Activity packet for teachers and schools:
Surfing human and mouse genomes on the Internet.

Most people have heard that "solving" human and mouse genomes will give us information useful in understanding disease and human health, but they may not know how scientists plan to use the 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 animal genomes accessible to the public on the Internet. 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 it, and how to ask your own questions about the genome.

Each cell of your body has a nucleus, a compartment that holds DNA. This is 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, C, and T. Genes, which are the recipes used to make protein molecules, make up only part of the DNA (about 2.5%) consists of genes. 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 sort of 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 about one in every thousand "letters" of the sequence is different.

The genome has already provided us with some fascinating insights into human evolution. Two 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."

From here you can examine the complete genome sequences of several animal species (you'll see the list in the column **Ensembl 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 isconstantly 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/

Now you see a series of bars on the left side 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.)

On the left you see cartoons of the 22 chromosomes and the X and Y chromosomes. (If you 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. Now you see a magnified picture of the chromosome on the left of the screen. The enlargement shows us some more detailed features of the DNA.
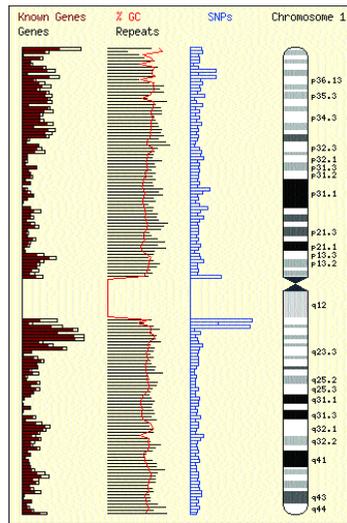
The brown bars in the left column show the "gene density" – how many genes are found in a particular region (remember that only a small percentage of DNA is actually used as genes). Some regions are very dense (long brown bars); there is also a region in the middle with no genes at all.

The column farthest to the right contains a close-up view of the chromosome. 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.

On the right side you'll find some additional information about chromosome 1. For example, "Known Ensembl Genes" means that 1855 genes have been directly found in experiments. A scientist may have found a protein in a cell, analyzed its chemical code, and discovered that the gene for the protein was located on Chromosome 1.

You'll also see something called "novel genes." Interestingly enough, until the human genome was decoded, no one had a good idea of how many genes human beings really have. It used to be that genes could only be discovered through difficult experiments. Along the way, though, people learned that the DNA code for most genes

contain some special features (a sort of text saying, "A gene starts here, and stops here"). Now that we have the whole sequence, we can use computers to try to read this code. "Novel genes" have been discovered in this way. It isn't a perfect method, and the computer program is being updated all the time. For example, by looking at the mouse genome we've found some new cues to identify genes, and this led to the discovery of over 200 new human genes. This is another reason why there will be a lot of new versions of the human genome for many years to come!

A "novel gene" is a prediction, and predictions can sometimes be wrong. So the only sure way to find a gene is still through an experiment.

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/contigview?chr=1&vc_start=78451177

Let's do that now. Some new boxes appear.
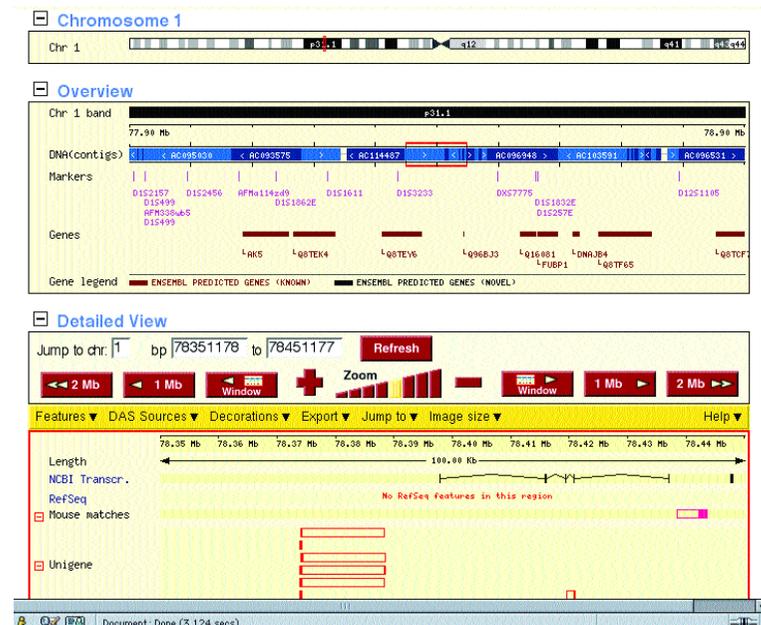
The upper box in "Overview" tells you the position you are at, like the coordinates on a map.

