



Micro Array Error Analysis

Protein Interaction Map Integration & Visualization

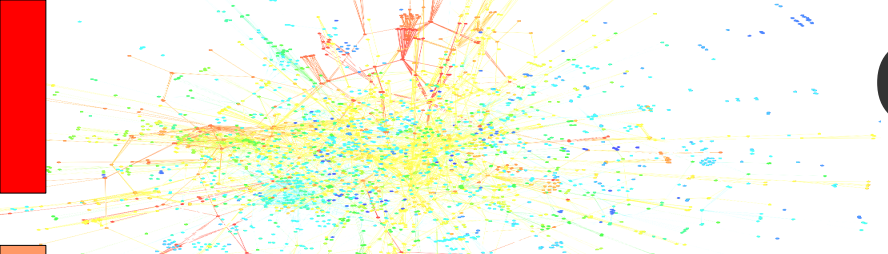
Dr. Werner Van Belle
werner.van.belle@gmail.com
werner@onlinux.be
Department of Virology
University of Tromsø



I. Measurements (I)

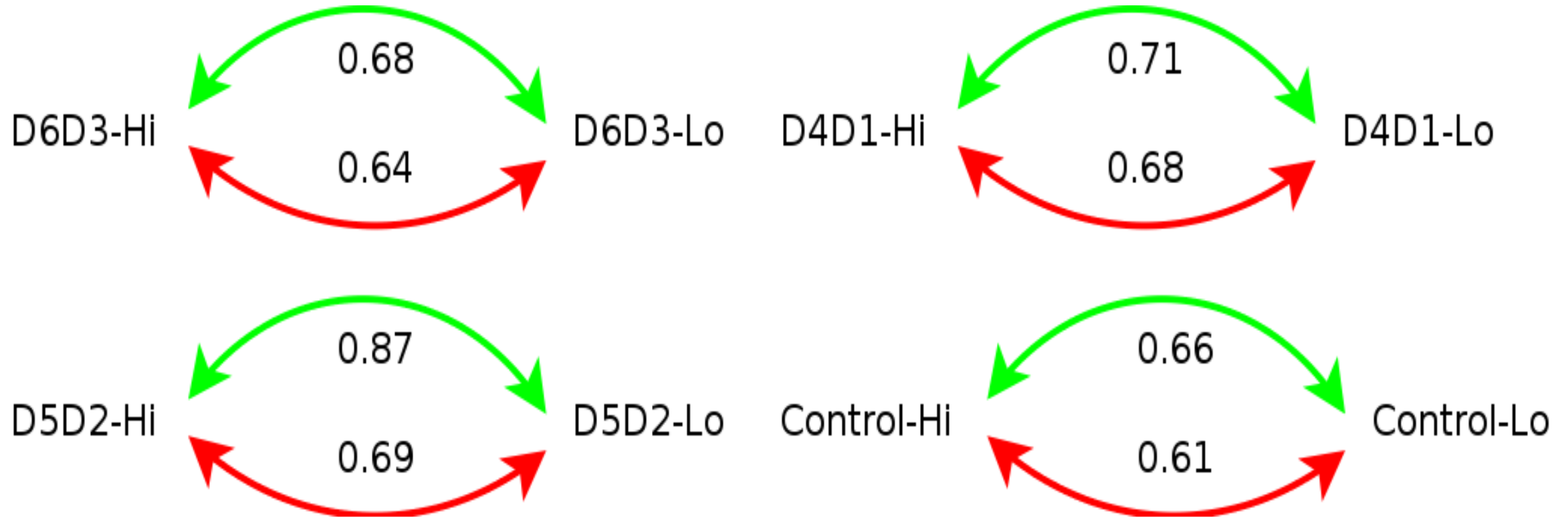
- **Rat Operon Oligo Array S5 [31 May]**
 - 1 plate measured with 1 gain
 - skip since non compatible with new measurements
- **KTH Rat 27k Oligo Micro Array**
 - 3 plates same sample, 2 gains
 - 1.12 microgram [[D4D1-Hi](#) & [D4D1-Lo](#)]
 - 3.02 microgram [[D5D2-Hi](#) & [D5D2-Lo](#)]
 - 2.44 microgram [[D6D3-Hi](#) & [D6D3-Lo](#)]
 - 1 control plate, 2 gains [[Control-Hi](#) & [Control-Lo](#)]

Correlations

- 
- Between gains
 - Between channels
 - Between plates
 - Towards the control

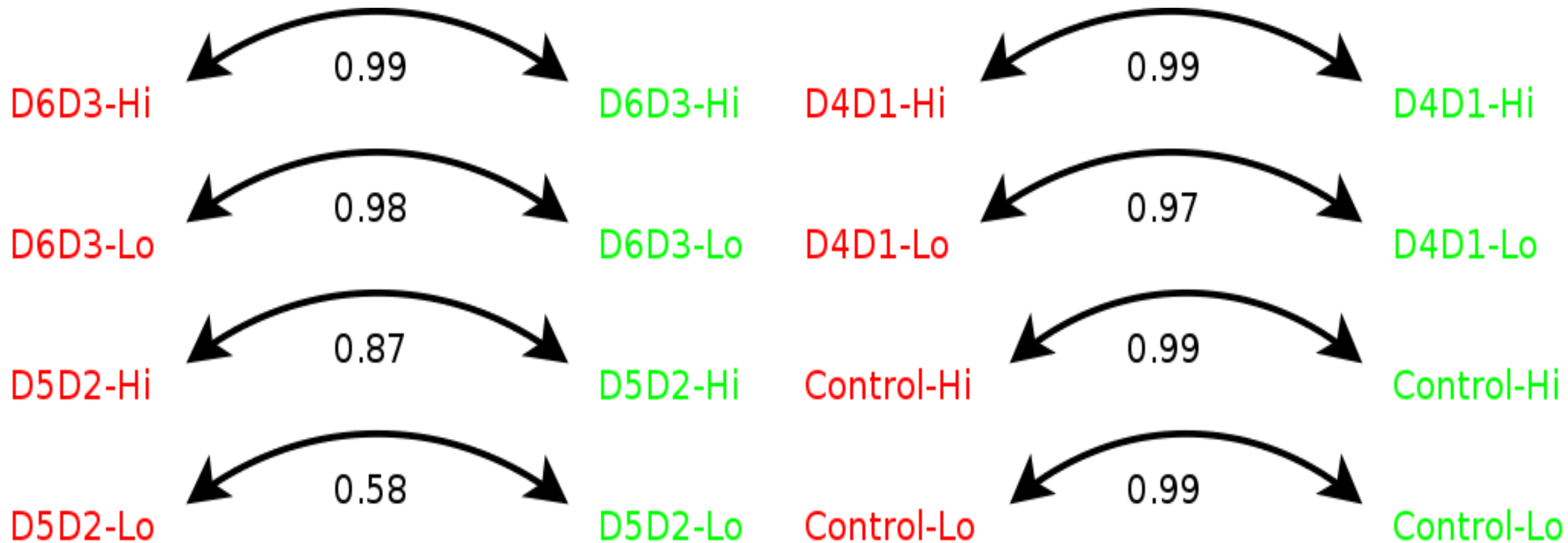
Correlations

- between hi and low gains



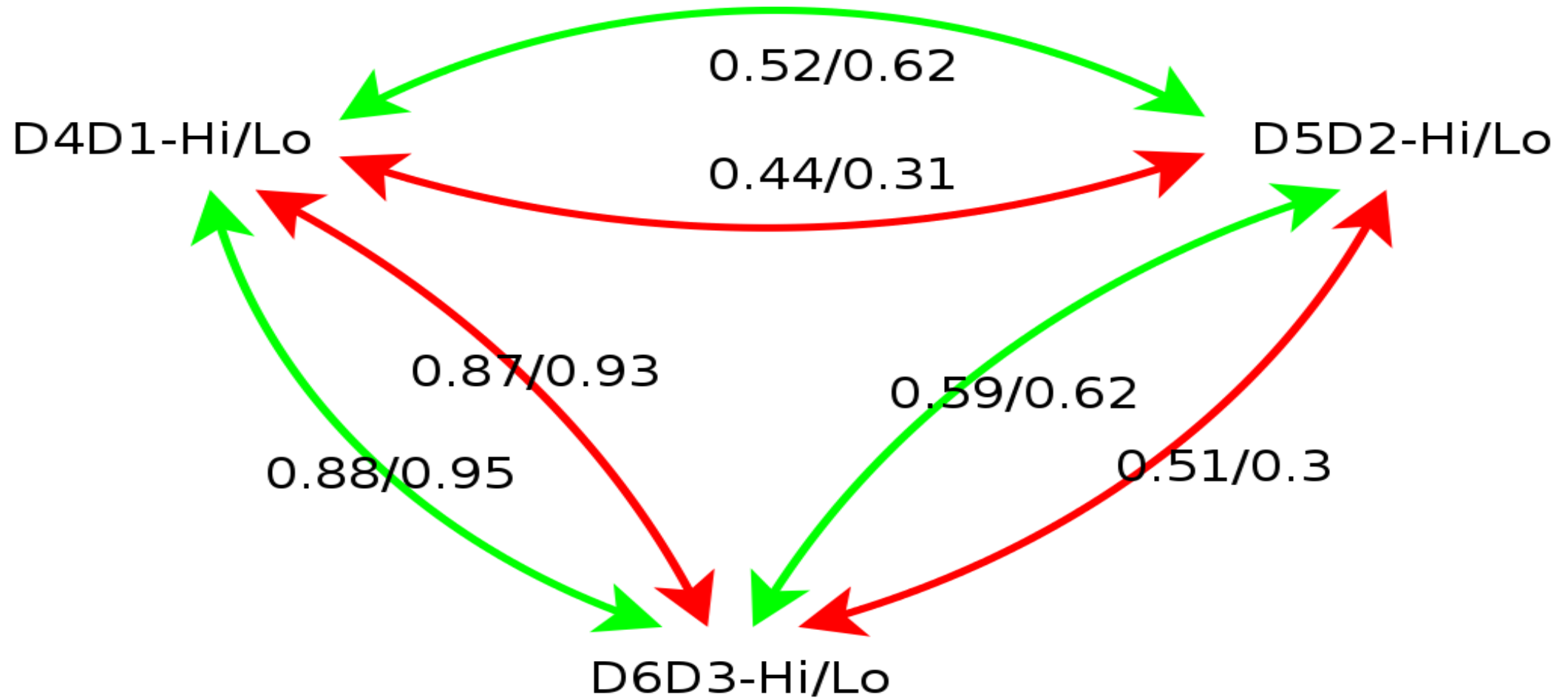
Correlations

- between red and green channels



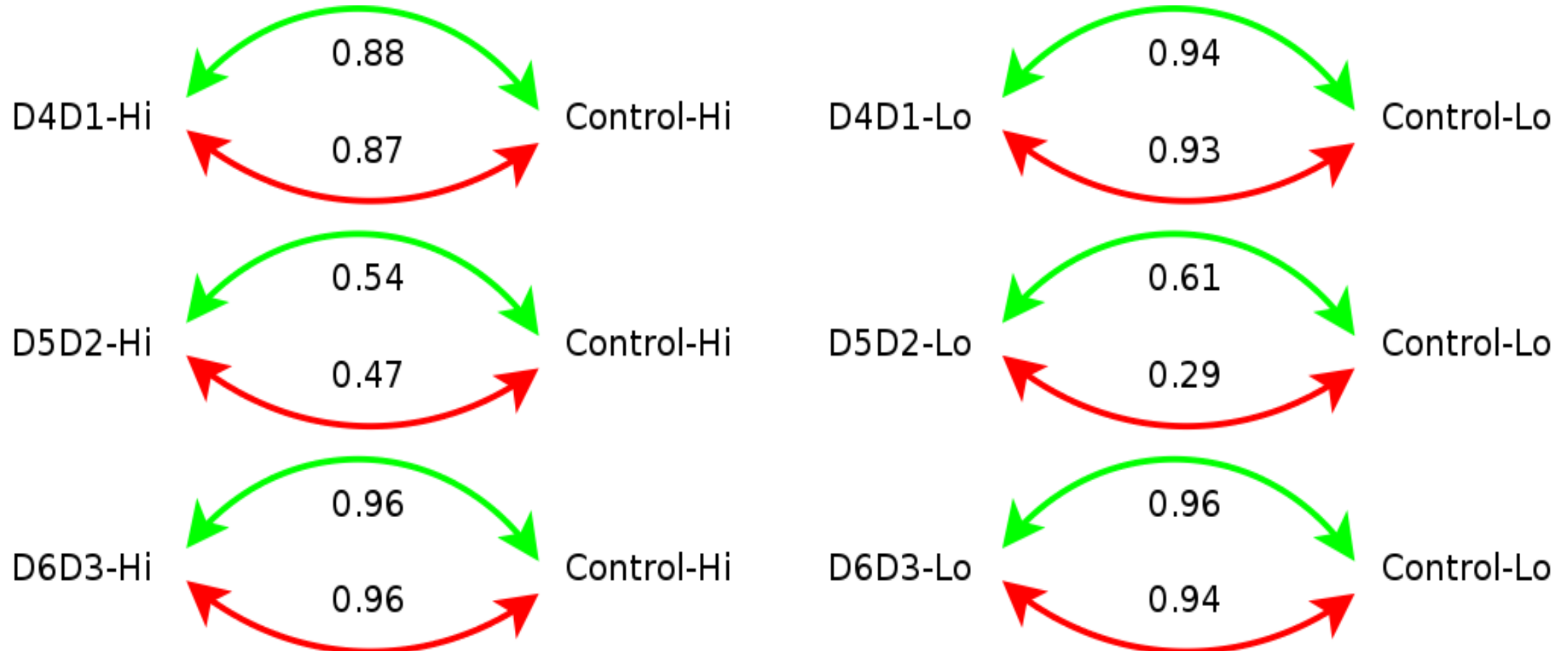
Correlations

- between plates



Correlations

- towards control



Measurement Errors

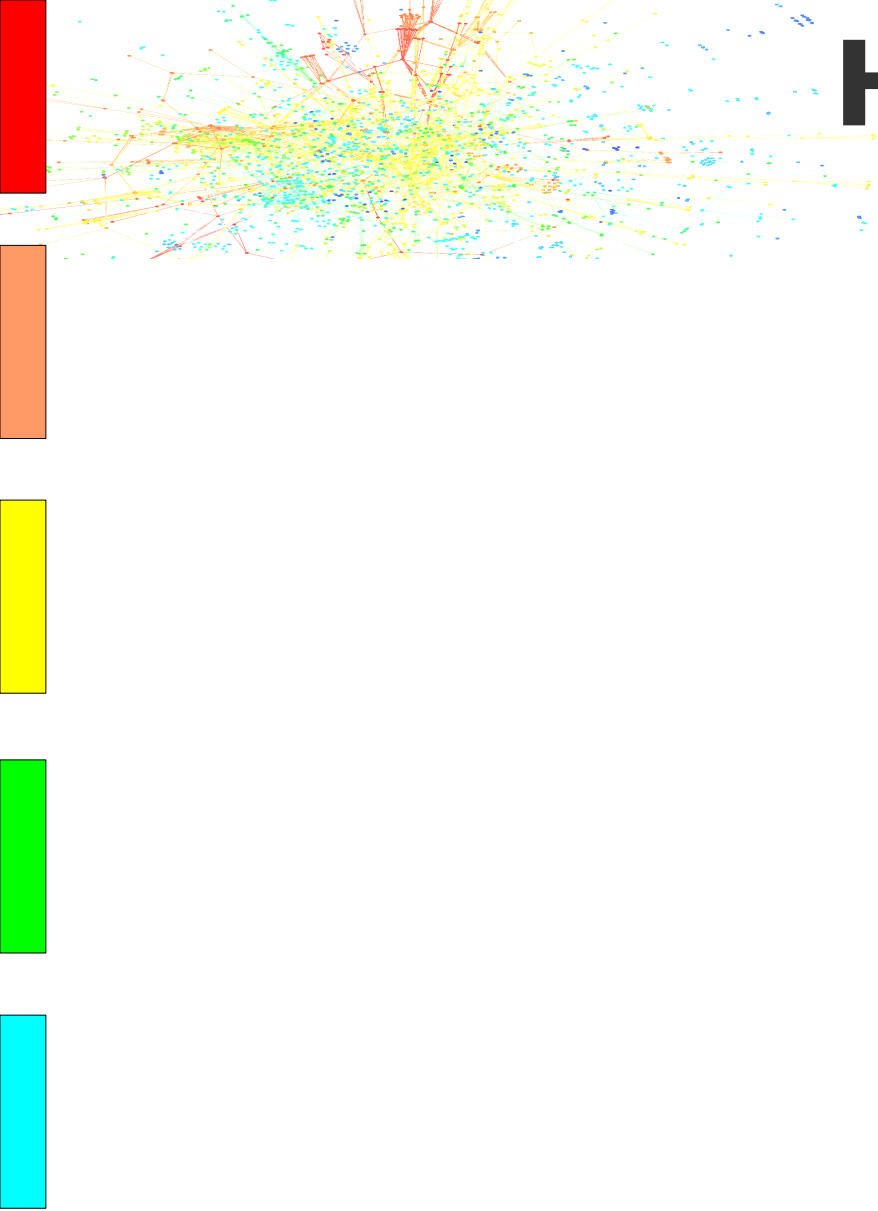




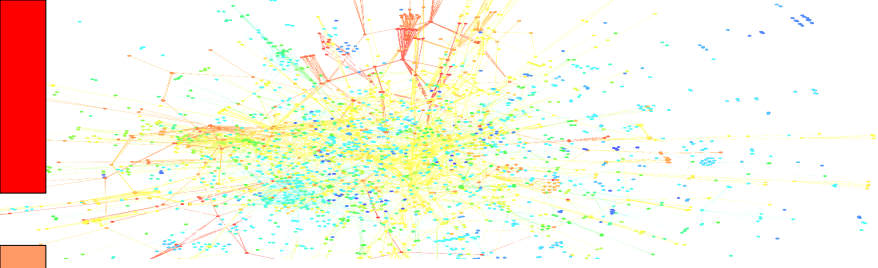
Sources of Errors

- Chemical/Physical
 - Hybridization
 - Quenching -> Normalisation
- Machine related
 - Dynamic range
 - Absolute errors
 - Relative errors

