Lecture - 2

NAPPA Technology and Protein Arrays-I

Applications of interactomics

Using genomics and proteomics technologies

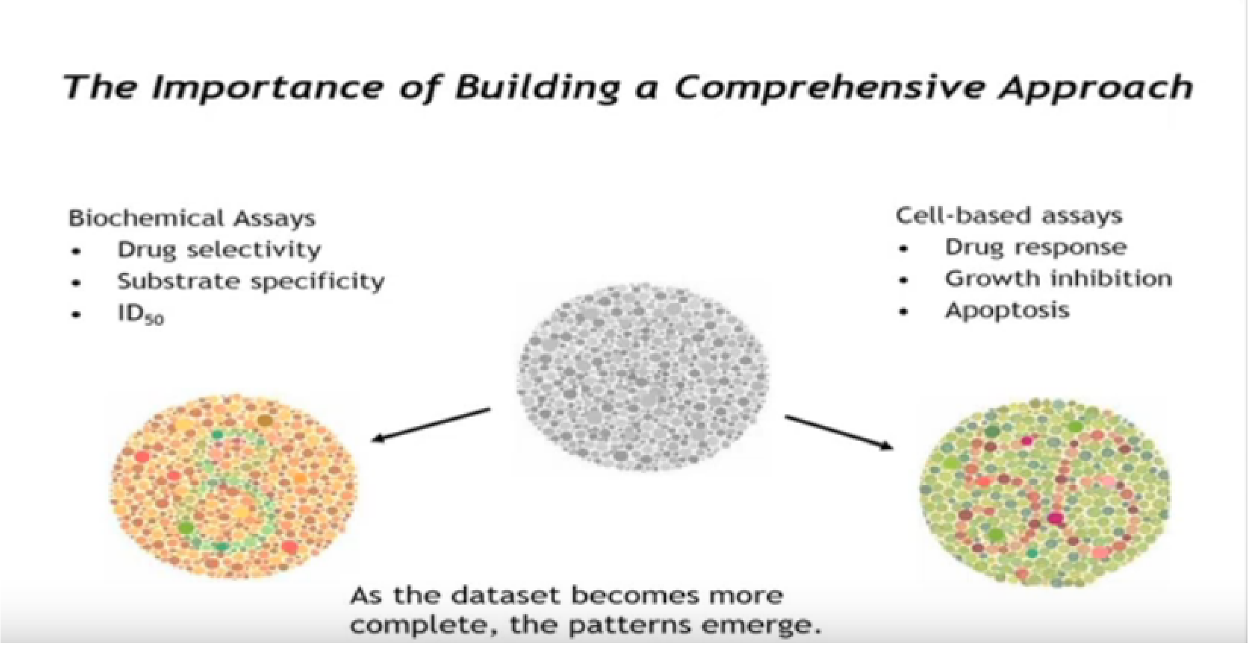
Welcome to more codes, on applications of interact omits, using genomics and proteomics technologies, it is my great pleasure to introduce, distinguished scientist procedural aware. Today and next few lectures, will be delivered by procedural aware, he's the executive director of Biodesign Institute, at Arizona State University and the director of Virginia G Piper bio design Center for personal Diagnostics. Dr. Ravin, has been one of the four most Investigators, in the rapidly evolving field of personal Diagnostics. Josh Ravin, has been instrumental in development of self 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. Ravin, 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. leper, 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, one to seven days after exposure, which is sponsored by biomedical Advanced Research and Development Authority. He’s also, the past president of us. Soup Ohio, and one of the convenience of last year conducted, human proteome Organization, World Congress in Orlando. Dr. Ravin is going to talk about, biomarker discovery based program, various considerations for statistical tools, which are required for biomarker evaluations and validation.

And how to make nappa arrays, using very simple lab based resources, then perform an antibody based screaming, 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, procedure Sheila Baird 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 regions, which you can generate using the normal protein technologies? Then later on you can simply transfer the genes of interest, into any vectors for your given experiment. I'm sure dr. Levine, 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's welcome Dr. Joshua LeBaer for his lecture.

Hi, I think we're ready to get started, yeah. All right? So, I'm gonna start a little bit at the beginning, we have, we have several lectures here to cover ,on terms of the Napa technology and so ,I thought it would be useful to sort of begin where, we begin. So, this is where biology was 10-15 years ago.

Refer Slide Time: (4:59)



The Importance of Building a Comprehensive Approach

Biochemical Assays

- Drug selectivity
- Substrate specificity
- ID_{50}

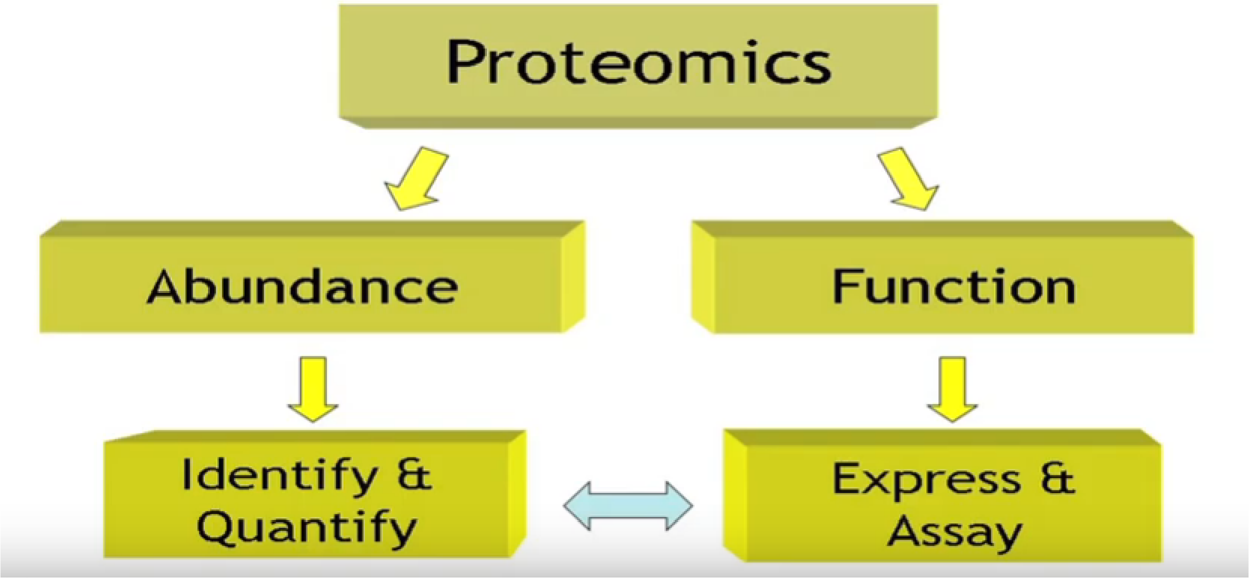
Cell-based assays

- Drug response
- Growth inhibition
- Apoptosis

As the dataset becomes more complete, the patterns emerge.

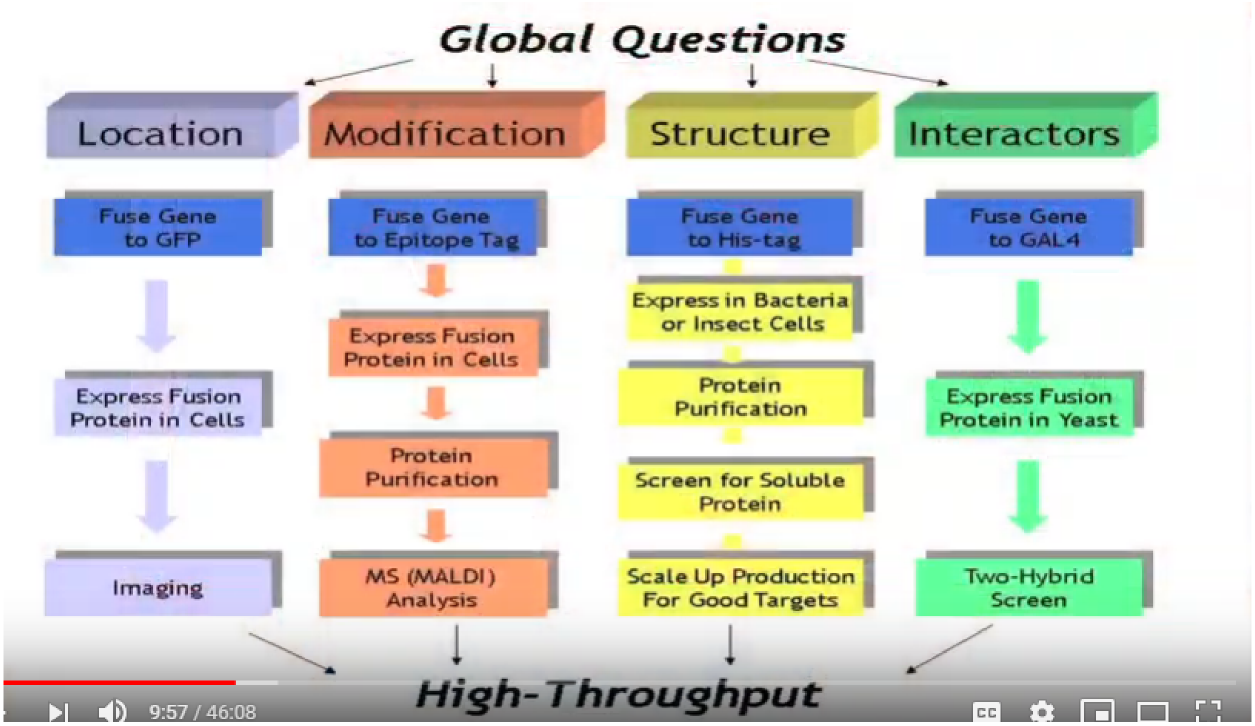
What I mean by that is, that we were studying proteins a few at a time? You know, maybe three or four or five proteins at a time, and that's you know, that's how much information we were getting, but what we were really trying to understand was the entire proteome? And we would take these proteins and 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 can't really get a full picture of what's there right? If you look at this you know, you don't know, what that color means? You look at that, you don't 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're trying to look at? And what it means and that's 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: (6:14)



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'll identify, which proteins are altered in disease, and that will provide useful information about what's causing? What's 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'll 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's 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, so what are the ways that we can look at the function of proteins? Right.

Refer slide time: (7:30)



Global Questions

Location
Fuse Gene to GFP
Express Fusion Protein in Cells
Imaging

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

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

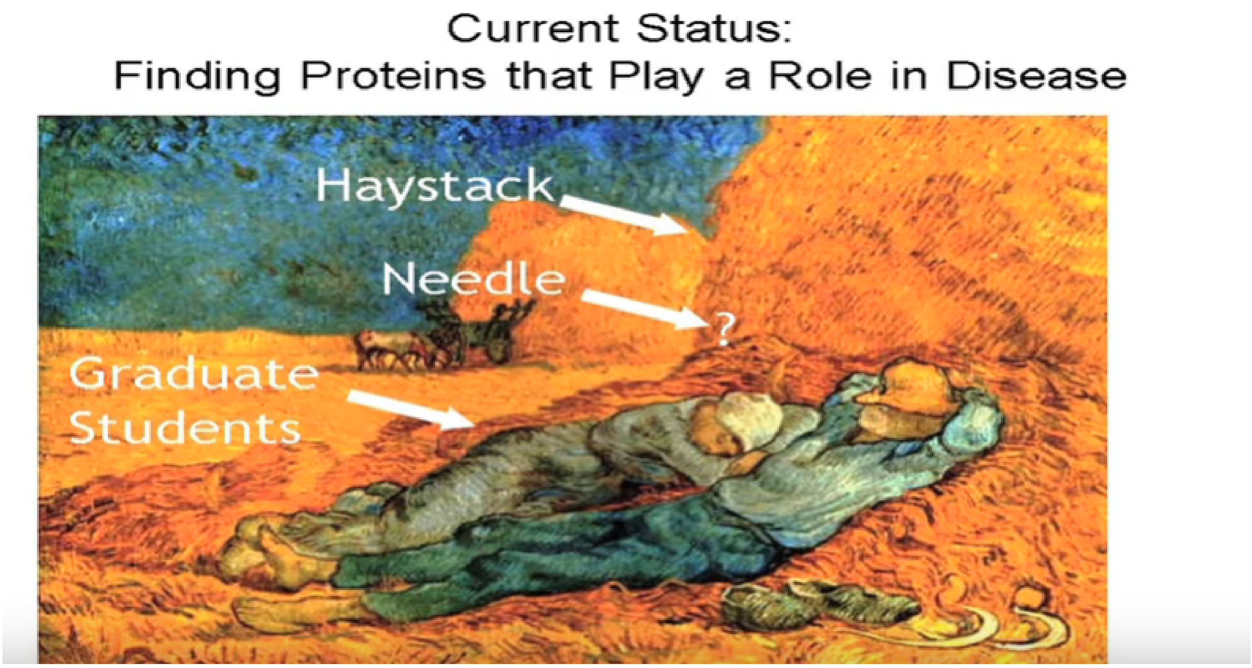
Interactors
Fuse Gene to GAL4
Express Fusion Protein in Yeast
Two-Hybrid Screen

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 set elated, is you know, is it ubiquitinated, modifications of 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, this topic that we're here today to talk about, 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 that? 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're 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, alright. 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's in cells, sometimes it's in cell in, in a cell-free extract, sometimes you're making it in vivo and the normal search circumstance, in other cases you're using a heterogonous 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's very cumbersome and you can't do that times thousands, and the goal here, is to be able to study proteins, hundreds of them at a time or thousands of them at a 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, all right?

