The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license and MIT OpenCourseWare in general is available at ocw.mit.edu.

GEORGE CHURCH: Use this slide to review how far the in-situ might go, and what its current limitations might be. And then we'll move on to arrays. One potential advantage to these sort of microscopic in-situ analyses is that if you use a non-destructive visualization rather than fixing the cells-- you actually monitor them real time-- you can sample this basically as quickly as a modern microscopic camera system can monitor, on the order of a millisecond.

You can obtain a sensitivity on the order of a single molecule of fluorescence. This is very challenging. It requires very small pixel sizes, but it is possible. And it's the basis of some of the sequencing methods that we discussed a couple of classes ago. And so that resolution itself is typically on the order of a micron or a quarter of a micron, sort of set by the limit of the optics in terms of the diffraction that can typically occur.

But you can get below that diffraction limit of 250 nanometers down as low as 10 nanometers by using tricks such as near-field optics and various deconvolution methods. Multiplicity is really the greatest limitation of the in-situ method right now, and it's certainly an opportunity for the creative ones in this group to address.

How can we get through multiplicity, looking at all of the RNA simultaneously as we can do in microarrays, which is essentially a microscopic method as well, But still have the spatial advantages of an in-situ? This is an unsolved problem. Multiplicity now is typically around one or two or three colors. The colors can be deconvolved by using band filters.

If you use combinations of colors, you can conveniently discriminate the 24 different types of human metaphase chromosomes. However, this depends, you should note-- don't get fooled into thinking you actually have 24 colors. These are combinations of ratios of color, considering them to be false-colored by the computer algorithm. But they depend on non-overlap or of being able to find objects in the visual field and extend them.

If where they overlap, you now have mixtures of mixtures, this no longer is simply deconvolved. So for all and practical purposes, we're limited to around four or five colors. So in situ, I would not call immediately genomics-compatible. The systems biology, it takes a vast number of in-situ experiments to get the kind of comprehensive data you can get out of microarray experiments. So let's focus on the kind of experiments like microarrays that can get us full genome scale information, and what are the limitations in quality for these sort of things.

Now, we can either lump or split various ways of measuring arrays in cells. The top two items on slide 28 either can be characterized-- microarrays might be associated with, sort of for historical reasons, longer probes, maybe the length of an entire gene or entire cDNA, messenger RNA. And affymetrix and other oligonucleotide-based methods typically use 25-nucleotide-long ligament oligomers.

Typically, the long microarray probes are used as single probes, one probe per gene, while the short ones typically have 24 probes per gene. These are not necessary differences. You can imagine various combinations. Another difference is in the long probes, typically, you'll do an experiment and control with different colors, and they're mixed together to control for some of the variations that might occur.

In the spotting of the long probes, this is typically spotted mechanically, while the short probes are developed by a photochemical method in which 100 mass, sort of black and white mass such as use in the Silicon Valley for making computer chips, which we introduced in the sequencing technology lecture, these kind of 100 photomass will allow you to make 25-mer, four possibilities per base. And you'll make, say, 20 of these scattered along the gene, and a mismatch control for each of those 20 perfect matches, PM and MM abbreviation.

Those mismatched controls help you get at the possible cross hybridization by related sequences or even distantly related sequences. And then what they typically do is subtract the mismatch controls from the perfect masses, and then average across all 20 of them, or some statistically good sampling of those 20.

OK. So you typically do ratios for the long probes, and you try to get absolute amounts from the short probes. And then there are two wildly different methods in the bottom of the slide. These are called SAGE, standing for Serial Analysis of Gene Expression, and NPSS, an acronym for a highly parallel bead-based method. Both of them essentially sequence, determine the sequence of somewhere between 14 and 22 nucleotides. This is the sort of minimum length sequence that's often but not always sufficient to identify an RNA molecule.

You're basically counting individual RNA molecules with a tag that's just long enough to be able to recognize it in a database. At a 14-mer, you can recognize it and say, a human cDNA database, but it's not unique enough to identify it in a human genomic database. The 22-mer is large enough to get an acceptable rate of false positives, false negatives in a human genomic library. So that's kind of the range which you can do this.