Hybridization



Controls

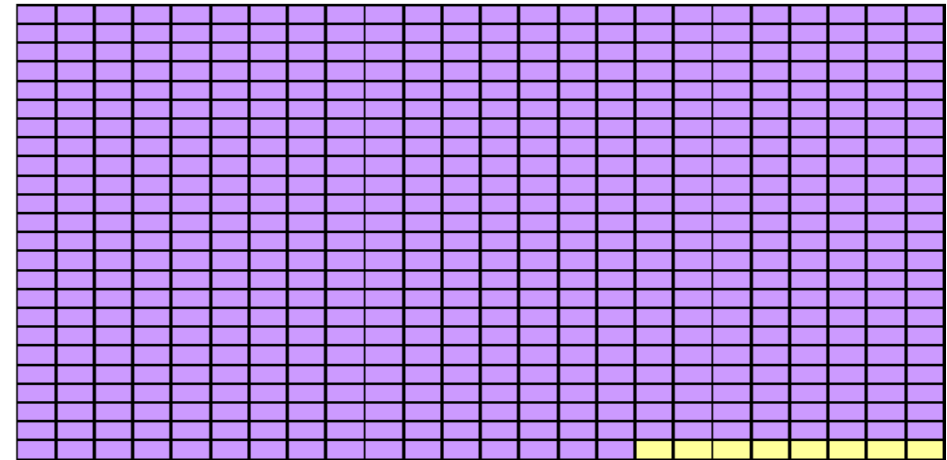
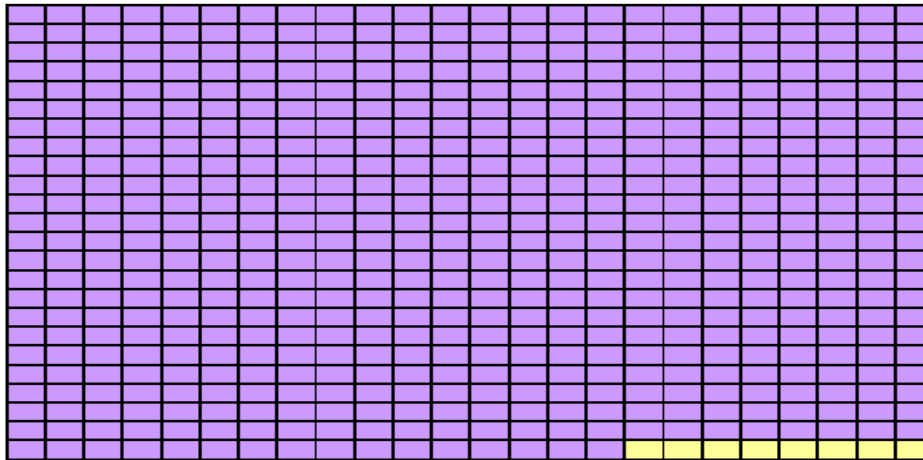
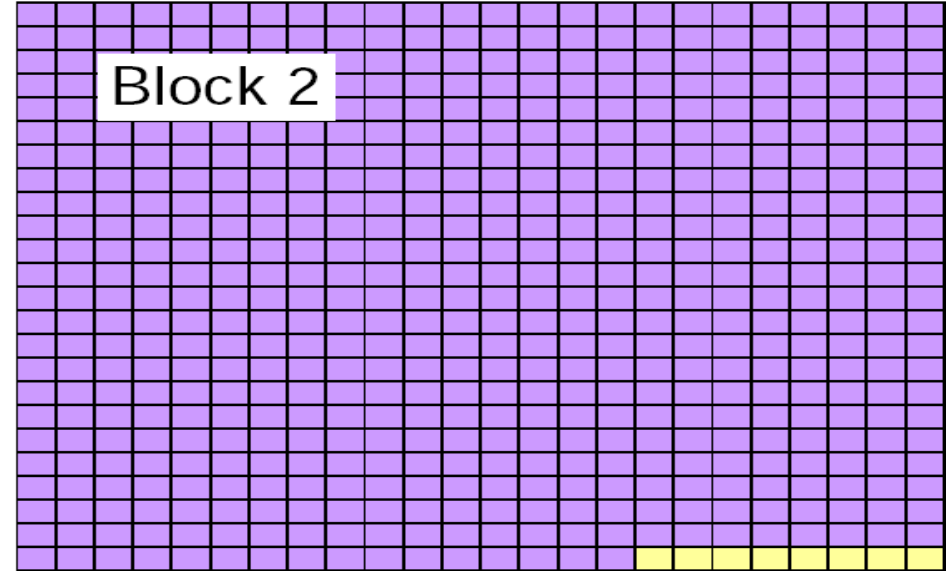
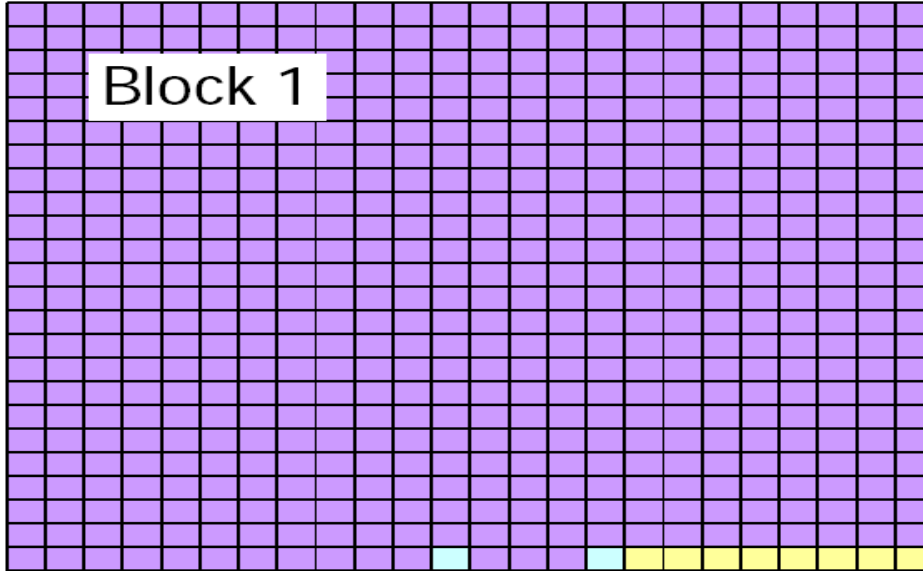


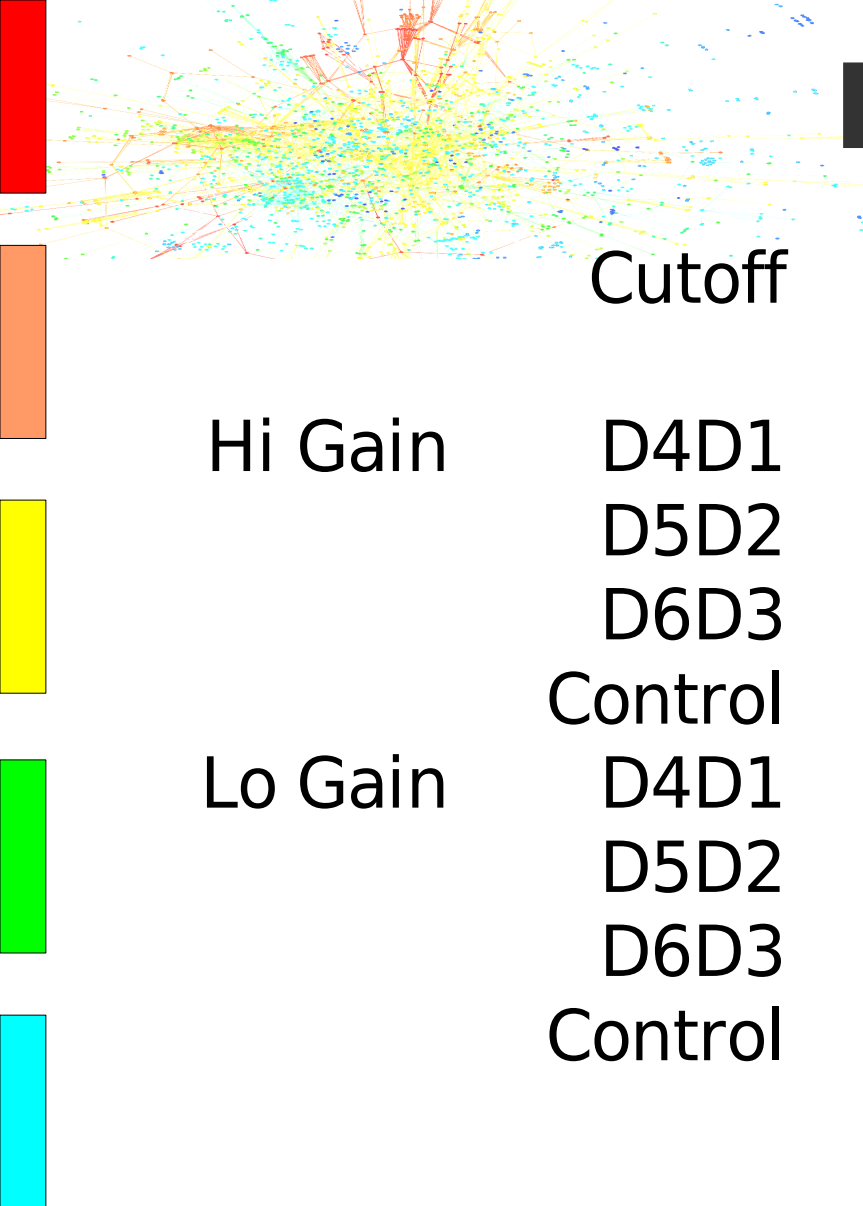
1 Coloumn 24

1
·
r
o
w
·
2
4

Block 1

Block 2



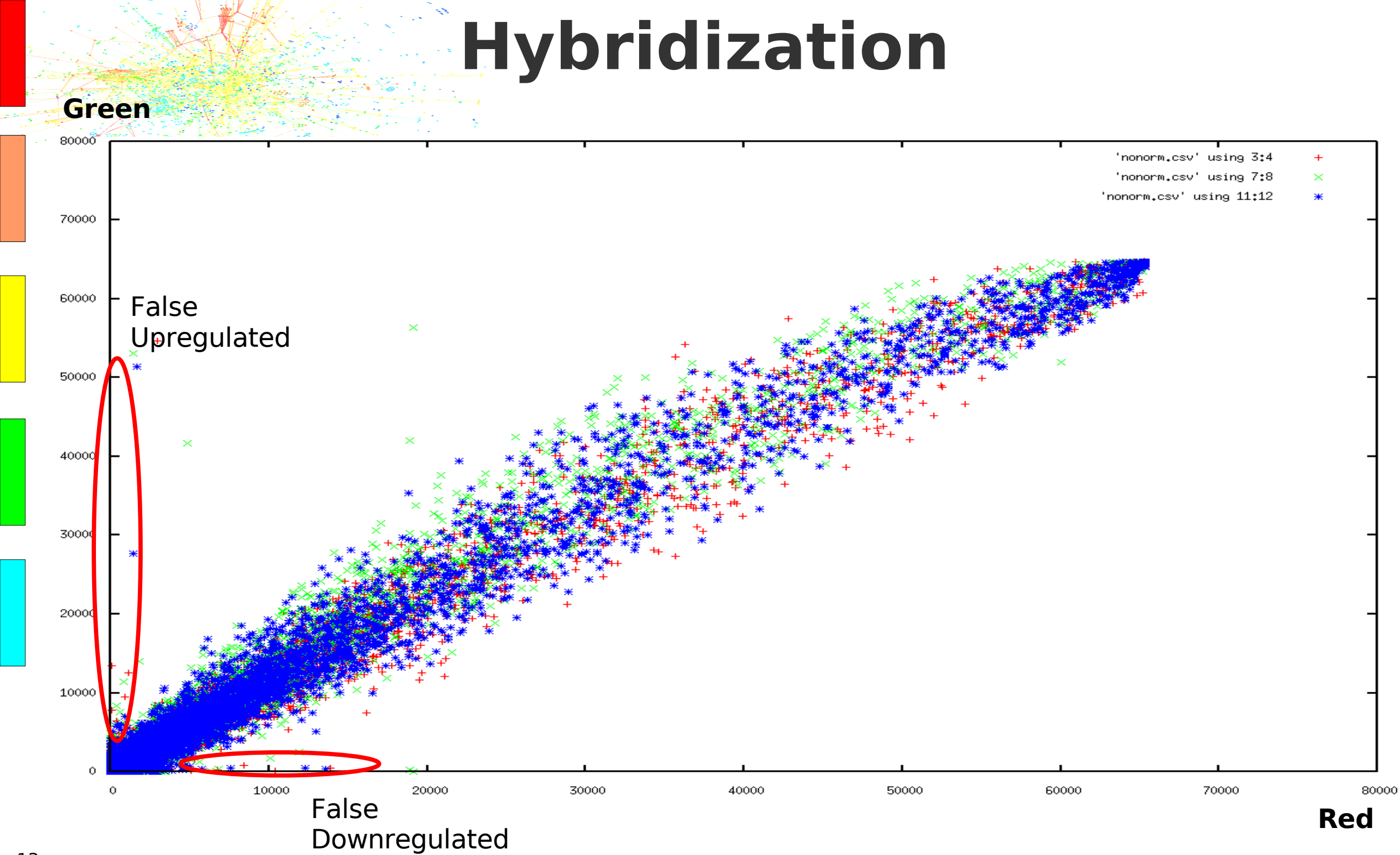


Hybridization

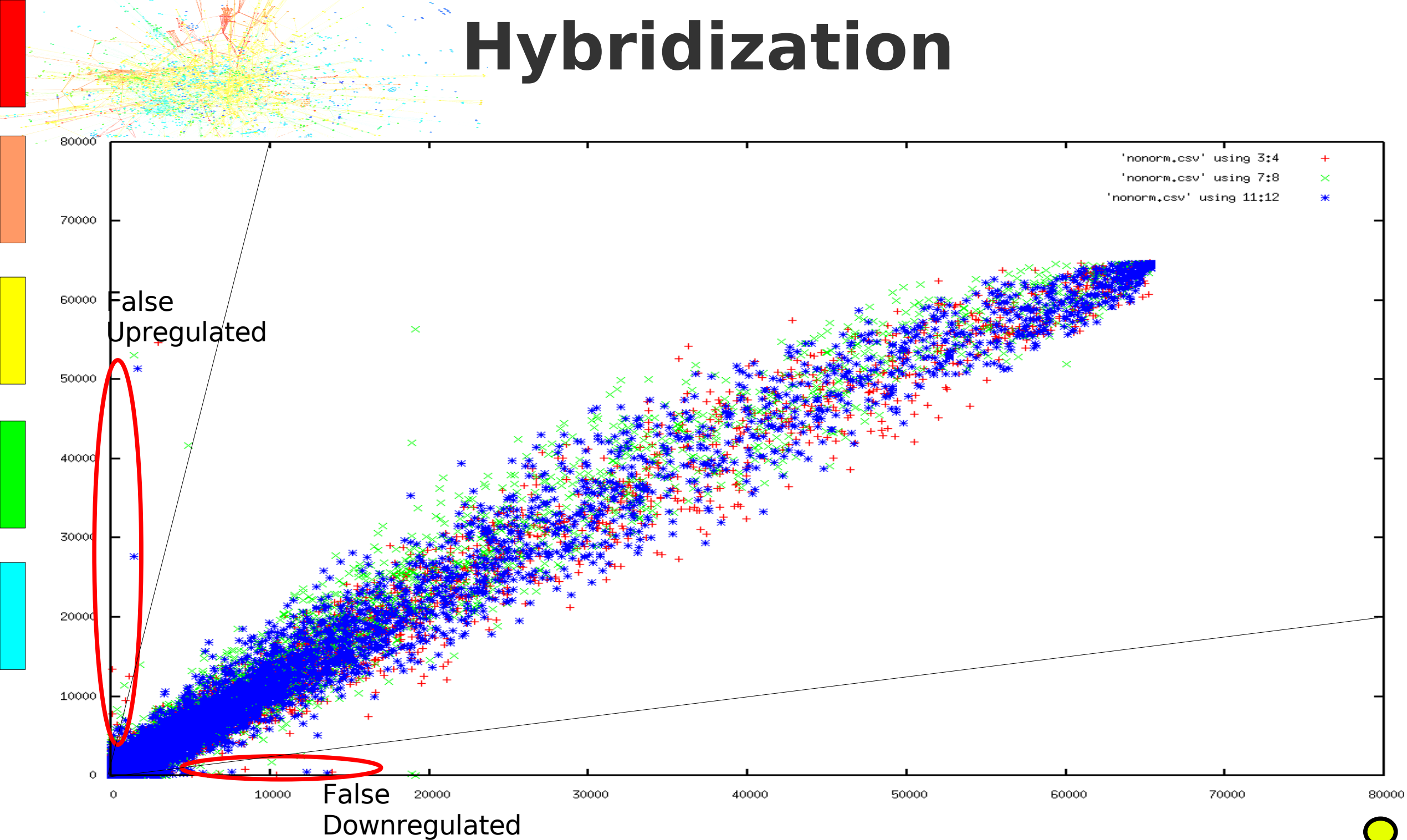
		Cutoff	Average		Median + 1 Stddev		
			Green	Red	Green	Red	
Hi Gain	D4D1		0.29	0.32	0.61	0.67	0.64
	D5D2		0.28	0.75	0.68	0.97	0.83
	D6D3		0.33	0.41	0.64	0.75	0.7
	Control		0.32	0.39	0.6	0.75	0.68
Lo Gain	D4D1		0.21	0.21	0.59	0.63	0.61
	D5D2		0.27	0.49	0.59	0.91	0.75
	D6D3		0.27	0.31	0.64	0.58	0.61
	Control		0.24	0.29	0.59	0.74	0.67
			0.28	0.4	0.62	0.75	0.68

