NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING (NPTEL)

Applications of Interactomics using Genomics and Proteomics technologies

Course Introduction by
Prof. Sanjeeva Srivastava

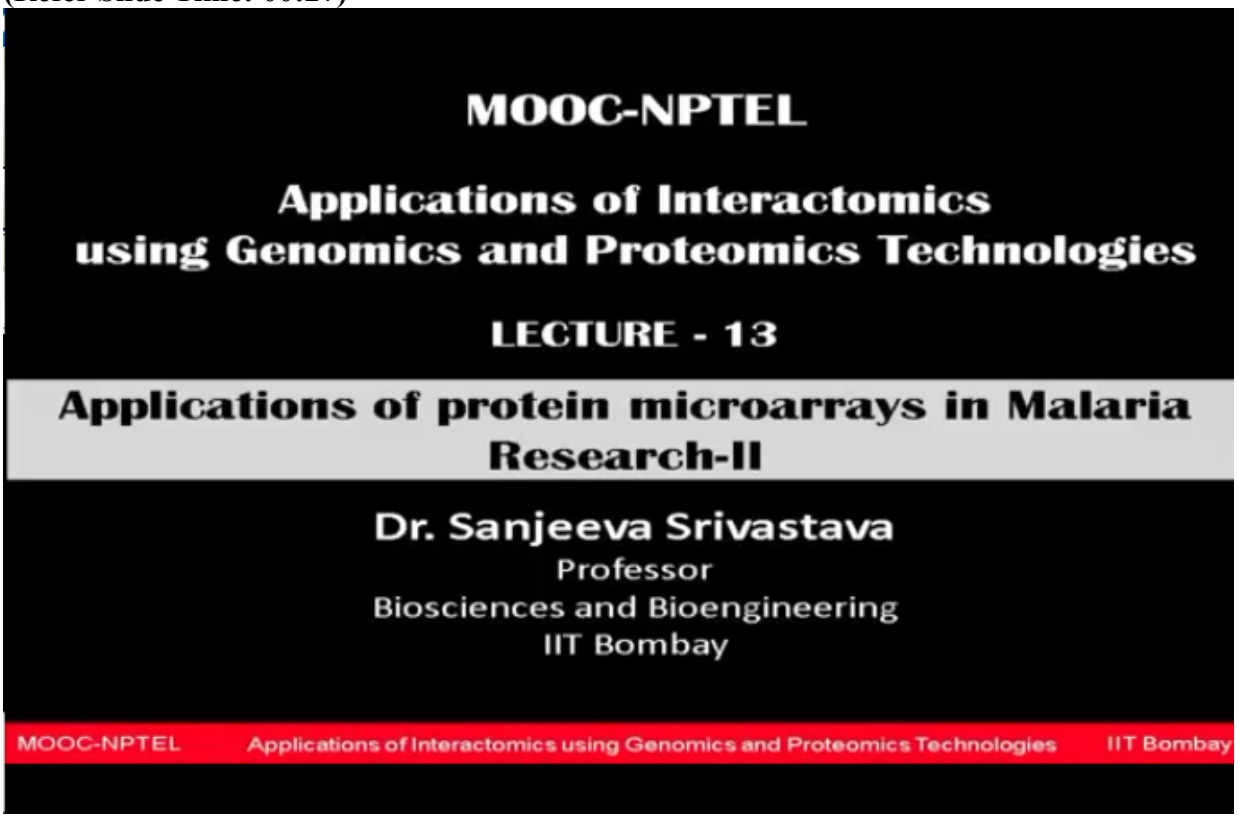
MOOC-NPTEL

Applications of Interactomics using Genomics and Proteomics Technologies

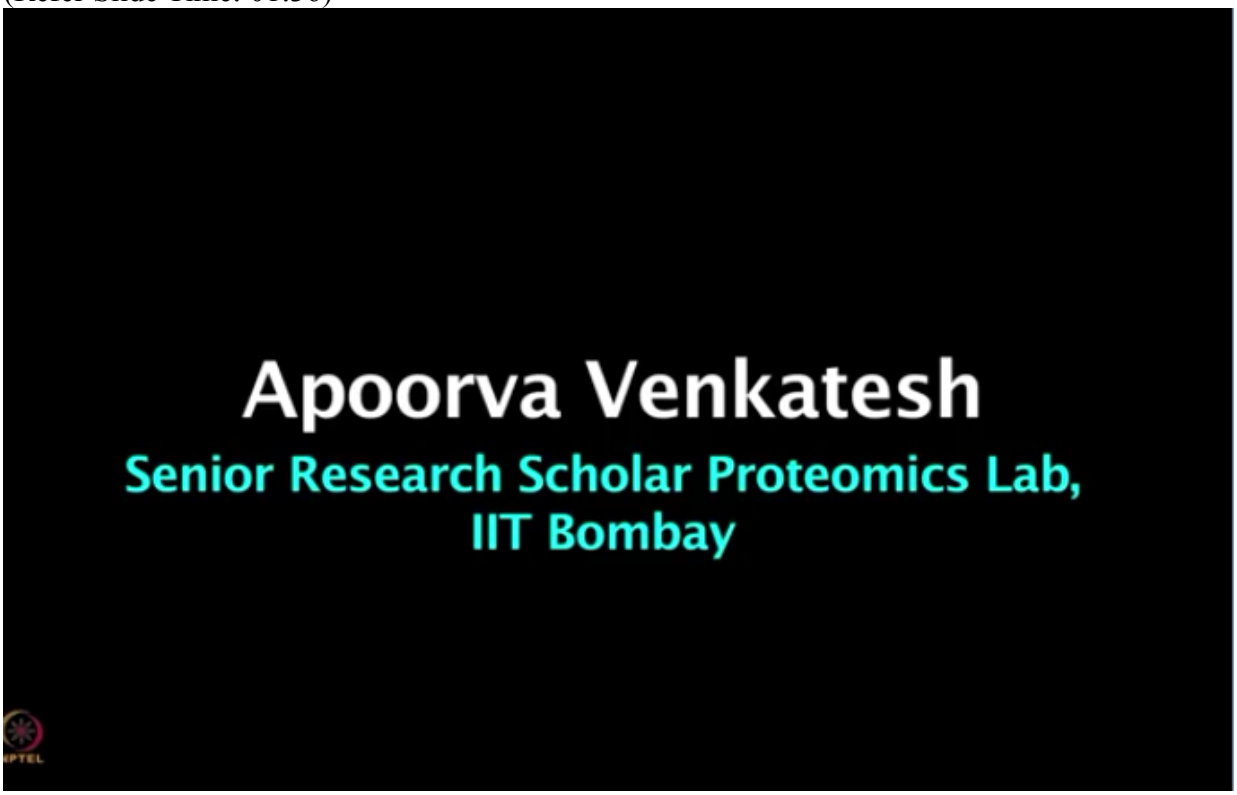
Lecture-13
Applications of Protein microarrays in Malaria Research-11

Dr. Sanjeeva Srivastava
Professor
Biosciences and Bioengineering
IIT Bombay

(Refer Slide Time: 00:27)



Welcome to MOOC course on applications of interactomics using genomics and proteomics technologies. We are discussing about different microarray based platforms and how to perform some biological applications on these chips.
(Refer Slide Time: 01:36)



In last lecture Ms. Apoorva Venkatesh showed you how to perform a microarray experiment using serum samples obtained from patients who are suffering from falciparum or vivax malaria.

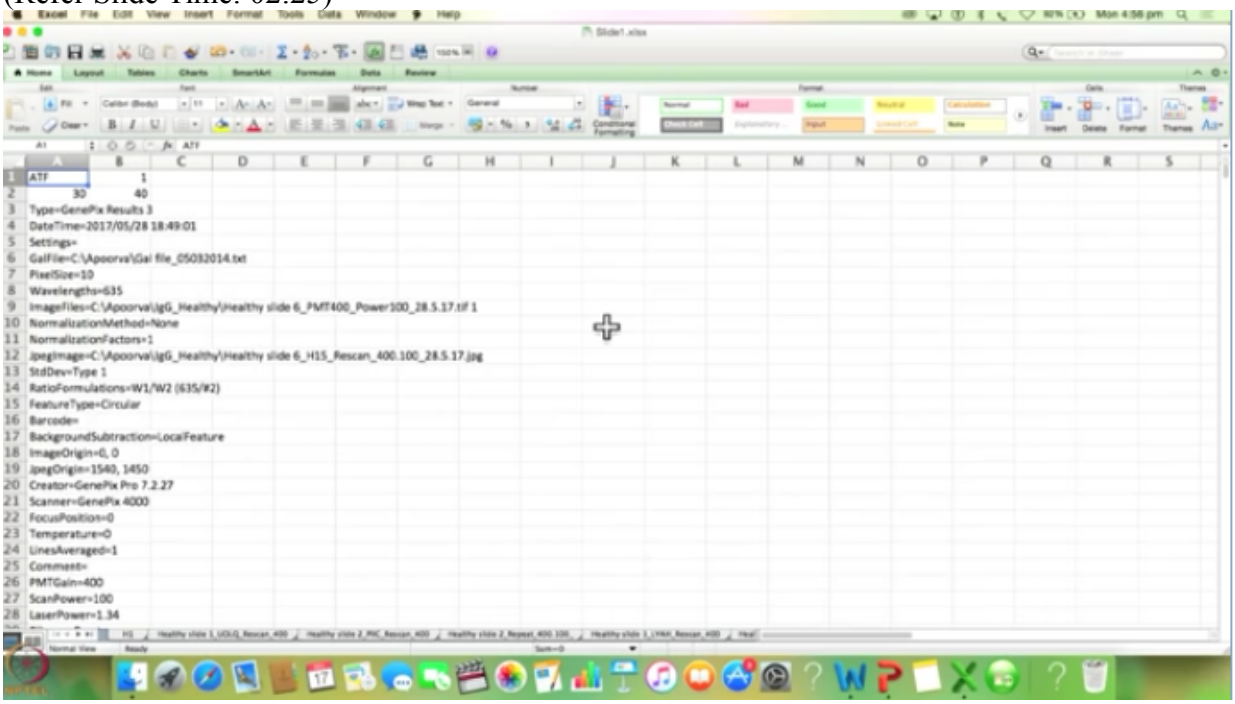
Today we are going to continue the demonstration and also show you the ways to do data normalization and how to do a microarray data analysis, specifically if you go worse to look for a biological question of question of interest.

In this case we are going to talk about several ways of how to make meaningful data from the patients who are suffering from malaria using protein microarray based platforms. So let's have this lecture and demonstration session today.

Welcome to the MOOC NPTEL course on applications of interactomics using genomics and proteomics technologies. I'm Apoorva Venkatesh, your TA for this course, and today we are going to talk about a microarray data normalization and analysis.

In the last lecture we're trying to profile humeral responses of malaria positive patients using microarray technology, so we are going to start from there, what we are going to do today is to see how to normalize microarray data using excel, so what we'll do is we'll start with the raw file you get from the microarray scanner, right, so once you take a slide and you scan it in the scanner you will extract the raw data and here is the excel sheet you see,

(Refer Slide Time: 02:25)



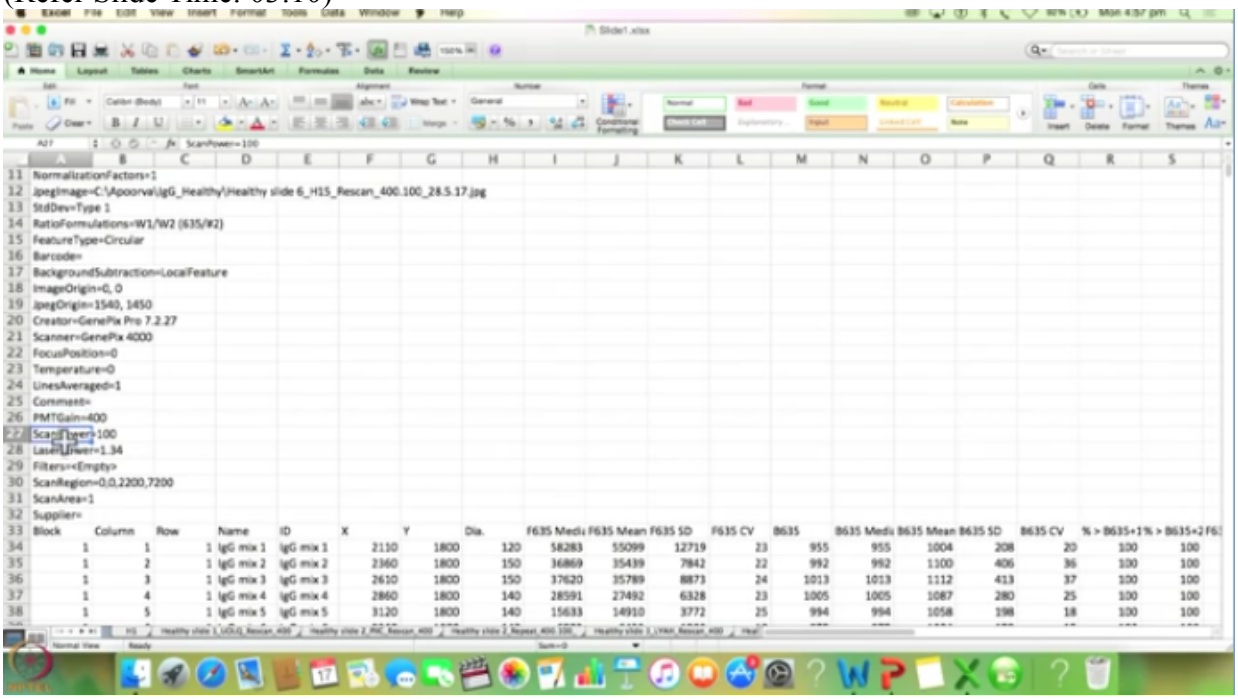
this is the type of data you get, I'm showing you this one particular slide, so I'll just like to repeat that one slide can probe 8 patient serum, so here in this one particular excel which you see here,

(Refer Slide Time: 02:35)

we actually had data for 8 patients.

So first of all I'm going to show you is how to reorganize this data, okay, so let's see first of all what kind of parameters are exported, and you will see that all important parameters are provided in this excel, for example start with pixel size is 10, the slide was canted a wavelength of 635 nanometer, then you go down normalization method this was not normalized yet, so it says none, then if you scroll down further you can see the PMT gain which is 400 scan per 100 laser power 1.34,

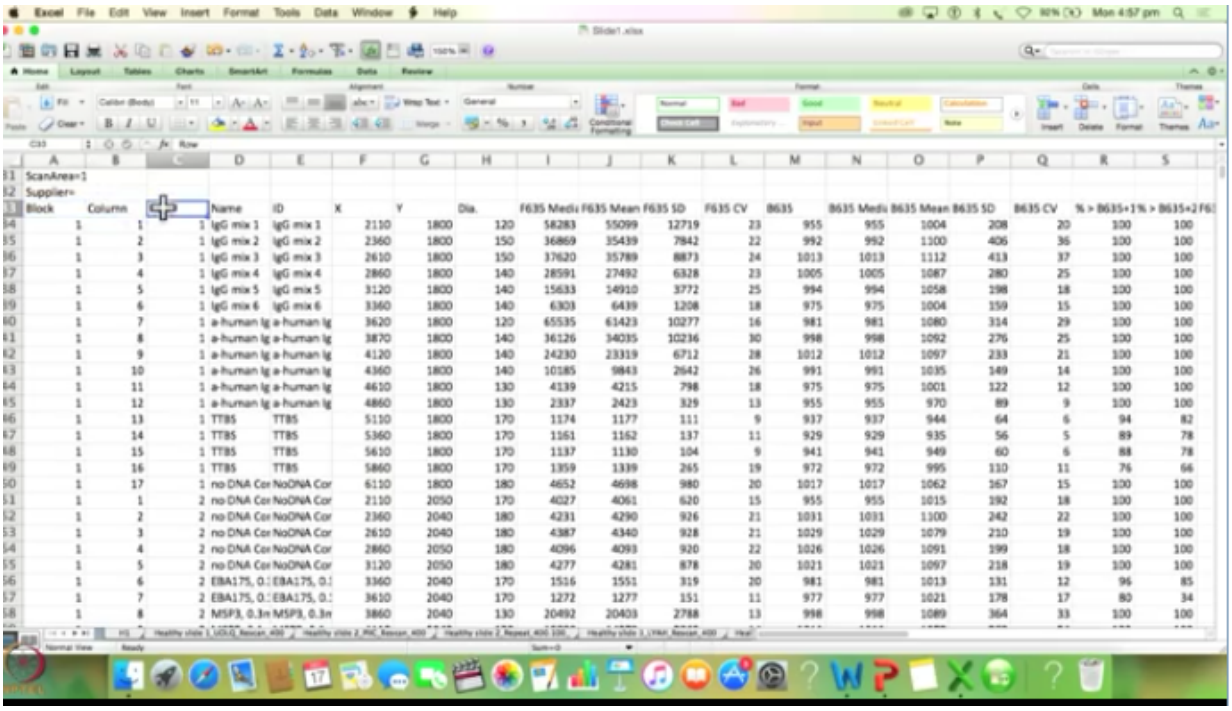
(Refer Slide Time: 03:10)



so basically later on if you want to go back and check this slides again, if example, for instance if you forget the parameters you used you can always go and open this excel to see what you had done, right.