And it's just, people tend to take shorter tags because the cost goes up with the length of the tag. And so these were conveniently short tags. So the top ones, you get quantitation by integration of the fluorescent signals. And the bottom two, you get quantitation by counting individual tags. The bottom two methods have the opportunity for discovery, while the top three, you basically can quantitate any gene or segment of the genome that you care to put on the array. But you won't necessarily discover anything outside of those features.

So these are four of the key methods that are used for quantitating RNAs right now on a genome scale, where you'll do hopefully multiple experiments for each type. Now, let's just zoom in a little bit more so you can have appreciation for where some of the systematic and random errors might creep in to these kind of experiments. And I'll just arbitrarily use the 25 oligonucleotide probe arrays as an example for the long arrays, the microarrays.

You might have, say, 1,000-nucleotide-long probes. You might have 10,000 of them on a glass slide about this big. With the photolithography, you can have a more like closer to a million features in a square centimeter. And in each of those features, each of those million positions on the array, in the square array, you will have maybe 10 to the 5th and 10 to the 6th molecules all identical in position 1, and then a new set of 10 to the 6th molecules in position 2, all aimed at a different RNA or different part of an RNA.

Each of those probe cells is ready to accept, fluorescently labeled or using biotin as an intermediate in order to get fluorescence. So you take your RNA and you directly [? biotinylate ?] it, or you make a cDNA copy. Or in one way or another, you introduce a fluorescent or bioitn molecule into a copy of your RNA. And then you apply that to the chip, and they will bind kinetically.

And the more the mass action that you get from the original RNA, the RNAs that are the most abundant will result in the largest number of biotins or fluorescent molecules on the array at a given element. Here, an indirect conjugate with a fluorescent stripped out of [INAUDIBLE] It's like a rift in this covalently attached [INAUDIBLE], and you get a fluorescent signal, which you quantitate, which tells you the amount of original messenger RNA.

If you have 20 different oligonucleotides per gene, you can scatter this about the array or you can have them in lines. This tends to be back from when they had them in lines. So you get the streakiness there. Now, one of the first things you want to do, much of the software from the companies is set up on the assumption that you will only do the experiment once.

Now, this may have been appealing in the early days from a cost standpoint, but it's not really cost-effective, in that you will make mistakes and you will draw incorrect conclusions that will require you to go back. But this is an example of an early experiment to establish the reproducibility from one experiment to another, possibly to reassure people that they didn't need to repeat the experiment.

But in any case, this is the thing that is now commonly done in order to assess that your experiments are indeed reproducible. And what you expect from this as you go along the horizontal axis towards higher and higher copies per cell, going off to the right or going up on the vertical axis, when you get high copies per cell, then you expect there to be very close similarity in the two measures from two different experiments done at two different days.

And then as you get to the very rare transcripts, you expect the various sources of noise in the experiment to start to dominate the light scattering and the array. The background fluorescence of the glass, the non-specific cross-hybridization between different RNAs, start to dominate over the true signal because, the true signal is going down and all those background signals are staying constant. So you start to get spread at a low number of copy numbers per cell.

You can see a huge fraction of the RNAs in the yeast cells that present a single copy, say one or less fewer RNAs per cell. Now, this can either indicate that most of the RNAs in the cell are not physiologically significant, or it could indicate that all it takes is a small burst of one or a few molecules of RNA to produce an even larger burst of proteins and even larger bursts of activities of those proteins. So you get this amplification.

And so the stochastics that we will study in the systems biology part of this becomes a more significant consideration. So looking at one molecule per cell, it's important to start thinking about what the implications of that might be for the systems biology, and asking, can we accurately measure it down there, and do we believe that it's biologically significant.

Now, there's a whole variety of microarray data analyses, ranging from the very hardware-oriented first data acquisition modules all the way up through analyzing single array data at a statistical level, to multiple related experiments, such as the one we showed in the previous slide, all the way up to clustering multiple examples from multiple different conditions to start asking the biological questions about why RNAs go up and down together.

