Interactomics Basics and Applications
Prof. Sanjeeva Srivastava
Prof. Apoorva Venkatesh
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 17
Applications of protein microarrays in Malaria Research-II

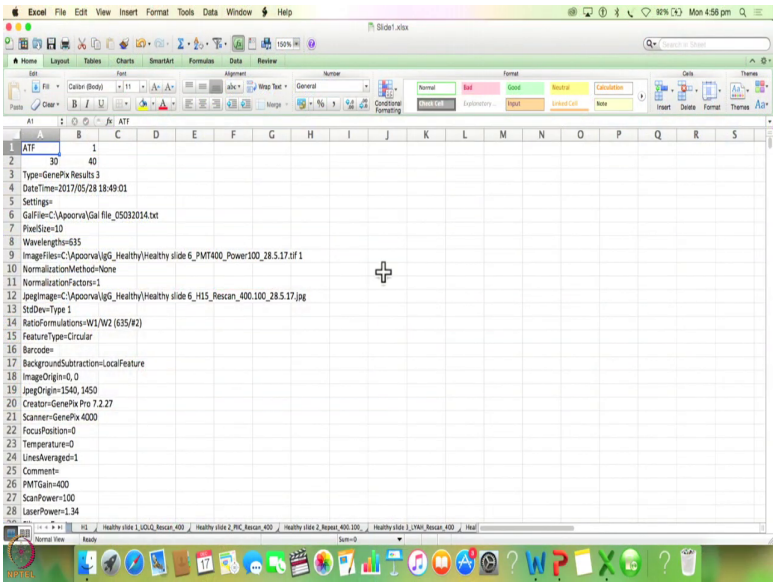
We are discussing about different microarray based platforms and how to perform some biological applications on these steps. In last lecture Ms. Apoorva Venkatesh showed you how to perform a microarray experiment using serum samples obtained from patients who had suffering from falciparum or vivax malaria.

Today we are going to continue the demonstration, and also show you the ways to do data normalization and how to do micro data analysis specifically if your goal was to look for a biological 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 us have that lecture and demonstration session today.

I am Apoorva Venkatesh your TA for this course. And today we are going to talk about microarray data normalization and analysis. In the last lecture we are trying to profile humeral responses of malaria positive patients using microarray technology. So, we are going to start from there.

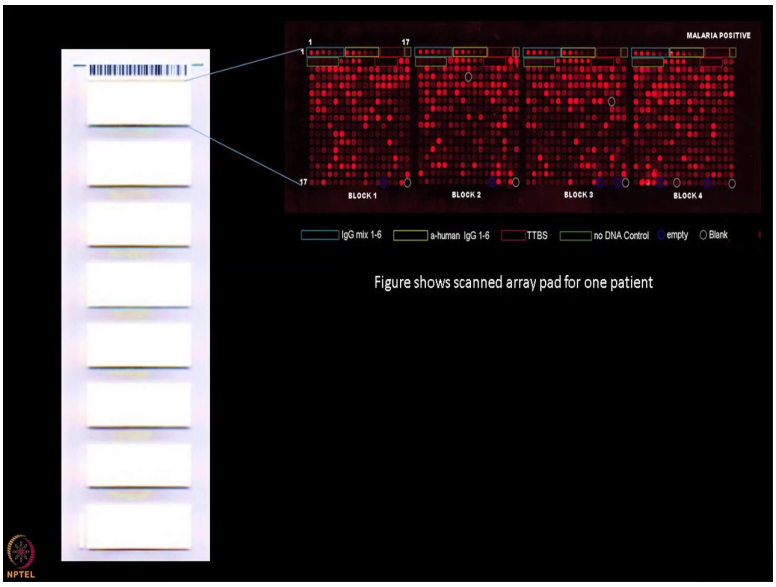
What we are going to do today is to see how to normalize microarray data using excel. So, what we will do is we will start with the raw file you get from the microarray scanner right. So, once you take your slide and you scan it in a scanner, you will extract the raw data.

(Refer Slide Time: 01:57)



And here is the excel sheet you see. This is the type of data you get, I am showing you this one particular slide. So, I will just like to repeat that one slide can probe 8 patients serum.

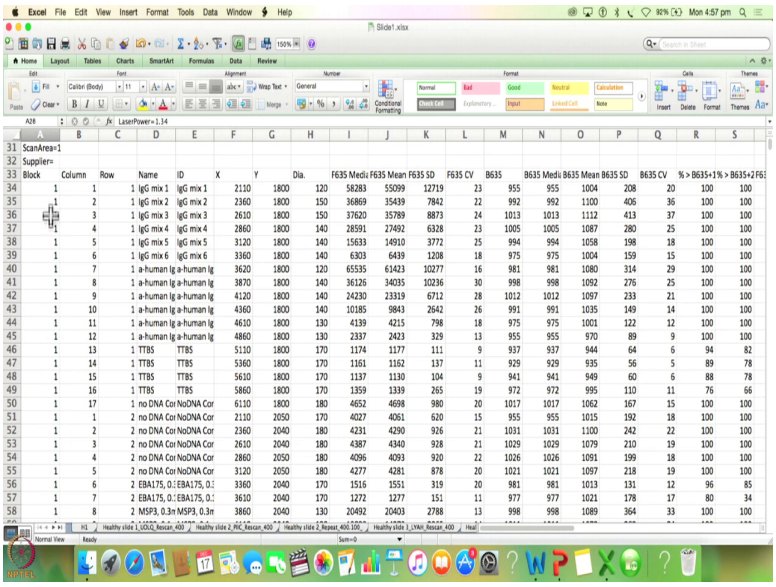
(Refer Slide Time: 02:07)



So, here in this one particular excel which you see here, we actually have data for 8 patients. So, first what I am going to show you is how to reorganize this data ok. So, let us 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 scanned at a wavelength of 635 nanometer.

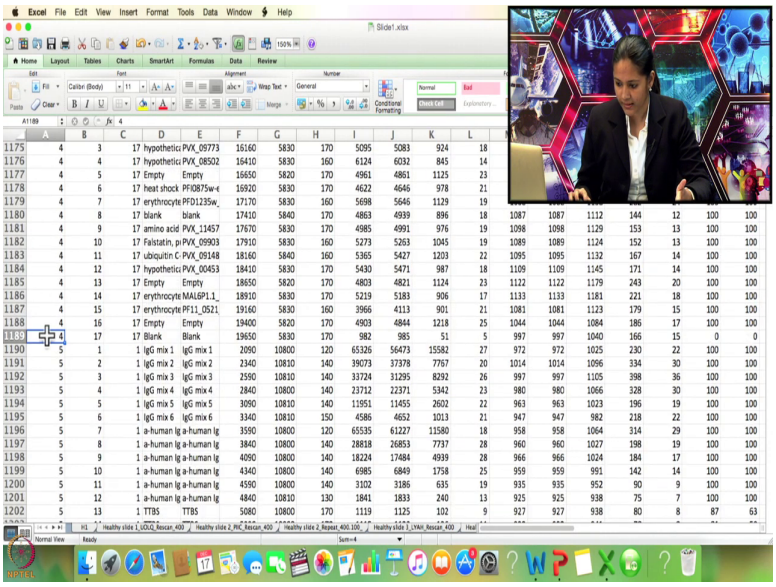
Then you go down normalization method this was not normalized yet, so it says none. Then if you scroll down further you can see the PMT gain which is 400, scan power 100, laser power 1.34. So, basically later on if you want to go back and check these slides again, if for example, differentiation if you forget the parameters we used, you can always go and open this excel to see what you had done right.

(Refer Slide Time: 02:59)



So, now let us scroll down further, you will see block column and row. So, this is very important again, let us go back to the slide layout one slide can probe 8 patients 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 1 patient data right.

(Refer Slide Time: 03:37)



So, then when I keep scrolling down and I go to block 5 a new patient begins, so that is what I am going to talk to you about how to reorganize this. So, for example, if you see here four ends here right with blank and you will see that there is an IgG 1, IgG mix 1 which starts again. So, this is basically your new patient.

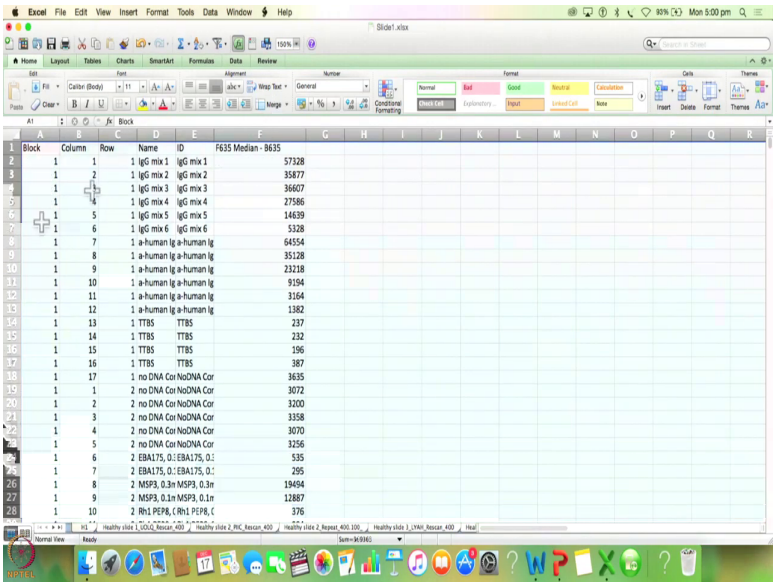
So, what we are going to do is, we are going to first reorganize this, but before this let me tell you which are the columns which are important for us. So, now I am going to scroll back up, and we are going to go through the columns which we have on this excel. So, now, apart from block column, row, the name, and the id basically this is your protein id, we do not need any of the other columns except for column.

So, you see column here. So, this is basically your F 635 median minus B 635 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 do not need any other column here. So, what I am going to do, I am going to first delete all unwanted columns to make this excel less complicated.

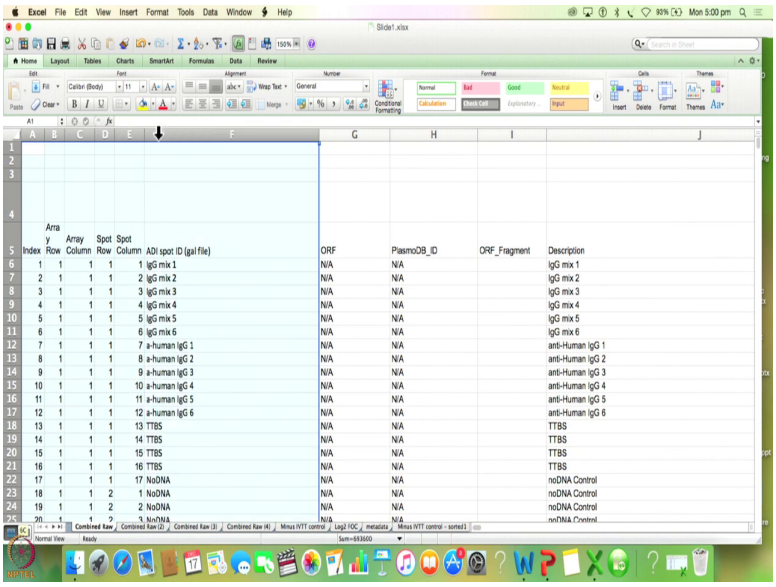
So, let us delete all of this, and then we go and delete this as well. We also do not need these parameters for the analysis, I am also going to delete this. So, finally, this is what you get.

(Refer Slide Time: 04:55)



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.

