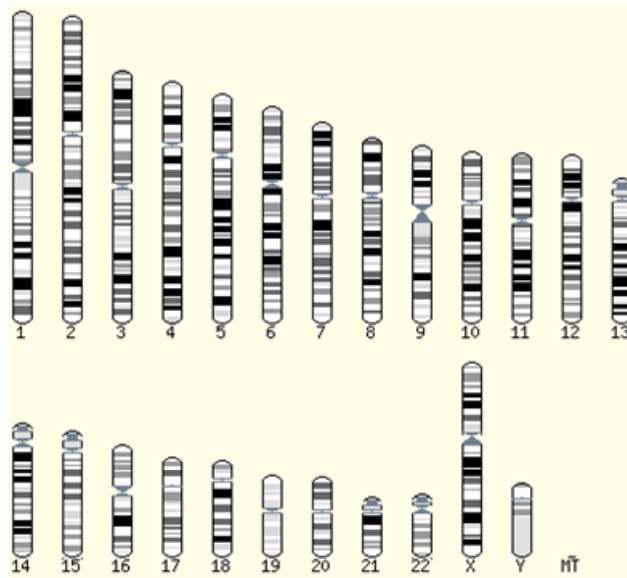


An activity pack for teachers and schools



Surfing Human and Mouse Genomes on the Internet



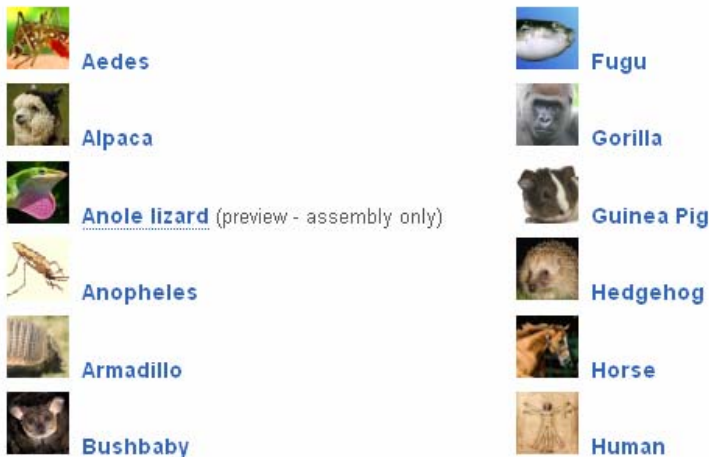
A project of the European Learning Laboratory for the Life Sciences, at the
European Molecular Biology Laboratory (EMBL)

Written by Russ Hodge and Miguel Andrade,
copyright 2002
Adapted to the new interface of Ensembl in 2009

Surfing human and mouse genomes on the Internet

Most people have heard that "solving" the human and mouse genome sequences gives us information useful in understanding disease and human health. How do scientists plan to use this information? This exercise will show you how to find the genomes on the internet and give you some idea of how scientists can use them to discover things about diseases.

You'll need access to the internet to do this. This activity will introduce you to *Ensembl*, which is a user-friendly browser for human and many other genomes accessible to the public on the internet. Below are just some of the species Ensembl has genomic information about!



You'll get a "walking tour" of the site and the information to be found there. Along the way you'll discover how scientists use Ensembl, and how to ask your own questions about the genome.

What is the genome, exactly? Each cell of your body has a nucleus, a compartment that holds DNA. This is the genome, the genetic material that you inherited from your parents. Each of the trillions of cells in your body has an identical copy of your genome. It's a huge amount of information - if you stretched out the DNA in a single cell, it would make a string about two meters long!

DNA is made up of four building blocks called *nucleotides*, represented by the letters *A*, *G*, *T*, and *C*. These letters stand for the bases Adenine (*A*), Guanine (*G*), Thymine (*T*) and Cytosine (*C*), which make up the genetic code.