=> At least 70% hybridisation

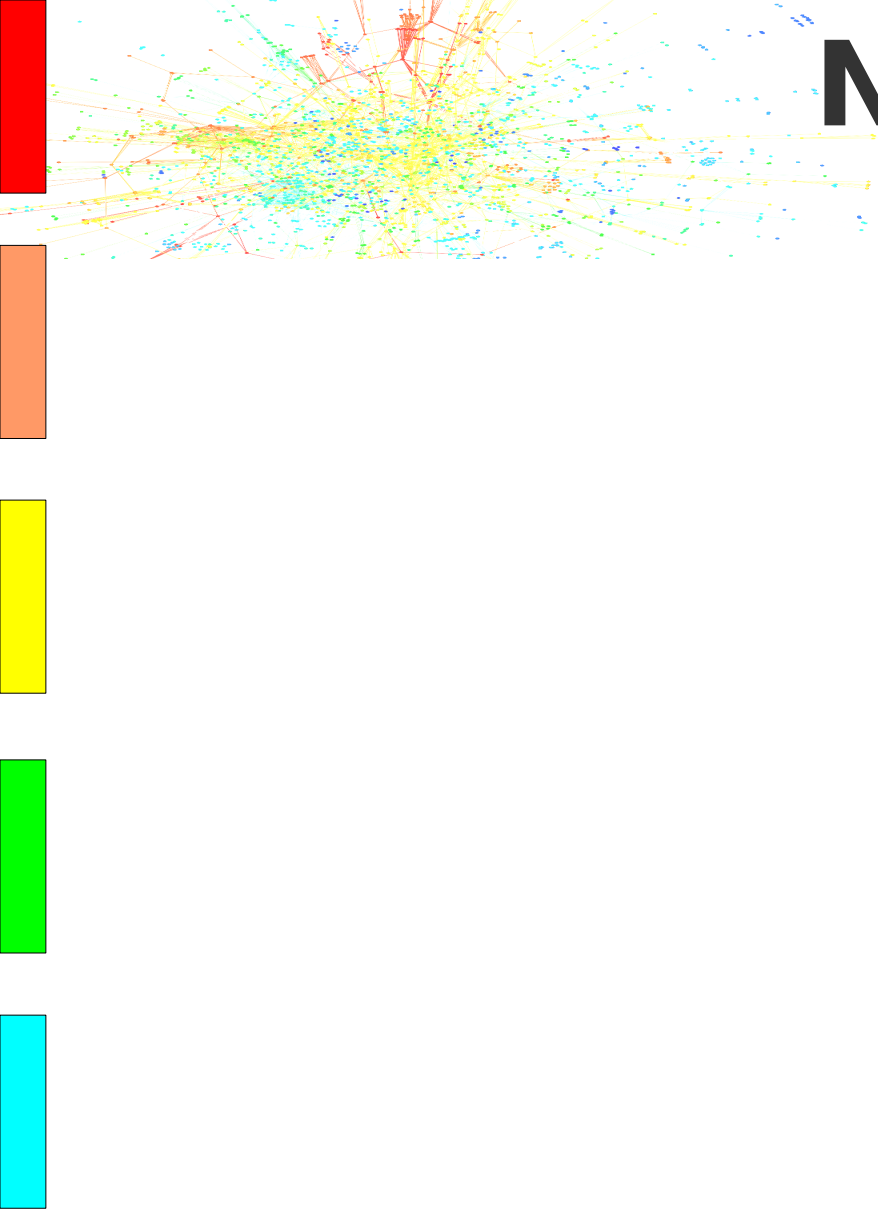
Hybridization



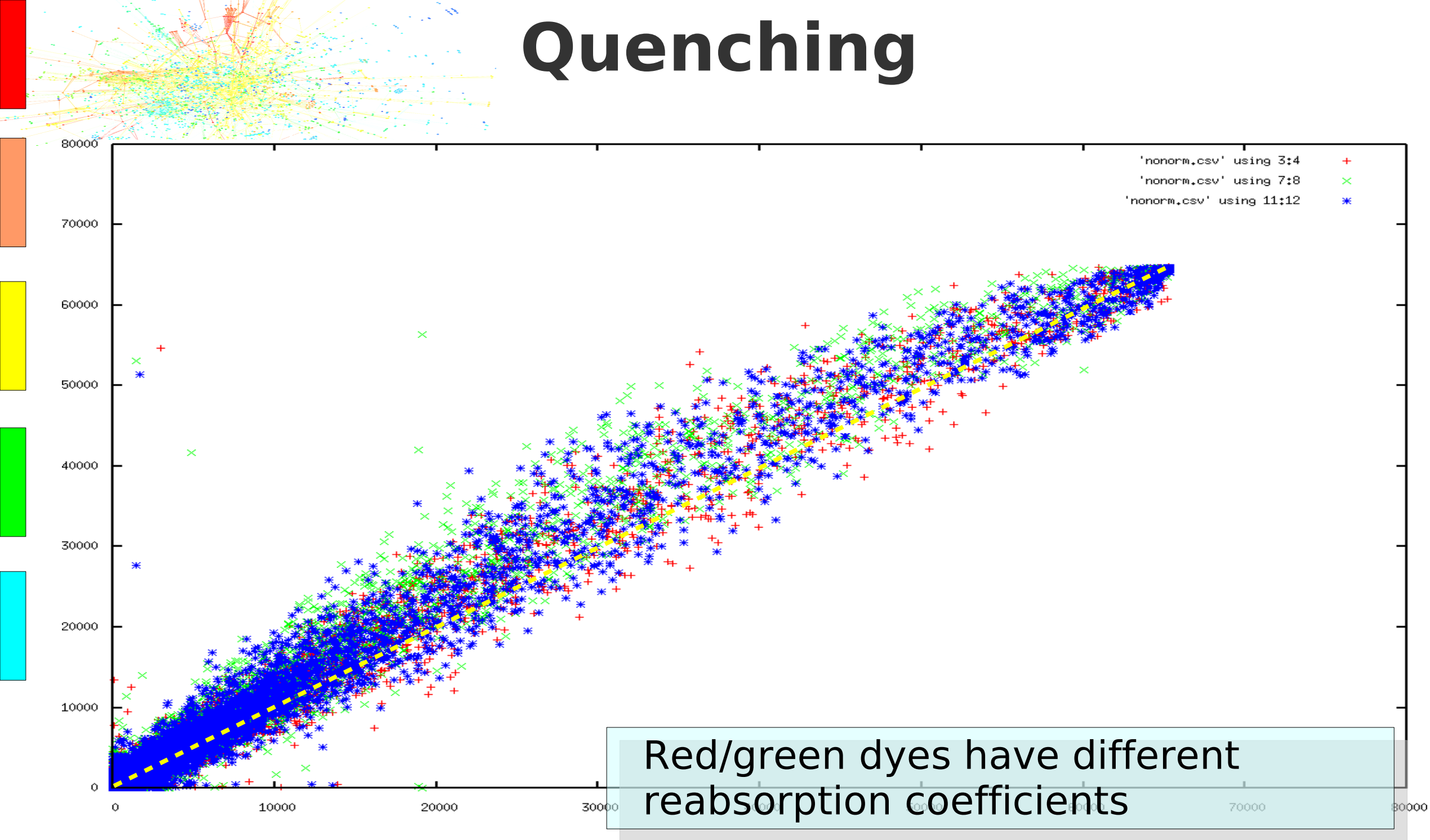
Hybridization



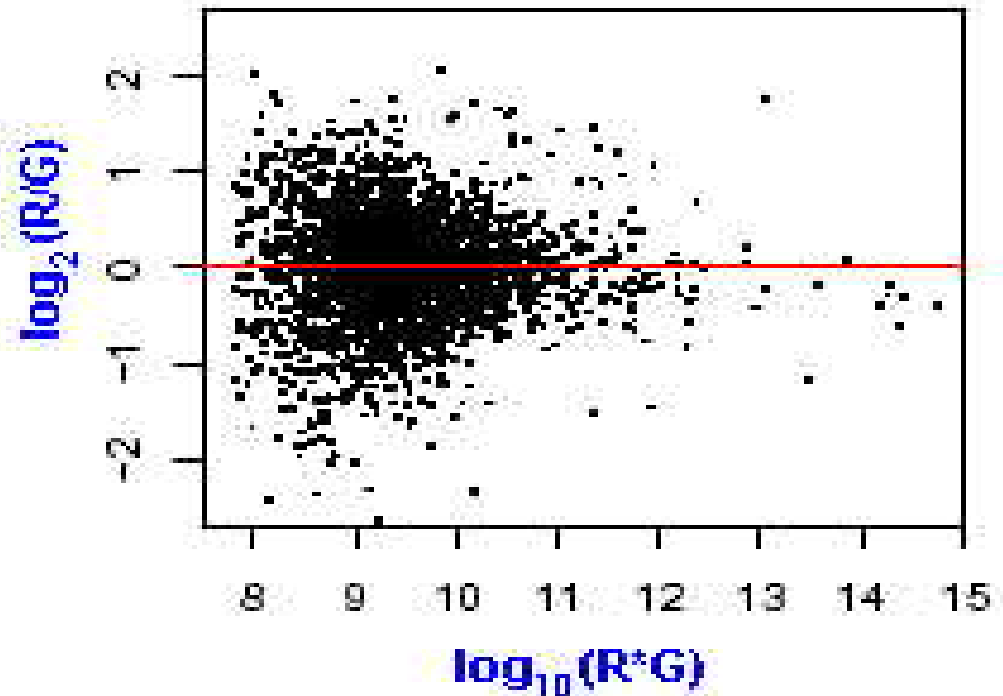
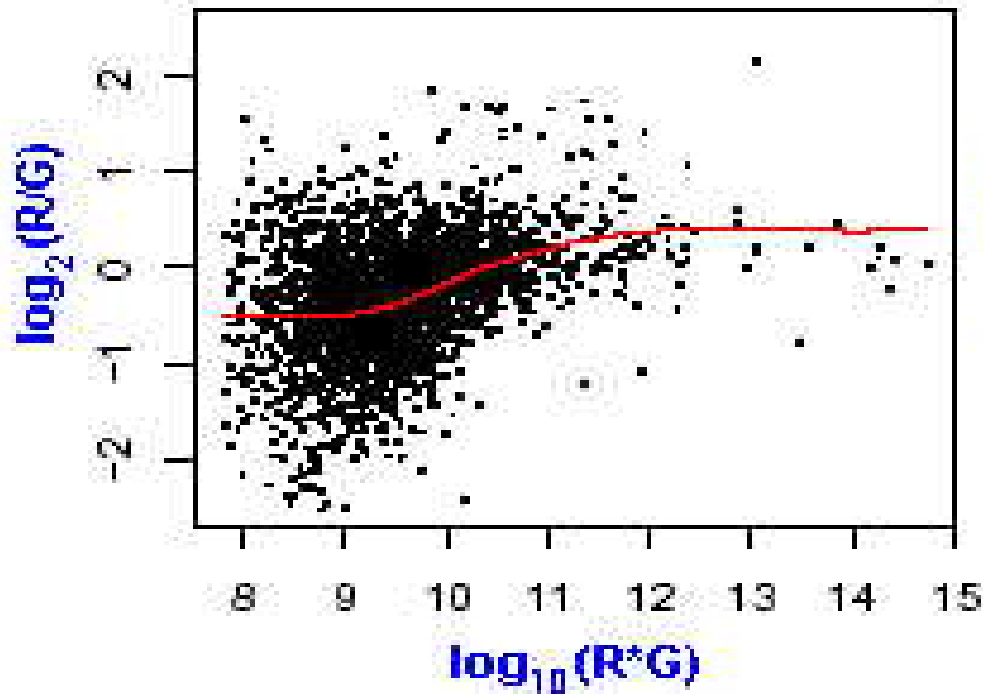
Normalization



Quenching

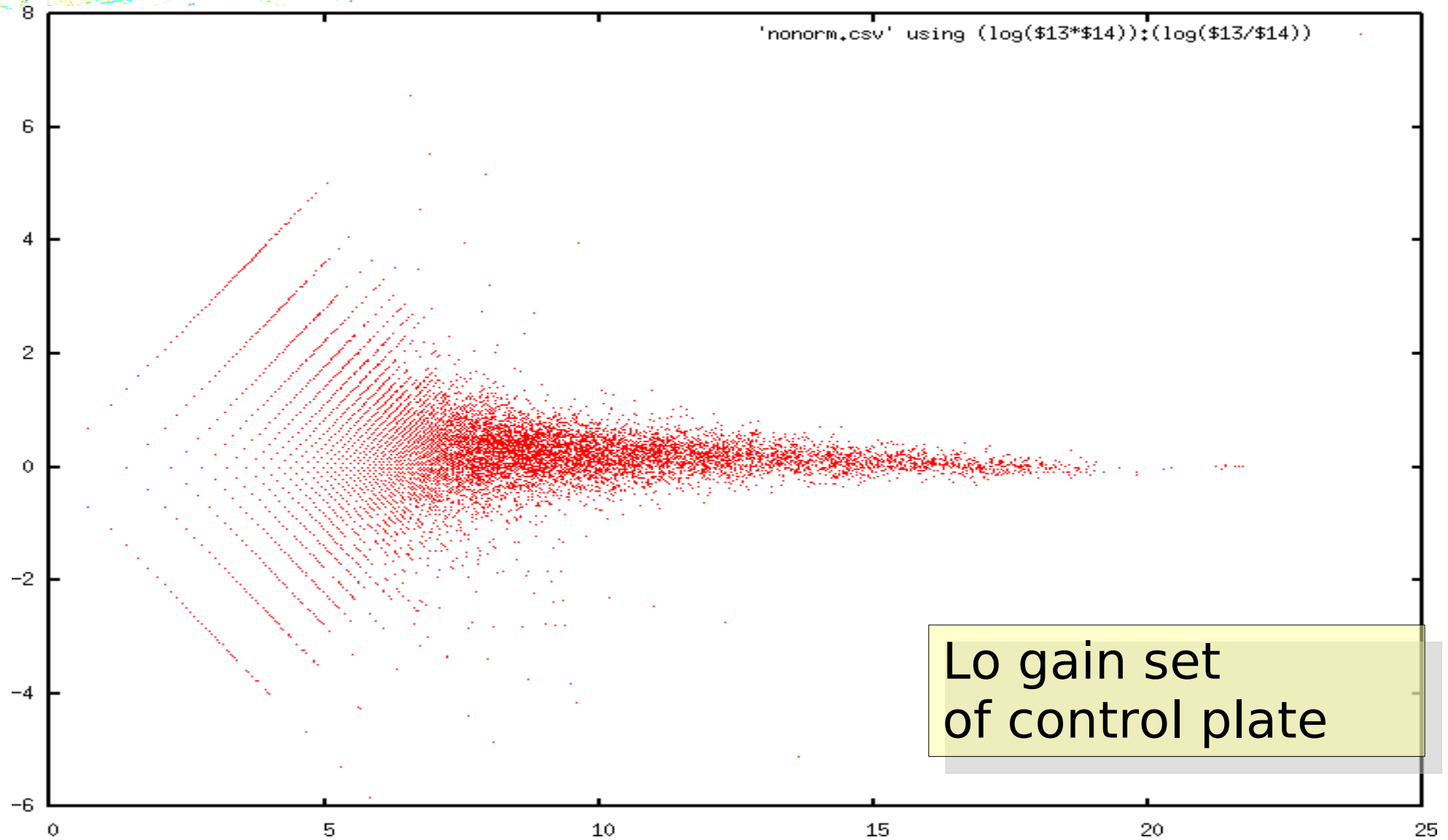


Lo[w]ess Normalization

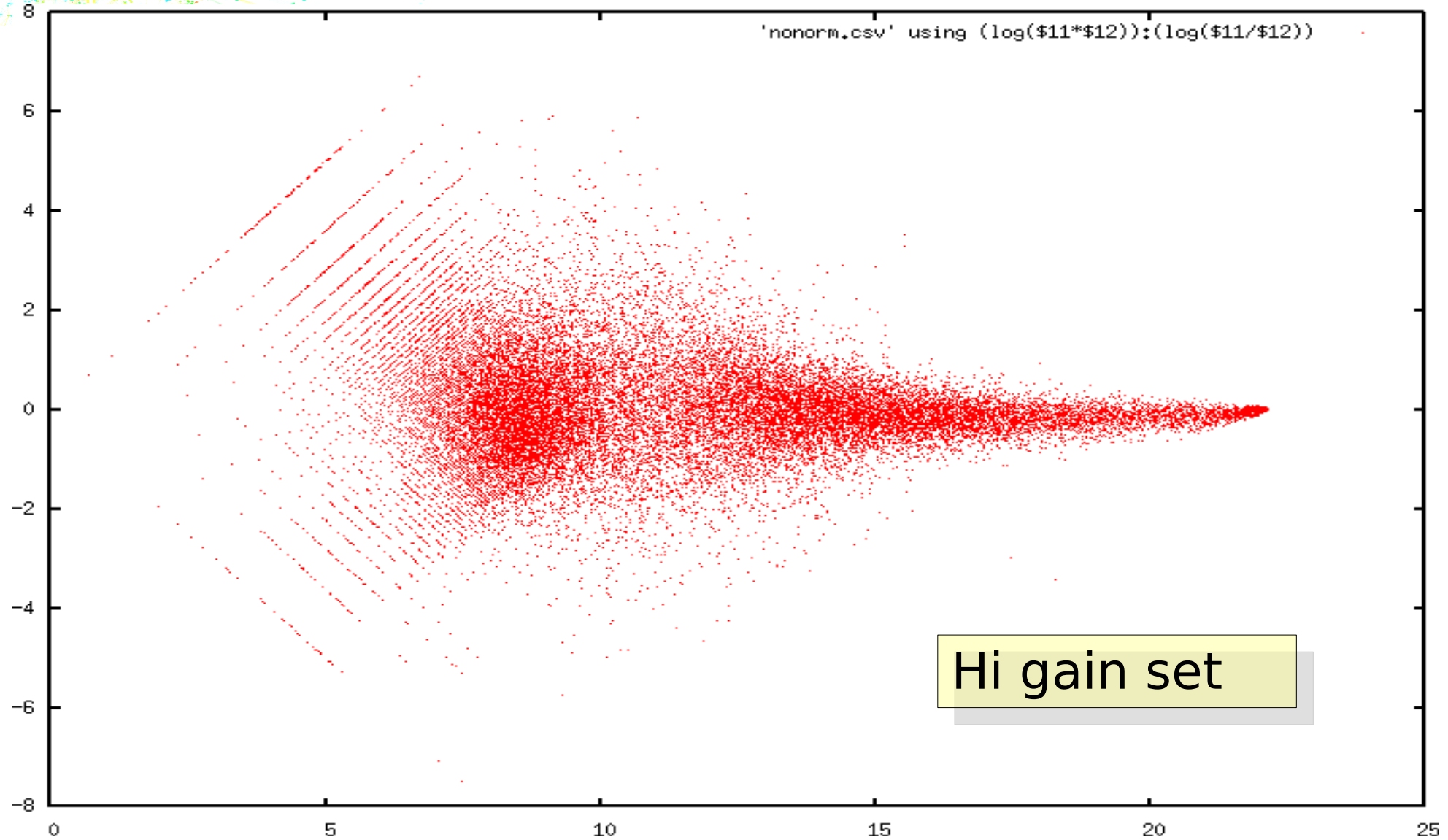


Reference: Taken from
http://www.ucl.ac.uk/oncology/MicroCore/HTML_Resource/Norm_Lowess1.htm

