Introduction to Biological Databases

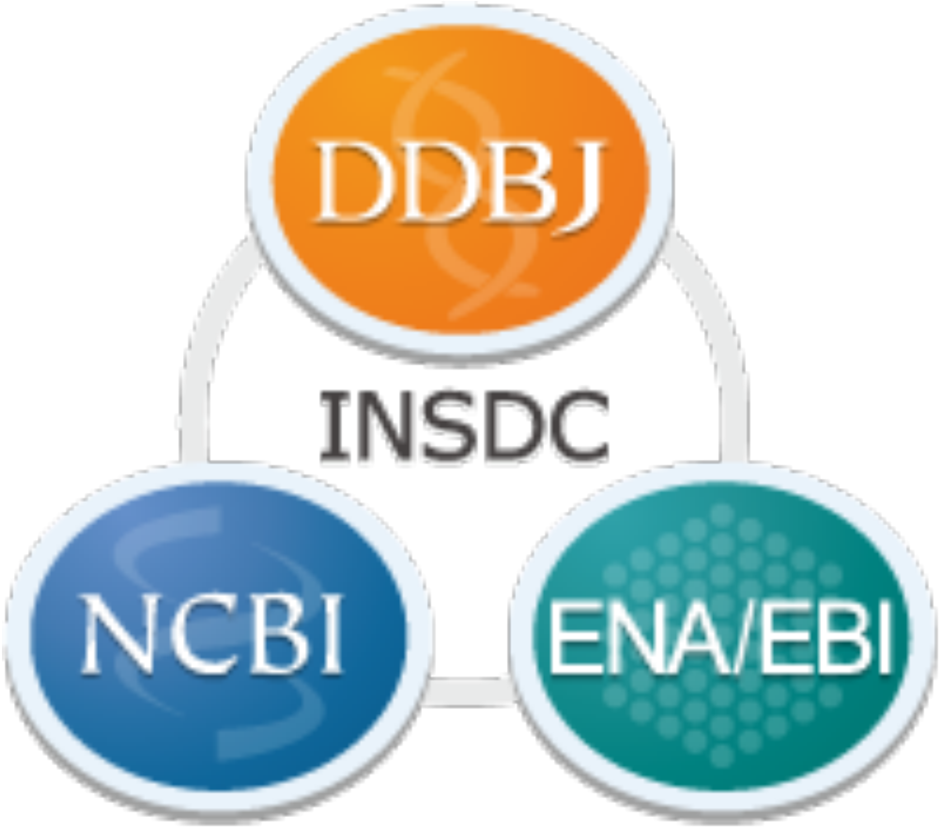
Biological Databases

a repository of data collected from scientific experiments, published literature, high-throughput technology, or computational analyses

- host of different research areas
- provide a single point of access

International Nucleotide Sequence Database Collaboration

- A longstanding global initiative
- Creating a comprehensive collection of public domain nucleotide sequences and metadata
- Worldwide synchronization of sequence data (submit anywhere; daily updates)
- Currently include sequence data from >160,000 species



EMBL-EBI

European Nucleotide Archive, established in 1980 at EMBL in Heidelberg, Germany

- Central database of DNA sequences
- EMBL-EBI established in UK on Wellcome Trust campus 1992; transition of two major bioinformatics services
- Provide a comprehensive range of molecular databases and offer extensive user training programme

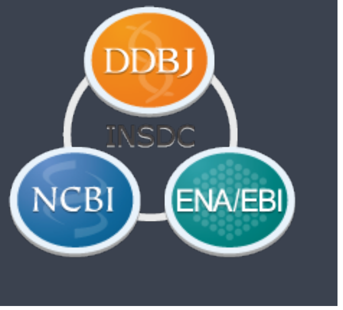
DDBJ
INSDC
NCBI
ENA/EBI

NCBI National Center for Biotechnology Information

In 1984 briefing sessions began on Capitol Hill; Dr. Allan Maxam germinated the idea of the Center

Created in 1988 as part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biological information



DDBJ

DNA Data Bank of Japan

EMBL and Genbank started international cooperation and invited Japan to participate

- trial data loading was started in 1983, and the next year DDBJ as up and running as part of the National Genetics Institute (NGI)
- To operate DDBJ more efficiently, the Center for Information Biology (CIB) was established within the NIG in 1995
- Services include biological database management and various software tools for data analysis (web services and NIG supercomputer)

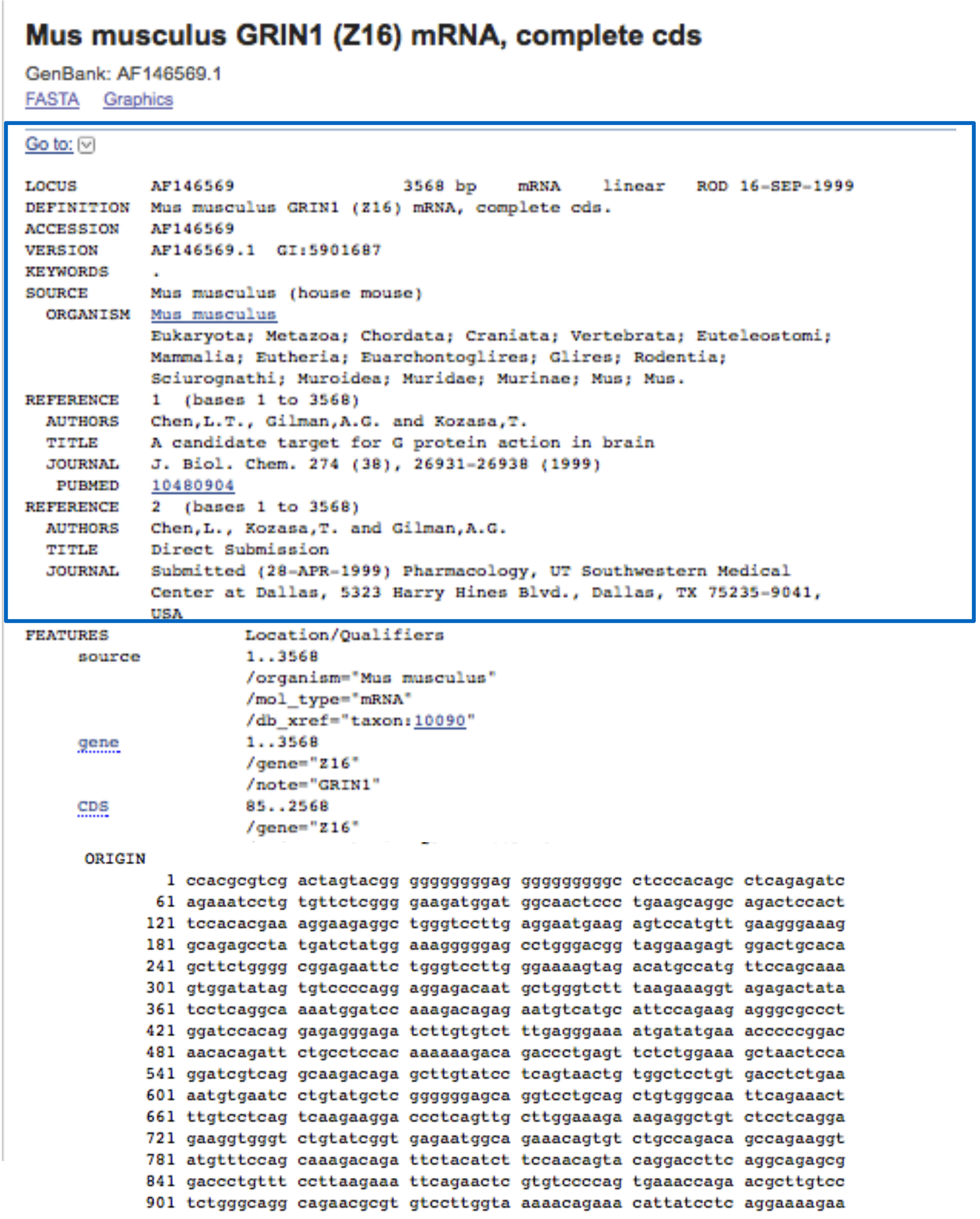
DDBJ
INSDC
NCBI
ENA/EBI

Sequence data

- Contain data from individual organisms, specific categories/ functions of sequences, or data generated by specific sequencing technologies. Examples:
  - Organism-specific
  - Sequence categories or functions
  - Data generated by specific sequencing technologies (EST, STS, HTG)

