

**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)

**MOOC-NPTEL**

**Applications of Interactomics  
using Genomics and Proteomics Technologies**

**LECTURE - 13**

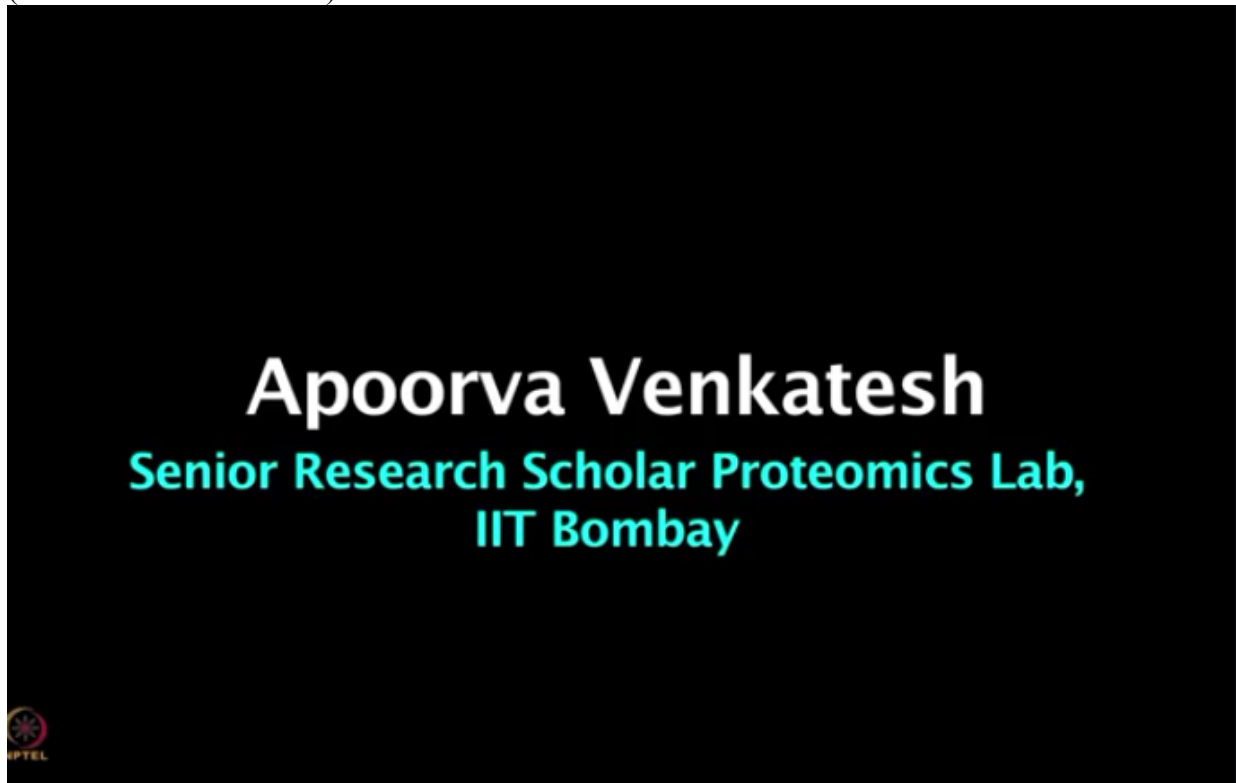
**Applications of protein microarrays in Malaria  
Research-II**

**Dr. Sanjeeva Srivastava**  
Professor  
Biosciences and Bioengineering  
IIT Bombay

MOOC-NPTEL Applications of Interactomics using Genomics and Proteomics Technologies IIT Bombay

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.

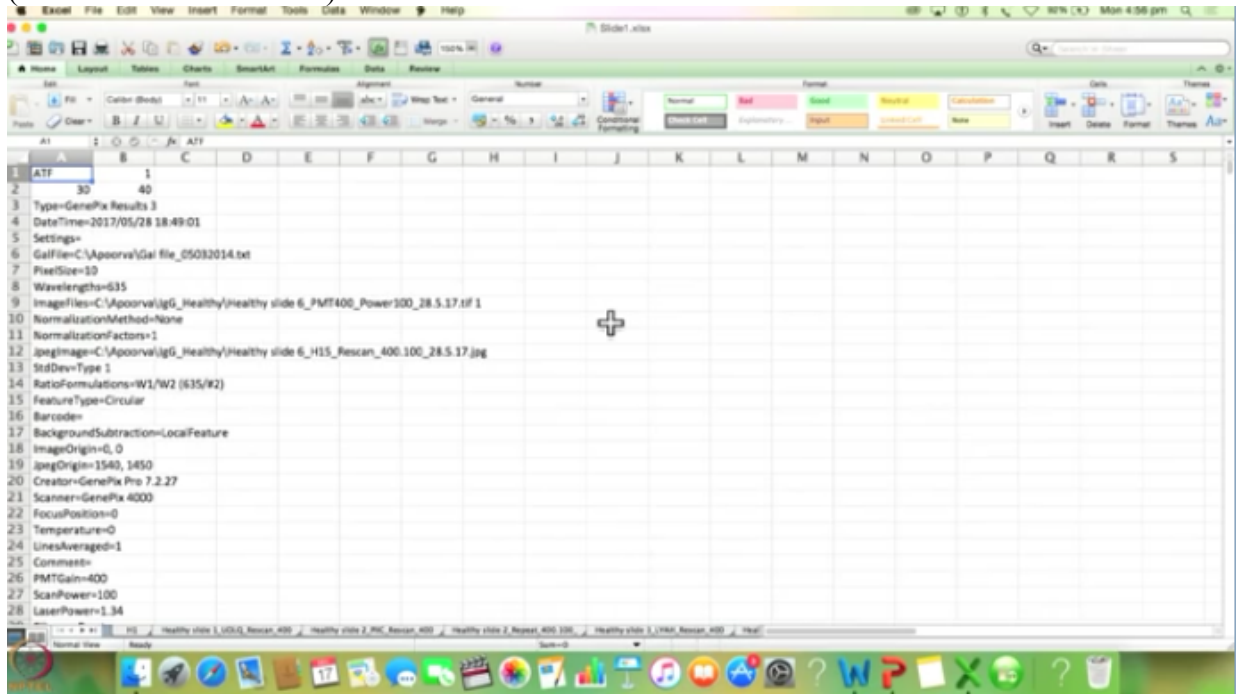
Today we are going to continue the demonstration and also show you the ways to do data normalization and how to do a microarray data analysis, specifically if you go worse to look for a biological question of question of interest.

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)



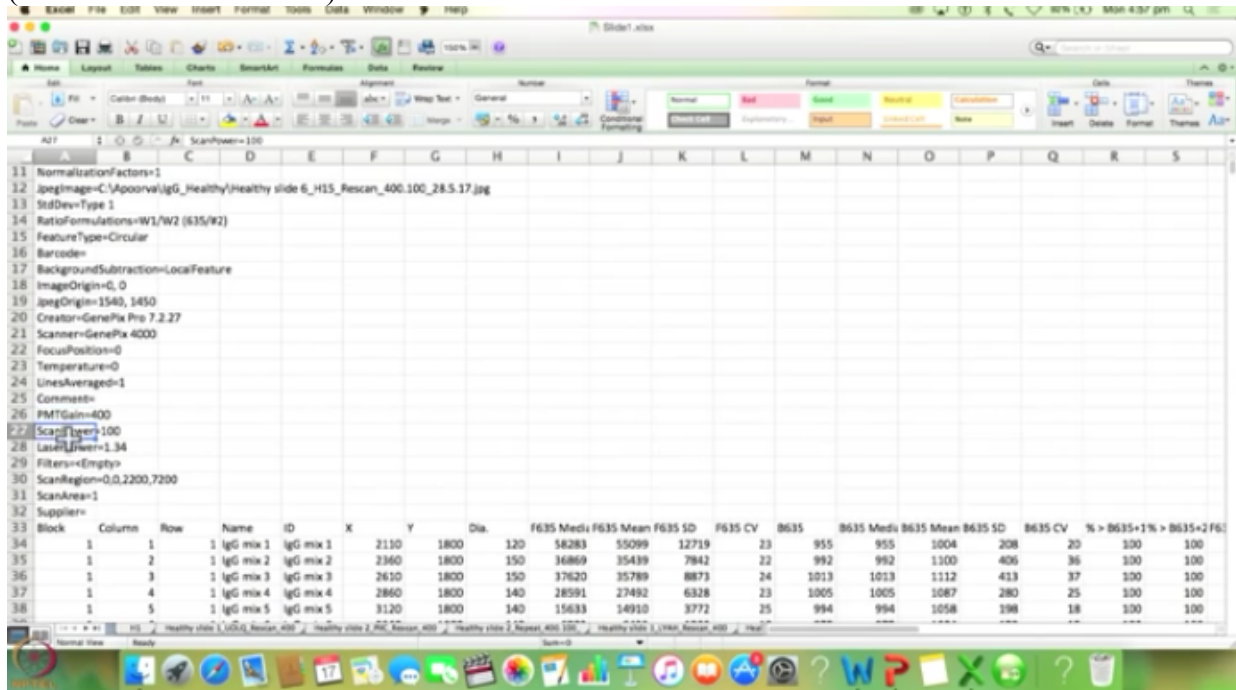
this is the type of data you get, I'm showing you this one particular slide, so I'll just like to repeat that one slide can probe 8 patient serum, so here in this one particular excel which you see here,

(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)