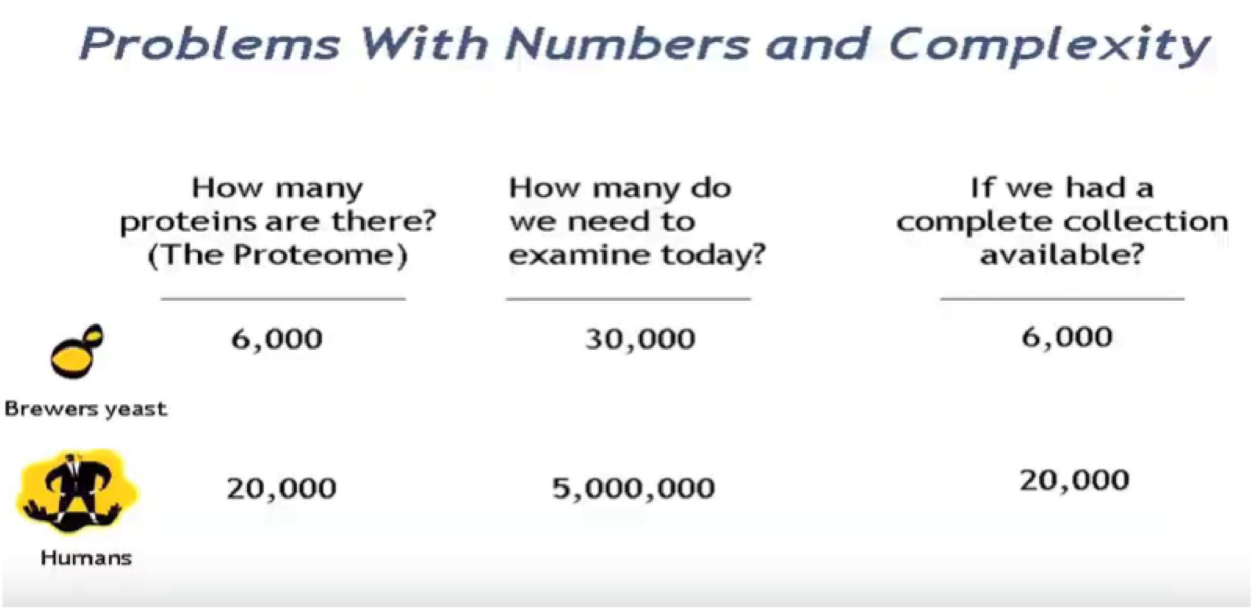
Refer slide time: (10: 45)



Current Status:
Finding Proteins that Play a Role in Disease

And when we began this work this was what? What? The field looked like. Right? So, what am I looking at? What we're looking at, a couple of graduate students. Who are exhausted? so, now why are they exhausted, well they've been looking through those haystacks, for the needle they're trying to find and it takes a long time to sift through the day 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: (11:10)

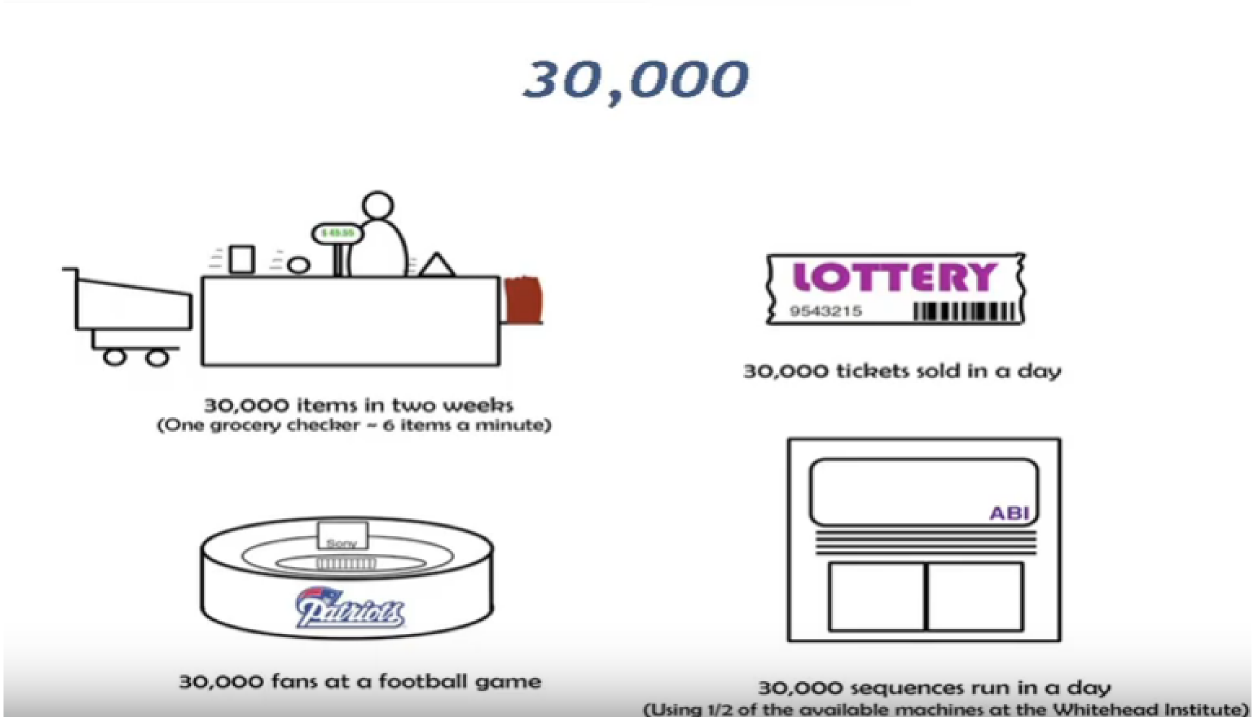


Problems With Numbers and Complexity

	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, you know, if you think about a simple organism like yeast, like sacrum Isis cerevisiae, there's around 6,000 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'd 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, for in, 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 post translational modification, that number expands dramatically but, let's 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 can't if you don't have cloned copies of those genes, if you have them in libraries ,like cDNA libraries or phage display libraries, you can't, if you want to test all proteins, in order to get past all the redundancy ,you would have to do five million assays, and that's 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. So, 20,000, 30,000 assays, that's a number.

Refer slide time: (13:10)



30,000

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

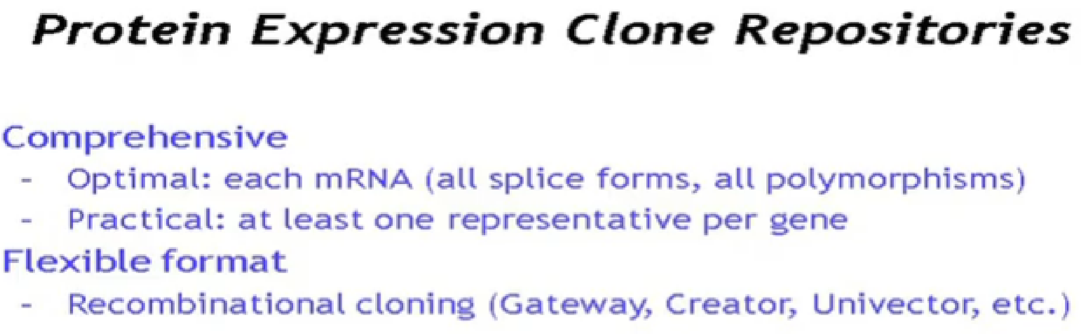
30,000 tickets sold in a day

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)

that i can imagine doing in a high-throughput biochemical setting, in a supermarket, in the united states, if, if you look at around six items a minute, when you're passing ,them that you could get that done in two weeks. Right? They sell 30,000 tickets, for a lottery in a single day and the city of Massachusetts. So, 30 thousands a number, that we could imagine, we could do that .right? And so, that's that was the goal.

Refer slide time: (13:33)

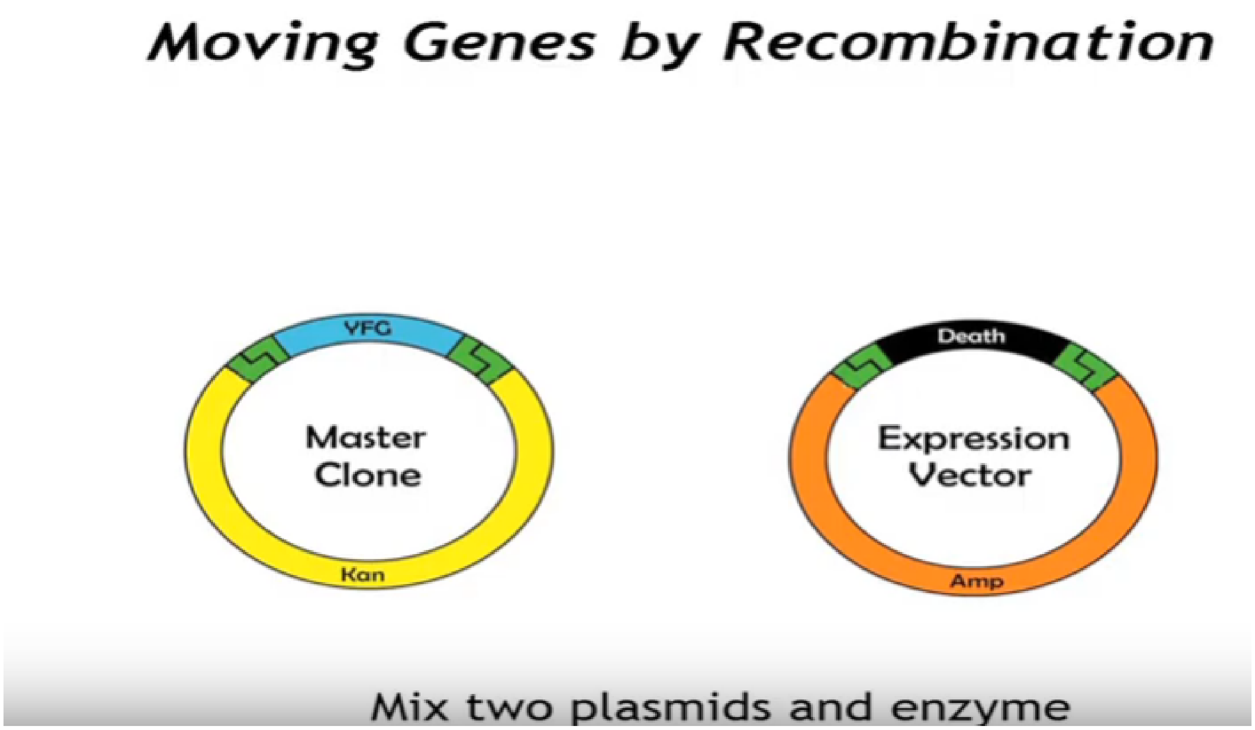


Protein Expression Clone Repositories

- 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'm trying to get you to protein microarrays but, before we can get to protein microarrays, we have to talk about, where the, the genes come from to make those arrays, how are you going to make all those proteins if you don't 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. Now, of course in the perfect world we'd have one copy of every splice form of every gene. But, at the very beginning let's 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 Rowling a cell setting. So, you had to, you had to have a format that was flexible and to get to flexible, we, we focused on this technology called gateway recombination.

Refer slide time: (14:39)

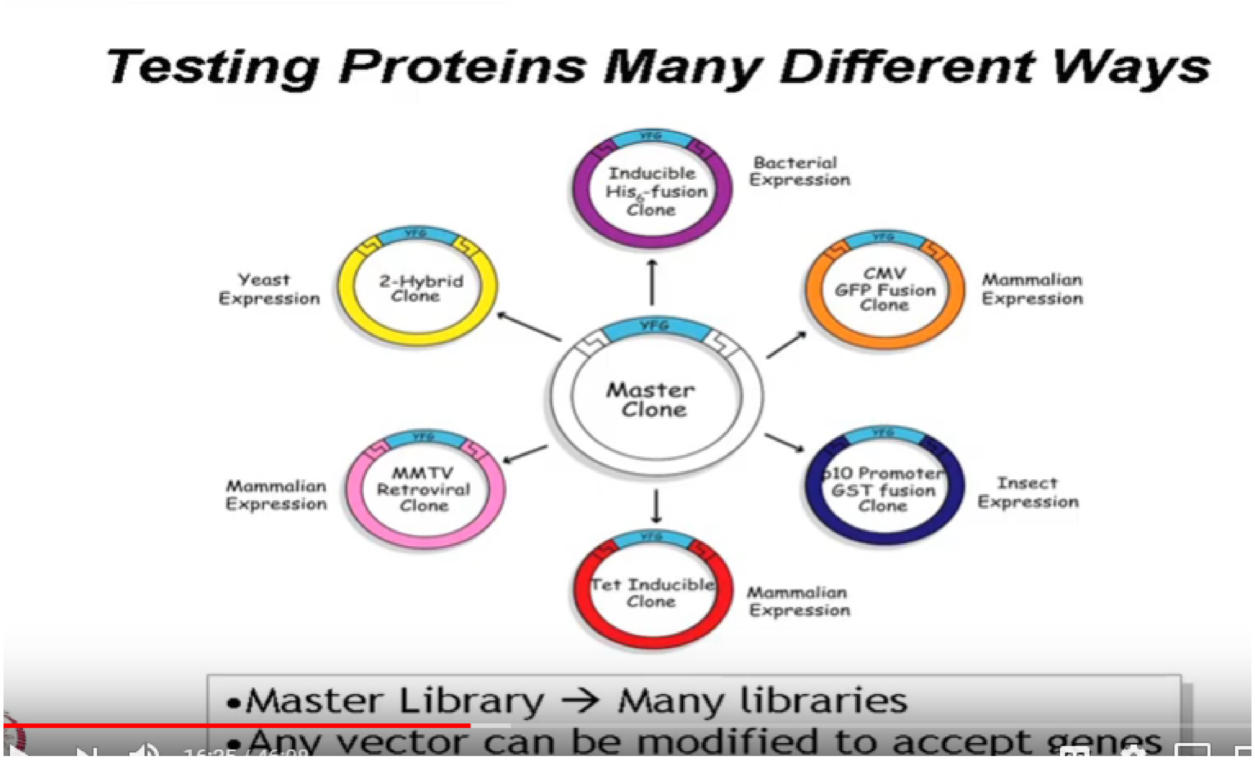


Moving Genes by Recombination

Mix two plasmids and enzyme

How many of you familiar with gateway? Not. So, many yet Okay? well now imagine doing restriction Digests, for every gene in the human genome, it gets to be a little complicated because, you'd 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 gonna start cutting up the proteins, into pieces. And then you're gonna have to reassemble them, or you're gonna 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's, it's essentially a type of recombination 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 recognized, by an enzyme system from phage lambda and so ,you can simply mix this plasmid plus that, plasmid in solute in, 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 death cassette in this guy, the only viable product is this one. It’s the only one that survives, and when that’s the only one that survives, now you can essentially develop a method for doing this operation in high-throughput, you can move thousands of genes, all, all, all by automation, and I'll show you that in a moment.