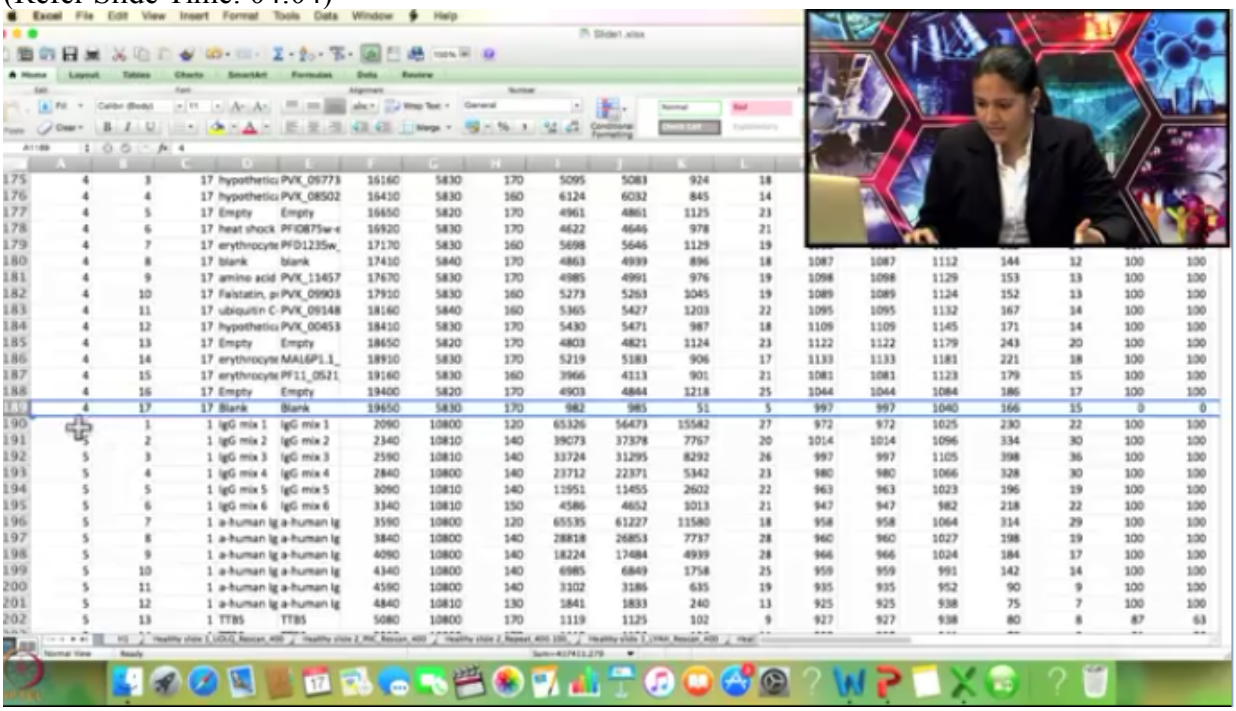
So now let's scroll down further you will see block, column, and row,

(Refer Slide Time: 03:28)



so this is very important, again let's go back to the slide layout, one slide can probe 8 patient sera, and one particular pad that is one pad which probes one patient sera has 4 blocks, so which means that if I keep scrolling down so every 4 blocks represents one patient data, right, so when I keep scrolling down and I go to block 5 a new patient begins, so that is what I'm going to talk to you about how to reorganize this, so if example if you see here, 4 ends here, right,

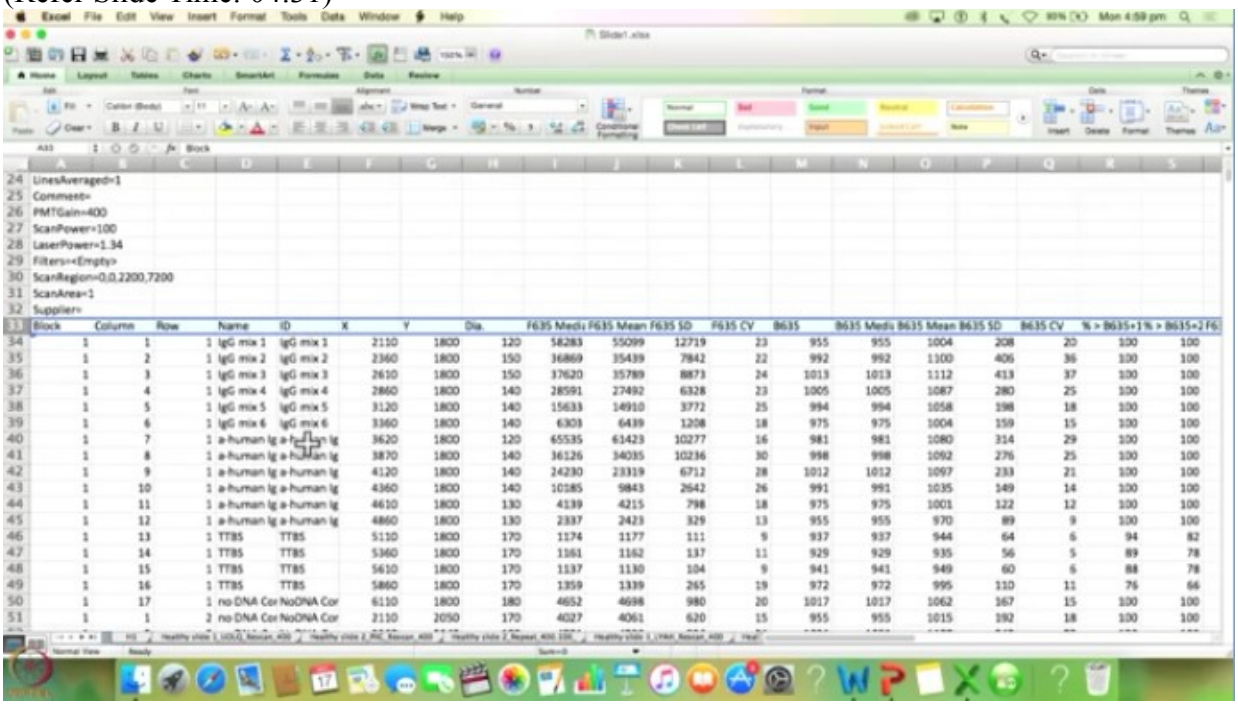
(Refer Slide Time: 04:04)



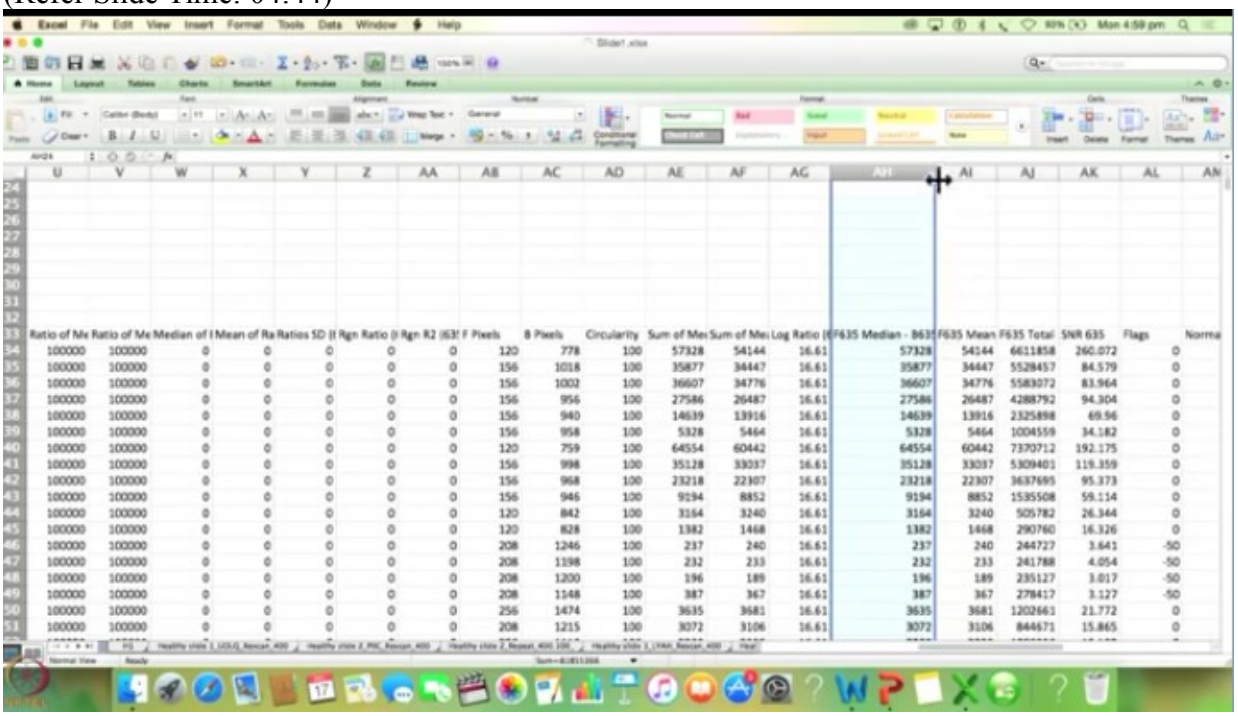
blank and you will see that is IgG1, IgG mix 1 which starts again, so this is basically your new patient, so what you are going to do is we are going to first reorganize this, but before this let

me tell you which are the columns which are important for us, so now I'm going to scroll back up and we are going to go through the columns which we have on this excel.

So now apart from block, column, row, the name, and the ID,
(Refer Slide Time: 04:31)

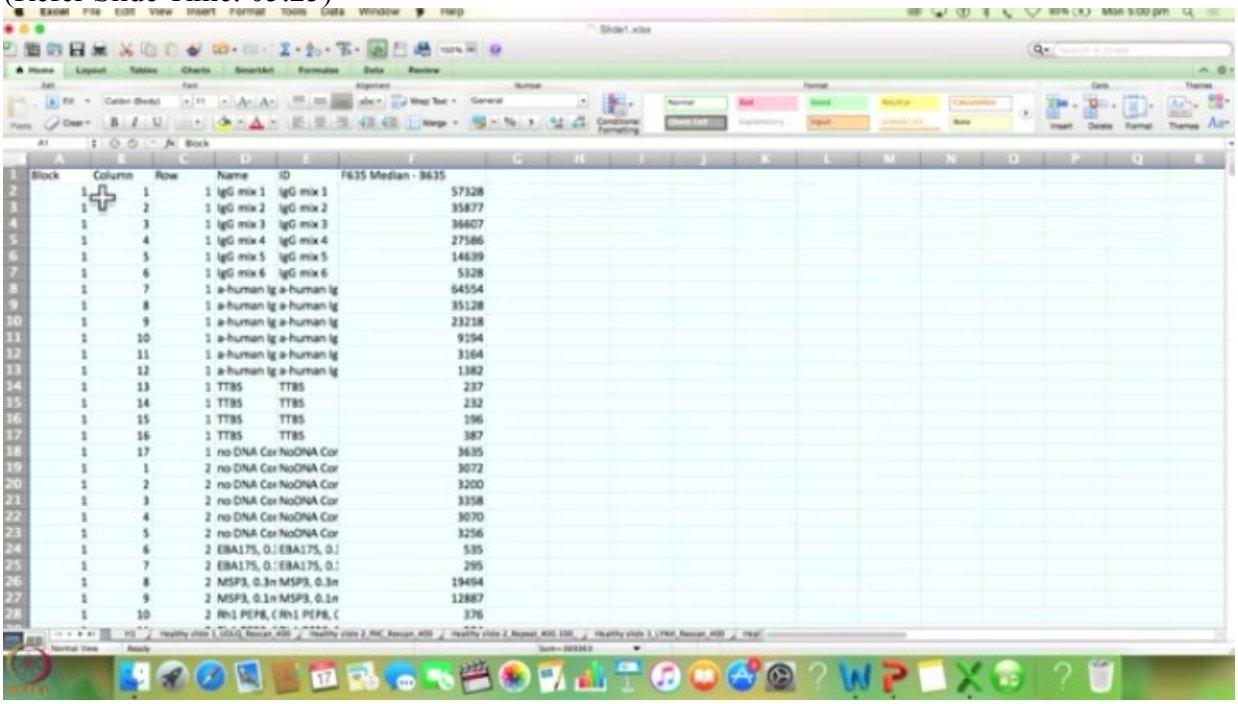


basically this is your protein ID, we don't need any of the other columns except for column AH, so you see column AH here,
(Refer Slide Time: 04:44)



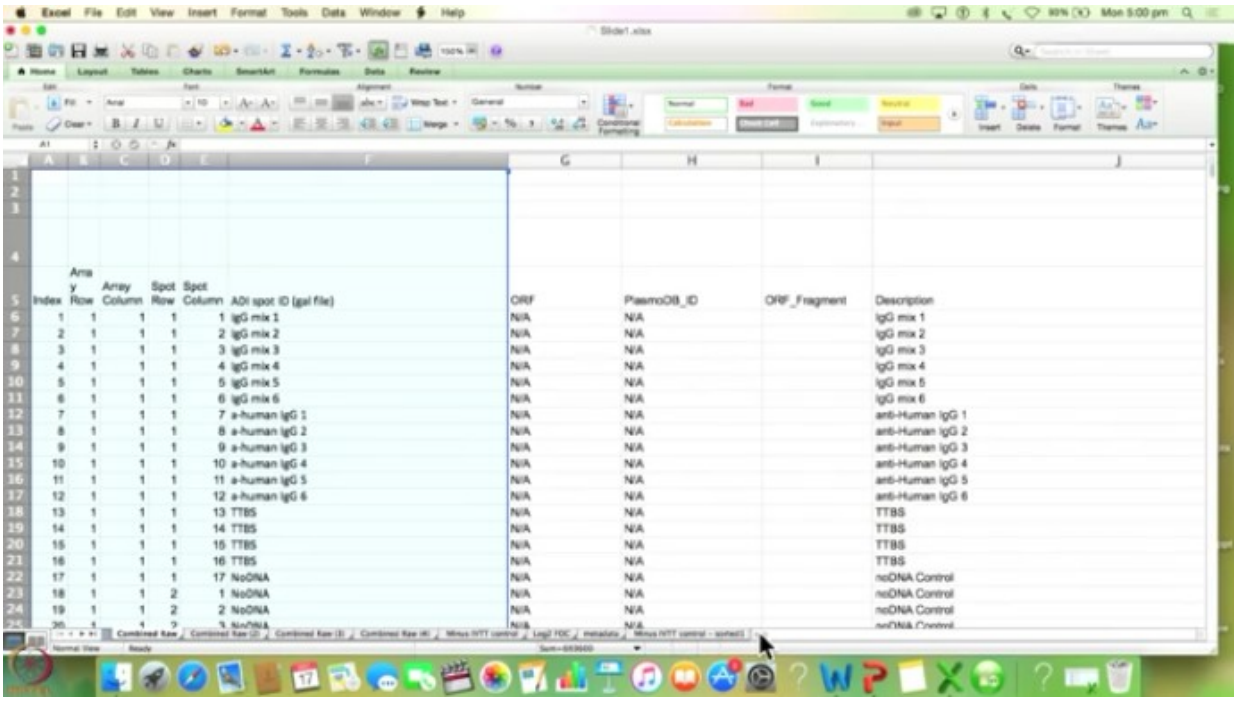
so this is basically your F635 medium minus B635 which is your basically your background signals right, so this is our the column which we actually extract and use for an analysis, and we don't need any other column here, so what I'm going to do, I'm going to first delete all unwanted columns to make this excel less complicated, so let's delete all of this, and then we go and delete this as well.

