

Bioinformatics – what is it and how can I access it?

Bioinformatics - definition

- ‘creation of algorithms, **computational and statistical techniques** and theory to solve formal and practical problems inspired from the management and analysis of biological data’ (Wikipedia)
- ‘ The sum of the **computational approaches to analyze, manage, and store biological data**. Bioinformatics involves the analysis of biological information using computers and statistical techniques, the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research.
- Bioinformatics is used in analyzing **genomes, proteomes (protein sequences)**, three-dimensional modelling of biomolecules and biologic systems, etc.
- Training in **informatics** requires backgrounds in molecular biology and computer science, including database design and analytical approaches.’ (www.medterms.com)

Bioinformatics – applications

- **Sequence analysis** – compare genes between species; identify proteins from cDNA or mRNA sequence
- **Comparative genomics**
- **Genome annotation** – biological characteristics of gene/protein sequences
- **Evolutionary biology** - measuring biodiversity (www.ubio.org)
- Analysis of **gene expression** – eg compare up-regulation and down-**regulation** of genes under disease vs healthy conditions
- **Mutations** (cancer, genetic disease)
- **Protein expression**
- Prediction of **protein structure and function**
- **Systems biology** (networks of enzymes and metabolism)
- Biomedical **imaging** and high-throughput image analysis

Why bioinformatics for epidemiologists?

- A paradigm shift
- Huge new technology may be analogous to 'DNA revolution' originating in 1953 (Watson & Crick discovery of the double helix)
- Genetic data increasingly an explanatory variable for disease / production characteristics
- Genetic variables an inextricable element of host/agent part of 'host/agent/environment' trio

Why bioinformatics for epidemiologists?

- ‘genetic epidemiology’ is itself a new sub-specialty field
- ‘epidemiological evaluation of the role of inherited causes of disease in families and in populations; it aims to detect the inheritance pattern of a particular disease, localize the gene and find a marker associated with disease susceptibility’ (M.Tevfik Dorak www.dorak.info)
- Genetic disorders may be **monogenic** – a single gene associated with disease
- Within monogenic diseases may have a **single mutation** causing disease, or **many different mutations** within a single gene causing varying disease phenotype
- Genetic disorders may be **polygenic** – multiple genes associated with disease or conferring susceptibility for disease given other triggering factors
- Important for epidemiologists to have some understanding of where the genomic/proteomic data comes from

Aim

- Introduce 'look' of web-based databases
- Give a few examples of basic functions and tools
- Bioinformatics databases cover many species – appreciation of vast quantity of data
- Give list of resources

A few definitions

- **Genome** – total genetic content of an organism
- **Proteome** – entire content of protein expressed by a genome
- **Genomic DNA** – DNA sequence with all introns, exons, coding and non-coding regions
- **mRNA** – messenger RNA which contains just exons (introns and non-coding regions are spliced out)
- **cDNA** – cloned DNA generated from mRNA transcripts – no non-coding DNA
- **Exon** – protein coding part of gene
- **Intron** – non-coding part of gene in-between exons

A few definitions

- DNAs based on nucleotide sequences AGCT
- RNAs based on nucleotide sequences AGCU
- Amino acid sequences represent peptides/proteins

- *.fasta file – standard text-based format for nucleic acid or protein (amino acid) sequence – a header indicated by ‘<’ followed by the nucleic acid or amino acid sequence

- *.GFF file – General Feature Format file – a text data file containing fields which describe attributes for a sequence – see sanger institute sanger.ac.uk. <seqname><source><feature><start><end><score><strand><frame>
- i.e. GFF file gives some information about the sequence – where it comes from, numbers of start and stop codons and splice sites, but doesn’t contain the sequence itself

- GFF file can be used, eg to process a DNA sequence through to mRNA to protein sequence using PERL software

Practical example – generalised glycogenosis (glycogen storage disease, Pompe's disease) in Brahman and shorthorn cows

- Monogenic – 5 mutations known with differing phenotype
- Acidic α -glucosidase (GAA, AAG - lysosomal enzyme) catabolises glycogen to glucose; gene defects cause glycogen storage disease
- Several different mutations:
 - 1057 \square TA – exon 7 – truncated protein (lethal)
 - 1783 C \blacklozenge T – exon 13 – truncated protein (lethal)
 - 1352T – silent polymorphism
 - 2223A – silent polymorphism
 - 2454 \square CA mutation (shorthorns) – (lethal)
 - + more..

