGAATATCGAAGATGAAGCAAAAGAACAAATTCAGTCCCTTCTCCATGA
AAATGGTACACGCAACATGTTGAGGGAGAAGACTTGCAACAAGAAGATGGACCCACAGGAGAACCACAACA
AGGATGATGTTTCTTATGGCGACTGATGTAGATGTAGATGTAGATTTTGAGACCTTGGAACCTGAGATGAAG
AGATACATACCAATGTCTAGCTTACGTGAACAGCTAAGCTACGTGAACAGCTAAGCTACGTGAACAGCTA
AGAACTGCTCCCCCTGAGGATAATCCTGTAGAAGATTCACAGGTAATTGTAGAAGAAGTAAGCATTTTTC
CCTGTGGAAGAACAGCAGGAAGTACCACCAGAAACAAATAGAAAAACAGATGATCCAGAACAAAAAGCAA
AAGTTAAGAAAAAGAAGCCTAAAACCTTTTTAAATAAATTTTGATAAGACTATAAAGCT GTAACCGTTAGTGCT
AGAAACTCTCCGTATAAAGGGAAAATTGAGAACGTGAAATCCTGTAAGATTACCTACGTTAAATGTGAGAA
GATAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTAGAGTA
Embedding BioInformatics Formats
Bioinformatic Sequence Markup Language (BSML)

[January 10, 2001] In January 2001, LabGnome 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 LabGnome 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, analyses, 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, 2006] The proposed Bioinformatic Sequence Markup Language (BSML) is a public-domain protocol for graphical genomic displays. The project goals are similar to those of the Data Description Language, according to the NCBI 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 aimed at conforming with the XML standard.” Goals in the evolving 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 SBR (Small Business Innovative Research) from the National Center for Human Genome Research, to develop the public domain protocol. The SBR, 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 Bioinformatics." Visual Genetics, 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 centered 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 dataminning is made easy by the intuitive "point and click"
Error: This page it is not active.
BSML

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/.
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 Modeling 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 Ugo Giannini]

Microarray and Gene Expression Markup Language (MAGE-ML): Description from the Reused 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 expert and the input of the gene expression data as illustrated below. All hoc queries, when the XML files are directly accessible, can take advantage of the suite of XML recommendations, including XSLT or XQuery. Queries against repositories could be specified a number of ways, including through an IDE interface that had as its query language either of the above choices or SQL 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 automatic and efficiency to the parsing. The vocabulary of MAGE-ML is organized into subvocabularies in such a way that the subvocabularies are independent of each other. These subvocabularies 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 subvocabularies, such as both the BioAssay and BioArrayData 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.
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 OMIM (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), MAGE-TAB (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 AffyDesigns before their Arrays, or they may require that they be exported to separate files.
Representation of microarray expression data that would facilitate the exchange or mining of information between different data systems.

**MAGE-TAB**


**Other Detailed Information**

- **Technical**
  - **XML**
  - **Gene Expression Markup Language (MAGE)**
  - **Microarray Gene Expression Markup Language (MAGE-ML)**
  - **MAGE-OM**
  - **MAGE-ML**

- **Guidelines**
  - **Mimicking**

Please subscribe to the MAGE mailing lists from [here](http://www.mged.org/mage-tab/).
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