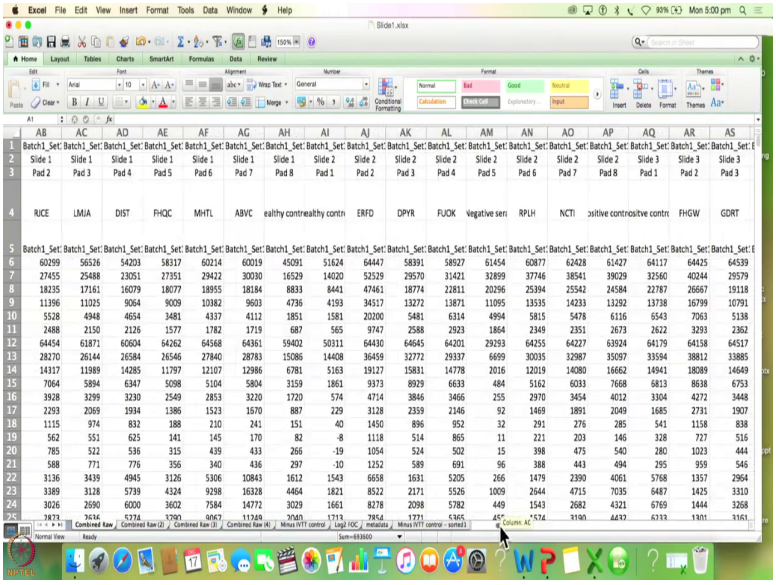
(Refer Slide Time: 05:07)



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
20	1	1	2	3	NoDNA	N/A	N/A		noDNA Control

So, you also see that there are additional columns here this is what you get from your gal file.

(Refer Slide Time: 05:13)

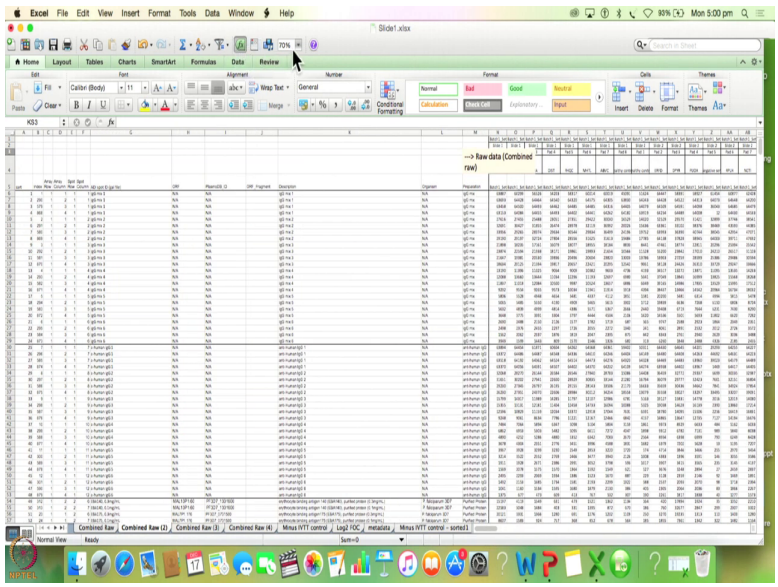


And when you scroll right, you see that you all the patients are now placed next to each other right. So, this is that kind of excel you get first. Now, what I am going to do is, I am going to reorganize this excel to make it easier. This is combined data for all your patients and now we will reorganize the proteins. So, as you will see the 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 will 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 4 blocks of the same pad.

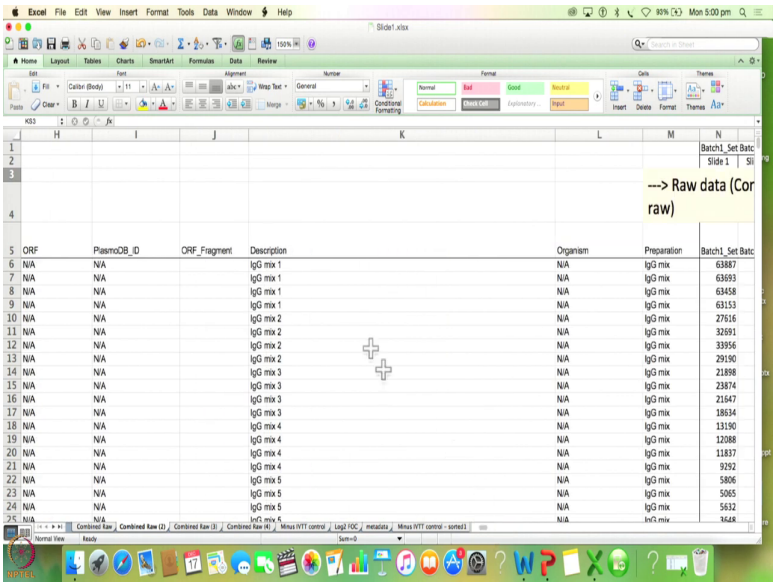
So, what I am trying to do here, I am trying to get all the IgG mix together. So, you will have 24 such spots one after the other. We will also do the same thing for your anti human IgG mix. And similarly we are also going to do the same thing for your next spots.

(Refer Slide Time: 06:19)



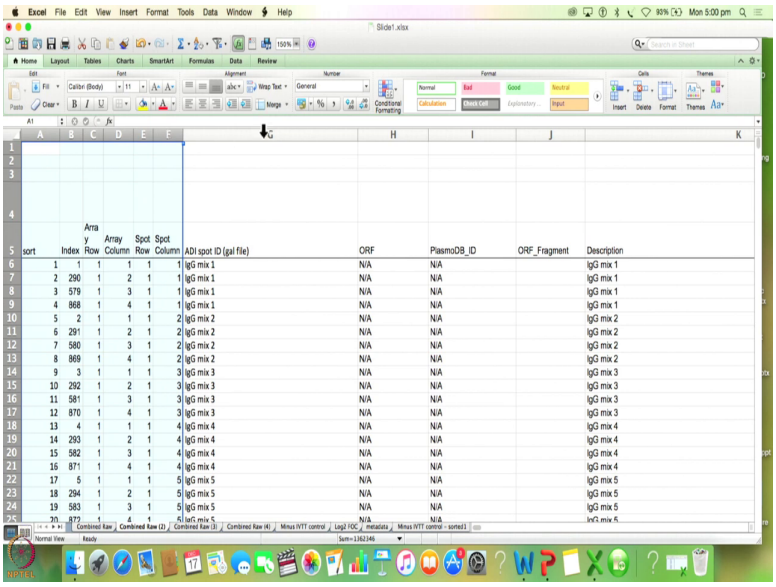
So, when we rearrange our excel, this is how your excel will look.

(Refer Slide Time: 06:27)



What I have done here I have put all the 24 IgG mix of 1 pad together right.

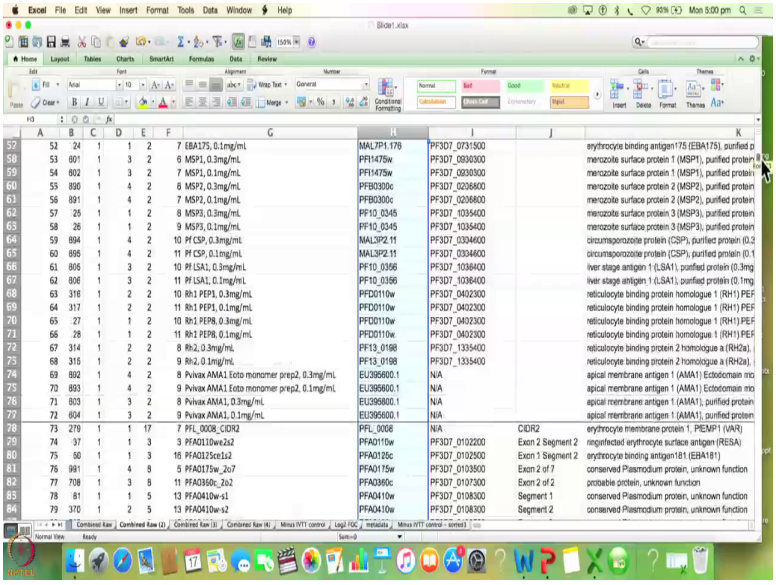
(Refer Slide Time: 06:35)



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
20	872	1	4	1	5	IgG mix 5	N/A	N/A		IgG mix 5

We will go through all the columns once more. For example, these are all just your spot details. Now, we will go to your ADI spot id which is your gal file. So, this is what you get from your gal file, I will come to this in a minute. Before that let us talk about ORF. So, this column here is basically your id.

(Refer Slide Time: 06:55)



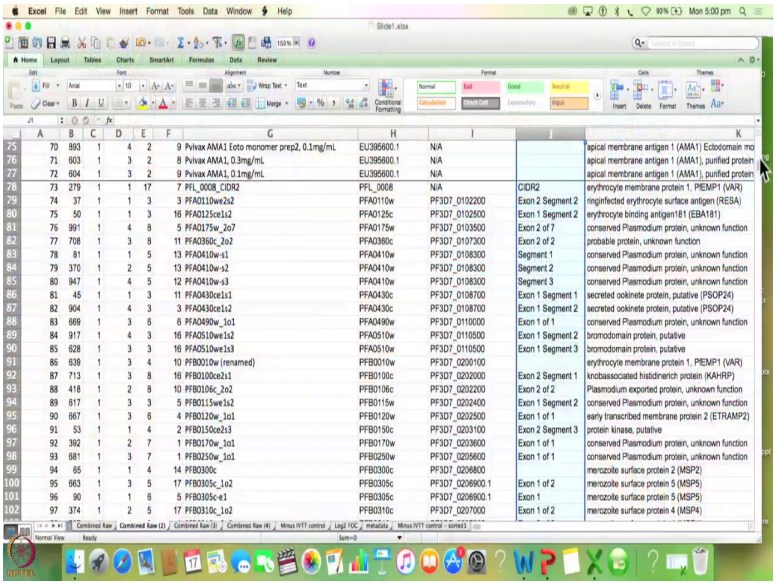
This is also going to give you details about the fragment which has been printed on the chip. So, let us go to a plasmo DB ID. Now, if you look at plasmo DB ID these are all basically each and every protein has a unique plasmo DB ID, so that is what is mentioned in this column here. If you go to the next column which is ORF fragment, 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 chip. So, basically as you know what is 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 is not really right to say that

proteins were expression on the chip, instead you be better to say that polypeptides were expressed on the chip.

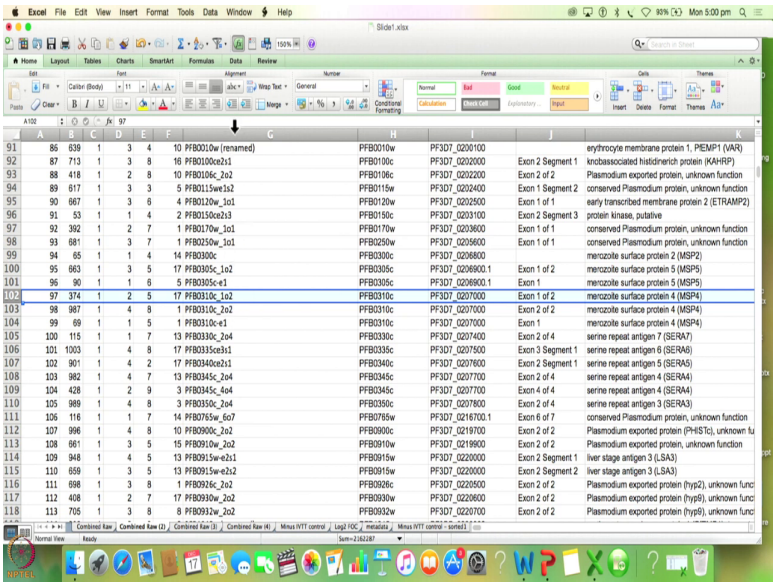
(Refer Slide Time: 07:55)



So, this particular column J gives us details about the polypeptide that was expressed in printed on the chip right. So, that is how you get this ADI spot spot id which is a 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 ad I sport id you will not find single do any single duplicate, because these are unique ids for each and every protein which takes into account the exon fragment which was printed on the chip, so that is what you see here.