Dennis JA et al Australian Veterinary Journal 2002;80(5):286

Dennis JA et al Mammalian Genome 2000;11:206

Ensembl Genome Browser - UQconnect

File Edit View Favorites Tools Help

Address <http://www.ensembl.org/index.html>

Google

e!Ensembl

Ensembl release 45 - Jun 2007

HOME · BLAST · BIOMART · SITEMAP **HELP**

Your Ensembl

- Login or Register
- About User Accounts

Help & Documentation

- About Ensembl
- Genomic Data
- Help & Information
- Software

Select a species

- Mammals
 - Armadillo
 - Bushbaby
 - Cat
 - Cow**
 - Dog
 - Elephant
 - Guinea Pig
 - Hedgehog
 - Human
 - Lesser hedgehog tenrec

Search Ensembl

Search: for

e.g. mouse chromosome 2 or rat X:10000..20000 or human gene BRCA2

Go

Ensembl tools

- Start a sequence search** →
Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.
- Mine Ensembl with BioMart** →
Extract information from the Ensembl database and export sequences or tables in text, html, or Excel format with BioMart
- Customise Your Ensembl** →
Register with Ensembl to bookmark your favourite pages, customise your home page and much more!
- Fetch data with the Ensembl API** →
Learn how to extract data from the public Ensembl database with this tutorial.

Important Notice




As part of the upgrade to Release 45, we have had to invalidate all login cookies. You will therefore need to log in again. Please accept our apologies for any inconvenience caused.

If you have any problems, please [contact our HelpDesk](#).

Ensembl 45

Pre! species

Popular genomes

-  **Human**
NCBI 38 | Vega
-  **Mouse**
NCBI m38 | Vega
-  **Zebrafish**
Zv6 | Vega

More genomes

- ▶ **Aedes** AaegL1
- ▶ **Anopheles** AgamP3
- ▶ **Armadillo** ARMA
- ▶ **Bushbaby** otoGar1
- ▶ **C.elegans** WS170
- ▶ **C.intestinalis** JGI 2
- ▶ **C.savignyi** CSAV 2.0
- ▶ **Cat** CAT

Internet

Start | Inboxes - Microsoft ... | Inboxes - UQconnect | Bioinformatics - w... | Microsoft Office P... | Ensembl Genom... | 9:54 AM

www.ensembl.org – Cow – Btau 3.1 – based on Hereford

Information about methods

Search Ensembl *Bos taurus*

Search: **Go**

e.g. chromosome 17 or 22:10000..200000 or Q59FM4.

Karyotype


Click on a chromosome for a closer view

1 2 3 4 5 6 7 8 9 10 11
12 13 14 15 16 17 18 19 20 21 22
23 24 25 26 27 28 29 X Y HT

Jump directly to sequence position

About the *Bos taurus* genome

Assembly

 Btau_3.1 is a preliminary assembly of the cow, *Bos taurus*, Hereford breed. The Btau_3.1 assembly used a combined strategy as developed for the rat genome (Nature 428:493-521). The combined strategy is a hybrid of the Whole Genome Shotgun (WGS) approach used for the mouse genome and the hierarchical (BAC clone) approach used for the human genome. The sequencing combines BAC shotgun reads with whole-genome-shotgun (WGS) reads from small insert libraries as well as BAC end sequences. The project coordination and genome sequencing and assembly is provided by the [Human Genome Sequencing Center at Baylor College of Medicine](#)

The N50 size is the length such that 50% of the assembled genome lies in blocks of the N50 size or longer. The N50 size of the contigs is 48.7 kb and the N50 of the scaffolds is 997.5 kb. The total length of all contigs is 2.73 Gb. When the gaps between contigs in scaffolds are included, the total span of the assembly is 2.87 Gb. The coverage is 7x.

Annotation

This is a full genebuild on the assembly 3.1 in the standard ensembl way. It also includes sheep BAC ends mapped to the genome

www.ensembl.org – Cow – Btau 3.1

Address: http://www.ensembl.org/Bos_taurus/index.html

sheep BAC ends mapped to the genome

Jump directly to sequence position

Chromosome: or region

From (bp):

To (bp):

What's New in Ensembl 45

Bos taurus News

There is no *Bos taurus*-specific news this release.

General News

- ▶ **Removal of viral genes**
Viral genes have been removed from all species with Ensembl genebuilds (i.e. [Read more...](#))
- ▶ **Schema changes**
Only minor changes to the database schema this release:
 - ▶ An index has been added to marker.display_marker_synonym_id
 - ▶ The "NOT NULL" constraint on external_db.db_release has been removed

[More news...](#)

Statistics

Assembly:	Btau_3.1, Aug 2006
Genebuild:	Ensembl, Sep 2006
Database version:	45.3b
Known genes:	16,938
Projected genes:	1,653
Novel genes:	3,164
Pseudogenes:	1,264
RNA genes:	1,671
Genscan gene predictions:	59,639
Gene exons:	217,983
Gene transcripts:	28,968
SNPs:	1,792,456
Base Pairs*:	3,247,285,296
Golden Path Length**:	3,033,353,239

* Total number of base pairs = sum of lengths of DNA table
** Reference assembly (Golden path) length = sum of non-redundant top level seq regions

© 2007 WTSI / EMBL. Ensembl is available to download for public use - please see the [code licence](#) for details.