Flatfile format for INDSC:

- Header



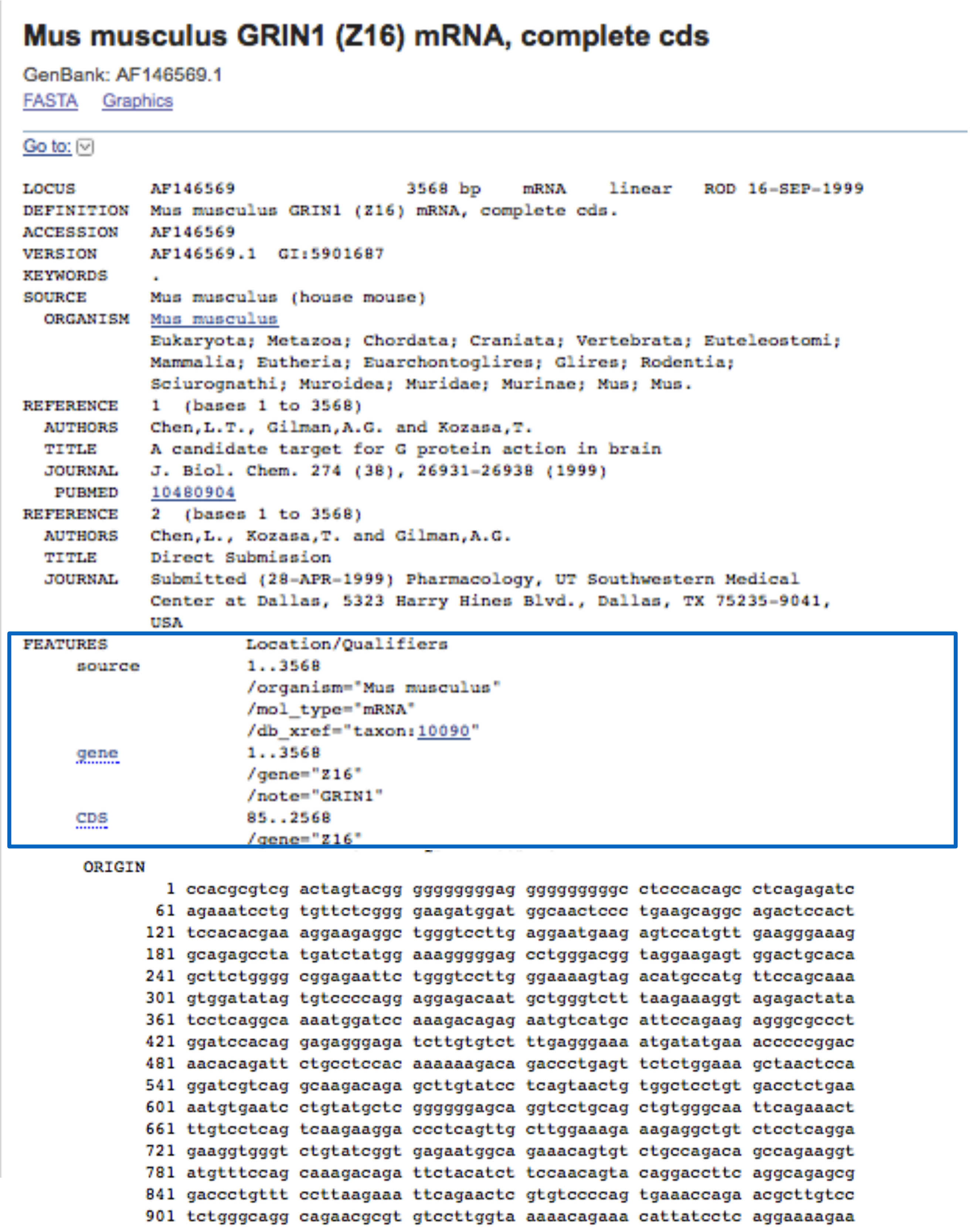
Mus musculus GRIN1 (Z16) mRNA, complete cds

GenBank: AF146569.1

FASTA Graphics

Go to:

```
LOCUS       AF146569                3568 bp    mRNA    linear   ROD 16-SEP-1999
DEFINITION  Mus musculus GRIN1 (Z16) mRNA, complete cds.
ACCESSION   AF146569
VERSION     AF146569.1  GI:5901687
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 3568)
  AUTHORS   Chen,L.T., Gilman,A.G. and Kozasa,T.
  TITLE     A candidate target for G protein action in brain
  JOURNAL   J. Biol. Chem. 274 (38), 26931-26938 (1999)
   PUBMED   10480904
REFERENCE   2  (bases 1 to 3568)
  AUTHORS   Chen,L., Kozasa,T. and Gilman,A.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-APR-1999) Pharmacology, UT Southwestern Medical
            Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235-9041,
            USA
FEATURES             Location/Qualifiers
     source          1..3568
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10090"
     gene            1..3568
                     /gene="Z16"
                     /note="GRIN1"
     CDS             85..2568
                     /gene="Z16"
      ORIGIN
        1 ccacgcgtcg actagtacgg gggggggag gggggggggc ctcccacagc ctcagagatc
       61 agaaatcctg tgttctcggg gaagatggat ggcaactccc tgaagcaggc agactccact
      121 tccacacgaa aggaagaggc tgggtccttg aggaatgaag agtccatgtt gaagggaaag
      181 gcagagccta tgatctatgg aaaggggag cctgggacgg taggaagagt ggactgcaca
      241 gcttctgggg cggagaattc tgggtccttg ggaaaagtag acatgccatg ttccagcaaa
      301 gtggatatag tgtccccagg aggagacaat gctgggtctt taagaaaggt agagactata
      361 tcctcaggca aaatggatcc aaagacagag aatgtcatgc attccagaag agggcgccct
      421 ggatccacag gagagggaga tcttgtgtct ttgagggaaa atgatatgaa acccccggac
      481 aacacagatt ctgcctccac aaaaaagaca gaccctgagt tctctggaaa gctaactcca
      541 ggatcgtcag gcaagacaga gcttgtatcc tcagtaactg tggctcctgt gacctctgaa
      601 aatgtgaatc ctgtatgctc gggggagca ggtcctgcag ctgtgggcaa ttcagaaact
      661 ttgtcctcag tcaagaagga ccctcagttg cttggaaaga aagaggctgt ctcctcagga
      721 gaaggtgggt ctgtatcggt gagaatggca gaaacagtgt ctgccagaca gccagaaggt
      781 atgtttccag caaagacaga ttctacatct tccaacagta caggaccttc aggcagagcg
      841 gaccctgttt ccttaagaaa ttcagaactc gtgtccccag tgaaaccaga acgcttgtcc
      901 tctgggcagg cagaacgcgt gtccttggta aaaacagaaa cattatcctc aggaaaagaa
```



