Good morning. Welcome back. So, the Red Sox won, it's pretty convincing, yeah, very good. Yay Red Sox.

So, as you can also tell, I have something of a cold, so I'll see if I, if my voice makes it through, but what I wanted to do today, if the voice allows, was to talk about genomics.

Now, this is a little bit different than what we normally do in the class because, I work on genomics, it's something I'm extremely interested in.

And so, what I wanted to do today, and I'll do it one more time before the end of the term, is to talk about research that's going on in genomics, give you a sense of what's really going on. I can assure you that what I say is not going to be in the text book, or any other text book. And, I'm not entirely sure how this might appear on an exam, so don't ask, because I'm really just going to talk about research that's going on today.

And part of the purpose in doing that is to a, show you that it's possible for you to understand the kind of research that's going on in this field, and b, to excite you about what's going on in this field. So each year I pick different things to talk about, and I've picked a few things, and we'll see. So feel free to interrupt and to ask questions, and all of that, but this is very much more, sort of the edge of genomics, including stuff that's going on, you know, right now as we speak. So, we'll fire away.

So a little introductory stuff. I call this, we can actually keep the lights up, I think people, can people read that? Yeah, it's fine, good, so we'll leave the lights up and I can see people.

So, I think the thing that sets apart this revolution of biology that we're looking through right now, is the transformation of biology, not just from being the study of living organisms, to the study of chemicals and enzymes, to the study of molecules, but to the study of biology as information. That is what's distinctive about this decade, is the idea that the information sciences have begun to merge with biology, or biology merged with information sciences, and that it's having a profound effect on driving biomedicine. In both of the two talks I'll give, this one and near the end of the term, that will be the common theme, because I think that's the most important thing that's going on right now. Now, just to remind you, of course, the idea that biology is about information is an old one, it goes back to my hero, Gregor Mendel, with the recognition that information was passed from parent to offspring, according to rules.

And, as you know, the history of biology in the 20th century can be read as the development of biology's information.

The first quarter of the 20th century was the development of the idea that the information lives in chromosomes. The next quarter of the 20th century, the idea that the information of the chromosomes resides in the DNA double-helix, and that information was contained in this molecule, and somehow in it's sequence, and you know all of this. And the next quarter of the 20th century, basically from 1950 to 1975, understanding how it is that the cell reads out that information, from DNA to RNA to protein, how it uses a genetic code to translate RNA's into proteins, and the development of the tools of recombinant DNA that made it possible for us to read out the information that the cell reads out.

So that brought us Â¾ of the way through the 20th century, with the ability to read out genetic information, at least in little ways, but they were little ways. You could write a PhD thesis, around that time, for sequencing 200 letters of DNA.

That would be, you know, considered amazingly exciting PhD thesis. The next quarter of the 20th century, the last quarter of the 20th century, was characterized by a veracious appetite to read as much of this information as possible.

It started, first, with trying to read out the sequence of individual genes, then sets of genes, then genomes of small organisms' bacteria, medium-sized organisms. And then, you know, in a wonderful closure to the 20th century, the reading out of the nearly complete genetic information of the human being in the closing weeks of the 20th century. When you remember that, that Mendel was rediscovered in January of 1900, that's when the papers rediscovering Mendel came out, and you figure you've got perfect bookends from the rediscovery of Mendel in January of 1900, to the sequencing of the human genome in around 2000.

You realize what a century can do. It's not bad, as centuries go, you know, to accomplish all that, and it gives you know, as students, you get a point estimate in time of what science knows, but you guys aren't old enough yet and haven't lived long enough yet, to measure the derivative, and see how rapidly it's changing.

But just look at what happened over the course of that century, and then just project forward to what that can mean for the next century. So what that's done is it's brought us to the next picture. I have a picture in my head, of biology as a vast library of information, a library of information in which evolution has been taking patient notes.

Evolution is a very good experimentalist, and it's a very patient note taker. It's notes, of course, are written in the genomes, and everyday evolution wakes up, changes a few nucleotides, sees how the organism works, if it was an improvement, evolution keeps the notes, if it was disadvantageous, evolution discards the notes.

That, by the way, for those of you working in labs, is no longer considered appropriate laboratory practice.

You're obliged to keep your laboratory notes from failed experiments, as well, but evolution got into this before those rules were codified, and so it discards the notes from unsuccessful experiments, and keeps the notes from the successful experiments. But nonetheless, we have all the notes from the successful experiments, and we can learn a tremendous amount from it. There's a volume on the shelf corresponding to each species on the planet. There's a volume on the shelf corresponding to each individual within each species, to each tissue within each individual within each species, and there's information there about the DNA sequence, about the RNA readouts, about the protein expression levels, and in principle, even if not yet in practice, we can pull down any volume we want, and interrogate it, and compare it for related species, for individuals within a species, some of whom might have a disease, some of whom might not, for different kinds of tissues treated in different ways.