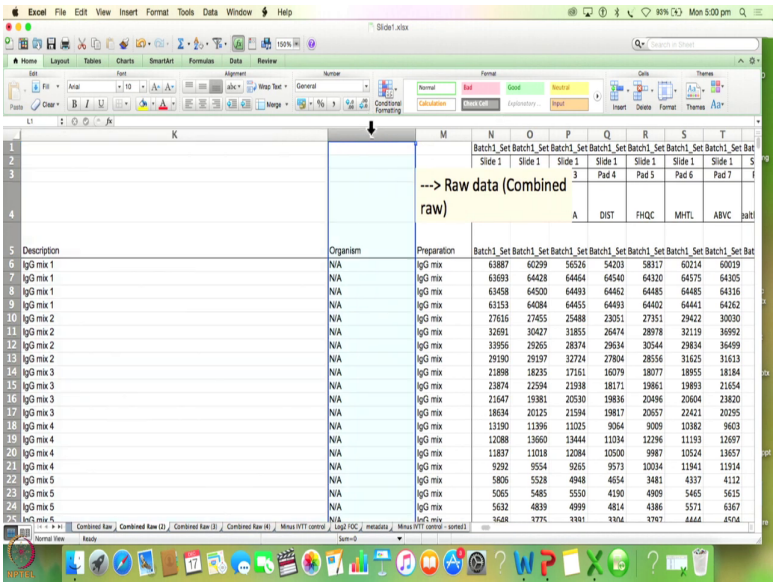
(Refer Slide Time: 08:35)



If you say for example, let us look at this particular row 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 am telling you all this is because this is very important for data analysis, for all sometimes you might just start with an analyzing your iref column, and then you will figure out later that there are a lot of duplicates and you do not know what you are actually doing.

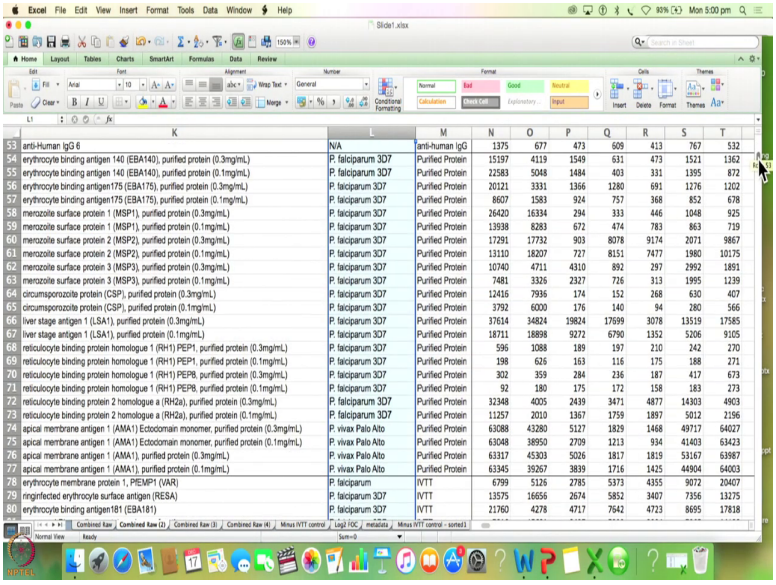
So, what we need to do is if you want to shortlist any antigens, we need to consider the G column for analysis right. So, now let us move onto the next column which is your description. So, this we all know, this basically describe what was printed on the chip right these are just basically the names of them, basically the names of the antigens.

(Refer Slide Time: 09:29)



Now, the next column is your organism. So, as you know you have two types of spots your plasmodium falciparum and plasmodium vivax.

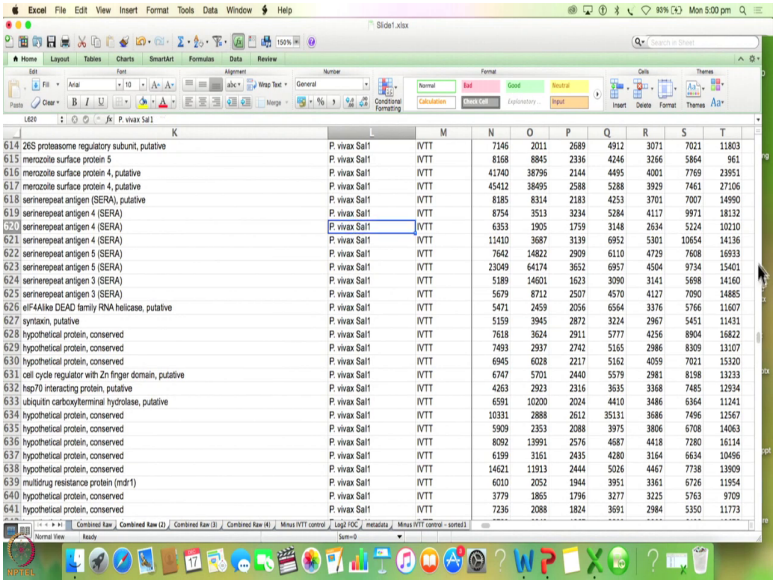
(Refer Slide Time: 09:39)



	K	L	M	N	O	P	Q	R	S	T
53	anti-Human IgG 6	N/A	anti-human IgG	1375	677	473	609	413	767	532
54	erythrocyte binding antigen 140 (EBA140), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	15197	4119	1549	631	473	1521	1362
55	erythrocyte binding antigen 140 (EBA140), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	22583	5048	1484	403	331	1395	872
56	erythrocyte binding antigen175 (EBA175), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	20121	3331	1366	1280	691	1276	1202
57	erythrocyte binding antigen175 (EBA175), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	8607	1583	924	757	368	852	678
58	merozoite surface protein 1 (MSP1), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	26420	16334	294	333	446	1048	925
59	merozoite surface protein 1 (MSP1), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	13938	8283	672	474	783	863	719
60	merozoite surface protein 2 (MSP2), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	17291	17732	903	8078	9174	2071	9867
61	merozoite surface protein 2 (MSP2), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	13110	18207	727	8151	7477	1980	10175
62	merozoite surface protein 3 (MSP3), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	10740	4711	4310	892	297	2992	1891
63	merozoite surface protein 3 (MSP3), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	7481	3326	2327	726	313	1995	1239
64	circumsporozoite protein (CSP), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	12416	7936	174	152	268	630	407
65	circumsporozoite protein (CSP), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	3792	6000	176	140	94	280	566
66	liver stage antigen 1 (LSA1), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	37614	34824	19824	17699	3078	13519	17585
67	liver stage antigen 1 (LSA1), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	18711	18898	9272	6790	1352	5206	9105
68	reticulocyte binding protein homologue 1 (RH1) PEP1, purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	596	1088	189	197	210	242	270
69	reticulocyte binding protein homologue 1 (RH1) PEP1, purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	198	626	163	116	175	188	271
70	reticulocyte binding protein homologue 1 (RH1) PEP8, purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	302	359	284	236	187	417	673
71	reticulocyte binding protein homologue 1 (RH1) PEP8, purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	92	180	175	172	158	183	273
72	reticulocyte binding protein 2 homologue a (RH2a), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	32348	4005	2439	3471	4877	14303	4903
73	reticulocyte binding protein 2 homologue a (RH2a), purified protein (0.1mg/mL)	P. falciparum 3D7	Purified Protein	11257	2010	1367	1759	1897	5012	2196
74	apical membrane antigen 1 (AMA1) Ectodomain monomer, purified protein (0.3mg/mL)	P. vivax Palo Alto	Purified Protein	63088	43280	5127	1829	1468	49717	64027
75	apical membrane antigen 1 (AMA1) Ectodomain monomer, purified protein (0.1mg/mL)	P. vivax Palo Alto	Purified Protein	63048	38950	2709	1213	934	41403	63423
76	apical membrane antigen 1 (AMA1), purified protein (0.3mg/mL)	P. vivax Palo Alto	Purified Protein	63317	45303	5026	1817	1819	53167	63987
77	apical membrane antigen 1 (AMA1), purified protein (0.1mg/mL)	P. vivax Palo Alto	Purified Protein	63345	39267	3839	1716	1425	44904	64003
78	erythrocyte membrane protein 1, PfEMP1 (VAR)	P. falciparum	IVTT	6799	5126	2785	5373	4355	9072	20407
79	ringinfected erythrocyte surface antigen (RESA)	P. falciparum 3D7	IVTT	13575	16656	2674	5852	3407	7356	13275
80	erythrocyte binding antigen181 (EBA181)	P. falciparum 3D7	IVTT	21760	4278	4717	7642	4723	8695	17818

So, basically this is going to tell you which organism does the antigen belong to. 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.

(Refer Slide Time: 09:51)



	K	L	M	N	O	P	Q	R	S	T
614	26S proteasome regulatory subunit, putative	P. vivax Sal1	IVTT	7146	2011	2689	4912	3071	7021	11803
615	merozoite surface protein 5	P. vivax Sal1	IVTT	8168	8845	2336	4246	3266	5864	961
616	merozoite surface protein 4, putative	P. vivax Sal1	IVTT	41740	38796	2144	4495	4001	7769	23951
617	merozoite surface protein 4, putative	P. vivax Sal1	IVTT	45412	38495	2588	5288	3929	7461	27106
618	serinerepeat antigen (SERA), putative	P. vivax Sal1	IVTT	8185	8314	2183	4253	3701	7007	14990
619	serinerepeat antigen 4 (SERA)	P. vivax Sal1	IVTT	8754	3513	3234	5284	4117	9971	18132
620	serinerepeat antigen 4 (SERA)	P. vivax Sal1	IVTT	6353	1905	1759	3148	2634	5224	10210
621	serinerepeat antigen 4 (SERA)	P. vivax Sal1	IVTT	11410	3687	3139	6952	5301	10654	14136
622	serinerepeat antigen 5 (SERA)	P. vivax Sal1	IVTT	7642	14822	2909	6110	4729	7608	16933
623	serinerepeat antigen 5 (SERA)	P. vivax Sal1	IVTT	23049	64174	3652	6957	4504	9734	15401
624	serinerepeat antigen 3 (SERA)	P. vivax Sal1	IVTT	5189	14601	1623	3090	3141	5698	14160
625	serinerepeat antigen 3 (SERA)	P. vivax Sal1	IVTT	5679	8712	2507	4570	4127	7090	14885
626	eIF4Alike DEAD family RNA helicase, putative	P. vivax Sal1	IVTT	5471	2459	2056	6564	3376	5766	11607
627	syntaxin, putative	P. vivax Sal1	IVTT	5159	3945	2872	3224	2967	5451	11431
628	hypothetical protein, conserved	P. vivax Sal1	IVTT	7618	3624	2911	5777	4256	8904	16822
629	hypothetical protein, conserved	P. vivax Sal1	IVTT	7493	2937	2742	5165	2986	8309	13107
630	hypothetical protein, conserved	P. vivax Sal1	IVTT	6945	6028	2217	5162	4059	7021	15320
631	cell cycle regulator with Zn finger domain, putative	P. vivax Sal1	IVTT	6747	5701	2440	5579	2981	8198	13233
632	hsp70 interacting protein, putative	P. vivax Sal1	IVTT	4263	2923	2316	3635	3368	7485	12934
633	ubiquitin carboxyterminal hydrolase, putative	P. vivax Sal1	IVTT	6591	10200	2024	4410	3486	6364	11241
634	hypothetical protein, conserved	P. vivax Sal1	IVTT	10331	2888	2612	35131	3686	7496	12567
635	hypothetical protein, conserved	P. vivax Sal1	IVTT	5909	2353	2088	3975	3806	6708	14063
636	hypothetical protein, conserved	P. vivax Sal1	IVTT	8092	13991	2576	4687	4418	7280	16114
637	hypothetical protein, conserved	P. vivax Sal1	IVTT	6199	3161	2435	4280	3164	6634	10496
638	hypothetical protein, conserved	P. vivax Sal1	IVTT	14621	11913	2444	5026	4467	7738	13909
639	multidrug resistance protein (mdr1)	P. vivax Sal1	IVTT	6010	2052	1944	3951	3361	6726	11954
640	hypothetical protein, conserved	P. vivax Sal1	IVTT	3779	1865	1796	3277	3225	5763	9709
641	hypothetical protein, conserved	P. vivax Sal1	IVTT	7236	2088	1824	3691	2984	5350	11773

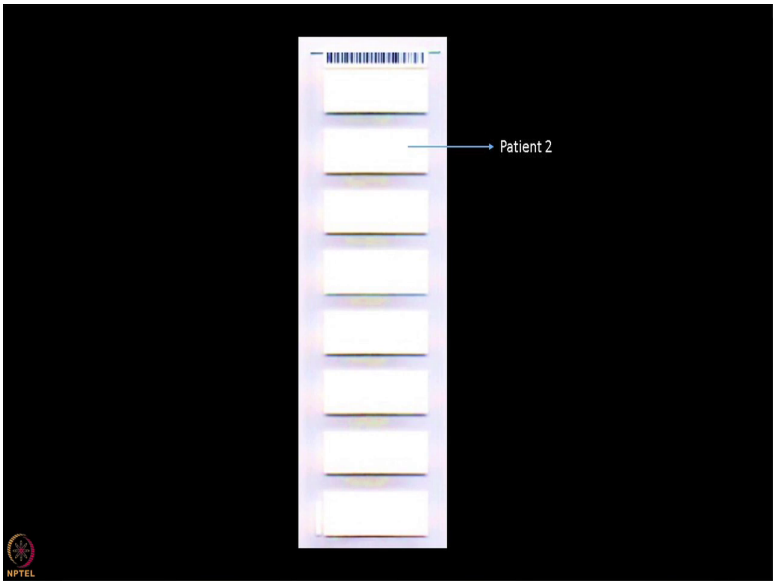
And then the next column which is m is going to talk to you about the preparation of the spot. Like for example, you have the first few spots are basically your IgG mix right. So, then the preparation is basically your IgG mix, it is not an IVTT spot. Now, if you scroll down further you will have similarly anti human IgG. Again you scroll down further you have certain purified proteins which are nothing but control proteins.