For intermediate analyses, where we'll be talking today, as introductory issues of data analysis, I'll illustrate dChip and a couple of other tools that indicate how reproducible experiments can be, and the kind of systematic errors that can creep in. The reproducibility helps you by repeating, helps you reduce the random errors.

And here are four papers recently that talk about measurements from multiple measures from the same experiment, or multiple measures by using two completely different microarray technologies. And I urge you to take a look at these. When we compare two distributions from microarray experiments, you can think of these. Even if they're not perfectly normal distributions, they're going to be roughly bell-shaped shaped curves.

So let's say that this is experiment 1 and this is experiment 2. You say, oh, they look the same. This is experiment 1 under condition 1. This is under condition 2. OK, now they look different. But how do you quantitate that? And the way you ask that is, the means of those two roughly bell-shaped distributions are far apart from one other. How far apart?

Well, they're farther apart from one another than the width of the distributions individually. And the distance between them, you can think of as the mean of the difference of the distributions. And then the width is a measure of the root mean square standard deviation, so [? versus ?] the combined width of the two. If one of them is wide and the other one is narrow, you have to have some way of combining those.

So that's sometimes called a student t-test. And the t statistic itself is simply the mean over the standard deviation. In other words, how many standard deviation widths apart are these two means? Or if you take the mean of the difference, take the distribution of the difference, then you want your null hypothesis. H0 here on slide 33 is the null hypothesis. If the mean value of the difference is 0, there's no difference between the two distributions.

If you can rule that out, then that would be the point of this test. So you can think, how many widths apart are the means of these two distributions. Now, this requires that actually, the distributions be very close to normal, not distinguishable from normal distribution with all its properties.

If you are in serious doubt or can prove that they're not normal, then you should go to a non-parametric. Normal means it's parametric. It has a mean and standard deviation that characterize it well. Then you can use a non-parametric. Whenever you see the word "ranks," that's a tip-off that you're going into something where you're making fewer assumptions. This has lower power. That means you might miss some significant differences. But on the other hand, if you can convince yourself with the Wilcoxon matched pair sign ranks test, then you don't need to worry about whether it's normally distributed.

In any case, we're going to look at some distributions and ask informally whether these are the same distribution or different. Yes.

AUDIENCE: [INAUDIBLE]

GEORGE CHURCH: So the question is, how do you deal with the multiple hypothesis testing. And this basically is exactly the same answer that we would have given in the last lecture on multiple hypothesis testing in genotyping. If you apply exactly the same-- it's a very good question, very appropriate here.

Just as before, where you would have multiple different phenotype-genotype combinations that you might want to test, essentially testing every possible single nucleotide polymorphism or combination in the genome, to a first approximation, whatever your significance is, it needs to be that much more significant if you have that many hypotheses.

AUDIENCE: [INAUDIBLE]

GEORGE CHURCH: You either have to improve your data. Allows you to test more hypotheses. Or you need to reduce the hypothesis, that number at the outset, by having a sharp biological question at the beginning. It's an excellent question, but there's no magic wand except those two that I know of.

OK, so here's some examples of independent experiments. Now, when someone says an independent experiment, you have to be clear about, is it that the same RNA sample split and then labeled independently. That's really not an independent experiment. On the other hand, you could take two completely, where you repeated the best of an independent experiment. If your objective is to ask how reproducible is the entire biological phenomenon, you should go back as early as possible, make a new cell line, try to get the conditions exactly the same, but completely independently executed, possibly by different researchers in different laboratories.

In that extreme, you expect to have more scatter. Here, these are the regression lines. The R squared is the number that pops out as an indication of deviation from the linear, just like the linear correlation coefficient, which is basically a squared term. You can see that as instead of splitting one sample, and doing kind of a trivial differentiable labeling, if you have more independent samples, you get more scatter and a lower figure of merit for the regression line.