That is, I think, going to be a tremendous theme of biology going forward, and that's why it's a particular pleasure to teach biology at MIT, where you guys understand what that could mean, that fusion could mean. Now, this idea of extracting genomic information in large-scale, is a relatively new one. In the mid-1980's, the scientific community began debating what was a pretty radical idea, sequencing the human genome.

This was floated in a couple of places, in 1984 at one meeting, somebody raised the idea, you've got to realize that sequencing itself, that sequencing DNA, only came from the late 70's, so within six, seven years of being able to sequence anything, people were now saying, let's sequence everything.

That was a reasonably audacious thing to do, and it was controversial. There were many people who felt that the human genome project was a terrible idea, and with good reason, because the initial version of the human genome project was, kind of, a blunderbuss approach.

It was, let's immediately mount a massive factory and start sequencing the human genome with the just horrible technologies of the mid-80's, with radioactive sequencing gels, and you know, lots and lots of people doing stuff.

And so, you know, many people in science were, were concerned that an entire generation of students would need to be chained to the bench, sequencing DNA. Sydney Brenner, a great molecular biologist, proposed the whole thing be done at institutions [LAUGHTER], because you know, people could be sentenced to, 20 million bases, with time off for accuracy, or things like that [LAUGHTER]. And so what happened was, the scientific community came together well, in it's best form.

Group, a group was put together by the National Academy of Sciences, who said, well look, this is a really good idea, but we also need a carefully thought-through program to do it.

We need intermediate goals that will get us things that will advance the science along the way, we need to improve the technologies, and laid out a plan. The goals of that plan, to develop a genetic map, a map showing the locations of DNA polymorphisms, sites of variation, genetic markers, just like Sturdiman did with fruit flies, but to do it with humans, and with DNA sequence differences, to be used to trace inheritance.

That, that genetic map could be used to map human diseases, and if all you accomplish was, got a human map of the human being, that would be a good thing. Then you could get a physical map of the human being, all the pieces of DNA overlapping each other, so that you would know if you had a genetic marker linked to cystic fibrosis, you would be able to get the piece of DNA that contains the gene. Then, if we managed to pull that off, we could get a sequence of the human genome, all three billion nucleotides, on the web, so that you could go to just any place on the genome, double-click, and up would pop the sequence. Now, you guys of course, don't laugh at that, but about eight years ago, when I would give talks about this, I would speak about, oh you'll be able to go double-click and up will pop the sequence, and of course, everybody thought that was really funny, and that, that was something people laughed at. But of course, you can just do that today, if anybody has a wireless you can just double-click, and up will pop the sequence. And then, of course, a complete inventory of all the genes within that sequence. And a very importantly, and from the very beginning, the notion that all this information should be completely, freely available to anybody, regardless of where they were, whether in academia, or industry, in first world, third world countries, that everybody should have free and unrestricted access to that information.

So a plan was laid out, I won't go into the details here, but the plan was laid out that involved work constructing genetic maps, physical maps, sequence maps, in the human, the mouse, and some model organisms, including the bacteria yeast, fruit flies, worms. And, quite remarkably, it largely went according to plan, over the course of about 15 years.

A lot of people in the scientific community came together and took up different tasks. I should say, with some pride, that MIT was by far, one of the leading contributors to this effort, having been involved in essentially every stage of this, the genetic mapping of human and mouse, the physical mapping of human and mouse, and the sequencing of human and mouse, and having been the leading contributor to the latter, and it's not an accident because MIT's a marvelous environment in which to undertake this kind of research.

It involved changing the way we do biology. Back in the mid-80's, when we sequenced DNA, we did it with radioactivity, remember I taught you how to sequence using radioactive label of a gel, and all that. That's how we did it, stood behind this plastic shield, and you loaded the gels. Of course, now it's done in a highly automated fashion. This is the production floor at the Broad Institute, which is here at MIT, where robots prepare all the DNA samples, so E. coli's grown up, and then you have to crack open the cells, purify the DNA, purify the plasmid, do a sequencing reaction, etc., etc. it's all done robotically there, and this is capable of processing, and does process, in a given day, about 200,000 samples per day. They then go, and this is all equipment designed by people here at MIT, and then commercially built for us. They then go to the back room where, actually, these are the previous generation of DNA sequencers, commercial detectors, those capillary detectors that have little lasers on them, there's a whole farm of them that sit there, and are able to get data out.