You can think of these bases as letters that form a word, if the word is a gene. Genes, which are the recipes used to make protein molecules, make up only part of the DNA (about 2.5%). Scientists still don't know much about what the other 97.5% of the genome does. (They sometimes call it "junk DNA," but it's becoming clear that it isn't worthless junk!)

Identical twins started out as a single fertilized egg cell which split in two, so they have the same genome. Everyone else has their own unique set of DNA (unless you have a clone somewhere!). The "human genome" everyone talks about is a standard example, drawn up from samples taken from several different people. If someone were to sequence your own personal genome, they would probably find that a lot of "letters" in the sequence are different.

The genome has already provided us with some fascinating insights into human evolution. A few years ago an EMBL research group used it to show that each of us, on the average, has about a hundred new mutations - errors that didn't come from our parents. We'll pass these altered genes along to our children. Most of these won't matter - they'll probably be in the "junk" somewhere. But mutations in genes have always happened. If they hadn't, the human species wouldn't be here!

The tour:

This will walk you through the genome site at Ensembl; we've included some screen shots that may help you find your way around.

A. Go to the website at <http://www.ensembl.org/>. This is one of the most important sites on the web where you can get direct access to information from genome projects, including the Human Genome Project.

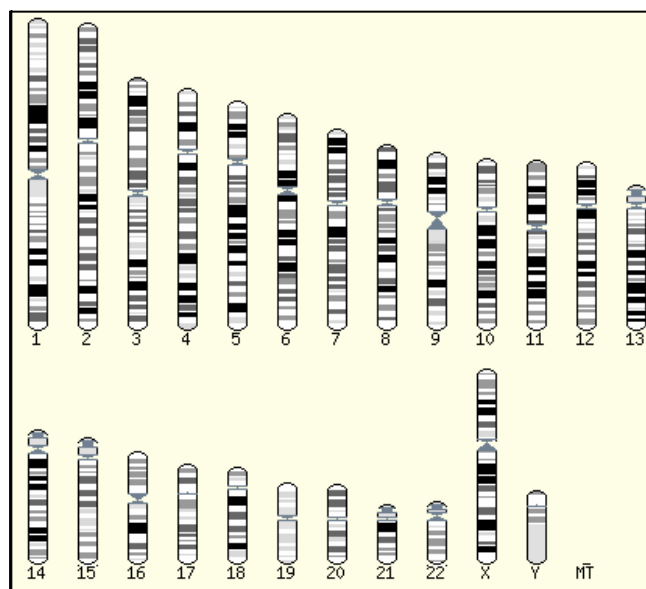
This page is called the "Ensembl Genome Browser."

The screenshot shows the Ensembl Genome Browser homepage. At the top, there is a navigation bar with links for 'Login / Register', 'BLAST/BLAT', 'BioMart', 'Docs & FAQs', and 'img'. Below this is a search bar with the text 'Search: All species for' and a 'Go' button. Below the search bar, there is a section titled 'Browse a Genome' with the text 'The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page.' Below this, there are three popular genomes listed: 'Human' (NCBI36), 'Mouse' (NCBIM37), and 'Zebrafish' (ZFISH). Below these, there is a section titled 'All genomes' with a dropdown menu to 'Select a species'. To the right of the search bar, there is a section titled 'New to Ensembl?' with several links: 'Add custom tracks', 'Upload your own data', 'Search for a DNA or protein sequence', 'Fetch only the data you want', 'Download our databases via FTP', and 'Mine Ensembl with BioMart'. Below this, there is a section titled 'What's New in Release 52 (9 December 2008)' with two bullet points: 'Homo sapiens core database (Human)' and 'Gorilla 2x assembly and genebuild (Gorilla)'.

From here you can examine the complete genome sequence of several animal species (you'll see the list if you click 'Select a Species'). Many more genomes have been sequenced than you see here, but animal genomes are especially important. Their genomes are very similar to ours, so they can be used as simpler models to perform experiments that will help us learn more about human biology and understand diseases.