NCBI

<subjectOf2>
  <geneticLocus>
    <component1>
      <individualAllele moodCode="EVN">
        <text>breast cancer 1, early onset</text>
      </individualAllele>
      <value code="83990" displayName="BRCA1" codeSystemName="NCBI Entrez">
        <translation code="20473" displayName="BRCA1" codeSystem="HGNC"/>
      </value>
    </component1>
    <component2>
      <sequence moodCode="EVN">
        <code code="BSMLcon3"/>
        <value mediaType="text/xml">
          <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">
                GCTCCCA CTCCATGAGG TATTTCTCA
                CATCCGTGTC CGGCCGCCGC CGCGGGGAGC CCGGCTTCAT CGCCGTGGGC
                TAGTGGGACG ACCAGCGATT CGTGGGTTCA GACAGCGACG CGCGAGGCA
                GAGGGTGGAG CCGGCGGCGC CCGGATTAGA GAGGGGGGC CCGGAGTTT
                GGGGCAAGA GACACCGGAAT GTGGAAGGCC AGTCACAGAC TGAGCGATGT
                GACCTGCGGA CCGTGCAGGC CTAATCAAC CAGAGCGAGG CCG
              </bsml:Seq-data>
            </bsml:Sequence>
          </bsml:Sequences>
        </value>
      </sequence>
    </component2>
  </geneticLocus>
</subjectOf2>
<subjectOf2>
  <geneticLocus>
    <component1>
      <individualAllele moodCode="EVI">
        <text>breast cancer 1, early onset</text>
        <value valueCode="83990" display_Name="BRCA1" codeSystemName="NCBI Entrez"/>
        <translation valueCode="20473" display_Name="BRCA1" codeSystemName="HGNC"/>
      </component1>
    </geneticLocus>
  </subjectOf2>
  <sequence moodCode="EVI">
    <code code="BSMLcon3">
      <value mediaType="text">
        <bsml:Definitions>
          <bsml:Sequence id="seq2" molecule="dna" ic-ackey="U14680 REGION: 101..199" title="BRCA1, exon 3" representation="raw" local-ackey="this could be used by the genetic lab">
            <bsml:Seq-data>
              GCTCCACCTCCAAGAGGTTTCTACA CATCCGTGTC CCAGCCCGGC CGCGGGAGAC CGCGCTCAT CGCCGTGGGC TACGTGGACG ACACGCAGTT GTGCGGGTCC GACAGCGACG CCGCGAGCCA GAGGATGGAG CGCGGAGGCC CGTGATAGA GCAGGAGGGG CCGGAGTATT GGGACAGGGA GACACGGAAT GGAAGGGCC AGTCACAAGAC TGACCGAGTG GACCTGGGGA CCGCTGCAGG CTACTACAAC CAGAGCGAGG CCG
            </bsml:Seq-data>
          </bsml:Sequence>
        </bsml:Definitions>
      </value>
    </code>
  </sequence>
</subjectOf2>
<table>
<thead>
<tr>
<th>LOCUS</th>
<th>HSU03109</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Human aspartyl beta-hydroxylase mRNA, complete cds.</td>
</tr>
<tr>
<td>ACCESSION</td>
<td>U03109</td>
</tr>
<tr>
<td>VERSION</td>
<td>U03109.1 GI:458031</td>
</tr>
<tr>
<td>KEYWORDS</td>
<td>.</td>
</tr>
<tr>
<td>SOURCE</td>
<td>Homo sapiens (human)</td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td></td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.</td>
</tr>
<tr>
<td>REFERENCE</td>
<td>1 (bases 1 to 2249)</td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Korioth,F., Gieffers,C. and Frey,J.</td>
</tr>
<tr>
<td>TITLE</td>
<td>Cloning and characterization of the human gene encoding aspartyl beta-hydroxylase</td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Gene 150 (2), 395-399 (1994)</td>
</tr>
<tr>
<td>PUBMED</td>
<td>7821814</td>
</tr>
<tr>
<td>REFERENCE</td>
<td>2 (bases 1 to 2449)</td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Korioth,F.</td>
</tr>
<tr>
<td>TITLE</td>
<td>Direct Submission</td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Submitted (03-NOV-1993) Korioth F., Fakultaet fuer Chemie-Biochemie II, Universitaet Bielefeld, Universitaetsstrasse 25, Bielefeld, 33615, Germany</td>
</tr>
<tr>
<td>FEATURES</td>
<td>Location/Qualifiers</td>
</tr>
<tr>
<td>source</td>
<td>1..2449</td>
</tr>
<tr>
<td></td>
<td>/organism=&quot;Homo sapiens&quot;</td>
</tr>
<tr>
<td></td>
<td>/mol_type=&quot;mRNA&quot;</td>
</tr>
<tr>
<td></td>
<td>/db_xref=&quot;taxon:9606&quot;</td>
</tr>
<tr>
<td></td>
<td>/clone=&quot;As-5&quot;</td>
</tr>
</tbody>
</table>
**Forget Format – Let’s Look at Content**

