Interactomics Basics and Applications

Prof. Sanjeeva Srivastava
Dr. Joshua LaBaer
Department of Biosciences and Bioengineering
Arizona State University, USA
Indian Institute of Technology, Bombay

Lecture - 06
NAPPA Technology and Protein Arrays-I

It is my great pleasure to introduce distinguished scientist Prof. Joshua LaBaer. Today and next few lectures will be delivered by Prof. Joshua LaBaer. He is the Executive Director of Biodesign Institute at Arizona State University and the Director of Virginia G Piper Biodesign Center for personalized diagnostics. Dr. LaBaer has been one of the foremost investigators in the rapidly evolving field of personalized diagnostics.

Dr. Joshua LaBaer has been instrumental in development of cell free expression based protein microarray platforms. One of the main contribution of his group has been development of Nucleic Acid Programmable Protein Arrays or NAPPA technology. Dr. LaBaer is particularly interested in advancing biomarker discovery based programs in particular to find out biomarkers for early detection of cancer and autoimmune disorders using protein microarrays.

He has built a fully sequenced verified clone sets for model organisms and pathogen genes which is one of the huge contribution for the whole society, and very important reagent resource for the researchers who want to perform high throughput biology. Dr. LaBaer is the principal investigator on a 36 million dollar contract to develop a blood based diagnostics that predicts absorb radiation dose received after a radiation event 1 to 7 days after exposure which is sponsored by Biomedical Advanced Research and Development Authority.

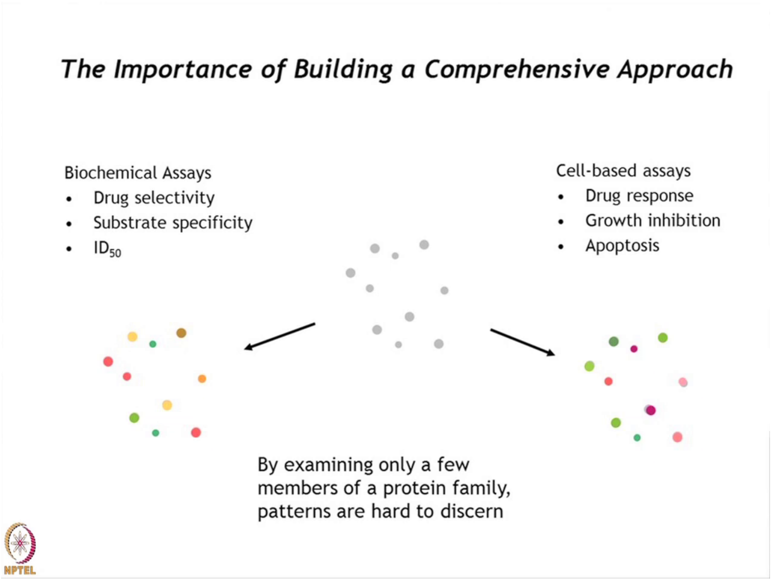
He is also the past president of US. Lupo. And one of the convenience of last year conducted Human Proteome Organization World Congress in Orlando. Dr. LaBaer is going to talk about biomarker discovery based program various considerations for in statistical tools which are required for biomarker evaluations and validation, and how to make NAPPA arrays using very simple lab based resources. Then perform auto antibody based screening for different cancers especially breast cancer and how to also utilize the protein microarray based platforms for functional studies especially the PTM based analysis.

In today’s lecture Prof. Joshua LaBaer will talk to you about the basics of proteomics, its significance for high throughput gene cloning experiments, and what are the steps required for gene cloning and generating clones which could be used for high throughput experiments even later on.

So, the kind of resources and reagents which you can generate using the novel cloning technologies then later on you can simply transfer the genes of interest into any vectors for your given experiment. I am sure Dr. LaBaer will introduce you not only the concepts of proteomics, but also the details about how to generate these high quality reagents which could be useful for your research. So, let us welcome Dr. Joshua LaBaer for his lecture.

Hi, I think we are ready to get started, yeah, all right. So, I am going to start a little bit at the beginning. We have we have several lectures here to cover on terms of the NAPPA technology, and so I thought it would be useful to sort of begin where we begin.

(Refer Slide Time: 04:38)



The Importance of Building a Comprehensive Approach

Biochemical Assays

- Drug selectivity
- Substrate specificity
- ID_{50}

Cell-based assays

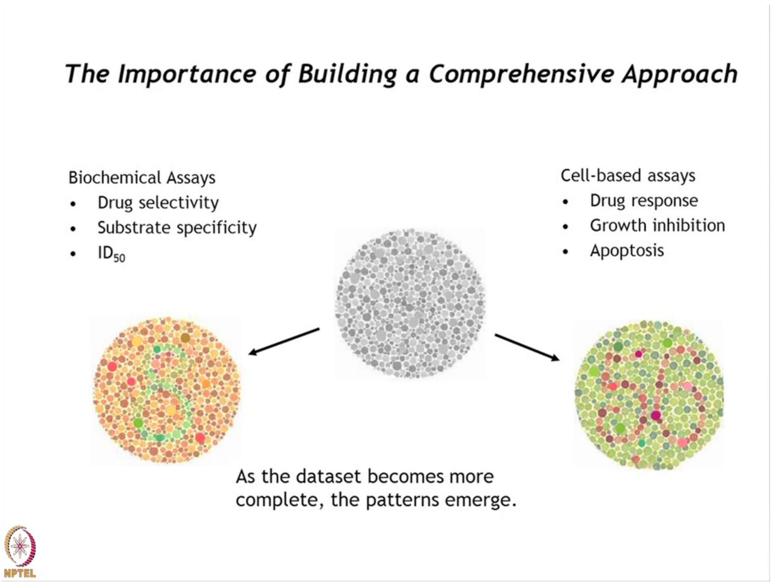
- Drug response
- Growth inhibition
- Apoptosis

By examining only a few members of a protein family, patterns are hard to discern

So, this is where biology was 10-15 years ago. What I mean by that is that we were studying proteins a few at a time you know maybe three or four or five proteins at a time and that is you know that is how much information we were getting, but what we were really trying to understand was the entire proteome, and we would take these proteins and we would do a certain set of assays on them maybe we look at drug selectivity, we might look at what the substrates were, we might do a variety of biochemical assays or we might do sort of cell-based assays on those proteins.

We would test them for a variety of features, and each one would get a different color sort of attached to it. But what we really wanted if you look at only a few things at a time, you cannot really get a full picture of what is there right.

(Refer Slide Time: 05:31)



The Importance of Building a Comprehensive Approach

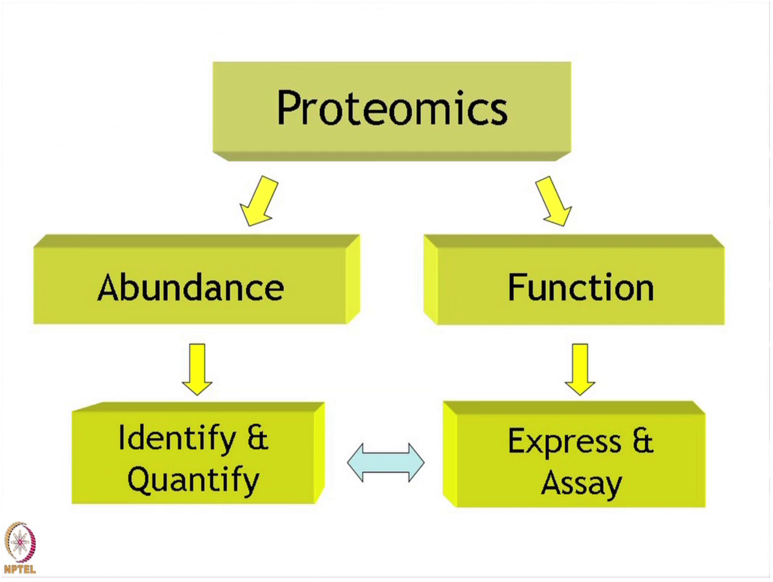
Biochemical Assays
- Drug selectivity
- Substrate specificity
- ID_{50}

Cell-based assays
- Drug response
- Growth inhibition
- Apoptosis

As the dataset becomes more complete, the patterns emerge.

If you look at this you know you do not know what that color means you look at that you do not get know what that color means. What you really want to do is everything, because when you do everything then you get to see the whole picture. You really understand what it is you are trying to look at, and what it means and that is really where proteomics comes in. Proteomics is the idea of not sending one or a few proteins at a time, but studying all of them trying to get a comprehensive study of everything.

(Refer Slide Time: 05:53)


Proteomics

Abundance

Function

Identify & Quantify

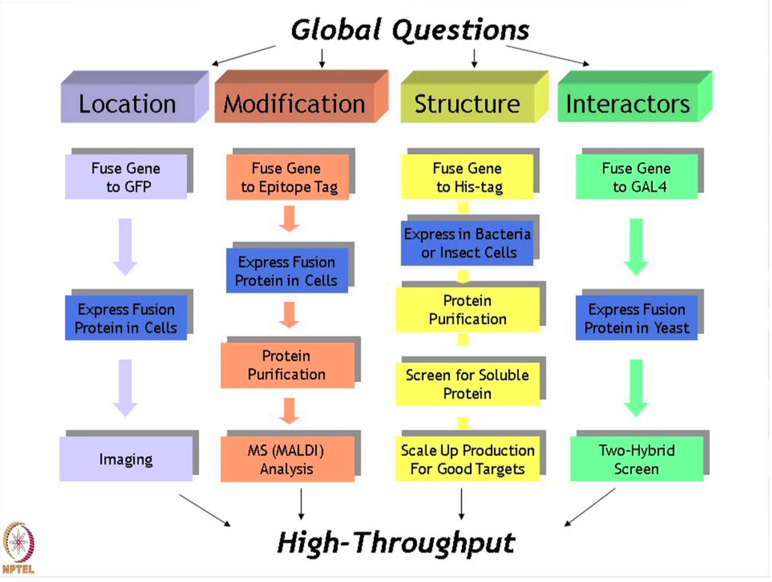
Express & Assay

So, there are two general approaches to proteomics. One approach here is looking at the abundance of specific proteins, how much protein is present. And what you typically do with the abundance approach is you compare the proteins in the disease to the proteins in the normal in the normal tissue, and you ask are there proteins that are changed in the context of disease relative to normal.

And then the hope would be that if you do this over and over again, you will identify which proteins are altered in disease and that will provide useful information about what is causing, what is causing illness. The typically this approach requires mass spectrometry or some type of technology that can measure the levels of proteins in a sample.

The other approach and the one that I will talk about today is what I call a function-based approach. And the goal here is to look at the individual proteins and ask what do they do, what is their role, how do they behave, who do they interact with, you know are they altered in disease. And obviously, these two approaches are complementary right, they support each other. So, what are the ways that we can look at the function of proteins right.

(Refer Slide Time: 07:07)



Global Questions

Location | Modification | Structure | Interactors

Fuse Gene to GFP → Express Fusion Protein in Cells → Imaging

Fuse Gene to Epitope Tag → Express Fusion Protein in Cells → Protein Purification → MS (MALDI) Analysis

Fuse Gene to His-tag → Express in Bacteria or Insect Cells → Protein Purification → Screen for Soluble Protein → Scale Up Production For Good Targets

Fuse Gene to GAL4 → Express Fusion Protein in Yeast → Two-Hybrid Screen

High-Throughput

So, here are a few of them. You can look at where proteins localized in cells or in the body, and that may tell you something about the role of that protein. You can look at how that protein is modified is it phosphorylated, is it as stellated, is you know is it ubiquitinated modifications or proteins tell you something about what they do. You can look at the structure of the protein.

So, what is its three-dimensional folding, how does it, how does, what shape does it take that will give you a clue about what its role is. And you can look at which other proteins that protein interacts with right. This topic that we are here today to talk about is interactomics.