Example – generalised glycogenosis in cattle – alpha glucosidase gene

The screenshot shows the Ensembl genome browser interface for *Bos taurus*. The browser window title is "Cow (Bos taurus) - UQconnect". The address bar shows the URL http://www.ensembl.org/Bos_taurus/index.html. The Ensembl logo and "Cow" are prominently displayed. A search bar is highlighted with a red circle and labeled "Alpha-glucosidase". The search bar contains the text "Search Ensembl *Bos taurus*" and a "Go" button. Below the search bar, there is a text input field with the placeholder "Search:" and a "Go" button. Below the search bar, there is a text input field with the placeholder "e.g. chromosome 17 or 22:10000..200000 or Q59FM4.1".

The page content includes a navigation menu on the left with sections: "Your EnSEMBL" (Login or Register, About User Accounts), "Help & Documentation" (About Ensembl, Genomic Data, Help & Information, Software), and "Select a species" (Mammals: Armadillo, Bushbaby, Cat, Chimpanzee, Cow, Dog, Elephant, Guinea Pig, Hedgehog, Human, Lesser hedgehog tenrec). The main content area is titled "Explore the *Bos taurus* genome" and features a "Karyotype" section with a diagram of chromosomes and a link to "Jump directly to sequence position". The "About the *Bos taurus* genome" section includes an "Assembly" subsection with a description of the Btau_3.1 assembly and an "Annotation" subsection with a description of the full genebuild.

Bovine alpha-glucosidase gene

Ensembl release 45: Bos taurus Gene report for ENSBTAG00000016021 - UQconnect

Address: http://www.ensembl.org/Bos_taurus/geneview?gene=ENSBTAG00000016021

Ensembl Gene Report for ENSBTAG00000016021

Gene	LYAG_BOVIN (UniProtKB/Swiss-Prot) To view all Ensembl genes linked to the name click here .												
Ensembl Gene ID	ENSBTAG00000016021												
Genomic Location	This gene can be found on Chromosome 19 at location 51,638,387-51,651,659 . The start of this gene is located in Contig AAFC03083061 .												
Description	Lysosomal alpha-glucosidase precursor (EC 3.2.1.20) (Acid maltase). Source: Uniprot/SWISSPROT Q9MYM4												
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned cDNAs followed by an ORF prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate UTRs (For more information see V. Curwen et al., Genome Res. 2004 14:942-50)												
Transcripts	<table border="0"><tr><td>ENSBTAT00000021325</td><td>ENSBTAP00000021325</td><td>LYAG_BOVIN</td><td>[Transcript info]</td><td>[Exon info]</td><td>[Peptide info]</td></tr><tr><td>ENSBTAT00000042592</td><td>ENSBTAP00000040228</td><td>novel transcript</td><td>[Transcript info]</td><td>[Exon info]</td><td>[Peptide info]</td></tr></table>	ENSBTAT00000021325	ENSBTAP00000021325	LYAG_BOVIN	[Transcript info]	[Exon info]	[Peptide info]	ENSBTAT00000042592	ENSBTAP00000040228	novel transcript	[Transcript info]	[Exon info]	[Peptide info]
ENSBTAT00000021325	ENSBTAP00000021325	LYAG_BOVIN	[Transcript info]	[Exon info]	[Peptide info]								
ENSBTAT00000042592	ENSBTAP00000040228	novel transcript	[Transcript info]	[Exon info]	[Peptide info]								
Alignments	This gene can be viewed in genomic alignment with other species view genomic alignment with 7 eutherian mammals Pecan view genomic alignment with 10 amniota vertebrates Pecan view genomic alignment with Homo sapiens												

Representation of gene

**Exon info
Protein info**

Chr. 19
Length 33.27 Kb
Forward strand
Ensembl trans. DDX48_BOVIN > USMG5_BOVIN >
Ensembl Known Protein Coding
DNA(contigs) AAFC03015967 > < AAFC03083061
Ensembl trans. < ENSBTAT00000042592 > < ENSBTAT00000042592 >
Ensembl Novel Protein Coding
< LYAG_BOVIN >
Ensembl Known Protein Coding
Length 33.27 Kb
Reverse strand

Bovine alpha-glucosidase gene

Ensembl release 45: Bos taurus Transcript Report for ENSBTAT00000021325 - UQconnect

Address: http://www.ensembl.org/Bos_taurus/transview?db=core&transcript=ENSBTAT00000021325&show=plain&number=off&submit=Refresh

Your EnsEMBL

- Login or Register
- About User Accounts

ENSBTAT00000021325

- Gene information
- Gene splice site image
- Genomic sequence
- Gene variation info.
- ID history
- Transcript information
- Exon information
- Protein information
- Export transcript data

Chromosome 19

51,638,387 - 51,651,659

- View of Chromosome 19
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region