Lo[w]ess normalization



Lo[w]ess normalization

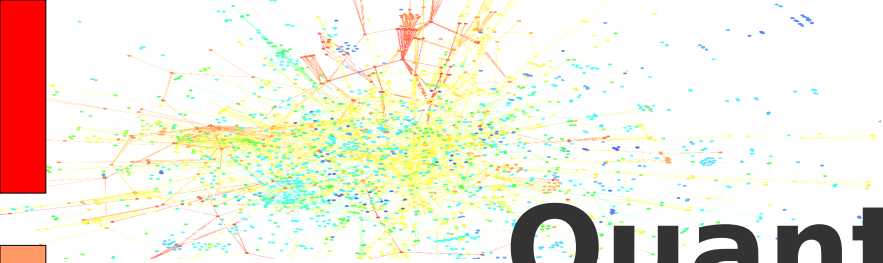




Lo[w]ess Normalization

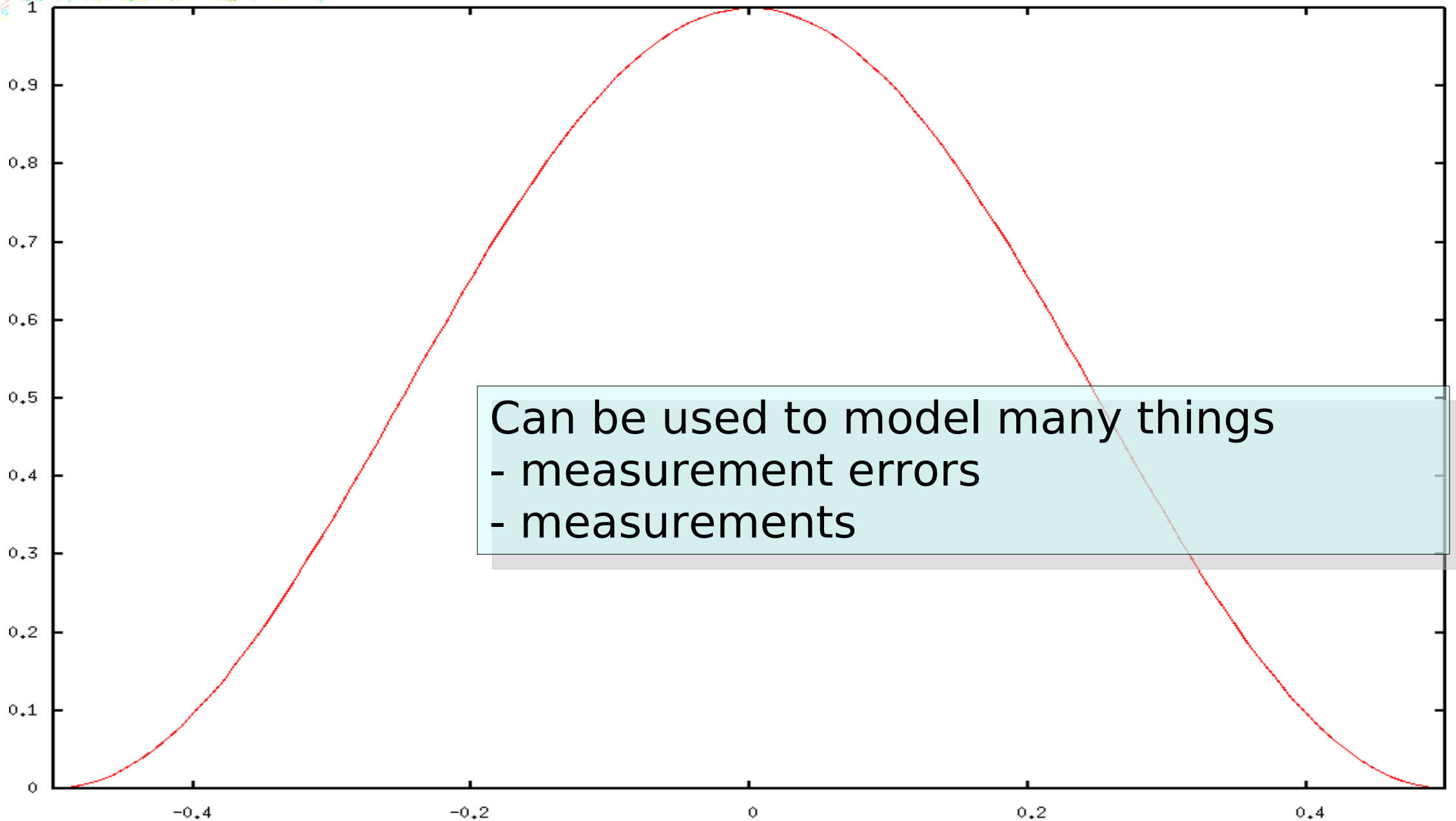
- Numerical problems ($\log \lll$)
- Ignores measurement errors
 - 2/1 has clearly more measurement errors involved than 2000/1000
- Normalizes the log ratio
 - Loose separate R and G values