So, who does, who do proteins interact with, who do they come in contact with that tells you something about what they do.

So, how do you do, how do you do those various studies? Well, if you want to look at the location of a protein, you might tag that protein with a fluorescent marker like the GFP put it in cells and ask where does it localized. If you want to look at its modification, you might purify the protein using an epitope tag and look at it under mass spectrometry, and ask what modifications can I observe on that tagged protein. If you want to look at the structure, you might purify the protein. And then after you purify the protein, you would crystallize it and you would do three-dimensional structures using X-ray crystallography.

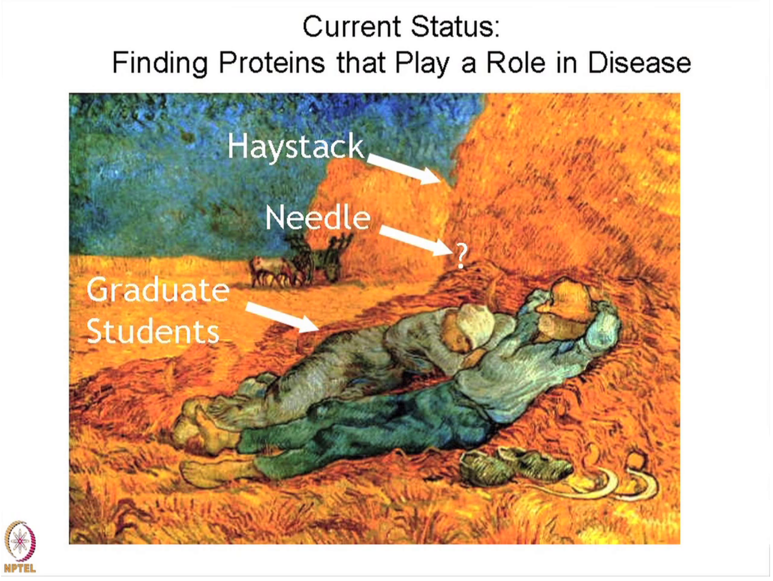
And, if you wanted to look at the interactors of that protein at least using traditional methods, you might tag that protein and then do like a yeast two hybrid assay or some kind of pull down assay to look at what proteins are attached to the protein that you are looking at right. And then yeah the goal of course is to do this in high throughput. What you want to do is look at these studies a thousand proteins at a time all right.

So, we looked at this kind of method when we began our work number of years ago. And what one of the first things we observed was that there are some things that all of these methods have in common. First of all you have to be able to make proteins. You have to be able to express them in some circumstance. Sometimes, it is in cells; sometimes it is in cell, in a cell free extract sometimes you are making it in vivo and the normal search circumstance; in other cases you are using a heterologous system alright.

The other thing that they all had is that to do things in high throughput. To study proteins in high throughput you most often need to put a tag on the protein. You all know what I mean by a tag an epitope tag, a chimeric tag. If you try to purify all proteins by their very biochemical nature, it is very cumbersome and you cannot do that times thousands. And the goal here is to be able to study proteins, hundreds of the many times or thousands of the many time.

And so the easiest way to do that is to put a GFP tag on them, a GST tag on them, a HIS tag on them some kind of tag that will allow you to have a biochemical hook to study the to study all the proteins in the same way alright.

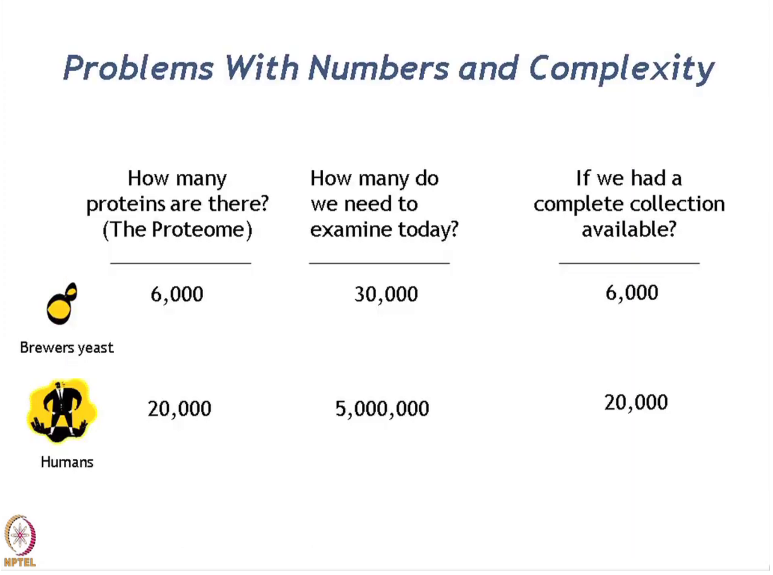
(Refer Slide Time: 10:26)



Current Status:
Finding Proteins that Play a Role in Disease

And when we began this work this was, what, what the field look like right. So, what am I looking at what we are looking at a couple of graduate students who are exhausted. So, now, why are they exhausted? Well, they have been looking through those haystacks for the needle they are trying to find, and it takes a long time to sift through the hay to find the needle. So, can we can we can we find a better way, is there is there a faster technology?

(Refer Slide Time: 10:47)



Problems With Numbers and Complexity

	How many proteins are there? (The Proteome)	How many do we need to examine today?	If we had a complete collection available?
Brewers yeast	6,000	30,000	6,000
Humans	20,000	5,000,000	20,000

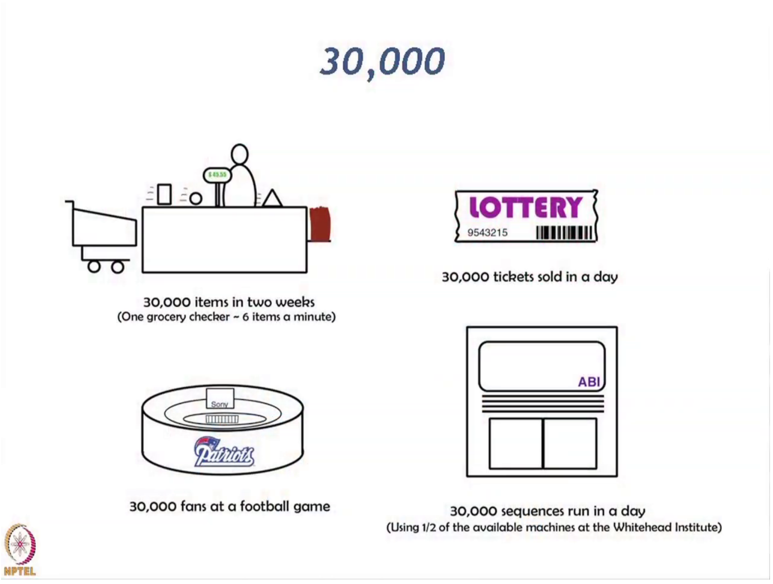
So, when you know if you think about a simple organism like yeast like (Refer Time: 10:55) here VCA, there is around 6000 unique proteins in yeast. So, if you were to do high throughput screening using CDNA libraries or phage display or something like that, you could look at around 30,000 different samples, and you would pretty much have sampled everything that would be you know a fivefold redundancy right.

You would look at everything five times to make sure that you with a Poisson distribution you would get everything. Of course, the simplest method would be to have a cloned gene for every gene in yeast, and then test it once and only once, and then you would do 6000 assays and that would be very easy right.

So, the same thing would be true for in the case of humans, it gets more complicated. So, we now know that there are roughly 20,000 give or take a few protein, unique protein species in humans. Obviously, once you start taking care splice variants, and posttranslational modification that number expands dramatically. But let us just say for the sake of simple simplicity, if we took each unique gene and tested it once and only once, there would be 20,000.

But you cannot if you do not have cloned copies of those genes, if you have them in libraries like CDNA libraries or phage display libraries, you cannot if you want to test all proteins in order to get past all the redundancy, you would have to do 5 million assays and that is just too many. Ideally what you want is a cloned collection of all of the genes in the human each one a perfect copy, so that you could test every gene once only once, and then you would be doing roughly 20,000 assays.

(Refer Slide Time: 12:47)



30,000

30,000 items in two weeks
(One grocery checker ~ 6 items a minute)

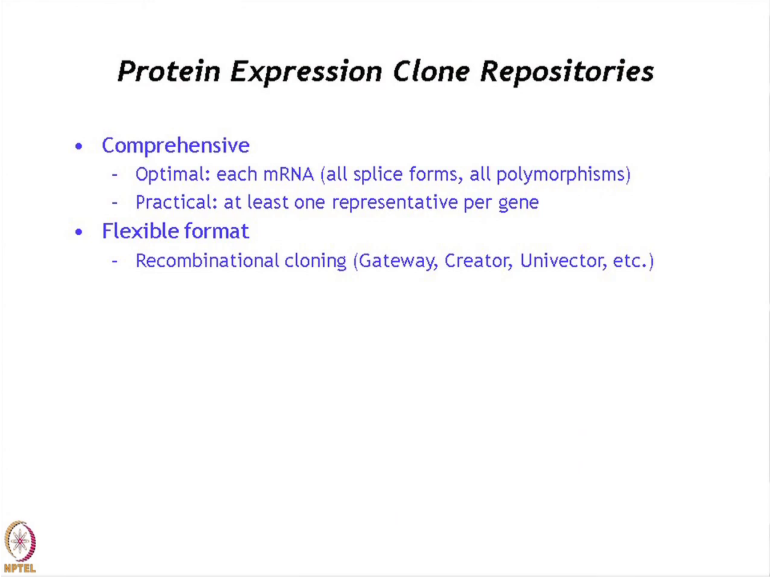
30,000 tickets sold in a day

30,000 fans at a football game

30,000 sequences run in a day
(Using 1/2 of the available machines at the Whitehead Institute)

So, I 20,000 30,000 assays that is a number that I can imagine doing in a in a high throughput biochemical setting. In a supermarket in the united states if you look at around 6 items a minute when you are passing them that you could get that done in 2 weeks right. They sell 30,000 tickets for lottery in a single day in the State of Massachusetts. So, 30,000 is a number that we could imagine, we could do that right and so that is that was the goal.

(Refer Slide Time: 13:11)



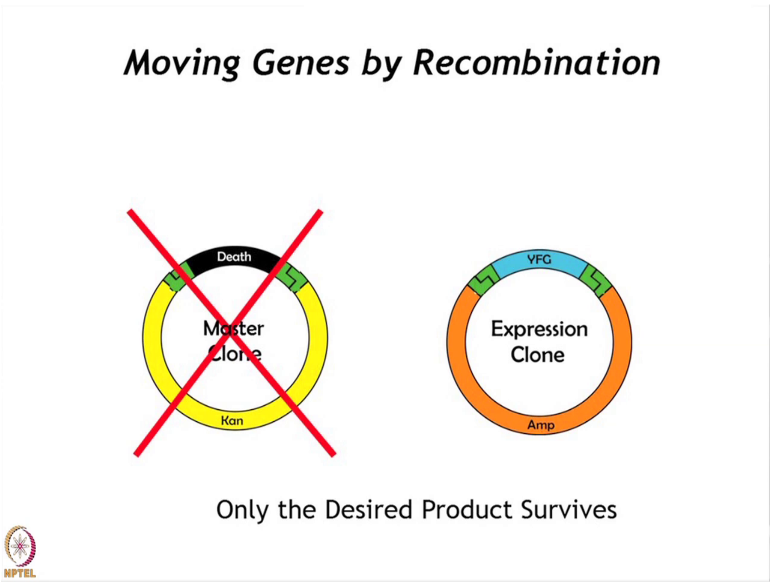
Protein Expression Clone Repositories

- Comprehensive
  - Optimal: each mRNA (all splice forms, all polymorphisms)
  - Practical: at least one representative per gene
- Flexible format
  - Recombinational cloning (Gateway, Creator, Univector, etc.)