Ensembl Archive

Ensembl Transcript Report

Transcript	LYAG_BOVIN (UniProtKB/Swiss-Prot) To view all Ensembl genes linked to the name click here .
Ensembl Transcript ID	ENSBTAT00000021325
Transcript information	Exons: 19 Transcript length: 3,787 bps Translation length: 937 residues This transcript is a product of gene: ENSBTAG00000016021
Genomic Location	This transcript can be found on Chromosome 19 at location 51,638,387-51,651,659 . The start of this transcript is located in Contig AAF03083061 .
Description	Lysosomal alpha-glucosidase precursor (EC 3.2.1.20) (Acid maltase). Source: UniProt/SWISSPROT Q9MYM4
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned cDNAs followed by an ORF prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate UTRs (For more information see V.Curwen et al., Genome Res. 2004 14:942-50)
Similarity Matches	This Ensembl entry corresponds to the following database identifiers: UniProtKB/Swiss-Prot: LYAG_BOVIN [Target %id: 100; Query %id: 100] [align] RefSeq peptide: NP_776338.1 [Target %id: 100; Query %id: 100] [align] RefSeq DNA: NM_173913.1 [Target %id: 100; Query %id: 98] [align] EntrezGene: GAA EMBL: AF171665 [align] AF171666 [align] Protein ID: AAF81636.1 [align] AAF81637.1 [align] UniGene: Bt.52221 [Target %id: 100; Query %id: 98] Bt.65694 [Target %id: 12; Query %id: 90] Bt.66219 [Target %id: 20; Query %id: 91]
Oligo Matches	This Ensembl entry corresponds to the following database identifiers: Affymx Microarray Bovine: Bt.4929.1.S2_at
GO	The following GO terms have been mapped to this entry via UniProt and/or RefSeq: GO:0004553 [hydrolase activity, hydrolyzing O-glycosyl compounds] IEA GO:0004558 [alpha-glucosidase activity] IEA GO:0005764 [lysosome] IFA

Aliases for other databases

Bovine alpha-glucosidase gene

Ensembl release 45: Bos taurus Transcript Report for ENSBTAT00000021325 - UQconnect

File Edit View Favorites Tools Help

Address: http://www.ensembl.org/Bos_taurus/transview?db=core&transcript=ENSBTAT00000021325&show=plain&number=off&submit=Refresh

View of Chromosome 19
Graphical view
Graphical overview
Export information about region
Export sequence as FASTA
Export EMBL file
Export Gene info in region
Export SNP info in region

Ensembl Archive

- View previous release of page in Archive!
- e! 44: Apr 2007 (Btau_3.1)
- e! 43: Feb 2007 (Btau_3.1)
- e! 42: Dec 2006 (Btau 2.0)
- e! 41: Oct 2006 (Btau 2.0)
- e! 40: Aug 2006 (Btau 2.0)
- e! 39: Jun 2006 (Btau 2.0)
- e! 38: Apr 2006 (Btau 2.0)
- e! 37: Feb 2006 (Btau 2.0)
- e! 36: Dec 2005 (Btau 2.0)
- e! 35: Nov 2005 (Btau 1.0)
- e! 34: Oct 2005 (Btau 1.0)
- e! 33: Sep 2005 (Btau 1.0)
- e! 32: Jul 2005 (Btau 1.0)

Stable Archive! link for this page

RefSeq DNA: [NM_173913.1](#) [Target %id: 100; Query %id: 98] [align]
EntrezGene: [GAA](#)
EMBL: [AF171665](#) [align] [AF171666](#) [align]
Protein ID: [AAF81636.1](#) [align] [AAF81637.1](#) [align]
UniGene: [Bt.52221](#) [Target %id: 100; Query %id: 98]
[Bt.65694](#) [Target %id: 12; Query %id: 90]
[Bt.66219](#) [Target %id: 20; Query %id: 91]

Oligo Matches

This Ensembl entry corresponds to the following database identifiers:
Affymx Microarray Bovine: [Bt.4929_1.S2_at](#)

GO

The following GO terms have been mapped to this entry via UniProt and/or RefSeq:

- [GO:0004553](#) [hydrolase activity, hydrolyzing O-glycosyl compounds] IEA
- [GO:0004558](#) [alpha-glucosidase activity] IEA
- [GO:0005764](#) [lysosome] IEA
- [GO:0005975](#) [carbohydrate metabolic process] IEA
- [GO:0005980](#) [from Mus musculus ENSMUSP00000026666] [glycogen catabolic process] IEA
- [GO:0008152](#) [metabolic process] IEA
- [GO:0016787](#) [hydrolase activity] IEA
- [GO:0016798](#) [hydrolase activity, acting on glycosyl bonds] IEA


InterPro

[IPR000322](#) Glycoside hydrolase, family 31 - [View other genes with this domain]
[IPR000519](#) P-type trefoil - [View other genes with this domain]

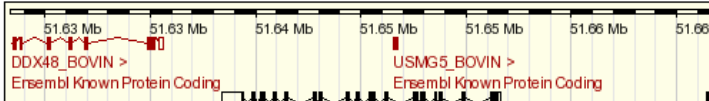
Protein Family

[ENSF00000000727](#): ALPHA GLUCOSIDASE PRECURSOR EC_3.2.1.20 MALTASE
This cluster contains 5 Ensembl gene member(s) in this species.