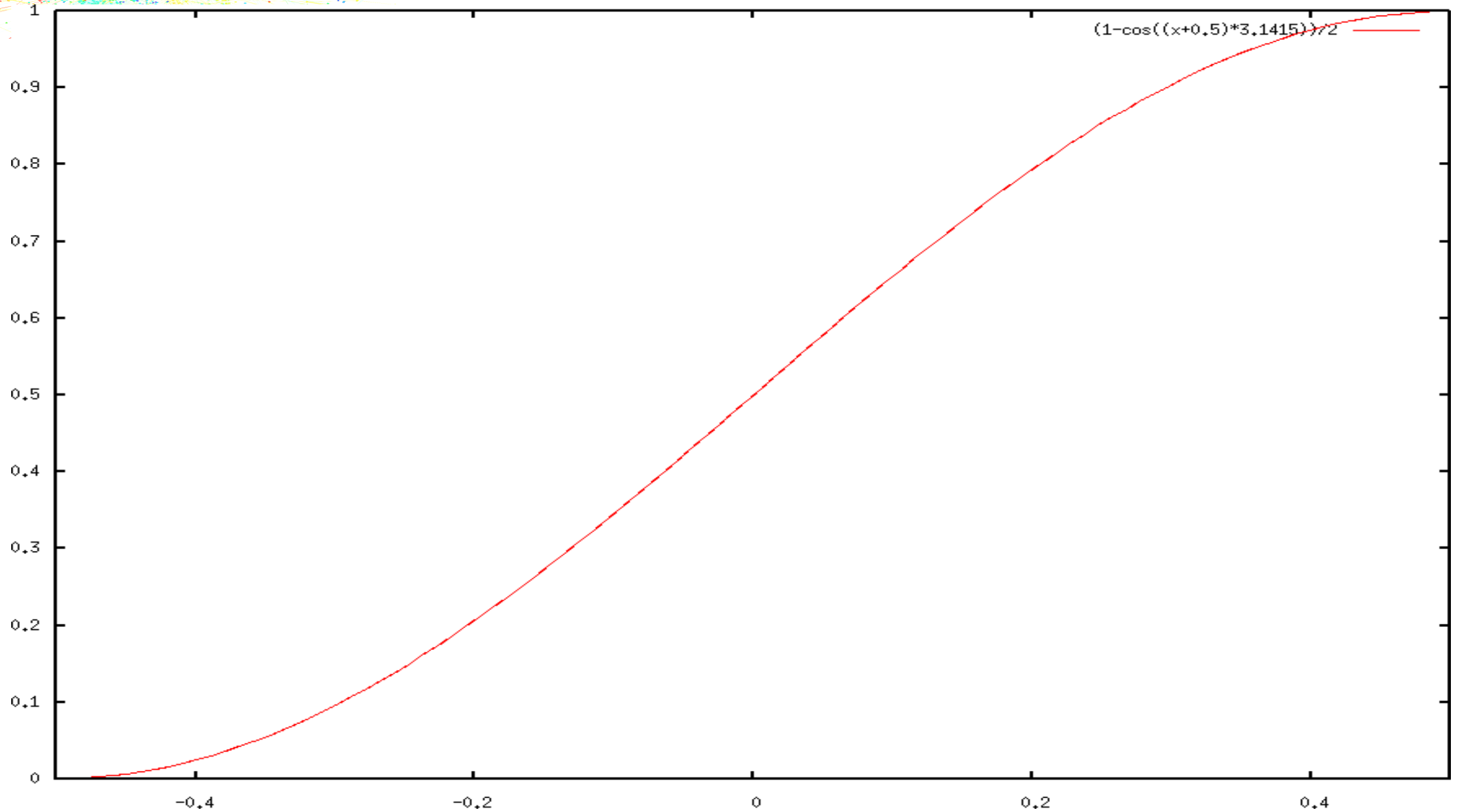


Quantile Normalization

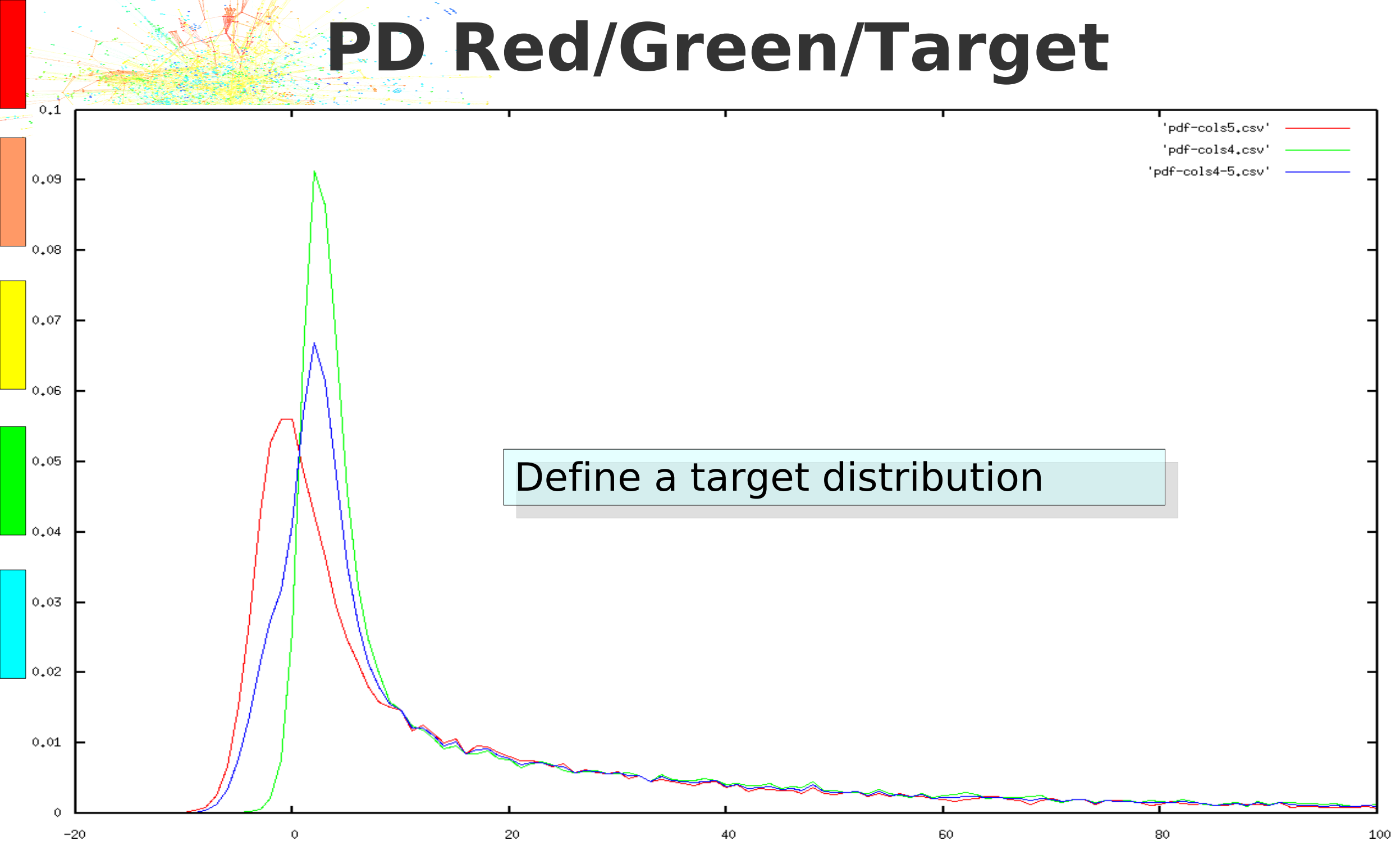
Gaussian PDF



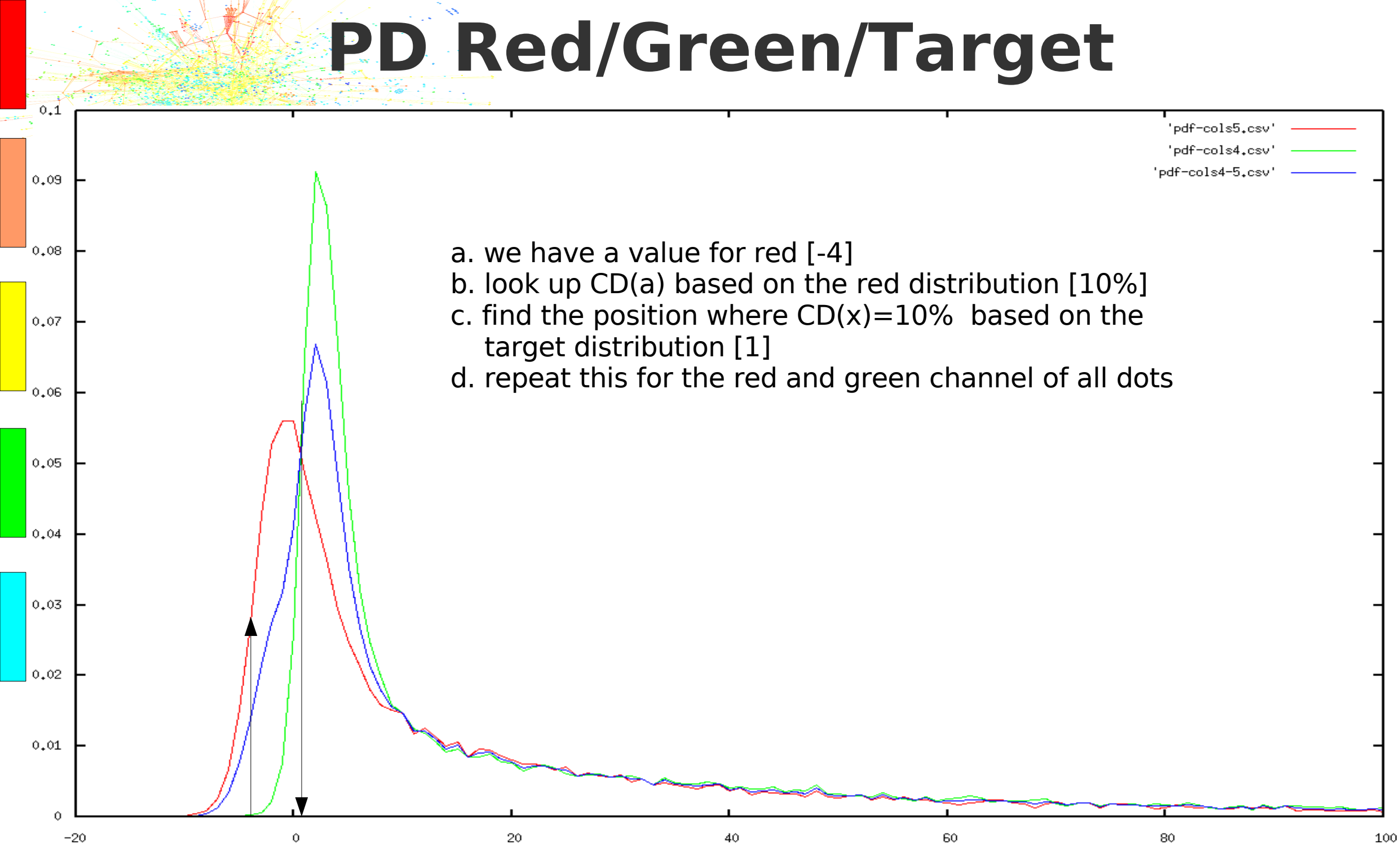
Gaussian CDF



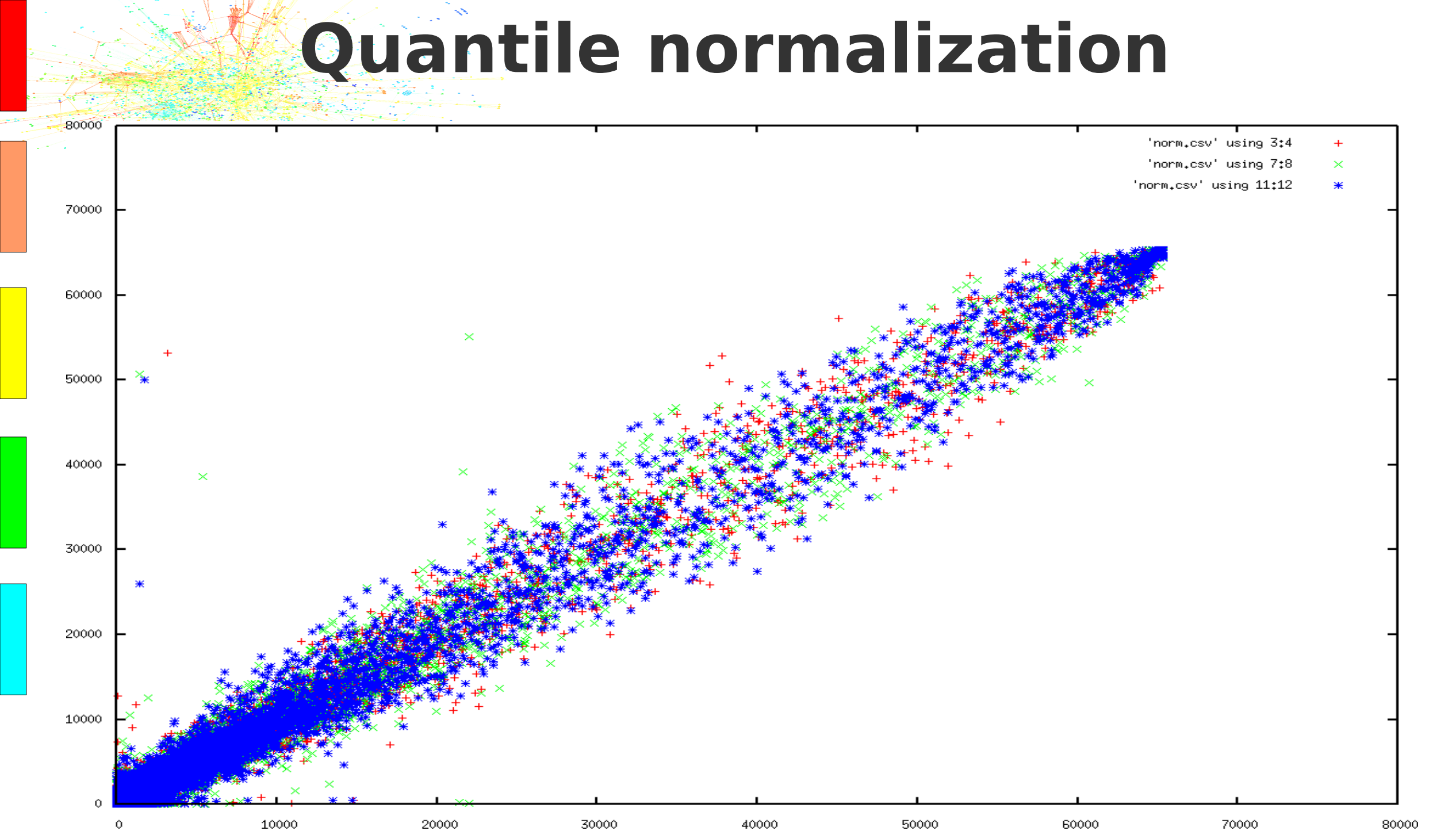
PD Red/Green/Target



PD Red/Green/Target



Quantile normalization





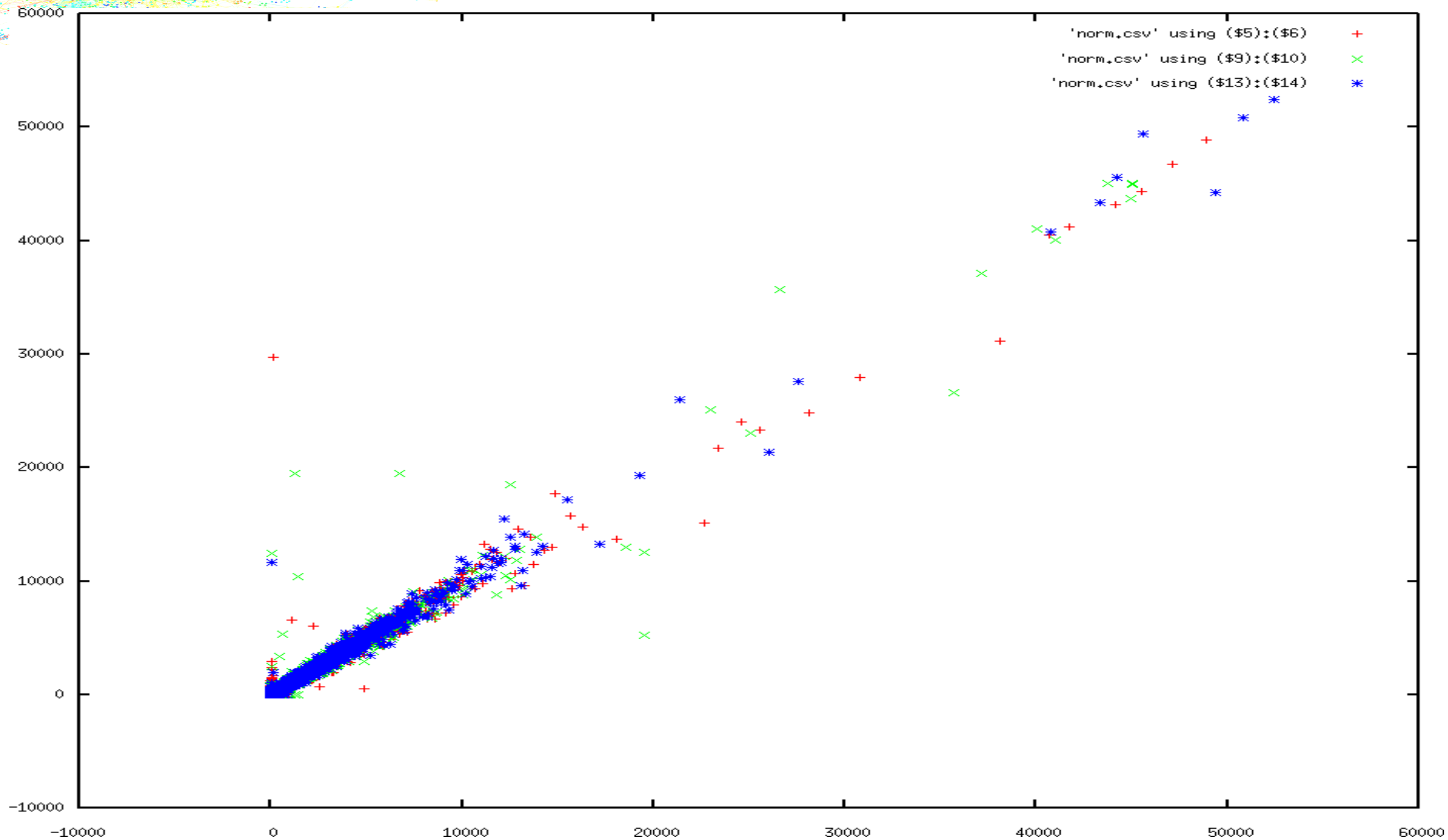
Dynamic Range & Clipping



Dynamic Range & Clipping

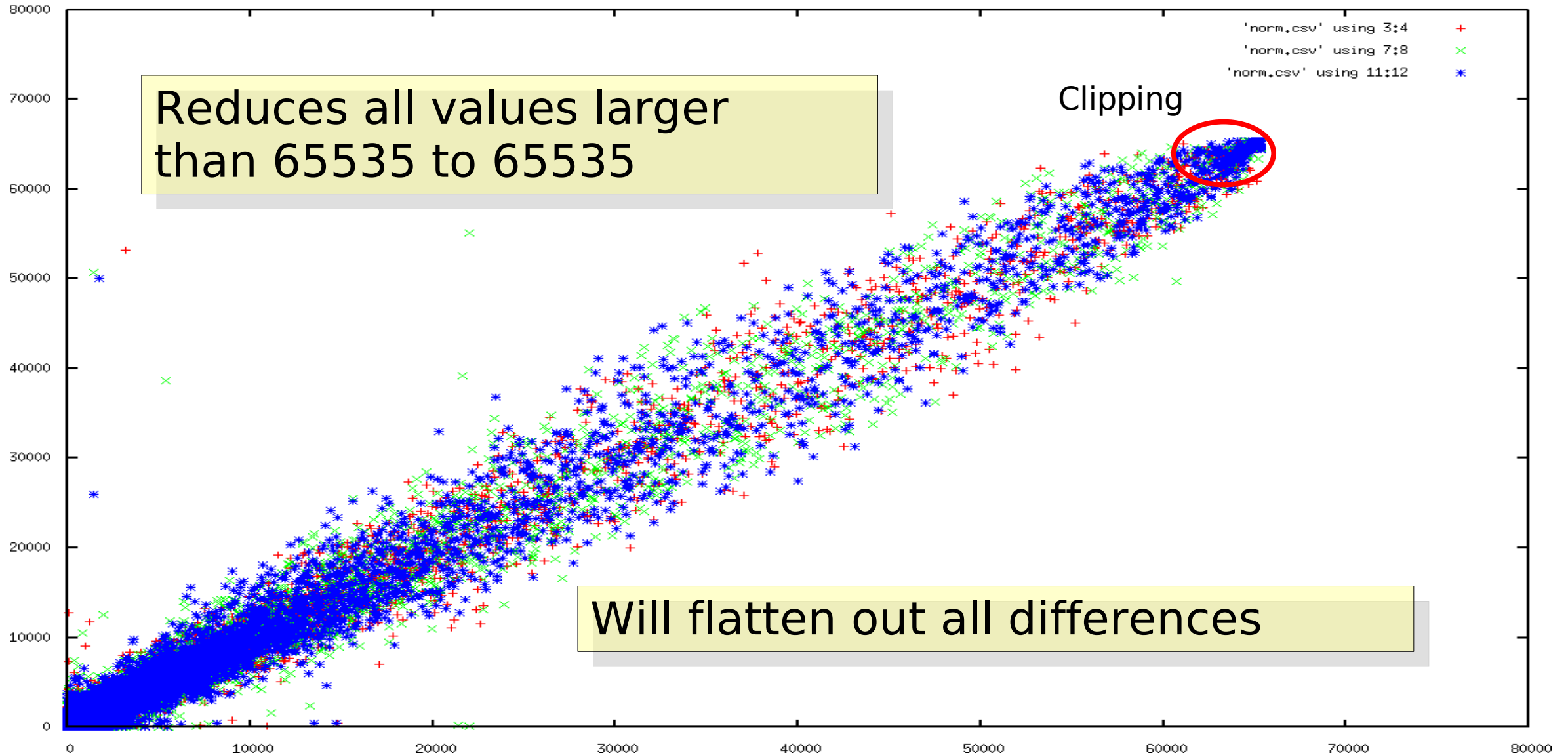
- Low gains
 - lower sensitivity for low intensity spots
 - proper measurement of high intensity spots
- High gains
 - higher sensitivity for low intensity spots
 - too high a sensitivity for higher intense spots
- High intensity spots -> properly measured with low gain
- Low intensity spots -> better measured with high gain

R/G Scatterplot Lo Gain



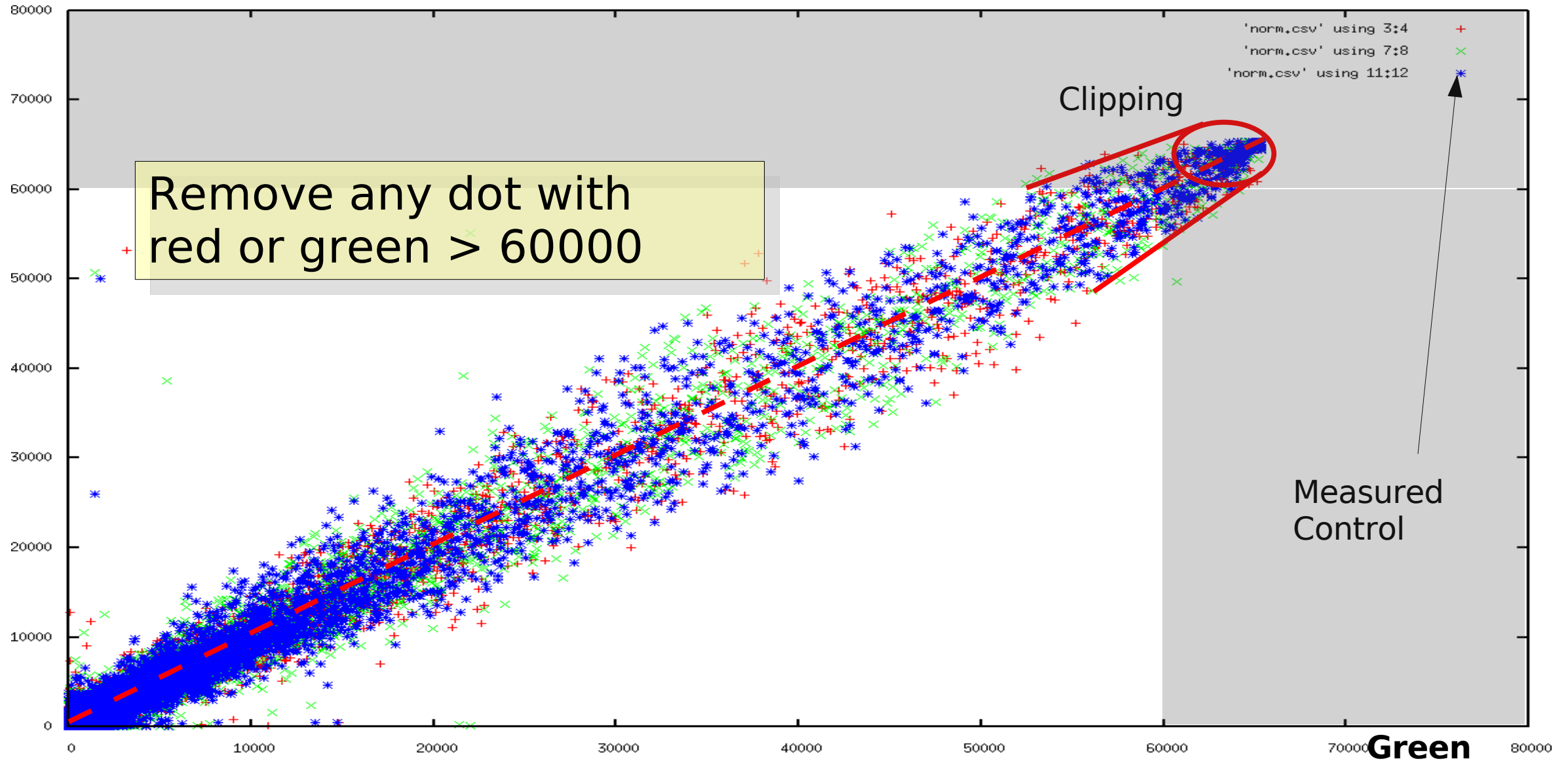
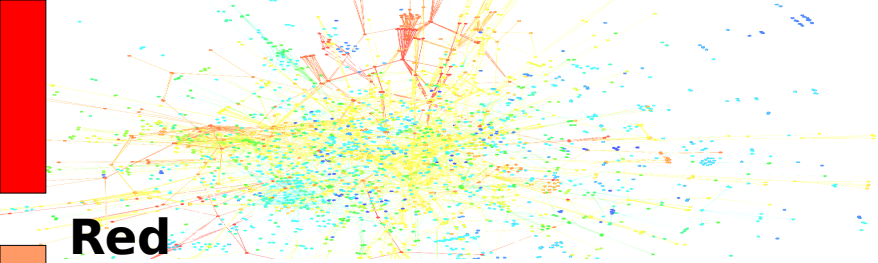
Clipping