Mus musculus GRIN1 (Z16) mRNA, complete cds

GenBank: AF146569.1

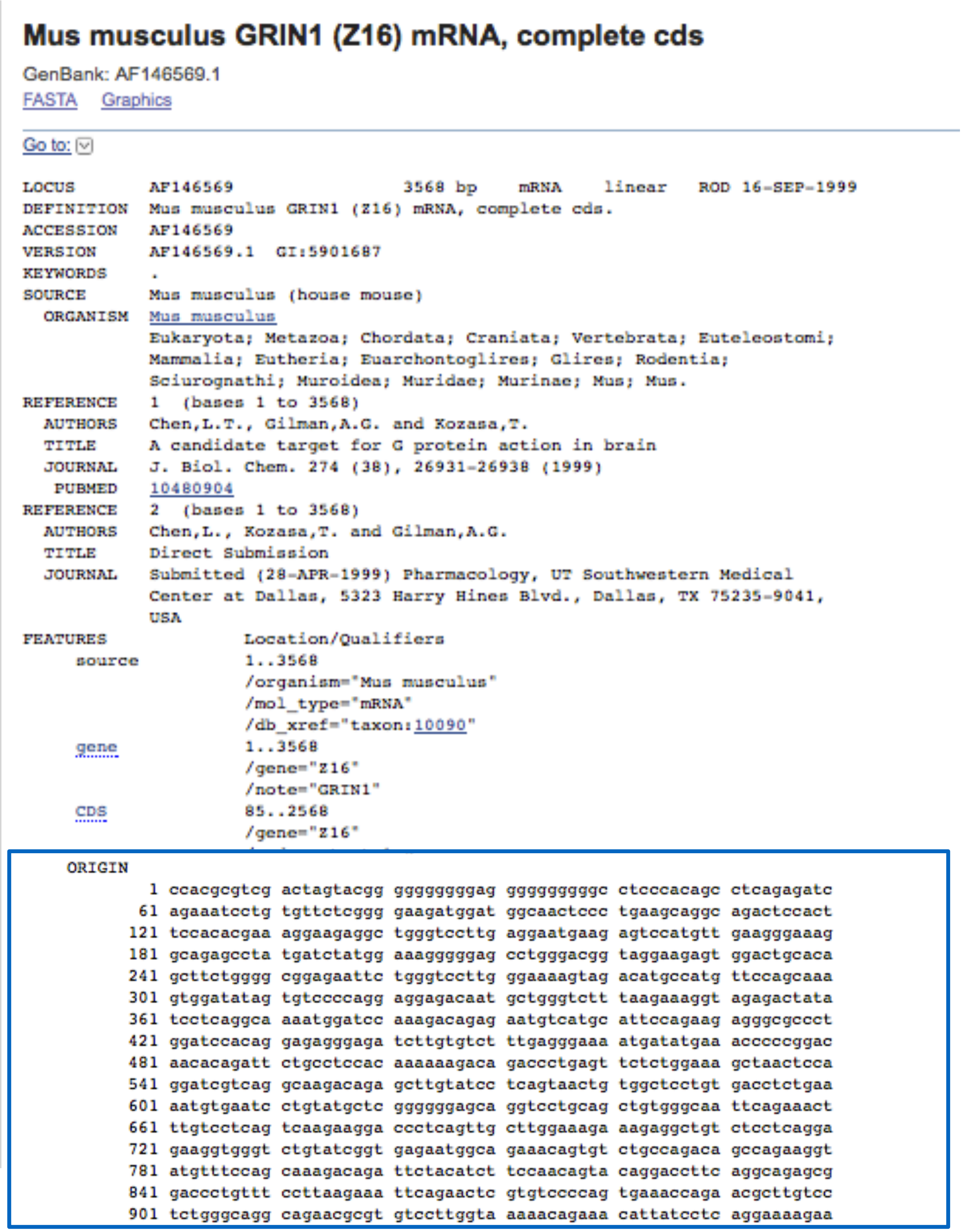
FASTA Graphics

Go to:

```
LOCUS       AF146569                3568 bp    mRNA    linear   ROD 16-SEP-1999
DEFINITION  Mus musculus GRIN1 (Z16) mRNA, complete cds.
ACCESSION   AF146569
VERSION     AF146569.1  GI:5901687
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 3568)
  AUTHORS   Chen,L.T., Gilman,A.G. and Kozasa,T.
  TITLE     A candidate target for G protein action in brain
  JOURNAL   J. Biol. Chem. 274 (38), 26931-26938 (1999)
   PUBMED   10480904
REFERENCE   2  (bases 1 to 3568)
  AUTHORS   Chen,L., Kozasa,T. and Gilman,A.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-APR-1999) Pharmacology, UT Southwestern Medical
            Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235-9041,
            USA
FEATURES             Location/Qualifiers
     source          1..3568
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10090"
     gene            1..3568
                     /gene="Z16"
                     /note="GRIN1"
     CDS             85..2568
                     /gene="Z16"
      ORIGIN
        1 ccacgcgtcg actagtacgg ggggggggag gggggggggc ctcccacagc ctcagagatc
       61 agaaatcctg tgttctcggg gaagatggat ggcaactccc tgaagcaggc agactccact
      121 tccacacgaa aggaagaggc tgggtccttg aggaatgaag agtccatgtt gaagggaaag
      181 gcagagccta tgatctatgg aaagggggag cctgggacgg taggaagagt ggactgcaca
      241 gcttctgggg cggagaattc tgggtccttg ggaaaagtag acatgccatg ttccagcaaa
      301 gtggatatag tgtccccagg aggagacaat gctgggtctt taagaaaggt agagactata
      361 tcctcaggca aaatggatcc aaagacagag aatgtcatgc attccagaag agggcgccct
      421 ggatccacag gagagggaga tcttgtgtct ttgagggaaa atgatatgaa acccccggac
      481 aacacagatt ctgcctccac aaaaaagaca gaccctgagt tctctggaaa gctaactcca
      541 ggatcgtcag gcaagacaga gcttgtatcc tcagtaactg tggctcctgt gacctctgaa
      601 aatgtgaatc ctgtatgctc gggggagca ggtcctgcag ctgtgggcaa ttcagaaact
      661 ttgtcctcag tcaagaagga ccctcagttg cttggaaaga aagaggctgt ctcctcagga
      721 gaaggtgggt ctgtatcggt gagaatggca gaaacagtgt ctgccagaca gccagaaggt
      781 atgtttccag caaagacaga ttctacatct tccaacagta caggaccttc aggcagagcg
      841 gaccctgttt ccttaagaaa ttcagaactc gtgtccccag tgaaaccaga acgcttgtcc
      901 tctgggcagg cagaacgcgt gtccttggta aaaacagaaa cattatcctc aggaaaagaa
```

Flatfile format for INDSC:

- Header
- Feature Table: details about the sequence



Mus musculus GRIN1 (Z16) mRNA, complete cds

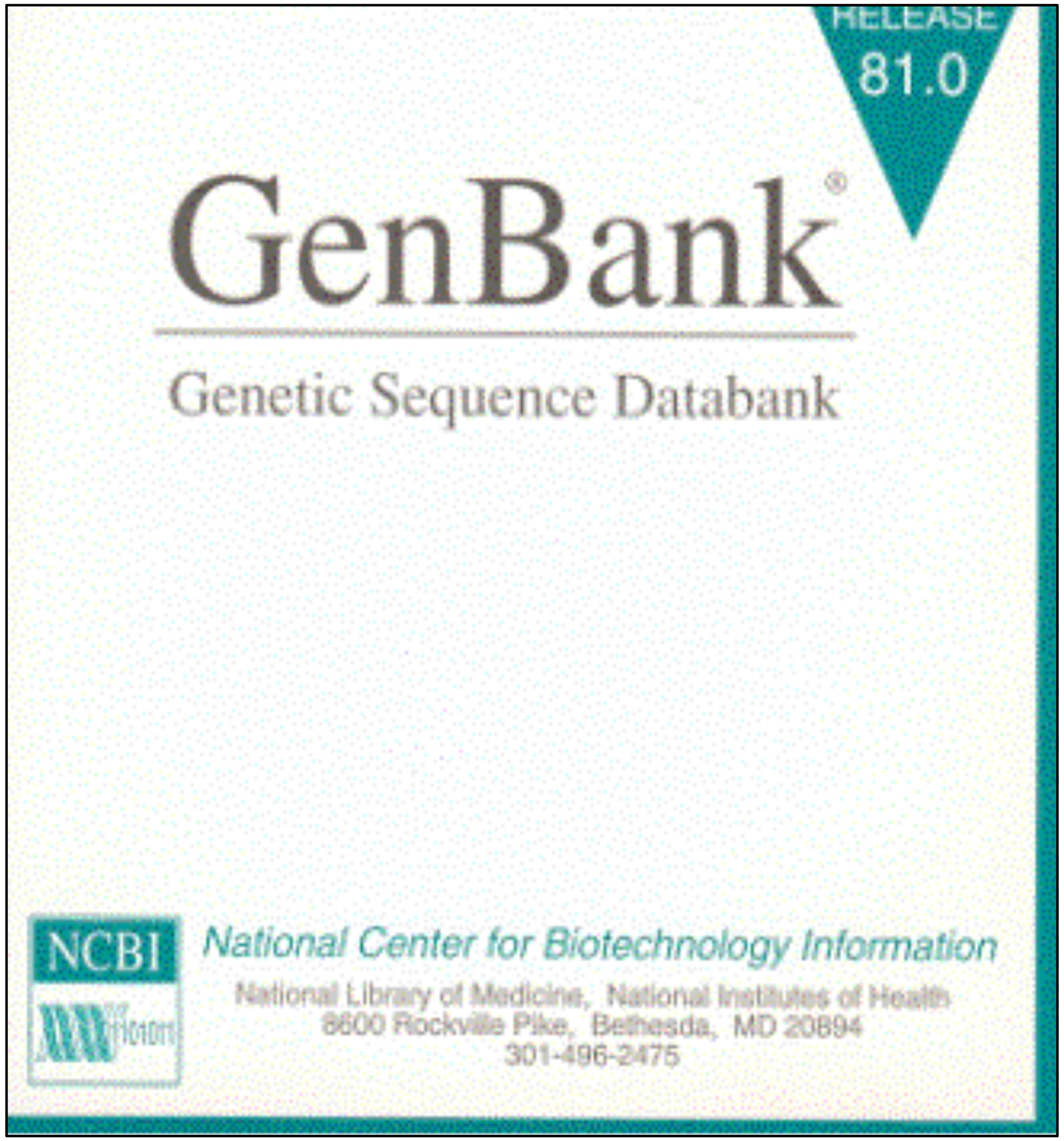
GenBank: AF146569.1

```
LOCUS       AF146569                3568 bp    mRNA    linear   ROD 16-SEP-1999
DEFINITION  Mus musculus GRIN1 (Z16) mRNA, complete cds.
ACCESSION   AF146569
VERSION     AF146569.1  GI:5901687
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 3568)
  AUTHORS   Chen,L.T., Gilman,A.G. and Kozasa,T.
  TITLE     A candidate target for G protein action in brain
  JOURNAL   J. Biol. Chem. 274 (38), 26931-26938 (1999)
   PUBMED   10480904
REFERENCE   2  (bases 1 to 3568)
  AUTHORS   Chen,L., Kozasa,T. and Gilman,A.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-APR-1999) Pharmacology, UT Southwestern Medical
            Center at Dallas, 5323 Harry Hines Blvd., Dallas, TX 75235-9041,
            USA
FEATURES             Location/Qualifiers
     source          1..3568
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10090"
     gene            1..3568
                     /gene="Z16"
                     /note="GRIN1"
     CDS             85..2568
                     /gene="Z16"
```

```
    ORIGIN
        1 ccacgcgtcg actagtacgg gggggggag ggggggggc ctcccacagc ctcagagatc
       61 agaaatcctg tgttctcggg gaagatggat ggcaactccc tgaagcaggc agactccact
      121 tccacacgaa aggaagaggc tgggtccttg aggaatgaag agtccatgtt gaagggaaag
      181 gcagagccta tgatctatgg aaagggggag cctgggacgg taggaagagt ggactgcaca
      241 gcttctgggg cggagaattc tgggtccttg ggaaaagtag acatgccatg ttccagcaaa
      301 gtggatatag tgtccccagg aggagacaat gctgggtctt taagaaaggt agagactata
      361 tcctcaggca aaatggatcc aaagacagag aatgtcatgc attccagaag agggcgccct
      421 ggatccacag gagagggaga tcttgtgtct ttgagggaaa atgatatgaa acccccggac
      481 aacacagatt ctgcctccac aaaaaagaca gaccctgagt tctctggaaa gctaactcca
      541 ggatcgtcag gcaagacaga gcttgtatcc tcagtaactg tggctcctgt gacctctgaa
      601 aatgtgaatc ctgtatgctc gggggagca ggtcctgcag ctgtgggcaa ttcagaaact
      661 ttgtcctcag tcaagaagga ccctcagttg cttggaaaga aagaggctgt ctcctcagga
      721 gaaggtgggt ctgtatcggt gagaatggca gaaacagtgt ctgccagaca gccagaaggt
      781 atgtttccag caaagacaga ttctacatct tccaacagta caggaccttc aggcagagcg
      841 gaccctgttt ccttaagaaa ttcagaactc gtgtccccag tgaaaccaga acgcttgtcc
      901 tctgggcagg cagaacgcgt gtccttggta aaaacagaaa cattatcctc aggaaaagaa
```

Flatfile format for INDSC:

- Header
- Feature Table: details about the sequence
- Sequence


RELEASE
81.0
GenBank®
Genetic Sequence Databank
NCBI
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD 20894
301-496-2475

Library Series
GenBank
Genetic Sequence Databank
GenBank Flat File Format
February 15, 1994
Release 81.0
COMPACT
DATA STORAGE
NCBI
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Bethesda, MD 20894
301-496-2475

An explosion of sequence data

ftp://ftp.ncbi.nih.gov/genbank/

Release 212.0	February 15, 2015
190,250,235	Reported Sequences
1,399,865,495,608	Total Bases

- full release every two months
- incremental updates daily
- available only via ftp