We'll start with the human genome. The information is constantly being revised; the version number and date show when this happened last. Corrections have to be made in the sequence itself, sometimes - because although sequencing technology is very good, it isn't perfect. Additionally, scientists sometimes must correct how the sequence is interpreted. The human genome project produced tiny puzzle pieces of the genome which had to be assembled in a whole map. Sometimes a piece has been put in the wrong place, or the "picture" it represents (such as whether it contains a gene or not) has to be corrected.

Click on **Human** under the headline **Ensembl Species**. You'll be taken to the address: http://www.ensembl.org/Homo_sapiens/. Click on 'Karyotype' at the left. Now you see a series of bars in the middle of the screen. These are chromosomes - or huge "knots" of DNA. To get two meters of DNA into a cell, it has to be wound up and knotted very tightly!



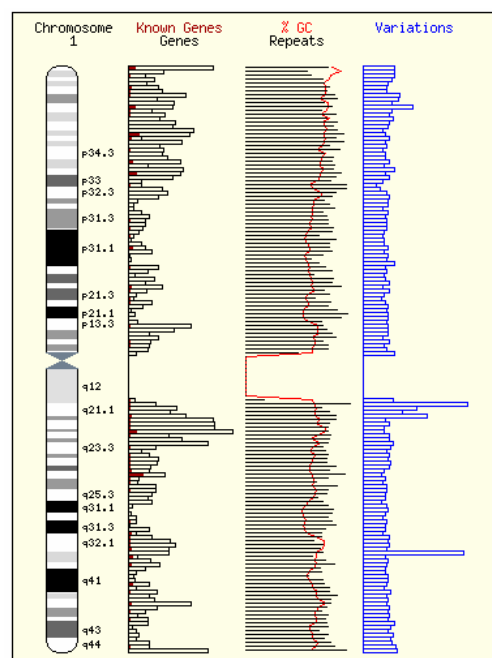
Above, you see a cartoon of 22 numbered human chromosomes and the X and

Y chromosomes. (If you had clicked on the mouse, you'd see that it has two fewer chromosomes!) The thick black stripes drawn on the chromosomes represent regions called "bands". These band regions have different physical properties than other parts, and they appear as a different color when scientists stain the chromosomes to see them better.

You can think of them as "landmarks," like key places of interest on a country's map. That's how scientists have used them in the past - as reference points to talk about different sections of the DNA.

Zoom in on chromosome 1 by moving your mouse over and clicking on it. Follow the link to 'Chromosome summary'. Now you see a magnified picture of the chromosome. The enlargement shows us some more detailed features of the DNA.

The chromosome shows more of the banding pattern, and some of the bands are labelled. These band names are like 'road signs' telling you where on the chromosome you are. You'll see the names scientists have given to some of the bands, starting with "p" for one of the arms of the chromosome, and "q" for the other.



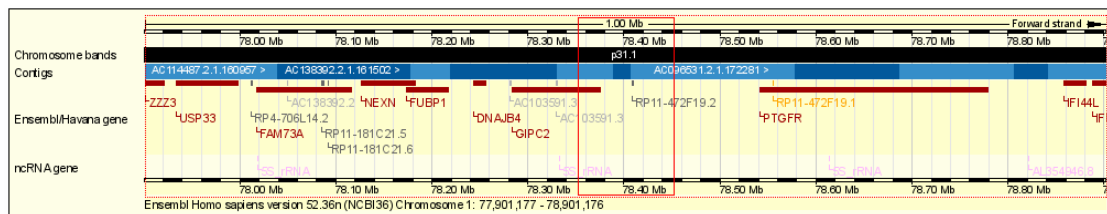
[Click on the image above to zoom into that point](#)

The bars just to the right of the chromosome show the "gene density" - how many genes are found in a particular region (remember that only a small percentage of DNA is actually used as protein-coding genes). Some regions are very dense (many bars); there is also a region in the middle with no genes at all. We also see the percent of G (guanine) and C (cytosine) nucleotides in the DNA sequence. Finally, 'Variations' are alleles that appear in a population for this chromosome.