And so our first goal in my laboratory was to build a repository of cloned copies of all human genes. So, obviously, I am trying to get you to protein microarrays, but before we can get to protein microarrays we have to talk about where the genes come from to make those arrays, how are you going to make all those proteins if you do not have the cloned copies of genes. So, the first thing we wanted was to get a comprehensive collection. We wanted at least one copy of every gene. Of course, in the perfect world we would have one copy of every splice form of every gene, but at the very beginning let us at least get one representative of each gene.

The second thing we wanted was a flexible format we recognized that different users might have different applications for these genes. And so some of them would need to make the proteins in cells as we talked about earlier, some of them would make them in vitro, some of them would make them in the natural cell setting, some of them would be in that in a header oolagah cell setting. So, you had to you had to have a format that was flexible.

(Refer Slide Time: 14:17)



Moving Genes by Recombination

Only the Desired Product Survives

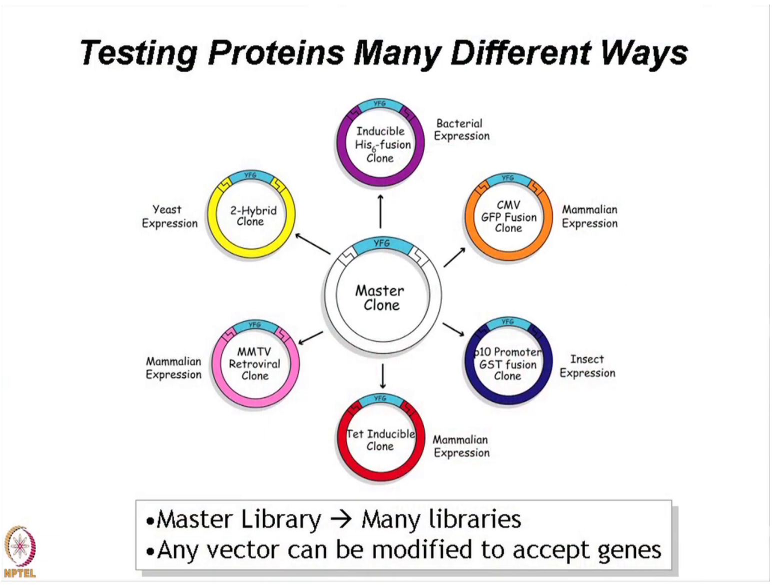
And to get too flexible we focused on this technology called gateway recombination how many of you familiar with gateway, not so many yet, ok, well. Now, imagine doing restriction digests for every gene in the human genome. It gets to be a little complicated because you have to look at which enzymes could this gene could I use for this gene, and which enzyme could I use for that gene, and for really long genes restriction enzymes are going to start cutting up the proteins into pieces, and then you are going to have to reassemble them or you are going to have to clone them in unique ways, it would be very complicated.

So, a number of years ago folks at what a company that was called Life Technologies developed a technology called gateway cloning, it is essentially a type of recombinational cloning. So, the idea is you have you have your favorite gene here and flanking that gene are these site specific recombination sites. And we want to be able to move this your favorite gene into some plasmid vector that allows me to make that protein.

And so by using a common system with gateway, these sites are recognized by an enzyme system from phage lambda. And so you can simply mix this plasmid plus that plasmid in solute in the same sample and add an enzyme, and these two fragments effectively swap locations.

And because these are on they have different selectable markers, and this has a def cassette in this guy, the only viable product is this one. It is the only one that survives. And when that is the only one that survives, now you can essentially develop a method for doing this operation in a high throughput. You can move thousands of genes all by automation.

(Refer Slide Time: 15:53)

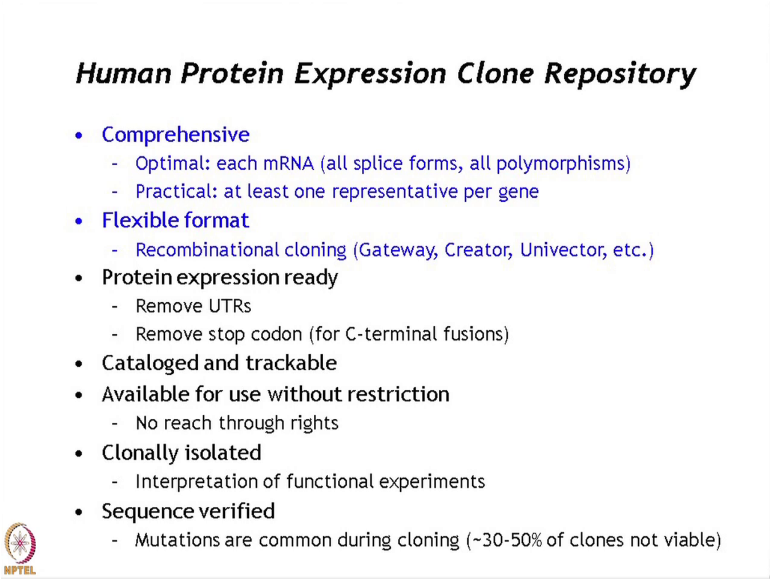


Testing Proteins Many Different Ways

- Master Library → Many libraries
- Any vector can be modified to accept genes

And I will show you that in a moment. So, this is the idea. You build a library of genes in this master vector here. And then the idea is to transfer that gene into any of these other vectors to do any kinds of studies to make protein in insect cells and in human cells, bacterial cells, just by putting the gene into any specific vector and you can do this in high throughput. And my laboratory does that a lot. We move thousands of genes from one vector to another ok.

(Refer Slide Time: 16:22)



Human Protein Expression Clone Repository

- Comprehensive
  - Optimal: each mRNA (all splice forms, all polymorphisms)
  - Practical: at least one representative per gene
- Flexible format
  - Recombinational cloning (Gateway, Creator, Univector, etc.)
- Protein expression ready
  - Remove UTRs
  - Remove stop codon (for C-terminal fusions)
- Cataloged and trackable
- Available for use without restriction
  - No reach through rights
- Clonally isolated
  - Interpretation of functional experiments
- Sequence verified
  - Mutations are common during cloning (~30-50% of clones not viable)

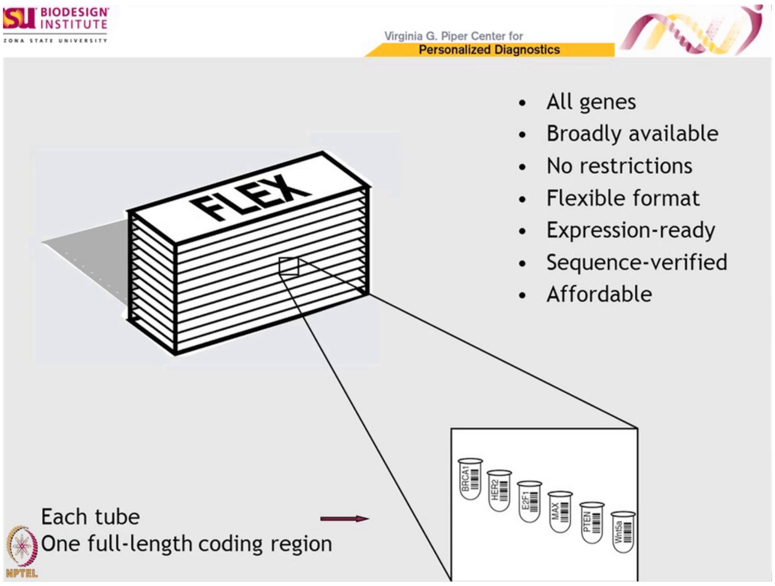
Another thing that you want if you are going to make these clones properly, so that you can do high throughput protein production is you need to make them protein expression ready. And what do I mean by that, well, we have to remove the untranslated sequences from their mRNAs, and we also have to remove the stop codon because if we want to put epitope tags remember we said we want to be able to put tags on these proteins, if there is a stop codon present then when you translate the protein it will stop at the stop codon and it would not allow you to add the epitope tag.

And so one of the things that we had to do was go through all of the genes in the human and remove the stop codons. Of course, it does not work at all if it is not cataloged and trackable. So, you have to build into the whole system a database, a tracking database and a storage system, so that when you want a gene you know where to find it. So, it is it is the molecular version of building a library right. You have to store the books in a place where you can find them same way with the genes here.

One of the things that we wanted in our system was that we wanted to make these clones available to everybody. So, if you are going to a library of all the genes in the human or any other organism, it should be a resource that we all share. And so when we built this we built this in such a way that we could share it with everybody. And then the last thing of course, if you have done molecular biology you know that when you make molecules, sometimes you get a mixture. And a mixture is useless if you are trying to do experiments where you know what you are testing.

And so one of the things we want to make sure we did was that we individually isolated each unique clone, so that when we sequenced it and used it we knew exactly what we were working with. There was no doubt about what it was. And that is the last thing I mentioned to you which is that we sequence verified everything we built that was key because we oftentimes what you get does not work ok.

(Refer Slide Time: 18:24)



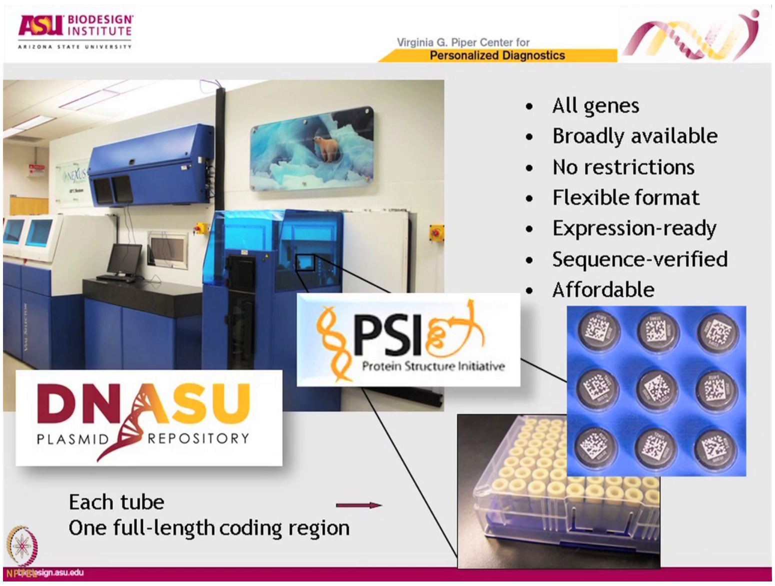
- All genes
- Broadly available
- No restrictions
- Flexible format
- Expression-ready
- Sequence-verified
- Affordable

Each tube
One full-length coding region

So, here is the goal of what we were trying to build. We called it flex to begin with for full length expression ready and it had a number of attributes to it right. It had the goal was to get all genes in it we want to make it broadly available. We want to use a flexible format, we wanted them to be protein expression ready, and we want them to be sequence verified, and of course, we wanted this to be affordable, so that people could use it.

And this is sort of a cartoon that we drew years and years ago about what this would look like, sort of this idea of a lot of tubes that had barcodes on them each one representing a unique gene and each one addressable.