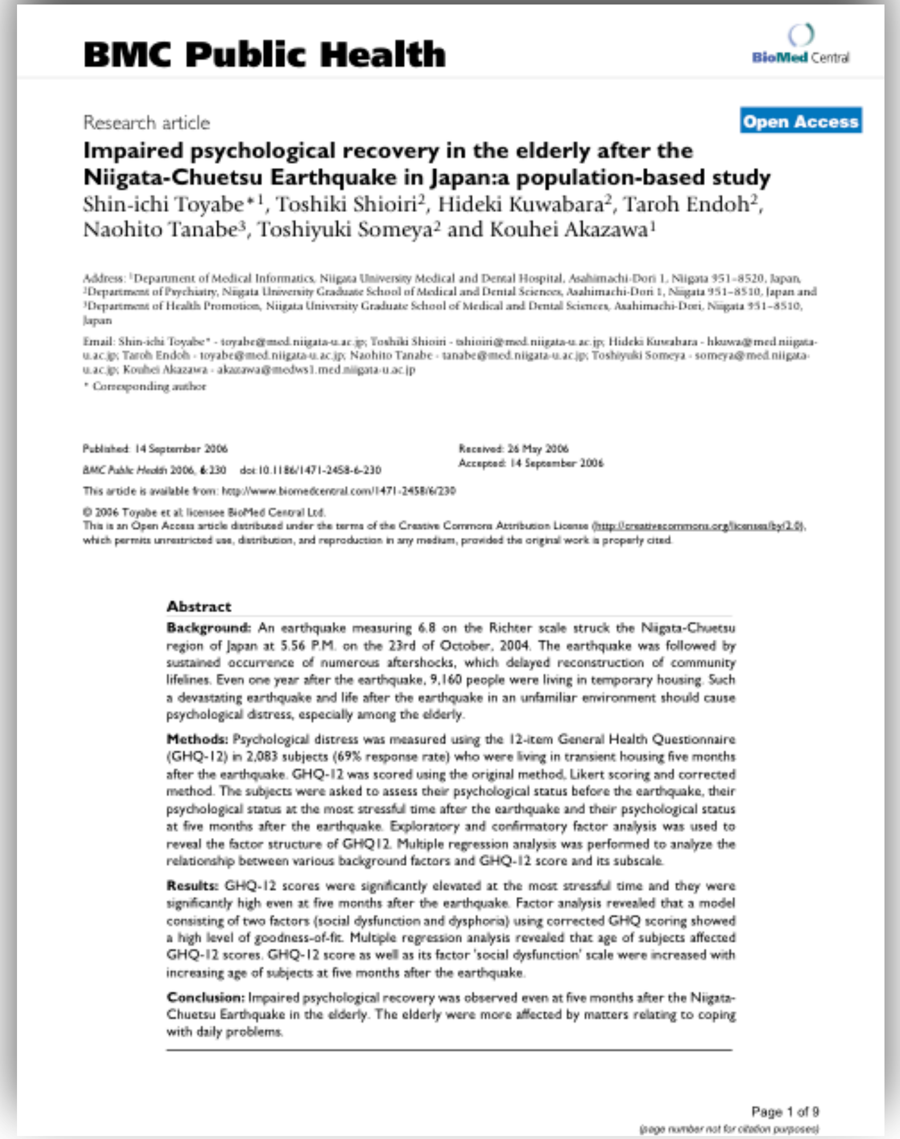
Archival data

- repository of information
- redundant; might have many sequence records for the same gene, each from a different lab
- submitters maintain editorial control over their records:
- what goes in is what comes out
- no controlled vocabulary
- variation in annotation of biological features

Curated data

- non-redundant; one record for each gene, or each splice variant
- each record is intended to present an encapsulation of the current understanding of a gene or protein
- records contain value-added information that have been added by an expert(s)

Archival data



BMC Public Health

BioMed Central

Research article

Open Access

Impaired psychological recovery in the elderly after the Niigata-Chuetsu Earthquake in Japan:a population-based study

Shin-ichi Toyabe*[1], Toshiki Shioiri[2], Hideki Kuwabara[2], Taroh Endoh[2], Naohito Tanabe[3], Toshiyuki Someya[2] and Kouhei Akazawa[1]

Address: [1]Department of Medical Informatics, Niigata University Medical and Dental Hospital, Asahimachi-Dori 1, Niigata 951-8520, Japan, [2]Department of Psychiatry, Niigata University Graduate School of Medical and Dental Sciences, Asahimachi-Dori 1, Niigata 951-8510, Japan and [3]Department of Health Promotion, Niigata University Graduate School of Medical and Dental Sciences, Asahimachi-Dori, Niigata 951-8510, Japan

Email: Shin-ichi Toyabe* - toyabe@med.niigata-u.ac.jp; Toshiki Shioiri - tshioiri@med.niigata-u.ac.jp; Hideki Kuwabara - hkuwa@med.niigata-u.ac.jp; Taroh Endoh - toyabe@med.niigata-u.ac.jp; Naohito Tanabe - tanabe@med.niigata-u.ac.jp; Toshiyuki Someya - someya@med.niigata-u.ac.jp; Kouhei Akazawa - akazawa@medws1.med.niigata-u.ac.jp

* Corresponding author

Published: 14 September 2006

BMC Public Health 2006, **6**:230 doi:10.1186/1471-2458-6-230

Received: 26 May 2006
Accepted: 14 September 2006

This article is available from: http://www.biomedcentral.com/1471-2458/6/230

Abstract

Background: An earthquake measuring 6.8 on the Richter scale struck the Niigata-Chuetsu region of Japan at 5.56 P.M. on the 23rd of October, 2004. The earthquake was followed by sustained occurrence of numerous aftershocks, which delayed reconstruction of community lifelines. Even one year after the earthquake, 9,160 people were living in temporary housing. Such a devastating earthquake and life after the earthquake in an unfamiliar environment should cause psychological distress, especially among the elderly.

Methods: Psychological distress was measured using the 12-item General Health Questionnaire (GHQ-12) in 2,083 subjects (69% response rate) who were living in transient housing five months after the earthquake. GHQ-12 was scored using the original method, Likert scoring and corrected method. The subjects were asked to assess their psychological status before the earthquake, their psychological status at the most stressful time after the earthquake and their psychological status at five months after the earthquake. Exploratory and confirmatory factor analysis was used to reveal the factor structure of GHQ12. Multiple regression analysis was performed to analyze the relationship between various background factors and GHQ-12 score and its subscale.

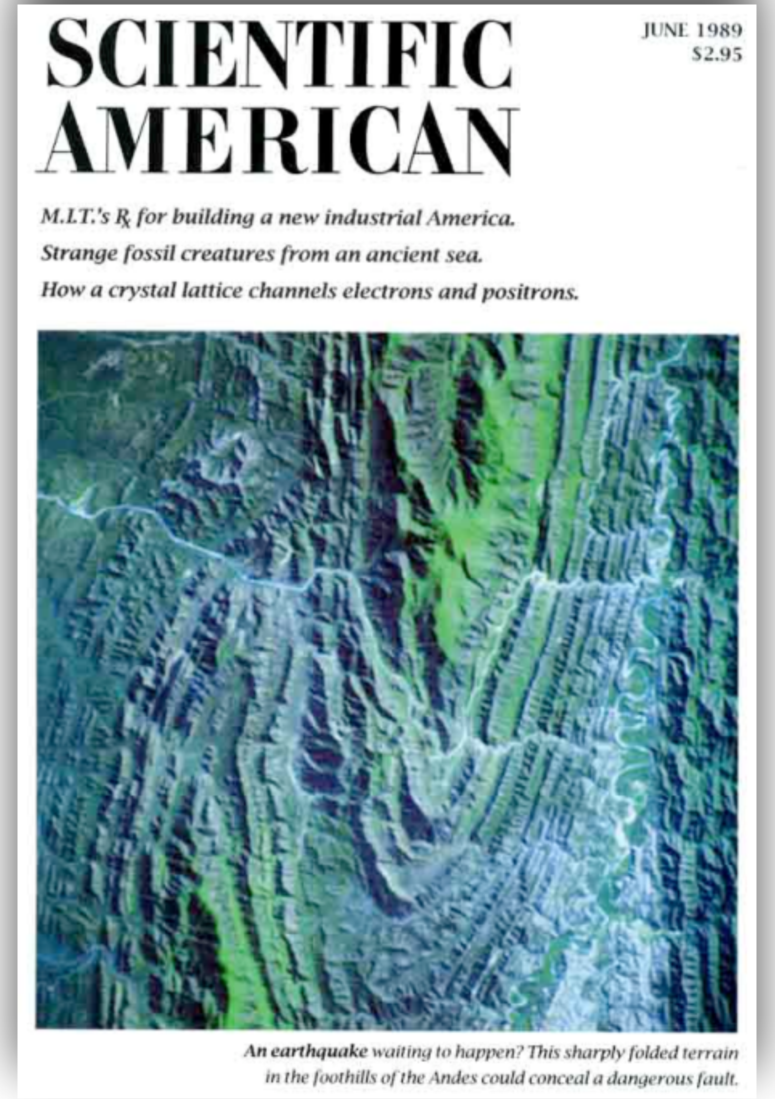
Results: GHQ-12 scores were significantly elevated at the most stressful time and they were significantly high even at five months after the earthquake. Factor analysis revealed that a model consisting of two factors (social dysfunction and dysphoria) using corrected GHQ scoring showed a high level of goodness-of-fit. Multiple regression analysis revealed that age of subjects affected GHQ-12 scores. GHQ-12 score as well as its factor 'social dysfunction' scale were increased with increasing age of subjects at five months after the earthquake.