Refer slide time: (16: 16)

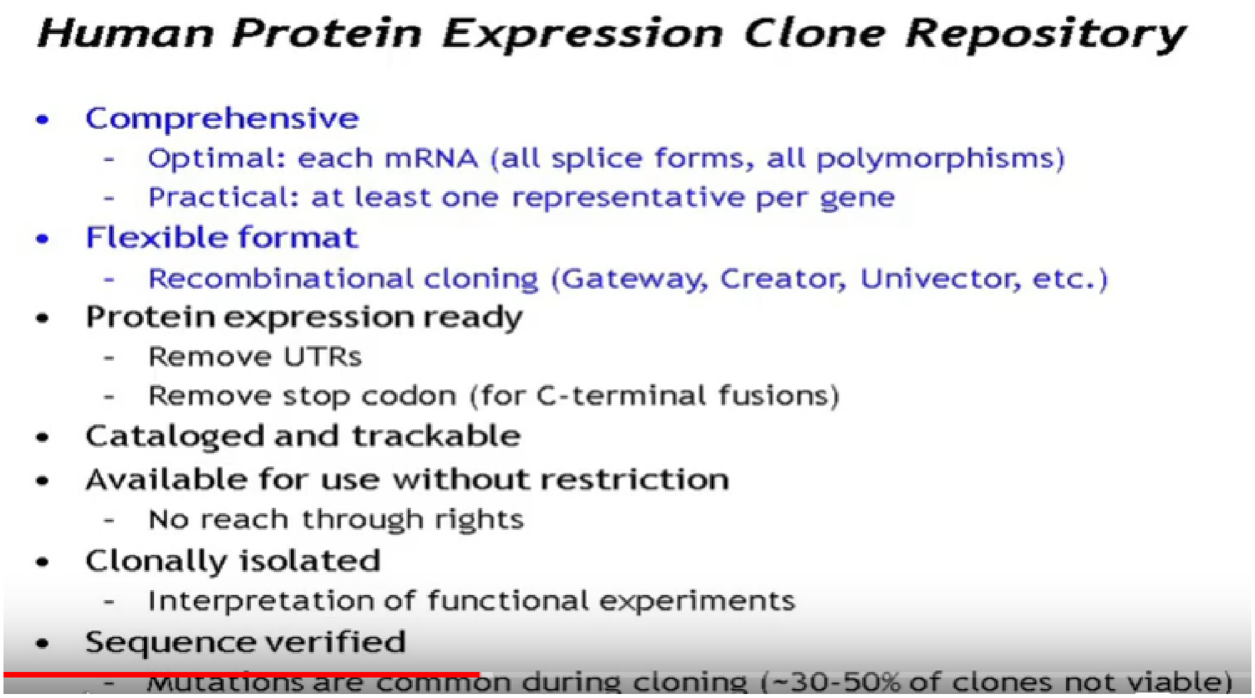


Testing Proteins Many Different Ways

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

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, we move thousands of genes from one vector to another. Okay? Another thing that you want if you're going to make these clones properly

Refer slide time: (16: 48)

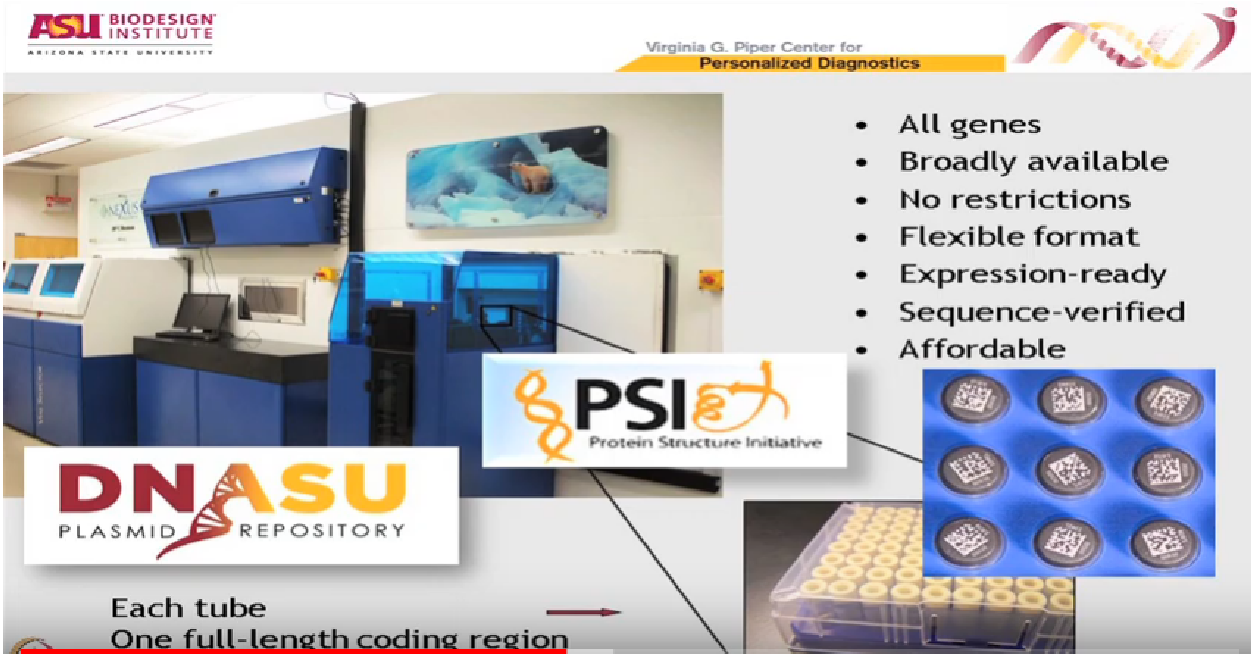


Human Protein Expression Clone Repository

- 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)

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 epitome tags, remember we said, we want to be able to put tags on these proteins ,if there's a stop codon present then, then when you translate the protein it will stop, at the stop codon and it won't 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 doesn't work at all if it's 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's, it's the molecular version of building a library. Right? U,u,u ,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 want to make these available to everybody. So, if you're gonna make a library ,of all of 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've done molecular biology you know, that when you make molecules, sometimes you get a mixture and a mixture is useless ,if you're trying to do experiments, where you know? what you're 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 Okay? And that's 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 doesn't work .Okay?

Refer slide time: (18:46)



ASU BIODESIGN INSTITUTE ARIZONA STATE UNIVERSITY

Virginia G. Piper Center for Personalized Diagnostics

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

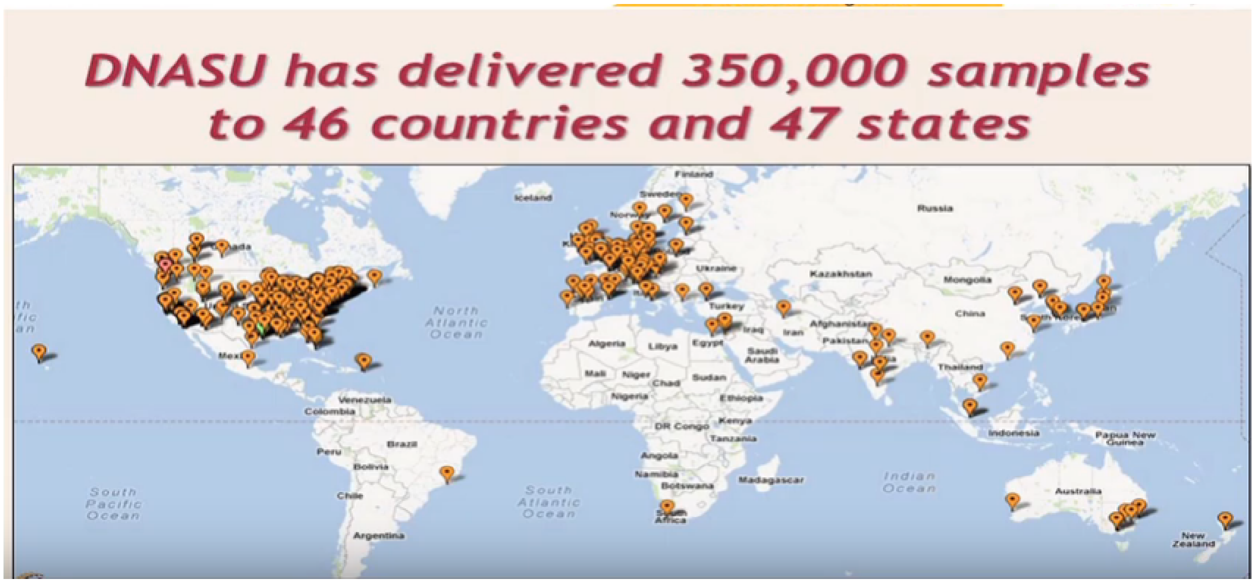
PSI Protein Structure Initiative

DNASU PLASMID REPOSITORY

Each tube
One full-length coding region

so, here, here's 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, well the good news is, that, that, that dream is now ,becoming a reality that's, this is what it looks like today .what you're looking at here? is a two-million-dollar freezer ,it's 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 bar 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 bar codes 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 DNA su, and I 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 three hundred and thirty thousand, unique plasmids in our collection. So, a very large collection of plasmids and they're all available to all of you, they're available to everybody on, on the planet we, we ship them every, we ship them every day.

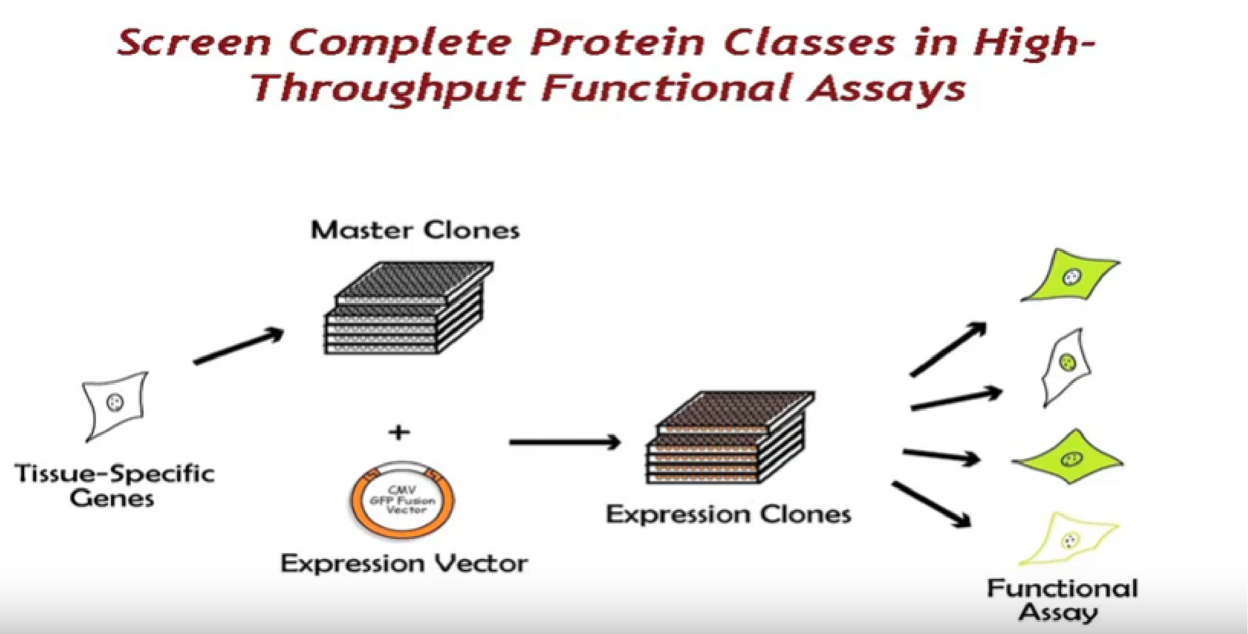
Refer slide time: (20:45)