In the course of a single day, we can now generate about 40 billion bases, I'm sorry, in the course of a single year we can generate about 40 billion bases of DNA sequence.

The genome project itself, was a collaboration involving 20 different groups around the world, groups in the United States, United Kingdom, France, Germany, and Japan, and China. They were of different sizes, they used different approaches, but everybody was committed to one common cause of producing this information, and making it freely available, and everybody worked together. And for the rest of my life, when it comes to Friday, at 11 o'clock, I will always think genome project, because we had a weekly conference call of all the groups in the world working on this Fridays, at eleven, and it was a fascinating experience, there were many, many years of that. So a draft sequence, a rough draft sequence of the human genome, was published in the year, in February of 2001, it was announced with some fanfare in June of 2000, but the real scientific paper came out in February of 2001.

This was not a perfect sequence of the human genome, by any means. We discovered about 90% of the sequence of the human genome. It still had about 150, 00 gaps in it, it had errors. But, it still did have 90% of the sequence of the human genome.

For the next three years, people worked very hard, and, as of last April, a finished sequence of the human genome was produced, and was published a couple weeks ago, and it contains, our best guess, about 99.

% of the human genome, and it still has about 343 gaps, they're, we know what they are, we know where they are, but they're not sequence able with current technology.

That's the Â“finished human genomeÂ”. What is it like? Well, this is a picture of the genome, do we have a pointer, yes, I see here we do have a pointer. This is your genome here, this is chromosome number 11, and I'll call attention to some interesting bits. So these colored lines here, represent genes, or gene-predictions, based on both, sequencing of the DNA, and mapping them back to the genome, as well as computer programs that analyze the genome.

And, right here, you have a big pileup of lots of genes, very few genes of here. Lots of genes, few genes. Notice the places where there are lots of genes, match up with these light-grey bands, which are the light-grey bands of the microscope, on chromosomes. The places with very few genes match up with the dark bands in the chromosome.

Do you know why that is, that the gene-rich regions are these light bands, and the gene-poor regions are the chromosome dark bands? Me neither. Nobody has a clue. It's really, it's really just one of these things. We had no reason to expect that we'd see these striking patterns, and other genomes, e-coli, doesn't have this dense, urban cluster, and these big, rural plains that are gene-poor. This is very weird, and it's distinctive to mammals. You'll also notice that the gene-rich regions, here, are rich in G's and C's, they have different distributions of some repeat elements, it's all sorts of weirdness that comes from just looking at the genome. The biggest weirdness was the number of genes, the count of genes is, our best guess, about 22, 00 genes, if I had to pick a number today, it would be our count of genes, and of course, that's down from the 100, 00 that was in some textbooks, and it's down from even 30 to 40, 00 that was in the genome paper of February, 2001.

Our best guess is that it's really just about that range.

Genes, themselves, are very interesting.

When you look at, you know, if we only have 22,000 genes we know of, how do we manage to run a human being with so few genes?

It is, by the way, probably fewer genes than the mustard weed, or Arabidopsis thaliana. So, what do we do? Well, humans, one thing we may take comfort in, is that we, although we only have about 22,000 genes, there's a lot of alternative splicing, on average the typical gene, on average, has about two alternative splice products.

Some have many, some have few, but probably, when you're all done, those 22, 00 genes may encode 70-80,000 different proteins, and it could be more than that because we don't know all the alternative splice products, and what they do. But, if you ask, humans get credit for being really inventive or creative, for having lots of new genes that make us human, the answer is, no.

Not only are humans not different in their gene complement from other mammals, mammals, as a group, really haven't invented that much, when you get down to it. Most of the recognizable sub-domains of proteins, proteins are built up of sub-domains, recognizable sequences that have certain motifs that fold up in certain ways, or carry out certain enzymatic functions.

And it looks like our genomes, our genes, are mixed-and-matched combinations of many domains that were invented a long time ago, in invertebrates and before, and that most of evolutionary innovation in the more complex, multi-cellular animals, has simply been mixing-and-matching these domains in new ways, to get slightly different functions.

You don't get a lot of points for creativity, but it does seem to work.

By far, the most derivative of all, and what characterizes our genome tremendously is, when a gene works, make extra copies of it, and let it diverge slightly, and take up new functions. Really, your genome is just characterized by large expansions of families, immunoglobulin-like genes, intermediate filament proteins holding together the cytoskeleton.

There are 111 different keratin-like genes in your genome.