We also don't need these parameters for the analysis, I'm also going to delete this, so finally this is what you get.
(Refer Slide Time: 05:25)



So now you keep scrolling down and then you arrange all patients side by side, so when you do that this is what you get, so you also see that there are additional columns here, this is what you get from your gal file, and when you scroll right you will see that all the patients are now place next to each other, right, so this is that kind of excel you get first.

Now what I'm going to do is that I'm going to reorganize this excel to make it easier,
(Refer Slide Time: 05:57)

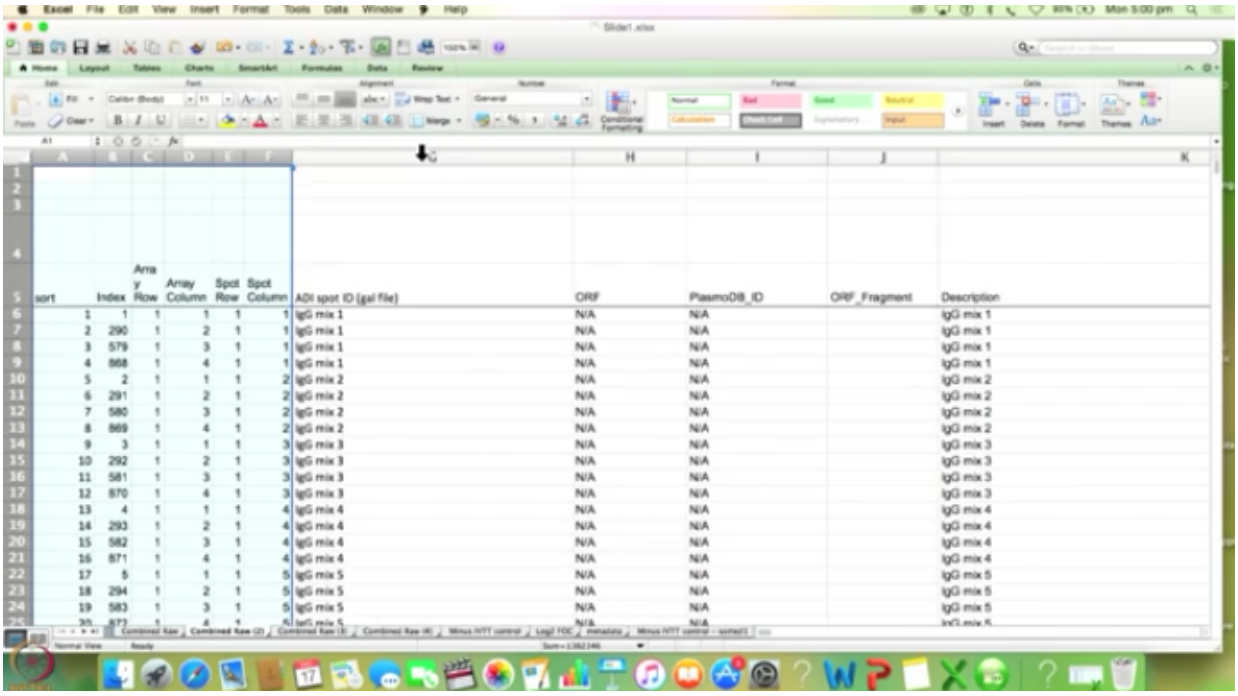


Index	Array Row	Array Column	Spot Row	Spot Column	ADI spot ID (gal file)	ORF	PlasmoDB_ID	ORF_Fragment	Description
1	1	1	1	1	IgG mix 1	N/A	N/A		IgG mix 1
2	1	1	1	2	IgG mix 2	N/A	N/A		IgG mix 2
3	1	1	1	3	IgG mix 3	N/A	N/A		IgG mix 3
4	1	1	1	4	IgG mix 4	N/A	N/A		IgG mix 4
5	1	1	1	5	IgG mix 5	N/A	N/A		IgG mix 5
6	1	1	1	6	IgG mix 6	N/A	N/A		IgG mix 6
7	1	1	1	7	a-human IgG 1	N/A	N/A		anti-Human IgG 1
8	1	1	1	8	a-human IgG 2	N/A	N/A		anti-Human IgG 2
9	1	1	1	9	a-human IgG 3	N/A	N/A		anti-Human IgG 3
10	1	1	1	10	a-human IgG 4	N/A	N/A		anti-Human IgG 4
11	1	1	1	11	a-human IgG 5	N/A	N/A		anti-Human IgG 5
12	1	1	1	12	a-human IgG 6	N/A	N/A		anti-Human IgG 6
13	1	1	1	13	TTBS	N/A	N/A		TTBS
14	1	1	1	14	TTBS	N/A	N/A		TTBS
15	1	1	1	15	TTBS	N/A	N/A		TTBS
16	1	1	1	16	TTBS	N/A	N/A		TTBS
17	1	1	1	17	NoDNA	N/A	N/A		noDNA Control
18	1	1	2	1	NoDNA	N/A	N/A		noDNA Control
19	1	1	2	2	NoDNA	N/A	N/A		noDNA Control

this is combine data for all your patients and now we will reorganize the proteins, so as you will see that proteins are present in the same order as they are present in your slide, so what we first do is now that this is common data for all patients we have already put them together, we are going to bring the IgG mix from all four blocks together. So I'll repeat this slide layout once more, now you have 6 IgG mix here this is present in your block 1.

Similarly you have the same spots present in all four blocks of the same pad, so what I'm trying to do here, I'm trying to get all the IgG mix together so we'll have 24 such spots one after the other, we will also do the same thing for your anti human IgG mix.

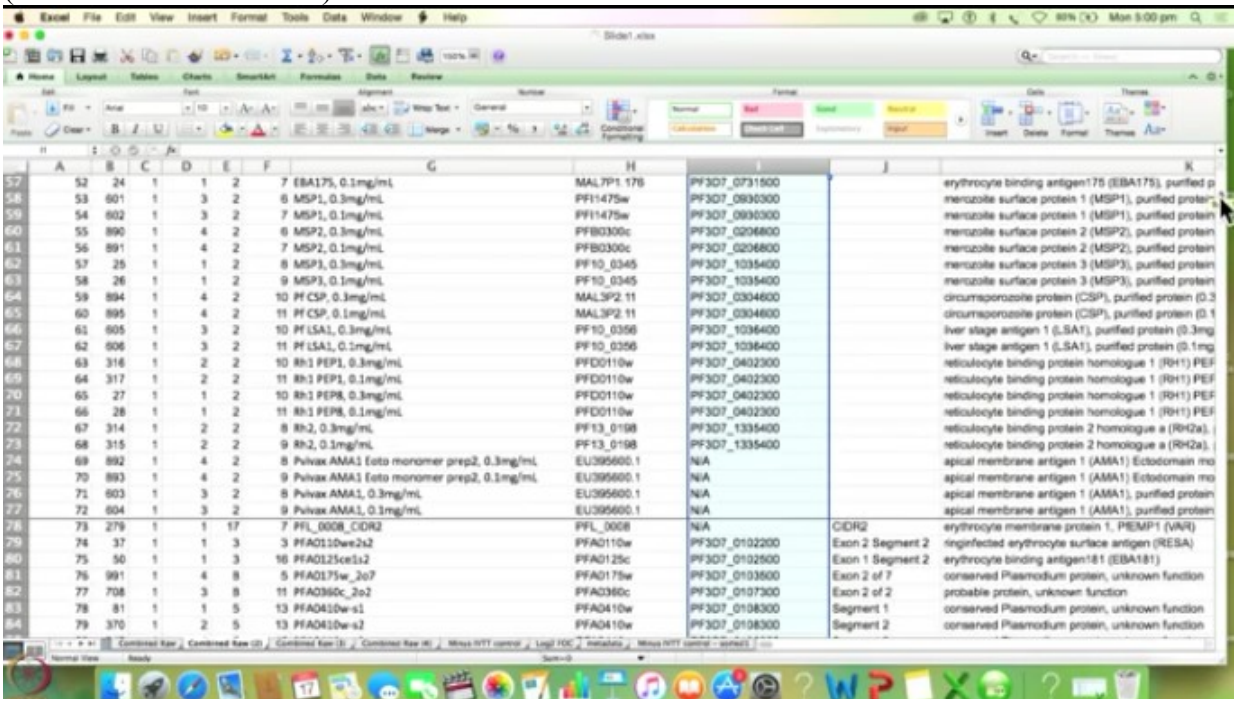
And similarly we also going to do the same thing for your next spots, so when we rearrange our excel, this is how your excel will look, what I have done here, I've put all the 24 IgG mix of 1 pad together, right, we will go through all the columns once more, for example these are all just your spot details,
(Refer Slide Time: 07:09)



sort	Index	Array Row	Array Column	Spot Row	Spot Column	ADI spot ID (gal file)	ORF	PlasmoDB_ID	ORF_Fragment	Description
1	1	1	1	1	1	IgG mix 1	N/A	N/A		IgG mix 1
2	290	1	2	1	1	IgG mix 1	N/A	N/A		IgG mix 1
3	579	1	3	1	1	IgG mix 1	N/A	N/A		IgG mix 1
4	868	1	4	1	1	IgG mix 1	N/A	N/A		IgG mix 1
5	2	1	1	1	2	IgG mix 2	N/A	N/A		IgG mix 2
6	291	1	2	1	2	IgG mix 2	N/A	N/A		IgG mix 2
7	580	1	3	1	2	IgG mix 2	N/A	N/A		IgG mix 2
8	869	1	4	1	2	IgG mix 2	N/A	N/A		IgG mix 2
9	3	1	1	1	3	IgG mix 3	N/A	N/A		IgG mix 3
10	292	1	2	1	3	IgG mix 3	N/A	N/A		IgG mix 3
11	581	1	3	1	3	IgG mix 3	N/A	N/A		IgG mix 3
12	870	1	4	1	3	IgG mix 3	N/A	N/A		IgG mix 3
13	4	1	1	1	4	IgG mix 4	N/A	N/A		IgG mix 4
14	293	1	2	1	4	IgG mix 4	N/A	N/A		IgG mix 4
15	582	1	3	1	4	IgG mix 4	N/A	N/A		IgG mix 4
16	871	1	4	1	4	IgG mix 4	N/A	N/A		IgG mix 4
17	5	1	1	1	5	IgG mix 5	N/A	N/A		IgG mix 5
18	294	1	2	1	5	IgG mix 5	N/A	N/A		IgG mix 5
19	583	1	3	1	5	IgG mix 5	N/A	N/A		IgG mix 5

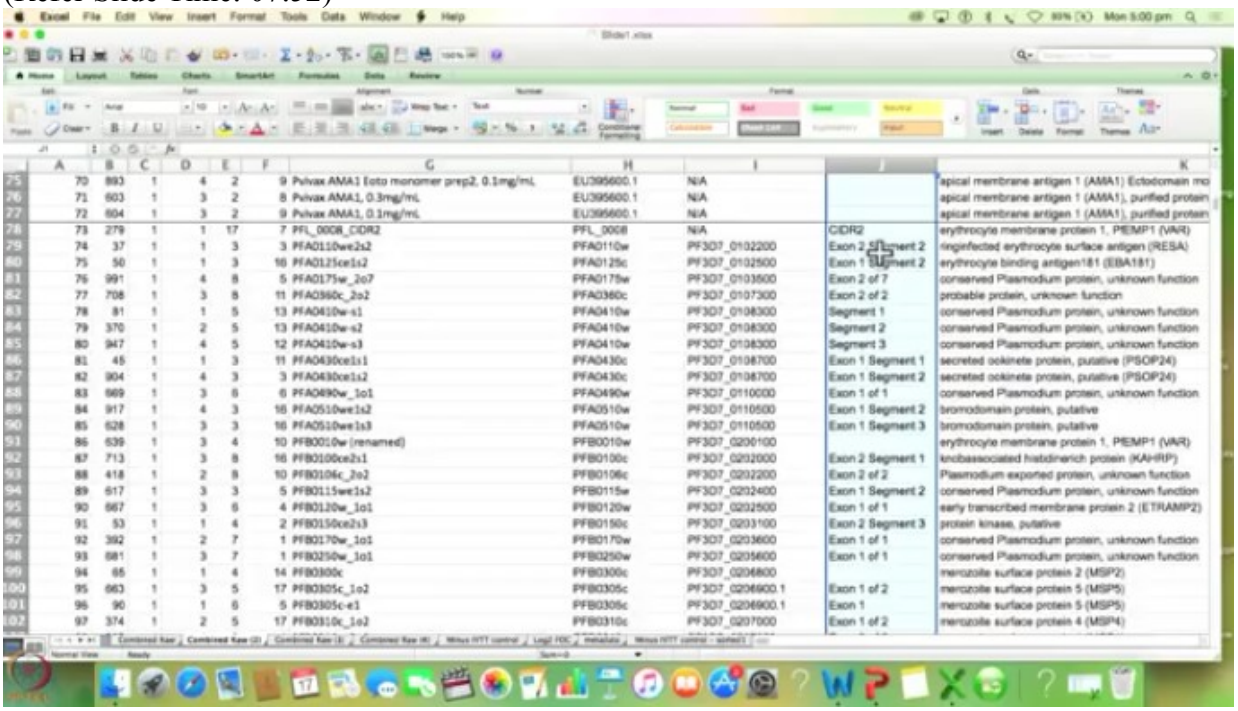
now we'll go to your ADI spot ID which is your gal file, so this is what you get from your gal file, I'll come to this in a minute, before that let's talk about ORF, so this column here is basically your ID, this is also going to give you details about the fragment which has been printed on the chip.

So let's go to plasmo DB ID, now if you look at plasmo DB ID these are all basically each and (Refer Slide Time: 07:35)