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

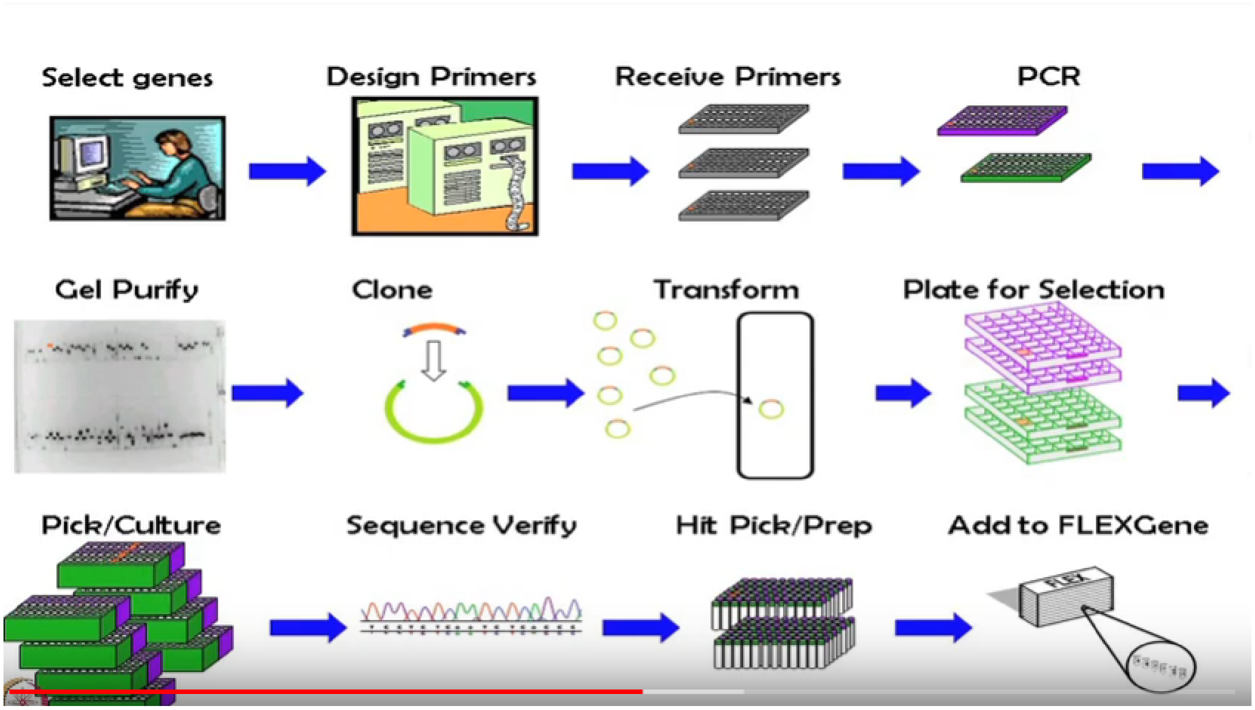
in fact, I think we have shipped ,over 350 thousand samples worldwide now ,now they're not all human some of them are other organisms, they're not all in gateway but, these are all plasmids that we are made, and where other people have made and given to us to share with them for uses in all kinds of experiments. So, what does this allow you to do? If you have all these different clones, for all these protein genes, well imagine that you wanted to look at it do a study of a set of genes.

Refer slide time: (21:14)



Screen Complete Protein Classes in High-Throughput Functional Assays that are unique to a particular tissue, maybe you're looking at neurological systems because, you're studying brain tumors, or you're looking at liver cells because, you and you want to look at genes expressed in livers in, 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's say you put them into cells and do some kind of functional assay ,and ask where do these proteins localized, 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: (22:03)



Select genes

Design Primers

Receive Primers

PCR

Gel Purify

Clone

Transform

Plate for Selection

Pick/Culture

Sequence Verify

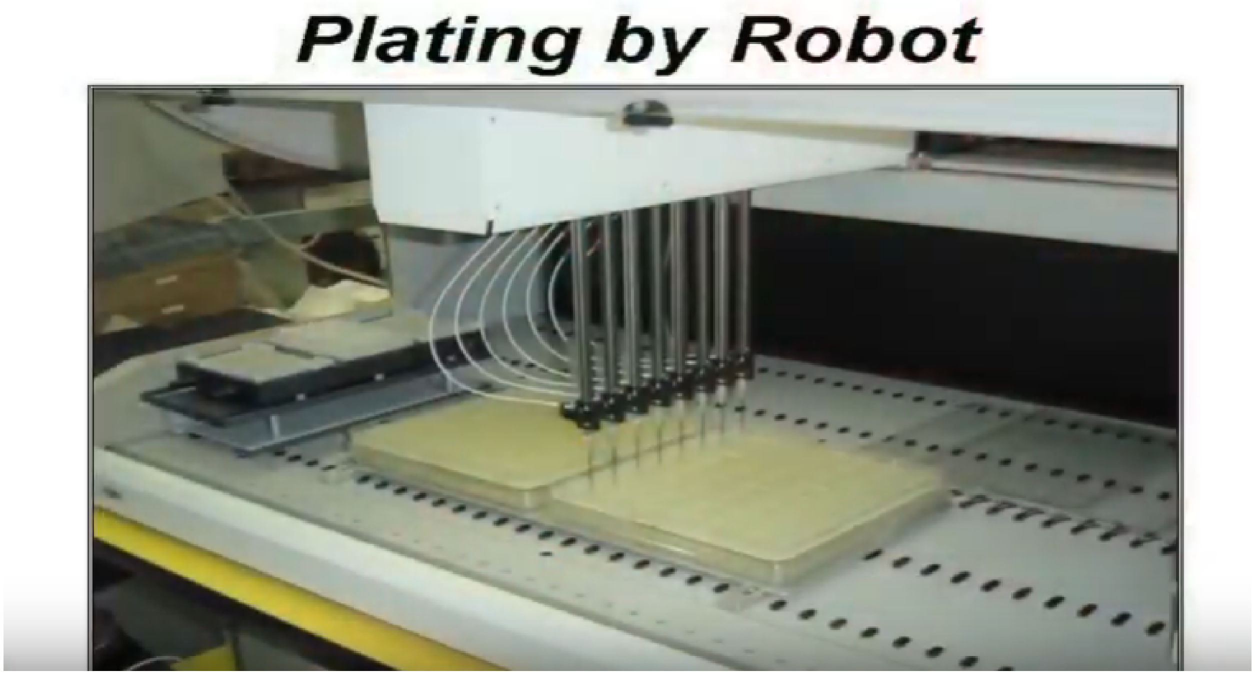
Hit Pick/Prep

Add to FLEXGene

So, I'll tell you a little bit about how we make these clones, we still do that, we're still trying to finish the human library, we've got now, almost 15,000 unique human gene clone ,that's well on the way to getting to the, the unique set that we're aiming for is around 18,000. So, we're very close to getting the full, the full set. The process looks a little bit like this is a an overview ,I will admit that it's altered a little bit in recent years, and I'll 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, 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 recombination cloning system ,transform them into bacterial plate them, pick them for culture and then sequence them to make sure that they're correct. Now, I will mention a couple of things that we do, nowadays a little bit differently ,one thing that we're 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-gen sequencing which wasn't 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 CVS, of the clones now there's a trick there's a problem with that. Right?

And the problem with that is, that when, when, when you, when you, do next-gen sequencing, you can't tell, which clone, a particular sequence comes from right. next-gen is just all the sequence that's 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 won't know, which clone that 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 won't know which it came from so, whenever you mix these clones you have to do so using informatics approaches, upfront that makes sure that they're 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 can't 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, 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're building a library, where you can go and get a specific gene, from a specific tube anytime, you want it. So, we spend a lot of time thinking about that .here's, some other the automation that we use.

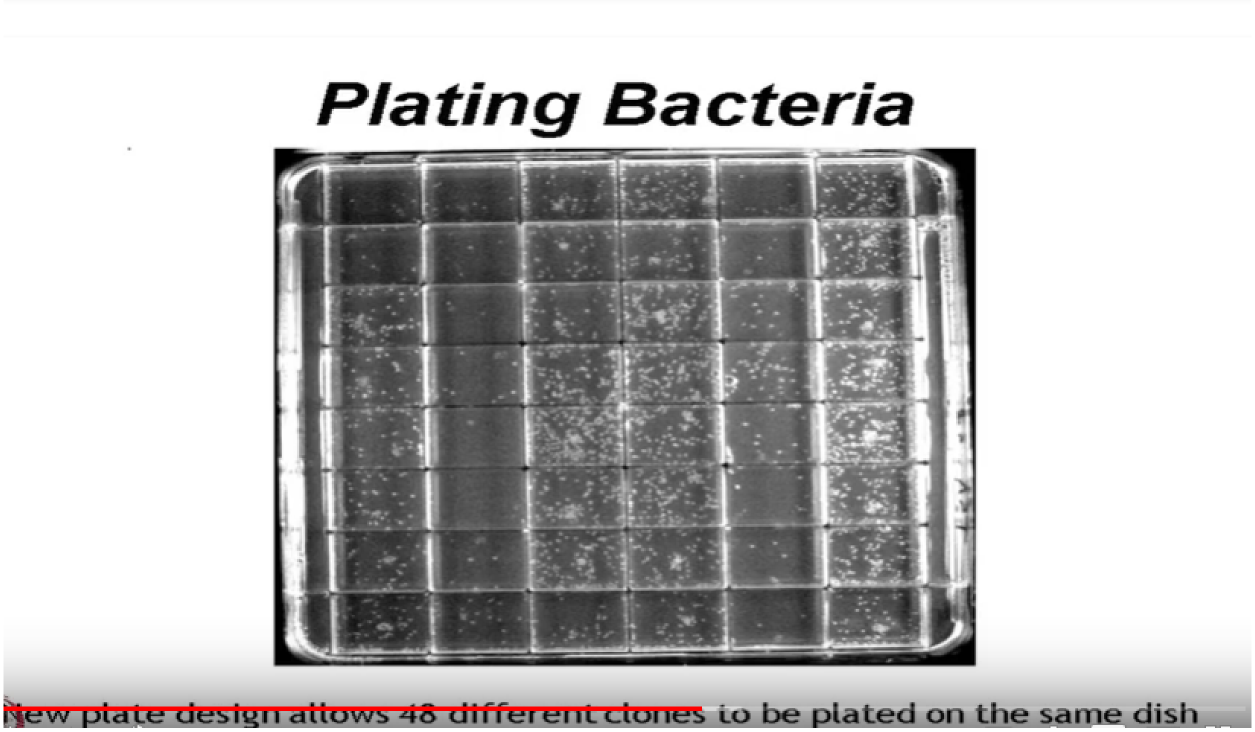
Refer slide time: (25:43)



Plating by Robot

this is a robot, it's we've transformed bacteria with DNA member, I told you ,we'd transform the bacteria with the DNA, we picked each of these different Wells and we've played them on these specialized plates.

Refer slide time: (25:55)

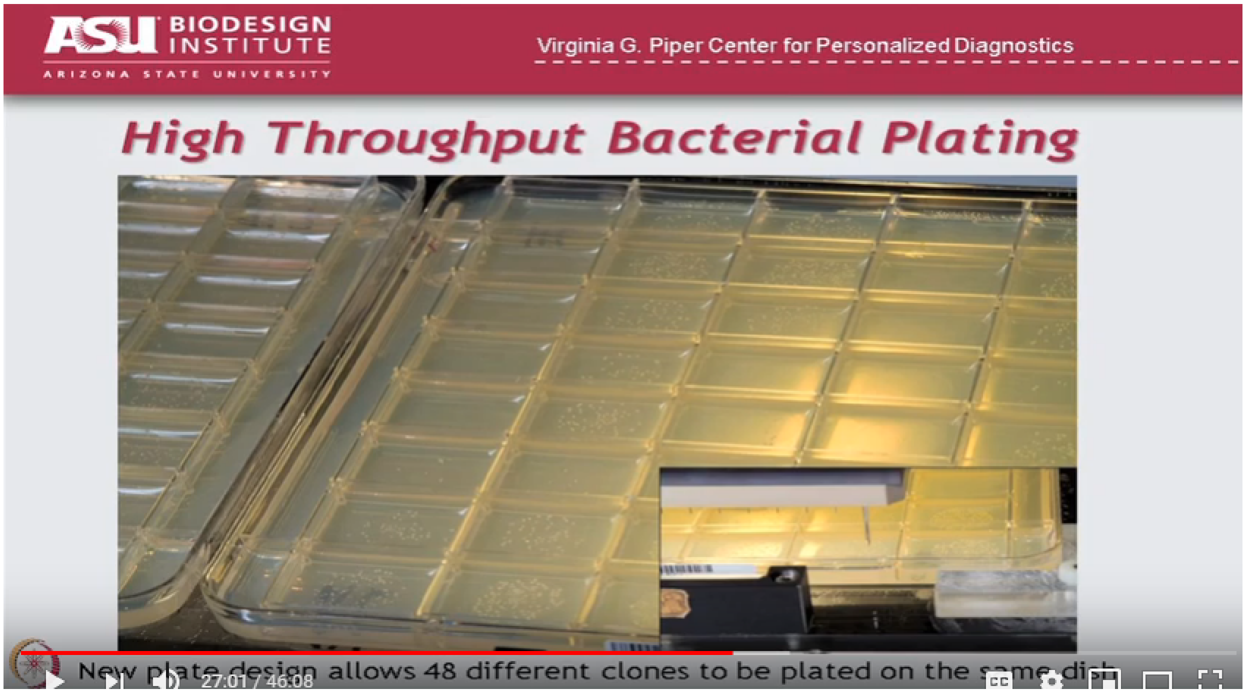


Plating Bacteria

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're these bioassay dishes ,they're 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.