Conclusion: Impaired psychological recovery was observed even at five months after the Niigata-Chuetsu Earthquake in the elderly. The elderly were more affected by matters relating to coping with daily problems.

Page 1 of 9
(page number not for citation purposes)

vs

Curated data



SCIENTIFIC AMERICAN

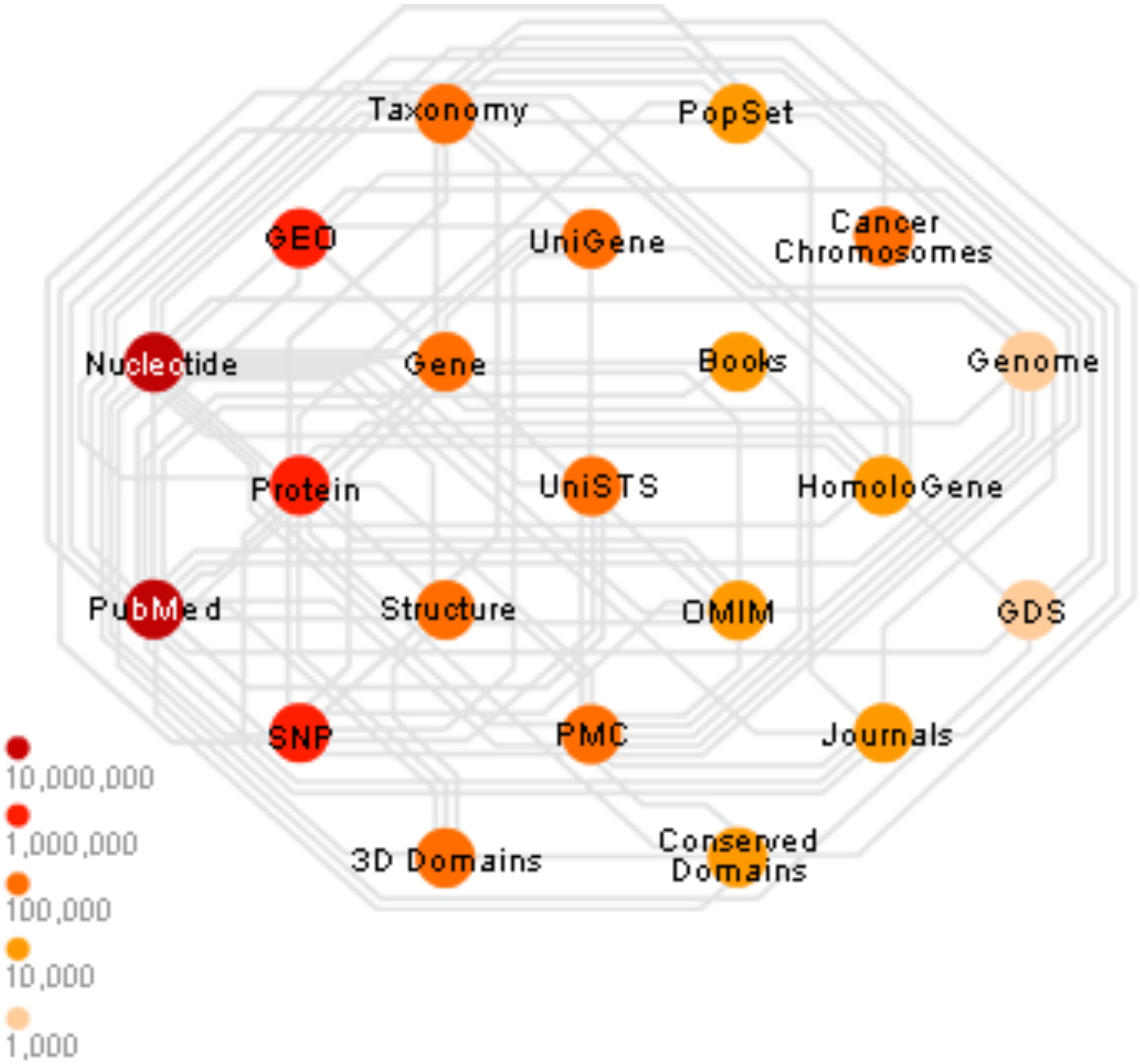
JUNE 1989
$2.95

M.I.T.'s ℞ for building a new industrial America.

Strange fossil creatures from an ancient sea.

How a crystal lattice channels electrons and positrons.

An earthquake waiting to happen? This sharply folded terrain in the foothills of the Andes could conceal a dangerous fault.



More than just sequence data

Genome assemblies

The genome sequence produced after chromosomes have been fragmented, sequenced and the resulting sequences have been put back together.

The **reference assembly** for a genome can be compiled from the DNA of one individual, a collection of individuals, a breed or a strain.

Genome Reference Consortium

- "…working to create assemblies that better represent diversity and provide more robust substrates for genome analysis."
    - novel assembly algorithm
    - correcting assembly errors (fix patches)
    - addition of new alternate loci (patches)
    - filling in gaps