Block	Column	Name	ID	X	Y	Dia.	F635 Mediu	F635 Mean	F635 SD	F635 CV	B635	B635 Mediu	B635 Mean	B635 SD	B635 CV	% > B635+1%	% > B635+2%
1	1	lgG mix 1	lgG mix 1	2110	1800	120	58283	55099	12719	23	955	955	1004	208	20	100	100
1	2	lgG mix 2	lgG mix 2	2360	1800	150	36869	35439	7842	22	992	992	1100	406	36	100	100
1	3	lgG mix 3	lgG mix 3	2610	1800	150	37620	35789	8873	24	1013	1013	1112	413	37	100	100
1	4	lgG mix 4	lgG mix 4	2860	1800	140	28591	27492	6328	23	1005	1005	1087	280	25	100	100
1	5	lgG mix 5	lgG mix 5	3120	1800	140	15633	14910	3772	25	994	994	1058	198	18	100	100
1	6	lgG mix 6	lgG mix 6	3360	1800	140	4303	6439	1208	18	975	975	1004	159	15	100	100
1	7	a-human lg a-human lg		3620	1800	120	65535	61423	10277	16	981	981	1080	314	29	100	100
1	8	a-human lg a-human lg		3870	1800	140	36126	34035	10236	30	998	998	1052	276	25	100	100
1	9	a-human lg a-human lg		4120	1800	140	24230	23319	6712	28	1012	1012	1097	233	21	100	100
1	10	a-human lg a-human lg		4360	1800	140	10185	9843	2642	26	991	991	1035	149	14	100	100
1	11	a-human lg a-human lg		4610	1800	130	4139	4215	798	18	975	975	1001	122	12	100	100
1	12	a-human lg a-human lg		4860	1800	130	2337	2423	329	13	955	955	970	89	9	100	100
1	13	TT85	TT85	5110	1800	170	1174	1177	111	9	937	937	944	64	6	94	82
1	14	TT85	TT85	5360	1800	170	1161	1162	137	11	929	929	935	56	5	89	78
1	15	TT85	TT85	5610	1800	170	1137	1130	104	9	941	941	949	60	6	88	78
1	16	TT85	TT85	5860	1800	170	1359	1339	265	19	972	972	995	117	11	76	66
1	17	no DNA Cor/NoDNA Cor		6110	1800	180	4652	4698	980	20	1017	1017	1062	367	15	100	100
2	1	no DNA Cor/NoDNA Cor		2110	2050	170	4027	4061	620	15	955	955	1015	192	18	100	100
2	2	no DNA Cor/NoDNA Cor		2360	2040	180	4231	4290	926	21	1031	1031	1100	242	22	100	100
2	3	no DNA Cor/NoDNA Cor		2610	2040	180	4387	4340	928	21	1029	1029	1079	210	19	100	100
2	4	no DNA Cor/NoDNA Cor		2860	2050	180	4096	4093	920	22	1026	1026	1051	199	18	100	100
2	5	no DNA Cor/NoDNA Cor		3120	2050	180	4277	4281	878	20	1021	1021	1097	218	19	100	100
2	6	EBA175, 0: EBA175, 0:		3360	2040	170	1516	1551	319	20	981	981	1013	131	12	96	85
2	7	EBA175, 0: EBA175, 0:		3610	2040	170	1272	1277	151	11	977	977	1021	178	17	80	34
2	8	MSP3, 0.3: MSP3, 0.3:		3860	2040	130	20482	20403	2788	13	998	998	1089	364	33	100	100

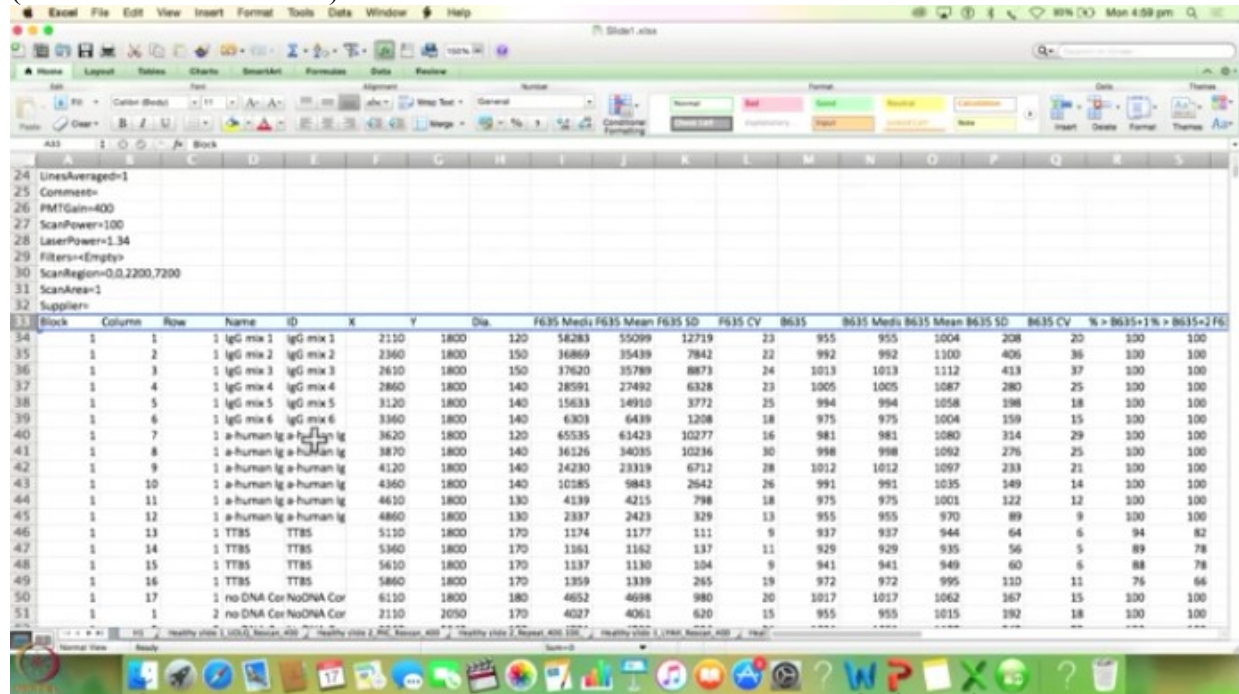
so this is very important, again let's go back to the slide layout, one slide can probe 8 patient sera, and one particular pad that is one pad which probes one patient sera has 4 blocks, so which means that if I keep scrolling down so every 4 blocks represents one patient data, right, so when I keep 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)

Block	Column	Name	ID	X	Y	Dia.	F635 Mediu	F635 Mean	F635 SD	F635 CV	B635	B635 Mediu	B635 Mean	B635 SD	B635 CV	% > B635+1%	% > B635+2%
4	3	17 hypotheitic PVK_05773		16160	5830	170	5095	5083	924	18	1087	1087	1112	144	12	100	100
4	4	17 hypotheitic PVK_08502		16410	5830	160	6124	6032	845	14	1098	1098	1129	153	13	100	100
4	5	17 Empty	Empty	16650	5820	170	4961	4861	1125	23	1089	1089	1124	152	13	100	100
4	6	17 heat shock PFD1235w-e		16920	5830	170	4622	4646	978	21	1095	1095	1132	167	14	100	100
4	7	17 erythrocyte PFD1235w		17170	5830	160	5698	5646	1125	19	1109	1109	1145	171	14	100	100
4	8	17 blank	blank	17410	5840	170	4863	4939	896	18	1122	1122	1179	243	20	100	100
4	9	17 amine acid PVK_11457		17670	5830	170	4985	4991	976	19	1089	1089	1124	152	13	100	100
4	10	17 Falstatin, pi PVK_09903		17910	5830	160	5273	5263	1045	19	1095	1095	1132	167	14	100	100
4	11	17 ubiquitin C PVK_09148		18160	5840	160	5365	5427	1203	22	1109	1109	1145	171	14	100	100
4	12	17 hypotheitic PVK_00453		18410	5830	170	5430	5471	987	18	1122	1122	1179	243	20	100	100
4	13	17 Empty	Empty	18650	5820	170	4803	4821	1124	23	1133	1133	1181	221	18	100	100
4	14	17 erythrocyte MAL6PL1_1		18910	5830	170	5219	5183	906	17	1081	1081	1123	179	15	100	100
4	15	17 erythrocyte PFD1_0521		19160	5830	160	3966	4113	901	21	1044	1044	1084	186	17	100	100
4	16	17 Empty	Empty	19400	5820	170	4903	4884	1218	25	997	997	1040	166	15	0	0
4	17	17 Blank	Blank	19650	5830	170	982	985	51	5	972	972	1025	230	22	100	100
5	1	1 IgG mix 1	IgG mix 1	2090	10800	120	65326	56473	15542	27	1014	1014	1096	334	30	100	100
5	2	1 IgG mix 2	IgG mix 2	2340	10810	140	39073	37378	7757	20	997	997	1105	398	36	100	100
5	3	1 IgG mix 3	IgG mix 3	2590	10810	140	33724	31295	8292	26	980	980	1066	328	30	100	100
5	4	1 IgG mix 4	IgG mix 4	2840	10800	140	23712	22371	5342	23	963	963	1023	196	19	100	100
5	5	1 IgG mix 5	IgG mix 5	3090	10810	140	11951	11455	2602	22	947	947	982	218	22	100	100
5	6	1 IgG mix 6	IgG mix 6	3340	10810	150	4586	4652	1013	21	958	958	1064	314	29	100	100
5	7	1 a-human lg a-human lg		3590	10800	120	65535	61227	11580	18	960	960	1027	398	19	100	100
5	8	1 a-human lg a-human lg		3840	10800	140	28818	26853	7737	28	966	966	1024	384	17	100	100
5	9	1 a-human lg a-human lg		4090	10800	140	18224	17484	4939	28	935	935	952	90	9	100	100
5	10	1 a-human lg a-human lg		4340	10800	140	6985	6849	1758	25	925	925	938	75	7	100	100
5	11	1 a-human lg a-human lg		4590	10800	140	3102	3186	635	19	927	927	938	80	8	87	63
5	12	1 a-human lg a-human lg		4840	10810	130	1841	1833	240	13	927	927	938	80	8	87	63
5	13	1 TT85	TT85	5080	10800	170	1119	1125	102	9	927	927	938	80	8	87	63