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>HSU03109</th>
<th>2449 bp mRNA</th>
<th>linear</th>
<th>PRI 30-NOV-1995</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESSION</td>
<td>U03109</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VERSION</td>
<td>U03109.1</td>
<td>GI:458031</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOURCE</td>
<td>Homo sapiens (human)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukaryota;</td>
<td>Metazoa; Chordata;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Craniata;</td>
<td>Vertebrata; Euteleostomi;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mammalia;</td>
<td>Eutheria; Euarchontoglires; Primates; Haplorrhini;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catarrhini;</td>
<td>Hominidae; Homo.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>1 (bases 1 to 2249)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Korioth, F., Gieffers, C. and Frey, J.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Cloning and characterization of the human gene encoding aspartyl beta-hydroxylase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Gene 150 (2), 395-399 (1994)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PUBMED</td>
<td>7821814</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REFERENCE</td>
<td>2 (bases 1 to 2449)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUTHORS</td>
<td>Korioth, F.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Direct Submission</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JOURNAL</td>
<td>Submitted (03-NOV-1993) Korioth F., Fakultaet fuer Chemie-Biochemie II, Universitaet Bielefeld, Universitaetsstrasse 25, Bielefeld, 33615, Germany</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEATURES</td>
<td>Location/Qualifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source</td>
<td>1..2449</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/organism</td>
<td>&quot;Homo sapiens&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/mol_type</td>
<td>&quot;mRNA&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/db_xref</td>
<td>&quot;taxon:9606&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/clone</td>
<td>&quot;As-5&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Why Versions Matter

U12345: position 10 is “C”
U12345.1: position 10 is “C”

U12345
U12345.1

AGCTGCCCGCGTCGCGTGACCCCCGGCGACGTGA
Why Versions Matter

U12345: position 10 is “C”
U12345.1: position 10 is “C”

U12345
U12345.1
AGCTGCCCCGCGTGTGCTGCTGCTGCTGCTGACCCCCCGCGCCTG

U12345
U12345.2
TAGCTGCCCCGCGTGTGCTGCTGCTGCTGCTGACCCCCCGCGCCTG
Why Versions Matter

U12345: position 10 is “C”
U12345.1: position 10 is “C”

U12345
AGCTGCCCGCGTCGCGTGTGTACCCCCCGCGCACTGA
U12345.1

U12345: position 10 is “G”

U12345
TAGCTGCCCCGCGTGCATTGTGTACCCCCCGCGCAGCTG
U12345.2
**Why Versions Matter**

U12345: position 10 is “C”
U12345.1: position 10 is “C”

U12345: position 10 is “G”
U12345.1: position 10 is still “C”

U12345: position 10 is “G”
U12345.2: position 10 is still “C”
Another Reason Why Live Archives Are Important
OBX|3|CWE|51958-7^Transcript reference sequence identifier^LN||NM_170707.1^^2.16.840.1.113883.6.280||||||F|20080702100909|||||||Laboratory for Molecular Medicine^L^22D1005307^^^CLIA&2.16.840.1.113883.4.7&ISO|1000 Laboratory Lane^Ste.123^Cambridge^MA^99999^USA^B
OBX|3|CWE|51958-7^Transcript reference sequence identifier^LN|NM_170707.1^2.16.840.1.113883.6.280|||F|20080702100909||Laboratory for Molecular Medicine^L^22D1005307^^CLIA&2.16.840.1.113883.4.7&ISO|1000 Laboratory Lane^Ste.123^Cambridge^MA^99999^USA^B
The Changing Genome

1,548,122
The Changing Genome

Insert 100 bp

1,548,222
RefSeqGene/LRG

Insert 100 bp
The Genome Reference Consortium

The Genome Reference Consortium consists of:

- The Wellcome Trust Sanger Institute
- The Genome Center at Washington University
- The European Bioinformatics Institute
- 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.

GRCs77 is now available in Map Viewer
Fri, 14 Aug 2009
NCBI has annotated and released the latest version of the public human genome assembly (GRCs77).

Resolved Issues

Mouse (MG-3729)
Sep 03, 2010
AC113082.6 has been updated to AC113082.7 (vector sequence removed). Fixing the half-dovetail joins between AC113082 and CT025561.

Human (HG-858)
Sep 22, 2010
Certificate has been submitted for the overlap between AC026399.21 and AC215219.3.
The Genome Changes
17q21.31 microdeletion syndrome
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

Clinical Lab

Individual Test

NCBI dbGaP

DeIdentified Individual Test

Clinical Lab

Individual Test

NCBI dbVar

Summary Data

Clinical Lab

Individual Test
More Steps Along the Way