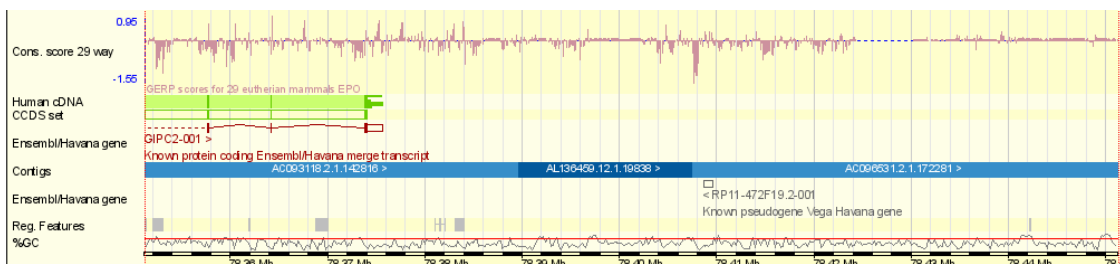
Let's zoom in even farther on the chromosome. Let's zoom in on a place where there are lots of genes - the regions without any aren't nearly as interesting! You can zoom in by clicking on a place with your mouse, or you can type in:

http://www.ensembl.org/Homo_sapiens/Location/View?r=1:78351178-78451177

Let's do that now. Some tracks are displayed. Let's take the panels one at a time. The first panel is shown below.



The blue regions make up the genome sequence of the chromosome. All the genes in this region are shown as filled red rectangles. The empty red rectangle expands to the region below:



This is a smaller region of the chromosome. We are looking at only half of one gene, the *GIPC2* gene. The graph near the top of the picture shows a basepair-per-basepair score for how well each nucleotide is conserved when many species are compared. For example, if a nucleotide 'G' scores high, that means in every species it's a G when you look at that position!

Go back to the human home page:

http://www.ensembl.org/Homo_sapiens/

Search for Alzheimer in the search box at the top of the page.

Ten genes were found. Scroll down the search results to click on **PSEN1**:

Ensembl protein coding Gene: ENSG00000080815 (HGNC (curated): PSEN1) [\[Region in detail\]](#)
Ensembl protein_coding gene ENSG00000080815 has 7 transcripts: ENST00000261970, ENST00000394164, ENST00000406768. associated peptides: ENSP00000261970, ENSP000003263

Clicking on the Ensembl gene ID (ENSG00000080815) brings us to the gene summary page.

http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000080815

There are seven transcripts in this gene! That means there are seven slightly different, though related, proteins encoded by this gene.

Gene: PSEN1 (ENSG00000080815)

Presenilin-1 (PS-1)([EC 3.4.23.-](#))(Protein S182) [Contains Presenilin-1 NTF subunit;Presenilin-1 CTF subunit;Presenilin-1 CTF12

Location [Chromosome 14: 72,672,908-72,756,862](#) forward strand.

Transcripts There are 7 transcripts in this gene: [hide transcripts](#)

PSEN1-001	ENST00000324501	ENSP00000326366	protein_coding
PSEN1-002	ENST00000357710	ENSP00000350342	protein_coding
PSEN1-003	ENST00000394164	ENSP00000377719	protein_coding
PSEN1-004	ENST00000394167	ENSP00000377712	protein_coding
PSEN1-005	ENST00000406768	ENSP00000385948	protein_coding
PSEN1-201	ENST00000261970	ENSP00000261970	protein_coding
PSEN1-202	ENST00000344094	ENSP00000339523	protein_coding

Click on the first transcript, ENST00000324501, to learn more.