They're all different, they do different things, but they all came from one gene that was copied, copied, copied, at random, randomly duplicated, and then diverged to take up new functions. Growth factors, flies and worms managed to get by just fine, thank you, with two growth factors of the TGF beta-class, whatever that is. You have 42 growth factors of this TGF beta-class, all of which help communicate, cells communicate, in different ways.

And then, of course, all the olfactory receptors.

In your genome, you have about 1, 00 genes for olfactory, for smell receptors. This is what Richard Axel and Linda Buck won a Nobel Prize for this year, was their work on the olfactory receptors. Sad to say though, out of all your olfactory receptors, genes, most of them are broken. They're most pseudo-genes.

It's not true in dogs and mice, who keep their olfactory receptor genes in pretty fine-working order, but it's very clear that in primates with color vision, our olfactory receptor genes have been going to seed. They've been piling up mutations, and there's no selective pressure to keep many of them.

And, in fact, we've now shown, in a paper that will come out soon, that this process is accelerating dramatically in the last 7 million years since we diverged from chimps. And so, humans have almost completely lost interest in smell, that's not totally true, some of these olfactory receptors surely matter for various processes, but most of them are probably irrelevant right now.

And so, anyway, that's the nature of the genes there.

Anyway, another interesting fact that's worth mentioning about your genome is half of your genome consists of transposable elements, elements that simply duplicate themselves, and hop around the genome. Elements that are like viruses, they make a copy, sometimes in RNA, the RNA is copied back into DNA and slammed elsewhere in your genome. These elements, well the, there are four classes.

Alo elements, Line elements, Retro-Virus like elements, all these go through RNA intermediates, and use reverse transcription.

And then there's certain DNA transposons, that go through DNA intermediate. The number of copies of the aloe element, the aloe element that's hopped around your genome, you have about a million, you have a million fossils of this element. You say, why is it there, and the answer is, because it's there. Because anything that knows how to make a copy of itself, and insert it itself in it's genome, you can't get rid of. You can consider it, if you wish, an infection, but half of your genome consists of an infection, with these kinds of transposable elements.

Now that's it, yes?

Well, it's very interesting, what's the effect? Well, they do, some of them are transcribed and, it's very interesting.

Sometimes it's bad, one of them will hop into a gene and mutate it, and that's bad, that person will have a lethal mutation, but the genome has probably begun to use them, and count on their being there. So, when a bunch, when a transposable goes in, and creates a spacing, if you, for example, if an engineering committee came in and cleaned up the genome by getting rid of all the transposable elements, it would surely not work.

Because we have evolutionarily come to count on the spacing there.

It's sort of like, if in some very, some very messy attic, you put a cup of coffee down on top of a stack of papers, those papers may be utterly irrelevant, but now they're holding up that cup of coffee that you put down on it. And if you were to just, poof, magically get rid of them, the cup of coffee would come crashing to the ground.

So, you know it, they're just there, taking up space. Now sometimes, even more than that, a few of them have actually been co-opted into being human genes.

We know that a few of these transposable elements have mutated into being our genes that do something for us.

And others of them may do things in affecting the general neighborhood with regard to transcription, and so, instead of it being a parasite, think of them as a symbiont, that's a genomic symbiont, which takes some advantage of us, and we may, you know, have worked out a compromise to take some advantage of it.

Every time a copy is made of these, and it hops in the genome, some mutations may happen in the master element, but when it lands in the new place, we have a record of that hop. And if you reconstruct the sequence of the million AluI elements, you can see which ones are very close relatives of each other, and had to have hopped recently, and which ones are somewhat more distant relatives.

And you can build an evolutionary tree connecting all of the repeat elements that have hopped around your genome, and thereby attaching a date to each of them, as to when they hopped.

So it really is a fossil record, and you can figure out how many of them have been hopping at different times over history.

And we can even make a plot of that, this is long ago, sometime here, some 30 million years ago, there was a huge explosion and in transposion, transposons, in our genome.

We don't know why that happened, but it's very interesting, it does correspond to very interesting periods of primate evolution.

And then, interestingly, there's been a huge crash, and transposition has dropped dramatically. We have no clue why this is, but we have a whole fossil record here of the rate of transposition of different kinds of repeat elements around our genome, and people are now starting to try to figure out what in the world this means. So all this is sort of there, inherent in the sequence, and if you want the sequence, as I say, you can go to the web and pull all this stuff now. So how do we understand the sequence? Well, I've told you a little bit about it, from the simple things that we've done, but there's a lot more that needs to be learned about the sequence, so what I really want to turn to, is how we're extracting information out of this sequence.