Red



Green

Clipping





Absolute and Relative Errors



Absolute Errors

- Every measured value is based on the real value, but with an unknown value added

$$v_m = v_r + \epsilon \qquad |\epsilon| \leq \beta$$

with error 0.1

$10 \in [0.9, 10.1]$

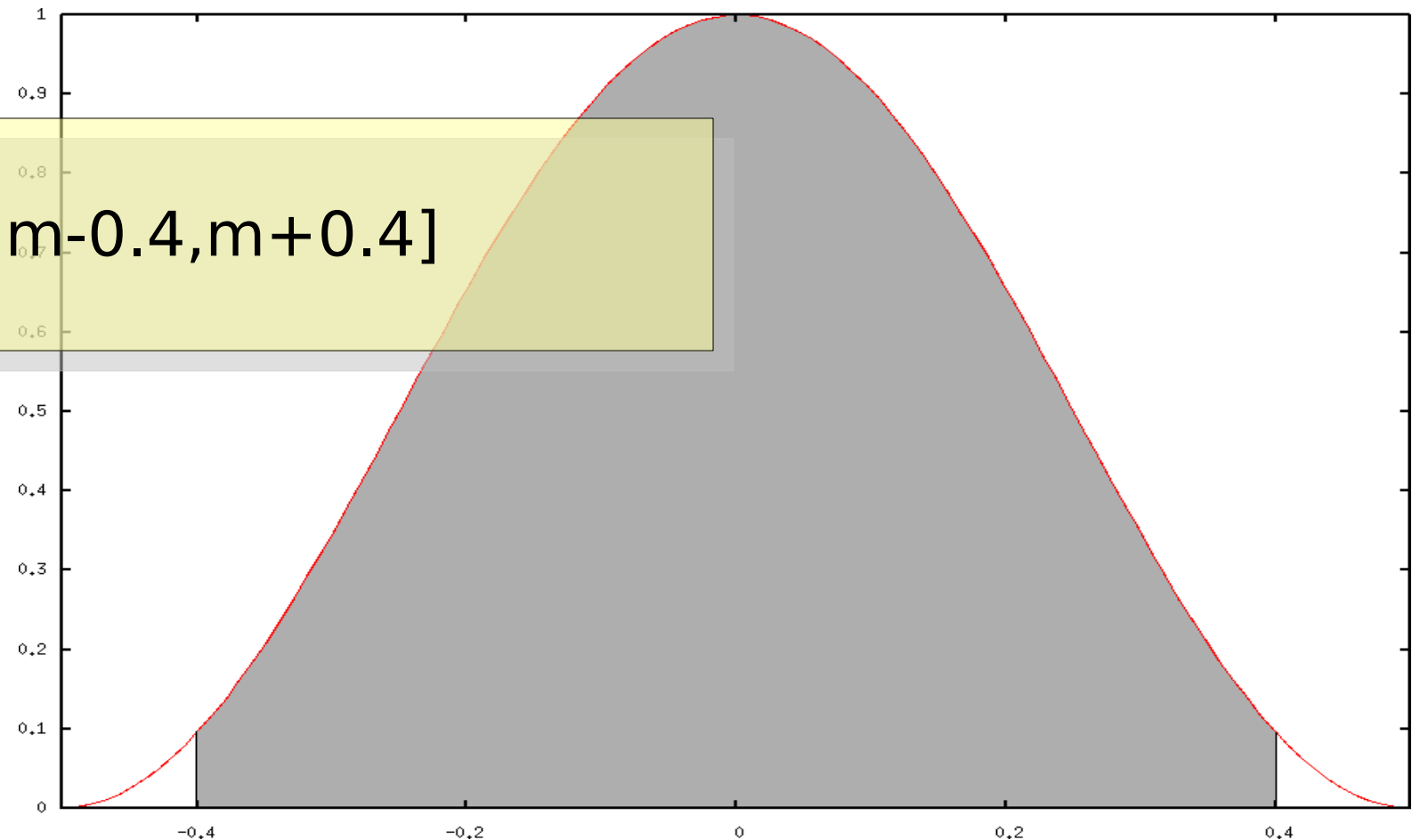
$100 \in [99.9, 100.1]$

$1000 \in [999.9, 1000.1]$

Absolute Error PD

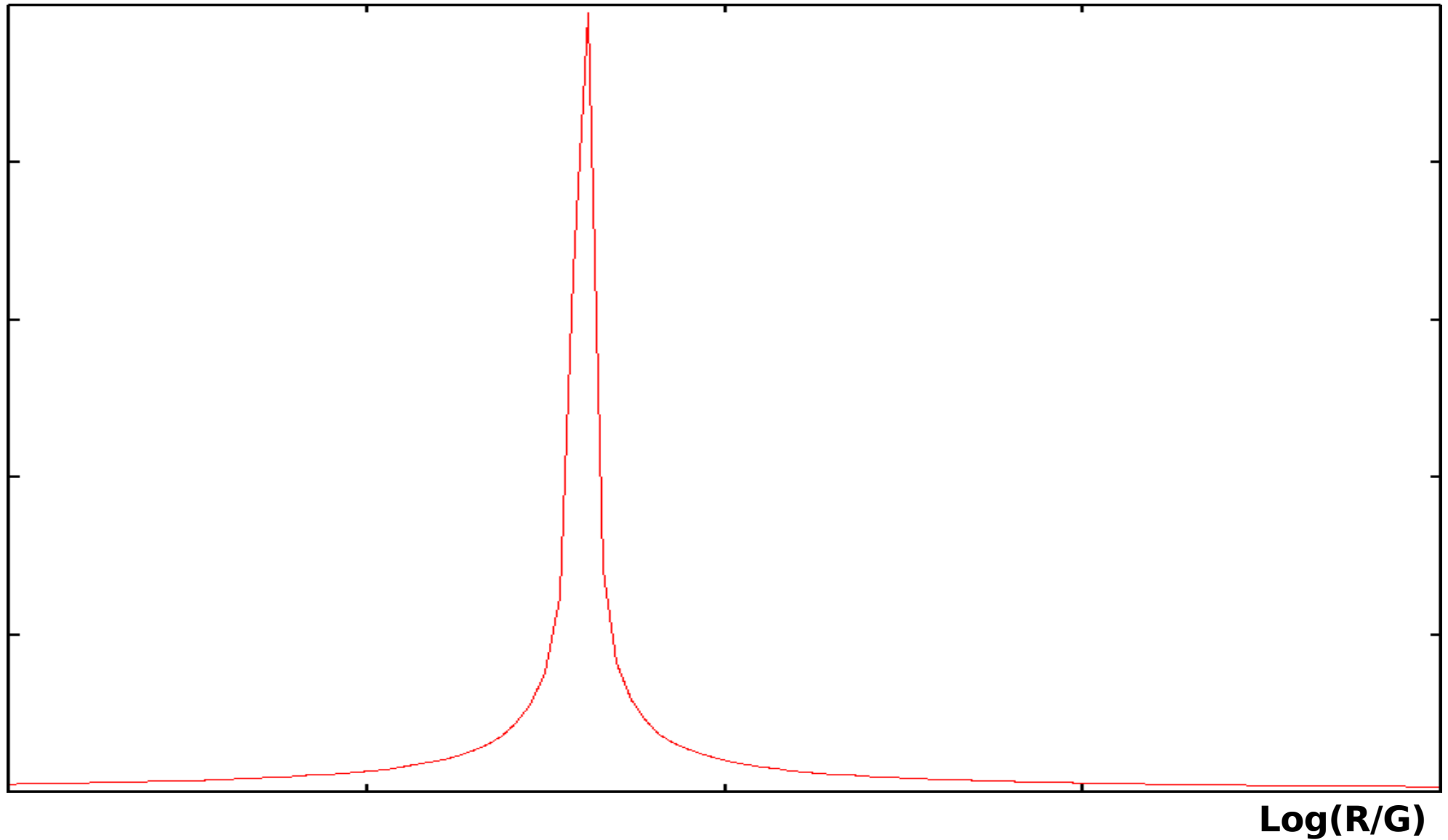
Based on the PD of the error, a confidence interval that will cover at least 95% of the real values can be created

If we measure r ,
then r will be within $[m-0.4, m+0.4]$
in 95% of the cases.



Micro Array Errors

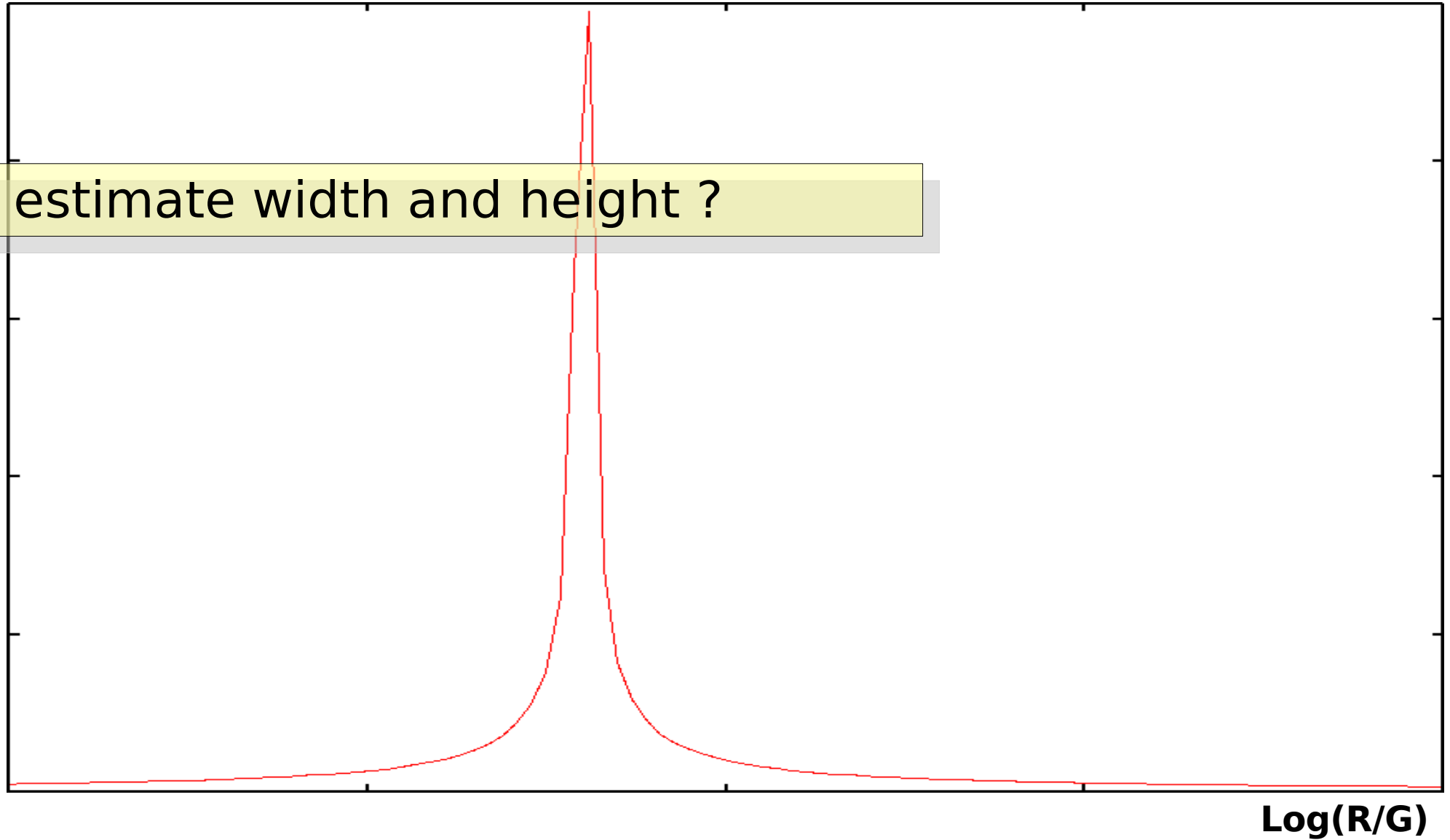
- Lorentz Distribution



Micro Array Errors

- Lorentz Distribution

How to estimate width and height ?



Micro Array Errors

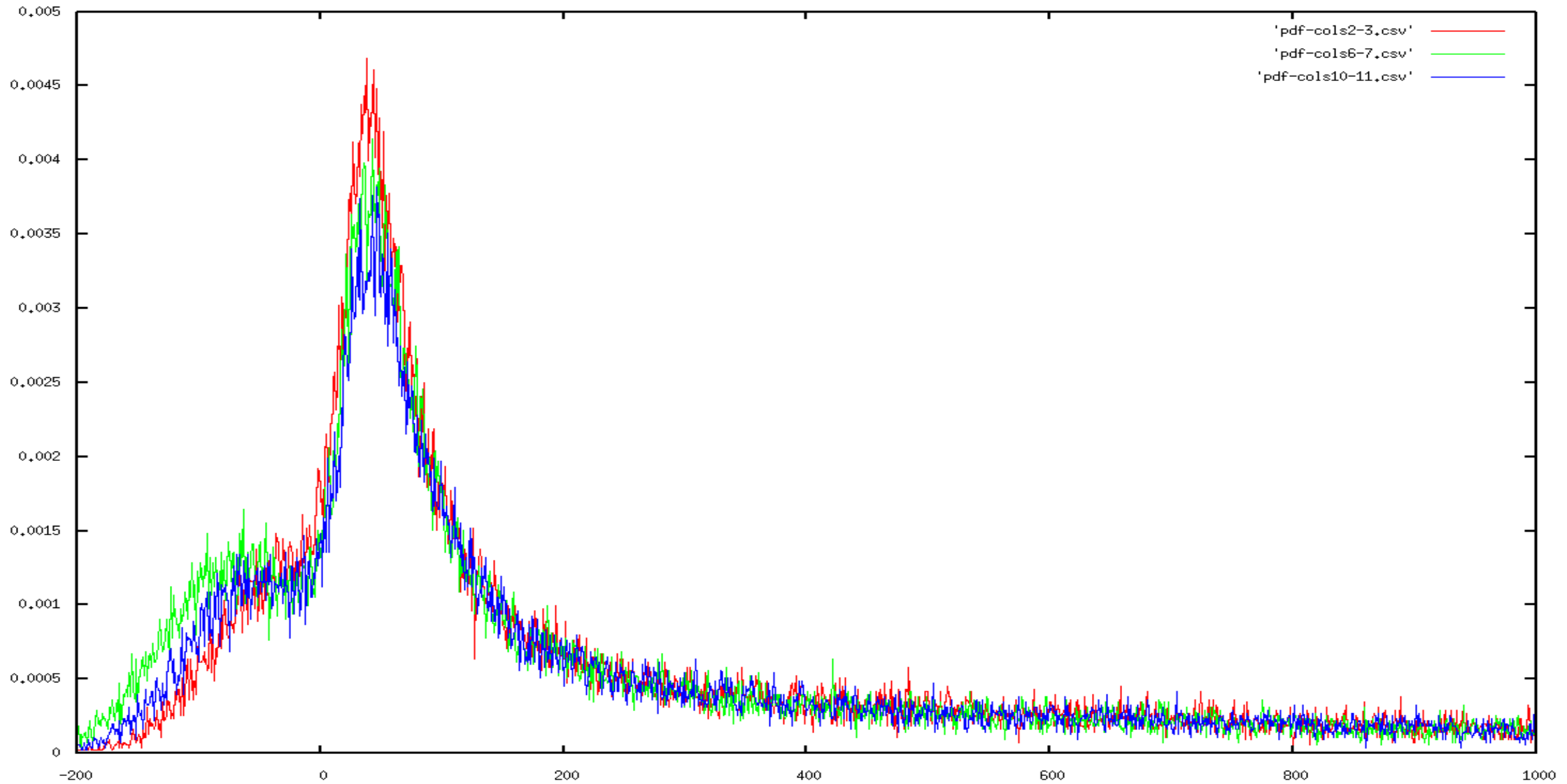
- Lorentz Distribution

How to estimate width and height ?

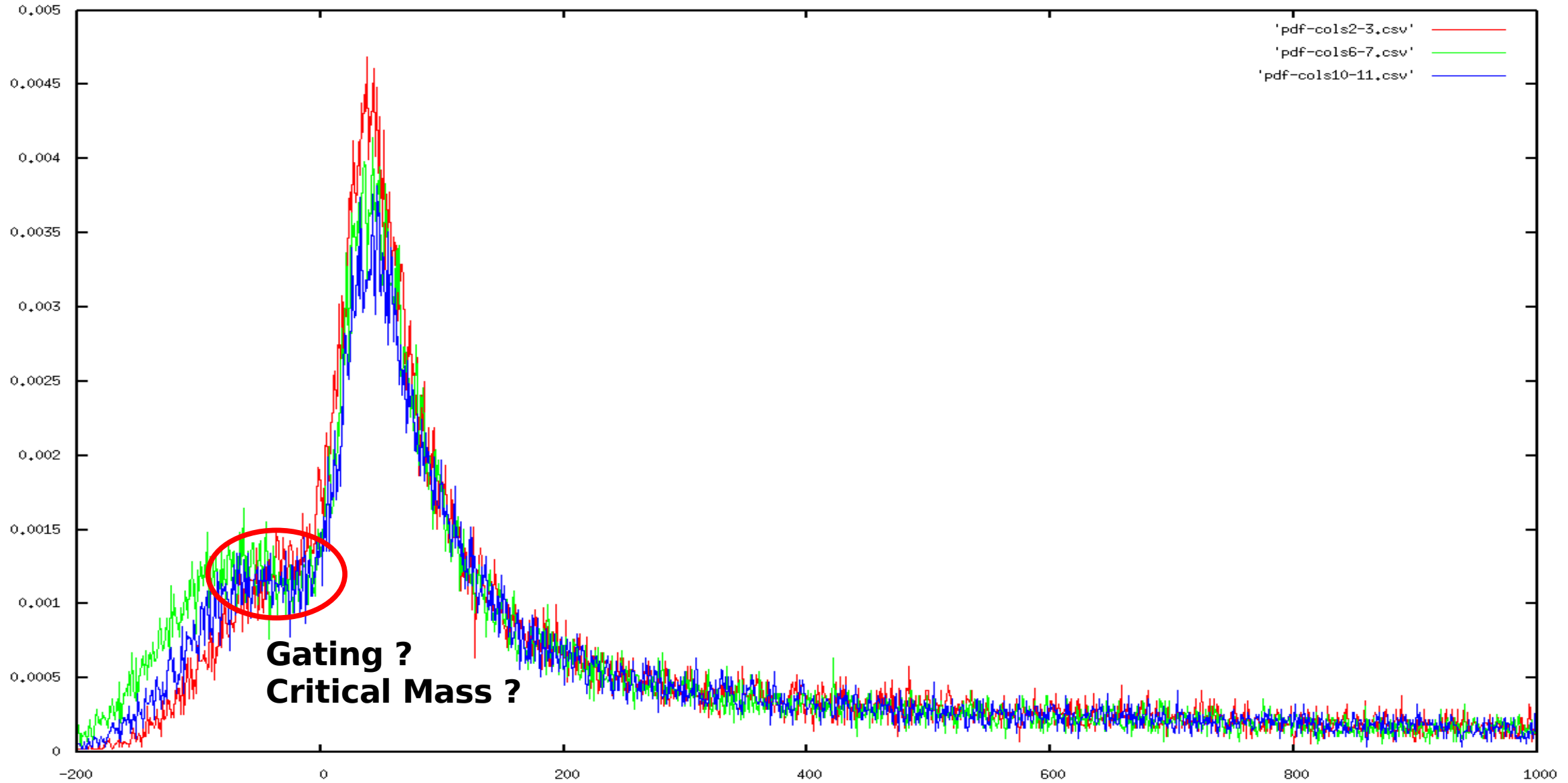
Go Back to the control sample, and measure the distribution

Log(R/G)