(Refer Slide Time: 19:03)



Virginia G. Piper Center for Personalized Diagnostics

- All genes
- Broadly available
- No restrictions
- Flexible format
- Expression-ready
- Sequence-verified
- Affordable

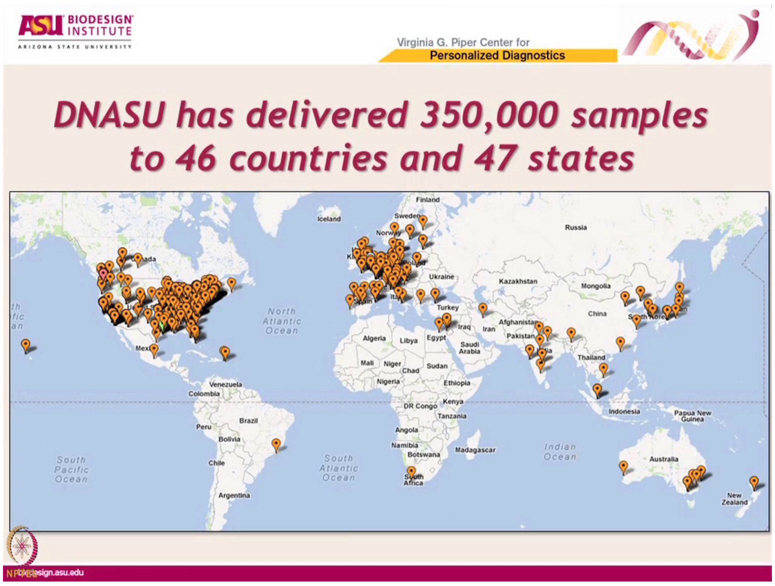
Each tube
One full-length coding region →

What the good news is that that dream is now becoming a reality that is this is what it looks like today. What you are looking at here is a 2 million dollar freezer, it is a very expensive freezer, but it stores tubes in this format here. This is what the tubes look like and on the bottom of these tubes here you have these 2D barcodes and those 2D barcodes are unique for each gene.

So, if we were to drop a rack of these tubes not that we ever drop racks of tubes, but if we dropped a rack of tubes, we could pick them up and put them in random order into a box, and then the barcode reader would read all those barcodes, and it would know exactly where every gene was because the barcodes are unique for every gene right.

And of course, all of this is available at this website DNASU, and I encourage you all to go to that website. all you need is those five letters and that is a list of all the genes that we have in our collection right. Now, we have over 330,000 unique plasmids in our collection, so a very large collection of plasmids. And they are all available to all of you. They are available to everybody on the planet. We ship them every, we ship them every day. ah

(Refer Slide Time: 20:23)



ASU BIODESIGN INSTITUTE
ARIZONA STATE UNIVERSITY

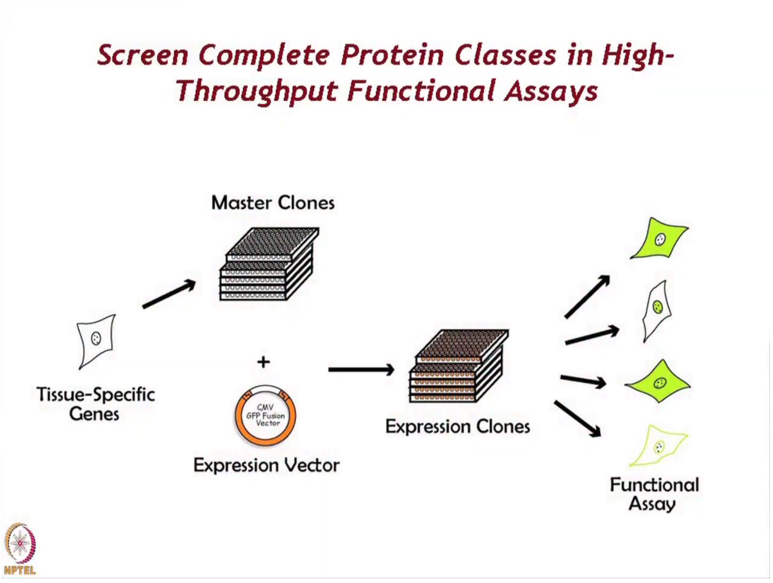
Virginia G. Piper Center for
Personalized Diagnostics

DNASU has delivered 350,000 samples to 46 countries and 47 states

biodesign.asu.edu

In fact, I think we have shipped over 350,000 samples worldwide now. Now, they are not all human, some of them are other organisms, they are not all in gateway, but these are all plasmids that we have made, and where other people have made and given to us to share with them for uses in all kinds of experiments. So, what does this allow you to do if you have all these different clones for all these protein genes.

(Refer Slide Time: 20:53)

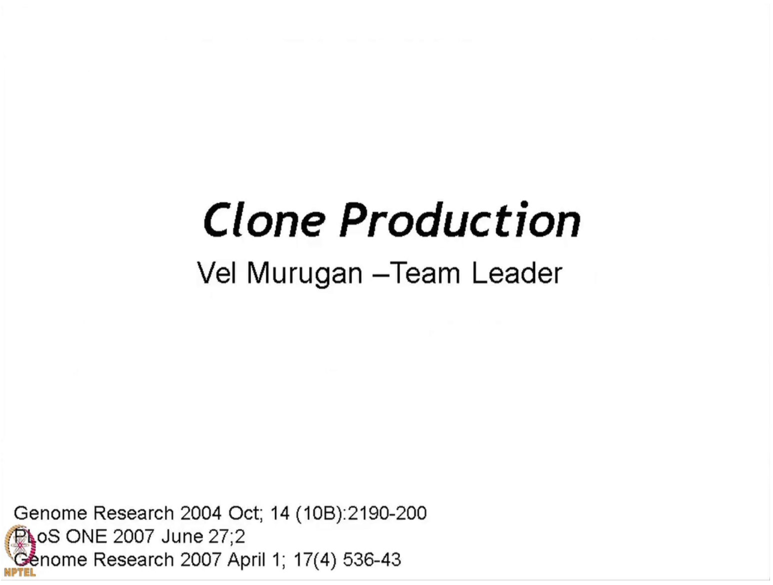


Screen Complete Protein Classes in High-Throughput Functional Assays

Well, imagine that you wanted to look at it do a study of a set of genes that are unique to a particular tissue maybe you are looking at neurological systems, because you are studying brain tumors or you are looking at liver cells because you and you want to look at genes expressed in livers in specifically in hepatic cells. You can go to the library that has the set of master clones.

You can take those master clones and mix them with this expression vector to make the expression clones, the ones that have the gene in the unique vector that will make proteins in the setting that you want to study. And let us say you put them into cells and do some kind of functional assay, and ask where do these proteins localize or what do these proteins interact with. So, the idea is to study proteins and high throughput and the key is to have genes for those proteins in a format that allows you to move them and study them in that setting.

(Refer Slide Time: 21:41)



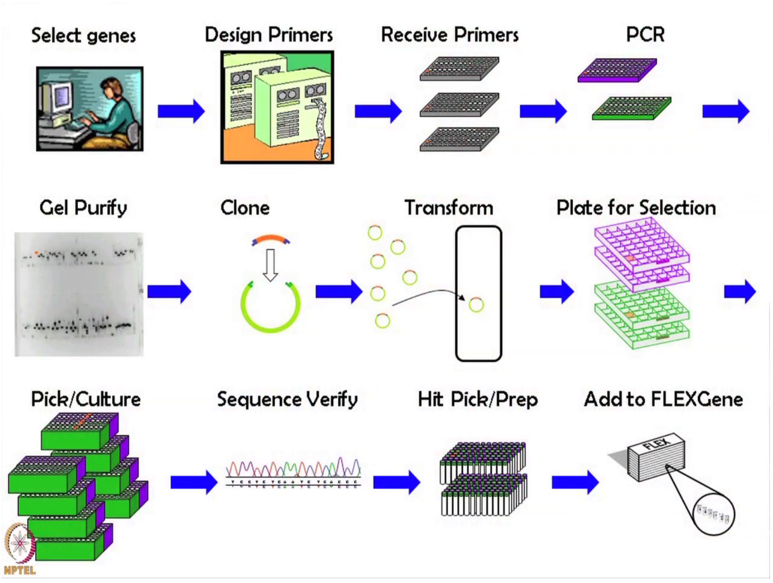
Clone Production

Vel Murugan –Team Leader

Genome Research 2004 Oct; 14 (10B):2190-200
PLoS ONE 2007 June 27;2
Genome Research 2007 April 1; 17(4) 536-43

So, I will tell you a little bit about how we make these clones. We still do that; we are still trying to finish the human library. We have got now almost 15,000 unique human genes cloned that is well on the way to getting to the unique set that we are aiming for is around 18,000. So, we are very close to getting the full the full set.

(Refer Slide Time: 22:10)



The process looks a little bit like this. This is a an overview I will admit that its altered a little bit in recent years and I will tell you where those changes have been made. But basically we start by identifying the genes of interest. We design PCR primers that will capture just the open reading frame for that gene we then do PCR with those primers in 96 well plate. So, high throughput PCR to capture inserts that are unique to the gene, we then capture them into the vector using a recombinational cloning system, transform them into bacterial plate them, pick them for culture and then sequence them to make sure that they are correct.

Now, I will mention a couple of things that we do nowadays a little bit differently. One thing that we are doing a little bit differently is that sometimes now instead of managing all of these unique clones as separate clones, sometimes we will work in batches of pools of clones do all the processing in the batch, and then individually pick them with a colony selector. So, we always colony select them as unique entities, but sometimes you can do some of the processing in batch mode.

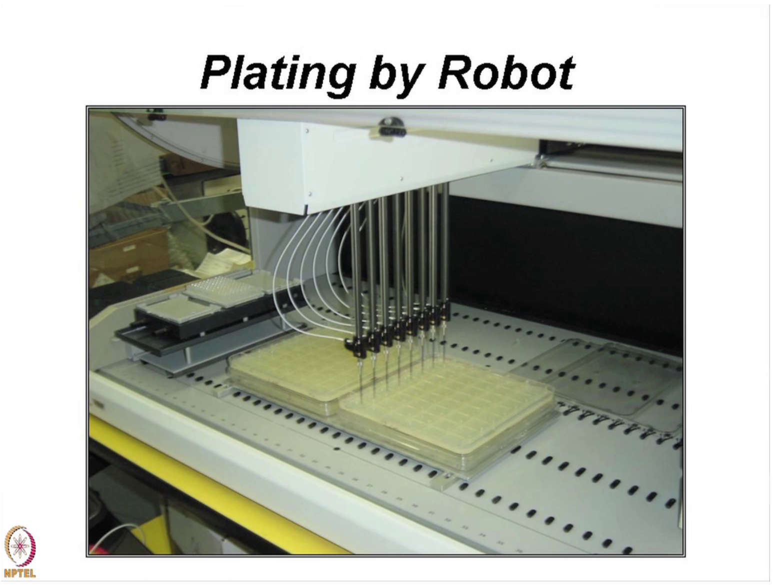
The other thing that we do is nowadays we can sequence them in batches as well using next generation sequencing which was not available when we began this process. So, you can actually pool clones, extract their DNA, do the sequencing as a batch and then and then use that to interpret the c v s of the clones. Now, there is a trick there is a problem with that right. And the problem with that is that when you do next gen sequencing, you cannot tell which clone a particular sequence comes from right. Next gen is just all the sequence that is in the tube, and so you have to be clever about how you set this up.

First of all you have to make sure that when you mix clones together that they are nothing like each other, because if you put two clones that are similar in sequence and you get a mutation, you would not know which clone that came from, that makes sense. So, if you have two genes that are almost identical, and in one of those identical regions you see an alteration, you would not know which it came from. So, whenever you mix these clones you have to do so using informatics approaches upfront that make sure that they are not at all alike.

The second thing that you have to do is you have to realize that when you sequence them on batch, you can tell what them the overall sequence of the gene was, but you cannot confirm that that gene is in that in its appropriate tube right. And we need to know that the correct gene is in the correct tube. So, in addition to the next gen sequencing of the whole batch, we also have to do at least one sequencing read for each gene uniquely from that too, so that we can confirm that we have the right gene in the right place, because this comes back to that library thing.

In the end you are building a library where you can go and get a specific gene from a specific tube anytime you want it.

