HANDOUT

LECTURE-31

MICROARRAY WORK-FLOW: IMAGE SCANNING AND DATA PROCESSING

Slide 1:

This module contains the summary of the discussion with Mr. Pankaj Khanna, an application specialist from Spinco Biotech, a distributor for the Molecular Devices product, the GenePix microarray scanner. Mr Khanna will provide an overview of microarray image scanning using the GenePix Microarray scanner and subsequent data analysis using the associated software programs, GenePix Pro Image Acquisition and Analysis and Acuity. The following basic concepts will be covered:

- Various scanning parameters including types of photomultiplier tubes, filter selection,
- Ratio image biology
- Comparison of test to the reference samples
- Data analysis including definitions of Gal files, gene array lists, .

Image processing involves various steps:

- Identification of the spots and distinguishing them from the background signal.
- Determining the spot area ,
- Designating a particular area as background.
- Reporting the summary statistics
- Assigning the spot intensity after subtracting it from background intensity.

An overview of a basic microarray experiment (be it protein or DNA) is as follows:

1. Manufacture of the array

2. Determination of experimental design: This step involves tailoring conditions to best address the biological question under study, designating the reference and control groups, etc
3. Sample collection and preparation.
4. Hybridization of sample to pre-planned chip
5. Slide scanning
6. Data analysis, which mainly involves bioinformatics and biostatical analysis

- In terms of maximal capability these microarray scanners scanners can go upto four internal lasers, 16 filters in resolutions of up to 2.5 micron, and accommodate a very wide variety of slide types and materials. They represent the best hardware in this market, for upto 2 dye (usually Cy3 and Cy5) scanning.
- Scanning can occur in 2 modes- non-confocal microarray scanning, or by inverted chemistry scanning
- While confocal scanning is useful for thick images, on the planar microarray slide, non-confocal scanning works well (as, in theory, the spots occupy a single plane), to basically perform a 2-D scan It should be noted that, while scanning thick sections, such as tissue sections, a technology called confocal spectroscopy is used, to generate a 3-D image.
- But for microarray slide scanning, where all the spots are on the same plane, confocal scanning, in combination with inverted chemistry, helps produce well resolved images with a good signal to noise ratio. .
- How does non-confocal scanning coupled with the inverted scanning help? It helps in dealing with deformities and helps deliver best signal to noise ratio so that your data is more valid for any analysis. In other words, this process can actually correct any misalignment which may happen due to the design of the slide and scanning procedure.
- Slides usable: A wide range of slides, available from various academic as well as commercial vendors are compatible with Genepix scanners. This list includes

Agilent, nimplegen, corning etc. Furthermore, tissue arrays are also available from Corning. Overall, this platform can support a wide range of applications.

- Scanner resolution range is from 5 micron/pixel resolution to 2.5 microns/pixel resolution. The basic rule of thumb is that the size of your spot, be protein or RNA, that should be 10 times that of the resolution chosen for scanning.
- DNA spots are less than 50 microns and most of the proteins usually stand at 200 microns so essentially the rule goes the 10:1.
- Another basic rule of thumb is to avoid white spots. The white spot essentially indicates saturation and to avoid saturation, the PMT voltage can be adjusted. It should be lowered in case of saturation or increased to raise spot intensity if needed. So this is how photomultiplier tube is very essential to control the different kind of natural variability within the chip.
- Optimal scan conditions: Variations arise by different means, including technical, biological or due to printing chemistry or assay variations. Technical variations can be controlled at the level of PMT. To obtain the best signal to noise ratio, PMT voltages, and to a far lesser extent, laser power, can be adjusted. Inverted scanning as well as the non-confocal chemistry is also very useful in this regard.
- So how to quantify signal to noise ratio? This is the difference between spot intensity and background variation i.e. Signal-Background.
- Note: Background variation is the Standard deviation of the background.
- Note: During scanning to obtain spot and background intensities, multiple scans are preferred, which are then averaged. 2-3 scans are usually done to maximally reduce the signal to noise ratio.