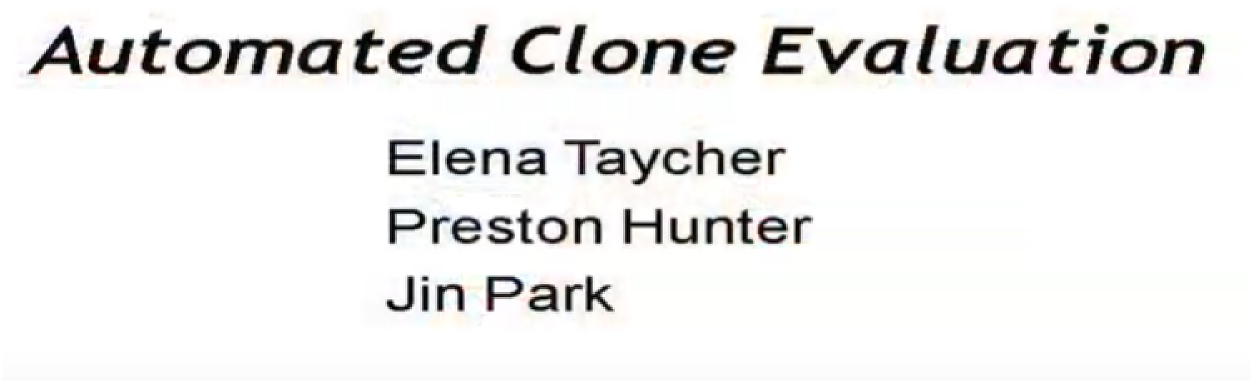
Refer slide time: (26:19)



High Throughput Bacterial Plating

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

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 ,and so, our error rate was around 15%, since then we now have robots to do this, robots don't take coffee breaks ,robots don't forget where they were and ,robots can work for many, many hours without getting tired .so, you see here's ,here's the robot and there's a little pin coming down here ,and that's gonna 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've 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're correct. How are you gonna make sure? that the gene that you have in that in that well is correct and all the sequences are right, or if they're not right, how can you document that they're 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?
Refer slide time: (27:48)

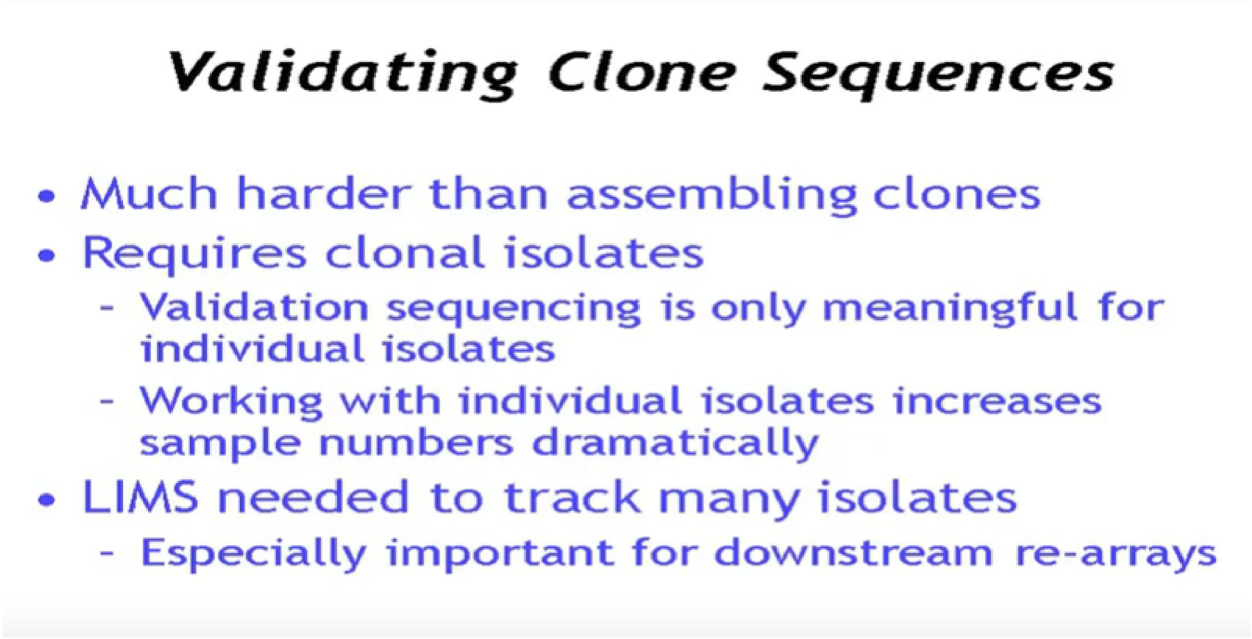


Automated Clone Evaluation

Elena Taycher
Preston Hunter
Jin Park

Or, you could get clever and you could develop a software tool to do that, and that's 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? So, I will tell you a few features of validating clone sequences.

Refer slide time: (28:06)

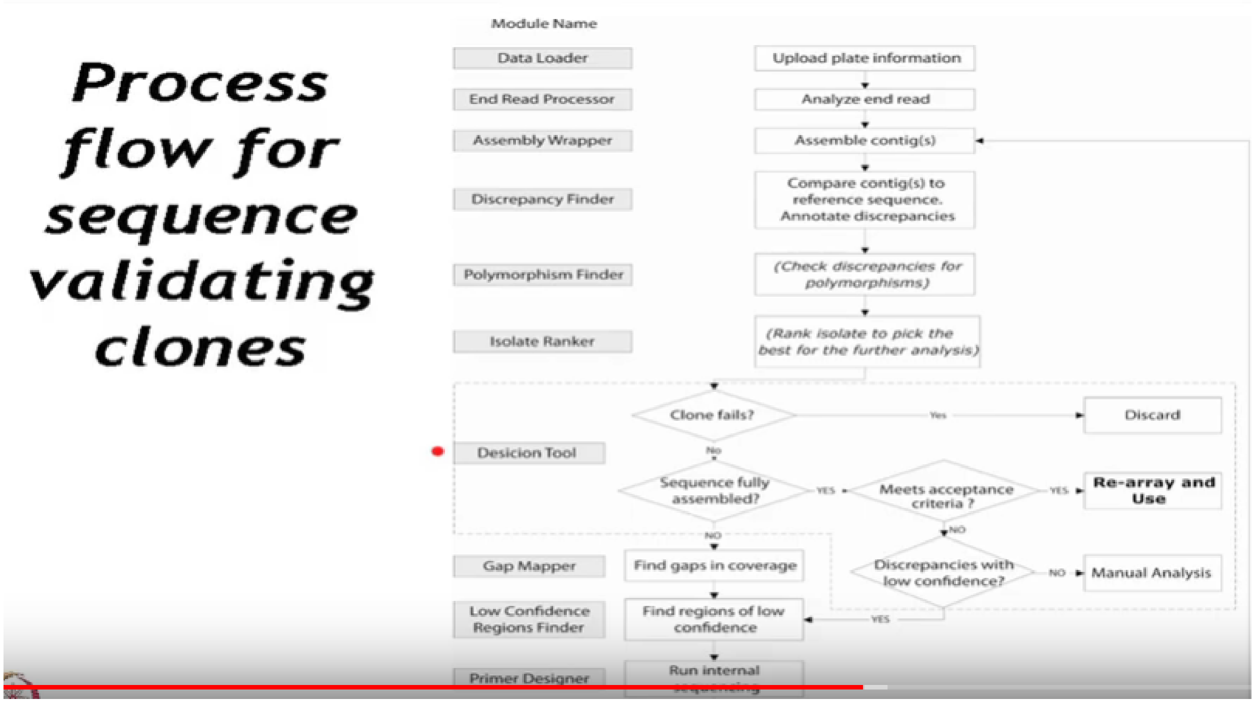


Validating Clone Sequences

- 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

first of all, much harder than actually making the clones, making the clothes is relatively straightforward, it's a lot of molecular biology steps, you can do it it's 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're sequencing a mixture of things because, as we said, earlier if there's a mixture you'll never know which one is correct and which one is wrong. Right? And so, but of course when you're 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's called a lim system, are you guys familiar with the term, lim system, li m laboratory information management system, what that does? is it's, it's an automated software application, that’s going to you manage all of the steps in your laboratory, it's 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’re, you're always tracking using informatics where things are located. So,

Refer slide time: (29: 28)



Process flow for sequence validating clones

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: Find regions of low confidence

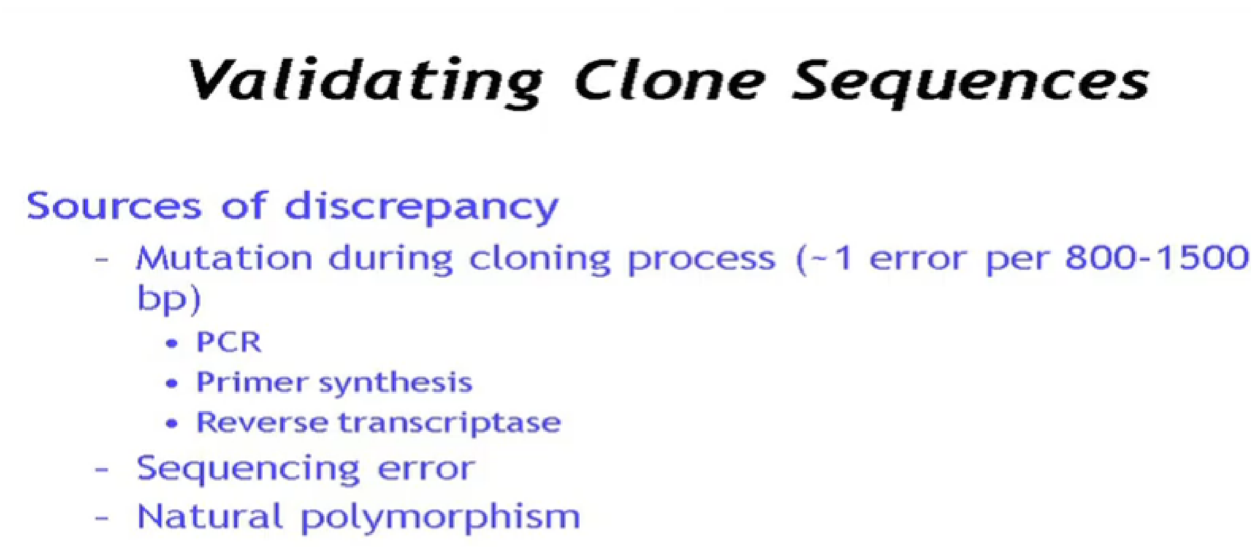
Gap Mapper — Find gaps in coverage

Low Confidence Regions Finder — Find regions of low confidence

Primer Designer — Run internal sequencing

So, this is the, the flow process, that we used for sequence validating our clones, it began, it begins by loading up the plate information that's 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, do you note an end, read is it's 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's the same primer for every gene in your collection, because it doesn't 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, 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, we, we look for what are called discrepancies, and I'll come back to what I mean by discrepancies in a moment, we then make sure that they're 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'd 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've 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. Okay? I won't go into too long. So, let me tell you about that just when I, when I mean by the discrepancy finder.