EMBL-EBI

NCBI

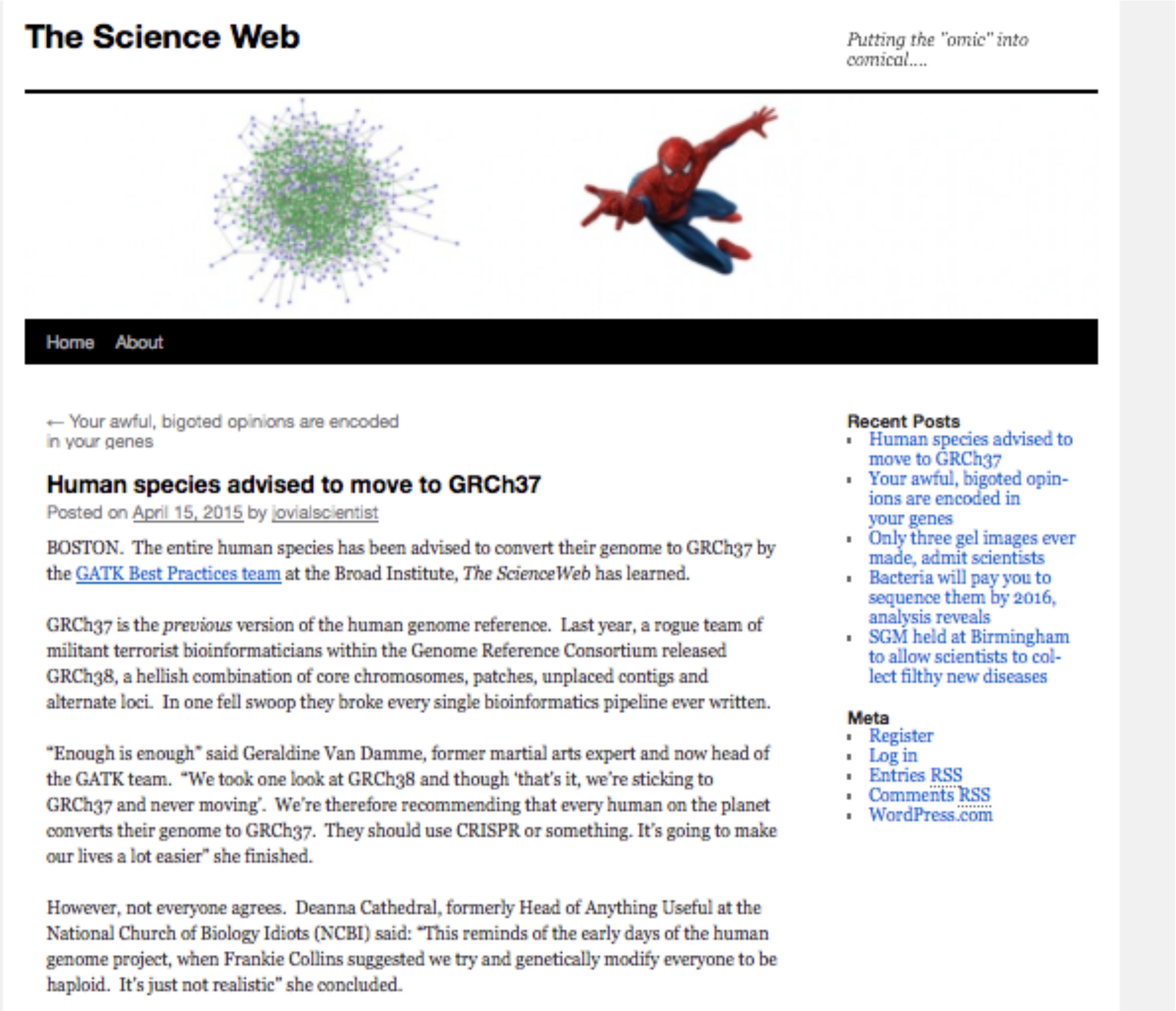
wellcome trust sanger institute

THE GENOME INSTITUTE at Washington University

GRCh37 or hg19?

GRCh37 or hg19?

- Ensembl/NCBI versus UCSC
- chromosomal coordinates are the same
- contig sequences are the same, but different naming convention (i.e. ‘chr1’ versus ‘1’)
- one-based coordinate system versus a zero-based coordinate system



The Science Web

Putting the "omic" into comical....

← Your awful, bigoted opinions are encoded in your genes

Human species advised to move to GRCh37

Posted on April 15, 2015 by jovialscientist

BOSTON. The entire human species has been advised to convert their genome to GRCh37 by the GATK Best Practices team at the Broad Institute, *The ScienceWeb* has learned.

GRCh37 is the *previous* version of the human genome reference. Last year, a rogue team of militant terrorist bioinformaticians within the Genome Reference Consortium released GRCh38, a hellish combination of core chromosomes, patches, unplaced contigs and alternate loci. In one fell swoop they broke every single bioinformatics pipeline ever written.

"Enough is enough" said Geraldine Van Damme, former martial arts expert and now head of the GATK team. "We took one look at GRCh38 and though 'that's it, we're sticking to GRCh37 and never moving'. We're therefore recommending that every human on the planet converts their genome to GRCh37. They should use CRISPR or something. It's going to make our lives a lot easier" she finished.

However, not everyone agrees. Deanna Cathedral, formerly Head of Anything Useful at the National Church of Biology Idiots (NCBI) said: "This reminds of the early days of the human genome project, when Frankie Collins suggested we try and genetically modify everyone to be haploid. It's just not realistic" she concluded.

Recent Posts

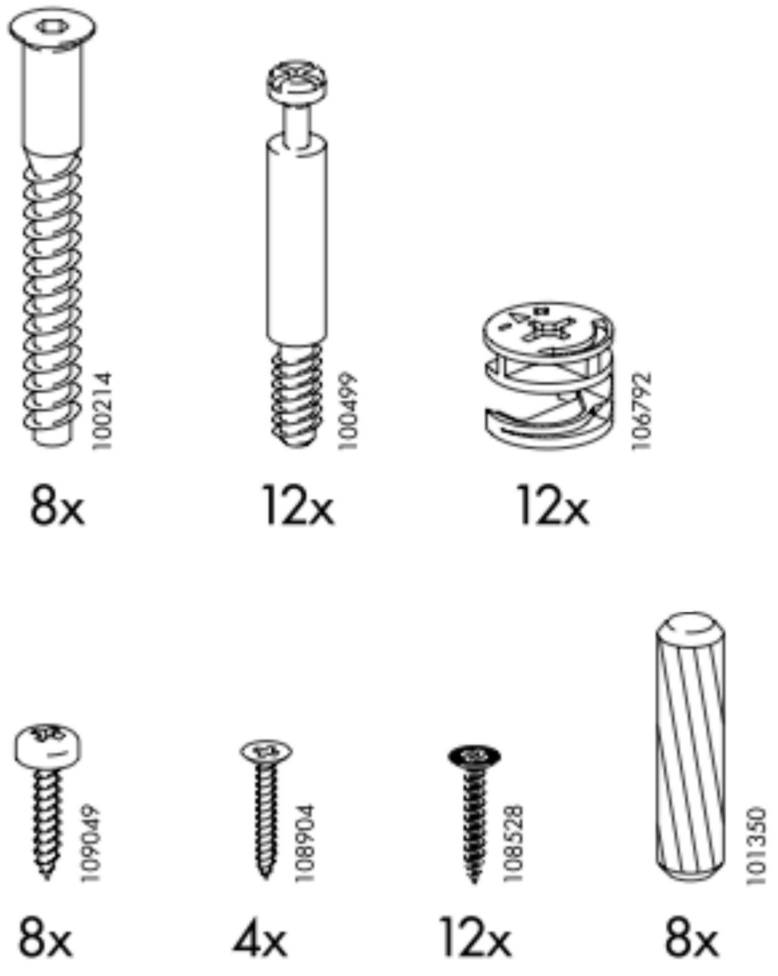
- Human species advised to move to GRCh37
- Your awful, bigoted opinions are encoded in your genes
- Only three gel images ever made, admit scientists
- Bacteria will pay you to sequence them by 2016, analysis reveals
- SGM held at Birmingham to allow scientists to collect filthy new diseases