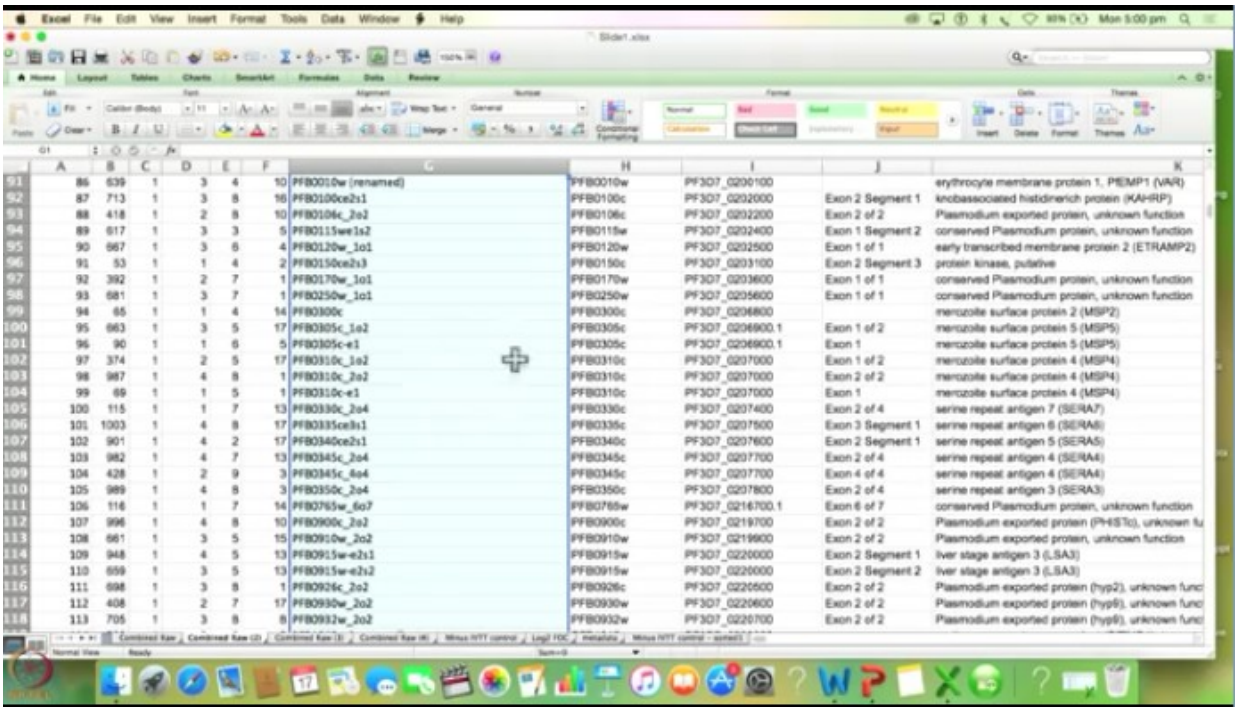
52	24	1	1	2	7	EBA175, 0.1mg/mL	MAL7P1.176	PF3D7_0731500		erythrocyte binding antigen175 (EBA175), purified p
53	601	1	3	2	6	MSP1, 0.3mg/mL	PF11475w	PF3D7_0930300		merozoite surface protein 1 (MSP1), purified protein
54	602	1	3	2	7	MSP1, 0.1mg/mL	PF11475w	PF3D7_0930300		merozoite surface protein 1 (MSP1), purified protein
55	890	1	4	2	6	MSP2, 0.3mg/mL	PFB0300c	PF3D7_0206800		merozoite surface protein 2 (MSP2), purified protein
56	891	1	4	2	7	MSP2, 0.1mg/mL	PFB0300c	PF3D7_0206800		merozoite surface protein 2 (MSP2), purified protein
57	25	1	1	2	8	MSP3, 0.3mg/mL	PF10_0345	PF3D7_1035400		merozoite surface protein 3 (MSP3), purified protein
58	26	1	1	2	9	MSP3, 0.1mg/mL	PF10_0345	PF3D7_1035400		merozoite surface protein 3 (MSP3), purified protein
59	894	1	4	2	10	Pf CSP, 0.3mg/mL	MAL3P2.11	PF3D7_0304600		circumsporozoite protein (CSP), purified protein (0.3
60	895	1	4	2	11	Pf CSP, 0.1mg/mL	MAL3P2.11	PF3D7_0304600		circumsporozoite protein (CSP), purified protein (0.1
61	605	1	3	2	10	Pf LSA1, 0.3mg/mL	PF10_0356	PF3D7_1036400		liver stage antigen 1 (LSA1), purified protein (0.3mg
62	606	1	3	2	11	Pf LSA1, 0.1mg/mL	PF10_0356	PF3D7_1036400		liver stage antigen 1 (LSA1), purified protein (0.1mg
63	316	1	2	2	10	Rh1 PEP1, 0.3mg/mL	PFD0110w	PF3D7_0402300		reticulocyte binding protein homologue 1 (RH1) PEF
64	317	1	2	2	11	Rh1 PEP1, 0.1mg/mL	PFD0110w	PF3D7_0402300		reticulocyte binding protein homologue 1 (RH1) PEF
65	27	1	1	2	10	Rh1 PEP8, 0.3mg/mL	PFD0110w	PF3D7_0402300		reticulocyte binding protein homologue 1 (RH1) PEF
66	28	1	1	2	11	Rh1 PEP8, 0.1mg/mL	PFD0110w	PF3D7_0402300		reticulocyte binding protein homologue 1 (RH1) PEF
67	314	1	2	2	8	Rh2, 0.3mg/mL	PF13_0198	PF3D7_1335400		reticulocyte binding protein 2 homologue a (RH2a),
68	315	1	2	2	9	Rh2, 0.1mg/mL	PF13_0198	PF3D7_1335400		reticulocyte binding protein 2 homologue a (RH2a),
69	892	1	4	2	8	Pvivax AMA1 Ecto monomer prep2, 0.3mg/mL	EU395600.1	N/A		apical membrane antigen 1 (AMA1) Ectodomain mo
70	893	1	4	2	9	Pvivax AMA1 Ecto monomer prep2, 0.1mg/mL	EU395600.1	N/A		apical membrane antigen 1 (AMA1) Ectodomain mo
71	603	1	3	2	8	Pvivax AMA1, 0.3mg/mL	EU395600.1	N/A		apical membrane antigen 1 (AMA1), purified protein
72	604	1	3	2	9	Pvivax AMA1, 0.1mg/mL	EU395600.1	N/A		apical membrane antigen 1 (AMA1), purified protein
73	279	1	1	17	7	PFL_0008_CIDR2	PFL_0008	N/A	CIDR2	erythrocyte membrane protein 1, PEMP1 (VAR)
74	37	1	1	3	3	PFA0110we2s2	PFA0110w	PF3D7_0102200	Exon 2 Segment 2	ringinfected erythrocyte surface antigen (RESA)
75	50	1	1	3	16	PFA0125ce1s2	PFA0125c	PF3D7_0102500	Exon 1 Segment 2	erythrocyte binding antigen181 (EBA181)
76	991	1	4	8	5	PFA0175w_2o7	PFA0175w	PF3D7_0103500	Exon 2 of 7	conserved Plasmodium protein, unknown function
77	708	1	3	8	11	PFA0360c_2o2	PFA0360c	PF3D7_0107300	Exon 2 of 2	probable protein, unknown function
78	81	1	1	5	13	PFA0410w-s1	PFA0410w	PF3D7_0108300	Segment 1	conserved Plasmodium protein, unknown function
79	370	1	2	5	13	PFA0410w-s2	PFA0410w	PF3D7_0108300	Segment 2	conserved Plasmodium protein, unknown function

every protein has a unique plasmo DB ID so that's what is mentioned in this column here, if you go to the next column which is ORF fragments, so this will explain your ORF your column H better, if you go here you will see that this specifies which exon segment is printed on the
(Refer Slide Time: 07:52)



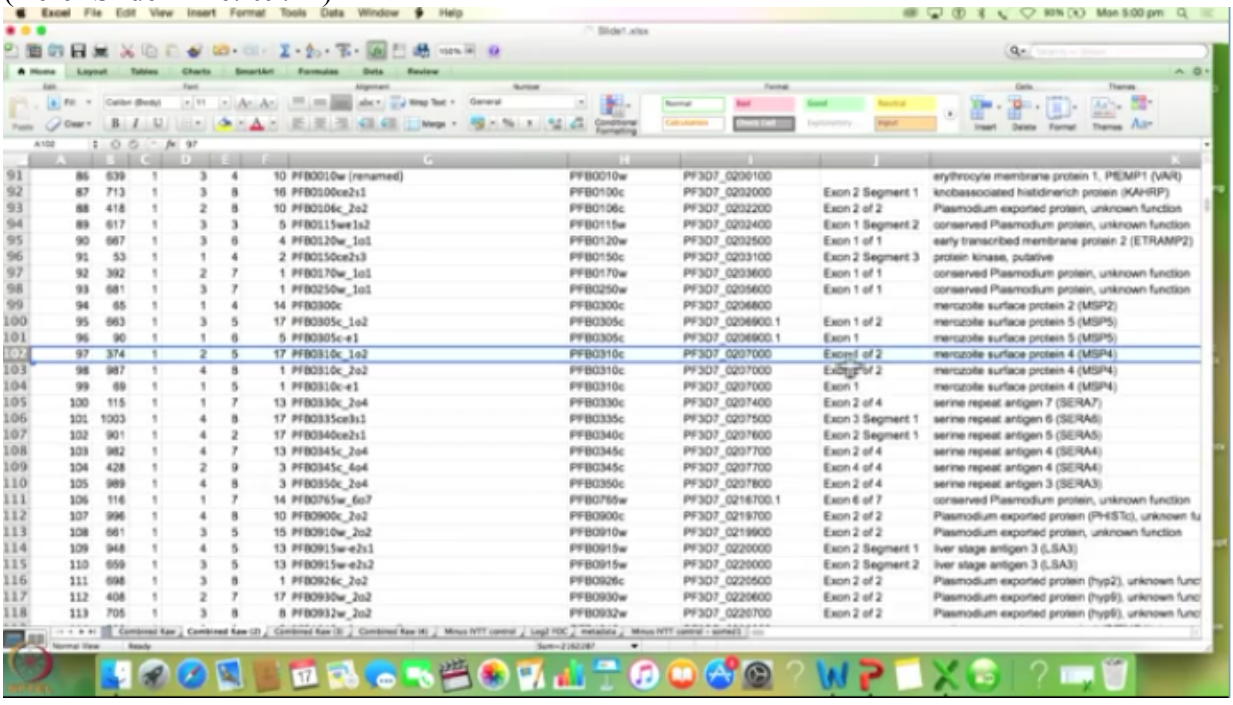
chip, so basically as you know spots which are printed on the chip were not purified proteins, they were IVTT spots and basically not, so what is IVTT? In vitro transcription translation, so what was expressed? The whole protein was not expressed here, only a certain segment of a particular exon of a protein was expressed, right, so basically it's not really right to say that proteins were expressed on the chip, instead it will be better to say that poly peptides were expressed on the chip, so this particular column J gives us details about the poly peptide that was expressed and printed on the chip, right, so that is how you'll get this ADI spot ID whichi s the unique ID for each and every protein.

What I mean here is that if you go to plasmo DB ID and then if you try to look for duplicates will actually find duplicates here because it could be, that for the same protein different exon fragments were printed on the chip, so you might get duplicates here, whereas if you go to your ADI spot ID you will not find single,
(Refer Slide Time: 08:55)



any single duplicate because these are unique ID's for each and every protein which takes into account the exon fragment which was printed on the chip, so that's what you will see here.

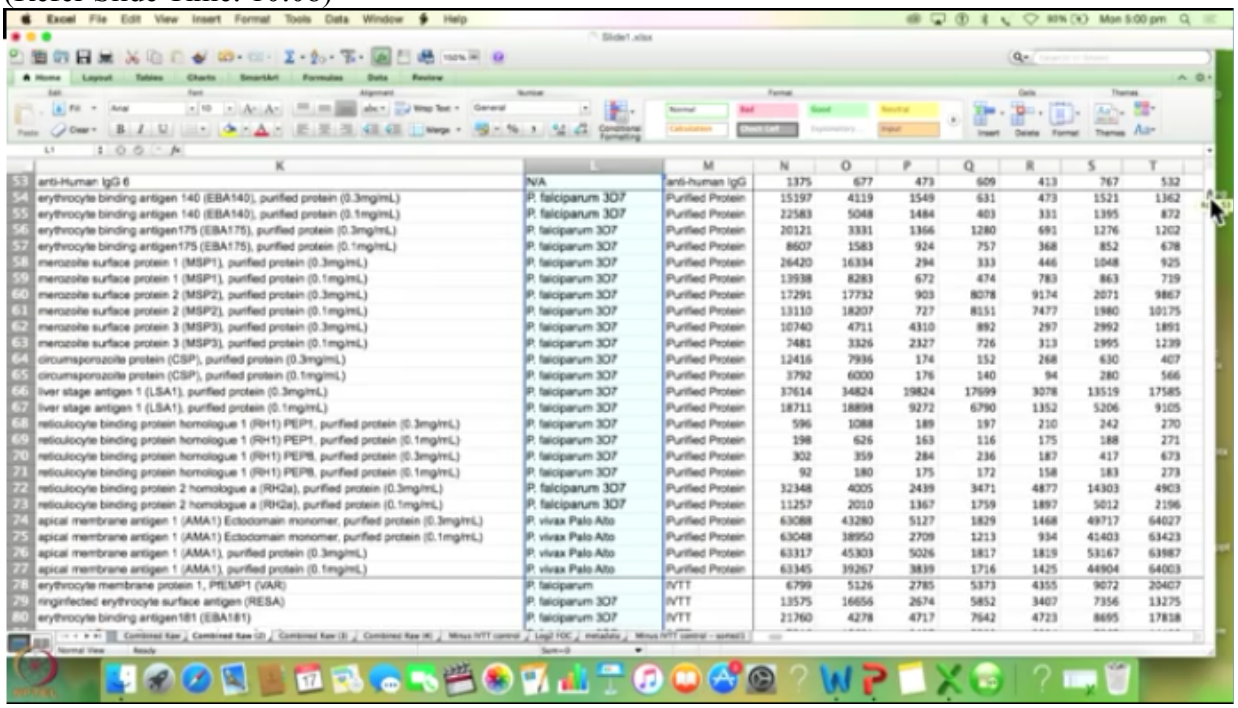
If you say for example, let's look at this particular row,
(Refer Slide Time: 09:12)



if you say that this was the ID and this is exon 1 of 2 you will actually see the ID here, and 1O2, so this becomes your unique ID for each and every protein, why I'm telling you all this is because this is very important for data analysis for all, sometimes you might just start with an analyzing your I column and then you will figure out later that there are lot of duplicates, you

don't know what you are actually doing, so what we need to do is if you want to shortlist any antigens we need to consider the G column for analysis, right.

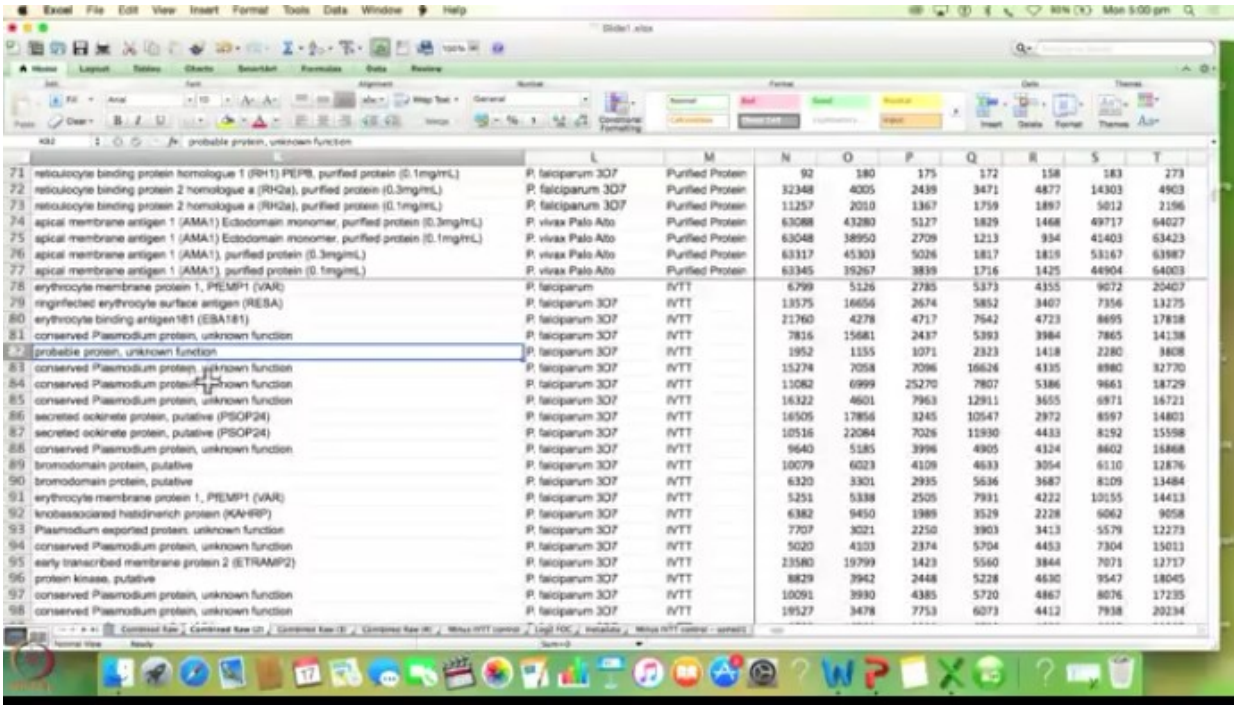
So now let's move on to the next column which is a description, so this we all know this basically describe what was printed on the chip, right, these are just basically the names of them, basically the names of the antigens, and the next column is your organism, so as you know you have two types of spots here plasmodium falciparum and plasmodium vivax, so basically this is going to tell you which organism does the antigen belong to,
(Refer Slide Time: 10:08)



so you have plasmodium falciparum 3D7 here for instance and probably and if you scroll down further you will see plasmodium vivax sal-1, right, so this is going to give you details of the organism.

And then the next column which is M is going to talk to you about deep preparation of the spot, like for example you have the first few spots are basically your IgG mix right, so then the preparation is basically your IgG mix, it's not an IVTT spot, now if you scroll down further you will have similar anti human IgG, again you scroll down further you have certain purified proteins which are nothing but a controlled proteins, so our control spots are were printed as purified proteins and not as IVTT spots.

(Refer Slide Time: 10:58)



Now if you scroll down further you will find all your other spots basically your antigens which you were trying to study are all printed as IVTT spots, so basically this entire column M gives you details about the spot preparation.