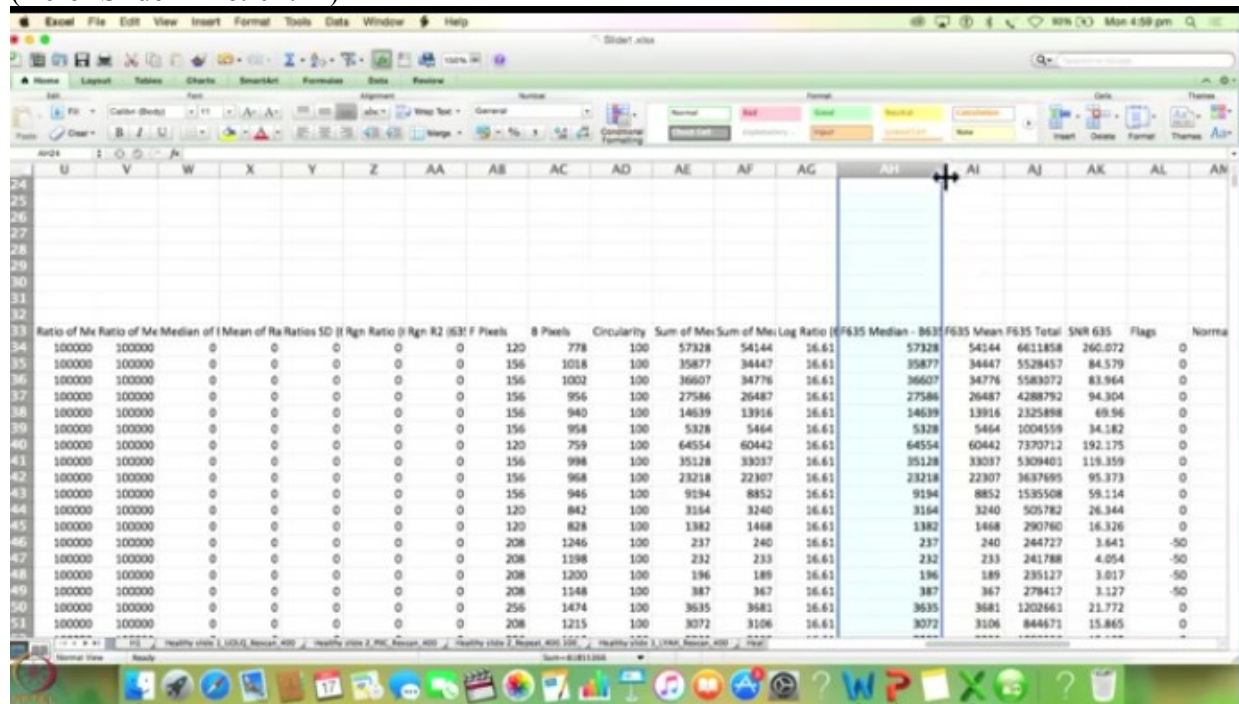
blank and you will see that is IgG1, IgG mix 1 which starts again, so this is basically your new patient, so what you are going to do is we are going to first reorganize this, but before this let

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)



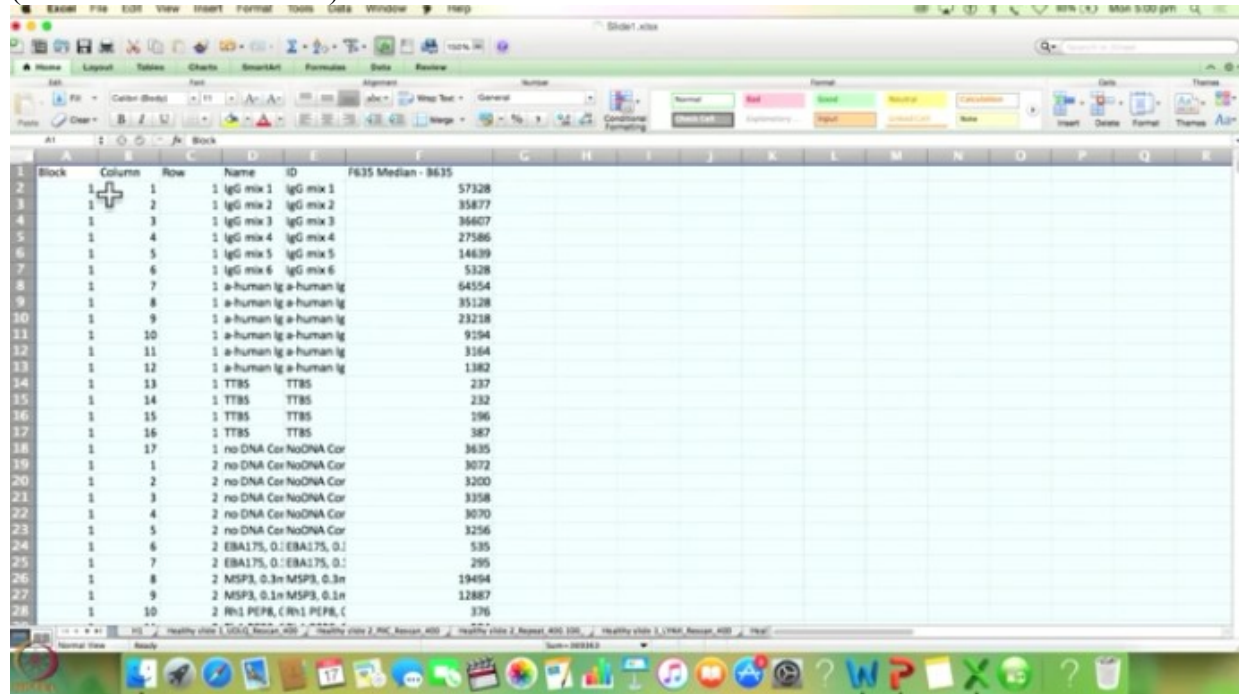
basically this is your protein ID, we don't need any of the other columns except for column AH,  
so you see column AH here,  
(Refer Slide Time: 04:44)



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)

Array Index	Array Row	Array Column	Spot Row	Spot Column	ADI spot ID (gel file)	ORF	PsymoOB_ID	ORF_Fragment	Description
1	1	1	1	1	1 IgG mix 1	N/A	N/A		IgG mix 1
2	2	1	1	1	2 IgG mix 2	N/A	N/A		IgG mix 2
3	3	1	1	1	3 IgG mix 3	N/A	N/A		IgG mix 3
4	4	1	1	1	4 IgG mix 4	N/A	N/A		IgG mix 4
5	5	1	1	1	5 IgG mix 5	N/A	N/A		IgG mix 5
6	6	1	1	1	6 IgG mix 6	N/A	N/A		IgG mix 6
7	7	1	1	1	7 a-human IgG 1	N/A	N/A		anti-Human IgG 1
8	8	1	1	1	8 a-human IgG 2	N/A	N/A		anti-Human IgG 2
9	9	1	1	1	9 a-human IgG 3	N/A	N/A		anti-Human IgG 3
10	10	1	1	1	10 a-human IgG 4	N/A	N/A		anti-Human IgG 4
11	11	1	1	1	11 a-human IgG 5	N/A	N/A		anti-Human IgG 5
12	12	1	1	1	12 a-human IgG 6	N/A	N/A		anti-Human IgG 6
13	13	1	1	1	13 TTBS	N/A	N/A		TTBS
14	14	1	1	1	14 TTBS	N/A	N/A		TTBS
15	15	1	1	1	15 TTBS	N/A	N/A		TTBS
16	16	1	1	1	16 TTBS	N/A	N/A		TTBS
17	17	1	1	1	17 NoDNA	N/A	N/A		noDNA Control
18	18	1	2	1	1 NoDNA	N/A	N/A		noDNA Control
19	19	1	2	2	2 NoDNA	N/A	N/A		noDNA Control
20	20	1	2	3	3 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	Array Index	Array Row	Spot Row	Spot Column	ADI spot ID (gal file)	ORF	Plasmid ID	ORF Fragment	Description
1	1	1	1	1	1	IgG mix 1	N/A		IgG mix 1
2	290	1	2	1	1	IgG mix 1	N/A		IgG mix 1
3	579	1	3	1	1	IgG mix 1	N/A		IgG mix 1
4	868	1	4	1	1	IgG mix 1	N/A		IgG mix 1
5	2	1	1	2	2	IgG mix 2	N/A		IgG mix 2
6	291	1	2	2	2	IgG mix 2	N/A		IgG mix 2
7	580	1	3	2	2	IgG mix 2	N/A		IgG mix 2
8	869	1	4	2	2	IgG mix 2	N/A		IgG mix 2
9	3	1	1	3	3	IgG mix 3	N/A		IgG mix 3
10	292	1	2	3	3	IgG mix 3	N/A		IgG mix 3
11	581	1	3	3	3	IgG mix 3	N/A		IgG mix 3
12	870	1	4	3	3	IgG mix 3	N/A		IgG mix 3
13	4	1	1	4	4	IgG mix 4	N/A		IgG mix 4
14	293	1	2	4	4	IgG mix 4	N/A		IgG mix 4
15	582	1	3	4	4	IgG mix 4	N/A		IgG mix 4
16	871	1	4	4	4	IgG mix 4	N/A		IgG mix 4
17	5	1	1	5	5	IgG mix 5	N/A		IgG mix 5
18	294	1	2	5	5	IgG mix 5	N/A		IgG mix 5
19	583	1	3	5	5	IgG mix 5	N/A		IgG mix 5
20	872	1	4	5	5	IgG mix 5	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 plasmid DB ID, now if you look at plasmid DB ID these are all basically each and (Refer Slide Time: 07:35)

Array Index	Array Row	Spot Row	Spot Column	ADI spot ID (gal file)	ORF	Plasmid ID	ORF Fragment	Description
52	24	1	1	7	EBA175, 0.1mg/ml	MAL7P1.178	PF3D7_0731500	erythrocyte binding antigen 175 (EBA175), purified p
53	601	1	3	2	8 MSP1, 0.3mg/ml	PF11475w	PF3D7_0930300	merozoite surface protein 1 (MSP1), purified protein
54	802	1	3	2	7 MSP1, 0.1mg/ml	PF11475w	PF3D7_0930300	merozoite surface protein 1 (MSP1), purified protein
55	890	1	4	2	8 MSP2, 0.3mg/ml	PF80300c	PF3D7_0206800	merozoite surface protein 2 (MSP2), purified protein
56	891	1	4	2	7 MSP2, 0.1mg/ml	PF80300c	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	PF00110w	PF3D7_0402300	reticulocyte binding protein homologue 1 (RH1) PEF
64	317	1	2	2	11 Rh1 PEP1, 0.1mg/ml	PF00110w	PF3D7_0402300	reticulocyte binding protein homologue 1 (RH1) PEF
65	27	1	1	2	10 Rh1 PEP8, 0.3mg/ml	PF00110w	PF3D7_0402300	reticulocyte binding protein homologue 1 (RH1) PEF
66	28	1	1	2	11 Rh1 PEP8, 0.1mg/ml	PF00110w	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 Pulvax AMA1 Ecto monomer prep2, 0.3mg/ml	EU395600.1	N/A	apical membrane antigen 1 (AMA1) Ectodomain mo
70	893	1	4	2	9 Pulvax AMA1 Ecto monomer prep2, 0.1mg/ml	EU395600.1	N/A	apical membrane antigen 1 (AMA1) Ectodomain mo
71	803	1	3	2	8 Pulvax AMA1, 0.3mg/ml	EU395600.1	N/A	apical membrane antigen 1 (AMA1), purified protein
72	804	1	3	2	9 Pulvax AMA1, 0.1mg/ml	EU395600.1	N/A	apical membrane antigen 1 (AMA1), purified protein
73	279	1	1	17	7 PFL_0008_CDR2	PFL_0008	N/A	erythrocyte membrane protein 1, PREMP1 (VAR)
74	37	1	1	3	3 PFA010w-s2	PFA0110w	PF3D7_0102200	Exon 2 Segment 2
75	50	1	1	3	10 PFA0125c-s2	PFA0125c	PF3D7_0102500	Exon 1 Segment 2
76	991	1	4	8	5 PFA0175w_3o7	PFA0175w	PF3D7_0103500	Exon 2 of 7
77	708	1	3	8	11 PFA0360c_3o2	PFA0360c	PF3D7_0107300	probable protein, unknown function
78	81	1	1	5	13 PFA0410w-s1	PFA0410w	PF3D7_0108300	Segment 1
79	370	1	2	5	13 PFA0410w-s2	PFA0410w	PF3D7_0108300	Segment 2

every protein has a unique plasmid 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)