OK, now, what are the guidelines for-- what are some of the considerations in RNA quantitation? I think we've touched upon this before, but I just want to drive it home, that some people will say, I'm only going to look at things that are more than a three-fold effect. This is sort of the ratio limits that you might perceive in the early RNA [? TIP ?] experiments.

But I think we're getting better at it and the biological motivation is high. We've seen that human trisomies, where you just have a 1.5-fold increase in dosage, every single one of them has a huge phenotypic consequence. Many of them result in lethality. We should set as a goal to be able to monitor most of the RNAs of biological significance down to this 1.5-fold effect, which can have these dramatic implications.

We mentioned the Oligonucleotides, we might be able to get more of them per gene. How cam we utilize this, not only the number that we can get, but the specificity? If you have a gene length oligonucleotide, or cDNA, then you're going to pick up not only the gene of interest, but every related gene, all the alternative splice forms, all the very, very close family members.

So with oligonucleotides, you can then go and target individual splice forms, but then when you apply your algorithms, you have to be careful not to lump them all together as if it's one gene. You have to say, OK, this is splice form number one, number two. And just having oligonucleotides aimed at particular exons is not sufficient to tell you which exons insist in particular RNAs.

You can have present in the population exons 1, 2, 4, 6 12, and so forth. But you don't know whether 1 and 12 are on the same molecule, however. That requires a more specialized method, possibly high-throughput method. There is another set of economic forces pushing towards just doing a subset of the genome.

Just like not repeating the experiment, you probably don't want to give in to economic forces unless you absolutely have to, because if one person studies a cancer subset, another studies a blood-related subset, and another one studies these little pieces of the genome, then when they want to pool their data in order to ask questions about what genes are cluster together because they're in their proliferative cells, and which ones cluster together because they're in this developmental stage or another, they can't do it because they don't share enough genes on their arrays to do this meta analysis.

So that's a consideration when you're in the experimental design phase. And hopefully, computational biologists are involved not only in the interpretation of the data, but in the design of the experiments as well. Here's yet another way of looking at the variation that you have in the experiment. We're introducing, I think, the coefficient of variation here, which is simply the standard deviation normalized to the mean.

So you can just phrase it. It's a way of sharing, in a generic sense, how much variation you have. So you can say the coefficient of variation is, say, 10%. And that's independent of what units you're measuring. And so we have is on the horizontal axis, the x-axis here, number of messenger RNAs per cell, and in the vertical axis, the coefficient of variation.

And you can see that when you get up above, say, 20% coefficient of variation, you getting less trustworthy, because here, we've used the algorithms that are built in to the [INAUDIBLE] software for asking whether it thinks an RNA is present or not if the intensity is very low, and a variety of other criteria.

For a single experiment, it will classify whether it thinks the RNA is present or not. But if you use a large number of different experiments-- each of these dots being a different RNA-- you use a large number of experiments, you can now beat the company software, because it's made the assumption that you're just doing one experiment.

And so here in the dark blue are examples where in 3 experiments, all three experiments was called f one by one. But you can see that even with cases where it's not called present in any of the three experiments, these magenta ones, you can still find very high reproducibility, that is to say, very low coefficient of variation, down around 10%. There's some pink dots all in this region around 10%, and these are just as reliable as the blue dots. Even though they're not called present by the software, collectively, they're very reproducible, and therefore, they're trustworthy.

So actually, reproducing your experiment is not just something you do to appease nature and lower your statistical noise. It actually allows you to get data for RNAs that otherwise might be inaccessible. So there's immediate gratification there, even at the slight expense.

So now let's broaden back out a little bit on a number of different methods and their advantages and disadvantages. Each one has a set of advantages. We've already talked about two of them, which is the immobilized genes, labeled RNA scenario. That's basically the microarrays or chips. And the advantage here is that in a very high throughput manner, you can manufacture large numbers of these. And you can get high multiplicity, all the RNAs that we know of monitored simultaneously. The in-situ, we've also talked about. The major advantage is retaining the spatial relationships.

Some of these other methods, if instead of immobilizing the probes on a solid surface, you immobilize the RNAs, and then one by one, you label the probes, this will allow you to first, say, separate the entire transcriptome of RNAs in electrophoretic separation.