So, DNA sequence is long and boring, it's only marginally more interesting than reading your hard disk, because it has four letters, instead of ones and zeros, but it's, you know, well, it's pretty really boring if you take a look at it. How do you attach meaning to all this stuff? One of the most powerful ways is by comparison with other genomes. And so, comparing the human genome to the mouse genome is very informative in many ways.

So, as soon as the human genome was far along, a portion of the international consortium, set to work getting a sequence of the mouse genome. And that was published in December of 2002. We have a nice map of the mouse genome, with all these things, it, too, shows these gene-rich regions, gene-poor regions, all sorts of funny things. And if we look closely at a portion of the human genome over here, I've picked about a million bases of the human genome, and we take any little spot in that million bases of the human genome, let's say over here.

And we take half the DNA sequence corresponding to this spot, and we run it in the computer against the mouse genome, and ask where in the mouse genome do we get the best match for this, the best match to this is here. Now let's do it for this piece, here. The best match anywhere in the mouse genome lands in the same million bases here as the mouse genome. In fact, for every single sequence that we pull out from this million bases in the human genome, the best match is in this million bases of the mouse genome. That's very interesting. Why is that? Sorry? No, people do know.

It, it was a good try, though. [LAUGHTER]. This million bases in the mouse genome, and this million bases in the human genome, represent the evolutionary descendents of a common million bases that occurred in our common ancestor 75-million years ago.

This is a clear evidence of the evolution here, because we can see that this is a segment of DNA from our common ancestor that really hasn't undergone much rearrangement, and we can just line up the sequences and see.

In fact, we can build a whole map across the mouse genome like this.

For any bit of the mouse genome, I don't know, here's a bit on mouse chromosome 17, this whole stretch corresponds to a portion of human chromosome number eight. This stretch here, I don't know, this green color here on chromosome number six, corresponds to chromosome four in the human. And so, we can build a look-up table that says, for any portion of the human genome, what's the corresponding portion of the mouse genome that came from the same ancestor, has basically the same complement of genes in it. And there's only about 330 such regions that we need to cut-and-paste the human genome order to the mouse genome order, roughly speaking. There's a lot of little local rearrangements, but at this gross level. So now, if we go back more closely and we look at this, and we say, OK, so now we look at this region, we now know these two regions descend from a common ancestor, if we do a careful evolutionary analysis, lining up all the sequences, and see how well-preserved the sequences are, some are much better preserved than others. Evolution has been much more lovingly conserving other sequences than others, and so, so let's now zoom-in on a gene, this is a gene that goes by the name, PP-Gama, I'm fond of this gene but, it doesn't matter. If we look, I've indicated all the regions here, in which there's a heightened degree of conservation. The sequence is well-conserved here, here, here, here, here, here, here, and here, here, here, here, here, here. These correspond to the exons of the PPR-Gama gene, they encode the protein of the gene, then the splicing goes like this, OK? These things here do not correspond to the exons. People have no idea what they are, in fact, this is not supposed to be here. The official textbook picture says, the vast majority of what matters for a gene, what evolution should preserve, is the exons plus the promoter.

Here's the promoter. But in fact, what we found is that an awful lot more is being preserved. In fact, across the genome, our best estimate is there are about 500,000 conserved elements across the genome, and only 1/3 of them are protein-coding exons.

That means 2/3 of the stuff evolution has been interested in, is not protein-coding exons, and the truth is, we do not know what it is, this was a very radical finding, when this mouse paper came out, about a year and a half, about two years ago now.

What it must be, I think, but we're guessing, are regulatory signals, the structural elements in chromosomes, RNA genes, but there's an awful lot more of it than we had imagined.

And we've, now we're in this fascinating situation, where computational analysis has told us what's on evolution's mind, and now we have to go to the lab and figure out what in the world it does.

But there's no doubt that it must do something, because evolution has preserved it quite well. Now, I oversimplified greatly in this discussion, let me first say, and I'll come back to that. We do know, if we take some of those elements, here's one, there's a 481 base-pair elements that's 84% identical between human and mouse.

You could write yourself a little statistical model to say that's way unusual to have something that's so well preserved. When Eddie Ruben and his colleagues from Berkley made a knockout mouse that deleted that segment, this knockout mouse loses regulation of three different genes in the neighborhood, saying that this must be a regulatory sequence that affects multiple genes in the neighborhood. That, that's one, with about 300, 00 such elements to go, in order to attach meaning to them. So doing this entirely by knocking out mice will be a slow process, one's going to need other ways to be able to attach meaning, but there's no doubt. Now, there's some other interesting papers where people have knocked some of these things out, and they've seen no effect on the mouse. They get a totally viable mouse. Can you conclude from that, that they have no function? Why not? The knockout mouse is viable.