So, our control spots here were printed as purified proteins and not as IVTT spots. Now, if you scroll down further you will find all your other spots basically your antigens which you are trying to study had 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 patient samples.

So, if I just move this a little bit, what do you see here for example, let us 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 2 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 is telling me this is telling me that this particular positive control was probed on batch 1 set 1 on slide 1 and pad 1 right. Now, similarly, so let us go to the next one which is a real sample that was just a positive control. So, this is basically probed on batch 1, set 1, slide 1 and pad 2.

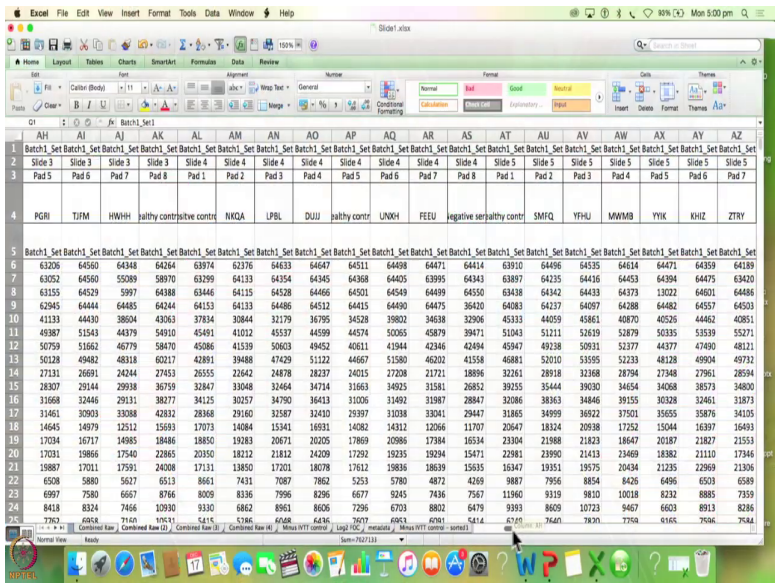
(Refer Slide Time: 11:41)



So, this is going to tell me my position of the sample. So, if ever I want to go back to the slides and to check the real spots right the images of the spots, then I will know exactly where

to go. So, for example, if some sample is b is not behaving well and I want to go and cross check the intensity of the spots. For example, some sample is giving me very high intensity signals and I want to go back and check whether it is real, then I know exactly which file to open because I have all the details here.

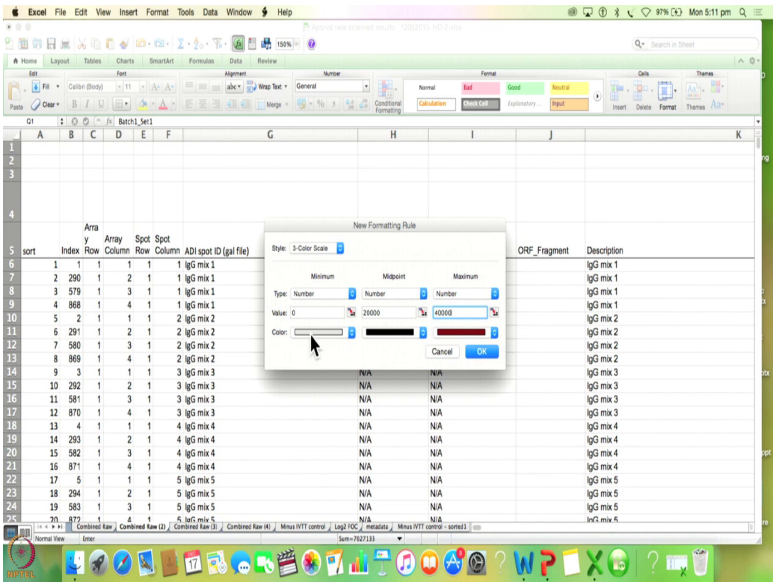
(Refer Slide Time: 12:07)



So, this is for the all other columns. So, this is I hope you have 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 color gradient to this excel right. And now I will tell you why we are going to apply the gradient. Let us first do that.

So, for which what I am going to do is I am going to go to conditional formatting, I am going to go to color scales more rules. And I am going to choose a three color scale and then I am going to choose number type ok.

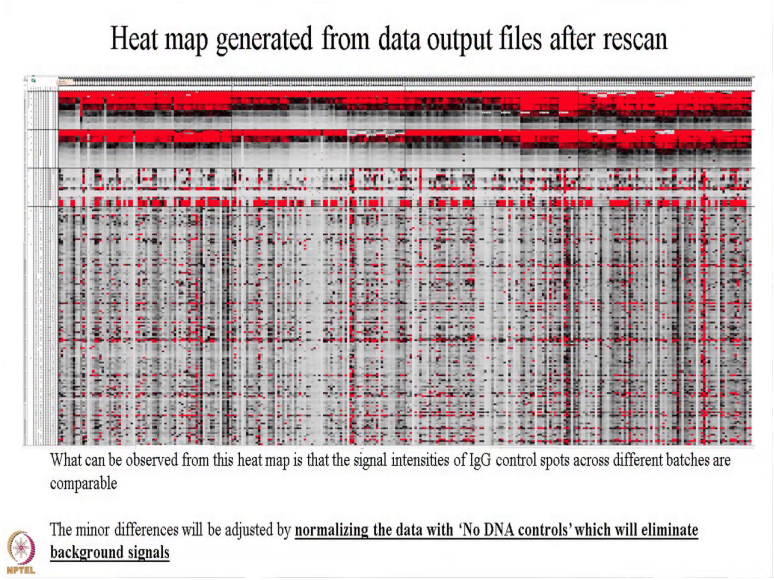
(Refer Slide Time: 12:53)



So, I am going to say 0. And I am going to assume that my entire data falls you know in between say certain negative values and maybe around 80000 is my maximum value. So, I am just going to assume that if my data falls in this range I am going to split my data be based on three numbers – 0, then my midpoint will be say 20000 and my highest will be 40000. And I am going to choose some colors here.

So, I say this is maybe gray, then I am going to keep this black, and I am going to keep this red. So, what this is going to do is all my values above 40000 are going to be in dark red, and then around 20000 will be black and the lowest or the least values will be gray and those which are in negatives will be almost white, so that is how I am going to apply a color gradient here.

(Refer Slide Time: 13:53)



So, you can see in this slide here basically what I have done is I have just minimized this excel a little bit. So, you can you will be able to see all four batches at once. So, I do not know if you can see a black line here. So, basically this is going to split your batches, so in fact, it is going to split your set, set 1 from set 2 of the first batch and set 1 and 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 color gradient, what you can see is that the signal intensities particularly for this batch, in fact this whole batch, but batch 1, 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 your IgG mix, and this particular line which you see here is your anti human igg.

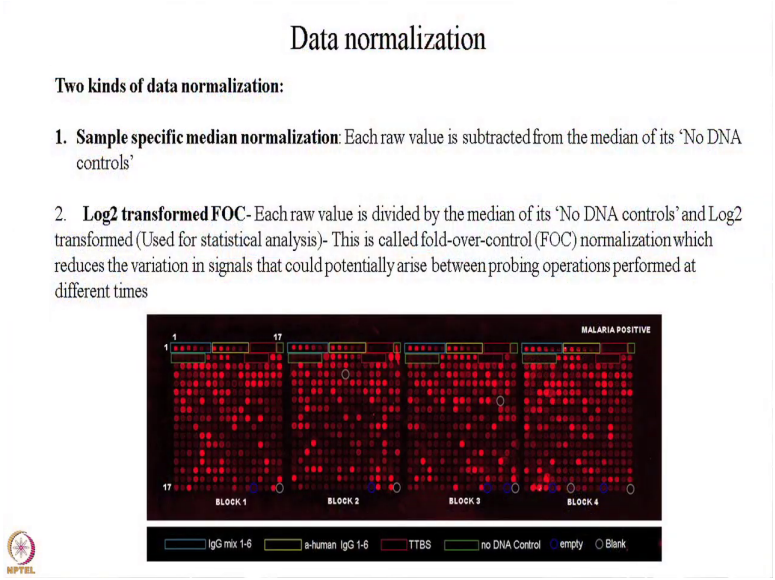
So, basically this is your control which is going to tell you where you need to rescan your slide or not. So, if this is very high, then all your signals by default for this particular batch will also be high right, so that is going to screw up your results a little screw up your results later because 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 the strip is going to basically help you in deciding whether you need to rescan your slides are different PMT settings and power 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 realize if 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 ok.

So, now having rescanned all the slides, as you can see in this slide the settings look pretty uniform, though it is still not very uniform and you still feel that batch 2 set 1 has higher signals, but overall it is because this will then be taken care of by normalization. So, now what we do is we will proceed with normalization using excel.

(Refer Slide Time: 16:21)



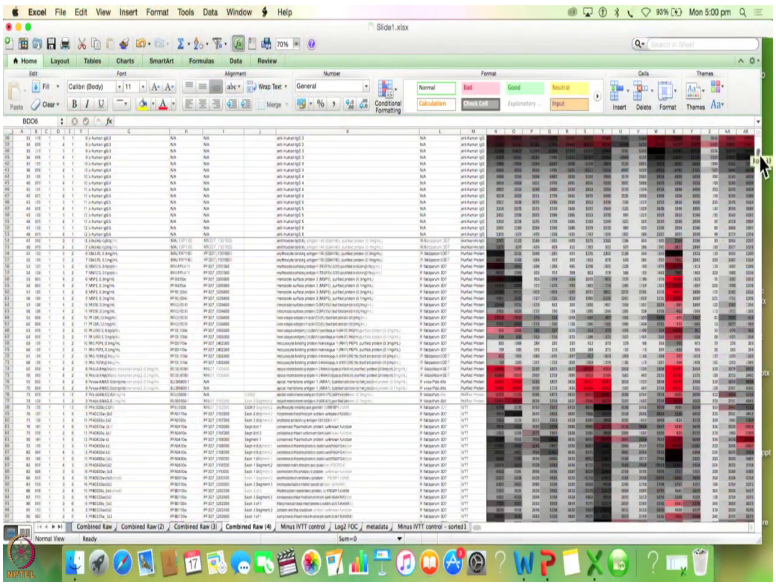
Now, there are two strategies I am going to talk to you about today. The first strategy is basically a very simple normalization method which we will use only for visualization. For example, if you want to prepare heat maps, then we will use this the first normalization method. However, if you want to perform statistical tests, then we will use the second kind of normalization which I will talk to you about.