Transcript structure



Transcript neighbourhood



GO terms – gene ontology database – describes gene and gene product (protein) attributes

Internet

Start | Inbox - Microsof... | Inbox - UQconnect | Bioinformatics - ... | 2 Internet Ex... | bioinformatics_p... | 11:30 AM

Bovine alpha-glucosidase gene

Ensembl release 45: Bos taurus Transcript Report for ENSBTAT00000021325 - UQconnect


File Edit View Favorites Tools Help

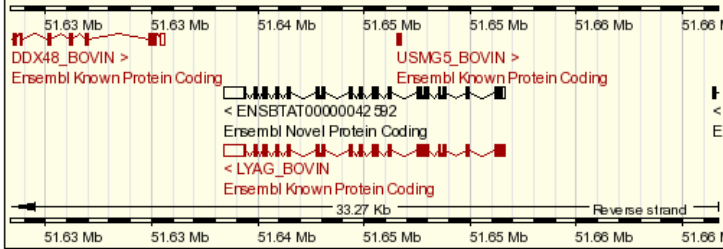
Address http://www.ensembl.org/Bos_taurus/transview?db=core&transcript=ENSBTAT00000021325&show=plain&number=off&submit=Refresh

Google

e! 38: Apr 2006 (Btau 2.0)
e! 37: Feb 2006 (Btau 2.0)
e! 36: Dec 2005 (Btau 2.0)
e! 35: Nov 2005 (Btau 1.0)
e! 34: Oct 2005 (Btau 1.0)
e! 33: Sep 2005 (Btau 1.0)
e! 32: Jul 2005 (Btau 1.0)
Stable Archive! link for this page

Protein Family
ENSF00000000727: ALPHA GLUCOSIDASE PRECURSOR EC_3.2.1.20 MALTASE
This cluster contains 5 Ensembl gene member(s) in this species.

Transcript structure


Transcript neighbourhood


Transcript sequence
ATGATGAGGTGGCCACCATGCTCCCGCCCCTGCTGGGGGTCTGCACCCTCCTCTCCTTG
GCGCTCCTGGGGACACATCCTGCTTCAIGACTTGGAGGTGGTCCCCCGAGAACTACGAGGC
TTCTCCCAAGACGAGATTACCAAGCTTGCCAGCCAGGAGCCAGCCAGCCAGAATGCCGT
GGCAGCCCCAGGGCAGCACCTACACAGTGCAGCTGCCCCCAACAGCCGCTTTGACTGC
GCCCCAGACAAGGGCATCACCCCGCAGCAGTGTGAGGCCCGTGGCTGCTACATGCGCT
GCAGAGTGGCCTCCGGATGCCAGATGGGGCAGCCCTGGTGCTTCTTCCCTCCAGCTAC
CCCAGTTACAGGCTGGAGAACCTGACCACCCTGAGACAGGGTACACAGCCACCCTGACC
CGTGCCGTCCCGACCTTCTTCCCAAGGACATCATGACCTTGAGGCTGGACATGTTGATG
GAGACCGAGAGCCGACTCCACTTCACGATCAAGATCCTGCCAACAGCCTATGAAGTG
CCCTTGGAGACCCCGGTGTCTATAGCCAGCGCCGTTACACTCTACAGCGTGGAGTTC
TCGGAGGAGCCCTTGGGGTGGTCTGCGGGGGAAGCTGGACGGACGGGTGCTGCTGAAC
ACCAGGTGGCCCCCTGTTCTTTGGGGACAGTTCCTGAGCTGTCCACTCCCTGCCA
TCCCAGCACATCACAGGCTTGGCGAACACCTTGGGTCCCTGATGCTCAGCACCAACTGG
ACCAAGATCACCTCTGGAAACGGGACATCGCCCTGAGCCCAACGTTAACTGTATGGA
TCTCACCTTTCTACCTGGTCTGGAGGATGGCGGGTTGGCTCACGGGGTCTTCTGCTG
AACAGCAATGCCATGGATGGTCTGCAGCCAGCCAGCCCTCAGCTGGAGGTCGACA
GGCGGGATCCTGGACGTGTACATCTTCTGGGCCGGAGCCCAAGAGCGTGGTGCAGCAG
TACCTGGACGTGTTGGCTACCCGTTTCATGCCCGTATGGGGCTGGGCTTCCACCTG
TGCCGCTGGGGTACTCCACCTCTGCCATCACCCGCCAGGTCGTGGAGAATATGACCAGG
GCTTACTGCCCTGACGTCAGTGGATGGACCTGACTACATGGATGCCAGGCGGAC