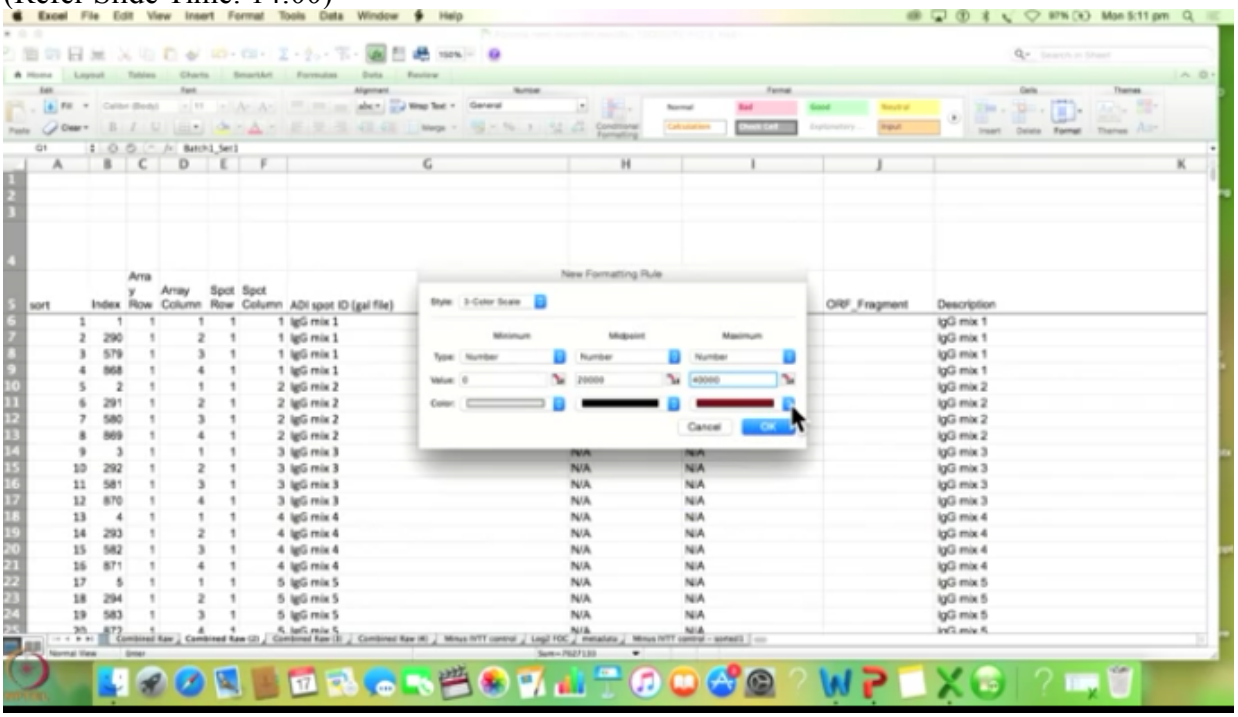
Now the all other columns here are basically your patients samples, so if I just move this a little bit, what you will see here, for example let's consider the first sample, this is basically a positive control which was part of batch 1 set 1, slide 1, and pad 1, so let me again take you back to the experiment, this experiment was performed in 4 sets of 2 batches or rather two batches of 4 sets, so you have batch 1 set 1, batch 1 set 2, then you have batch 2 set 1 and batch 2 set 2, so basically what is this telling me? This is telling me that this particular positive control was probed and batch 1 set 1 on slide 1 and pad 1, right.

Now similarly so let's go to the next one which is a real sample that was just a positive control, so this is basically probe 1 batch 1 set 1, slide 1 and pad 2, so this is going to tell me my position of the samples, so if ever I want to go back to the slides and check the real spots, right, the images of the spots then I have no exactly where to go, so for example if some samples is not behaving well and I want to go and cross check the intensity of this spots, for example some sample is giving me very high intensity signals and I want to go back and check whether its real, then I'll know exactly which file to open, because I have all the details here, so this is for the all other columns.

So this is, I hope you now understood how the excel sheet looks, right, so now what we are going to do the next thing is we are going to apply a colour gradient to this excel, right, and now I'll tell you why we are going to apply the colour gradient, so let's first do that, so for which what I'm going to do is I'm going to go to conditional formatting, I'm going to go to colour scales, more rules, and I'm going to choose it three colour scale, and then I'm going to choose number type, in case I'm going to say 0, and I'm going to assume that my entire data

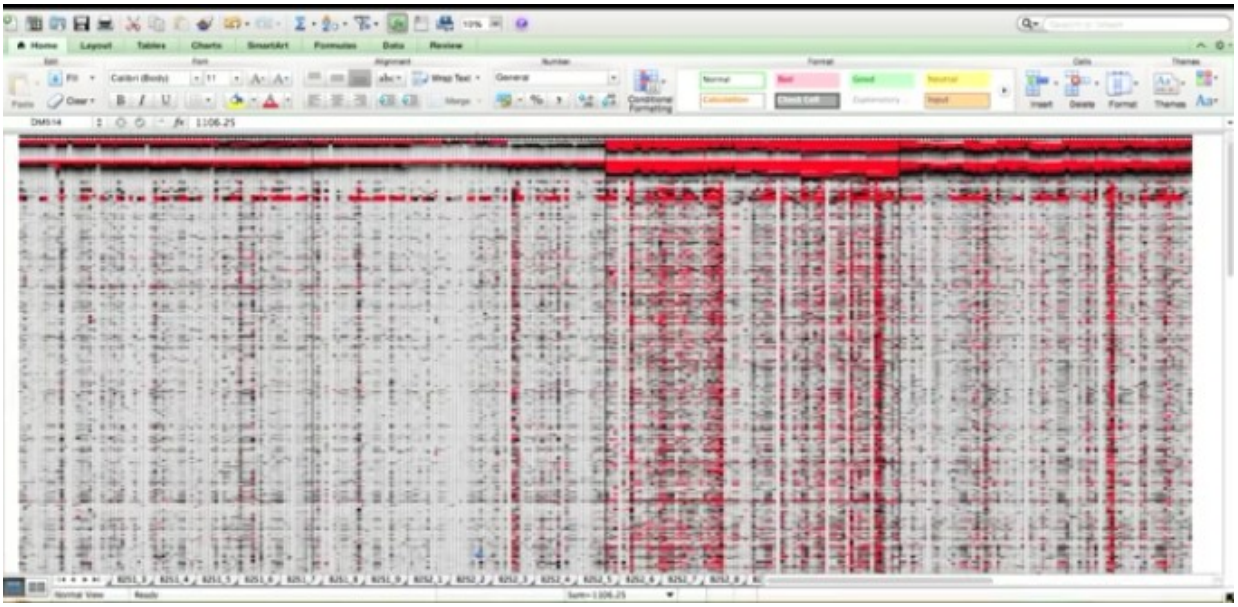
falls you know in between say certain negative values and may be around 80,000 is my maximum value, so I'm just going to assume that if my data falls in this range, I'm going to split my data based on three numbers 0, then my midpoint will be say 20,000, and my highest will be 40,000, and I'm going to choose some colours here, so I say this is maybe grey, then I'm going to keep this black, and I'm going to keep this red,

(Refer Slide Time: 14:00)



so what this is going to do? Is all my values above 40,000 are going to be in dark red, and then around 20,000 will be black, and the lowest or the least values will be grey, and those which are in negatives will be almost white, so that's how I'm going to apply a colour gradient here, so you can see in the slide here, basically what I have done is I've just minimize this excel a little bit,

(Refer Slide Time: 14:22)



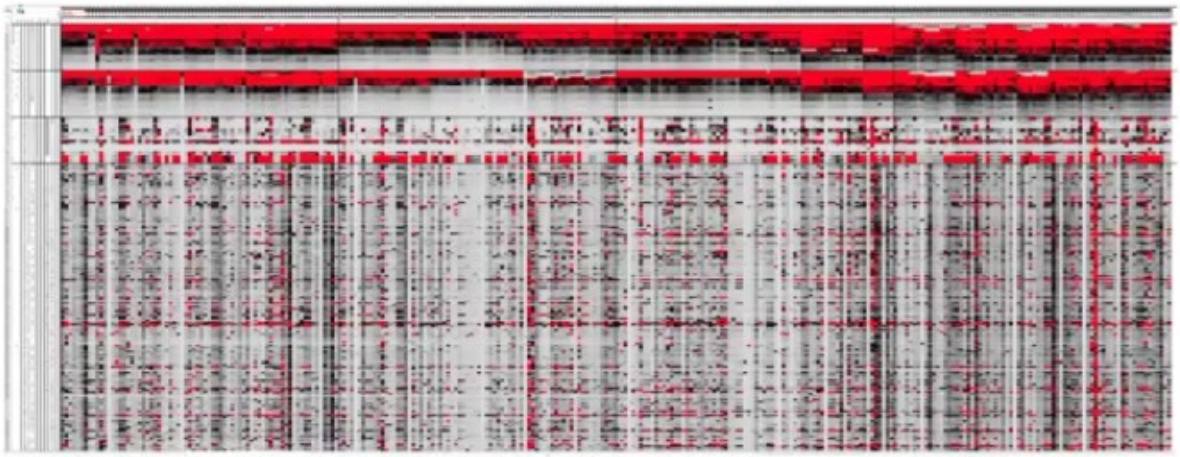
so you will be able to see all four batches at once, so I don't know if you can see a black line here, so basically this is going to split your batches, so in fact it's going to split your set, set 1 from set 2 of the first batch, and set 1 from set 2 of the second batch, so basically this is batch 1 set 1, batch 1 set 2, batch 2 set 1 and batch 2 set 2, so when you minimize this excel a little bit and you apply this colour gradient what you can see? Is that the signal intensities particularly for this batch, in fact this whole batch, but batch 2 set 1 is really high compared to the rest of the batches.

So you will know that mainly from the IgG signals here, so this particular line which you see here are your IgG mix, and this particular line which you see here is your anti human IgG, so basically this is your control which is going to tell you whether you need to rescan your slide or not, so if this is very high then all your signals by default for this particular batch will also be high, right, so that is going to screw up your results a little, screw up your results later because all the patients in this batch are going to show high signals which will be, which is not correct, so this IgG mix printed on this chip is going to basically help you in deciding whether you need to rescan your slides at different PMT settings and PAR settings, right.

So what we will do here is we will rescan the slides once more, bring these settings down a little bit and bring, these settings not as low as this, but a little lower because this is also a little high compared to this if you see, right, so later on we realized this is because of the membrane thickness of the slides, there could be other issues also which you might encounter later, so to avoid this you need to first bring down the signals and then any changes after that will be corrected by normalization, okay.

(Refer Slide Time: 16:23)

Heat map generated from data output files after rescan



What can be observed from this heat map is that the signal intensities of IgG control spots across different batches are comparable

The minor differences will be adjusted by **normalizing the data with 'No DNA controls' which will eliminate background signals**

So now having rescanned all the slides as you can see in the slide, the settings look pretty uniform though a still not very uniform and you will still feel that batch 2 set 1 has higher signals, but overall it's okay because this will then be taken care of by normalization.

So now what we will do is we will proceed with normalization using excel, now there are two strategies I'm going to talk to you about today, the first strategy is basically a very simple normalization method which we will use only for visualization, for example if you want to prepare heat maps,

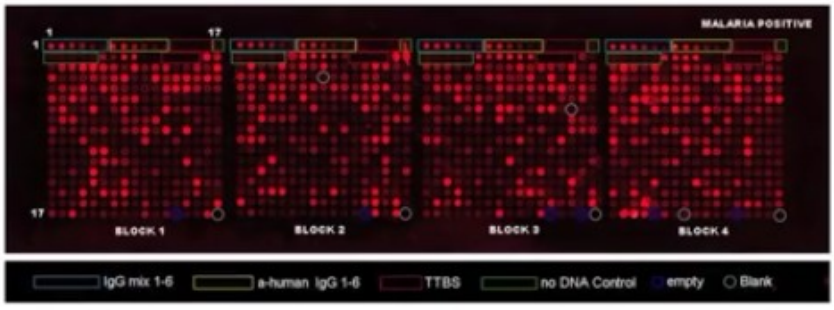
(Refer Slide Time: 16:51)

Data normalization

Two kinds of data normalization:

1. **Sample specific median normalization**: Each raw value is subtracted from the median of its 'No DNA controls'

2. **Log2 transformed FOC**- Each raw value is divided by the median of its 'No DNA controls' and Log2 transformed (Used for statistical analysis)- This is called fold-over-control (FOC) normalization which reduces the variation in signals that could potentially arise between probing operations performed at different times



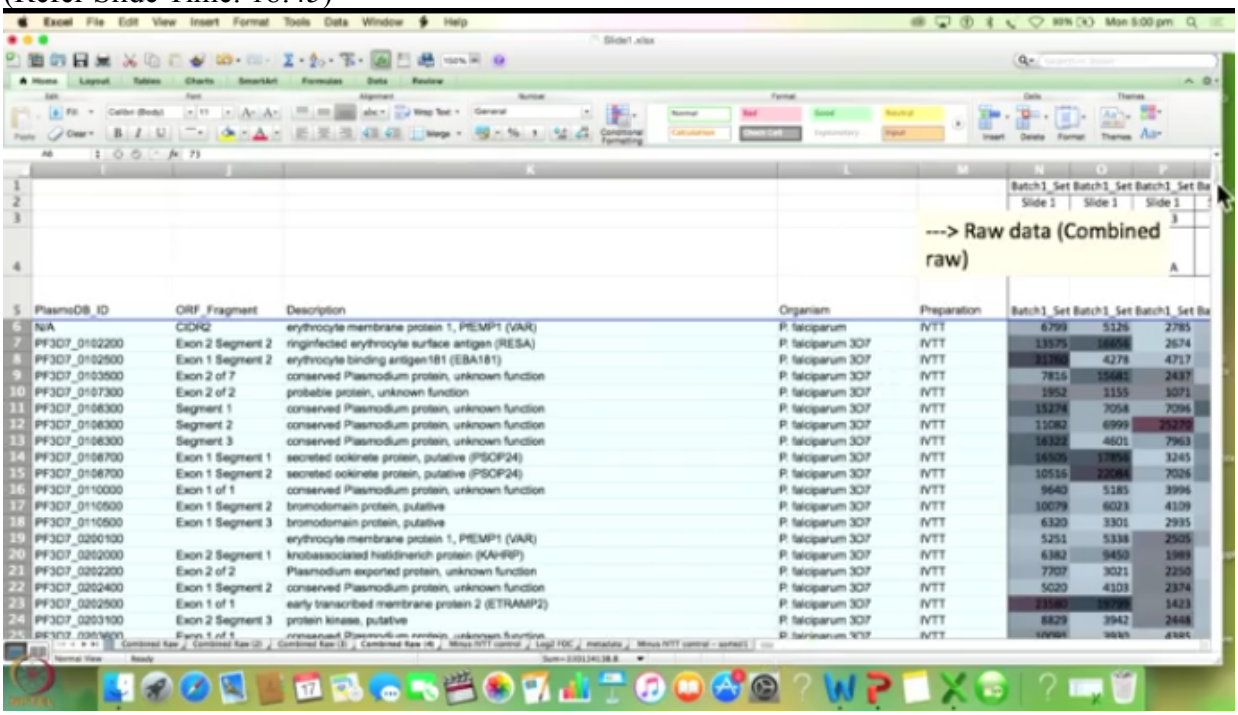
then we will use this the first normalization method, however if you want to perform statistical tests, then we will use the second kind of normalization which I'll talk to you about.

So let's first go through the first normalization method, so what we are going to do in the first normalization is we are going to subtract the raw values for each of the IVTT spots from the samples specific medium value of the no DNA controls, so I'm sure that this is a really confusing, so what we'll do is we will go step by step, first I'm going to show you what raw values are and then I'm going to show you what the no DNA controls are, right, so again we are going to come back to the same excel, it is colour coded and we've reach the stage, you also know that this, now in this data we have IgG mix, we have anti human IgG, we have purified proteins, we don't need any of those right now for our analysis, we are going to directly go down to the IVTT spots, so in fact what we'll do is we will probably just delete those rows to avoid confusion.