And so in a highly parallel method, you've now immobilized them after they've been separated by size. So if you want to know the size of the RNA, which is a big hint as to its exon composition and so on, measuring the size of RNA, this is one of the few ways to do it. Very hard to do with arrays or in situ. OK.

If, on the other hand, you want sensitivity, where you want to really detect at the noise level, which say, for mammalian veterbrate RNAs is around 10 to the minus 4 copies per cell, that's the level at which if you look for almost any part of the genome, any kind of RNA, even things that shouldn't be expressed, you will find them down in the minus 4 [? per cell. ?] That probably is not biologically significant, but it's a biological fact.

Getting down to that level, or if you have a mixed tissue and you want to detect 1 part in 10 to the 10th, you might have 5 times 10 to the 5th messenger RNAs per cell. But if you have 10 to the 5th cells, then a single copy messenger RNA would be down to the minus 10. That's feasible with reverse transcriptase, quantitative reverse transcriptase, [INAUDIBLE]. And it has it is the standard which all the others can barely match.

Reporter constructs are something we do not consider generally a high throughput method, although there are genomic constructs of reporter constructs for an entire genome like yeast. But here, the real advantage of that method is there's no worry of cross-hybridization. With in-situs, with northerns, with arrays, there is a chance that if you probe for RNA x, it will happen to hybridize, especially if it's present in high abundance, it will happen to hybridize to one or the other ones.

But with a reporter construct, we will take a fluorescent protein array, a luminescent protein and hook it up to the gene, and insist with the gene you're interested in. And that will directly or indirectly monitor the expression of your favorite gene. That has no possibility for cross-hybridization. We've talked about the advantages of counting. The disadvantage is of course cost. It allows you to do gene discovery. It doesn't address alternative splicing.

Here's an example of comparing two of those methods. As microarrays are being introduced, one needs to validate them to ask whether you're measuring one RNA or multiple RNAs of different sizes to ask whether quantifying a northern blot correlates with quantitating an array. And here, you can see a fairly acceptable linear relationship between the two quantitative measures. And this has been played out many times.

The opportunity that you have when you make an array, we said that the SAGE and NPSS allow you to do discovery. But another way of doing it is putting down lots of oligonucleotides, even oligonucleotides in regions where your genome annotation may not have indicated that there's a gene. So you can see here the bottom 60% of this array was in so-called non-protein-coding regions.

And you can just see what you get when you do that. It doesn't cost that much more to put down some of these non-coding regions. And you can ask in these untranslated regions whether there are maybe antisense RNAs that will overlap in the translated regions. Or you can look for DNA protein interactions in certain kinds of experiments. And you can look for RNA fine structure. Where does the gene actually end?

You may annotate that the RNA ends here, but you need ways of actually measuring them. So there's a lot of uses for nucleic acid probes in so-called non-protein-coding regions, which can range from 12% of the genome in simple prokaryotes to up to 98% of the genome in humans. So what are the sources of random systematic errors?

We have a secondary structure that we talked about at the beginning of this lecture. It can cause different parts of the array to have different hybridization efficiencies. The position on our array to have an effect, for example, poor mixing, if you're making your array by a non-reproducible method, the amount of target nucleic acid immobilized on the array can vary. And you need to control for that, for example by having an internal standard. Cross-hybridization hybridization, we've talked about. The unanticipated transcripts, you can handle by tiling, by basically putting oligonucleotides throughout the genome.

So here's an example of spatial effects. What you do is spike in known amounts of known RNAs which are present throughout the array. And so these are internally spiked in addition to your unknown fluorescently labeled arm probes. And you can ask whether you're getting a perfectly uniform, edge-to-edge hybridization with the known for the answer.

And if you're getting peaks and troughs, then you can use these internal standards to calibrate that particular hybridization experiment and correct for this kind of systematic error. This could occur again and again. Here's two different experiments giving roughly similar edge effects.