Once we're in the transcript summary page, we can learn more specific things about this protein.

http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core:g=ENSG00000080815;r=14:72672908-72756862;t=ENST00000324501

Click on '**General identifiers**' at the left to learn more about PSEN1 (Presenilin-1). This link is circled in the picture below.

Location: 14:72,672,908-72,756,862 | Gene: PSEN1 | Transcript: PSEN1-001

Transcript: PSEN1-001 (ENST00000324501)

Presenilin-1 (PS-1)([EC 3.4.23.-](#))(Protein S182) [Contains Presenilin-1 NTF subunit;Presenilin-1 CTF subunit;Presenilin-1 CTF12(PS1-CTF1 Prot;Acc:P49768) source: 3.4.23.]

Location [Chromosome 14: 72,672,932-72,756,862](#) forward strand.

Gene This transcript is a product of gene [ENSG00000080815](#) - There are 7 transcripts in this gene: [hide transcripts](#)

PSEN1-001	ENST00000324501	ENSP00000326366	protein_coding
PSEN1-002	ENST00000357710	ENSP00000350342	protein_coding
PSEN1-003	ENST00000394164	ENSP00000377719	protein_coding
PSEN1-004	ENST00000394157	ENSP00000377712	protein_coding
PSEN1-005	ENST00000406768	ENSP00000385948	protein_coding
PSEN1-201	ENST00000261970	ENSP00000261970	protein_coding
PSEN1-202	ENST00000344094	ENSP00000339523	protein_coding

Now we see a lot of information from scientific experiments that were done all over the world!

Click on one of the 'MIM disease' IDs to see what the gene is associated with in the 'Online Mendelian Inheritance in Man' database.

MIM disease: [600274](#) [\[view all locations\]](#)
[607822](#) [\[view all locations\]](#)

The first hit shows an association of this gene, this particular protein, to Frontotemporal Dementia!

[#600274](#)

FRONTOTEMPORAL DEMENTIA; FTD

Alternative titles; symbols

DEMENTIA, FRONTOTEMPORAL
 FRONTOTEMPORAL LOBAR DEGENERATION; FTLD
 DEMENTIA, FRONTOTEMPORAL, WITH PARKINSONISM
 FRONTOTEMPORAL DEMENTIA WITH PARKINSONISM
 FRONTOTEMPORAL LOBE DEMENTIA; FLDEM
 FTDP17

Congratulations! You just used the genome browser to find a disease-related gene.

If you were a biologist working on Alzheimer's disease, you might want to know if there is a similar gene in the mouse. Finding such a gene could permit you to do experiments in mice which might tell you something about Alzheimer's disease. You can use Ensembl to give you that information, too.

Go back to the gene summary page here:

http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;q=ENSG00000080815

Click on 'Orthologues' at the left. This is a list of similar genes to human PSEN1 in a lot of different species.

Opossum (<i>Monodelphis domestica</i>)	1-to-1	ENSMODG00000006792 Target %id: 87; Query %id: 88 [Align]	PSEN1 Presenilin-CTF12] [S
Mouse (<i>Mus musculus</i>)	1-to-1	0.06994 ENSMUSG00000019969 Target %id: 92; Query %id: 92 [Align]	Psen1 No descrip
Microbat (<i>Myotis lucifugus</i>)	1-to-1	0.11892 ENSMLUG00000015451 Target %id: 52; Query %id: 47 [Align]	PSEN1 Presenilin-CTF12] [S
Pika (<i>Ochotona princeps</i>)	1-to-1	0.09214 ENSOPRG00000001989 Target %id: 88; Query %id: 86 [Align]	PSEN1 Presenilin-CTF12] [S
Platypus (<i>Ornithorhynchus anatinus</i>)	1-to-1	ENSOANG00000001055 Target %id: 80; Query %id: 88 [Align]	PSEN1 Presenilin-CTF12] [S

And we've just discovered a very important piece of information: there is a gene in Mouse that is very similar to the human gene - in fact, it's so similar that it has the same name (preselin 1, or PSEN 1 for short)! A similar gene is found in Opossum, Microbat, Pika, and even Platypus!