Transcript –
can be
exported – as
.fasta file is
good

Bovine alpha-glucosidase gene

Ensembl release 45: Bos taurus Transcript Report for ENSBTAT00000021325 - UQconnect

Address: http://www.ensembl.org/Bos_taurus/transview?db=core&transcript=ENSBTAT00000021325&show=plain&number=off&submit=Refresh

```
GCCATGAGGAAGGCCCTCACCCCTGCGCTACGTGCTACTGCCCCTAICTCTACACACTGTTCCACAGGGCCACGTCAGAGGCGAGACAGTGGCCCGCCCTCTTCTCTGGAGTCCCCGAGGACCCAGCACCTGGACTGTGGACCCGACGCTCCTGTGGGGGAGGCTCTGCTCATCACC  
CCGGTGTCTGAGGCTGAGAAGGTTGAAGTCACTGGTACTTCCCCAGGGCAGCTGGTACGACCTGACAGCGGTACCAATGGAGGCCCTTGGCAGCCTCCCACCTCCTGCACCCCTCAGCTGTCTATCCACAGCAAGGGGCGAGTGGGTGACGCTGTCCGCCCCGCTGGACACCATCAAC  
GTCCACCTCCGGGCGGGCACAATCATCCCTATGCAGGGCCCTGCCCTCAGACCACAGAGTCCCAGCAAGCAGCATGGCCCTGGCTGTGGCCCTGACAGCCAGTGGGGAGGCCCAAGGGGAGCTGTCTGGGACGACGGGAGAGCCTGGGAGTCTGGATGGTGGGGACTACACACAG  
CTCATCTTTCTGGCCAAGAACAACACCTTTGTAACAAGCTGGTGCAGCTGAGCAGTGGGGCCAGCGCTGACCTGCTGACAGCTCCCAAGTGGCCAGCACTTGCACCCCTGTGCGGGTGTAGGGGCTTGGGTGAGAAGTGTCTACCCCTGAGTCACATGGCACTGACGCTCCTC  
GAACACCCAGATCCGCTTGAAGTCTGTCTCCCGGGCTCTGGGCCAGCAGCTCGGCTGAGTCACTGCCAGATGCCAGTCAGATCAATGCCCTGAGCGGTGCCCTGTACTGGAATGTACTGGAAGCTTGTCTCCTGTAGCCTGTCACTGCAATGTGTGCACAGCCAGCCACCTAC  
TGCCCGCTATGACGCAGAGGTGGTGCCTGGACAGGGTGGTACCTGCATTTGCAAGGCCCCACACCTCGGGAGGCCCTGCTGCTGGGACCCGATGACAGCACAGAATGGGGGCTGTGCA  
GACTGCTTTCTGGGCGCTGCCCCCAACTCAGCAGAGTTCATAGGACTCAGGGAACCTC  
AATCTGAAGTGCAGTATTTTCAATAAAGGATGCCCTGAAGGTCTTGTATGACAGGAGGATCCCATCCTTGTATGGGCCCTGGACATTTCTGGGCAAGTGTGCGGCTTCACTGTGTACTCA  
GAGCAGCACCCTCACAGCCCATCGCCTCACACCCAGTATCCGCTCTCAGGTGACTGGAGAACAGCCAGCCAGGCCCTGGTGTGATCACTGGTGGGAGATGCGGCCCCGCCCAACCT  
GCCCAAGGCTCTGTGGACTAAAGTGTACACCCACACCCGTTCCAGACAAACCACTGGTCTGCTTTGGTCACTATAAATCAGTTTACATTTTTTGGAAATTTCTATAAATGGAATCA  
TACAGCTTTTTTATTTGGGGGGGGGCTAATAATTAATCTAACTGCTCTGAGATTCACTGGTTGTATGTCTTTTTCATTGCCACATAGTTCTCTGTGGTATGGAGGTGCCACATGG  
TTTATCA
```

Show the following features: Exons
Number residues: Exons
Exons and Codons
Exons, Codons and Translation
Exons, Codons, Translations and SNPs
Exons, Codons, Translation, SNPs and Coding sequence

Options for viewing and export as text file

© 2007 [WTSI](#) / [EBI](#). Ensembl is available to [download for public use](#) - please see the [code licence](#) for details.