Refer slide time: (31:18)



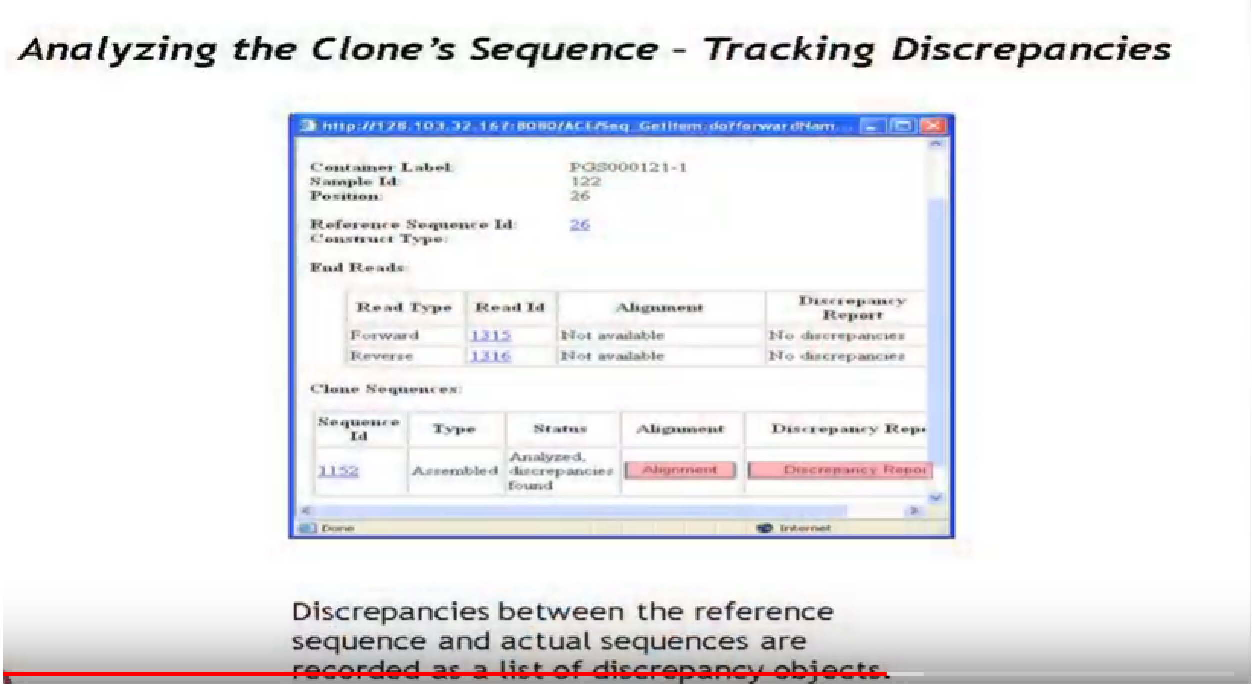
Validating Clone Sequences

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, what are the reasons, that a clone Sequence, doesn't match the correct or the expected sequence, turns out that there's 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's no longer useful. Right? Because, now we're not looking at biology, we're looking at mutants, but a much more common reason, why the clone sequence doesn't 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's incorrect but, it's not the clones problem, it's the sequencing problem, it turns out that sequencing errors can occur, as often is one in a hundred bases. So, if it's happening 100 bases ,and your clone is a thousand bases long, there's a good chance you're gonna 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 don't 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: 22)



Analyzing the Clone's Sequence - Tracking Discrepancies

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:

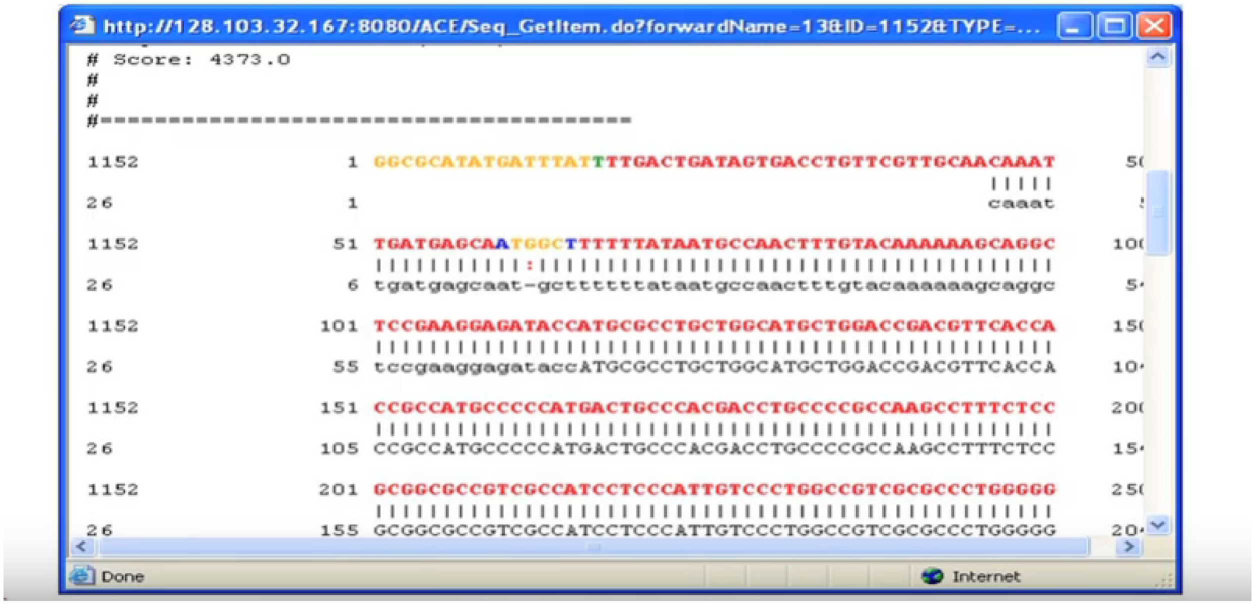
Sequence Id	Type	Status	Alignment	Discrepancy Repo
1152	Assembled	Analyzed, discrepancies found	Alignment	Discrepancy Repor

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: (33:35)

Analyzing the Clone's Sequence - Tracking Discrepancies

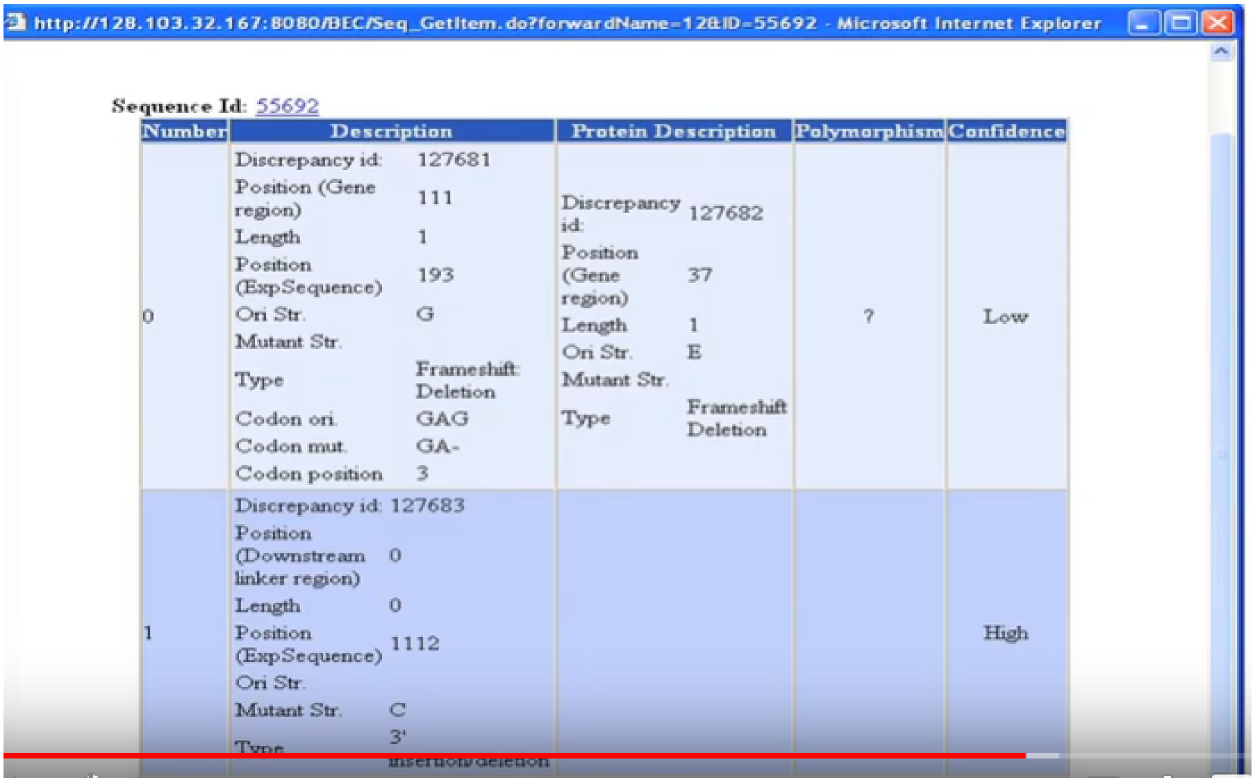


http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardName=13&ID=1152&TYPE=...

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

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

1152                 51 TGATGAGCAATGGCTTTTTTATAATGCCAACTTTGTACAAAAAAGCAGGC    10
                        |||||||||||:||||||||||||||||||||||||||||||||||||||
26                    6 tgatgagcaat-gcttttttataatgccaactttgtacaaaaaagcaggc     5

1152                101 TCCGAAGGAGATACCATGCGCCTGCTGGCATGCTGGACCGACGTTCACCA    15
                        ||||||||||||||||||||||||||||||||||||||||||||||||||
26                   55 tccgaaggagataccATGCGCCTGCTGGCATGCTGGACCGACGTTCACCA    10

1152                151 CCGCCATGCCCCCATGACTGCCCACGACCTGCCCCGCCAAGCCTTTCTCC    20
                        ||||||||||||||||||||||||||||||||||||||||||||||||||
26                  105 CCGCCATGCCCCCATGACTGCCCACGACCTGCCCCGCCAAGCCTTTCTCC    15