So, let us 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 this sample specific median value of the no DNA controls. So, I am sure that this is a little confusing. So, what we will do is we will go step by step.

First I am going to show you what raw values are and then I am going to show you what the no DNA controls are, right. So, again we are going to come back to the same it is color coded

and you have view we have reached the stage. You also know that this now in this data, we have IgG makes, we have anti human IgG, we have purified proteins, we do not need any of those right now for our analysis, we are going to directly go down to the IVTT spots.

(Refer Slide Time: 17:19)

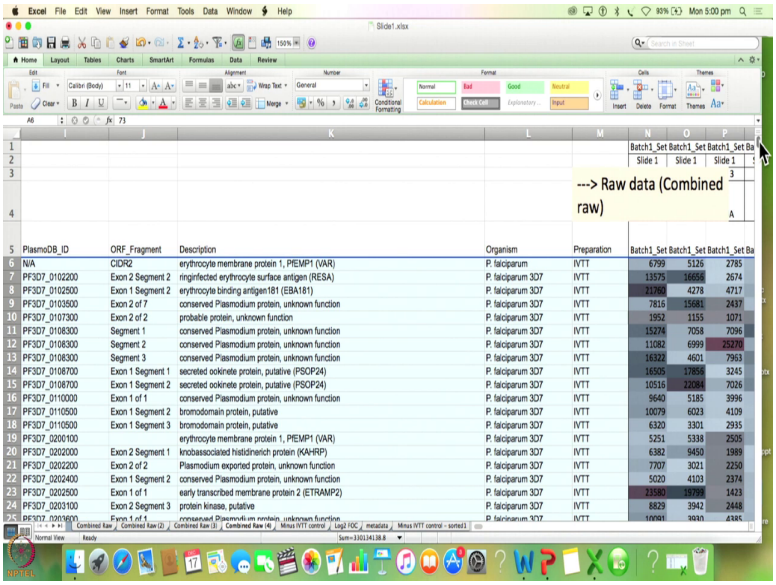


So, in fact, what we will do is we probably just delete those rows to avoid confusion. So, let us start from here. I am going to delete the first few maybe what I will do is I will just zoom this a little bit. So, I have just zoomed this a little bit. What we are going to do is we are going to delete unwanted rows right now. So, we do not want IgG mix, we do not want to anti human IgG, we do not want purified proteins right now.

Again let me tell you the purified proteins basically you do not require in the analysis, but it is important when for example, your slide is not worked at all. And or you have not got the

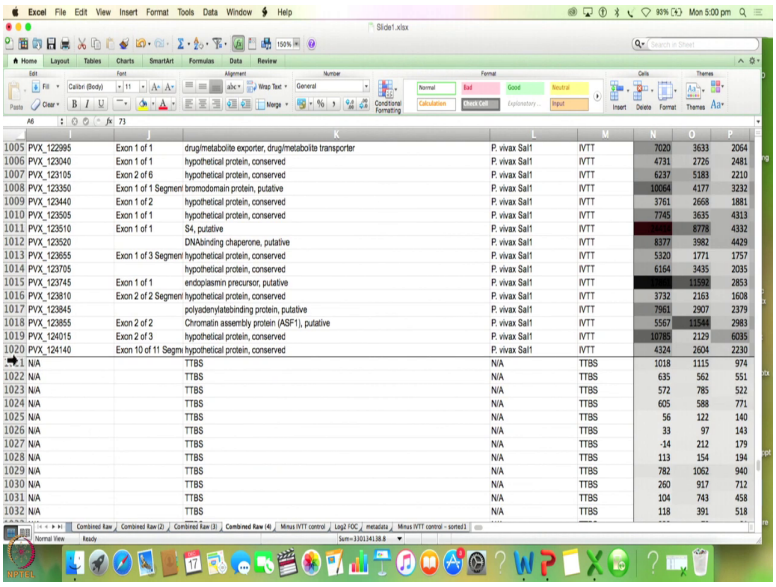
signals you required. You can always go back to the positive control spots to see what they signals were right. So, this is basically used for such you know analysis just as controls.

(Refer Slide Time: 18:13)



So, right now we are going to delete those rows. And we are going to only keep rows which are IVTT mix right that is what this is. So, now we are going to have this wave 500 plasmodium falciparum IVTT spots, and 515 plasmodium falciparum vivax IVTT spots. So, we are going to go down, you have deleted unwanted rows. There are a few more rows below which we do not need.

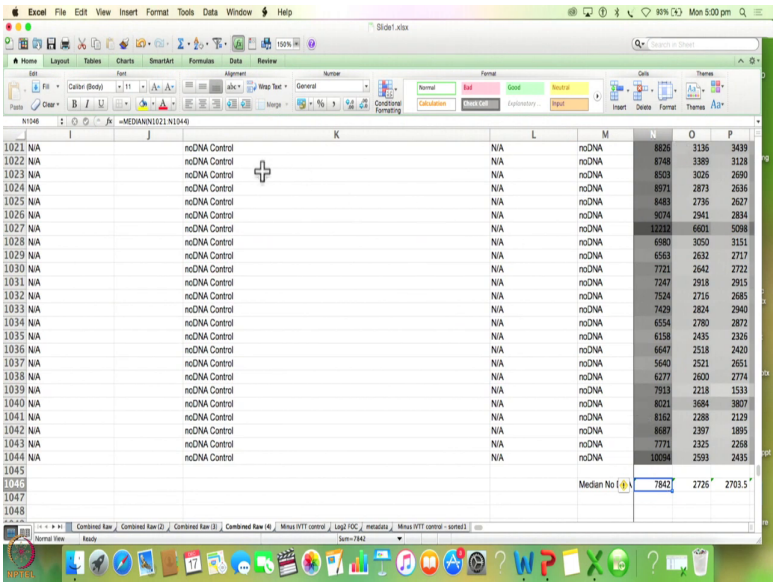
(Refer Slide Time: 18:37)



1005	PVX_122995	Exon 1 of 1	drug/metabolite exporter, drug/metabolite transporter	P. vivax Sal1	IVTT	7020	3633	2064
1006	PVX_123040	Exon 1 of 1	hypothetical protein, conserved	P. vivax Sal1	IVTT	4731	2726	2481
1007	PVX_123105	Exon 2 of 6	hypothetical protein, conserved	P. vivax Sal1	IVTT	6237	5183	2210
1008	PVX_123350	Exon 1 of 1 Segment	bromodomain protein, putative	P. vivax Sal1	IVTT	10064	4177	3232
1009	PVX_123440	Exon 1 of 2	hypothetical protein, conserved	P. vivax Sal1	IVTT	3761	2668	1881
1010	PVX_123505	Exon 1 of 1	hypothetical protein, conserved	P. vivax Sal1	IVTT	7745	3635	4313
1011	PVX_123510	Exon 1 of 1	S4, putative	P. vivax Sal1	IVTT	[illegible]	8778	4332
1012	PVX_123520		DNAbinding chaperone, putative	P. vivax Sal1	IVTT	8377	3982	4429
1013	PVX_123655	Exon 1 of 3 Segment	hypothetical protein, conserved	P. vivax Sal1	IVTT	5320	1771	1757
1014	PVX_123705		hypothetical protein, conserved	P. vivax Sal1	IVTT	6164	3435	2035
1015	PVX_123745	Exon 1 of 1	endoplasmin precursor, putative	P. vivax Sal1	IVTT	[illegible]	11592	2853
1016	PVX_123810	Exon 2 of 2 Segment	hypothetical protein, conserved	P. vivax Sal1	IVTT	3732	2163	1608
1017	PVX_123845		polyadenylatebinding protein, putative	P. vivax Sal1	IVTT	7961	2907	2379
1018	PVX_123855	Exon 2 of 2	Chromatin assembly protein (ASF1), putative	P. vivax Sal1	IVTT	5567	11544	2983
1019	PVX_124015	Exon 2 of 3	hypothetical protein, conserved	P. vivax Sal1	IVTT	10785	2129	6035
1020	PVX_124140	Exon 10 of 11 Segm	hypothetical protein, conserved	P. vivax Sal1	IVTT	4324	2604	2230
1021	N/A		TTBS	N/A	TTBS	1018	1115	974
1022	N/A		TTBS	N/A	TTBS	635	562	551
1023	N/A		TTBS	N/A	TTBS	572	785	522
1024	N/A		TTBS	N/A	TTBS	605	588	771
1025	N/A		TTBS	N/A	TTBS	56	122	140
1026	N/A		TTBS	N/A	TTBS	33	97	143
1027	N/A		TTBS	N/A	TTBS	-14	212	179
1028	N/A		TTBS	N/A	TTBS	113	154	194
1029	N/A		TTBS	N/A	TTBS	782	1062	940
1030	N/A		TTBS	N/A	TTBS	260	917	712
1031	N/A		TTBS	N/A	TTBS	104	743	458
1032	N/A		TTBS	N/A	TTBS	118	391	518

So, after these 15000 spots, 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.

(Refer Slide Time: 18:57)



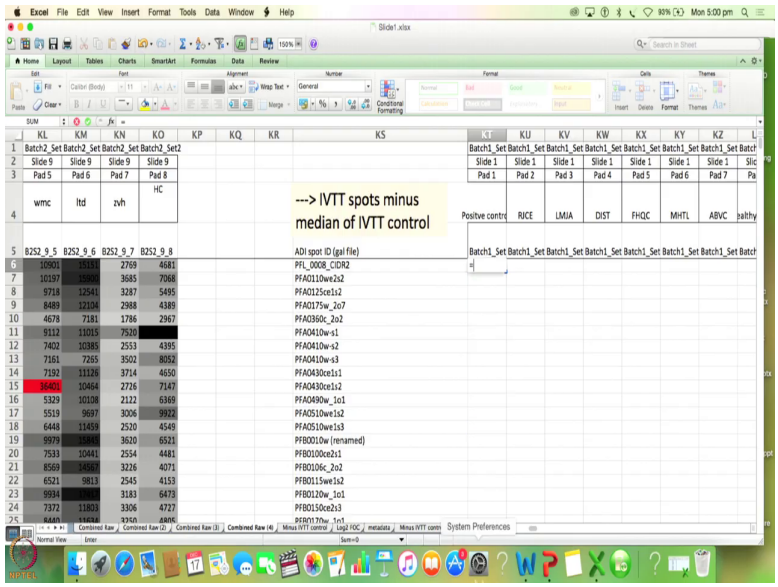
So, now what we have are 15000 IVTT spots and 24 no DNA control spots. So, now what are these no DNA control spots? So, basically these pots have the entire IVTT mix except the plasmid. So, basically what you expect here is no expression, because you do not even have the plasmid here, whereas the IVTT 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 do not have that here.

So, what is this going to provide is going to provide your background signal. So, what we are trying to do in the first type of normalization is, we are subtracting our raw signals from background. So, now there are 24 such spots which you remember we have rearranged and that is why it is come together group 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 have just done this the whole thing in excel. So, now we have a median value here. So, the first thing what I am going to do is for this particular sample which is in column n, this is one sample, I am going to subtract the values for each and every IVTT spot from that particular median right.

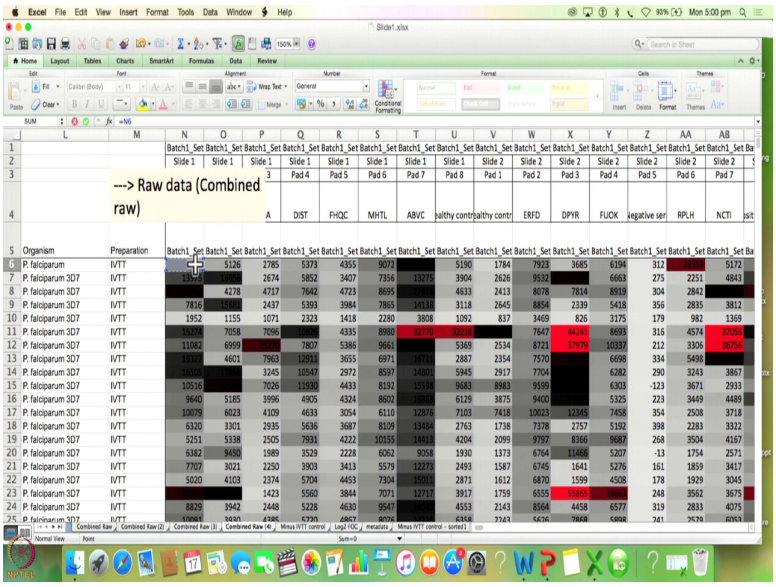
So, each of these spots which you see here, I am going to subtract it and that is what is my median sample specific median normalization. So, let us scroll, so probably we will do this in the same excel ok. I have kept placed by that here.