Meta

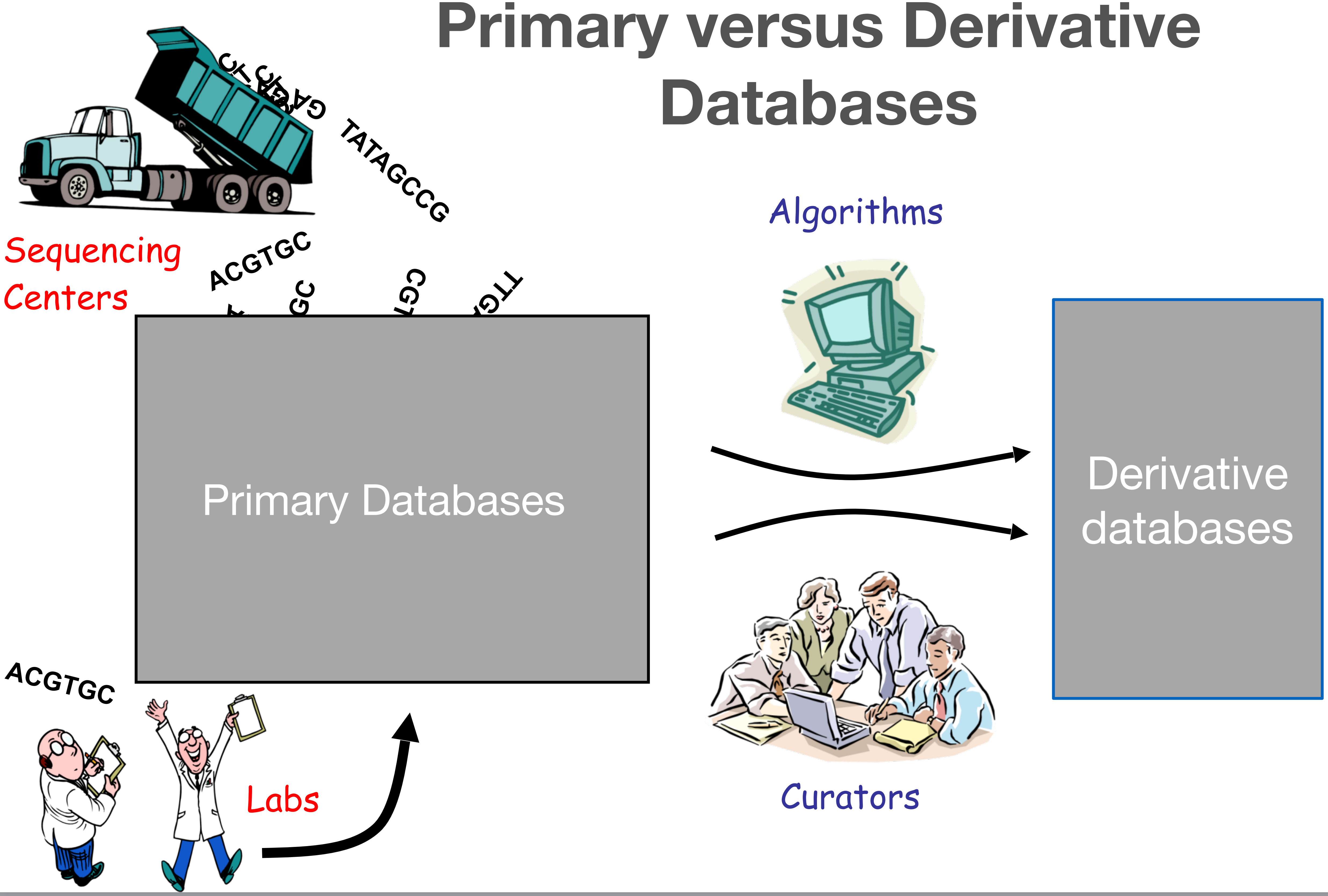
- Register
- Log in
- Entries RSS
- Comments RSS
- WordPress.com

Gene Builds
(not to be confused with *genome* builds)

- A set of annotations for the assembled genome
- Database specific
- Predicted genes based on varying levels of evidence



Primary versus Derivative Databases



Sequencing Centers

Labs

Primary Databases

Algorithms

Curators

Derivative databases

ID conversions: Biomart

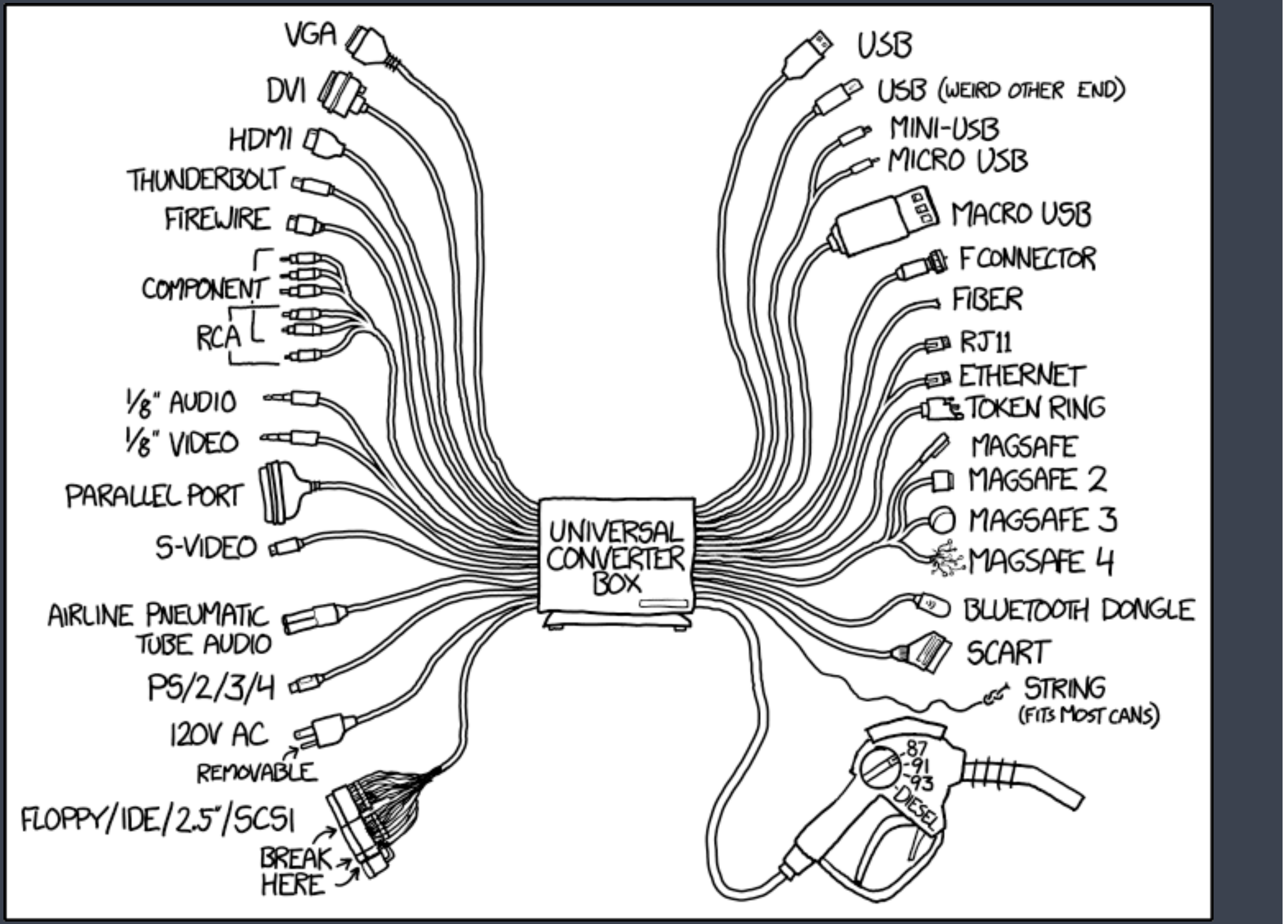
- a search engine (part of Ensembl) that can map terms across multiple domains and output them into table format

- No programming required

    1. Choose database

    2. Apply filters (which genes do we want to look at)

    3. Choose attributes (determine output columns)



https://xkcd.com/1406/

Genome Browsers

- a graphical interface for display of information from a biological database for genomic data
  - General use (UCSC, IGV)
  - Human-specific; Other vertebrate genomes (Ensembl, Entrez)
  - Non-vertebrate genome (FlyBase, WormBase, TAIR)