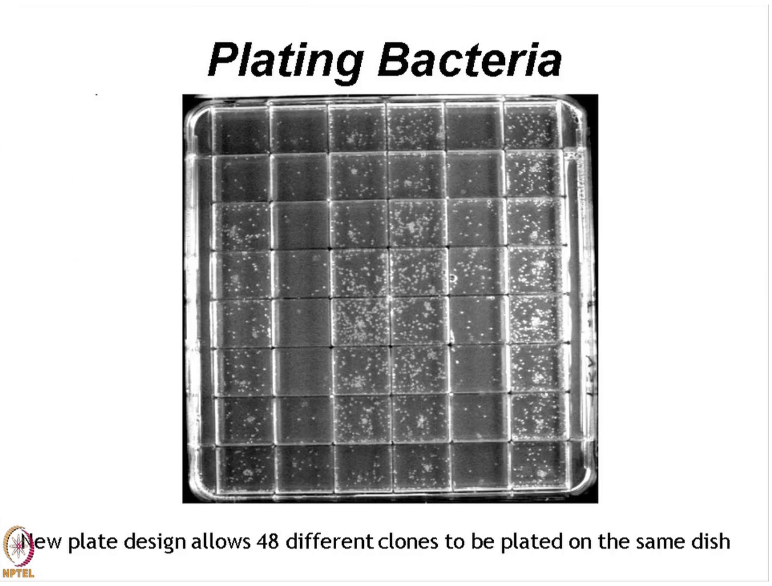
(Refer Slide Time: 25:22)



Plating by Robot

So, we spend a lot of time thinking about that. Here is some of the automation that we use. This is a robot. It is we have transformed bacteria with DNA member, I told you we would transform the bacteria with the DNA. We picked each of these different wells and we have plated them on these specialized plates.

(Refer Slide Time: 25:34)

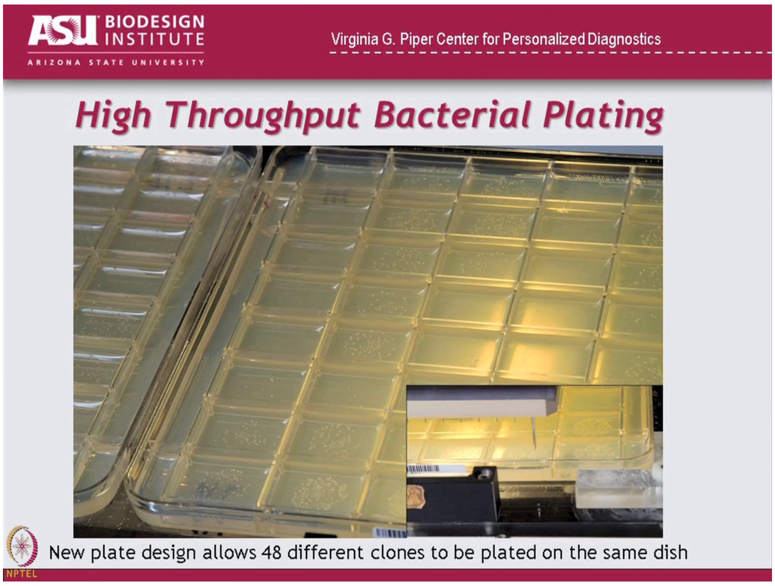


Plating Bacteria

New plate design allows 48 different clones to be plated on the same dish

And these are plates that we actually invented in our laboratory. You now see them widely used in the field. What they are is, they are these bioassay dishes, they are shaped like this. And they have columns and rows. And each of these little areas here is a different clone a different gene. And you can see I hope you can see the different bacterial colonies collecting there.

(Refer Slide Time: 25:58)



ASU BIODESIGN INSTITUTE ARIZONA STATE UNIVERSITY

Virginia G. Piper Center for Personalized Diagnostics

High Throughput Bacterial Plating

New plate design allows 48 different clones to be plated on the same dish

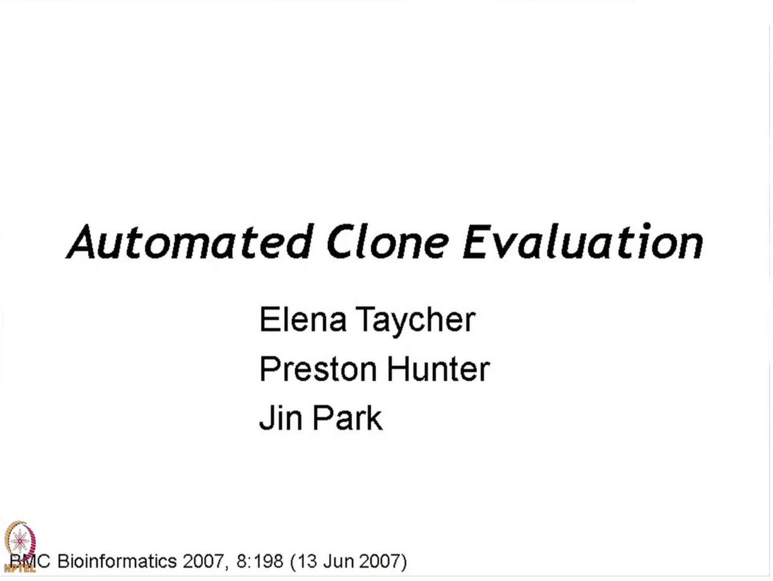
NPTEL

And of course, this is then addressable by robots that can pick individual colonies. So, we used to use undergraduates to pick colonies, and they were very well meaning, but believe it or not human beings make a lot of errors when they have to spend a lot of time using toothpicks to pick colonies and put them in wells. And so our error rate was around 15 percent. Since then we now have robots to do this.

Robots do not take coffee breaks; robots do not forget where they were, and robots can work for many many hours without getting tired. So, you see here is here is the robot, and there is a little pin coming down here and that is going to pick the colony. And hopefully I think you can see the little colonies on the augur there. So, so we do a lot of the colony picking by this method all right.

So, now you get all these clones right, you have made this library of clones. And you have them all in these tubes and you even done some DNA sequencing how do you know that they are correct, how are you going to make sure that the gene that you have in that in that well is correct, and all the sequences are right, or if they are not right how can you document that they are wrong. Well, you could hire lots and lots of people to spend lots and lots of time reading the sequences and assembling the sequences for all these clones right, or you could get clever and you could develop a software tool to do that.

(Refer Slide Time: 27:25)



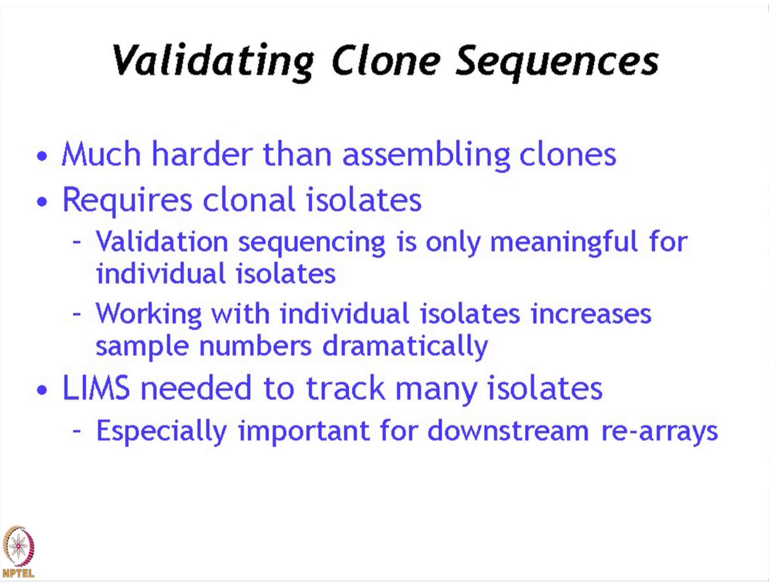
Automated Clone Evaluation

Elena Taycher
Preston Hunter
Jin Park

BMC Bioinformatics 2007, 8:198 (13 Jun 2007)

And that is what we did. We developed software that actually goes through and evaluates the clone sequence, compares it to the correct sequence, and lets us know where there are differences all right.

(Refer Slide Time: 27:44)



Validating Clone Sequences

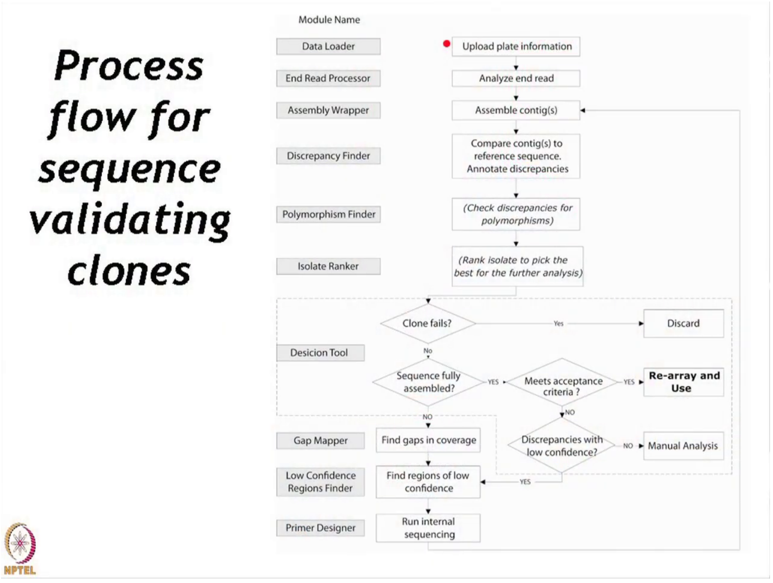
- Much harder than assembling clones
- Requires clonal isolates
  - Validation sequencing is only meaningful for individual isolates
  - Working with individual isolates increases sample numbers dramatically
- LIMS needed to track many isolates
  - Especially important for downstream re-arrays

So, I will tell you a few features of validating clone sequences. First of all much harder than actually making the clones; making the clones is relatively straightforward it is a lot of molecular biology steps you can do it, it is not terrible, but actually making sure that the sequences are correct is takes a lot of time. The first thing is of course, you have to you have to pick individual colonies, I mentioned that before.

Sequencing has no value if you are sequencing a mixture of things, because as we said earlier if there is a mixture, you will never know which one is correct and which one is wrong right and so, but of course, when you are working with individual clones you have a lot more work to do, because you have lots more of those. And then of course, you need what is called a LIM system. Are you guys familiar with the term LIM system? L I M: - Laboratory Information Management System.

What that does is it is an automated software application that is going to you manage all of the steps in your laboratory. It is going to track each gene, each clone from well to well as it moves through all the various robotic steps. Of course, this in this implies that all of your steps are going to be done on 96 well dishes with barcodes on them; so that you are you are always tracking using informatics where things are located.

(Refer Slide Time: 29:07)



Process flow for sequence validating clones

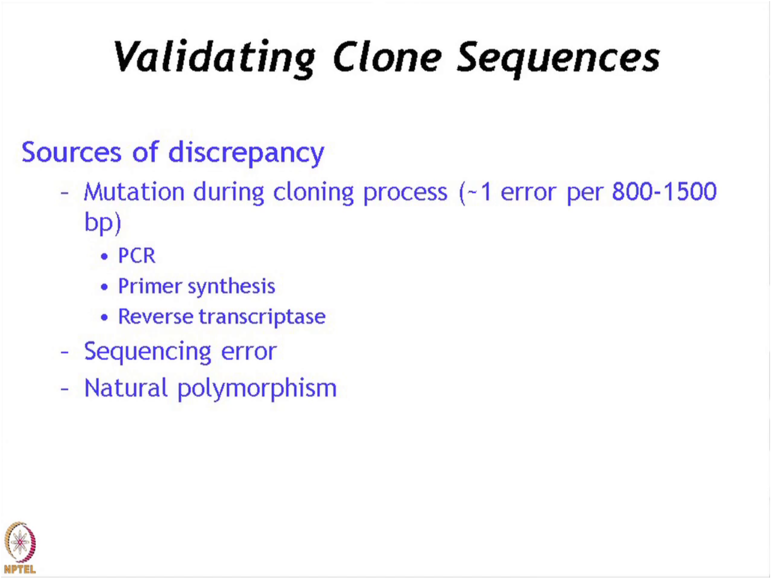
So, so this is the flow process that we used for sequence validating our clones. It began it begins by loading up the plate information that is the information of your plate that has all the clones on it, and what genes are supposed to be in there. We then read end reads, we do you note an end read is that is just the very end of the gene. The nice thing about an end read is that the primer the sequencing primer that you use can be in the plasmid vector.