GenePix classic 4000 B:

- This is popular since it offers 6.5 minutes of simultaneous scan. The meaning of simultaneous scan is both lasers can act at the same time. So you have very less time for scanning. Apart from that it has got two lasers, at 635 and 532 nm which is classically used for cy3 and cy5 and they are compatible dyes.

- So in view of this we have standard green and standard red filters to accommodate all cy3 and cy5 applications.
- The laser power can also be adjusted from 100 % to 33% to 10% based on what application and what intensity desired.
- So adjusted focal position is from 50 to 200 microns so this allows you to focus in different ways so we can have a different slides compatibility coming up with the scanners- 5 micron resolution maximum it allows to go for and it can go upto 100 micron resolution.
- It uses a non-confocal optical design.
- Any standard slide can be used for scanning so a wide range of application is possible.

GenePix Pro Software

- Imaging is done in the form of multi-imaging or single image tif format.
- Images are created in the Tif format and then exported as colored images in 24-bit JPEG format.
- The JPEG format is only for visualization purposes, while the basic data processing will be done on tif format.
- Lasers for image acquisition usually range from 2-4, however many applications may only use a single laser as well.
- Alignment with gal file:The features(ie the printed DNA or protein) on the array blocks need to be aligned
- Features are actual genes or proteins or the representative of biological material what you are checking.
- Now these have to be aligned with annotation information.
- Automated block and feature alignment is possible. Even the different sizes and shapes of the spot can be handled, right from the circular, square to irregular feature which can be handled at the level of image alignments.
- To sum up, in a microarray, after the image acquisition, first thing comes alignment. Then you have a background subtraction so that you get a true signal coming in.

- GenePix Pro does help in background subtractions in the different format and also in the normalization features,.
- Importance of control features: Background subtraction can be done in the form of local, global, negative and morphological controls so negative and morphological controls are subjected to the design of the slide type.
- Negative controls should be some spots which should not be bind to anything and leave blank whereas the local and global can be calculated in the general space where there are no spots available and the area which is not being spotted nearby your particular feature so in this fashion background subtraction can happen and then feature viewer and feature pixel plot so basically the major thing comes after the acquisition is visualization of the data and these visualization comes in the form of pixels and plots the graph so the graph helps us in understanding globally what is happening in short so it gives you a real image how the things are happening so this can also be done.
- And there are multiple ways of calculating in the form of ratio calculations after normalization or during the normalization of the data so analyses immediately after those involve few of the normalization process which GenePix Pro very well handles and other important feature is the flagging of the spot.
- In biology we see some spots are not really good or because of some artifacts they should not be taken for the analysis so essentially we need to flag that spot as good, bad or absent so these can also be done by GenePix Pro software. Lastly normalization of the images and the different formats is also allowed to happen in GenePix Pro.

I think you rightly mentioned the need and importance of the control features because many times the quality of the experiment depends on how well your controls are performing?

- That applied to both positive and negative response and when you talk about DNA microarray technology obviously we are talking about very high density arrays here where lot of controls and lot of spots are built in place, many times when I talk about

protein microarrays specially when we talk about functional arrays so we actually put different type of controls just based on that one particular experiment, for example I am looking for some biomarkers response I need to have certain positive controls some biomarker which need to light up on the array which will guide me as a positive control then if I am looking at immune response I need to have some sort of IgG and some of those type of control features which will guide me just how non-specific response could be so again empty spot we need some type of spots where there is no DNA or no protein is printed or no biological material is there so usually these type of controls are part of the design and that helps further for background subtraction as well as how good quality data we are obtaining.True. That is the major calculation for background work.