You need to account for these things to avoid that particular source of systematic errors, especially if you put all of your oligonucleotides for a particular gene near one another. A better strategy for statistical experimental design is to put your oligos randomly throughout the array.

Here's another one, unanticipated RNAs. Two examples, one an open reading frame of unknown function. You can sometimes mis-annotate. If you have two open reading frames on opposite strands, it could be one is used, generally speaking. One is used and one isn't. And you could pick the wrong one. You might pick the big one, and it could be the little one is the one that's actually used. And that's what happened in this case.

And another one is-- so that was a translated RNA. We just happened to pick the wrong strand. Here is an untranslated RNA, such as the snow RNAs we saw before. This one is an untranslated RNA that was discovered in a so-called intergenic region.

If you have a statistical test for the goodness, the quality of an individual oligonucleotide hybridization, based on, say, its reproducibility or its relative intensity that you expect-- if you have 20 different oligonucleotides all for one gene, and you expect number 1 typically is stronger than number 2, and then you find a case where number 1 is weaker than number 2, then you can flag that.

You can say, I don't believe that particular spot. And if you color-code them all-- see, here is white spots, things that don't fit your statistical model for the array hybridization. This is the advantage of having a statistical model of the entire process. Then you can mark those as white, and you can look to see whether they have a [? statistically ?] significant spatial distribution, which they do in this case. They all seem to be clumping at this corner.

Now, what could cause that? Well, we've already illustrated that there are ways that you can use internal standards to calibrate. This was not a case where we had poor hybridization efficiency or strong hybridization efficiency around the edges. This was something where the alignment of the grid is done by these little squares along the edge, and the computer algorithm that finds these spots was distracted by this little spot off the side, which is not part of the checkerboard.

And once you manually correct that error, now you snap in. On the right-hand side of slide 43 is now the statistical model of this after getting the alignment right. Here, you had been associating the wrong oligos with a particular signal. It didn't fit the model. Now it fits the model, and you see the little scattered strips of gray where you have individual genes which are misbehaving, rather than the entire corner of the array.

OK, so now we get the very interesting interpretation issues where we're using the same kind of information. Once you have a model, a very sophisticated model of how the individual oligos in the array behave, here, what we do is we take genomic DNA as an example of a fairly equimolar calibration standard.

If you take genomic DNA and label it, you expect every segment in the genome to be present at the same molarity, with the exception of repetitive elements, which we'll put aside for the moment. And so that means that any place, that is, any oligonucleotide that doesn't hybridize with the genomic DNA, such as these ones that go close to baseline here at 0-- remember this is perfect mass versus [? mismatch ?] that we're plotting on this.

When you get close to 0 for the genomic DNA in black, that means that really doesn't hybridize well. It's not that it's missing from the genome. It's that it has some secondary structure. So this is the secondary structure that's been a theme for this plot.

And that sort of secondary structure is actually a piece of data that you can do data mining on. You can go through the entire genome and you can look for secondary structures. And you can ask for those. Secondary structures depend on what part of the genome is transcribed. Now, here is a messenger RNA. This is one of the few messenger RNAs for which you have a plausible secondary structure. Most secondary structures are on structural RNAs or enzyme-related RNAs, the ribosomal RNAs, [INAUDIBLE] RNAs, and so forth.

This is the messenger RNA for this gene product, LTT. And if you look where this black arrow is coming from the right-hand 0 you'll see a long helix. And that helix is at the 3' end of the messenger RNA, and it's very well characterized both structurally and functionally. And it's known to be involved in at least one important biological process, which is the termination of transcription.

When you get close to the end of the RNA, that [? pair ?] can reform, and it sends a signal to the transcription apparatus to stop. So that is a believable hairpin of known function. And the interesting feature of this microarray is that's one of the places where both the genomic DNA in black and two completely different RNA samples fail to hybridize, consistent with it being a very strong hairpin with a dozen G, C and A base pairs.

Another thing you can derive from this detailed model of the array, here, you have 60 different oligonucleotides along the gene and adjacent intergenic regions. The question is, where does the RNA transcription stop. Well, if you look in the places where the DNA control is high, you'll find the RNA is high, going from right to left.