The second box shows the genes (known from experiments or predicted by computer) in that region.

You can click on one of the genes in that box, and it will slide the chromosome map down to where that gene is in the center. For example, zoom to or type in:

http://www.ensembl.org/Homo_sapiens/contigview?highlight=&chr=1&tvc_length=1000000&vc_start=78351178&vc_end=78451177&vc_pix=596&vc_left=98&click_to_move_window.x=353&click_to_move_window.y=127

When you do that, you've zoomed in as far as it goes – to the level of the DNA code, the list of bases (A,C,G,T) that make up DNA. You've now arrived at the code. It appears in the box at the bottom. This is the recipe

for a gene. If you had the right equipment, this information would let you "manufacture" part of this gene in your laboratory just by gluing single nucleotides together! (This used to be very difficult, but scientists are getting better and better at synthesizing DNA all the time.)

There's another line called "Ensembl Trans." To the right of the text you'll see squares and bars connected by a wiggly line. Click on one of the squares.

For example, http://www.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000077254

takes you to what's called the "Gene View." To understand what this page tells you, you have to know a little bit about the architecture of genes.

We said before that genes make up only a small part of the genome. Scientists are still trying to learn what the rest of the DNA does. Often genes are separated by long stretches of such "mystery" sequences. And when you finally get to a gene, it is often scattered in bits and pieces.

The end result is like opening the wrong kind of a file in your word processing program. For example, try opening an "html file" – the special format used to make up web pages – in a word processor. You may see the text, but it will be scattered out among a lot of instructions which are commands that tell another program how to turn it into a web document. In genes, the parts used to make proteins (the "text", called *exons*) are often interrupted by other DNA sequences called *introns*, which often contain the biological equivalent of formatting instructions. (Or like advertisements stuck in the middle of a recipe book.)

If you look in the box called "predicted transcripts," you'll see small squares representing the exons. You'll also see a link to a "protein sequence." While the DNA code consists of four letters, proteins are made up of a more complex chemical code with 20 letters. The cell has a machine that translates the DNA code into protein chemistry.

If you click on the "view protein information" link, you might find information about the function of the protein. A scientist might have done an experiment with it somewhere; he might have discovered what job it does in cells. Let's try it by clicking on

ENSP00000263184

In this case scientists have discovered that the protein is something called a PVHL-INTERACTING DEUBIQUITINATING ENZYME 1. Whew. Sorry, you'll have to ask a biologist who is a specialist in enzymes what that is!

Zooming in on chromosomes like this might be fun, but you can also ask questions of the genome the other way around – which is one of the most important ways that scientists actually use the data. Let's try to find a gene that has been linked to a disease... We'll take Alzheimer's Disease as an example. Are any genes are known to play a role in this illness? If so, where can they be found in the genome?

Go all the way back to http://www.ensembl.org/Homo_sapiens/

There's a drop-down menu at the top (the grey button, which now says "Anything"). If you hold your mouse button down on this menu, you'll see various ways to search for things. Choose "Disease" by pulling your mouse down to that item.