Visualization

- Data can be presented in the form of histograms and the scatter plots.
- These scatter plots can be plotted once against the channel types- laser 1 versus laser-2 or wavelength 1 versus wavelength 2, in classical say cy3 versus cy5. How these two things behave for me so again kind of graphs and also the images can also be exported to PDF as well as being visualized in GenePix Pro for your further screening for your QC applications.
- The basic GUI which we are going to see in few minutes, actually contains three different areas the first one is image control so there you wanted to see which kind of laser I am going to use and second one is a different feature which are being used for controlling of the image and towards the right hand side we have a pen which is allowed or helpful in hardware control.
- So the basic one in the hardware control is first as we said we can look to the auto-PMT and other adjustments which is being done for that we use preview scan.
- In preview scan there are different tabs which allows you to do true scan or preview scan and based on the particular laser which scan you are using. If you start with a preview scan you decide which pixilation suits me, which different power of laser suits me once you are able to do these decisions made you can go for your own

data scan. So it's like general scanning, even if you want to scan a sheet, first of all you want to preview it that how the overall image looks like and then since you know your experiment, you know your requirements like what wavelength need to be used, what type of fluorophore you have used then one need to optimize and correct those things using the data scan which actual scan performed and then one need to review whole things and how the slide look like.

- So there is a button towards the side where you have the control for the scanning time so as it is duo-channel that means two lasers are present in that. It is allowed to select whether you want to use one or two and then based on the one or the user application you select both the laser and then at the level of live scan you control for the PMT and whichever resolution you want to use for.
- So these are in the same software towards the right hand you see the pen where you have hardware control button there also you can look for different images which is now suiting for your own biological application.

Software Scan

- When we are acquiring the image what the intensity histogram tell us while scanning is in process and even after the scanning is done?
- How one can really ensure that the scan is good and what type of balance one need to make in that?
- So basically the preview scan when you are scanning your live data you can just switch on to histogram graph.
- There what it gives you how much green and red channels are contributing towards the intensity. So you really want that they are overlapping so they are really balancing, there could be a small variability in the beginning owing to the fact that they are just background and then the spots coming in. there you want that they are really overlapping after little bit of lag that is few seconds of lag that is it.
- So once you are able to do then see and select whether yes this PMT is good for me so this Is the way you check which PMT is more suitable. So you select the PMT,

look at them, see the overlapping where there is best overlapping without the saturation you want to go over those statistics.

- Running the Graphic user interface. So what you see now is GUI (graphical user interface) of GenePix Pro software, on top of it is in form of different tab buttons which allows you in different work groups for example 'image' allows you in different ways encountering of the image acquisition and histograms looks at how that image has performed.
- So this is what we were speaking about in the earlier slide where you can see live kind of demo which is happening and then lab book actually gives you what all have you done in different stepwise so every movement of yours is in this software is being logged in and then analysis can be done in the form of batch form which allows multiple slides so that you do alignment and do analysis which can be performed as batch analysis.
- Once analysis is over results can be seen and the scatter plot can be plotted at the level of this graphical user interface.
- Once you are through, you can look at the reports as well. So let's look at major function of imaging and how one can control for image acquisition.
- Let's quickly go through different kind of buttons here. Now the imaging can be at different wavelengths and preview can be done at 635 and 532 nm in a single laser based the wavelength can be done at 635 or wavelength at 532, even the ratio of the imaging how both has performed together can be looked at looking at button of ratio of imaging so this one allows you to see how the image is being done after the scan.
- You can look at one channel, preview channel or different channels. Now let's look at different tools which are available to you while or after the scanning so the major ones are here where you can move across the chip in form of this hand tool, plus indicates the zoom tool, and the other tools are- this is for the unzoom and you can also look at the whole image button.
- So once you have the image these tools becomes activated. So these two are for blocks and looking and controlling blocks and these ones are the features. So many

a times what happens usually get a GAL files is a feature information file. Gal stands for gene array list so actually it gives the X and Y coordinates where each array usually is being present in the form of block which each block are in terms of the feature so these blocks and feature positions are being recorded in GAL file then the information or annotation is given to each spot so GAL file essentially contains the X and Y and the number of the columns and so the information of each spot how they are being annotated and placed on the chip so by chance if you don't know or you prepared by yourself these buttons here allows to make own block and create your own GAL file with the help of tool called gene array generator.