So, it is the same primer for every gene in your collection, because it does not begin in the gene, it begins outside the gene in the neighboring DNA sequence. And it and the nice thing about that is it tells you that you have the right gene. We then have to assemble all the different reads and this is typically for sequencing where you had to do multiple reads per gene. We then compare the sequences to make sure that they are correct.

So, we look for what are called discrepancies. And I will come back to what I mean by discrepancies in a moment. We then make sure that they are not just common polymorphisms, and then we rank the isolates. And then we have this decision tool here which basically goes and asks if you have a discrepancy, is that discrepancy likely to be a mutation. And if it is a mutation, do I would reject this clone or not. Because at the end we have to decide do we keep it or do we fail to clone.

And then in addition to all of that we have to make sure that we have got the complete sequence. So, when we assemble the sequences, we compare the sequence of the gene to the expected sequence, and we ask do we have it all, have we sequenced everything or do we need to go back and get more sequence ok. I would not go into too long. So, let me tell you about that is when I when I mean by the discrepancy finder.

(Refer Slide Time: 30:58)



Validating Clone Sequences

Sources of discrepancy

- Mutation during cloning process (~1 error per 800-1500 bp)
  - PCR
  - Primer synthesis
  - Reverse transcriptase
- Sequencing error
- Natural polymorphism

So, what are the reasons that a clone sequence does not match the correct or the expected sequence; turns out that there is more than one reason why that could happen of course. So, obviously, one source the one that were most worried about is that the clone underwent mutation, that during the process of amplifying the DNA or capturing it or making the primers mute errors were introduced. And of course, if we have too many errors in a clone, it is no longer useful right because now we are not looking at biology we are looking at mutants.

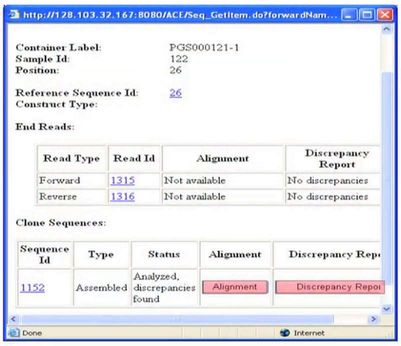
But a much more common reason why the clone sequence does not match is sequencing error, it turns out the actual process of doing the sequencing in itself has errors. And so therefore, we may get a sequence that is incorrect, but it is not the clones problem, it is the sequencing problem it turns out that sequencing errors can occur as often as one in a hundred bases. So, if it is happening one hundred bases, and your clone is a thousand bases long, there is a good chance you are going to have errors in there.

So, how do you fix that you? You go back and you read it again and sometimes you have to get multiple reads to make sure that you have the right clone. Of course, another reason why your clone might not match the natural the clone sequence that you have in your database is, it could be a natural polymorphism right. If we were to sequence the genes of everybody in this room, I guarantee you will find differences all over the place.

And those differences do not reflect that your mutants, it just reflects the natural variation that occurs within a population. We all have sequence variants in our in our students. In fact, I just had my genome sequenced, this fall as part of a project at ASU, and sure enough I found all kinds of sequence variation, and I have no idea what it means.

(Refer Slide Time: 33:01)

Analyzing the Clone's Sequence - Tracking Discrepancies



http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardNam...

Container Label: PGS000121-1
Sample Id: 122
Position: 26

Reference Sequence Id: 26
Construct Type:

End Reads:

Read Type	Read Id	Alignment	Discrepancy Report
Forward	1315	Not available	No discrepancies
Reverse	1316	Not available	No discrepancies

Clone Sequences:

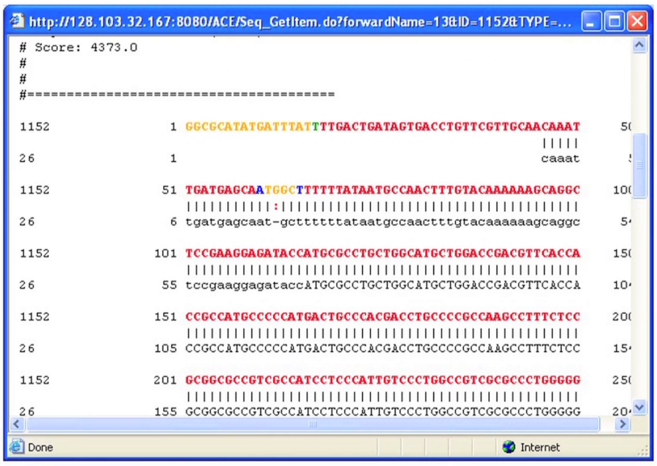
Sequence Id	Type	Status	Alignment	Discrepancy Rep
1152	Assembled	Analyzed, discrepancies found	Alignment	Discrepancy Repor

Discrepancies between the reference sequence and actual sequences are recorded as a list of discrepancy objects.

So, this is how we track sequences. This is the forward read the reverse read of a clone. And this is the assembled sequence, and then we can look at its alignment, and we can look at all the discrepancies that we find.

(Refer Slide Time: 33:13)

Analyzing the Clone’s Sequence - Tracking Discrepancies

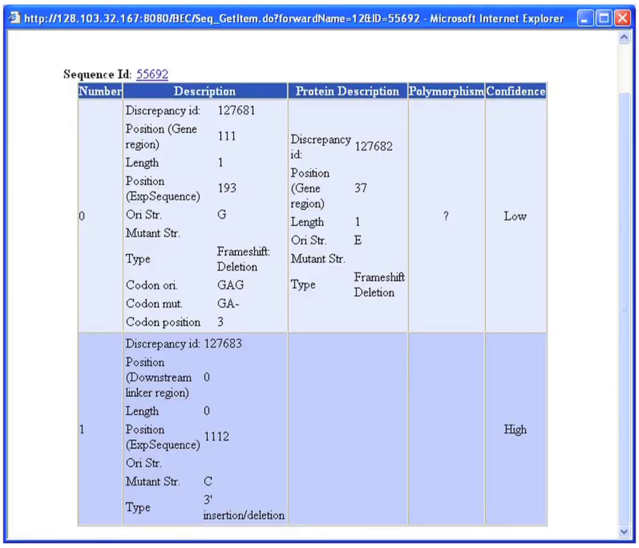


http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardName=13&ID=1152&TYPE=...

```
# Score: 4373.0
#
#
#========================================

1152              1 GGCGCATATGATTTATTTTGACTGATAGTGACCTGTTCGTTGCAACAAAT      50
                                                                |||||
26                1                                             caaat       5

1152             51 TGATGAGCAATGGCTTTTTTATAATGCCAACTTTGTACAAAAAAGCAGGC     100
                    ||||||||||:|||||||||||||||||||||||||||||||||||||||
26                6 tgatgagcaat-gcttttttataatgccaactttgtacaaaaaagcaggc      54

1152            101 TCCGAAGGAGATACCATGCGCCTGCTGGCATGCTGGACCGACGTTCACCA     150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
26               55 tccgaaggagataccATGCGCCTGCTGGCATGCTGGACCGACGTTCACCA     104

1152            151 CCGCCATGCCCCCATGACTGCCCACGACCTGCCCCGCCAAGCCTTTCTCC     200
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
26              105 CCGCCATGCCCCCATGACTGCCCACGACCTGCCCCGCCAAGCCTTTCTCC     154

1152            201 GCGGCGCCGTCGCCATCCTCCCATTGTCCCTGGCCGTCGCGCCCTGGGGG     250
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
26              155 GCGGCGCCGTCGCCATCCTCCCATTGTCCCTGGCCGTCGCGCCCTGGGGG     204
```

If you click on the alignment button, then you get something that looks like this which is showing the alignment of the sequence with the expected sequence. And obviously, these colors indicate where we see discrepancies right here for example, there are some discrepancies. Now, you will notice that these discrepancies are occurring very close to the end of the gene, and that that could be a sign that there are sequencing errors, because usually at the beginning and end of reads you get some mistakes that come up.

(Refer Slide Time: 33:45)


http://128.103.32.167:8080/BEC/Seq_GetItem.do?forwardName=12&ID=55692 - Microsoft Internet Explorer

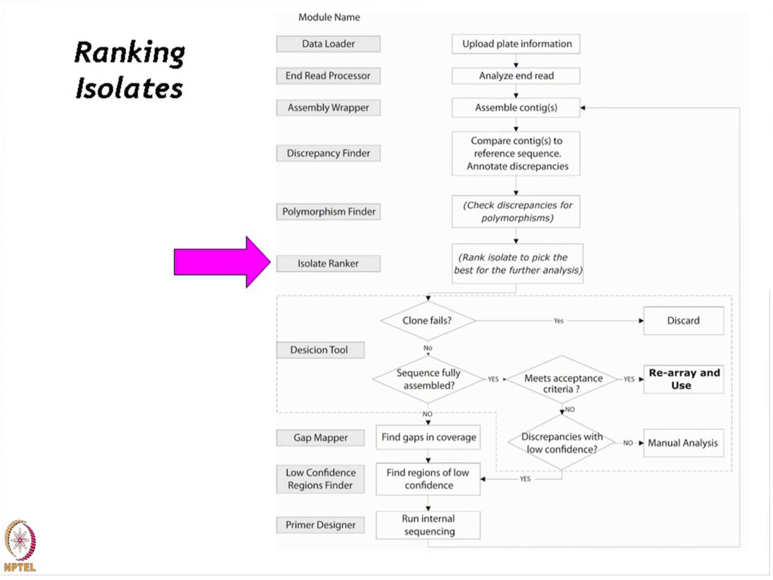
Sequence Id: 55692

Number	Description	Protein Description	Polymorphism	Confidence
0	Discrepancy id: 127681 Position (Gene region) 111 Length 1 Position (ExpSequence) 193 Ori Str. G Mutant Str. Type Frameshift: Deletion Codon ori. GAG Codon mut. GA- Codon position 3	Discrepancy id: 127682 Position (Gene region) 37 Length 1 Ori Str. E Mutant Str. Type Frameshift Deletion	?	Low
1	Discrepancy id: 127683 Position (Downstream linker region) 0 Length 0 Position (ExpSequence) 1112 Ori Str. Mutant Str. C Type 3' insertion/deletion			High

And then and then here is what we this is if you click on the discrepancy button, you will get this report. And it will tell you every time there is a difference between our sequence and the expected sequence. What that difference is, what kind of difference it is, and then what implication it has on the protein. In this case, there is a frame shift deletion; that means, that where we have gone out of sync from the triplet codons that you expect in DNA.

When you go out of sync, you have the increased your opportunity to run into a stop codon and cause an aberrant truncation of the protein and that is what happened in this case right. Obviously, mutations that cause profound changes like that are much more deleterious in our clones than simple substitution mutations.