The RNA tracks the DNA. The red and blue tracks the black until you get to position -33. And there, the red and blue drop to baseline, and the Black stays going up and down at a higher level. And that happens to coincide with the known transcriptional start. And so that would be another way of mapping the transcriptional start.

You'll notice that some of the hybridization intentially drops below 0. This is just an artifact of having the perfect match minus the mismatch. If it happens to be the case that your mismatch control is cross-reacting with some other DNA, say repetitive DNA in the genome or RNA, then it can get actually more intense than the perfect match. And so you can get a negative value. But otherwise, the negative intensity would be meaningless.

Now, splice domains. In principle, you can go through the whole human genome and you predict where all the exons are, where all the splice junctions are, and in principle, even all the alternative splicing. In practice, is not that easy. And you can use all the hidden Markov models and so forth we've been developing. You can do multi-sequence alignments to get these motifs here, where two bits is a full scale. And you can find donors and acceptors in this kind of pattern, GT donor, AG acceptor.

But when you come right down to it, you want to have some way of going through this empirically as well. And so what you can do is you can basically ignore or look sort of independently, do a tiling of the genome with oligonucleotides, as was done here by Shoemaker et al.

And this was, I think, one of the nicer papers that came out in the*Nature* issue on the human genome [INAUDIBLE] sequence. Here, as the sequence was coming out, chromosome 22 was one of the first chromosomes nearly completed. At the top of slide 47, you see how the metaphase chromosome is banded and labeled. If you take a little [? 113-kit ?] kilobase chunk of that, it's the next line down. Then you blow that up further, and all the way down to oligonucleotide 60-mers, tiled every 10 base pairs as a starting point all the way along this 100-kilobit chromosome 22.

And then you hybridize it with RNAs from a variety of different human tissues. Then you ask, in the vertical axis, what is the log of a normalized signal intensity for these various RNAs. And you'll get a little histogram here, where purple spike means there's a lot of hybridization under at least some of the conditions. And then there'll be a zone where there's almost no hybridization.

And that's because those introns that we had, that we showed in the previous slide, are spliced out, and they're in low abundance. They're displaced out of the nucleus before the RNAs accumulate, so they tend to be in low abundance. And they're not found in the mature messenger RNA. And so when you label these up, you're selectively labeling the exons in [? CNA. ?]

And if you can see, they coincide well with the little green exons in the annotation, except every now and then, you'll find something-- here's a case for exon 3 where the green annotation in the original sequence is too short. And here's a blow-up near the bottom where the purple region clearly extends beyond the green annotation in the [INAUDIBLE] from sequence algorithm, sequence analysis algorithm, where 102 base pairs should be extended five times to that exon to make it a slightly larger exon.

But when you extend it by that, you ask, well, does it still have the splice site, or does it have a new splice site that we can recognize. And sure enough, it does. It has an AG and a fairly good match to the motif we had in the previous slide. And you can see that the purple intensity drops close to 0 here as soon as you get out of the exon as now properly defined. So this is a way of including additional data in addition to the sequence by tiling and by quantitative hybridization.

Now, the last topic today is time series. This connects the quantitative data that we're collecting, where you're not just collecting an isolated condition and comparing it some other condition. It actually matters the order of the different conditions that you have. And this is a great advantage in analyzing causality, and we'll illustrate it in the context of messenger RNA decay, and finally in ways of aligning different time series data.

Now, why do we want time [? courses? ?] If we do a gene knockout or we do a gene deletion, by the time you isolate that mutant and characterize and do the RNA, you've now gotten not just the primary effects, but all the downstream effects of that knockout. So the best would be to have some kind of conditional control of the transcription, so that when you first either turn it on or turn it off, the first events that occur are likely to be primary events.

Now, the way that you control that needs to be, not have too many perturbing forces on the whole system. So temperature shift, it's an easy class of mutation to get, but it's not suitable because there's a huge temperature effect on the entire system. Chemical knockouts can be more specific, but you need to prove that.