(Refer Slide Time: 20:23)



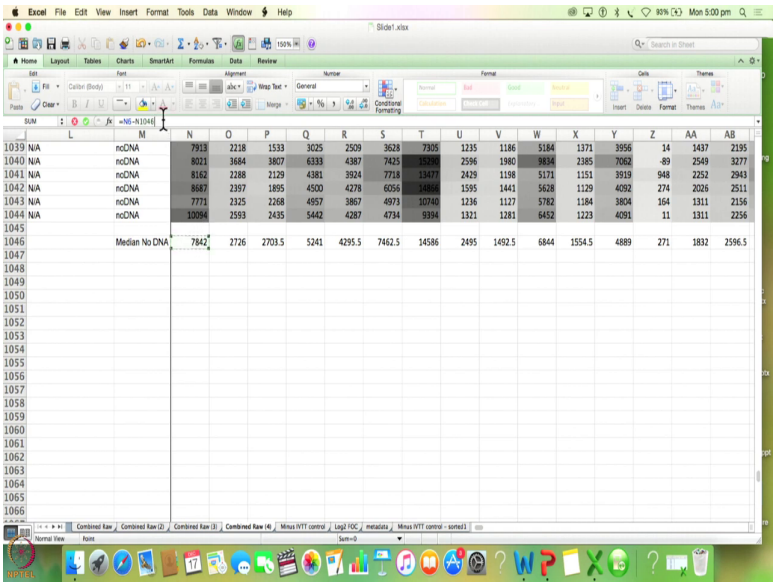
So, this is called IVTT sports minus median of IVTT control, so that is exactly what we have to do we are going to say is equal to, then we are going to go to that particular spot. So, say let us take the first patient.

(Refer Slide Time: 20:45)



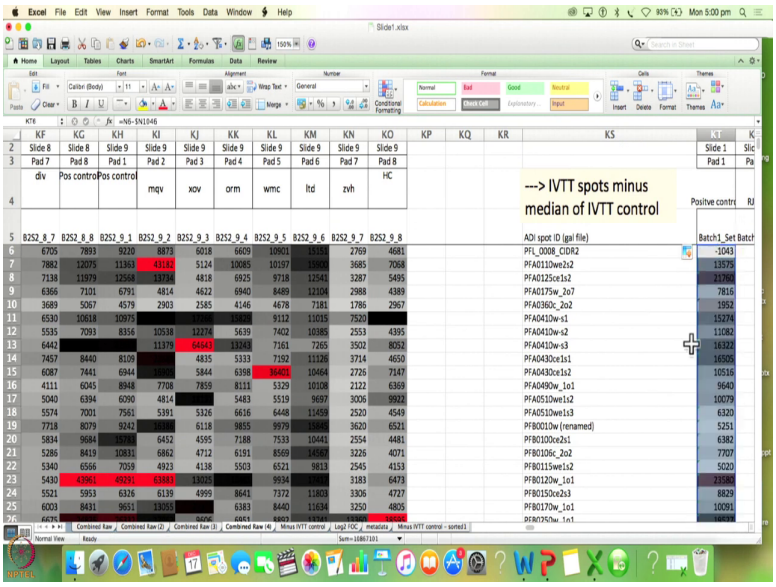
And then we see minus and we go to the median value, which is 7842.

(Refer Slide Time: 20:55)



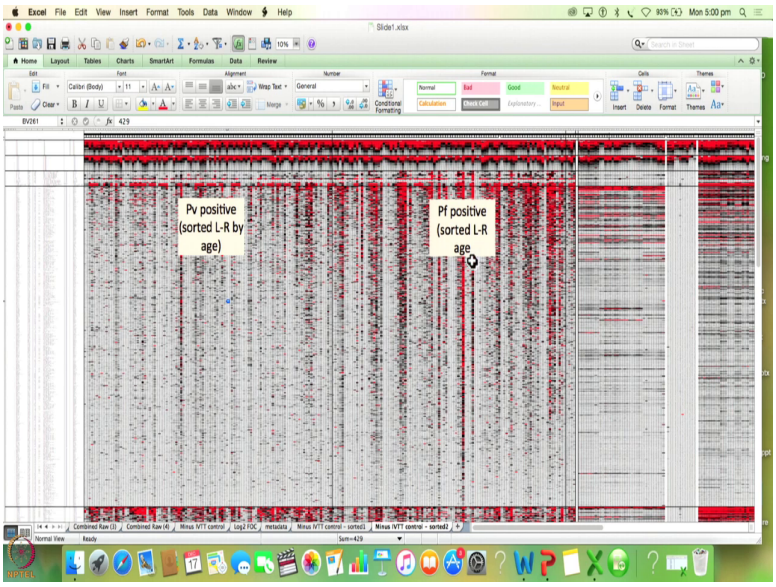
So, now, because I want this row to remain constant throughout I am going to put a dollar sign in front of the row.

(Refer Slide Time: 21:15)



So, this is what we get here. And now I am just going to drag this across as well as down.

(Refer Slide Time: 21:25)



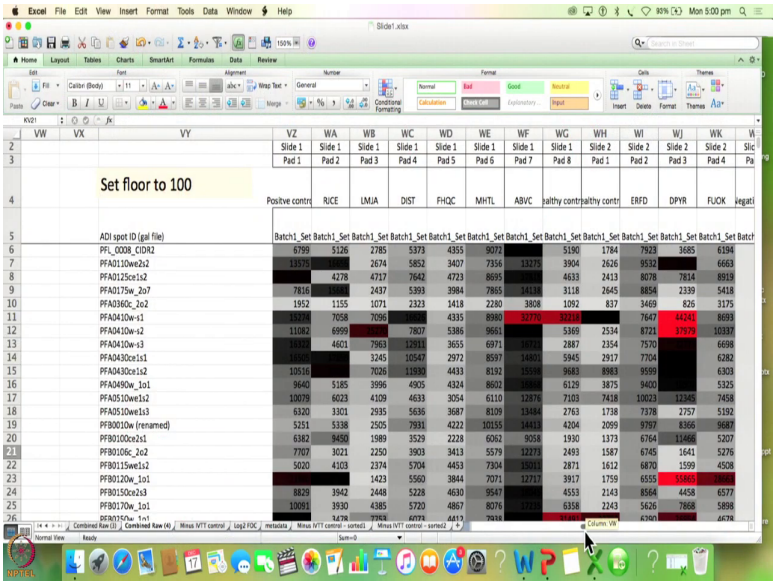
So, once you drag and drop, this is the kind of excel you get. I have just minimized this, but if you apply your color gradient this is how it looks overall. So, this is what you can use now to make your heat maps and what I have also done is I have sorted this based on the antigens as well as the patients who were a falciparum positive and vivax positive. I have 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. So, I have put all your P v positive patients together and P f positive together and I have also sorted based on age. So, you have P v positive, P f positive, as well as sorted by age. So, this way you can sort your excel in different ways.

You can also use other softwares to make a heat map, but basically this kind of not once you normalize it in this way, you do not perform any statistical analysis with this data. For

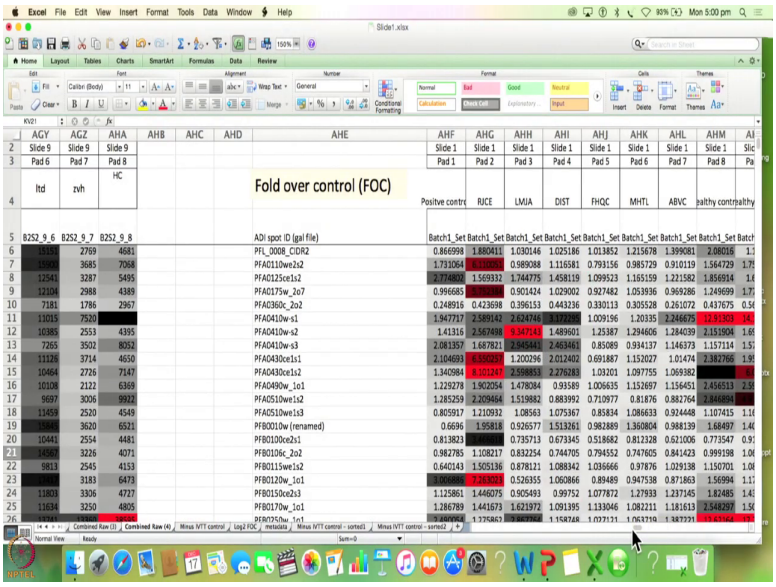
statistical analysis, I am going to now show you the next normalization method which is your log 2 transform fold over control normalization.

(Refer Slide Time: 22:33)



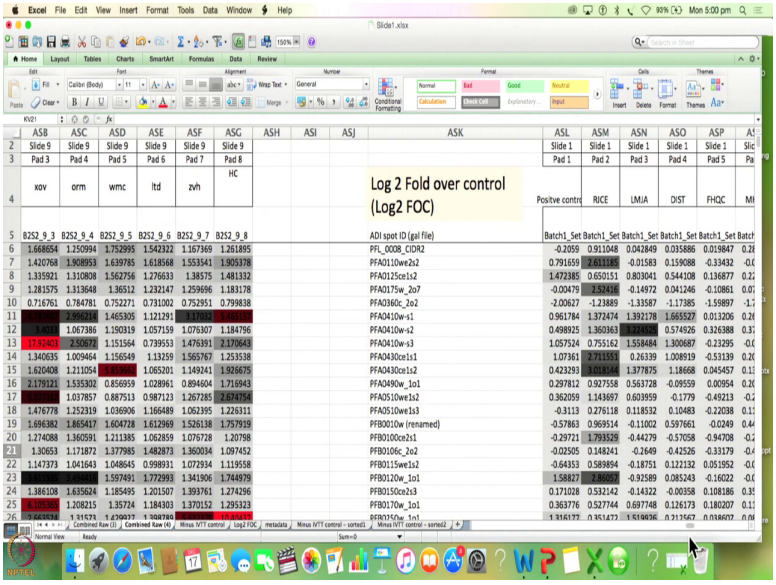
So, for this I am not going to show you the entire method again because now you know how to do it on excel, I am sure you all know. For this I am only going to show you the steps, the first thing what we do is we want to set a floor of 100. So, what with what we are trying to say here is that all the 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 is the first thing I have done it here for you and we are going to keep scrolling right.

(Refer Slide Time: 23:05)



The next step what we are going to do is to divide each and every raw value by the median of the IVTT controls control spots. So, just like 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 is what is called fold over control.

(Refer Slide Time: 23:29)



So, once you set up flow of 100, then you divide it. And the next thing you are going to do is to convert this whole data into log values. So, you log to transform this entire data right and that is why it is 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 ok.

So, now either you can use programming to do your statistical analysis or you can use different softwares which provide your statistical tests, 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 which software you can you want to you want to use.