Could be redundant, it could even not be redundant, but yes, it could be redundant, but you couldn't knock out both of two things. It turns out, suppose knocking it out affected the mouse's viability by part, ten to the third, it was only 99.9% as fertile, would you be able to see that in the laboratory? No. Would that matter to evolution?

It would be lethal, in an evolutionary sense.

Such mutation could never propagate through a population.

One part, and ten to the third, is massive selection against, from an evolutionary point of view, but almost undetectable in a laboratory batch. Evolution has a far more sensitive assay than we do. Now, I won't go into detail, but for the mathematically inclined here, showing that there really were about 5% of the human genome under, under evolutionary selection, it was a complicated affair, because with only two genomes, what we really had to do, and if this doesn't make sense, ignore it.

We looked at the background distribution of conservation of the genome in unimportant elements, in those repeat elements that we knew to be functionally broken. We looked at the overall conservation of the genome, and found that the overall genome has this rightward tail, by subtracting the distributions we were able to see how much excess conservation there was.

That's because we only had two genomes, we had to draw inferences.

If we had more genomes, like the mouse and the rat, and the dog and the-this-and-the-that, we would be able to extract signal from noise.

We would be able to see right away, which bits were well-conserved, and we wouldn't have to do this as a sensitive statistical analysis.

So, in fact, we need more mammalian genomes, so, so right now there's been a sequence of the rat genome in the past year or so, there's a sequence of the dog genome, we're writing up that paper now, but it's on the web already. There's a sequence of the chimpanzee genome we're writing up a paper on that, in collaboration with our friends in the genome-sequencing community.

We're currently sequencing a variety of other organisms, as well. And if you had enough organisms, you ought to be able to just line it up and say, what has evolution preserved, and figure out exactly which nucleotides matter, and which nucleotides don't, are allowed to drift freely, at the background rate. How far could you go with this?

Well, we decided to try an interesting experiment.

We said, since mammals are very big, then we're going to need a lot of genome sequences, how about we try a small organism, like yeast? What if we were to try to do this, this kind of evolutionary, genomic analysis on something like the yeast genome? And so, this is work that I'll describe, that was between a bunch of people here at MIT who do genome-sequencing, and a student in computer science, Manolis Kellis, was PhD student in computer science, he now just joined the faculty here at MIT in computer science. But it was a really great example of how biology and computer science could come together.

So, the genome-sequencing folks sequenced three related species, through our friend, the baker's yeast, Saccharomyces cerevisiae, workhorse of geneticist. These three different species are separated by different evolutionary distances, from Saccharomyces cerevisiae. When you line up their genomes, just like with human and mouse, you find the genes occur largely in the same order, and it's not hard to pick out, oh there's this gene there, there, it's all lined up, you've got these evolutionary segments, and very few rearrangements have occurred across these species, despite the fact that they're about 20 million years apart in history.

But here's an interesting thing. When the yeast genome, Saccharomyces cerevisiae, was first published in 1995, the paper describing it reported 6, 00 genes. Now, how did they know there were 6,200 genes? They ran a computer program looking for open reading frames. Any open reading frame, consecutive codons without a stop sufficiently long, was called a gene.

But statistically, you could, by chance, just have a long stretch of codons without a stop codon.

And so, if I saw 100 codons in a row, without a stop, they called it a gene, but it might just be chance.

And they knew that, of course, they wrote that in the paper, but for many years, people then had 6, 00 open reading frames, which were the yeast's genes.

Could evolution now tell us which one of them were real and which weren't? Well, it turns out that evolution was tremendously powerful in doing that.

If you take something that's a well-known gene that has been extensively studied by yeast geneticists, you line it up across all four species, you almost never see deletions.

And when you do see the lesions, here in grey, they're always a multiple of three. Why are they a multiple of three?

They preserve the reading frame. By contrast, if I take some clear, intergenetic DNA, that's not protein-coding, and I compare it across these four species, I see lots and lots of frame shifting deletions that occur, Evolution tolerates frame shifting deletions, and if I juts write down the rates, frame shifting deletions are 75x more common in intergenic DNA, than genic DNA. This provides a very powerful test.