An example of a fairly time-honored chemical knockout is rifampicin, which fairly specifically affects just the RNA polymerase. And so this is an interesting case, where the effect is to stop initiation of transcription. And so then, as we do our time series, what we see is the RNAs for LPP, which we showed a few slides ago, is very stable.

It basically lasts longer than the lifetime, the doubling time of the cell, possibly many cell generations. And other RNAs, such as CSPE, have extremely short half-lives in the order of 2 and 1/2 minutes. And you can compare various methods for quantitation. Then you've come up with different half-lives here.

OK, so that's an example of a very significant class of chemically manipulated knockouts. So you can precisely phase them. You have very few other consequences, and then you can measure a time series. It'd be nice to be able to do that for any particular RNA and see what the downstream consequences are.

Now, whenever you do a perturbation where you have two time series, you want to know how all the RNAs occurred during heat shock or some other pulse of some chemical relative to pulse of a different chemical, or the time series as it would have occurred without any. You can see how they won't necessarily align up point by point. You can't just start them at time 0 and expect them all to line up.

In fact, you can't even expect them necessarily to line up where you have a uniform stretch. You might have to have piecewise stretch, where certain parts go faster than others. Now, this may hopefully click in your mind a connection to the dynamic programming, where we had two sequences of bases or amino acids. And you wanted to expand or contract different sections of those by inserting a placeholder.

Well, it doesn't make quite as much sense here with time series to insert a placeholder. So you can do that. You can have a discrete block diagram. Just this, here's series A and B in middle upper diagram. Or you can have a more continuous function, where you've tried to more smoothly warp.

Both of these are dynamic programming algorithms. The smooth warping is slightly more a little more complicated. The insertion deletion one is exactly the same three conditions that we went through for pairwise alignments in dynamic programming. But this is partly to drive home how many different ways you can use dynamic programming. You can use it in HMMs, which is that multi-sequence alignment, and now for time series and gene expression.

And you can see, from the literature on cell cycle, almost all of the data time series that we have so far actually don't align perfectly point by point, because you use wildly physiologically different conditions to get cells to synchronize, say, for cell division, or to start an event here using a mating pheromone, a small peptide that's released in the media that kind of controls the cell cycle and allows you to arrest and then release from arrest, or a temperature-sensitive mutant, even though I malign temperature-sensitive in just a moment. It is one of the most precise ways of getting synchrony of cell division.

Cell division is a particularly good illustrative notion, partly because we mentioned it earlier in the course. But also, if you think of any dividing set of cells, many of the cell types that you'd be interested are dividing-- stem cells, microbial cells and so on. That automatically is a mixture of cells. If you mush them up and extract RNA, you're kidding yourself if you think this was a homogeneous population.

If on the other hand, you synchronize the cells, then you've removed one major variable that could confound. Now they are much more homogeneous cells that are in the same state, and the cell cycle can be synchronously isolated as a population. There may be other sources of heterogeneity, but you've eliminated a big one. In any case, you take these two data series. They have different time constants, different lengths, and even different warps.

Now, you want to take the x's and superimpose them on the o's. And here's an example of that now. They're both put together, and even though there may be little deviations for any particular gene, when you talk about the thousands of different genes, very rich pattern. Lots of information, plenty of opportunity for smoothing out individual variations. But here, you get superimposed patterns. And here's the traceback that tells you exactly where the insertions and deletions or smooth warping might occur to align these two different cell cycles' data sets.

So in summary, we've connected the multi-sequence alignments from last class to allow you to model RNA structure, how RNA structure helps you model it. An interesting class of RNA guide sequences involved in methylation as an illustration of finding genes that don't encode proteins.

And we talked about various quantitation methods, errors that present and solutions of errors, statistical methods for asking whether two distributions are related or have no difference in their means, interpretation errors about where RNAs start and stop, how you get alternative splicing, and finally time series data, which we will find very useful for connecting RNA and protein measures over time series for analyzing causality and systems biologies. OK, thank you very much. See you next time. Be sure to get your problem sets in to your teaching fellows.