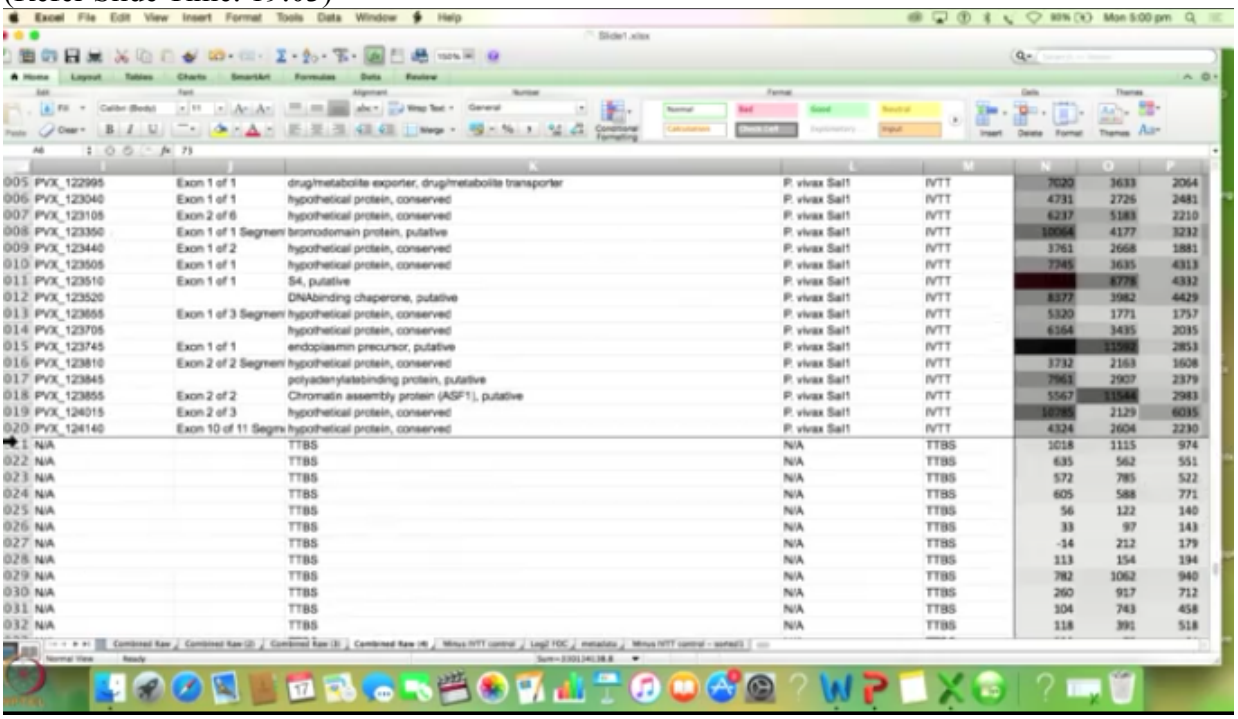
So let's start from here, I'm going to delete the first few maybe what I will do is I'll just zoom this a little bit, so I have just zoomed this a little bit, what we are going to do is we are going to delete unwanted rows right now, so we don't want IgG mix, we don't want anti-human IgG, we don't want purified proteins right now.

Again let me tell you the purified proteins basically we don't require in the analysis, but it's important when for example your slide is not worked at all and or you have not got the signals you required, you can always go back to the positive control spots to see what the signals were, right, so this is basically used for such you know analysis just as controls, so right now we are going to delete those rows and we are going to only keep rows which are IVTT mix, right, that's what this is.

(Refer Slide Time: 18:45)

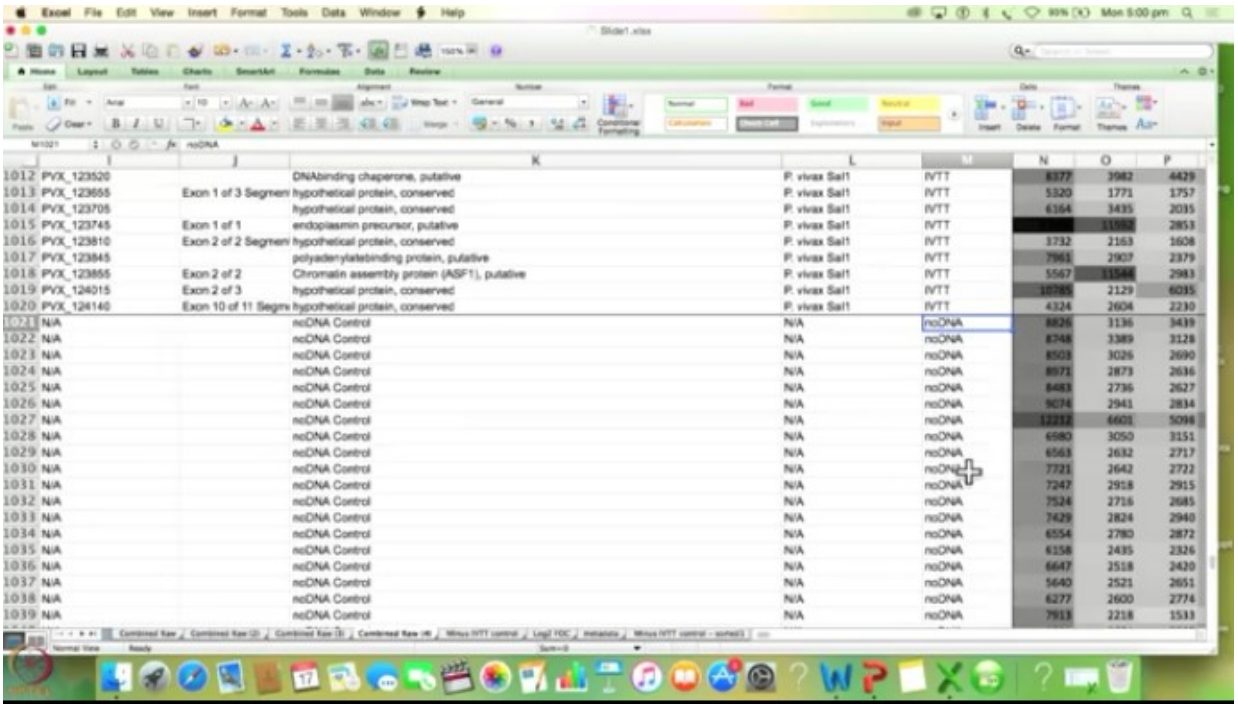


So now we are going to have this way 500 plasmodium falciparum IVTT spots, and 515 plasmodium vivax IVTT spots, so we are going to go down, you have deleted unwanted rows, there are few more rows below which we don't need, so after these 1015 spots
(Refer Slide Time: 19:03)



there are few more like TTBS which is nothing but your buffer spot, where only buffer is spotted and then you have some empty spots data, then we have data for blank, so this is also unwanted we are going to delete that as well.

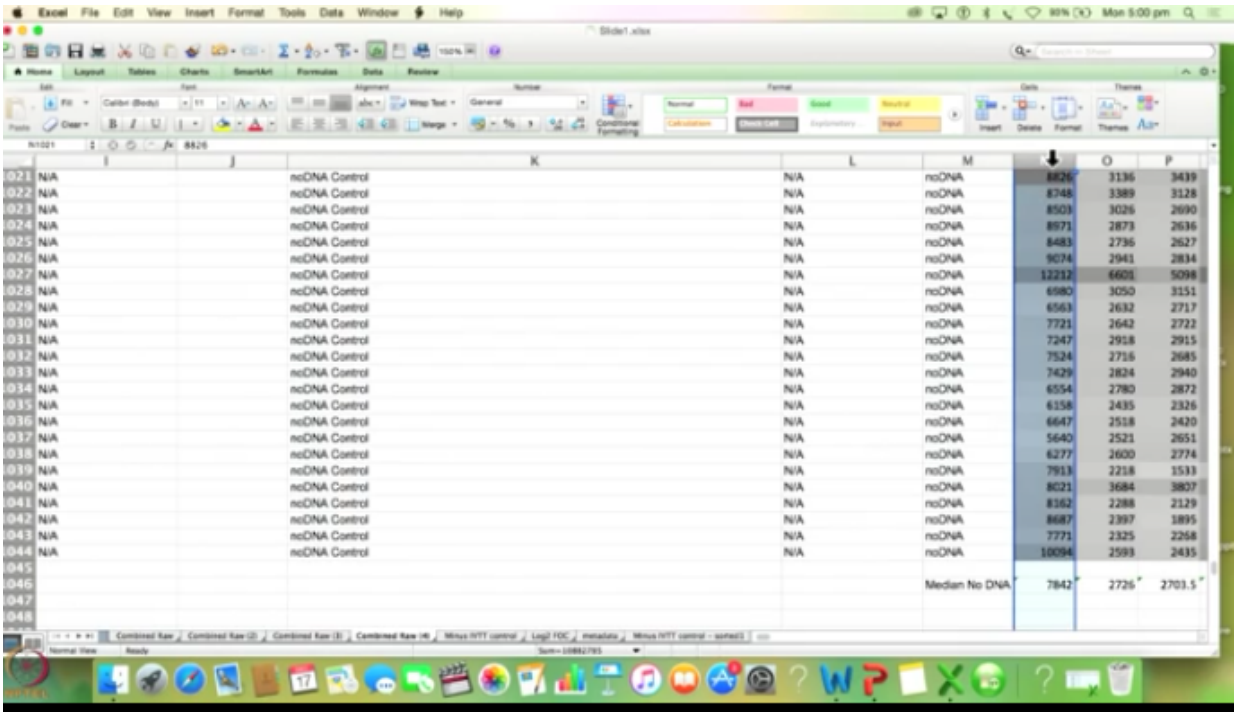
So now what we have are 1015 IVTT spots and 24 no DNA control spots, so now what are these no DNA control spots?
(Refer Slide Time: 1:31)



So basically these spots have the entire IVTT mix except the plasmid, so basically what you expect here is no expression because you don't even have the plasmid here, whereas the IVTT spots have the entire IVTT machinery just like no DNA but they also have the plasmid where you are going to express your gene of interest, whereas you don't have that here, so what is this going to provide? This is going to provide your background signal, so what we are trying to do in the first type of normalization is we are subtracting a raw signals from background.

So now there are 24 such spots which you remember we have rearranged and that's why it's come together, grouped together like this, the first thing we are going to do is take a median of this which I have already provided you here, so this is the formula for it, I've just done this the whole thing in excel, so now we have a median value here, so the first thing what I'm going to do is for this particular sample which is in column N,

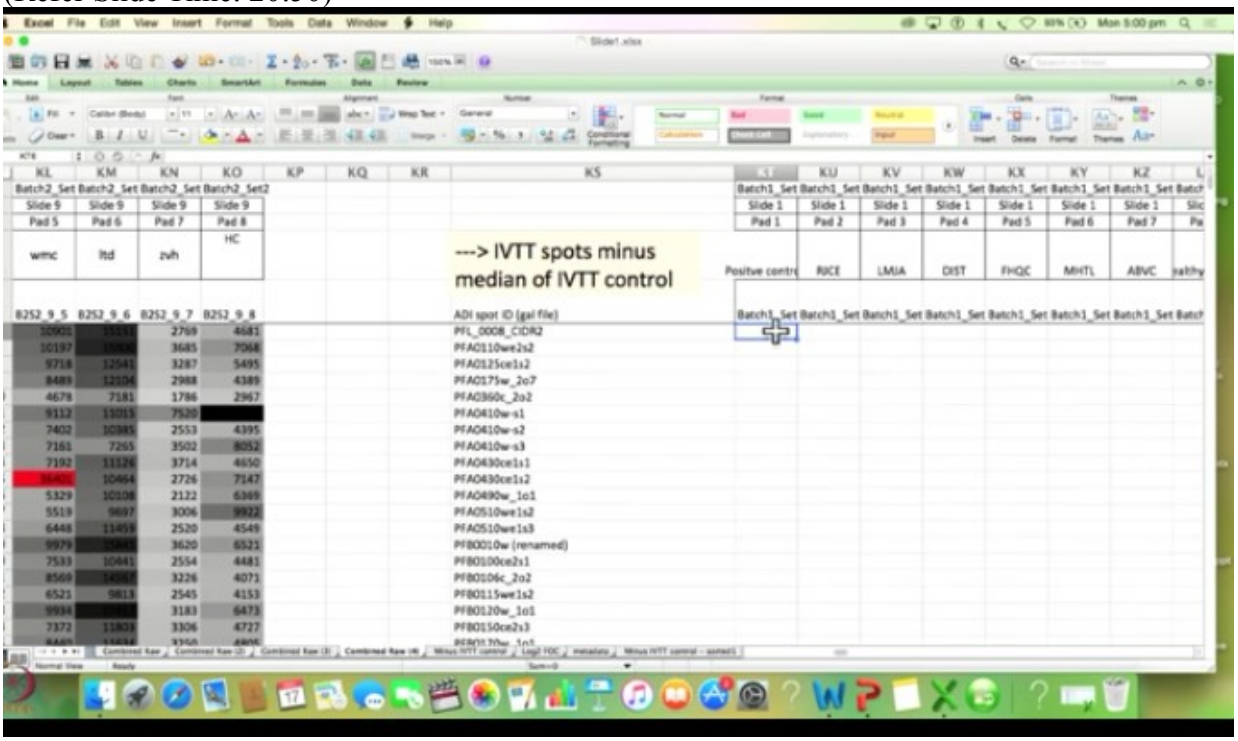
(Refer Slide Time: 20:20)



this is one sample, I'm going to subtract the values for each and every IVTT spot from that particular medium, right, so each of this spots which you see here, I'm going to subtract it and that is what is my median sample specific median normalization.

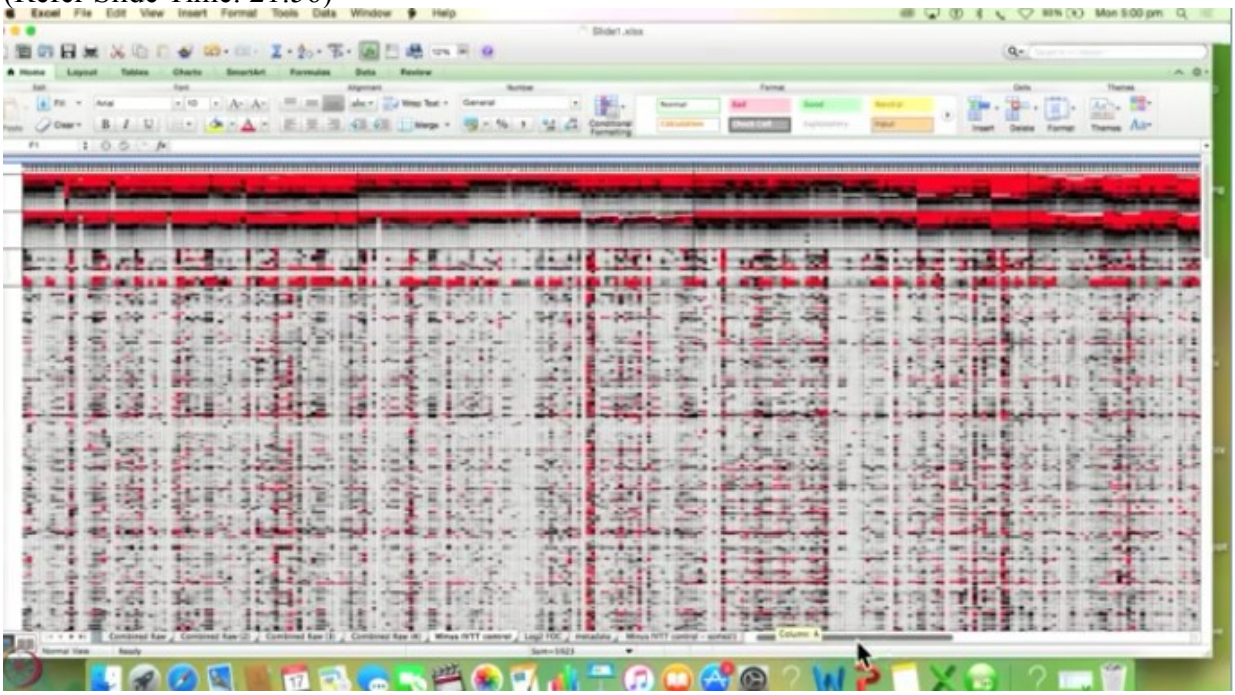
So let's scroll, so probably we will do this in the same excel, okay I have kept place for that here,

(Refer Slide Time: 20:50)



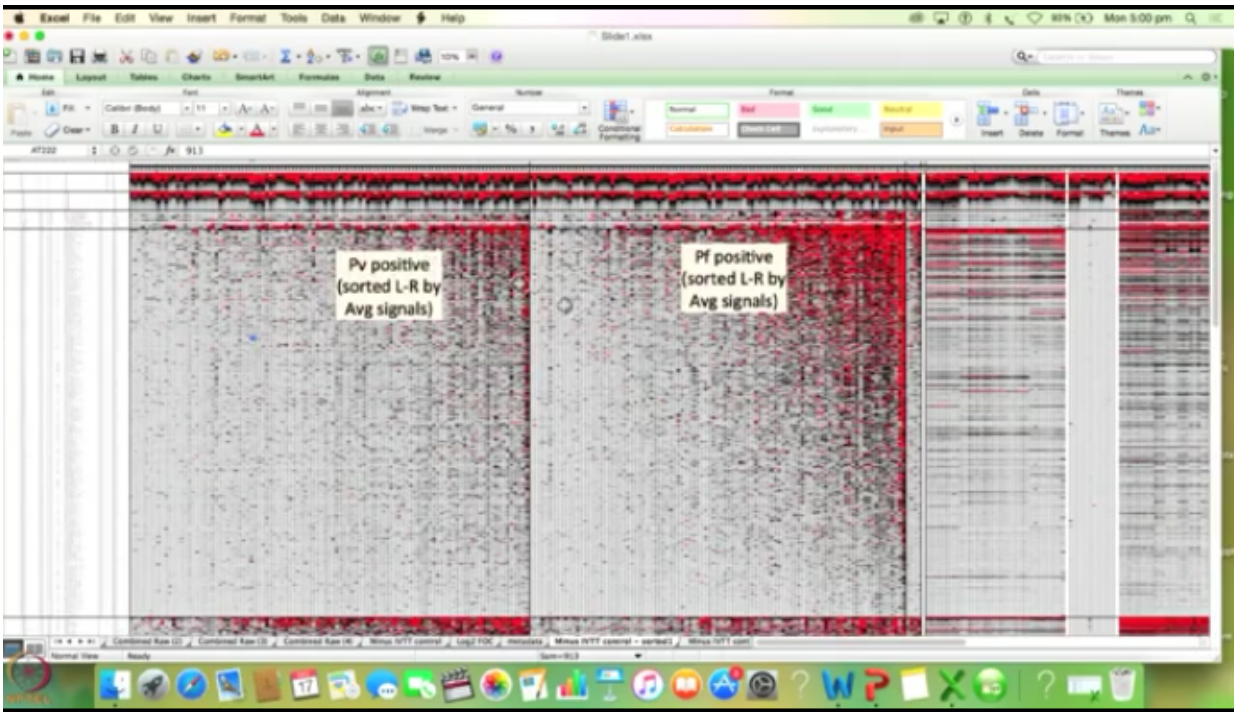
so this is called IVTT spots minus median of IVTT control, so that's exactly what we have to do, we are going to say is equal to, then we are going to go to that particular spot, so say let's take the first patient, and then we will see minus and we go to the median value which is 7842, so now because I want this row to remain constant throughout, I'm going to put it dollar sign in front of the row, so this is what we get here, and now I'm just going to drag this across as well as down, so once you drag and drop,

(Refer Slide Time: 21:56)



this is the kind of excel you get, I have just minimized this, but if you apply a colour gradient this is how it looks overall, so this is what you can use now to make your heat maps and what I have also done is that I have sorted this based on the antigens as well as the patients who were falciparum positive and vivax positive, I've split them completely, and I have also made another excel sheet based on age, you can also split them based on age, so this is how I have sorted them.