(Refer Slide Time: 34:36)



Ranking Isolates

Module Name

Data Loader — Upload plate information

End Read Processor — Analyze end read

Assembly Wrapper — Assemble contig(s)

Discrepancy Finder — Compare contig(s) to reference sequence. Annotate discrepancies

Polymorphism Finder — *(Check discrepancies for polymorphisms)*

Isolate Ranker — *(Rank isolate to pick the best for the further analysis)*

Desicion Tool

Clone fails?
Yes
Discard
No
Sequence fully assembled?
YES
Meets acceptance criteria ?
YES
Re-array and Use
NO
Discrepancies with low confidence?
NO
Manual Analysis
YES
NO

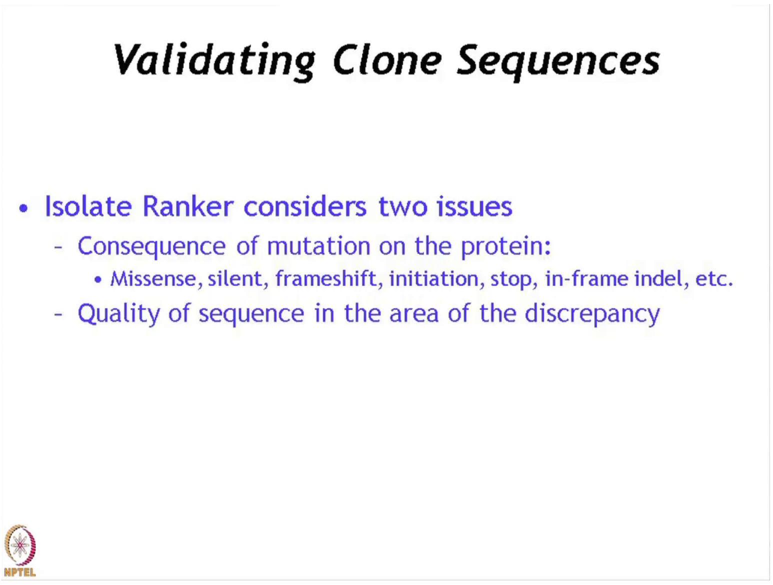
Gap Mapper — Find gaps in coverage

Low Confidence Regions Finder — Find regions of low confidence

Primer Designer — Run internal sequencing

This isolate ranker is just a tool that basically considers two issues. First as I indicated a moment ago, what are the consequences of the mutations, if the consequences are going to profoundly affect the protein then that would make on isolate much less likely to be interesting.

(Refer Slide Time: 34:56)



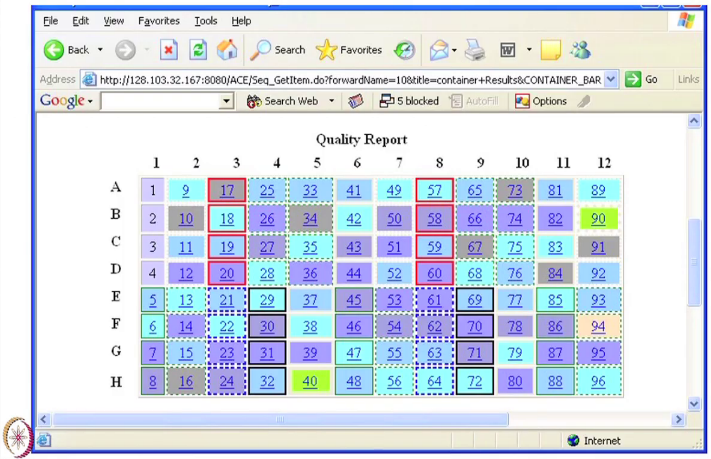
Validating Clone Sequences

- Isolate Ranker considers two issues
  - Consequence of mutation on the protein:
    - Missense, silent, frameshift, initiation, stop, in-frame indel, etc.
  - Quality of sequence in the area of the discrepancy

And then we need to know is the quality of the sequence in the area good quality sequence. Because, if the sequence quality is bad, then a much less likely to believe the mutation. If the sequence quality is bad, I am going to there is a very good chance that the mutation is due to bad sequencing and I could not actual mutation.

(Refer Slide Time: 35:18)

Ranking Individual Isolates



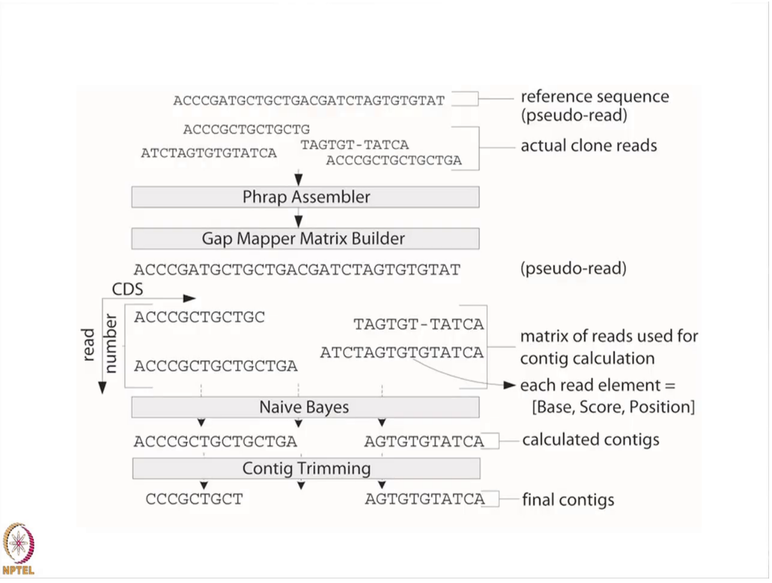
Address http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardName=10&title=container+Results&CONTAINER_BAR

Quality Report

	1	2	3	4	5	6	7	8	9	10	11	12
A	1	9	17	25	33	41	49	57	65	73	81	89
B	2	10	18	26	34	42	50	58	66	74	82	90
C	3	11	19	27	35	43	51	59	67	75	83	91
D	4	12	20	28	36	44	52	60	68	76	84	92
E	5	13	21	29	37	45	53	61	69	77	85	93
F	6	14	22	30	38	46	54	62	70	78	86	94
G	7	15	23	31	39	47	55	63	71	79	87	95
H	8	16	24	32	40	48	56	64	72	80	88	96

So, in the end you will get a chart that looks like this. And these various color codes indicate to us which clones are better than which other ones. So, we can pick the best clone for a gene. And then this last tool I will mention here is the gap mapper. And I remember I told you ideally we have sequence for the entire gene; if we do not have sequence for the entire gene, we need to go back and get an additional read to fill in the gap; otherwise we cannot say with certainty that we have a good clone.

(Refer Slide Time: 35:51)



ACCCGATGCTGCTGACGATCTAGTGTGTAT

reference sequence
(pseudo-read)

ACCCGCTGCTGCTG

ATCTAGTGTGTATCA

TAGTGT-TATCA

ACCCGCTGCTGCTGA

actual clone reads

Phrap Assembler

Gap Mapper Matrix Builder

ACCCGATGCTGCTGACGATCTAGTGTGTAT

(pseudo-read)

CDS

read number

ACCCGCTGCTGC

ACCCGCTGCTGCTGA

TAGTGT-TATCA

ATCTAGTGTGTATCA

matrix of reads used for
contig calculation

each read element =
[Base, Score, Position]

Naive Bayes

ACCCGCTGCTGCTGA

AGTGTGTATCA

calculated contigs

Contig Trimming

CCCGCTGCT

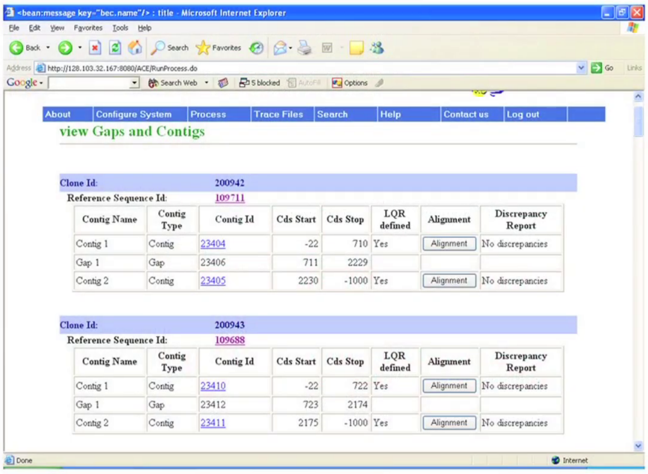
AGTGTGTATCA

final contigs

And so this gap mapper takes all the different reads from a particular gene, it assembles them by overlapping them, and then looks for any areas using essentially Bayesian mathematics. It looks for areas where there are missing areas, and then we trim back the ends a little bit, and then suggest that we have to go back and clone that do another sequence read for that missing area. So, we can get a better clone.

(Refer Slide Time: 36:19)

Finding Gaps and Low Quality Regions


<bean:message key="bec.name"/> : title - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://128.103.32.167:8080/ACE/RunProcess.do
Go Links
Google Search Web 5 blocked AutoFill Options
About Configure System Process Trace Files Search Help Contact us Log out
Done
Internet

view Gaps and Contigs

Clone Id: 200942

Reference Sequence Id: 109711

Contig Name	Contig Type	Contig Id	Cds Start	Cds Stop	LQR defined	Alignment	Discrepancy Report
Contig 1	Contig	23404	-22	710	Yes	Alignment	No discrepancies
Gap 1	Gap	23406	711	2229			
Contig 2	Contig	23405	2230	-1000	Yes	Alignment	No discrepancies

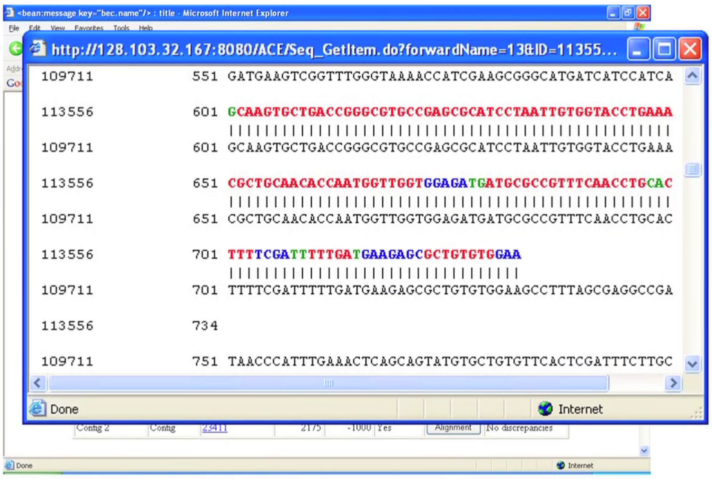
Clone Id: 200943

Reference Sequence Id: 109688

Contig Name	Contig Type	Contig Id	Cds Start	Cds Stop	LQR defined	Alignment	Discrepancy Report
Contig 1	Contig	23410	-22	722	Yes	Alignment	No discrepancies
Gap 1	Gap	23412	723	2174			
Contig 2	Contig	23411	2175	-1000	Yes	Alignment	No discrepancies

(Refer Slide Time: 36:25)