PD of High Gain



PD of High Gain







Relative Errors

- Every measured value is based on the real value, but multiplied with an unknown value

$$v_m = v_r \cdot \epsilon \qquad |1 - \epsilon| \leq \alpha$$

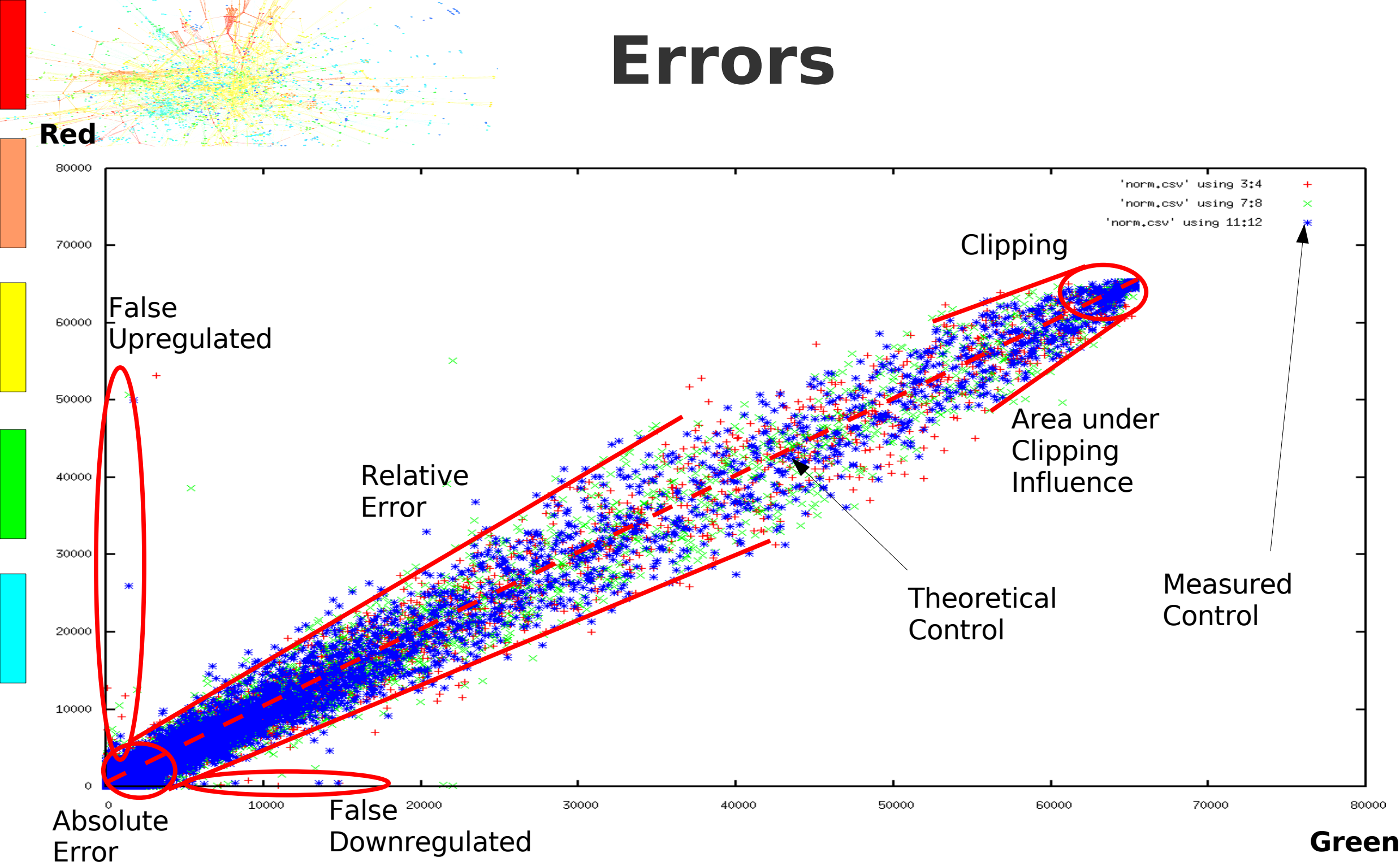
With $r=0.1$

10 > [9,11]

100 > [90,110]

1000 > [900,1100]

Errors





Combined Error

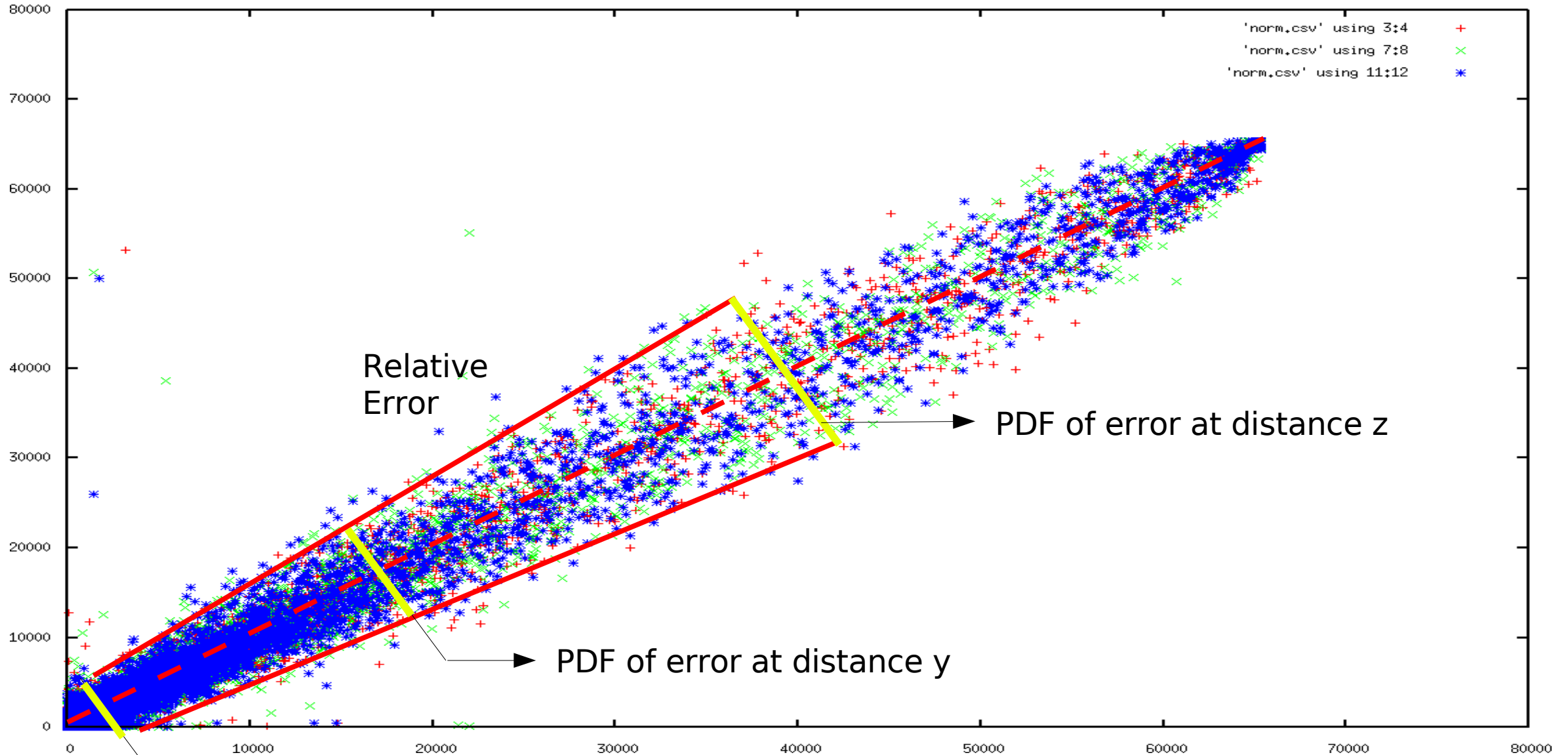
- Relative / Absolute errors

$$v_m = v_r \cdot \epsilon_\alpha + \epsilon_\beta \quad |\epsilon_\beta| \leq \beta \wedge |1 - \epsilon_\alpha| \leq \alpha$$

- Estimating model parameters
 - not straightforward
 - model might not be correct
- Express the error PD in function of the norm