A	B	C	D	E	F	G	H	I	J	K
75	893	1	4	2	9	Pulvax AMA1 Ecto monomer prep2, 0.1mg/ml	EU395600.1	NA		apical membrane antigen 1 (AMA1) Ectodomain mo
76	71	803	1	3	2	8 Pulvax AMA1, 0.3mg/ml	EU395600.1	NA		apical membrane antigen 1 (AMA1), purified protein
77	72	804	1	3	2	9 Pulvax AMA1, 0.1mg/ml	EU395600.1	NA		apical membrane antigen 1 (AMA1), purified protein
78	73	279	1	1	17	7 PFL_0008_CDR2	PFL_0008	NA	CDR2	erythrocyte membrane protein 1, PEMP1 (VAR)
79	74	37	1	1	3	3 PFA0110we2s2	PFAD110w	PF307_0102200	Exon 2 Segment 2	ringinfected erythrocyte surface antigen (RESA)
80	75	50	1	1	3	16 PFA0121ce1s2	PFAD129c	PF307_0102500	Exon 1 Segment 2	erythrocyte binding antigen181 (EBA181)
81	76	991	1	4	8	5 PFA0175w_2o7	PFAD175w	PF307_0103600	Exon 2 of 7	conserved Plasmodium protein, unknown function
82	77	708	1	3	8	11 PFA0360c_2o2	PFA0360c	PF307_0107300	Exon 2 of 2	probable protein, unknown function
83	78	81	1	1	5	13 PFA0410w-e1	PFAD410w	PF307_0108300	Segment 1	conserved Plasmodium protein, unknown function
84	79	370	1	2	5	13 PFA0410w-e2	PFAD410w	PF307_0108300	Segment 2	conserved Plasmodium protein, unknown function
85	80	947	1	4	5	12 PFA0410w-e3	PFAD410w	PF307_0108300	Segment 3	conserved Plasmodium protein, unknown function
86	81	45	1	1	3	11 PFA0430ce1s1	PFAD430c	PF307_0108700	Exon 1 Segment 1	secreted ookinete protein, putative (PSOP24)
87	82	904	1	4	3	3 PFA0430ce1s2	PFAD430c	PF307_0108700	Exon 1 Segment 2	secreted ookinete protein, putative (PSOP24)
88	83	989	1	3	8	6 PFA0490w_1o1	PFAD490w	PF307_0110000	Exon 1 of 1	conserved Plasmodium protein, unknown function
89	84	917	1	4	3	16 PFA0510we1s2	PFA0510w	PF307_0110500	Exon 1 Segment 2	bromodomain protein, putative
90	85	628	1	3	3	16 PFA0510we1s3	PFA0510w	PF307_0110500	Exon 1 Segment 3	bromodomain protein, putative
91	86	639	1	3	4	10 PFB0010w (renamed)	PFB0010w	PF307_0200100		erythrocyte membrane protein 1, PEMP1 (VAR)
92	87	713	1	3	8	16 PFB0100ce2s1	PFB0100c	PF307_0202000	Exon 2 Segment 1	knob-associated histidine-rich protein (KAHRP)
93	88	418	1	2	8	10 PFB0104c_2o2	PFB0106c	PF307_0202200	Exon 2 of 2	Plasmodium exported protein, unknown function
94	89	617	1	3	3	5 PFB0115we1s2	PFB0115w	PF307_0202400	Exon 1 Segment 2	conserved Plasmodium protein, unknown function
95	90	667	1	3	6	4 PFB0120w_1o1	PFB0120w	PF307_0202500	Exon 1 of 1	early transcribed membrane protein 2 (ETFMMP2)
96	91	53	1	1	4	2 PFB0150ce2s3	PFB0150c	PF307_0203100	Exon 2 Segment 3	protein kinase, putative
97	92	392	1	2	7	1 PFB0170w_1o1	PFB0170w	PF307_0203600	Exon 1 of 1	conserved Plasmodium protein, unknown function
98	93	881	1	3	7	1 PFB0250w_1o1	PFB0250w	PF307_0205600	Exon 1 of 1	conserved Plasmodium protein, unknown function
99	94	85	1	1	4	14 PFB0300c	PFB0300c	PF307_0206800		merozoite surface protein 2 (MSP2)
100	95	963	1	3	5	17 PFB0305c_1o2	PFB0305c	PF307_0206900.1	Exon 1 of 2	merozoite surface protein 5 (MSP5)
101	96	90	1	1	6	5 PFB0305c-e1	PFB0305c	PF307_0206900.1	Exon 1	merozoite surface protein 5 (MSP5)
102	97	374	1	2	5	17 PFB0310c_1o2	PFB0310c	PF307_0207000	Exon 1 of 2	merozoite surface protein 4 (MSP4)

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 which is the unique ID for each and every protein.

What I mean here is that if you go to plasmid 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)

Row	Column A	Column B	Column C	Column D	Column E	Column F	Column G	Column H	Column I	Column J	Column K
91	86	539	1	3	4	10	PF80010w (renamed)	PF80010w	PF3D7_0206100		erythrocyte membrane protein 1, PEMP1 (VAR)
92	87	713	1	3	8	16	PF80100c2s1	PF80100c	PF3D7_0202000	Exon 2 Segment 1	knob-associated histidrich protein (KAHRP)
93	88	418	1	2	8	10	PF80104c_2a2	PF80106c	PF3D7_0202200	Exon 2 of 2	Plasmodium exported protein, unknown function
94	89	817	1	3	3	5	PF80115w1a2	PF80115w	PF3D7_0202400	Exon 1 Segment 2	conserved Plasmodium protein, unknown function
95	90	987	1	3	8	4	PF80120w_1a1	PF80120w	PF3D7_0202500	Exon 1 of 1	early transcribed membrane protein 2 (ETRAPP2)
96	91	53	1	1	4	2	PF80150c2s3	PF80150c	PF3D7_0203100	Exon 2 Segment 3	protein kinase, putative
97	92	392	1	2	7	1	PF80170w_1a1	PF80170w	PF3D7_0203600	Exon 1 of 1	conserved Plasmodium protein, unknown function
98	93	881	1	3	7	1	PF80250w_1a1	PF80250w	PF3D7_0205600	Exon 1 of 1	conserved Plasmodium protein, unknown function
99	94	85	1	1	4	14	PF80300c	PF80300c	PF3D7_0206800		merozoite surface protein 2 (MSP2)
100	95	963	1	3	5	17	PF80305c_1a2	PF80305c	PF3D7_0206900.1	Exon 1 of 2	merozoite surface protein 5 (MSP5)
101	96	90	1	1	8	5	PF80305c-e1	PF80305c	PF3D7_0206900.1	Exon 1	merozoite surface protein 5 (MSP5)
102	97	374	1	2	5	17	PF80310c_1a2	PF80310c	PF3D7_0207000	Exon 1 of 2	merozoite surface protein 4 (MSP4)
103	98	987	1	4	8	1	PF80310c_2a2	PF80310c	PF3D7_0207000	Exon 2 of 2	merozoite surface protein 4 (MSP4)
104	99	89	1	1	5	1	PF80310c-e1	PF80310c	PF3D7_0207000	Exon 1	merozoite surface protein 4 (MSP4)
105	100	115	1	1	7	13	PF80330c_2a4	PF80330c	PF3D7_0207400	Exon 2 of 4	serine repeat antigen 7 (SERA7)
106	101	1003	1	4	8	17	PF80335c8s1	PF80335c	PF3D7_0207500	Exon 3 Segment 1	serine repeat antigen 6 (SERA6)
107	102	901	1	4	2	17	PF80340c2s1	PF80340c	PF3D7_0207600	Exon 2 Segment 1	serine repeat antigen 5 (SERA5)
108	103	982	1	4	7	13	PF80345c_2a4	PF80345c	PF3D7_0207700	Exon 2 of 4	serine repeat antigen 4 (SERA4)
109	104	428	1	2	9	3	PF80345c_4a4	PF80345c	PF3D7_0207700	Exon 4 of 4	serine repeat antigen 4 (SERA4)
110	105	989	1	4	8	3	PF80350c_2a4	PF80350c	PF3D7_0207800	Exon 2 of 4	serine repeat antigen 3 (SERA3)
111	106	116	1	1	7	14	PF80765w_6a7	PF80765w	PF3D7_0216700.1	Exon 6 of 7	conserved Plasmodium protein, unknown function
112	107	996	1	4	8	10	PF80900c_2a2	PF80900c	PF3D7_0219700	Exon 2 of 2	Plasmodium exported protein (PfESTc), unknown func
113	108	861	1	3	5	15	PF80910w_2a2	PF80910w	PF3D7_0219900	Exon 2 of 2	Plasmodium exported protein, unknown function
114	109	948	1	4	5	13	PF80915w-e2s1	PF80915w	PF3D7_0220000	Exon 2 Segment 1	liver stage antigen 3 (LSA3)
115	110	859	1	3	5	13	PF80915w-e2s2	PF80915w	PF3D7_0220000	Exon 2 Segment 2	liver stage antigen 3 (LSA3)
116	111	898	1	3	8	1	PF80924c_2a2	PF80926c	PF3D7_0220500	Exon 2 of 2	Plasmodium exported protein (hyp2), unknown func
117	112	408	1	2	7	17	PF80930w_2a2	PF80930w	PF3D7_0220600	Exon 2 of 2	Plasmodium exported protein (hyp3), unknown func
118	113	705	1	3	8	8	PF80932w_2a2	PF80932w	PF3D7_0220700	Exon 2 of 2	Plasmodium exported protein (hyp3), unknown func

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)