For example, graph pad prism is an excellent tool for preparing 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,

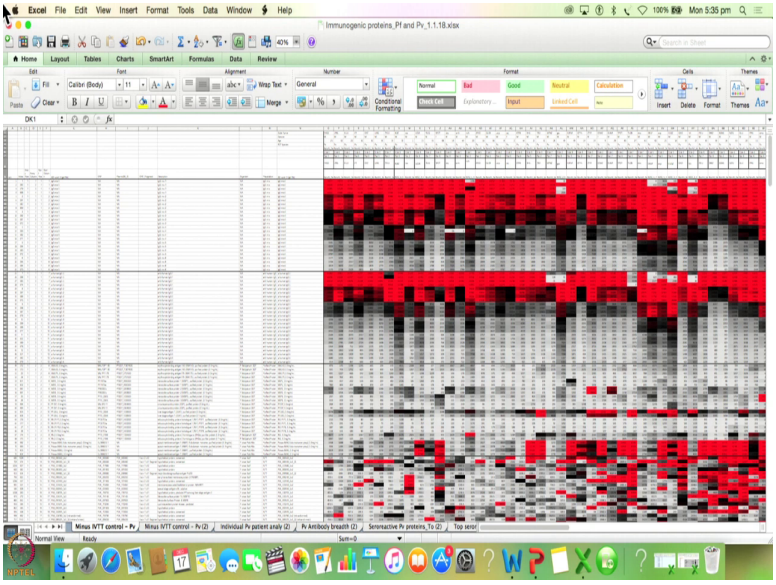
but still it is not very small for graph pad prism. So, for example, you can have graph pad prism can get stuck in the middle if you are using data even this size. So, of course, if you have bigger data sets, then it is very difficult to use software slipped off graph pad prism.

However, if you are going to have only 10 patients or 20 patients with 40 proteins or something graph pad prism does offer you a lot of statistical tests. Apart from that there are other softwares as well. You have metaboanalyst. Though it is for metabolomics data, you still have a module called significance analysis of 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 save a lot of time as well. So, what I am going to do towards this is we are coming to the end of the lecture, I am only going to show you very small analysis you do one I have done on excel.

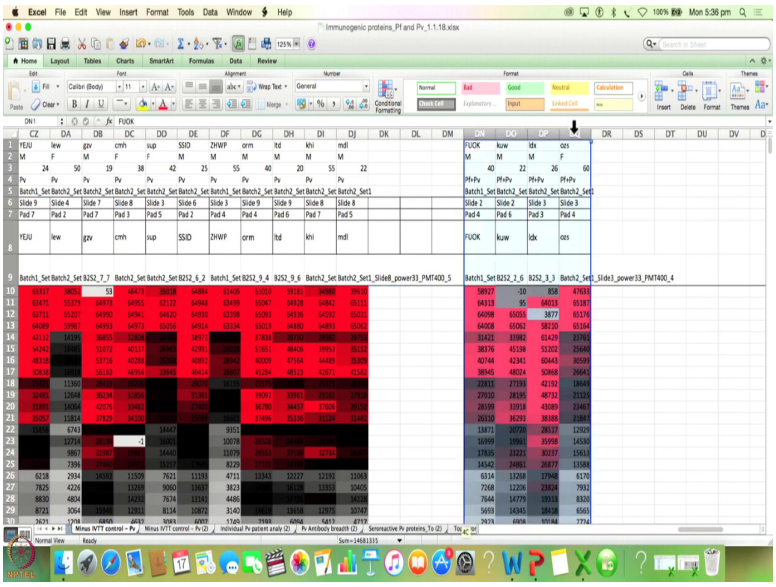
So, basically what my aim here is to identify most zero reactive proteins in my from my chip which means that the proteins which elicit the maximum antibody response in malaria patients right, so that is my aim. So, now just to get this whole list of best zero reactive proteins, I you can also do this on excel using a particular formula which I will show you now.

(Refer Slide Time: 25:47)



So, let us go back to the excel. This is how our excel was right. Before we removed all of these rows which are I have IgG and anti human igg. So, I have written 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 do not need this, but I have still retained the entire sheet from the beginning.

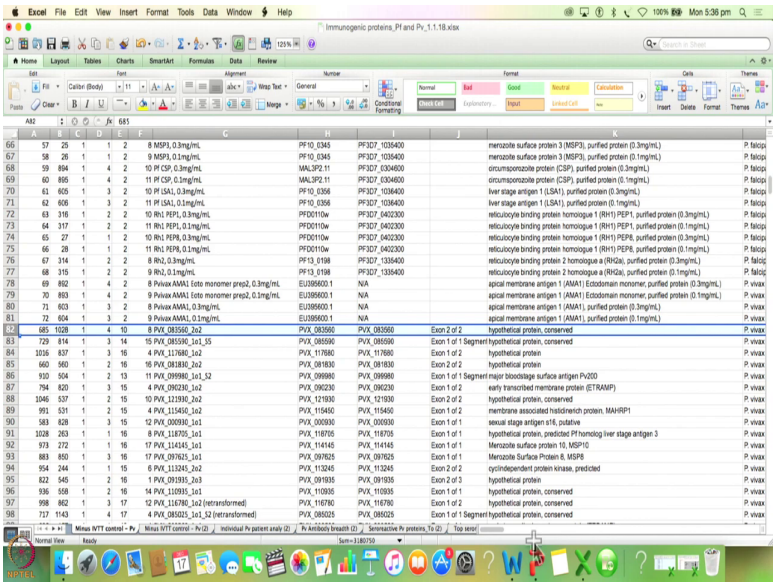
(Refer Slide Time: 26:13)



You will see that there are these four patients which are deliberately kept out of the analysis. For example, there are so if you see here there is P f plus P v and everywhere right. So, basically these are my patients who were diagnosed with mixed infection. So, I do not want any such patients in my analysis.

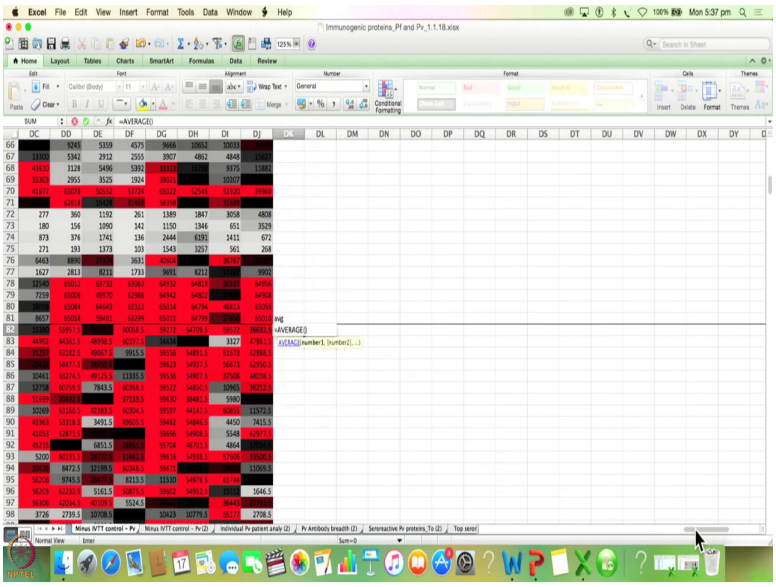
So, I am going to purely have groups which are plasmodium falciparum and plasmodium vivax, and I am going to look at look for their response to plasmodium falciparum antigens and plasmodium vivax respectively. So, I am not going to have any of these mixed patients, I have going to have kept them out. So, if you want, we can also delete them right.

(Refer Slide Time: 26:51)



So, we need to basically start from row number 82 that is what we are interested in, because these are the IVTT spots. So, the first thing I am going to do is I am going to take an average for each and every spot. So, for example, let us write here average, and I am going to say is equal to.

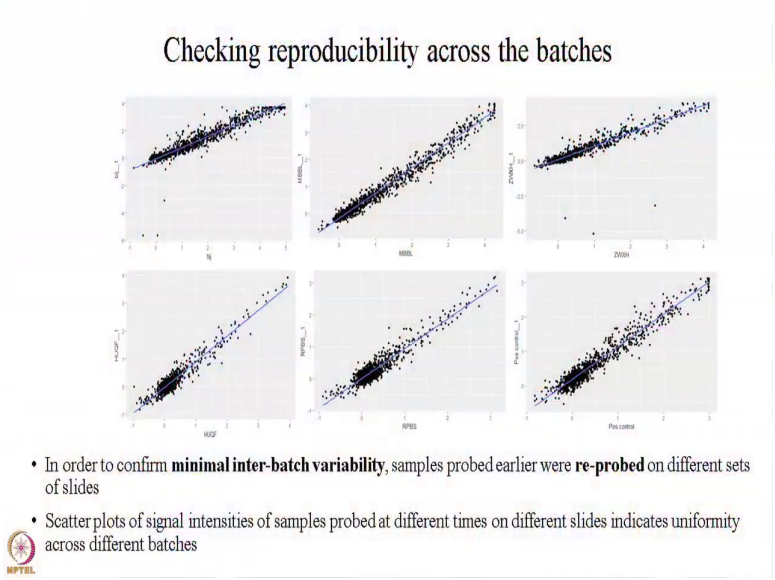
(Refer Slide Time: 27:07)



So, I get an average value, and I am going to just drag this down. So, you will have an average value for each and every spot for all the patients. What I am is telling you before is that in the previous sheet had many more columns here that is because they we had a lot more samples which have probed on the chips.

For example, we had positive controls which are nothing but samples from taken from a 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, do not get confused between positive control samples and positive control spot, they are totally different.

(Refer Slide Time: 28:13)

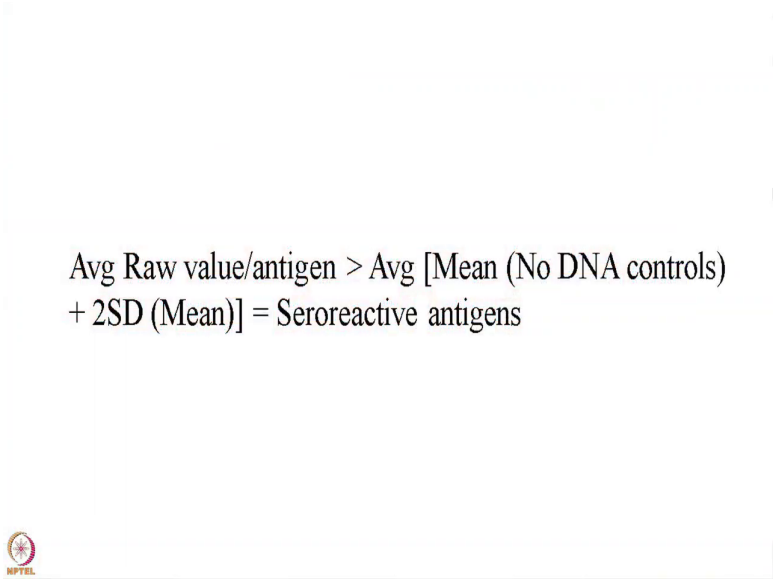


So, these positive control samples I have excluded them from this from this particular analysis. We also had healthy controls which are basically malaria naive individuals means patients who were detected, were not detected with malaria at the time of admission. So, they were malaria negative, those patients were also taken and program the chip just to see there is a difference in response, such patients have also removed from the analysis.

There are 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 are some scatter plots you can see, where I am showing patient-patient reproducibility. So, basically I am showing a reproducible it between my badge runs. All of these patients also have removed from the analysis.

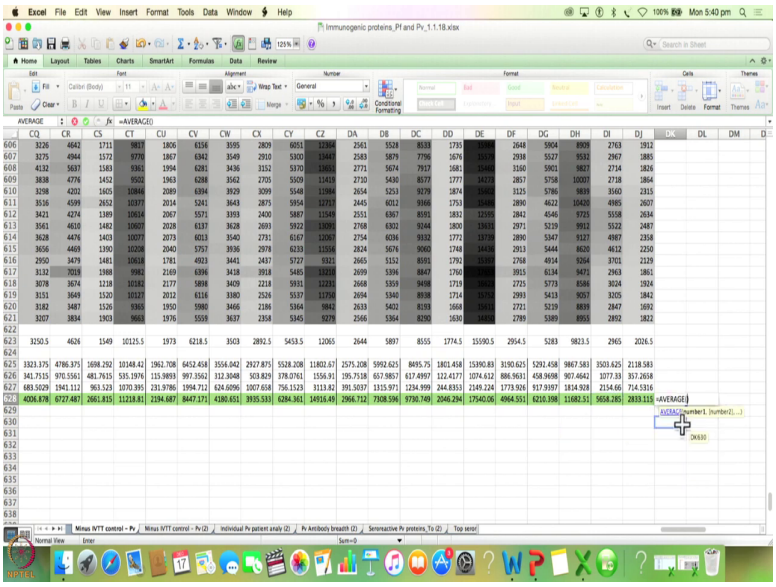
I have basically now in this excel 200 patients; 100 plasmodium vivax, and 96 plasmodium falciparum patients and 4 which a mixed infection also have removed. So, in this way you can choose to remove rows and columns based on what you want to study and you can make a excel less complicated right. So, that is what I had missed mentioning, but now that is done.