NCBI database - www.ncbi.nlm.nih.gov/sites/entrez?Db=gene

Entrez Gene: GAA glucosidase, alpha; acid [*Bos taurus*] - UQconnect

Address: http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=280798&ordinalpos=2&tool=EntrezSystem2.PEntrez.Gene.Gene_ResultsPanel.Gene_RVDocSum

NCBI Entrez Gene

Search: Gene for [] Go Clear

Display: Full Report Show 20 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

1: GAA glucosidase, alpha; acid [*Bos taurus*]
GeneID: 280798 updated 14-Jun-2007

Summary

Gene description	glucosidase, alpha; acid
Gene type	protein coding
RefSeq status	Provisional
Organism	Bos taurus
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos
Also known as	GAA

Genomic regions, transcripts, and products

(minus strand) Go to [reference sequence details](#)

NC_007317.2

51651659 5' 51638388 3'

NL_173913.1 NP_776338.1

Entrez Gene Home

Table Of Contents

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links

Links Explain

- Conserved Domains
- Genome
- Map Viewer
- Nucleotide
- Protein
- PubMed
- SNP
- SNP: GeneView
- Taxonomy
- ArkDB

Go to another database – similar but not identical information

Tools examples: BLAST, BLAT, BioMart

Online software tools for genomics analysis

- BLAST – **B**asic **L**ocal **A**lignment **S**earch **T**ool (available on ensembl, NCBI)
- BLAT – **B**last-**L**ike **A**lignment **T**ool - sequence identification tool from USCS (UCSC, University of California Santa Cruz Genome Project)
- BLAST and BLAT - **compare sequences of interest** with previously characterised genes
- **BioMart** – a query-oriented data management system developed by European bioinformatics Institute (EBI) and Cold Spring Harbour Laboratory (CSHL)

BLAST screen

Ensembl COW BlastView

Search: e! Cow EBI Sanger - **Go**

e.g. [AAFC03063255](#), [ENSBTAT0000028690](#)

Ensembl release 45 - Jun 2007 [HOME](#) · [BLAST](#) · [BIOMART](#) · [SITMAP](#) **HELP**

Your EnsEMBL

- Login or Register
- About User Accounts

Pufferfish
Takifugu rubripes

Now in Ensembl **Pre!**

new **SETUP** CONFIG RESULTS DISPLAY refresh **Online Help**

Summary

- ▶ **setup** Not yet initialised
- ▶ **configure** Not yet initialised
- ▶ **results** Not yet initialised
- ▶ **display** Not yet initialised

Enter the Query Sequence

Either Paste sequences (max 30 sequences) in FASTA or plain text:

Or Upload a file containing one or more FASTA sequences:
 Brows...

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq):
 Retrieve

Or Enter an existing ticket ID:
 Retrieve

dna queries
 peptide queries

Select the databases to search against

Select species:
Use 'ctrl' key to select multiple species

dna database
 peptide database

Done

Start Entrez Gene: GAA glucos... **BlastView - UQconnect** Bioinformatics - what is i... Internet 9:25 AM

Results from sequence query using BLAST

Address: http://www.ensembl.org/Bos_taurus/blastview/BLA_9bOURcYAM

Alignment Summary (click arrow to hide)

Select rows to include in table, and type of sort
(Use the 'ctrl' key to select multiples)

refresh

Links	Query Start	Query End	Ori	Chromosome Name	Start	End	Ori	Stats Score	E-val	%ID	Length
[A] [S] [G] [C]	4821	8070	+	Chr:19	51643590	51646839	-	3250	0.	100.00	3250
[A] [S] [G] [C]	10535	13035	+	Chr:19	51638625	51641125	-	2501	0.	100.00	2501
[A] [S] [G] [C]	2137	3936	+	Chr:19	51647724	51649523	-	1800	0.	100.00	1800
[A] [S] [G] [C]	1	894	+	Chr:19	51650766	51651659	-	894	0.	100.00	894
[A] [S] [G] [C]	3971	4806	+	Chr:19	51646854	51647689	-	836	0.	100.00	836
[A] [S] [G] [C]	9567	10401	+	Chr:19	51641259	51642093	-	835	0.	100.00	835
[A] [S] [G] [C]	8243	9056	+	Chr:19	51642604	51643417	-	814	0.	100.00	814
[A] [S] [G] [C]	9100	9553	+	Chr:19	51642107	51642560	-	454	0.	100.00	454
[A] [S] [G] [C]	1619	2069	+	Chr:19	51649591	51650041	-	451	0.	100.00	451
[A] [S] [G] [C]	4821	5130	-	Un	233326144	233326453	+	294	1.4e-190	98.71	310
[A] [S] [G] [C]	4921	5130	-	Un	243045521	243045730	+	194	3.3e-147	98.10	210
[A] [S] [G] [C]	4821	5112	-	Un	304572697	304572987	-	181	9.7e-166	90.44	293
[A] [S] [G] [C]	583	774	-	Chr:13	22629369	22629560	+	128	4.2e-145	91.33	196
[A] [S] [G] [C]	2321	2539	-	Chr:2	23357364	23357583	+	126	9.0e-137	89.19	222
[A] [S] [G] [C]	2324	2520	+	Chr:20	59289170	59289367	-	124	7.8e-142	90.50	200
[A] [S] [G] [C]	2321	2538	-	Chr:14	23214179	23214398	-	121	5.0e-135	88.69	221
[A] [S] [G] [C]	2319	2497	-	Chr:23	47905302	47905478	+	120	6.2e-115	91.67	180
[A] [S] [G] [C]	2320	2540	-	Chr:10	78615750	78615969	+	119	2.3e-103	88.34	223
[A] [S] [G] [C]	9666	9827	+	Chr:13	21262970	21263131	+	118	7.8e-141	93.21	162
[A] [S] [G] [C]	2322	2494	+	Chr:8	29436107	29436278	-	118	7.9e-107	91.95	174
[A] [S] [G] [C]	2321	2493	+	Chr:17	50436104	50436275	+	117	1.6e-147	91.91	173
[A] [S] [G] [C]	2322	2493	+	Chr:23	47959529	47959699	+	117	4.8e-139	91.91	173
[A] [S] [G] [C]	9666	9829	-	Chr:1	9024174	9024337	+	116	1.7e-145	92.68	164
[A] [S] [G] [C]	2322	2493	-	Chr:16	52683822	52683992	-	116	2.3e-141	91.86	172
[A] [S] [G] [C]	2323	2518	-	Chr:14	15011412	15011608	+	116	1.0e-140	89.50	200
[A] [S] [G] [C]	9666	9825	+	Chr:28	8165340	8165499	+	116	2.0e-138	93.12	160
[A] [S] [G] [C]	2321	2512	+	Chr:20	61596667	61596860	+	115	1.8e-139	89.74	195
[A] [S] [G] [C]	2318	2497	+	Un	126203764	126203943	+	114	2.7e-142	90.66	182
[A] [S] [G] [C]	2322	2520	+	Chr:1	47473445	47473343	-	114	4.5e-108	89.44	202