Row	Column A	Column B	Column C	Column D	Column E	Column F	Column G	Column H	Column I	Column J	Column K
102	97	374	1	2	5	17	PF80310c_1a2	PF80310c	PF3D7_0207000	Exon 1 of 2	merozoite surface protein 4 (MSP4)
103	98	987	1	4	8	1	PF80310c_2a2	PF80310c	PF3D7_0207000	Exon 2 of 2	merozoite surface protein 4 (MSP4)
104	99	89	1	1	5	1	PF80310c-e1	PF80310c	PF3D7_0207000	Exon 1	merozoite surface protein 4 (MSP4)
105	100	115	1	1	7	13	PF80330c_2a4	PF80330c	PF3D7_0207400	Exon 2 of 4	serine repeat antigen 7 (SERA7)
106	101	1003	1	4	8	17	PF80335c8s1	PF80335c	PF3D7_0207500	Exon 3 Segment 1	serine repeat antigen 6 (SERA6)
107	102	901	1	4	2	17	PF80340c2s1	PF80340c	PF3D7_0207600	Exon 2 Segment 1	serine repeat antigen 5 (SERA5)
108	103	982	1	4	7	13	PF80345c_2a4	PF80345c	PF3D7_0207700	Exon 2 of 4	serine repeat antigen 4 (SERA4)
109	104	428	1	2	9	3	PF80345c_4a4	PF80345c	PF3D7_0207700	Exon 4 of 4	serine repeat antigen 4 (SERA4)
110	105	989	1	4	8	3	PF80350c_2a4	PF80350c	PF3D7_0207800	Exon 2 of 4	serine repeat antigen 3 (SERA3)
111	106	116	1	1	7	14	PF80765w_6a7	PF80765w	PF3D7_0216700.1	Exon 6 of 7	conserved Plasmodium protein, unknown function
112	107	996	1	4	8	10	PF80900c_2a2	PF80900c	PF3D7_0219700	Exon 2 of 2	Plasmodium exported protein (PfESTc), unknown func
113	108	861	1	3	5	15	PF80910w_2a2	PF80910w	PF3D7_0219900	Exon 2 of 2	Plasmodium exported protein, unknown function
114	109	948	1	4	5	13	PF80915w-e2s1	PF80915w	PF3D7_0220000	Exon 2 Segment 1	liver stage antigen 3 (LSA3)
115	110	859	1	3	5	13	PF80915w-e2s2	PF80915w	PF3D7_0220000	Exon 2 Segment 2	liver stage antigen 3 (LSA3)
116	111	898	1	3	8	1	PF80924c_2a2	PF80926c	PF3D7_0220500	Exon 2 of 2	Plasmodium exported protein (hyp2), unknown func
117	112	408	1	2	7	17	PF80930w_2a2	PF80930w	PF3D7_0220600	Exon 2 of 2	Plasmodium exported protein (hyp3), unknown func
118	113	705	1	3	8	8	PF80932w_2a2	PF80932w	PF3D7_0220700	Exon 2 of 2	Plasmodium exported protein (hyp3), unknown func

if you say that this was the ID and this is exon 1 of 2 you will actually see the ID here, and 102, 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)

	K	L	M	N	O	P	Q	R	S	T
53	anti-human IgG	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 antigen 175 (EBA175), purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	20121	3331	1366	1280	691	1276	1202
57	erythrocyte binding antigen 175 (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	598	625	163	116	175	188	271
70	reticulocyte binding protein homologue 1 (RH1) PEPB, purified protein (0.3mg/mL)	P. falciparum 3D7	Purified Protein	302	359	284	236	187	417	673
71	reticulocyte binding protein homologue 1 (RH1) PEPB, 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	4025	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) Ectococain monomer, purified protein (0.3mg/mL)	P. vivax Palo Ato	Purified Protein	63088	43280	5127	1829	1468	49717	64027
75	apical membrane antigen 1 (AMA1) Ectococain monomer, purified protein (0.1mg/mL)	P. vivax Palo Ato	Purified Protein	63048	38950	2709	1213	934	41403	63423
76	apical membrane antigen 1 (AMA1), purified protein (0.3mg/mL)	P. vivax Palo Ato	Purified Protein	63317	45303	5026	1817	1819	53167	63987
77	apical membrane antigen 1 (AMA1), purified protein (0.1mg/mL)	P. vivax Palo Ato	Purified Protein	63345	39267	3839	1716	1425	44904	64003
78	erythrocyte membrane protein 1, PEEMP1 (VAR)	P. falciparum	IVTT	6799	5126	2785	5373	4355	9072	20407
79	ring-infected erythrocyte surface antigen (RESA)	P. falciparum 3D7	IVTT	13575	16656	2674	5852	3407	7356	13275
80	erythrocyte binding antigen 181 (EBA181)	P. falciparum 3D7	IVTT	21760	4278	4717	7642	4723	8695	17818

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)

	L	M	N	O	P	Q	R	S	T
71	reticulocyte binding protein homologue 1 (RH1) PEPR, 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	2438	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, PEMPT1 (VAR)	P. falciparum IVTT	6799	5326	2785	5373	4355	9072	20407
79	ring-infected erythrocyte surface antigen (RESA)	P. falciparum 3D7 IVTT	13575	16656	2674	5852	3407	7356	13275
80	erythrocyte binding antigen 181 (EBA181)	P. falciparum 3D7 IVTT	21760	4278	4717	7542	4723	8695	17838
81	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	7816	15681	2437	5393	3984	7865	14138
82	probable protein, unknown function	P. falciparum 3D7 IVTT	1952	1355	1071	2323	1418	2280	3808
83	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	15274	7058	7096	16626	4335	8980	32730
84	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	13082	6999	25270	7807	5386	9661	18729
85	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	16322	4601	7963	12911	3655	6871	16721
86	secreted oocyst protein, putative (PSOP24)	P. falciparum 3D7 IVTT	16505	17856	3245	10547	2972	8597	14801
87	secreted oocyst protein, putative (PSOP24)	P. falciparum 3D7 IVTT	10516	22084	7026	11930	4433	8152	15598
88	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	9640	5385	3996	4905	4324	8602	16868
89	bromodomain protein, putative	P. falciparum 3D7 IVTT	10079	6023	4109	4633	3054	6110	12876
90	bromodomain protein, putative	P. falciparum 3D7 IVTT	6320	3301	2935	5636	3687	8109	13484
91	erythrocyte membrane protein 1, PEMPT1 (VAR)	P. falciparum 3D7 IVTT	5251	5338	2505	7931	4232	10155	14413
92	knob-associated histidine-rich protein (KAHRP)	P. falciparum 3D7 IVTT	6382	9450	1989	3529	2228	6062	9058
93	Plasmodium exported protein, unknown function	P. falciparum 3D7 IVTT	7707	3021	2250	3903	3413	5579	12273
94	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	5020	4303	2374	5704	4453	7304	15011
95	early transcribed membrane protein 2 (ETRAP2)	P. falciparum 3D7 IVTT	23580	19799	1423	5560	3844	7071	12737
96	protein kinase, putative	P. falciparum 3D7 IVTT	8829	3942	2448	5228	4630	9547	18045
97	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	10091	3930	4385	5720	4867	8076	17235
98	conserved Plasmodium protein, unknown function	P. falciparum 3D7 IVTT	19527	3478	7753	6073	4412	7918	20234

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.

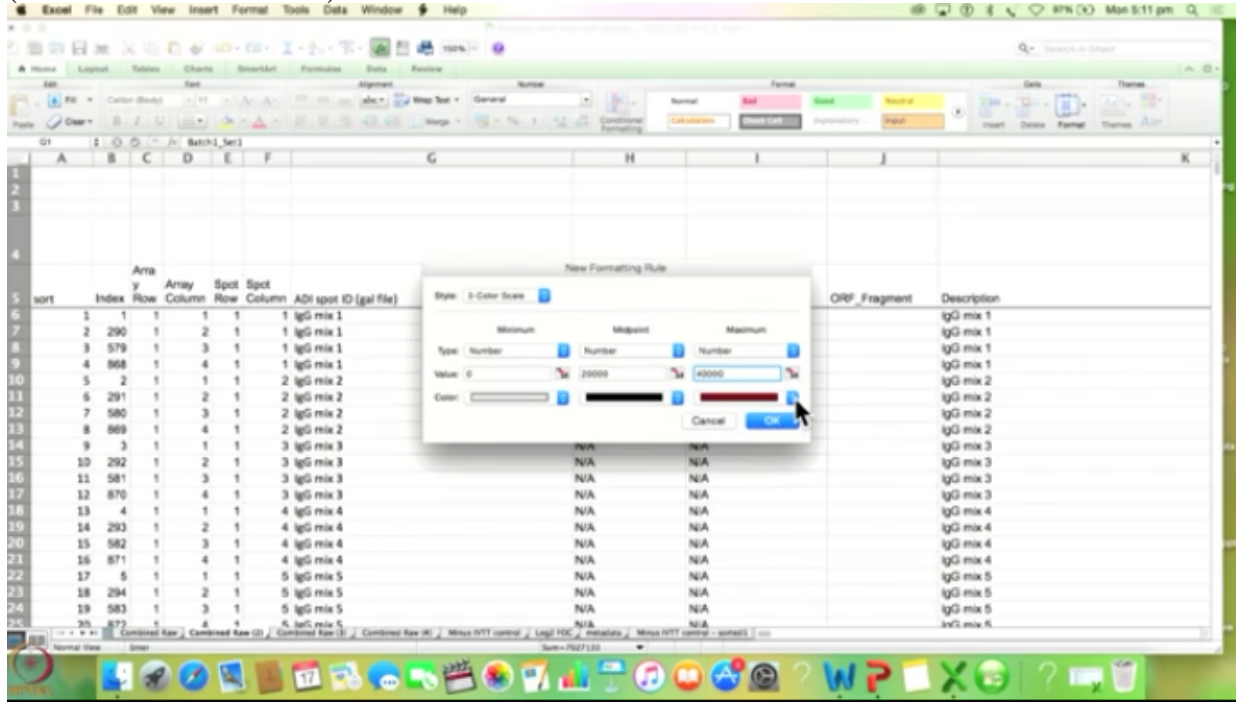
Now the all other columns here are basically your patients samples, so if I just move this a little bit, what you will see here, for example let's consider the first sample, this is basically a positive control which was part of batch 1 set 1, slide 1, and pad 1, so let me again take you back to the 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