Next to this menu button (which should now read "Disease") is the word **with**, and next to that is a blank text area. Type "alzheimer" in this area. Make sure you spell it correctly, or you won't get any results!

To start the search, click on the button marked "lookup".

You'll receive a list of several links. There are so many because scientists have discovered several different forms of Alzheimer's disease, and they are probably related to different genes.

Click on <u>OMIM Disease: Alzheimer disease-3</u>

There are several genes related to this disease.

Click on the link below Alzheimer disease-3. Choose the one marked "ENSG00000080815"

Find Disease ▾ Alzheimer disease-3 (3)   Lookup   [e.g. cancer]   Help

Search for "Alzheimer disease-3 (3)" (4 OMIM diseases)   Display all OMIM diseases on Chr: 1 ▾ Lookup

Ensembl id	HUGO synonyms	OMIM id	Chromosome	Cytolocation
Alzheimer disease-1, APP-related (3)				
ENSG00000142192	APP, AAA, CVAP, AD1	104760	21	q21.3 - q22.05
Alzheimer disease-3 (3)				
ENSG00000080815	PSEN1, AD3	104311	14	q24.3
Alzheimer disease-4 (3)				
ENSG00000143801	PSEN2, AD4, STM2	600759	1	q31 - q42
{Alzheimer disease, susceptibility to} (3)				
No ensembl prediction	BMH	602403	17	q11.2

Date : 2002-09-09 14:39:44   Help Desk / Suggestions

We are looking at the "transcript view" – again, this means the protein code for the gene. You've found a protein called *Presenilin 1*. Somewhere, sometime, a researcher discovered that this molecule plays some role in Alzheimer's disease.

e! Ensembl Human GeneView   The Wellcome Trust Sanger Institute   EBI

Home ▸ Human ▴ What's New ▴ BLAST ▴ SSAHA ▴ MartView ▴ Export Data ▴ Download ▴ Disease Browser ▴

Find Gene ▾ ENSG00000080815   Lookup   [e.g. ENSG00000139618, BRCA2]   Help

Ensembl Gene Report

Ensembl gene ID	ENSG00000080815
Genomic Location	View gene in genomic location: 71108203 - 71192133 bp (71.1 Mb) on chromosome 14. This gene is located in sequence: AC004858.2.1.125202
Description	PRESENILIN 1 (PS-1) (S182 PROTEIN). [Source:SWISSPROT;Acc:P49768]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases.
Predicted Transcripts	1: ENST00000261970 View supporting evidence View protein information 2: ENST00000261971 View supporting evidence View protein information 3: ENST00000286500 View supporting evidence View protein information

ENST00000286500 (NM_007319)
ENST00000261971 (PSEN1)
ENST00000261970 (PSEN1)
83.93 Kb

Researchers do exactly what you've just done all the time. Two years ago an EMBL team discovered that a molecule known to be involved in Alzheimer's disease was probably being cut in an unexpected way by another molecule. This was a surprise because the other molecule was well-known and had been heavily studied. In fact, there had been medical studies because this protein is affected by a drug often used to treat stroke victims. Suddenly there was a connection between a drug, strokes, and Alzheimer's disease - potentially of enormous value to medical researchers. If that information isn't stored very carefully in a database, the person who needs it may not find it. It's a huge challenge to keep all this information up-to-date.

Maybe someone has a piece of information about the Preselin 1 gene that can be helpful to us. Let's see where the gene can be found in the genome.

In the **genomic location** box, click on this:

<u>View gene in genomic location: 71108203 - 71192133 bp (71.1 Mb) on chromosome 14</u>

This takes us back to the genomic view, with the PSEN 1 gene (the abbreviation for Preselin 1) in the overview.

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.

The bottom box of the page you are on right now has a line called "Mouse matches." To the right of this text is a bar. Click on it; this will take you over to the Mouse genome.

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)!