(Refer Slide Time: 29:33)



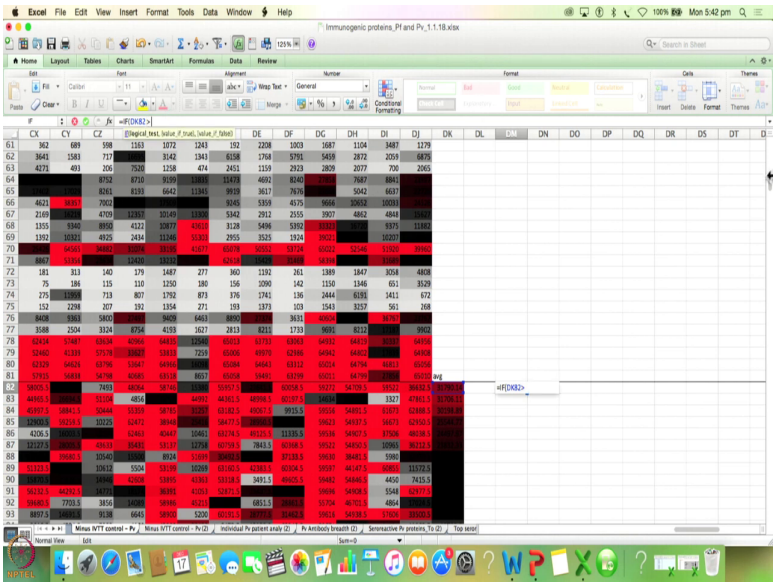
So, I have taken an average right now. Now, I am going to apply this particular formula which you see here. If the average value for a particular spot is more than twice the standard deviation of the mean of the no DNA controls, then that particular spot, then that particular antigen is zero reactive. So, what this means is that if this is the average, if this particular number is greater than this particular number which I am going to show you right now.

(Refer Slide Time: 30:01)



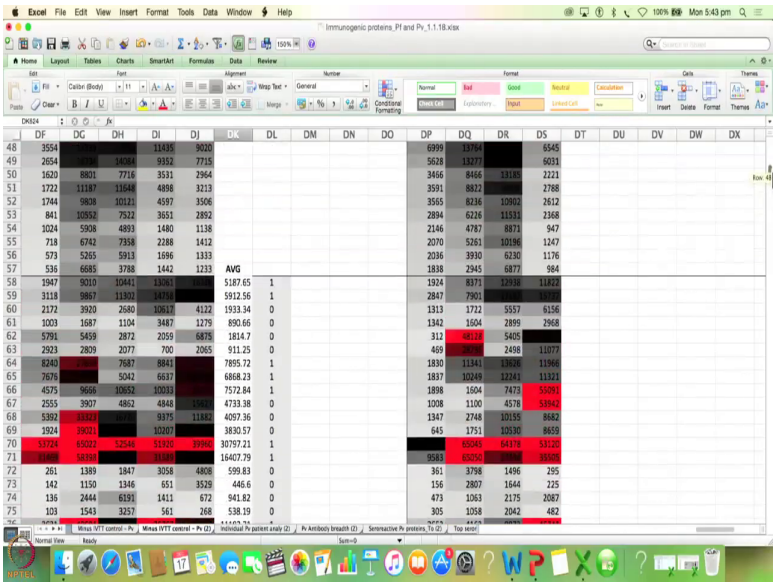
If you take an average of the mean plus 2 times standard deviation of the no DNA spots, this is my number. So, if that raw value or if any raw value is greater than this value, then that spot is basically zero reactive, then that antigen is sorry then that antigen is basically zero reactive.

(Refer Slide Time: 30:41)



So, I am going to say if this is equal to if function, this spot is greater than this, then 1 else 0, so I get 1 here.

(Refer Slide Time: 31:11)



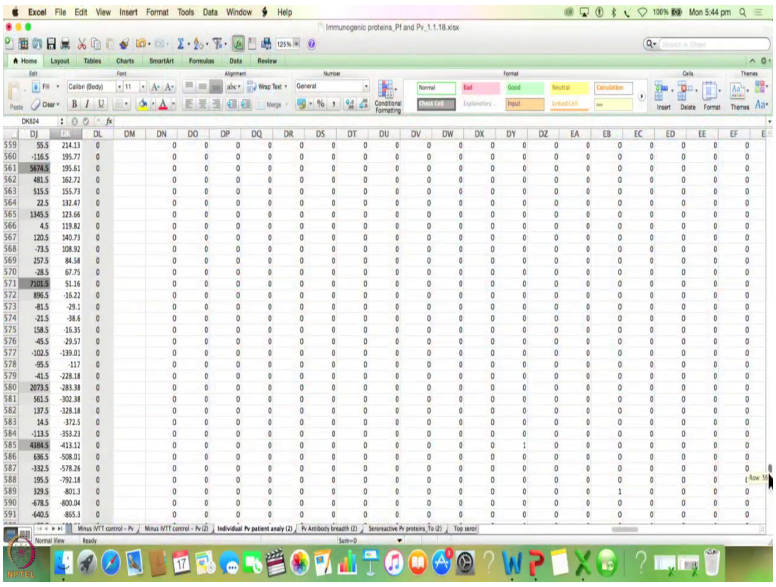
Then what I finally, get is an excel sheet like this where I have random ones and zeros right. So, all of these ones I am going to now say are my zero reactive proteins because they are greater than twice the standard deviation of my control spots. Now, a lot of people may also use healthy controls in their analysis right, but we do not have them.

So, what they do is they compare the signals in signal intensities in a malaria group versus a healthy population, but since we do not have all that I am going to simply say that this is mine, these are the list of my zero reactive proteins which I can now take forward for further analysis.

So, this is not a great start, this is not a statistical test this is only short listing my proteins, and here I am only short listing my proteins from 1500 to a handful which I can then take forward and study. So, this is what that sheet is. Now, what I have done here is I have taken

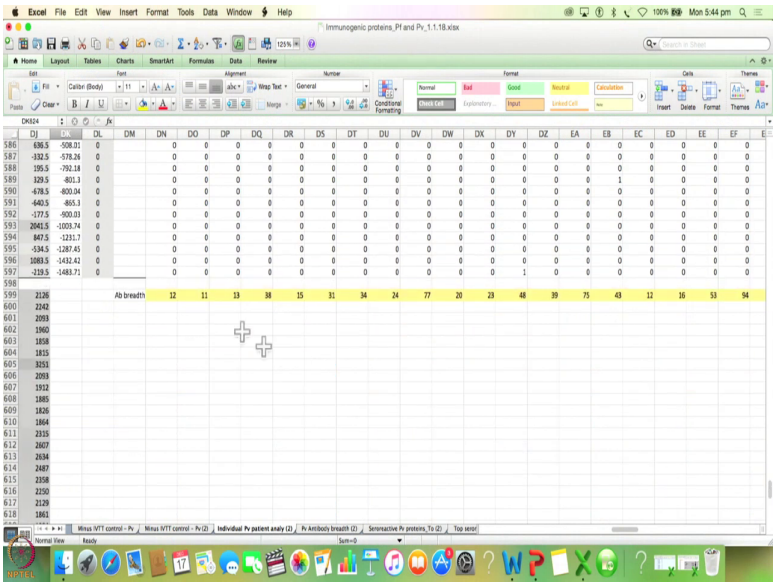
this for a group of patients, but now what if I want to check this for a particular patient, so that is what is my antibody breath which you see here. I have done this individually for every single patient, maybe I will zoom this a little bit.

(Refer Slide Time: 32:21)



So, what you see here is that of zoom this for every single patient. So, basically the previous one which I showed you was it 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 a sample specific right.

(Refer Slide Time: 32:45)



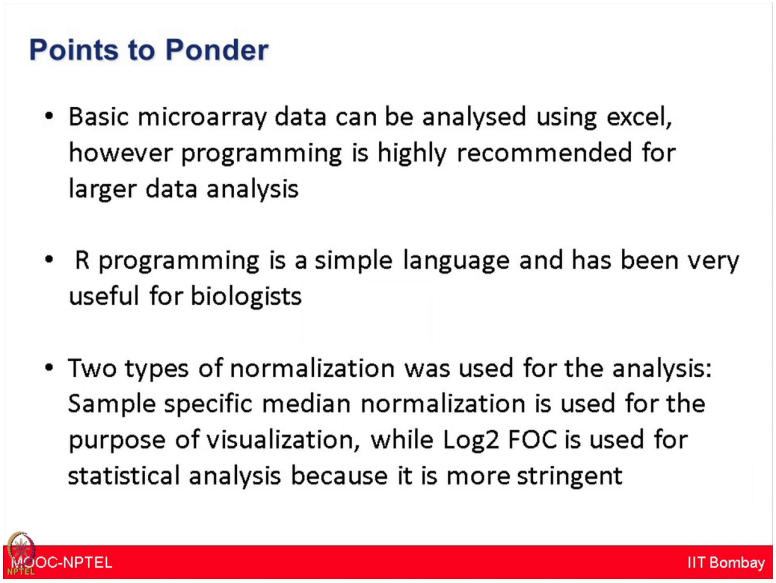
So, in this way if I if I scroll down what you will get here, this you will you will know the number of zero reactive antigens per patient which means that if one patient for example, here a zero reactive to only 12 antigens, whereas, there are some other patients which is zero reactive to 77 antigens. 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 of 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 of microarray data to each and every patient clinical information that you have. And then you need to you can perform any kind of statistical analysis and you can generate several results from the same single experiment so that is the beauty of this.

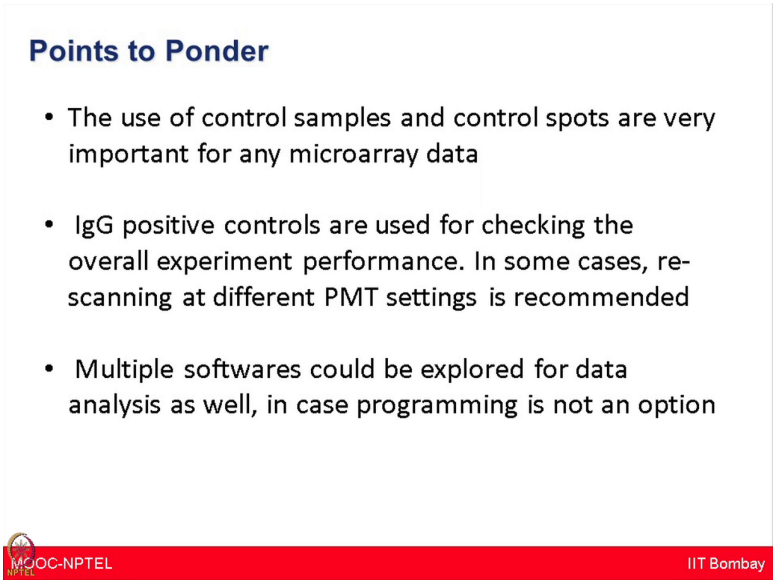
I hope you have got a glimpse of how to perform data analysis and how and how this basic statistics can be done, and how this is not the only way to do statistics at all you can do use softwares and programming, and I would still recommend that people do programming because if you want to make even a single small change, you do not have to repeat the entire analysis.

Also tomorrow if somebody provides you some other clinical information of the same patient population, you do not 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 is all for now. Thank you.

(Refer Slide Time: 34:07)



(Refer Slide Time: 34:23)



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. There are 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 checks and it is properly normalized. In such experiments your control features becomes very crucial. Both the positive controls and negative controls they guide you about how accurate and real a 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 there we will shift the gears to the cancer research and also the platforms. So, far we have talked about cell free expression based protein microarray platform. We will not 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 extra mental demonstrations as well as the theoretical concepts involved in performing such biological experiments.

See you in next lecture.

Thank you.