1152                201 GCGGCGCCGTCGCCATCCTCCCATTGTCCCTGGCCGTCGCGCCCTGGGGG    25
                        ||||||||||||||||||||||||||||||||||||||||||||||||||
26                  155 GCGGCGCCGTCGCCATCCTCCCATTGTCCCTGGCCGTCGCGCCCTGGGGG    20
```

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'll 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, some, some mistakes that come up, and then and then here's.

Refer slide time: (33:07)



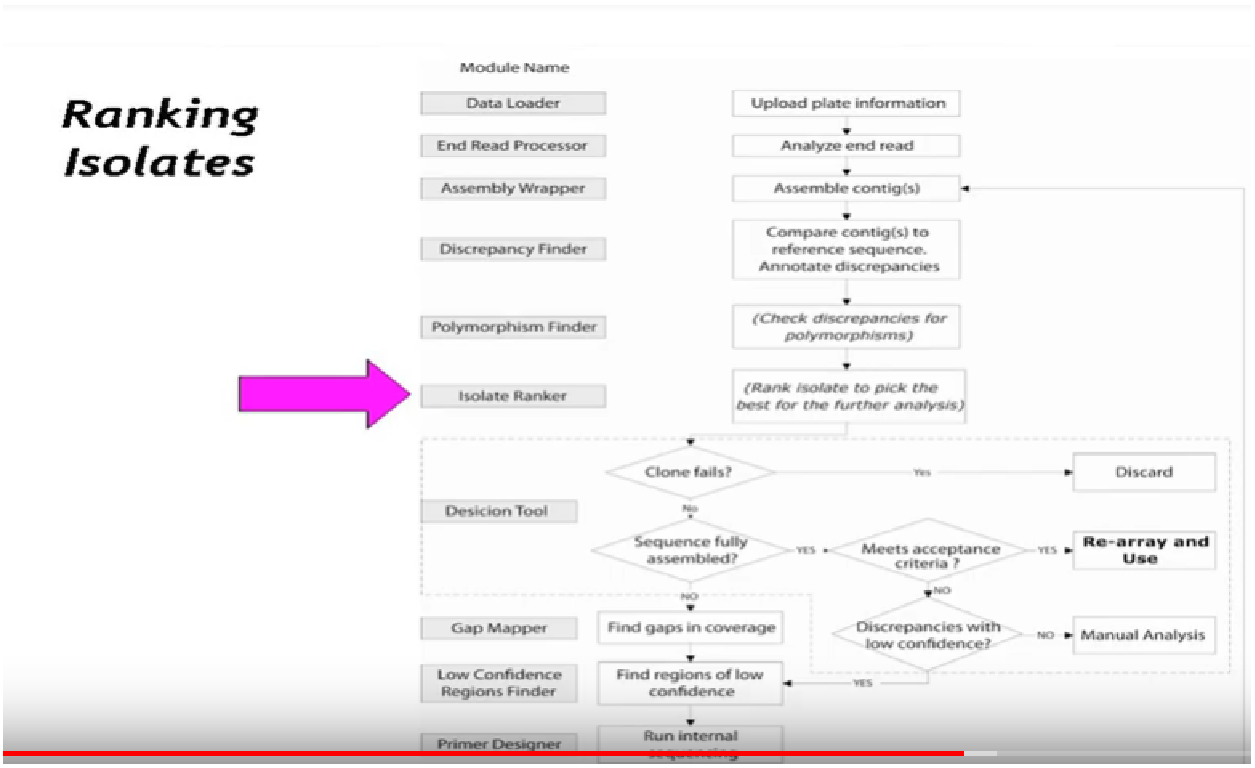
http://128.103.32.167:8080/BEC/Seq_GetItem.do?forwardName=12&ID=55692 - Microsoft Internet Explorer

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

What we this is? if you click on the discrepancy button, you'll get this report, and it will tell you every time there's 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's a frame shift deletion, that means that where we've 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's what happened in this case. Right? Obviously mutations that cause, profound changes like that, are much more Dilla tyria sin our clones than, then simple substitution mutations.

Refer slide time: (34:58)



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

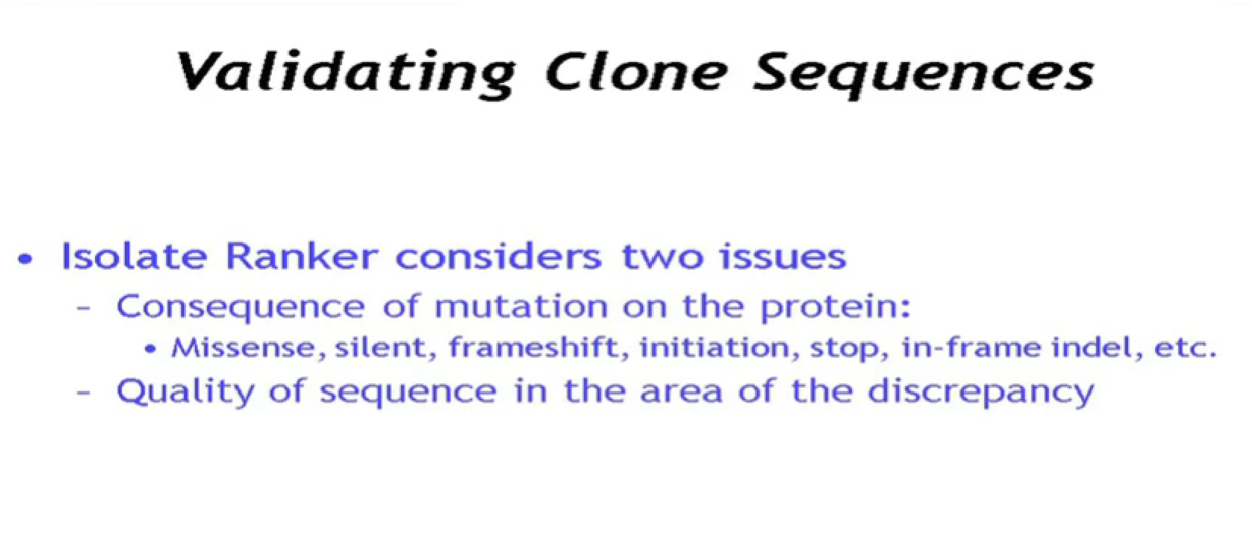
Gap Mapper — Find gaps in coverage

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, 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: (35:17)



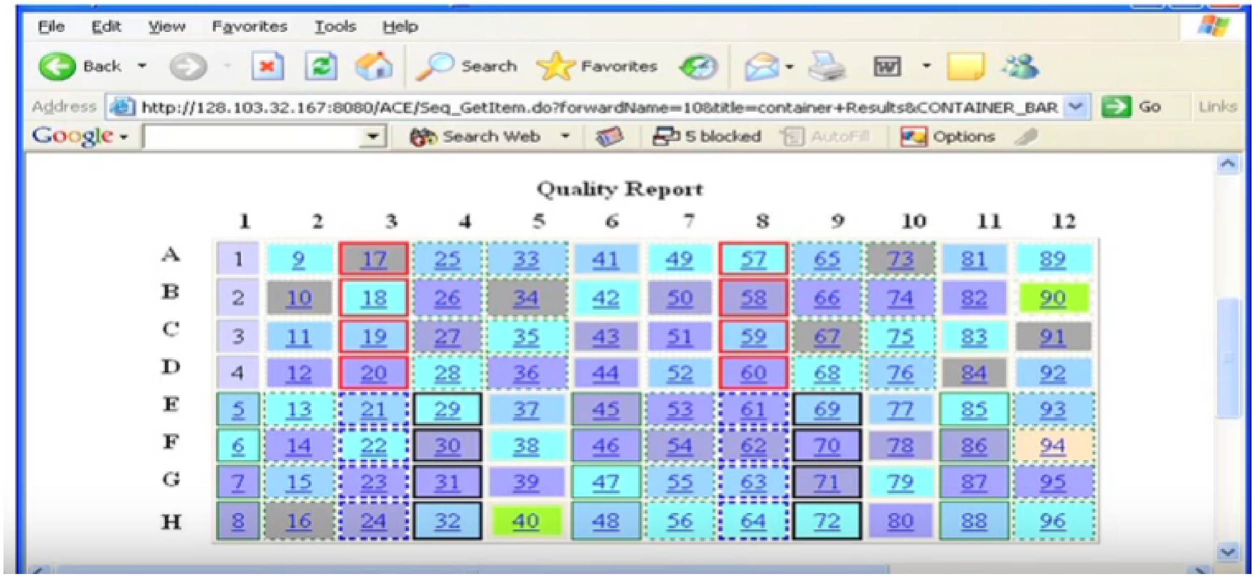
Validating Clone Sequences

- 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'm gonna there's a very good chance that the mutation is due to bad sequencing and not could, not, not the actual mute.

Refer slide time: (35:40)

Ranking Individual Isolates



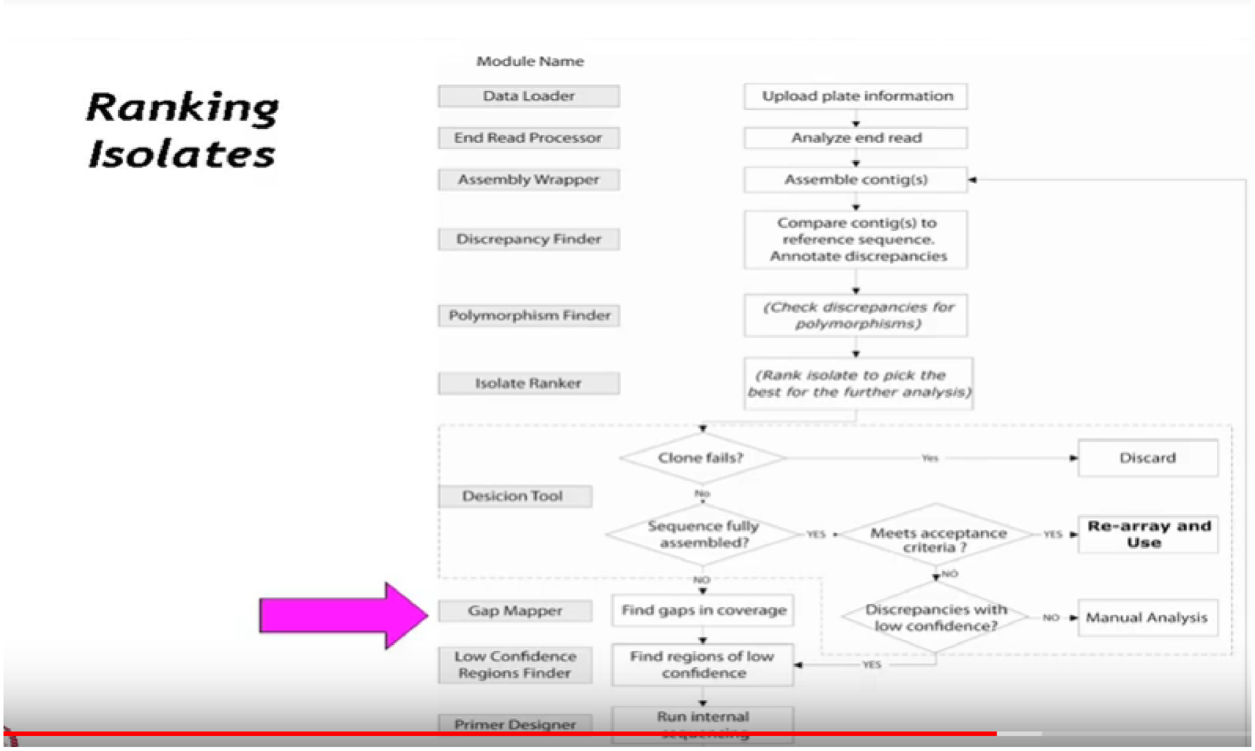
http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardName=10&title=container+Results&CONTAINER_BAR

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'll 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.

Refer slide time: (35:51)



Ranking Isolates

Module Name

Data Loader

End Read Processor

Assembly Wrapper

Discrepancy Finder

Polymorphism Finder

Isolate Ranker

Desicion Tool

Gap Mapper

Low Confidence Regions Finder

Primer Designer

Upload plate information

Analyze end read

Assemble contig(s)

Compare contig(s) to reference sequence. Annotate discrepancies

(Check discrepancies for polymorphisms)

(Rank isolate to pick the best for the further analysis)

Clone fails?

Yes

Discard

No

Sequence fully assembled?

YES

Meets acceptance criteria ?

YES

Re-array and Use

NO

NO

Discrepancies with low confidence?

NO

Manual Analysis

YES

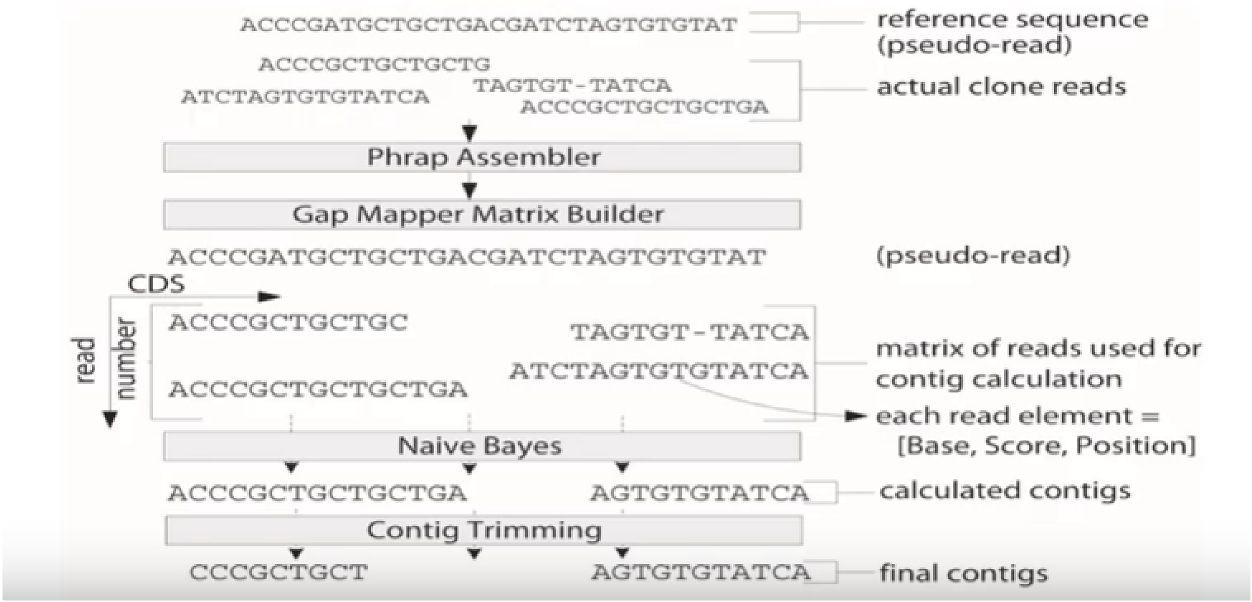
Find gaps in coverage

Find regions of low confidence

Run internal sequencing

and then this, this last tool I'll mention here, is the gap mapper and I remember, I told you ideally, we have sequence for the entire gene, if we don't have sequence for the entire gene, we need to go back, and get an additional read to fill in the gap otherwise, we can't say with certainty, that we have a good clone.

Refer slide time: (36: 11)



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

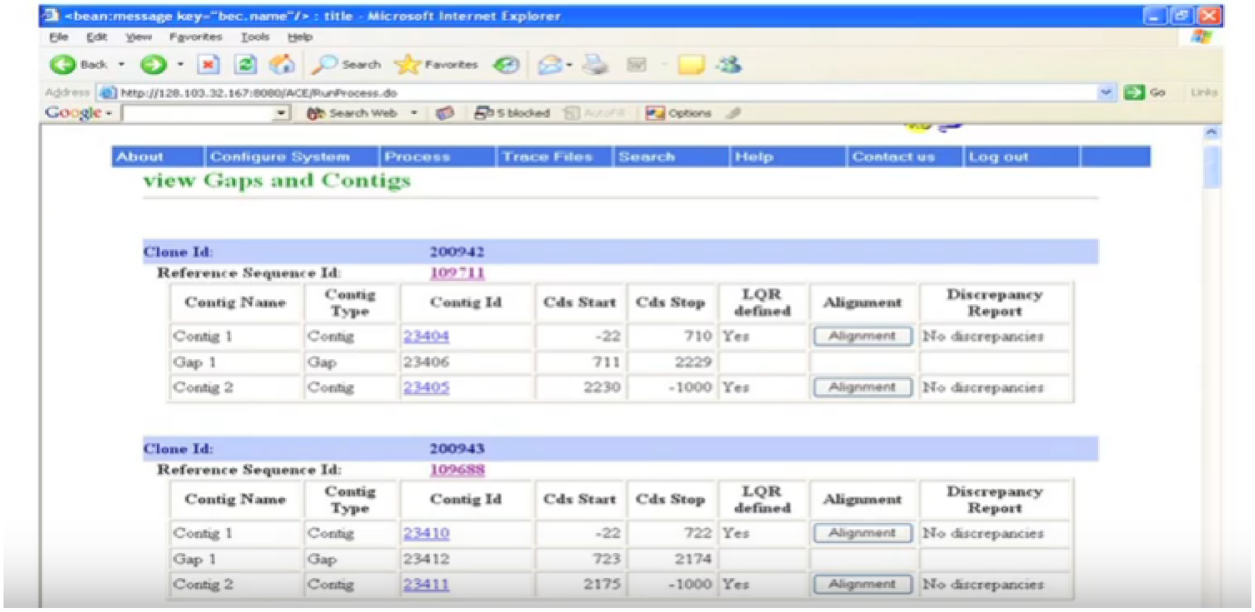
AGTGTGTATCA

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 our 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: (36:42)

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

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

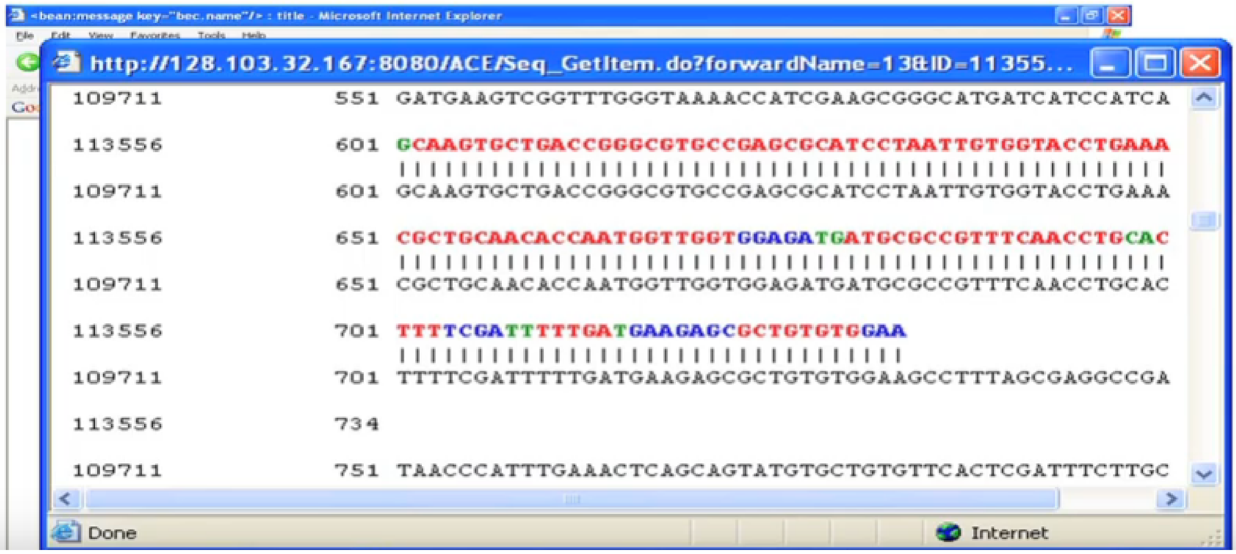
Clone Id: 200943
Reference Sequence Id: 109688

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

And then this is what that this is? What it looks like in our software.

Refer slide time: (36:46)

Finding Gaps and Low Quality Regions

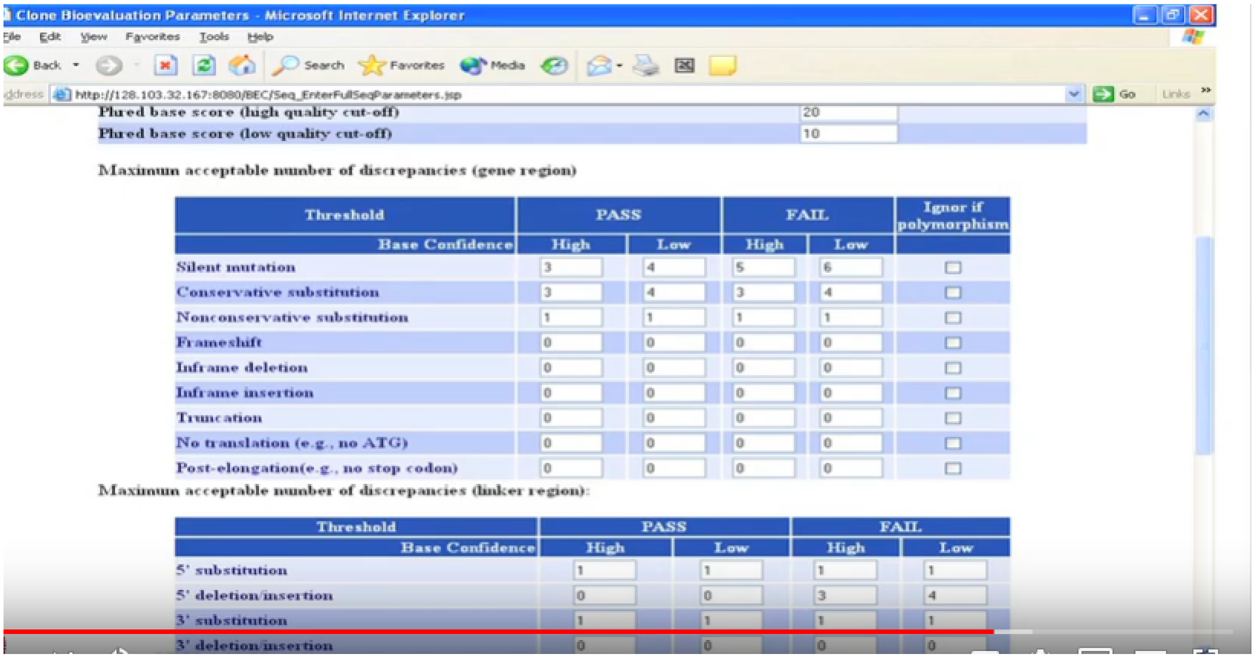


<bean:message key="bec.name"/> : title - Microsoft Internet Explorer
File Edit View Favorites Tools Help
http://128.103.32.167:8080/ACE/Seq_GetItem.do?forwardName=13&ID=11355...
Done
Internet