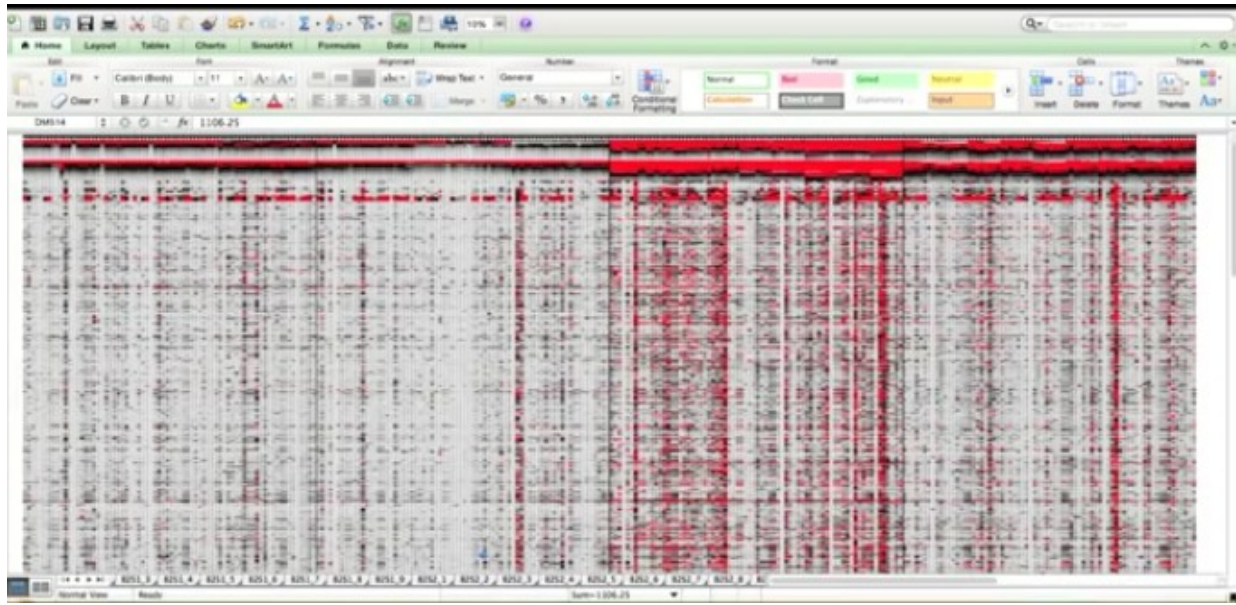
falls you know in between say certain negative values and may be around 80,000 is my maximum value, so I'm just going to assume that if my data falls in this range, I'm going to split my data based on three numbers 0, then my midpoint will be say 20,000, and my highest will be 40,000, and I'm going to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



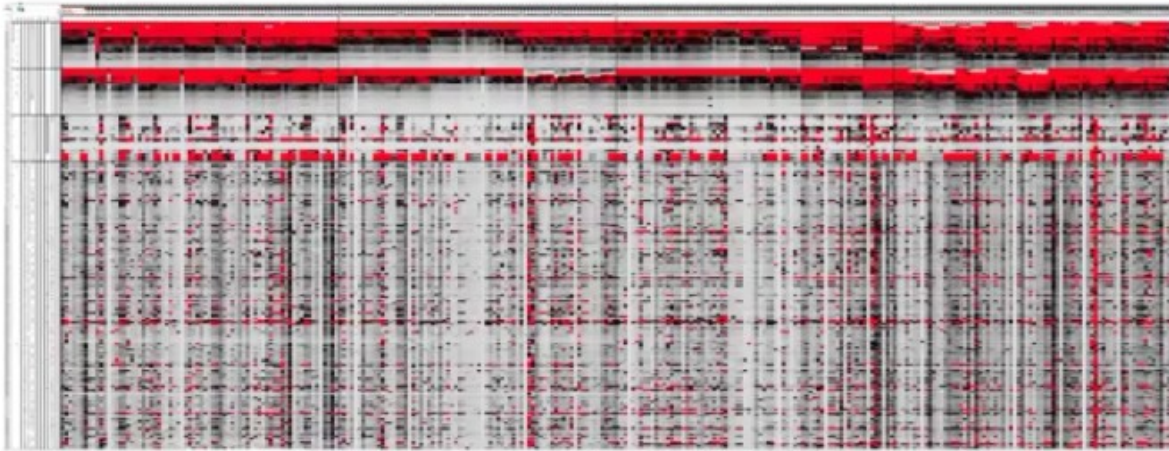
so you will be able to see all four batches at once, so I don't know if you can see a black line here, so basically this is going to split your batches, so in fact it's going to split your set, set 1 from set 2 of the first batch, and set 1 from set 2 of the second batch, so basically this is batch 1 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

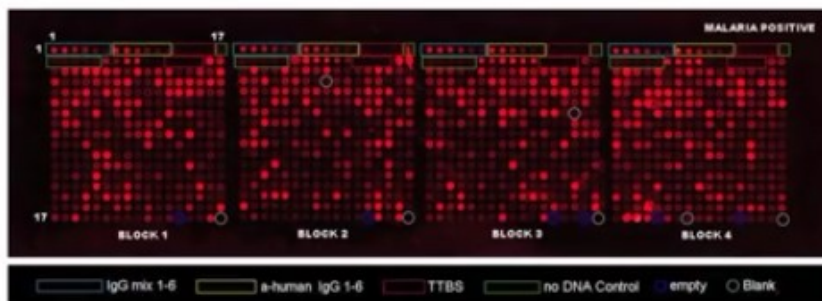
So now having rescanned all the slides as you can see in the slide, the settings look pretty uniform though a still not very uniform and you will still feel that batch 2 set 1 has higher signals, but overall it's okay because this will then be taken care of by normalization.

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. Log<sub>2</sub> transformed FOC-** Each raw value is divided by the median of its 'No DNA controls' and Log<sub>2</sub> 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)

Plasmid ID	ORF Fragment	Description	Organism	Preparation	Batch1	Set Batch1	Set Batch1	Set Batch1
6	NA	CDF2	erythrocyte membrane protein 1, P1EMP1 (VAR)	P. falciparum	IVTT	6799	5126	2785
7	PF3D7_0102200	Exon 2 Segment 2	ringinfected erythrocyte surface antigen (RESA)	P. falciparum 3D7	IVTT	13575	16956	2674
8	PF3D7_0102900	Exon 1 Segment 2	erythrocyte binding antigen181 (EBA181)	P. falciparum 3D7	IVTT	31762	4278	4717
9	PF3D7_0103900	Exon 2 of 7	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	7815	15681	2437
10	PF3D7_0107300	Exon 2 of 2	probable protein, unknown function	P. falciparum 3D7	IVTT	1952	1155	1071
11	PF3D7_0108300	Segment 1	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	15274	7054	7094
12	PF3D7_0108300	Segment 2	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	11082	6999	2527
13	PF3D7_0108300	Segment 3	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	16322	4601	7963
14	PF3D7_0108700	Exon 1 Segment 1	secreted ookinete protein, putative (PSOP24)	P. falciparum 3D7	IVTT	16326	17950	3245
15	PF3D7_0108700	Exon 1 Segment 2	secreted ookinete protein, putative (PSOP24)	P. falciparum 3D7	IVTT	10516	22064	7026
16	PF3D7_0110000	Exon 1 of 1	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	9640	5185	3996
17	PF3D7_0110500	Exon 1 Segment 2	bromodomain protein, putative	P. falciparum 3D7	IVTT	10079	6023	4109
18	PF3D7_0110500	Exon 1 Segment 3	bromodomain protein, putative	P. falciparum 3D7	IVTT	6320	3301	2935
19	PF3D7_0200100	Exon 1 Segment 1	erythrocyte membrane protein 1, P1EMP1 (VAR)	P. falciparum 3D7	IVTT	5251	5338	2905
20	PF3D7_0202000	Exon 2 Segment 1	knot-associated histidinerich protein (KAHRP)	P. falciparum 3D7	IVTT	6382	9450	1983
21	PF3D7_0202200	Exon 2 of 2	Plasmodium exported protein, unknown function	P. falciparum 3D7	IVTT	7727	3021	2250
22	PF3D7_0202400	Exon 1 Segment 2	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	5620	4103	2374
23	PF3D7_0202900	Exon 1 of 1	early transcribed membrane protein 2 (ETRAMP2)	P. falciparum 3D7	IVTT	13390	10008	1423
24	PF3D7_0203100	Exon 2 Segment 3	protein kinase, putative	P. falciparum 3D7	IVTT	8829	3942	2658
25	PF3D7_0203100	Exon 1 of 1	conserved Plasmodium protein, unknown function	P. falciparum 3D7	IVTT	10004	3630	4301

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)

ID	Name	Description	Species	O	P		
005	PVX_122995	Exon 1 of 1 drug/metabolite exporter, drug/metabolite transporter	P. vivax Sall	IVTT	7620	3633	2064
006	PVX_123040	Exon 1 of 1 hypothetical protein, conserved	P. vivax Sall	IVTT	4731	2726	2481
007	PVX_123105	Exon 2 of 6 hypothetical protein, conserved	P. vivax Sall	IVTT	6237	5183	2210
008	PVX_123350	Exon 1 of 1 Segment bromodomain protein, putative	P. vivax Sall	IVTT	10966	4177	3232
009	PVX_123440	Exon 1 of 2 hypothetical protein, conserved	P. vivax Sall	IVTT	3761	2668	1881
010	PVX_123505	Exon 1 of 1 hypothetical protein, conserved	P. vivax Sall	IVTT	7385	3635	4313
011	PVX_123510	Exon 1 of 1 S4, putative	P. vivax Sall	IVTT		8779	4312
012	PVX_123520	DNA-binding chaperone, putative	P. vivax Sall	IVTT	8377	3982	4429
013	PVX_123655	Exon 1 of 3 Segment hypothetical protein, conserved	P. vivax Sall	IVTT	5320	1771	1757
014	PVX_123705	hypothetical protein, conserved	P. vivax Sall	IVTT	6164	3435	2035
015	PVX_123745	Exon 1 of 1 endoplasmic precursor, putative	P. vivax Sall	IVTT		11599	2851
016	PVX_123810	Exon 2 of 2 Segment hypothetical protein, conserved	P. vivax Sall	IVTT	3732	2163	1608
017	PVX_123845	polyadenylation-binding protein, putative	P. vivax Sall	IVTT	7961	2907	2379
018	PVX_123855	Exon 2 of 2 Chromatin assembly protein (ASF1), putative	P. vivax Sall	IVTT	5567	11546	2983
019	PVX_124015	Exon 2 of 3 hypothetical protein, conserved	P. vivax Sall	IVTT	10785	2129	6035
020	PVX_124140	Exon 10 of 11 Segments hypothetical protein, conserved	P. vivax Sall	IVTT	4324	2604	2230
021	NIA	TTBS	NIA	TTBS	1018	1115	974
022	NIA	TTBS	NIA	TTBS	635	562	551
023	NIA	TTBS	NIA	TTBS	572	785	522
024	NIA	TTBS	NIA	TTBS	605	588	771
025	NIA	TTBS	NIA	TTBS	56	122	140
026	NIA	TTBS	NIA	TTBS	33	97	143
027	NIA	TTBS	NIA	TTBS	-14	212	179
028	NIA	TTBS	NIA	TTBS	113	154	194
029	NIA	TTBS	NIA	TTBS	782	1062	940
030	NIA	TTBS	NIA	TTBS	260	917	712
031	NIA	TTBS	NIA	TTBS	304	743	458
032	NIA	TTBS	NIA	TTBS	118	391	518