Finding Gaps and Low Quality Regions



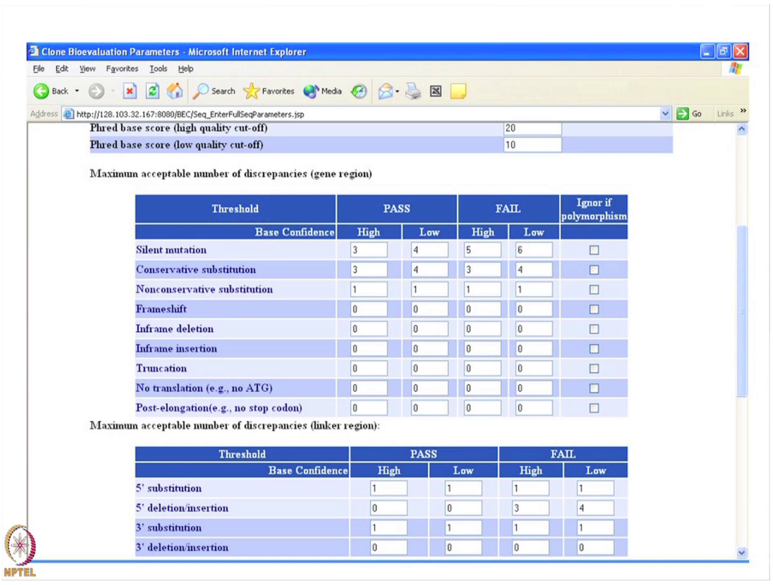
<bean:message key="bec.name"/> : title - Microsoft Internet Explorer
File Edit View Favorites Tools Help
http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardName=13&ID=11355...
Done
Internet

```
109711            551 GATGAAGTCGGTTTGGGTAAAACCATCGAAGCGGGCATGATCATCCATCA

113556            601 GCAAGTGCTGACCGGGCGTGCCGAGCGCATCCTAATTGTGGTACCTGAAA
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
109711            601 GCAAGTGCTGACCGGGCGTGCCGAGCGCATCCTAATTGTGGTACCTGAAA

113556            651 CGCTGCAACACCAATGGTTGGTGGAGATGATGCGCCGTTTCAACCTGCAC
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
109711            651 CGCTGCAACACCAATGGTTGGTGGAGATGATGCGCCGTTTCAACCTGCAC

113556            701 TTTTCGATTTTTGATGAAGAGCGCTGTGTGGAA
                      ||||||||||||||||||||||||||||||||||
109711            701 TTTTCGATTTTTGATGAAGAGCGCTGTGTGGAAGCCTTTAGCGAGGCCGA

113556            734

109711            751 TAACCCATTTGAAACTCAGCAGTATGTGCTGTGTTCACTCGATTTCTTGC
```

Contig 2	Contig	23411	2175	-1000	Yes	Alignment	No discrepancies

And then this is what that this is what it looks like in our software. And so you can see it basically predicts that there is a gap here that needs to be filled in. And then you can see these other these colors here are indicating that the quality of sequence in that area is not great.

(Refer Slide Time: 36:43)



Clone Bioevaluation Parameters - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://128.103.32.167:8080/BEC/Seq_EnterFullSeqParameters.jsp
Go Links
NPTEL

Phred base score (high quality cut-off)	20
Phred base score (low quality cut-off)	10

Maximum acceptable number of discrepancies (gene region)

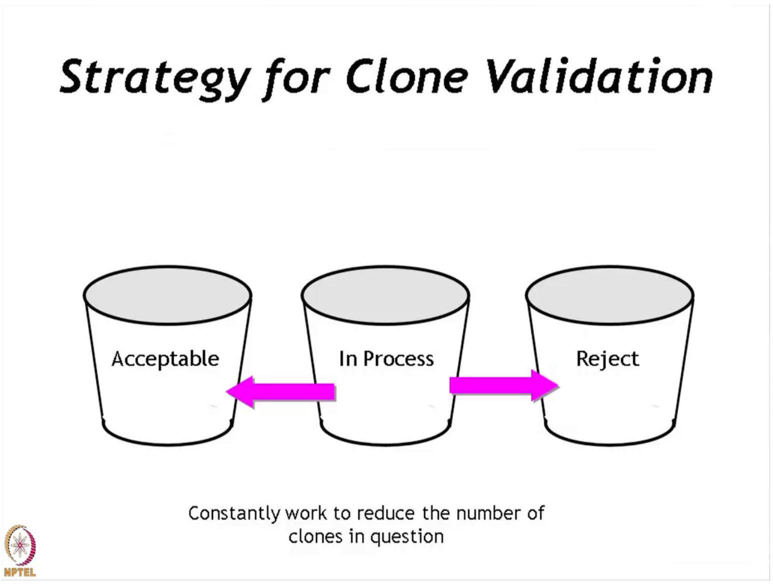
Threshold	PASS		FAIL		Ignor if polymorphism
Base Confidence	High	Low	High	Low	
Silent mutation	3	4	5	6	□
Conservative substitution	3	4	3	4	□
Nonconservative substitution	1	1	1	1	□
Frameshift	0	0	0	0	□
Inframe deletion	0	0	0	0	□
Inframe insertion	0	0	0	0	□
Truncation	0	0	0	0	□
No translation (e.g., no ATG)	0	0	0	0	□
Post-elongation(e.g., no stop codon)	0	0	0	0	□

Maximum acceptable number of discrepancies (linker region):

Threshold	PASS		FAIL	
Base Confidence	High	Low	High	Low
5' substitution	1	1	1	1
5' deletion/insertion	0	0	3	4
3' substitution	1	1	1	1
3' deletion/insertion	0	0	0	0

This is our decision tool. This is how do we decide whether or not to keep a clone. Our goal is always to either eliminate clones or keep them obviously. And so here we set the criteria that will make a pass or a fail. And we allow this is if the sequence is good if this sequence is not so good, then we can also ignore if there are polymorphisms.

(Refer Slide Time: 37:09)



Strategy for Clone Validation

Constantly work to reduce the number of clones in question

And so as I say as we run through our clone list at any given time, we are always trying to move clones either into the reject category or the acceptable category all right. So, that let me stop there, and see if there are any questions on the cloning process of making clones for collections. There any questions I can answer? Yeah.

Student: (Refer Time: 37:37) sequencing error. So, what is the mechanism of sequencing error.

It, it so question was: what is the mechanism a sequencing error that depends a little bit on what platform that you are using to do your sequencing. A lot of what we do is using traditional single clone sequencing you know set what they call Sanger sequencing. And in that case, it can vary what the causes are oftentimes, this Sanger sequencing involves different colors for different bases, and sometimes you get a region where you get a little bit more red than you should and so you cannot really tell is it an A or is it a T, I am not sure.

Sometimes it is just that you do not get adequate coverage, so you do not read as many times pass that base. So, there is a lot the method, the chemistry themselves have errors. Now, lately we are using alumina which is next gen sequencing. It also has an error frequency, but typically with lumina sequencing you get around that by doing so many reads, you cover it 30 times that you are less likely to have an error, but there is it is the process itself is error prone. Other questions?

Student: Sir when we have gene database that itself has much many errors that what it wants sir. So, when we compare our clone.

Yeah.

Student: How do we clone for the errors from the starting back up things?

Yeah, oh, you mean the database that has the gene sequences in it, no, that is a very good point. The gene sequences that are in you know the database is at NCBI in the US in the unit protein sequences all that stuff they have errors in them. And that is and so if we disagree with that it is not always clear that it is us that is at fault. Typically in a lot of cases in our in our circumstance well there is two let me say there is two ways that we have dealt with that.

The first is oftentimes we start making our genes from existing clones where we actually know their sequence. In that case we know what we are trying to achieve and we try to match that sequence. In the case you are referring to where we are trying to match a sequence in a database, we actually did develop a polymorphism tool and I had slides on that and I took them out because it was going to get too long.

But basically what that polymorphism tool does is it goes out to all the existing databases where there have been gene sequences uploaded for all these human genes, collects all the sequences from those genes and lines them up, and looks at the frequency at any given position and asks are there existing examples of other clones that have the sequence I have. And, if there are examples of that sequence, then I am more likely to accept the sequence. It is not perfect, but it does help.

Student: Sir also.

Ok.

Student: For this master group also you do the validation before like making the individual isolated of proteins.

Say, say it again.

Student: For the master group also you people do validation of this protein sequence before (Refer Time: 41:16).

Well, you know once we have once we have done that sequence validation, we think most people do not have to do it again, I mean it is certainly reasonable if you want to be extra careful if it is a very special clone for a research project of yours. But for high throughput of materials we have done a pretty good job of sequencing these, so I do not think you have to repeat that.

And I should point out that one of the qualities of the gateway process which is to transfer the insert from one master clone to an expression clone that is a conservative molecular process. So, once do you know that this sequence is correct, then you know that this sequence is correct. So, you do not have to resequence them both.

Student: So, like you people have your own database of whatever sequencing you have done so far for individual isolates and all.

Yes. In fact, are all of our clones if you go to our website, the DNAs website, we list the actual sequence of that clone. So, we have done the sequence and we have loaded that up on the database. I think there was one over there

Student: Hi sir, you told about the mismatches, mismatches in cloning (Refer Time: 42:25), how many mismatches are allowed you know or from the.

Yeah.

Student: (Refer Time: 42:30).

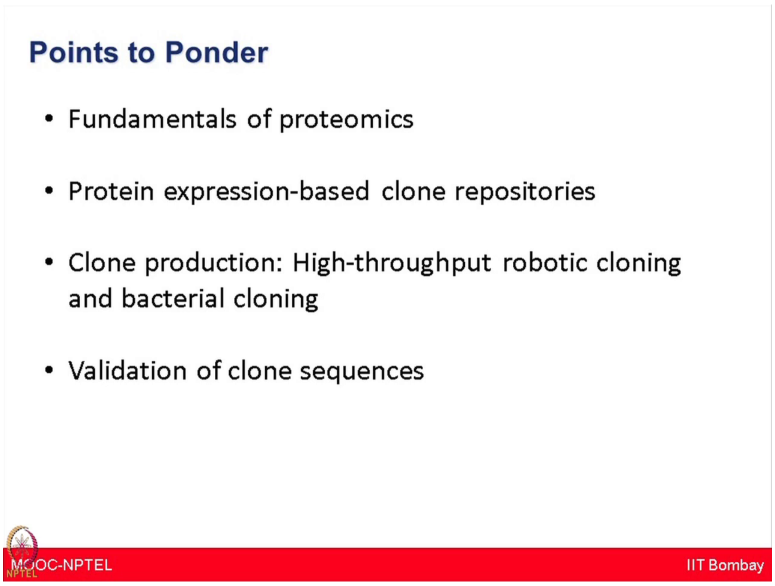
So, in the clone collection that we distribute, we for the most part not every case, but for the most part we try to limit it to no more than one amino acid difference. So, if there is more than two amino acids difference that we do not load it. I will say that at the very minimum we always load the actual sequence. So, you can always look at the actual sequence and ask is this agree enough with what I want to do to user, but most of the time its either 100 percent accurate or we allow one amino acid change.

Student: (Refer Time: 43:04) what is the genome (Refer Time: 43:06).

Well, the genome has not been sequenced, it is very hard to make the clones right. In fact, we learned that the hard way years ago we did a clone collection for an organism called Francis salatullah rensis just causes this illness called tularemia. And we were working with collaborators, and those collaborators were intimately involved in the genome sequence of that organism, and they said will get you an early copy of the genome. So, they gave us an early copy of the genome and we use that to design our clone collection, and then we built all those clones, and it was a disaster.

We our success rate which is usually in the 90 plus percent range was like 50 percent, it was horrible. And then about a year later they came out with the official sequence of the organism, and it was very different from the sequence that they gave us originally, there was a lot of changes in the sequence and. So, when we rebuilt the collection using the correct sequence. Now, we had like 96 percent accuracy. So, you really have to have a good quality genome sequence to do this kind of work.

(Refer Slide Time: 44:19).



Points to Ponder

- Fundamentals of proteomics
- Protein expression-based clone repositories
- Clone production: High-throughput robotic cloning and bacterial cloning
- Validation of clone sequences

MOOC-NPTEL IIT Bombay

So, today you have learned about fundamentals of proteomics. I am sure you are mesmerized, but all you can achieve using proteomic technologies, here also provided a glimpse of different protein expression based clone repositories. You also studied how to do the clone production especially in high throughput manner using robotic plating and high throughput bacterial plating. Finally, you learned how to validate these close sequences which is one of the most important step in the entire high throughput gene cloning pipeline.

We will continue more discussions in the next lecture.

Thank you.