- So now let's look at the control button, which is at the right side so first one is preview scan and then you also have a data scan. One stands for one wavelength so it allows you to take image from only one wavelength.
- And then you also have a multiple scan so you do a preview scan and then you do a multiple scan with this but you have other buttons which will light up when you have image in your hand and this is for the analysis. If you click this button the analysis will be performed after the alignment, this is actually open button so normally you have your file where you want to open and save your images.
- And this one is actually flag button as we discussed the features can be flaged, when image is available to you can look at good, bad or absent and you can give them different ratings. Here again looking at different zoom buttons which allows you to that which view you want to focus feature name and feature IDs for which particular one you want to go for. This one here allows different work flow controls.
- Now quickly go through a particular scan which is a simultaneous scan so both lasers would acquired at the same time so if I press on data scan button the image after putting your inverted image in the data hardware. It's scanning so you just see on the top which is very less visible as you see on the top.
- So let's try to zoom inside. If I press this button and allow zooming, you can see particularly how the scanning is happening. So you are looking at different image type so if I click on only one wavelength because it's live after scanning you can see it's going for ratio image scanning.

- So now quickly add a histogram you see it's started coming up because the scanning is going on live so it's stat reducing. Basically as we discussed it should be overlapping, my settings are looking very nice in this particular one.
- So as scanning is progressing one need to keep looking at the histogram, to determine how cy3 and cy5 are aligned.
- So how well aligned with the help of auto-PMT so cy3 and c4 you can adjust auto-PMT you can adjust laser power so that you can see this one.
- So if we see some variation then we need to come back here and adjust these parameters so that they are super-aligned.
- So in this fashion, image acquisition is happening and we give you a power that in between the slides usually people keep barcodes and our system is compatible with internal barcode which is being done.
- So nowadays, each slide is coming with multiple arrays because of the variable densities people are focusing on the custom type so this can also be done with the new software operation developments.
- So now as the scanning is being performed that's look like I have saved the image once I have saved the image I would liked to see how the different processes are being tagged.
- So say for example I have saved this image in the form of example this EST so I just want to open an image which I have just saved so basically as we discussed. Each particular array can be divided into the blocks so particular array of EST contains four blocks and each block is having features so number of feature information is given in GAL so basic terminology is array, block and features so I need to align that GAL information of positions on top of this. So I have to put a GAL file and do my further analysis so what I am going to do now is open a GAL file which allows me for alignment so best feature of GenePix pro is its capability of finding features by itself, it's a little clumsy so just show how the zoom button looks like.
- One needs to fine tune that alignment for overall proper image extraction. The only thing you have to do here is just take block and allow it to move it to the first alignment and then what you can do is click the button overhere which is for the

align, align can be done in different ways I recommend to you the first which finds all feature, all blocks and do automated fashion so you click once you see software automatically finds all the features wherever the features are by chance absent there is some physical deformity it says it is not present or flagged as bad.

- So once thing are good here it is automatically adjust according to the spot size, the overall width it is adjusting because some spots are not uniform. So as you can see you can move it but it does not affect life as long as you have kept burns and data is being stored but usually people ask me is it good idea if by chance move it once again so it's not a bad idea it takes a few idea to it. Once you have done this alignment let's look at these two slide which I said you can zoom out so you can see whole particular slide is now being scanned and aligned as well so it is very quick process which software performs very easily for doing the job. And once you have done this you can always hit a button of results. Now if I go to results it is empty if I click on results, results are being calculated and there 40 different columns which will be output in the form which GenePix understand different ways so just quickly looking at the major ones.
- The major ones are here looking for this F measure intensity for different channel 635 or 532 and this background calculation is being done accordingly in the same laser range so once you do corrections what happens you want to correct your intensity mean values with the values of the background so this is usually most important which people use for further calculations apart from ration of means or ratio of medians which can be calculated again and being presented to you in a different column format.
- So each column signifies different ones for example SD- standard deviation, CV- coefficient of variations and then different channels coming up. So in this fashion the results will be outputted if your image acquisition first controlling the part then allowing you to align and then do the analysis. So these are basic steps which anybody or everybody wants to do in microarray steps. So once you see the images the people end up in the form of results you have different columns available to you.