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)

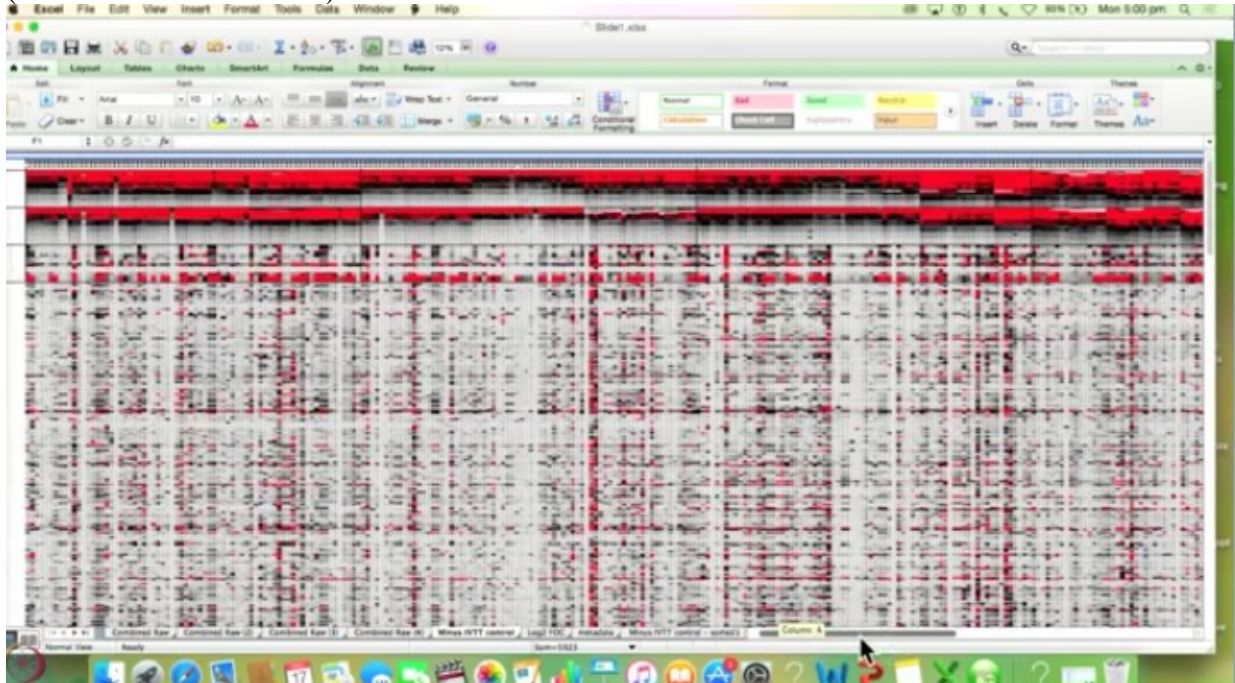
	I	J	K	L	M	N	O	P	
1012	PVX_123520		DNA binding chaperone, putative	F. vivax	SalI	IVTT	8377	3982	4429
1013	PVX_123655	Exon 1 of 3 Segmen	hypothetical protein, conserved	F. vivax	SalI	IVTT	5320	1771	1757
1014	PVX_123705		hypothetical protein, conserved	F. vivax	SalI	IVTT	4164	3435	2035
1015	PVX_123745	Exon 1 of 1	endoplasmic precursor, putative	F. vivax	SalI	IVTT		3102	2853
1016	PVX_123810	Exon 2 of 2 Segmen	hypothetical protein, conserved	F. vivax	SalI	IVTT	1732	2163	1608
1017	PVX_123845		polyadenylate binding protein, putative	F. vivax	SalI	IVTT	7961	2907	2379
1018	PVX_123855	Exon 2 of 2	Chromatin assembly protein (ASF1), putative	F. vivax	SalI	IVTT	5567	3308	2983
1019	PVX_124015	Exon 2 of 3	hypothetical protein, conserved	F. vivax	SalI	IVTT	10780	2129	6035
1020	PVX_124140	Exon 10 of 11 Segm	hypothetical protein, conserved	F. vivax	SalI	IVTT	4324	2604	2230
1021	NIA		noDNA Control	NIA		noDNA	8826	3136	3439
1022	NIA		noDNA Control	NIA		noDNA	8748	3389	3128
1023	NIA		noDNA Control	NIA		noDNA	8503	3026	2690
1024	NIA		noDNA Control	NIA		noDNA	8971	2873	2636
1025	NIA		noDNA Control	NIA		noDNA	8483	2736	2627
1026	NIA		noDNA Control	NIA		noDNA	9074	2941	2834
1027	NIA		noDNA Control	NIA		noDNA	12212	6601	5098
1028	NIA		noDNA Control	NIA		noDNA	6980	3050	3151
1029	NIA		noDNA Control	NIA		noDNA	6561	2632	2717
1030	NIA		noDNA Control	NIA		noDNA	7721	2642	2722
1031	NIA		noDNA Control	NIA		noDNA	7247	2928	2915
1032	NIA		noDNA Control	NIA		noDNA	7524	2716	2685
1033	NIA		noDNA Control	NIA		noDNA	7429	2824	2940
1034	NIA		noDNA Control	NIA		noDNA	6354	2780	2872
1035	NIA		noDNA Control	NIA		noDNA	6356	2435	2326
1036	NIA		noDNA Control	NIA		noDNA	6647	3518	2420
1037	NIA		noDNA Control	NIA		noDNA	5640	2521	2651
1038	NIA		noDNA Control	NIA		noDNA	6277	2600	2774
1039	NIA		noDNA Control	NIA		noDNA	7933	2218	1533

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.

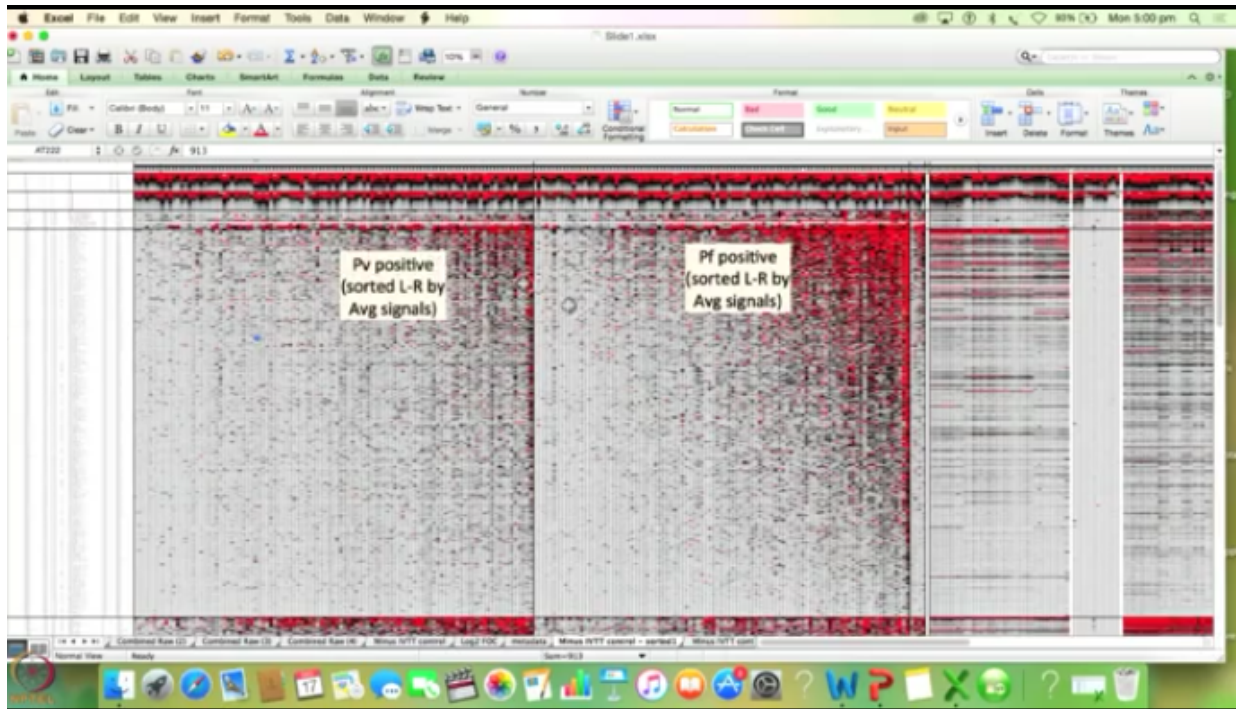
So now there are 24 such spots which you remember we have rearranged and that's why it's come together, grouped together like this, the first thing we are going to do is take a median of this which I have already provided you here, so this is the formula for it, I've just done this the whole thing in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



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.