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 "Ensembl trans" to have more information about the mouse gene. This is what you would do to find out what experiments have told us about the mouse gene. This link will let you dip into other databases that hold that kind of information. The best protein database in the world is called SWISSPROT – created twenty years ago by a charming man (from Switzerland) named Amos Bairoch. Amos still personally approves every piece of information about a protein that gets into SWISSPROT. So let's go over to the links box and find out what he can tell us about preselin 1. In the links section click on SWISSPROT: PSN1_MOUSE.

This gives us a lot of information about the gene and the protein, including important scientific articles that people have written about their experiments with it. In the list you'll see an article called, "Molecular cloning and tissue distribution of presenilin-1 in senenscence 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. The database contains such a huge amount of information that there is a lot of uncharted territory on the map! Anyone can get to it directly on the Inernet. 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.

11. Reading:

Molecular biology and the future of drug development

Over the past century, biology has gone from a focus on whole animals and organisms to a close-up look at what single molecules do in cells and organisms. Thanks to this new type of science, researchers no know how many medications work and they also have completely new ways to design and test new drugs.

In the past, most drugs were created when a researcher took a substance – often from a plant – and test it on animals and humans. If it seemed to have an important effect, chemists would search for the "active ingredient." They might then modify this substance in many ways to try to make it more powerful, or to remove side effects.

They nearly always did these things without a detailed knowledge of how the drug worked. You might assume that it somehow changed the chemistry of cells, but there was no way to get a close-up view of what was happening. It was a little like dropping an object into the top of a huge machine and waiting for something random to come out the bottom. If you were lucky, something good would drop out, and you would learn what to drop in at the top to get what you wanted at the bottom. But you couldn't look at the mechanism inside to discover the logic of what was happening, and sometimes there were nasty side-effects or breakdowns that you couldn't understand.

That situation is changing thanks to our understanding of single molecules – especially proteins. These "machines" perform a lot of different jobs in cells, usually by linking up to each other. Some proteins take in substances from food and process it so that cells can use it for energy. Others pass along information, for example the news that a virus is banging on the outer wall of the cell, or that it has gotten a lot colder outside. Still others sit on the surface of the cell and "taste" the outside environment, looking for information.

Like a piece of a machine, a protein works a certain way because its shape and physical characteristics let it interact with other parts. If something (often another molecule) changes those physical properties, it will function differently. Drugs often have their effects because they grab onto one of the cell's molecules and change its shape or chemistry. Plugging up a sticky

hole on a protein's surface, for example, might shut the door on another molecule that needs to worm its way through the gap. Or a drug might cut up a misbehaving protein, scattering its parts so that it can no longer function.

To find out which of these things is happening, scientists like to obtain detailed building plans of a molecule's structure, and then you can watch how the plan changes when the drug comes along. You can't see a single molecule through a microscope, but there are other ways to capture very high-resolution pictures of how molecules are structured. One of the most important techniques is called X-ray crystallography.

Such methods can give scientists exact maps of proteins. Such images often explain key aspects of a molecule's behavior. They might reveal exactly where the hole is on a protein's surface, the one you're trying to seal up, and then you can make an artificial adaptor plug to stick into the hole. Structural studies can also give us pictures of how whole machines, composed of several molecules, are put together.

This type of research is now being combined with other new technologies to improve medical research. An experiment with DNA chips, for example, can give researchers an idea of genes and proteins that help the malaria parasite invade cells. The next step is to find a substance, or design an entirely new drug, which alters the activity of these critical molecules.

X-ray studies used to be difficult and time-consuming. But thanks to the efforts of physicists, mathematicians, engineers, computer experts, and biologists, the process is speeding up. Drug companies are gearing up to test hundreds or thousands of alterations of single molecules and get a precise view of how they affect key proteins. This can eliminate years of testing that, in the past, had to be done with cells, animals, and human subjects. It promises to revolutionize the process of creating new drugs.