(Refer Slide Time: 22:28)



So I have put all your PV positive patients together and PF positive together, and I have also sorted based on age, so you have PV positive, PF positive as well as sorted by age, so this way you can sort your excellent different ways, you can also use other software to make your heat map, but basically this kind of, once you normalize it in this way you do not perform any statistical analysis with this data.

First statistical analysis I'm going to now show you the next normalization method which is your log 2 transform fold over control normalization, so for this I'm not going to show you the entire method again because now you know how to do it on excel I'm sure you all know, for this I'm only going to show you the steps, the first thing what we do is we are going to setup floor of 100, so what we are trying to say here is that all the samples which are below 100 is going to have a value of 100, so this is going to remove all my negative values from my data, so that's the first thing and I have done it here for you and we are going to keep scrolling right.

The next step what we are going to do is to divide each and every raw value by the median of the IVTT control spots, so just like how we did previously we subtracted raw values from the median of the IVTT control spots, this time we are going to divide it so that's what is called fold over control.

So once you setup floor of 100 then you divided, and the next thing you are going to do is to convert this whole data into log values, so your log to transform this entire data, right, and that's why it's called log 2 fold over control, so once you do this, this data can be used for any statistical analysis, so this because this normalization is known to be more stringent, okay.

So now either you can use programming to do your statistical analysis or you can use different softwares which provide you statistical test, but what you need to know is which type of test you need to use which is beyond the scope of this lecture, but you can always read about what

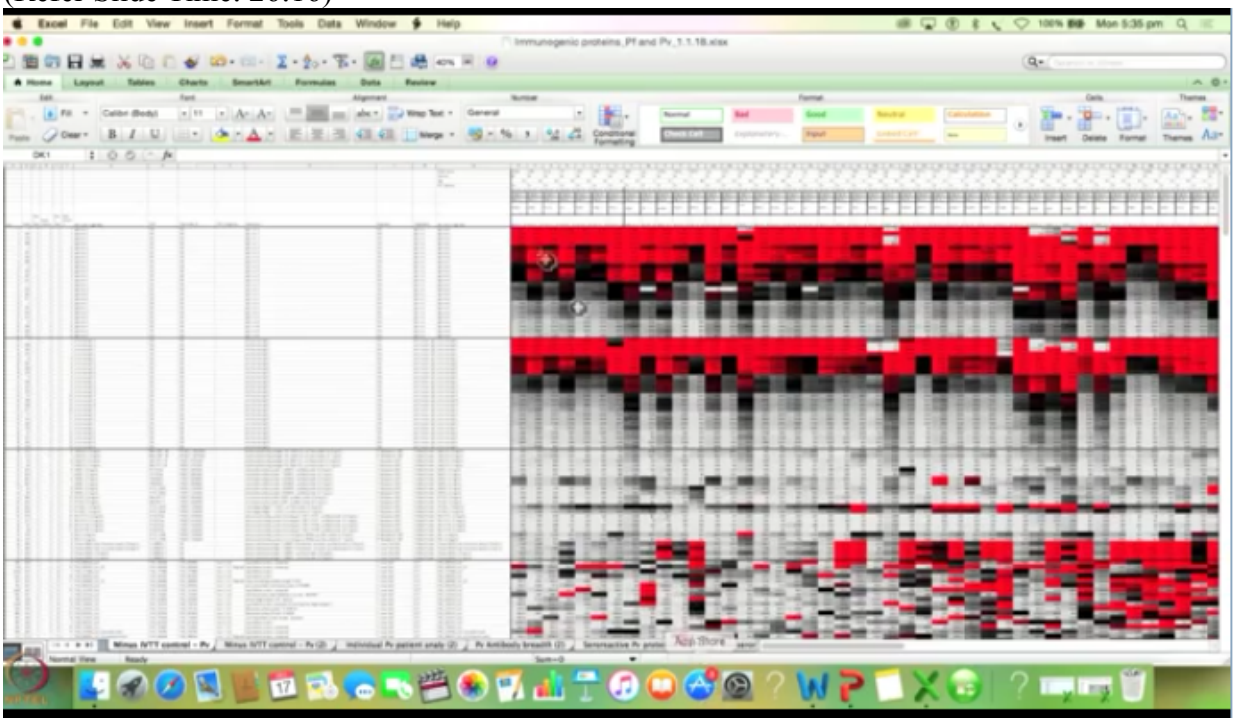
you want to do, and you can also decide on which software you can, you want to use, for example graphpad prism is an excellent tool for preparing for graphs, it also helps you in a lot of statistical analysis, but if your data is really huge like the one we have is not very huge, but still it is not very small for graphpad prism, so for example you can have graphpad prism can get stuck in the middle if you are using data if even this size, so of course if you have bigger data sets then it's very difficult to use softwares like graphpad prism.

However if you are going to have only 10 patients or 20 patients with 40 proteins or something, graphpad prism does offer you a lot of statistical test. Apart from that there are other softwares as well, you have metabol analyst though it is metabolomics data, you still have the module called significance analysis or microarrays in it which you can explore for your microarray data analysis, but there are of course our programming and python and other things will definitely be much better for your analysis, as you will save a lot of time as well.

So what I'm going to do towards this, we're coming to the end of the lecture, I'm only going to show you very small analysis you do an, I have done on excel, so basically what my aim here is to identify most zero reactive proteins from my chip, which means that the proteins which elicit the maximum antibody response in malaria patients right, so that is my aim.

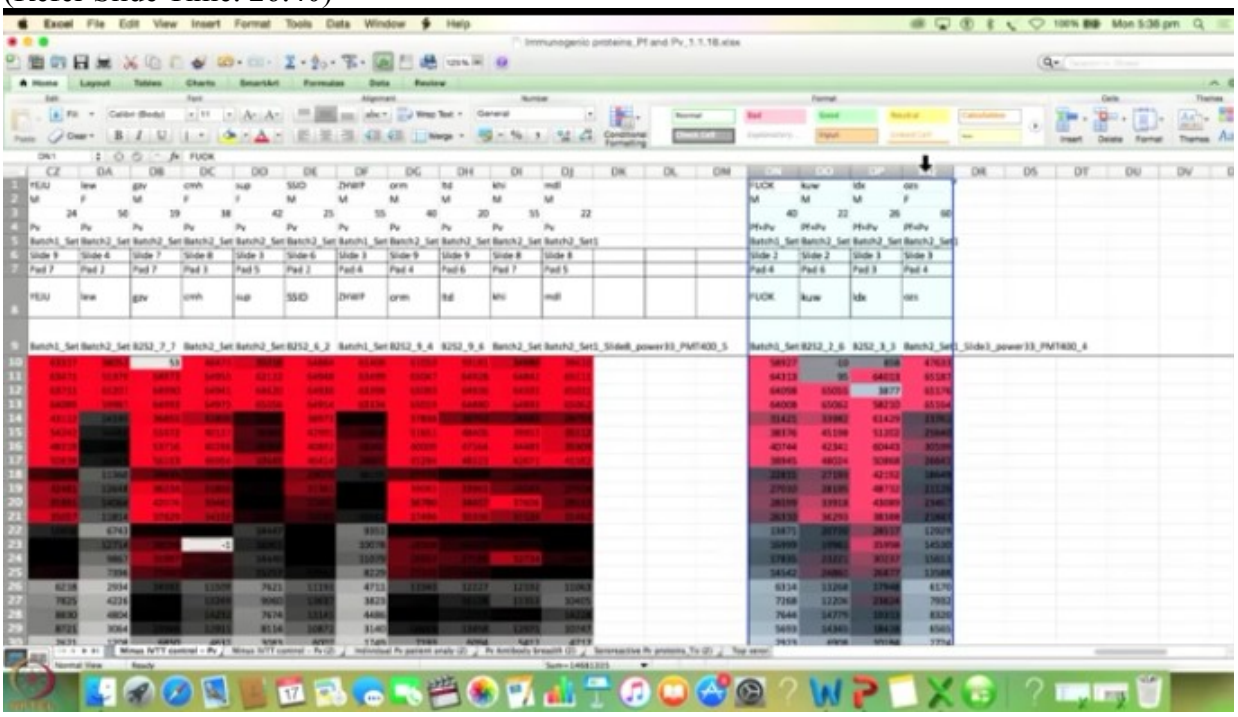
So now just to get this whole list of best zero reactive proteins, you can also do this on excel using a particular formula which I will show you now, so let's go back to the excel, this is how our excel was,

(Refer Slide Time: 26:16)



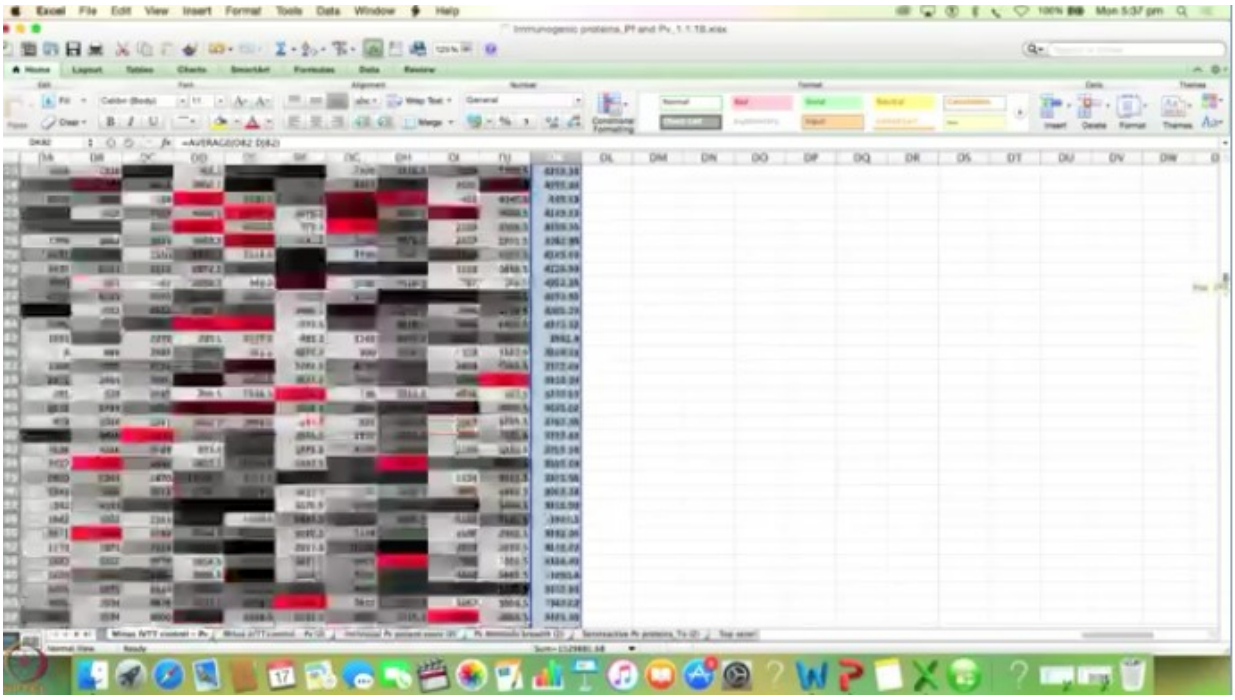
right, before we removed all of this rows which are IgG and anti-human IgG, so greeting them for now, and probably zoom this a little bit.

So now if you see that I have retained all the rows, so the first thing what we need to do of course we don't need this, but I have still retain the entire sheet from the beginning you will see that there are these 4 patients which are deliberately kept out of the analysis,
(Refer Slide Time: 26:40)



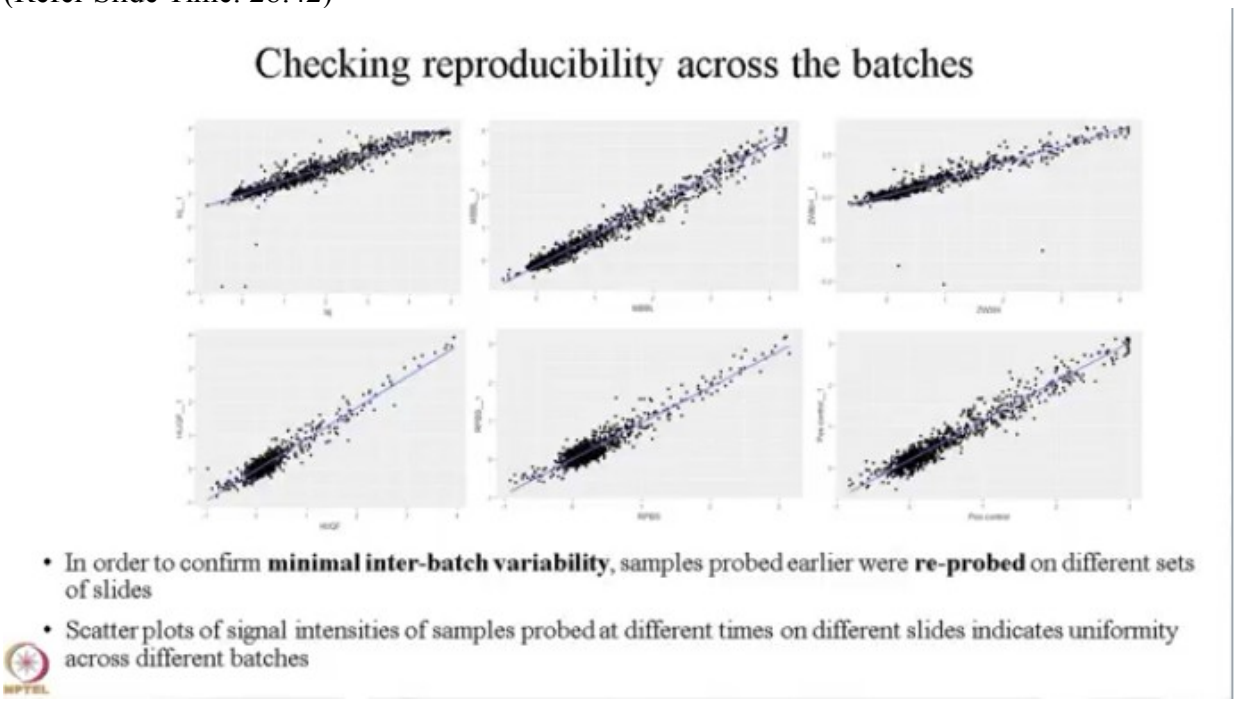
for example there are so if you see there is PF + PV everywhere right, so basically these are my patients who were diagnosed with mixed infection, so I don't want any such patients in my analysis so I'm going to purely have groups which are plasmodium falciparum and plasmodium vivax, and I'm going to look at, look for their response to plasmodium falciparum antigens and plasmodium vivax respectively, so I'm not going to have any of these mixed patients, so I have kept them out, so if you want we can also delete them, right.