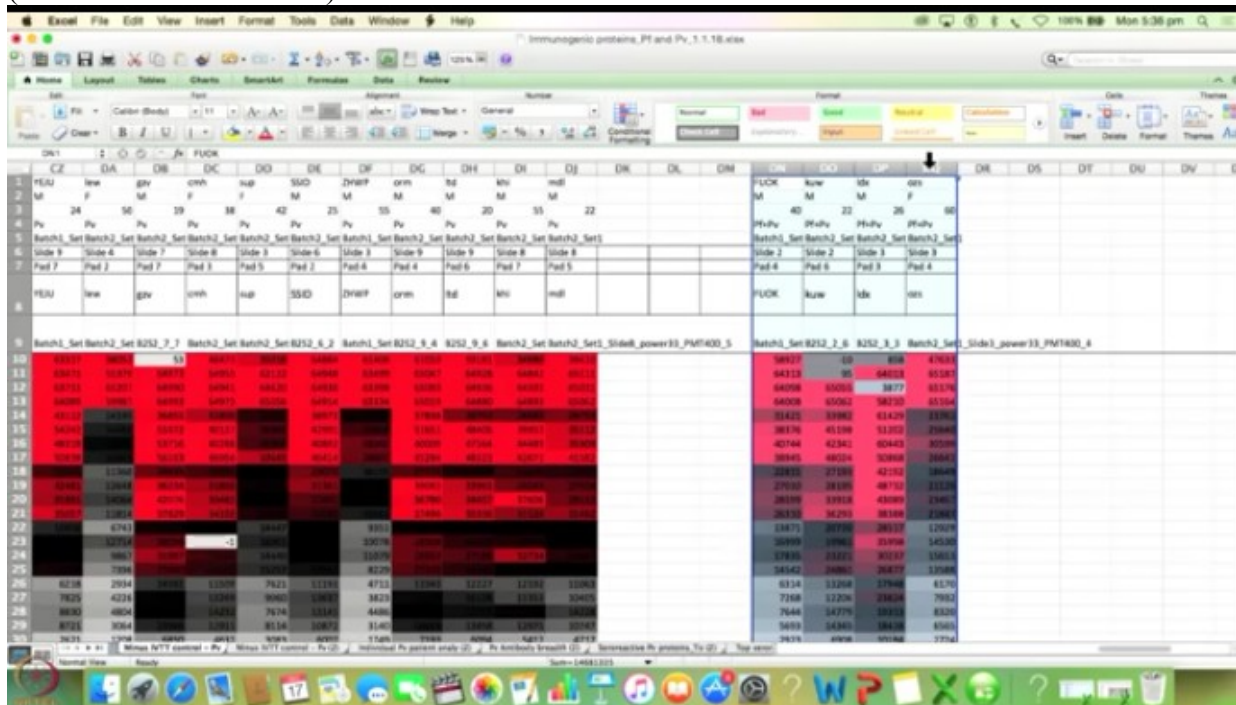
The next step what we are going to do is to divide each and every raw value by the median of the IVTT control spots, so just like how we did previously we subtracted raw values from the median of the IVTT control spots, this time we are going to divide it so that's what is called fold over control.

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



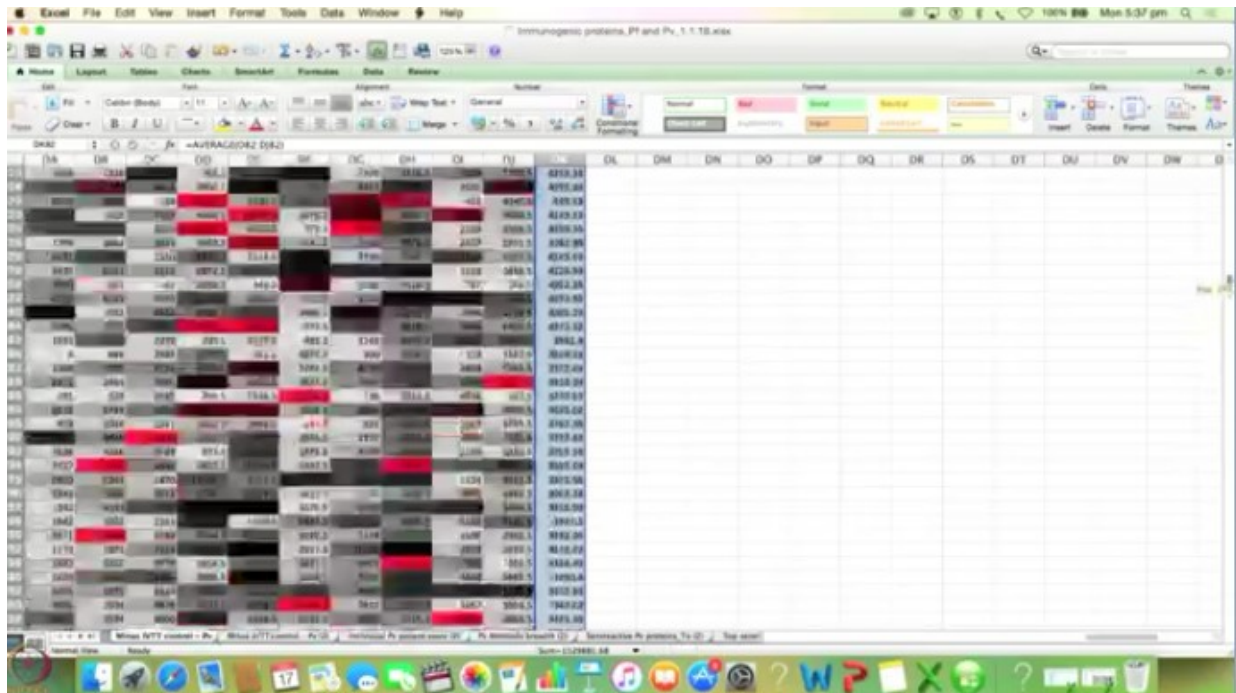
So now if you see that I have retained all the rows, so the first thing what we need to do of course we don't need this, but I have still retain the entire sheet from the beginning you will see that there are these 4 patients which are deliberately kept out of the analysis, (Refer Slide Time: 26:40)



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.

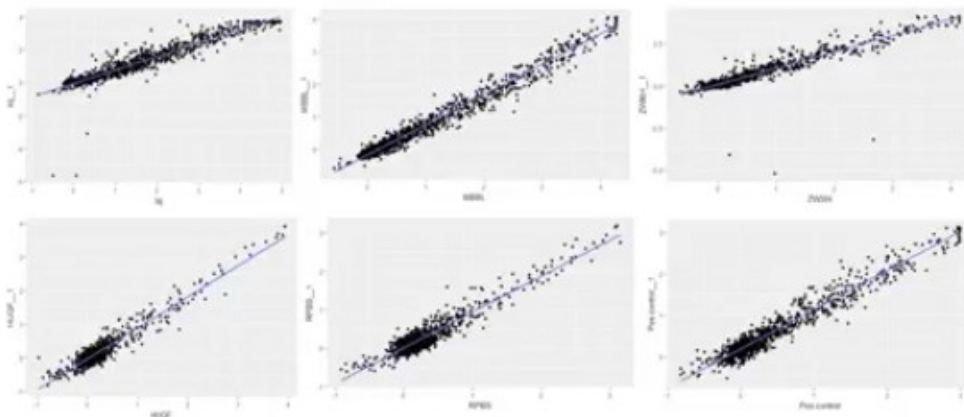
So we need to basically start from row number 82 that's what we are interested in, because these are the IVTT spots, so the first thing I'm going to do is I'm going to an average for each and every spot, so example let's write here average, and I'm going to say is equal to, so I get an average value 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)

### Checking reproducibility across the batches



- In order to confirm **minimal inter-batch variability**, samples probed earlier were **re-probed** on different sets of slides
- Scatter plots of signal intensities of samples probed at different times on different slides indicates uniformity across different batches



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)

$$\text{Avg Raw value/antigen} > \text{Avg [Mean (No DNA controls) + 2SD (Mean)]} = \text{Seroreactive antigens}$$



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.

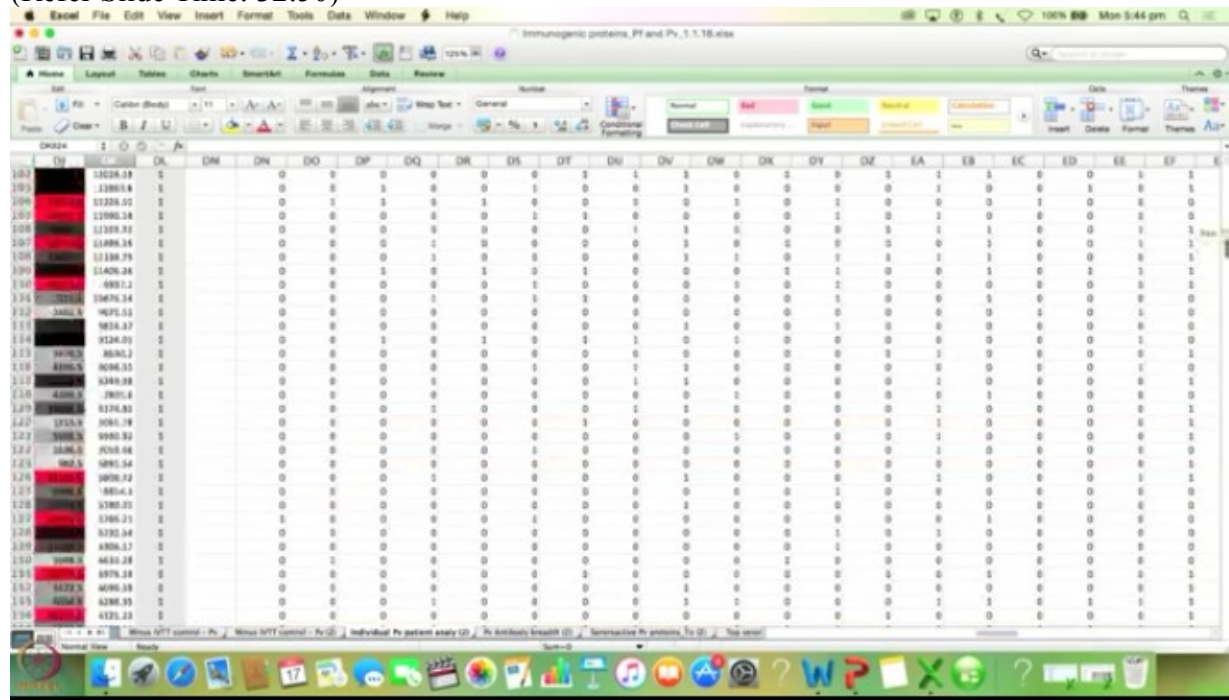
So what this means is that if this is the average if this particular number is greater than this particular number which I'm going to show you right now, if you take an average of the mean + 2 times standard deviation of the no DNA spots, this is my number, so if that raw value or if any raw value is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

### **Points to Ponder**

- **Basic microarray data can be analysed using excel, however programming is highly recommended for larger data analysis**
- **R programming is a simple language and has been very useful for biologists**
- **Two types of normalization was used for the analysis: Sample specific median normalization is used for the purpose of visualization, while Log2 FOC is used for statistical analysis because it is more stringent**

## Points to Ponder

- The use of control samples and control spots are very important for any microarray data
- IgG positive controls are used for checking the overall experiment performance. In some cases, re-scanning at different PMT settings is recommended
- Multiple softwares could be explored for data analysis as well, in case programming is not an option



**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)

# *Next lecture ....*

## Applications of protein microarrays in Cancer Research-I

**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**

**Apoorva Venkatesh**

**Teaching Assistants**  
**Shalini Aggarwal**

**Nikita Gahoi**

**Bharati Sakpal  
Project Manager**

**Bharati Sarang**

**Project Research Associate**

**Nisha Thakur**  
**Sr. Project Technical Assistant**

**Vinayak Raut**  
**Project Assistant**

**Copyright NPTEL CDEEP, IIT Bombay**