Why do humans and mice have some of the same genes? We evolved from a common ancestor about 75 million years ago, and both species have inherited that ancestor's genes. Genomes allow us to directly see how the codes have evolved over time, and so we can make some very good guesses about how long ago our common ancestor lived, and what it looked like.

Over time, mutations and various other things have made human and mouse genes different. The latest comparison of the mouse and human genomes show that we share about 99% of our genes! So there's a real chance that by finding the mouse relative of a human gene, we can learn something about diseases.

Click on the mouse gene identifier showed in the screenshot on the previous page ([ENSMUSG00000019969](#)).

http://www.ensembl.org/Mus_musculus/Gene/Summary?g=ENSMUSG00000019969

Now we're surfing the mouse gene!

Location	Chromosome 12: 85,029,152-85,076,149 forward strand.		
Transcripts	There are 2 transcripts in this gene: hide transcripts		
	Psen1-201	ENSMUST00000041806	ENSMUSP00000048363 protein_coding
	Psen1-202	ENSMUST00000101225	ENSMUSP00000098786 protein_coding

Gene summary [help](#)

Name	Psen1 (MGI (automatic))
CCDS	This gene is a member of the Mouse CCDS set: CCDS26030
Gene type	Known protein coding
Prediction Method	Transcripts were annotated by the Ensembl genebuild .

Again, there are multiple transcripts. Two similar proteins can be encoded by the mouse PSEN gene. Click on one, **ENSMUST00000041806**, and click 'General identifiers' at the left of the page, as before. Let's see what **UniProt** has to say.

This Ensembl gene entry corresponds to the following database identifiers:

MGI Symbol:	Psen1 [view all locations]
UniProtKB/Swiss-Prot:	PSN1_MOUSE [Target %id: 100; Query %id: 100] [align] [view all locations]
RefSeq peptide:	NP_032969.1 [Target %id: 100; Query %id: 100] [align] [view all locations]
UniProtKB/TrEMBL:	Q3TDW2_MOUSE [Target %id: 72; Query %id: 100] [align] [view all locations] Q3UYK2_MOUSE [Target %id: 100; Query %id: 100] [align] [view all locations]
UniProtKB/SpliceVariant:	P49769-2 [Target %id: 54; Query %id: 98] [align] [view all locations]
EntrezGene:	Psen1 [view all locations]

Follow the link to the UniProtKB/Swiss-Prot record (**PSN1_MOUSE**).

Then click '**Retrieve P49769**'

UniProt knowledgebase is one of the best, most comprehensive database of protein information in the world! In the UniProt record, we see a lot of information about the gene and protein.

<http://www.uniprot.org/uniprot/P49769>

★ Reviewed, UniProtKB/Swiss-Prot **P49769** (PSN1_MOUSE)
 Last modified January 20, 2009. Version 92. [History...](#)

Clusters with 100%, 90%, 50% identity | Documents (3) | Third-party data | Customize display

[Names and origin](#) · [Protein attributes](#) · [General annotation \(Comments\)](#) · [Ontologies](#) · [Binary interactions](#) · [Alter Entry information](#) · [Relevant documents](#)

This record gives us a lot of information, including important scientific articles that people have written. In the list near the end of the record, you'll see an article called, "Molecular cloning and tissue distribution of presenilin-1 in senescence accelerated mice (SAM P8) mice". Those mice have a disease called SAM which is similar to Alzheimer's in humans. So researchers are using these mice as a model to understand the human disease, and they are doing the same thing with malaria and many other diseases.

What you've now done with Ensembl is the same thing that scientists do every day to discover new things about the genome. Genome browsers contain such a huge amount of information and there is a lot of uncharted territory on the map! Anyone can get to them directly on the internet. If you browse around long enough, and learn to ask new types of questions using its functions, you'll eventually discover things about our genome that no one else has ever seen.