Run this test across the genome, looking for the density of frame shifting deletions, any place that doesn't tolerate frame shifting deletions is probably a real gene, anything that does tolerate it is probably not. When you sorted through all this, it turned out that 528 of the official yeast genes were clearly not real, not real genes. They were just chock-a-block full of these frame shifting deletions. And, and a bunch of others could be confirmed. So the yeast gene count, and I won't tell you all the experimental and other that shows this is right, but the yeast genome has now been revised downward to 5, 00 genes, and we have great confidence that almost all of those are real genes, there are 20 whose origins that we're not sure of, and new genes could be found in this way. Here's a really audacious thing.

This graduate student in computer science said, I think, based on these other species, there was a mistake made in the sequencing of the first yeast, and that the reason these things are called two separate genes, is that somebody made a sequencing error that got a stop codon here, but I think these are really part of one gene. And so, somebody went back and re-sequenced some of these, and sure enough, he had correctly predicted that there had been a mistake made at that letter, and that these were in fact, a single gene.

The computational analysis was incredibly powerful in this regard, it could go further than this, you could ask, could I also figure out the way genes are regulated in this fashion, could I work out the intergenic signals in the promoter regions? Remember that lac repressor to a certain operator site, well, all of these regulatory proteins bind to different sequences, could we figure out what the sequences were, computational? Well, if we look closely at a genic, intergenic region, here's one where there's two genes being transcribed in opposite directions, gal-1 and gal-10, both involved in galactose metabolism, and there's a particular protein, a transcription factor here, called Gal-4, in this region, and it has a particular sequence that it likes, CCG, 11 bases, GGC. So, that Gal-4 we see, is very well preserved across all of the species.

So, in no regulatory sequence is well-preserved, now let's look at that closely. This Gal-4 binding site is a measly, crummy, six nucleotides of information. At random, it's going to occur in many places in the yeast genome, but not be a real, important Gal-4, right? Some of them matter, some of them don't. How do we figure out which of these occurrences are real Gal-4, well, if we look across all four species, what we find is that those occurrences that occur in promoter regions, are much more likely to be conserved by evolution than those that don't. So there's a special property here, conservation of the motif and the motor regions.

In fact, this particular sequence is four times more likely to be preserved when it occurs in a promoter region, than when it occurs in a coded region. And for a typical control region, the opposite is true. Since genes, since coding sequences are better preserved in general, for a randomly chosen sequence, I don't know, ATGGCAT, it's more likely to be preserved in coding regions than non-coding regions.

So this Gal-4 motif has a very funky property that, on average, it's 12x more likely than background, to be preserved when it occurs in a promoter. Now, that's a test you apply to another motif, and another motif.

In fact, you could, by computer, test all possible motifs, and ask which ones have that property? Make a scatter plot, most motifs are better conserved when they occur in promoter regions, than when they occur in coding regions, some however, are better preserved in promoter regions than in coding regions.

Our friend, Gal-4, is up there, but there are a lot more things like it, that are better preserved by evolution than promoters are. You can make a list of them. You can get about 72 well-conserved, regulatory motifs and it turns out that 20 years of yeast work produced knowledge about things like the Gal-4 site, and other sites. Almost all the known regulatory sites that had been discovered over the course of 20 years of experimental work appear on this list that falls out of the computer analysis of evolutionary comparison of genomes.

You can actually go a step further, I'll hesitate to tell you, but I'll try anyway. If you wanted to find out, without knowing in advance, what these motifs were doing, what their biological function was, you can do that informationally, too. It turns out that if I take my motif, Gal-4, and I ask, which chains does it occur in front of? Well, across Saccharomyces cerevisiae, you find this crummy little motif in many, many places because, as I said, most of it's just noise. But if I ask, which genes have this motif in all four species, these genes, there's a huge overlap with a class of genes involved in carbohydrate metabolism.

So, if I didn't know in advance that the Gal-4 motif was involved in regulating genes in carbohydrate metabolism, I could tell, just from the fact that the genes that'd conserved it, are genes involved in carbohydrate metabolism.

You can do that using all sorts of tricks, expression of genes, protein mass spec, blah, blah, blah, and the short answer is, for almost all of those motifs that you can find in the computer, by consulting public data bases of sets of genes that are co-expressed, or have similar properties and all that, the computer can also offer you a pretty good hypothesis about what that motif is associated with.