So we need to basically start from row number 82 that's what we are interested in, because these are the IVTT spots, so the first thing I'm going to do is I'm going to an average for each and every spot, so example let's write here average, and I'm going to say is equal to, so I get an average value and I'm going to just drag this down, so you will have an average for each and every spot for all the patients.
(Refer Slide Time: 28:08)



What I missed telling you before is that the previous sheet had many more columns here, that's because we had a lot more samples which are probed on the chips, for example we had positive controls which were nothing but samples from taken from place which is a highly malaria endemic region, so you know that those spots have to give you a signal, right, so those are my positive control samples. So don't get confused between positive control samples and positive control spot, they are totally different, so these positive control samples I have excluded them from this particular analysis,

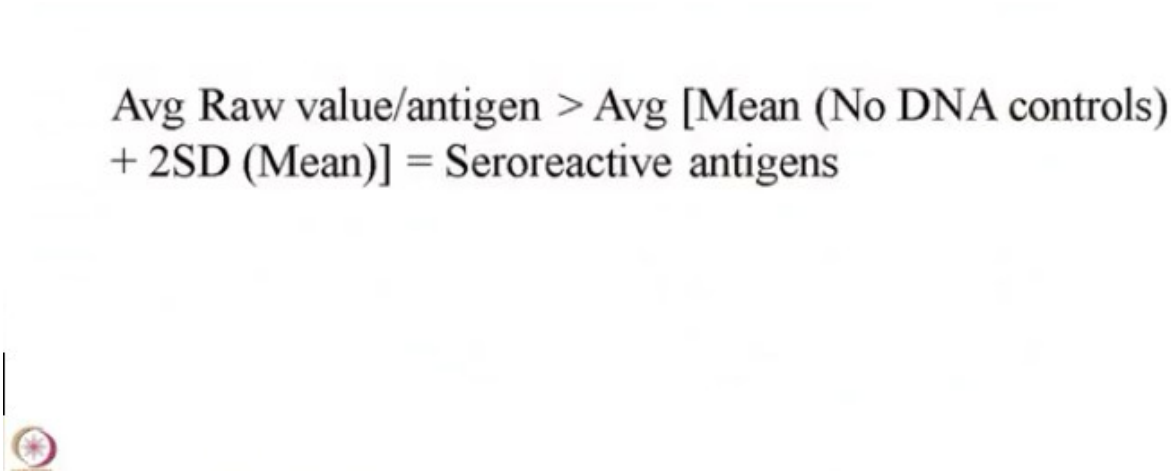
(Refer Slide Time: 28:42)



we also had healthy controls which are basically malaria nyu individuals, means patients were not detected with malaria at the time of admission, so they were malaria negative, those patients were also taken a probe on the chip just to see there is a difference in response, such patients have also removed from the analysis.

There is also certain samples which I have probed repeatedly in probably in duplicates or four times in all, you know once in all the sets, just to check for reproducibility, so here is some scatter plots you can see where I'm showing patient to patient reproducibility, so basically I'm showing reproducibility between my batch runs, so all of these patients also have removed from the analysis.

I have basically now in this excel 200 patients, 100 plasmodium vivax and 96 plasmodium falciparum patients, and four which are mixed infection also have removed, so in this way you can choose to remove rows and columns based on what you want to study and you can make a excel less complicated, right, so that's what I have missed mentioning, but now that is done, so I have taken an average right now.
(Refer Slide Time: 29:59)



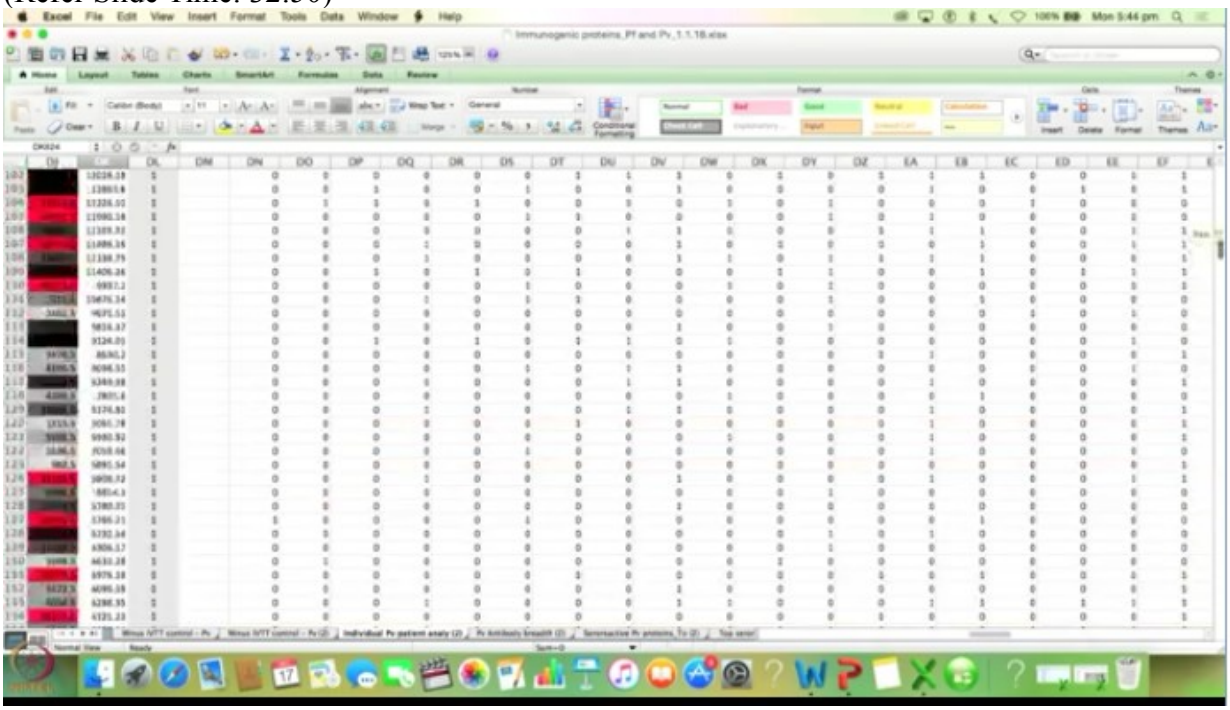
Now I am going to apply this particular formula which you see here, if the average value for particular spot is more than twice the standard deviation of the mean of the no DNA controls then that particular spot, then that particular antigen is zero reactive.

So what this means is that if this is the average if this particular number is greater than this particular number which I'm going to show you right now, if you take an average of the mean + 2 times standard deviation of the no DNA spots, this is my number, so if that raw value or if any raw value is greater than this value then that spot is basically zero reactive, then that antigen is, sorry then that antigen is basically zero reactive.

So I'm going to say if this is equal to, if function this spot is greater than this, then one is zero, so I get 1 here, and what if finally get is an excel sheet like this where I have random ones and zeros, right, so all of this ones I'm going to now say are my zero reactive proteins because they are greater than twice the standard deviation of my control spots.

Now a lot of people may also use healthy control of their analysis, right, but we don't have them, so what they do is they compare the signal intensities in a malaria group versus a healthy population, but since we don't have all that I'm going to simply say that this is my, these are the list of my zero reactive proteins which I can now take forward for further analysis, so this is not a great, this is not a statistical test, this is only shortlisting my proteins and I'm here only shortlisting my proteins from 1500 to handful which I can then take forward in study.

So this is what that sheet is, now what I have done here is that I have taken this for a group of patients, but now what if I want to check this for a particular patients, so that is what is my antibody breath which you will see here, I have done this individually for every single patient maybe I'll zoom this a little bit,
(Refer Slide Time: 32:50)

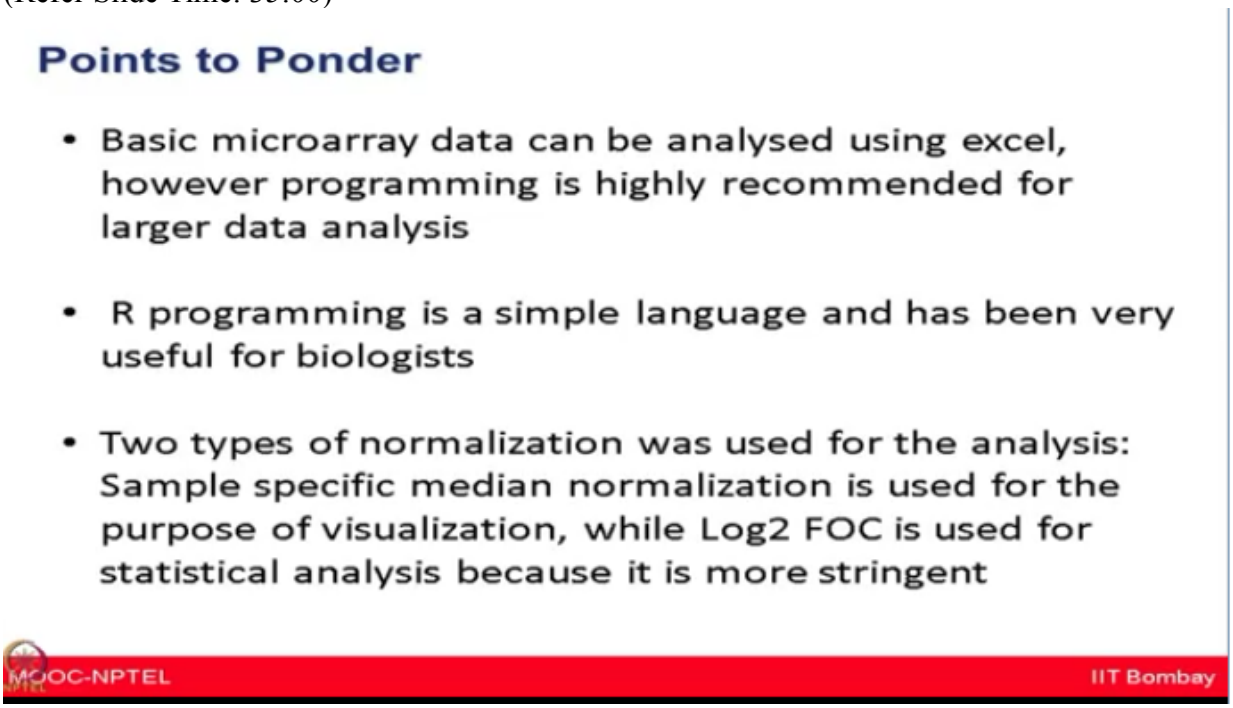


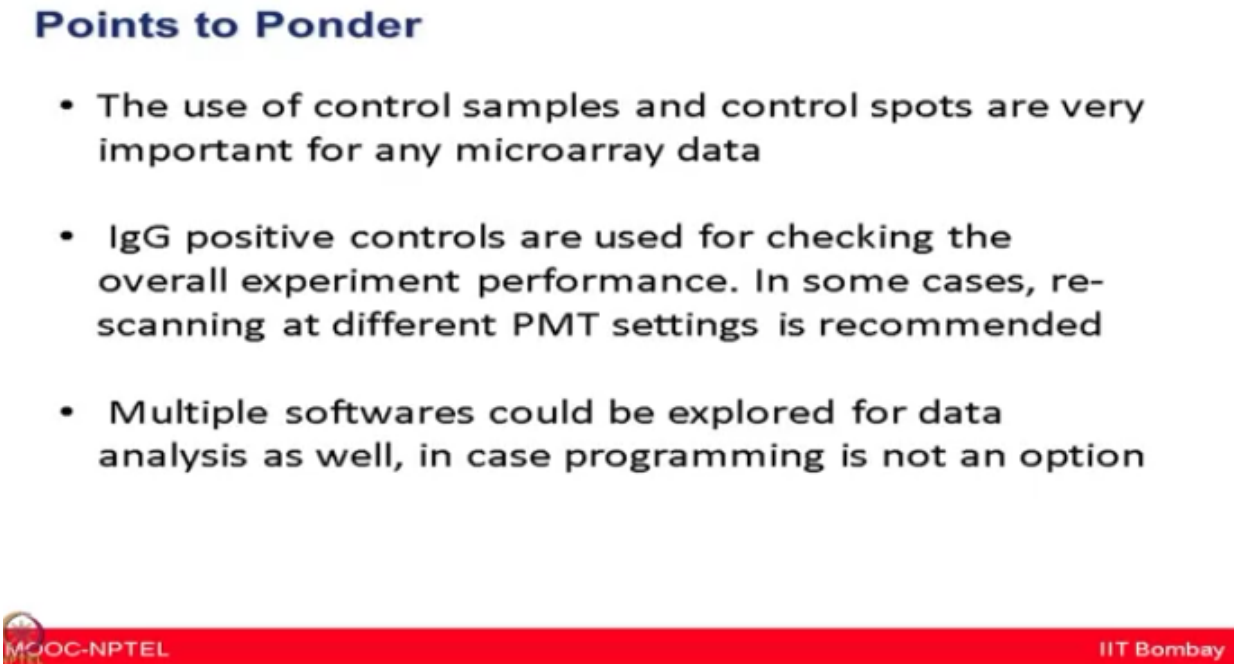
so what you see here is that I zoom this for every single patient so basically the previous one which I showed you was the average for a single spot for a group of patients, as well as the average for the no DNA control, here I have done it for each patient which means it is sample specific, right, so in this way if I scroll down what you will get here this, you will know the number of zero reactive antigens per patient, which means that if one patient, for example here is zero reactive to only 12 antigens, whereas there are some other patients which is zero reactive to 77 antigens, so this is basically my antibody breath.

So these are the two basic kind of analysis which I can show you in excel for now, power if microarray technology is basically the fact that you can perform this experiment very fast, probably in a day or two and then using any kind of patient data, all you need to do is map this

whole data which microarray data to each and every patient, clinical information that you have, and then you can perform any kind of statistical analysis and you can generate several results from the same single experiment, so that's the beauty of this.

I hope you have got a glimpse of how to perform data analysis and how basic statistics can be done, and how this is not the only way to do statistics at all, you can do use software and programming and I will still recommend that people do programming because if you want, even a single small change you don't have to repeat the entire analysis, also tomorrow if somebody provides you some other clinical information of the same patient population you don't have to repeat the analysis in excel, you can simply write a code for it and then in a few minutes you will get results for that as well, so that's all for now. Thank you.
(Refer Slide Time: 35:00)





Sanjeeva Srivastava: After going through this demonstration session and the insights of doing microarray based data analysis, you must have realized that there are many ways of analyzing and representing microarray data, of course there is no single way, no correct way of telling you what is the best way of doing it analysis, then many considerations you have to keep in mind when you are thinking about how to make meaningful information out of this high-throughput data.

There are several questions that can be answered using microarray data provided your data passes, the quality control chips and it is properly normalized, in such experiments you will control features becomes very crucial, both the positive controls and negative controls they guide you about how accurate and reload the data is, they could distinguish between real signals and background noise after proper analysis methods.

In the next class you will see another application of protein microarrays in a different application they will shift the gears to the cancer research and also the platforms. So far we have talked about self-array expression based protein microarray platform, we will now talk about how to take purified proteins printed on the chip using human proteome arrays and then apply those to investigate a deadly disease cancer, and try to talk to you about both excremental demonstrations as well as the theoretical concepts involved in performing such biological experiments. See you in lecture, thank you.

(Refer Slide Time: 36:53)

Prof. Sridhar Iyer

NPTEL Principal Investigator
&
Head CDEEP, IIT Bombay

Tushar R. Deshpande
Sr. Project Technical Assistant

Amin B. Shaikh
Sr. Project Technical Assistant

Vijay A. Kedare
Project Technical Assistant

Ravi. D Paswan
Project Attendant

Teaching Assistants

Apoorva Venkatesh

Nikita Gahoi

Shalini Aggarwal

Bharati Sakpal
Project Manager

Bharati Sarang

Project Research Associate

Nisha Thakur
Sr. Project Technical Assistant

Vinayak Raut
Project Assistant