Result - 100% homology with sequence in cow chromosome 19!

Start | Entrez Gene: GAA glucos... | BlastView - UQconnect | Bioinformatics - what is ... | 9:17 AM

Example of biomart query – use filters to narrow query to ‘all genes expressed in pancreatic ductal cells’

BioMart Project

BioMart is a query-oriented data management system. The system can be used with any database according to requirements. The system supports 'mining' like searches of complex data.

BioMart has built-in support for querying multiple datasets distributed programmatically using web services.

BioMart is completely Open Source.

Pancreatic Expression Database

Dataset
PANCREATIC EXPRESSION (Build 36)

Filters
[None selected]

Attributes
Gene Symbol
Ensembl Gene ID
Differential Expression Analysis
Direction of Regulation

Enter query using filter – ‘all genes expressed by pancreatic ductal cells’

Done
Ensembl
HapMap
Wormbase
Gramene
Drospege
ArrayExpress DW
PRIDE
PepSeeker

Internet

Start | Perl API Installation - UQ... | BioMart - UQconnect | BioMart - MartView - ... | Bioinformatics - what is i... | 9:50 AM

Results from BioMart query 'all genes expressed in normal pancreatic ductal cells'

Pancreatic Expression Database
Barts and The London
Queen Mary's School of Medicine and Dentistry
Institute of Cancer
MolDiag-PaCa

Home | Access the database | Datasets | News | Contact us

XML | Perl | Help

Dataset 1523 / 31206 Entries
PANCREATIC EXPRESSION (Build 36)

Filters
Normal pancreas ND (microdissected normal ductal cells) : Only

Attributes
Gene Symbol
Ensembl Gene ID
Differential Expression Analysis
Direction of Regulation

Export all results to: File | TSV | Unique results only [Go]

Email notification to: [Yellow box]

View: 10 rows as HTML | Unique results only

Gene Symbol	Ensembl Gene ID	Differential Expression Analysis	Direction of Regulation
CENPM_HUMAN	ENSG00000100162	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
CEP55_HUMAN	ENSG00000138180	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
CEP55_HUMAN	ENSG00000138180	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
Q53EZ4-2	ENSG00000138180	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
NR_002734.1	ENSG00000178717	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
ATOX1_HUMAN	ENSG00000177556	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
TRIB2_HUMAN	ENSG00000071575	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
TRIB2_HUMAN	ENSG00000071575	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
Q86VB1_HUMAN	ENSG00000136231	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated
NP_006538.2	ENSG00000136231	Pancreatic adenocarcinoma (PDAC) / Normal pancreas (normal ductal cells) (both microdissected)	Up-regulated

biomart version 0.6

Start | Perl API Installation - UQ... | Ensmart genomics - Goo... | BioMart - MartView - ... | Bioinformatics - what is i... | 9:47 AM

List of genes

Resources

- www.ensembl.org (ensembl)
- www.ncbi.nlm.nih.gov (US national library of medicine)
- www.hgsc.bcm.tmc.edu (Baylor College Texas cow genome project)
- www.animalgenome.org (US National Animal Genome Research Program)
- www.sanger.ac.uk (Sanger Institute)
- www.biomart.org (Biomart)

Summary

- Important for epidemiologists to have overview of bioinformatics/genomics databases and analysis
- Enhance our understanding of pathobiology of disease and agent-host-environment interactions at a level not previously possible