```
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
```

And so, you can see it basically predicts that there's 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: (37: 02)



Clone Bioevaluation Parameters - Microsoft Internet Explorer

http://128.103.32.167:8080/BEC/Seq_EnterFullSeqParameters.jsp

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

Maximum acceptable number of discrepancies (gene region)

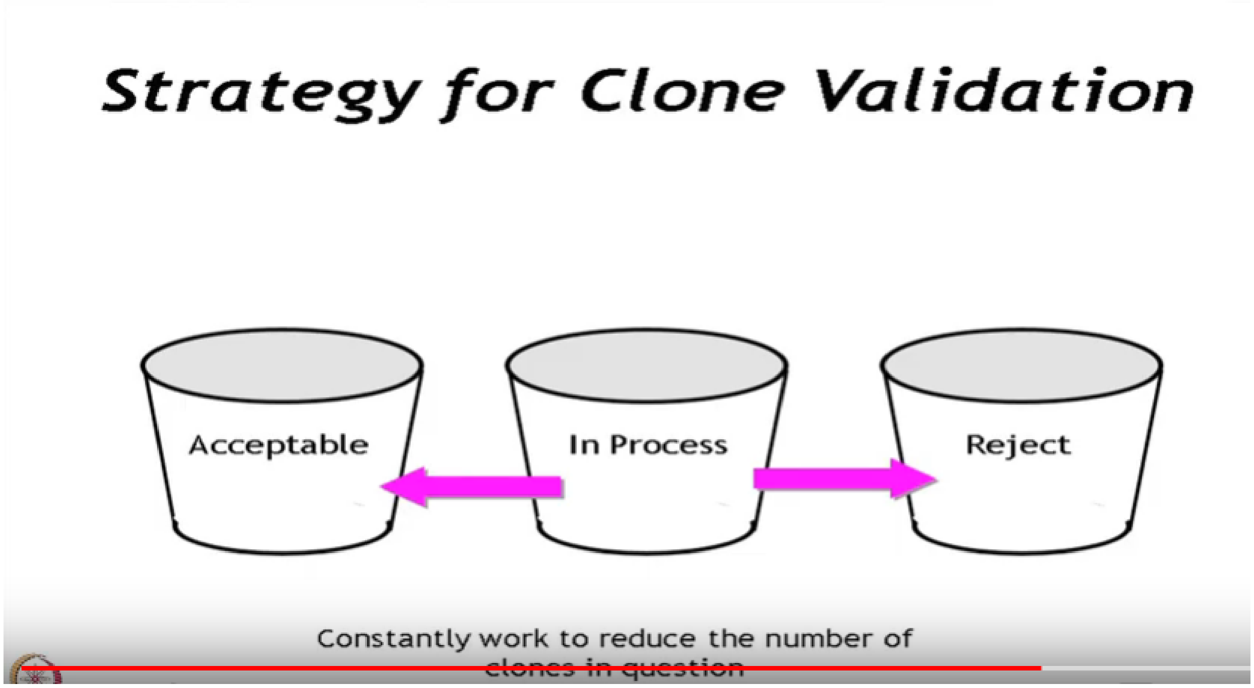
Threshold	PASS		FAIL		Ignor if polymorphism
Base Confidence	High	Low	High	Low	
Silent mutation	3	4	5	6	□
Conservative substitution	3	4	3	4	□
Nonconservative substitution	1	1	1	1	□
Frameshift	0	0	0	0	□
Inframe deletion	0	0	0	0	□
Inframe insertion	0	0	0	0	□
Truncation	0	0	0	0	□
No translation (e.g., no 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, is always to, 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: (37:32)



Strategy for Clone Validation

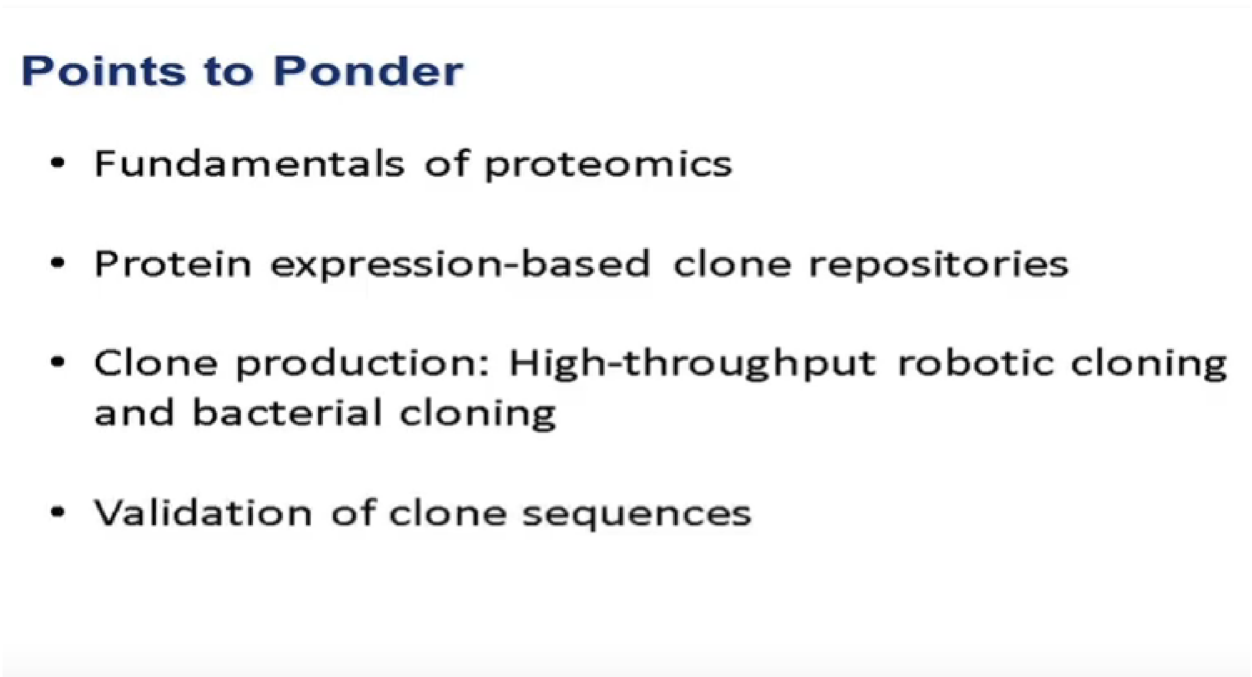
Constantly work to reduce the number of clones in question

And so, as I say as we run through our clone list, at any given time, we're always trying to move clones either into the reject category, or the acceptable category. Alright? So, that let me stop there and see if there are any questions on, on the cloning process of making clones for collections, there are any questions? I can, yeah .the question was, what’s the mechanism a sequencing error? That depends a little bit on what platform, that you're 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, you can't really tell is, it an A or is, it a T I'm not sure sometimes ,it's just that you don't get adequate coverage. So, you don't read as many times past that base. So, there's a lot the men then, method the, the chemistry themselves have errors, lately we're using alumina, which is next-gene sequencing, it also has an error frequency but ,it typically with Illumina sequencing, you get around that by doing. So, many reads you cover it 30 times, that you're less likely to have an error. But, there’s, it’s the process itself is error-prone. Other questions?

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, at NCBI in Butte, in the US ,in the unipro gene sequences all that stuff, they have errors in them and ,and that is and so, if we disagree with that it's not always clear, that it's us, that's at fault, typically in a lot of cases in our in our circumstance well there's - let me just say there's two ways that we've 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're trying to achieve and we try to match that sequence, in the in the case you're referring to where we're trying to match a sequence in a database, we actually did develop a polymorphism, tool and I ,I had slides ,on that and I took them out because, it was gonna get too long, but basically what the polymorphism tool does is, it goes out to all the existing databases where there have been gene sequences uploaded for, 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'm more likely to accept the sequence, it's not perfect but it does help. Okay? Sit, sit again, well you know once ,we've once ,we've done that sequence validation ,we think most people don't have to do it again, I mean it's certainly reasonable if you want to be extra careful if it's a very special clone for a research project of yours but, for high-throughput of materials, we've done a pretty good job of sequencing these so, I don't think you have to repeat that ,and I should point out that one of the qualities of the ,the Gateway process which is to transfer ,the insert from one master clone ,to an expression clone, that's a conservative molecular process. So, once do you know that this sequence is correct, and then you know that this sequence is correct? So, you don't have to resequence them both, yes. in fact are all of our clothes if you go to our website, the DNA z website, we list the actual sequence, of that clone.

So, we've done the sequence and we've loaded that up on the database, I think there was one over there, yeah. so ,in the 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's more than two amino acids difference, that we don't 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 use it. But, most of the time it's either 100% accurate or we allow one amino acid change .Okay? Well the genome hasn't been sequenced; it's 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 we'll 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% ,it was horrible and, 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% accuracy. so ,you ,you really have to have a good quality genome sequence, to do this kind of work.

Refer slide time: (44:54)



Points to Ponder

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

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, we 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 steps in the entire high-throughput gene cloning pipeline will continue more discussions in the next lecture, thank you.