- Scatter plot allows you in many ways, what you plotting on X and Y axis and if here you see here I am just plotting actually towards F635 median over the F635 median so you are comparing two different channels how they have behave so essential rule is this mostly the microarray assumes all the chips are having the spots and which are not vary too much so you expect most of them to stand nearby the origin of the centre so this what you want to look at.
- The next challenge is to rally obtain some useful information from this whole data which we have already acquired.

True. As we discussed GenePix Pro is acquisition software and the molecular devices recommend acuity software for further data analysis, which can be at the level of secondary or tertiary based on that. So you do statistics as well as visualizations on single as well multiple data to handle.

Thanks to molecular devices, they have provided me with this video which actually takes you from the basic process of biology in very brief to importance software parts and so also the hardware design which is being emphasized to the level of results so let's watch that video.

Animation

- Molecular devices introduces the worlds simplest, most reliable automatic slide scanner. Now you can walk away from scanning while GenePix autoloader 4200 AL automatically loads, scans, analyzes and saves results for up to 37 slides.
- The autoloader accommodates microarrays on standard glass microscope slides labeled with four fluorescent dyes.
- These microarrays can contain just a few hundred spots or tens of thousands of spots representing an entire genome.
- As many as 37 slides can be loaded in convenient slide carrier as the carrier is inserted into the scanner sensors detect the location of each slide indicated by a blue bar on the slide carrier map.

- On the batch scan tab in GenePix pro you have complete flexibility to define the most appropriate and analysis parameters for each slide or for groups of slides.
- You can also choose to automate scanning, analysis and file saving steps. Enter an email address and GenePix pro will notify you remotely when your batch is complete.
- Using the defined scanning parameters the precision robot arm leads in the action and moves to the first slide. It securely clamps the slide and carries it to the scanning area.
- A barcode reader records the barcode.
- And then the slide is positioned for scanning. The GenePix autoloader 4200 AL can be configured with up to 4 lasers.
- A neutral density filter wheel can be used to attenuate the laser power if necessary for especially bright samples.
- The laser excitation beam is delivered to the surface of microarray slides, the beam scans rapidly across the short access of the slide as the robot arm moves the slides more slowly down the long access.
- Fluorescence signal emitted from the sample is collected by a photomultiplier tube. As the scan proceeds, sensors detect any non-uniformity in the slide surface and the robotic arm adjust the slide position accordingly to ensure to the array's surface is always in perfect focus.
- Each channel is scanned sequentially and the developing images are displayed on the monitor. The multi-channel tif images are saved automatically according to file naming conventions specified by the user.
- After the slide has been scanned, the precision robot arm replaces it carefully in the slide carrier before picking the next slide.
- As each slide is scanned a list of each saved image with associated settings and analysis files accumulates in the batch analysis tab until the batch is complete. GenePix pro automatically finds the spots and calculates up to 108 different measurements for each spot, results are saved as GenePix results or GPR file, GPR

files can be saved automatically to the acuity database for statistical analysis, clustering and other advanced investigations.

- Be more productive let the molecular device's GenePix autoloader 4200 AL system scan and analyze microarray leaving you free for great scientific discoveries.

Slide 5:

So in summary today we talked about microarray scanning and processing. We had a discussion and demonstration to over basic concepts to microarray image scanning and processing. One need to look at various parameter while scanning images and image processing because that is very crucial for doing the further data analysis before you want to obtain any biological meaningful information you need to very carefully process the image and then further perform the data analysis. In the next lecture we will continue our lecture on microarray workflow and how to analyze the microarray data obtained from these images, which we have discussed in today's lecture.