You can even go a step further than that. You can begin to look at pairs of motifs, you can say, if I have a certain regulatory sequence, number one, and a second regulatory sequence, number two, do they tend to be preserved in front of the same genes as each other? Is their conservation correlated? And you can build a map of these two guys tend, when this guy's correlated, this guy tends to be correlated. And you can say, oh those proteins must be talking to each other, and you can read that off from the patterns of evolution, as well. There are two regulators, one called Sterile 12, one called Tec1. This computational analysis shows that they tend to co-occur in a conserved fashion, far more often then you'd expect by chance. And when you do the analysis, you find that those genes that just have a conserved Sterile 12, those genes tend to be involved in mating. Genes that just have a conserved instance of Tec1 tend to be involved in the budding of the yeast, and those genes that have conserved the occurrences of both tend to be involved in fillamentation. Now all that can be read out, which is way cool, this is not the way we used to do biology.

Now don't get me wrong, there's a ton of experiments that underlay creating these databases, and there's a ton of experiments that have to be done to check any of these things. But what we have is one of the most powerful hypothesis generators that's ever been seen here. Evolution, by telling us what to focus on, is giving us, on a silver platter, hundreds of hypothesis about who's interacting with whom, and sending us back to the lab then, to test these hypotheses. Now, what are the implications of all of this for the human genome?

Could we do this for the human genome? Well, these species, Saccharomyces cerevisiase, S.

paradoxus, S. mikatae and S. bayanus, are they a good model for mammals? Well it turns out that their evolutionary distance from each other is the same as the distance of human to lemur, to dog, to mouse.

So they were chosen with a purpose. Those are actually fairly good models for the human. So could we do exactly the same analysis for the human, for the entire human genome?

If we had, human, lemur, dog, and mouse, are basically four species, human, mouse, rat, and dog.

Well, there's one little fly in the ointment. The human genome is 20x bigger than the yeast genome. If I want to analyze the whole human genome, I have a problem of signal-to-noise.

The genome is 20x bigger, I've got 20x as much noise to get rid of. I won't walk you through it, but I need more evolutionary information to get rid of all that noise. And, you can do a simple calculation that says, my evolutionary tree needs to be bigger, it's branch length needs to be bigger by about the natural log of 20, to get rid of 20 fold more noise.

And that would mean I'd need more species, I'd need about 16 species, or something like that to be able to do that. But if I built an evolutionary tree that had a branch length of four, that is, four substitutions per base across this evolutionary tree, as indicated by these colored lines here, I should have enough power to analyze the entire human genome, the way we just did the yeast genome.

So we currently have human, chimp, mouse, rat, dog. As of this fall, during in fact, right at the beginning of this term, the National Institute of Health signed off on the sequencing of these additional eight mammals. These mammals are now in process, and in fact, the elephant is done, and the armadillo is in process, and the tree shrew, I think, is being caught at the moment.

[LAUGHTER]. The ten-, don't talk about the tree shrews. The tenrec is actually being tested right now, etc, and all this is going on right now, as we speak, and I think that by next summer, we should have much of, and by certainly, by a year from now, we should have all this information to do such an analysis. That said, we're of course, very impatient people, you could just take the human, the mouse, the rat, and the dog. And I said that's not enough if you wanted to analyze the whole genome, but suppose you just wanted to analyze a portion of the genome, maybe about a yeast-size piece of the genome, well let's see, at 20,000 genes, I don't know, suppose I take, I don't know, two kilo bases around each 20, 00 genes, well that's you know, 40 mega bases of DNA, it's only a couple-fold more than yeast. Maybe, if I just focus on a limited region around each promoter, I could start reading out these regulatory signals, with just four species.

So in fact, the post-doctorate fellow is, has been working on this problem over the summer, and a little bit, too, through the spring and summer, together with Manolis Kellis, who's now in the computer science department. And I think we have a preliminary list for the human genome that's fallen out over the course of the past couple of months, and we're in the process, right now, of finishing up a paper that we're hoping to get submitted by Friday, with a preliminary list of regulatory signals in the human genome, read out from evolution of human, mouse, rat, and dog.

It won't be everything, we don't have full power to pick up all possible signals, but we're picking up a lot of the signals, we're picking up a very large fraction of previously discovered signals, and lots more new signals, as well, are falling out of that analysis. So anyway, I can assure you that that's not in the textbooks because, actually, it hasn't been submitted yet. This other stuff I've described about the yeast analysis, this, you do want to look it up, there's a paper in nature about a year and change ago, Kellis et. al. describes this yeast work. This is what's going on.

This is what's fun about teaching at MIT, as I can tell you this stuff, and you guys have a sense for the convergence that's going on in our field. Much of what I've tried to make the biology, you know, in making the biology clear, I've talked about how the different directions, genetics, biochemistry, have converged together. What we're really seeing now is information sciences converging with that as well, and I've got to say, it's a tremendous amount of fun. See you on Monday, good luck on the quiz.