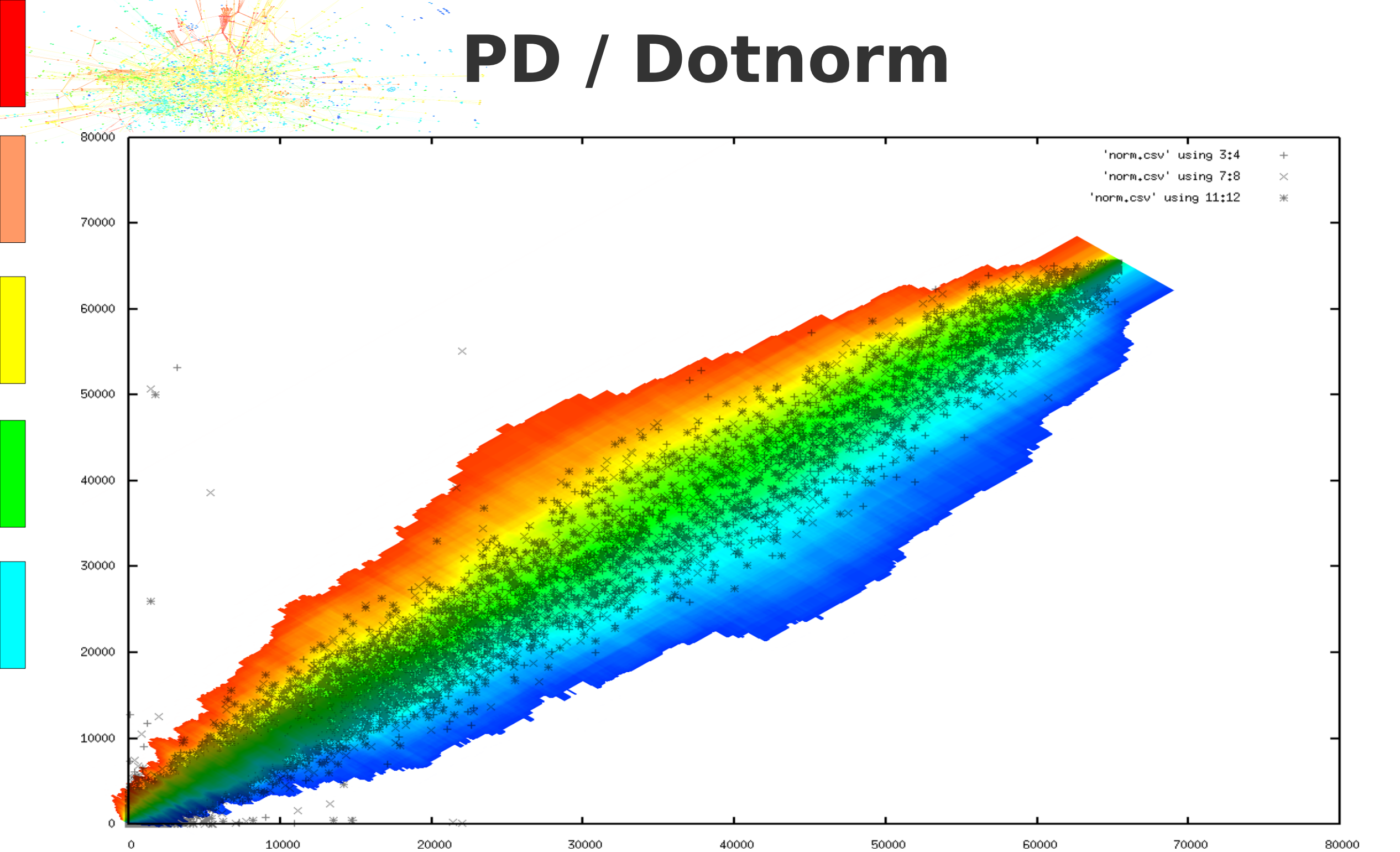
Relative Errors

Green



Red

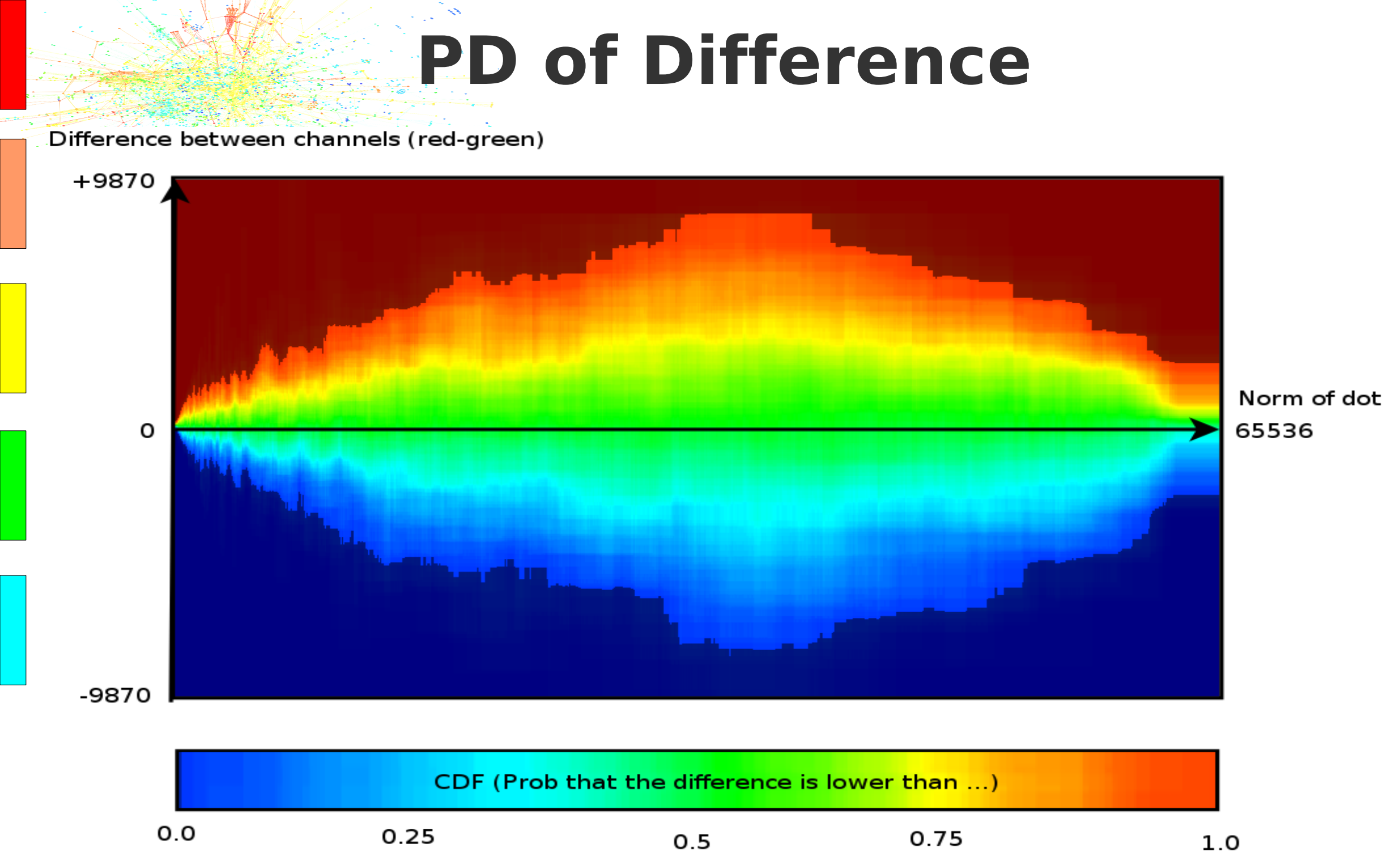
PD / Dotnorm



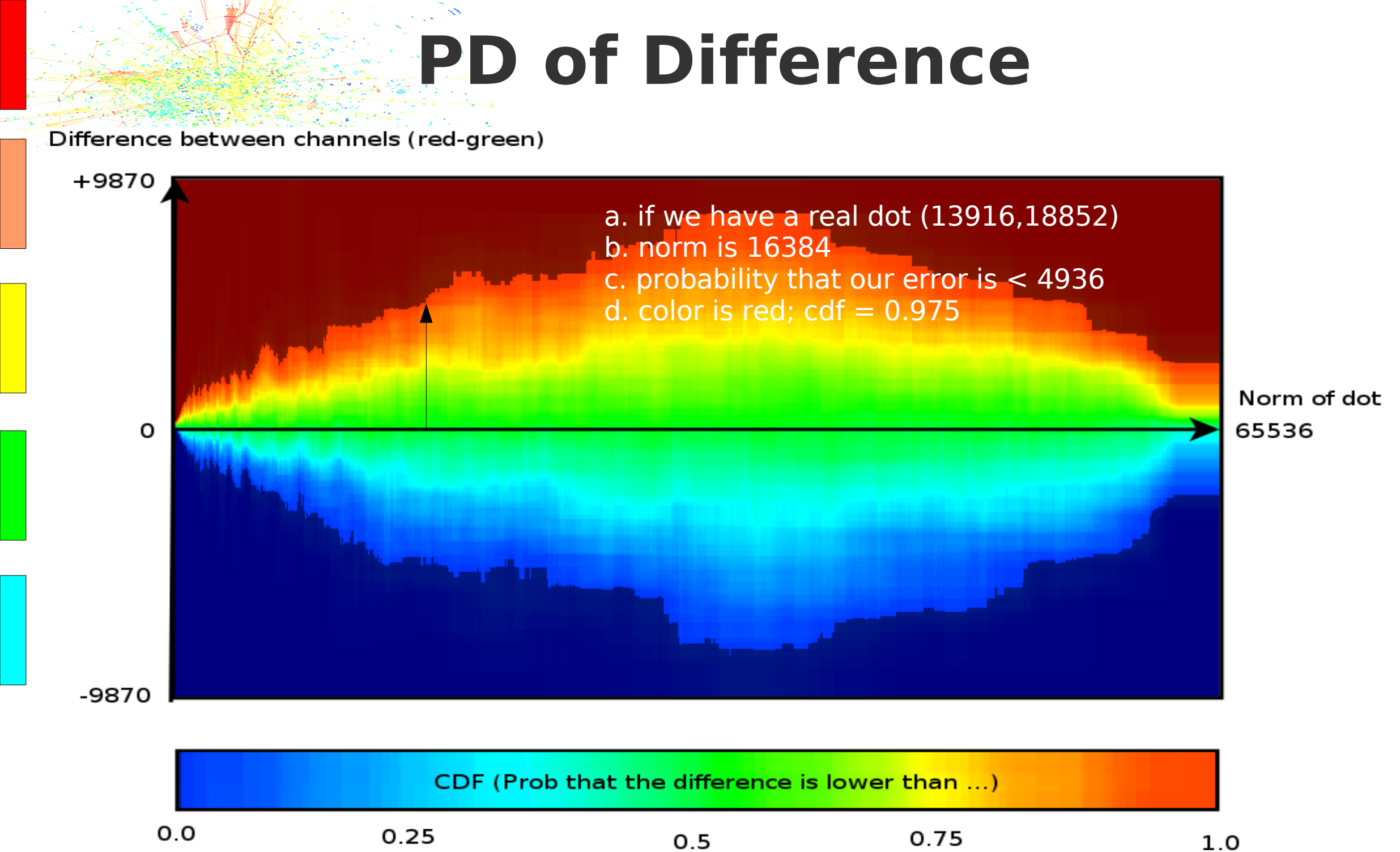
Up/Down Regulation



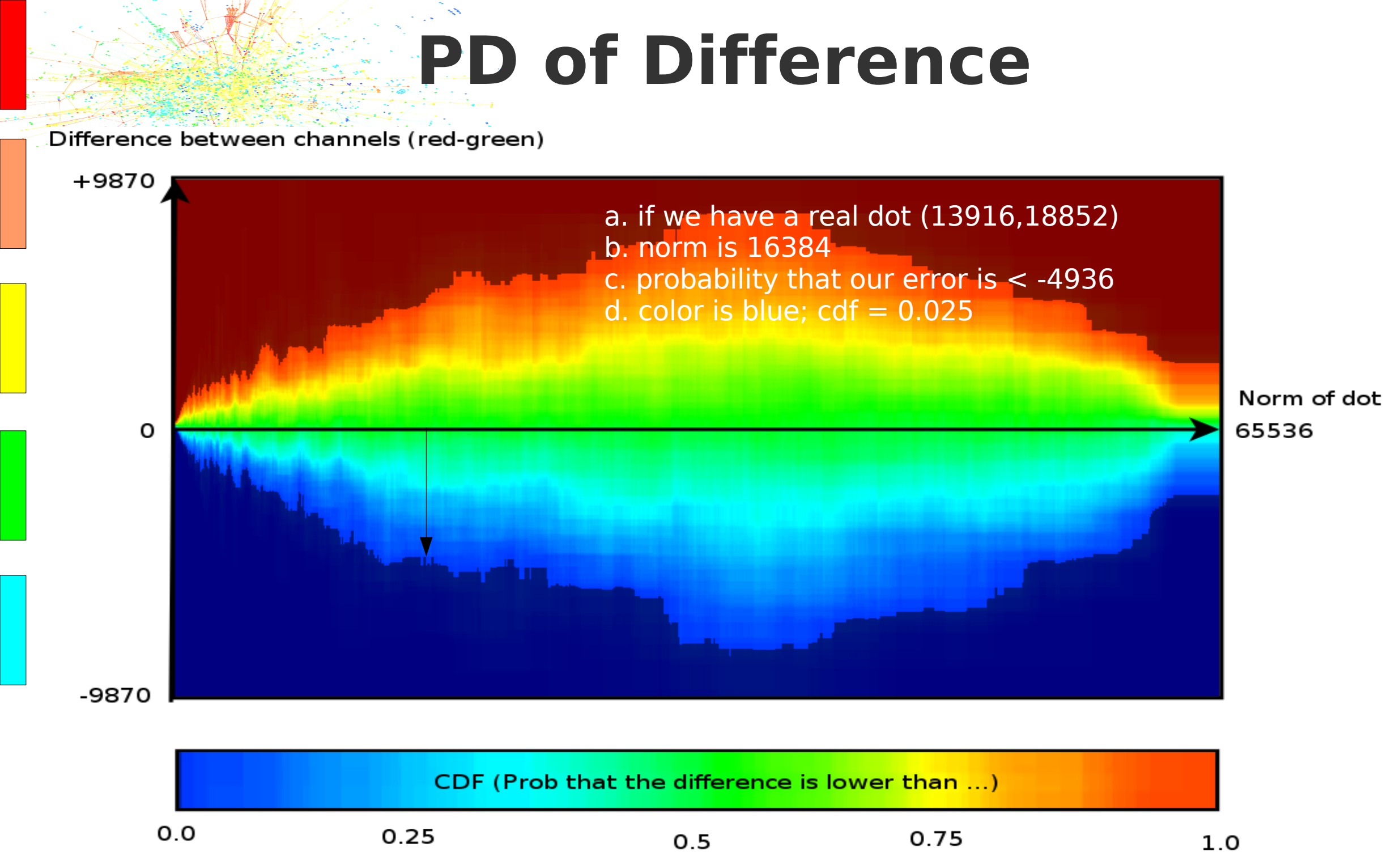
PD of Difference



PD of Difference

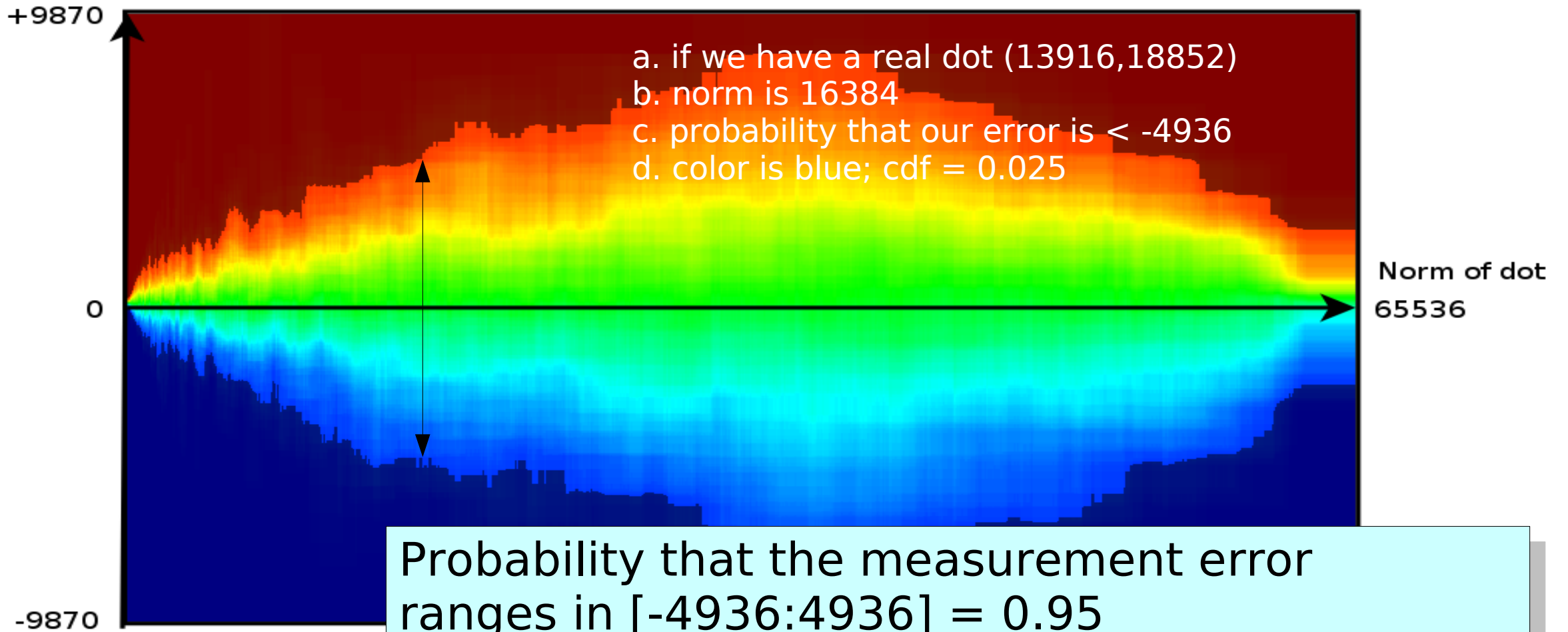


PD of Difference



PD of Difference

Difference between channels (red-green)



CDF (Prob that the difference is lower than ...)

0.0

0.25

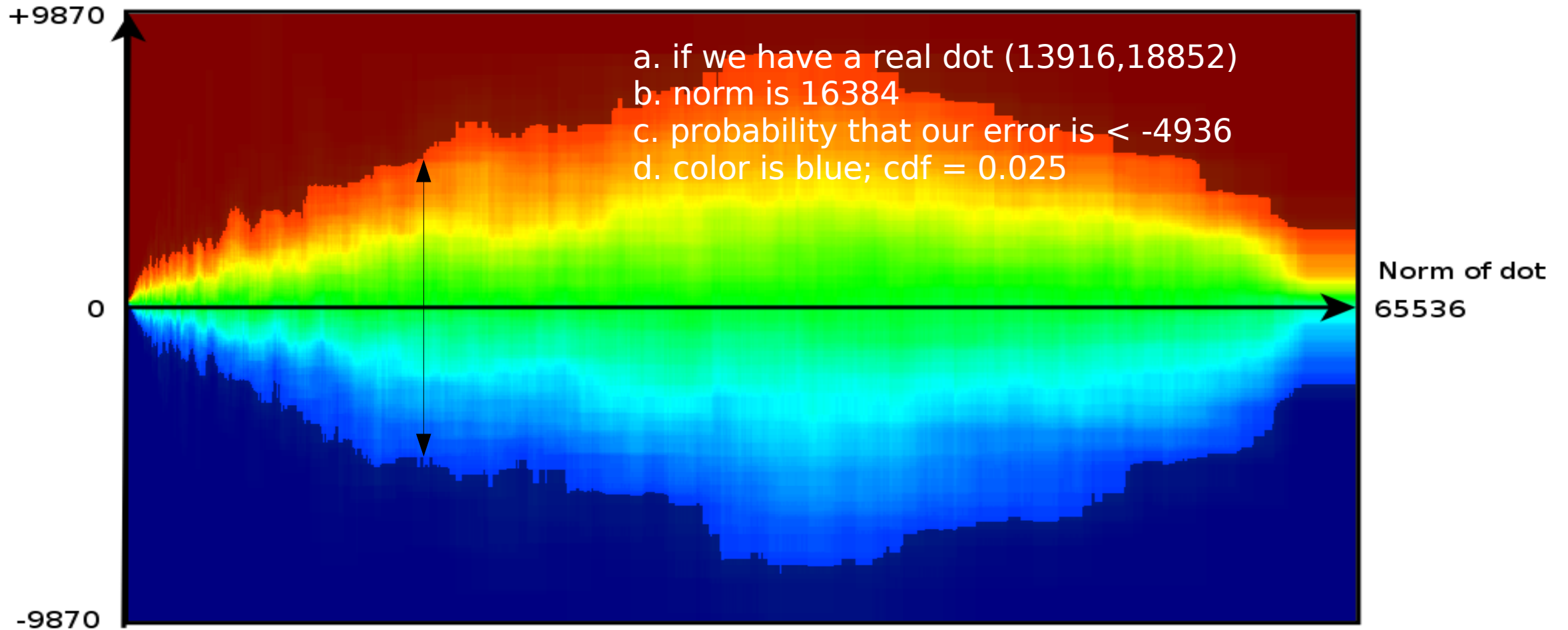
0.5

0.75

1.0

PD of Difference

Difference between channels (red-green)



Measured difference between R and G is 4936
Real difference is $4936 + [-4936:4936] = [0:4936]$

PD of Difference

- Confidence interval expresses the possible real values based on the measurement.

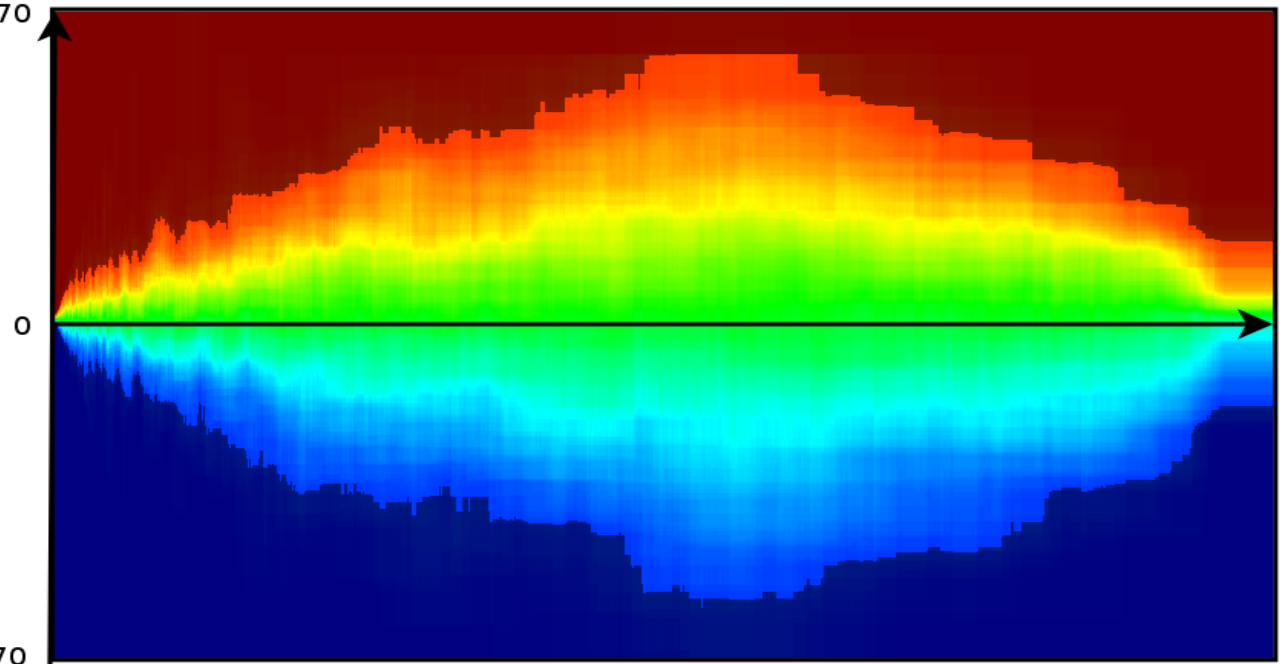
Dotnorm

95% C.I.

245	-399	399
5652.48	-2877	2877
38917	-6912	6912
39649	-8488.5	8488.5

Difference between channels (red-green)

+9870



Norm of dot
65536

CDF (Prob that the difference is lower than ...)

0.0

0.25

0.5

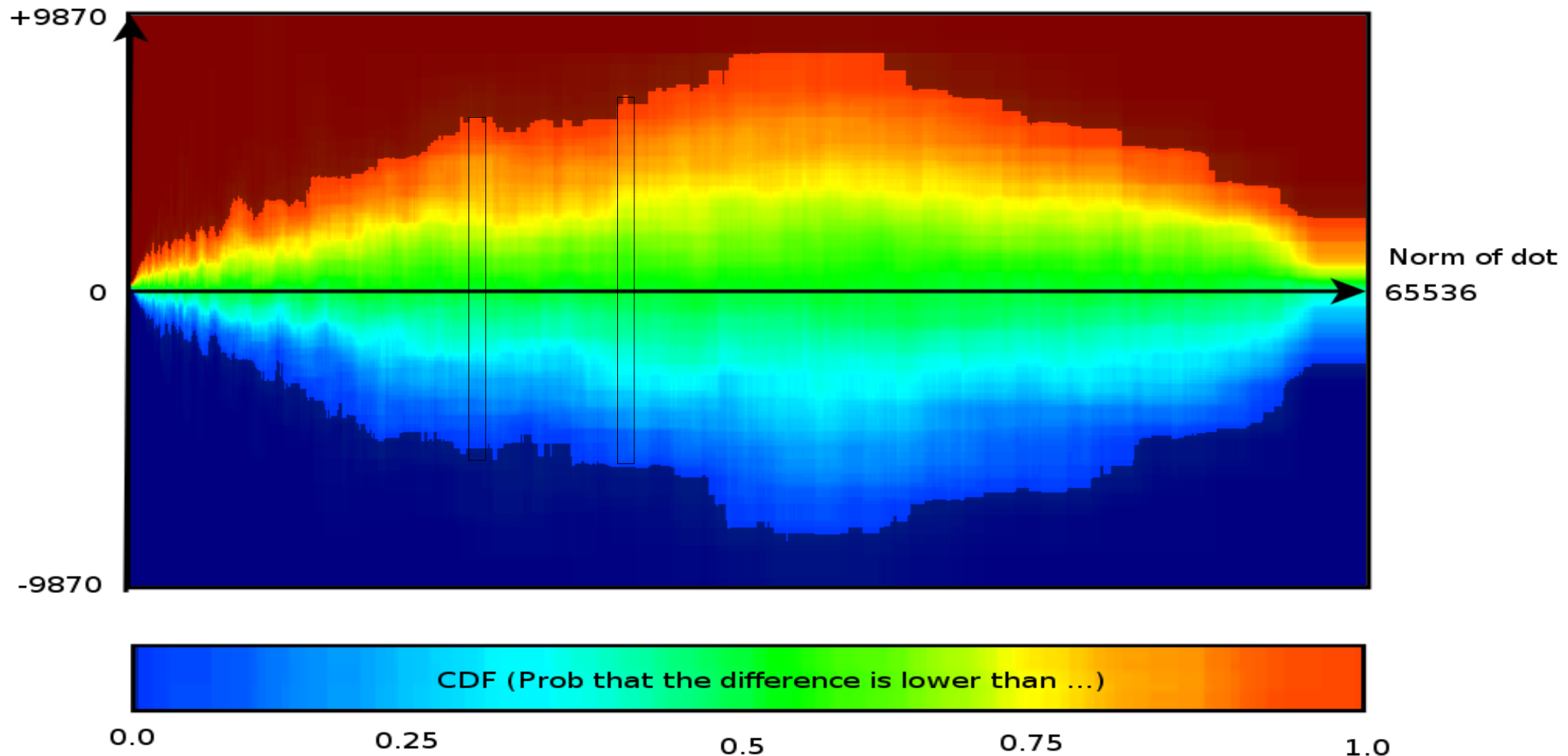
0.75

1.0

PD of Difference

- Multiple measurements lead to better estimate and smaller confidence interval

Difference between channels (red-green)

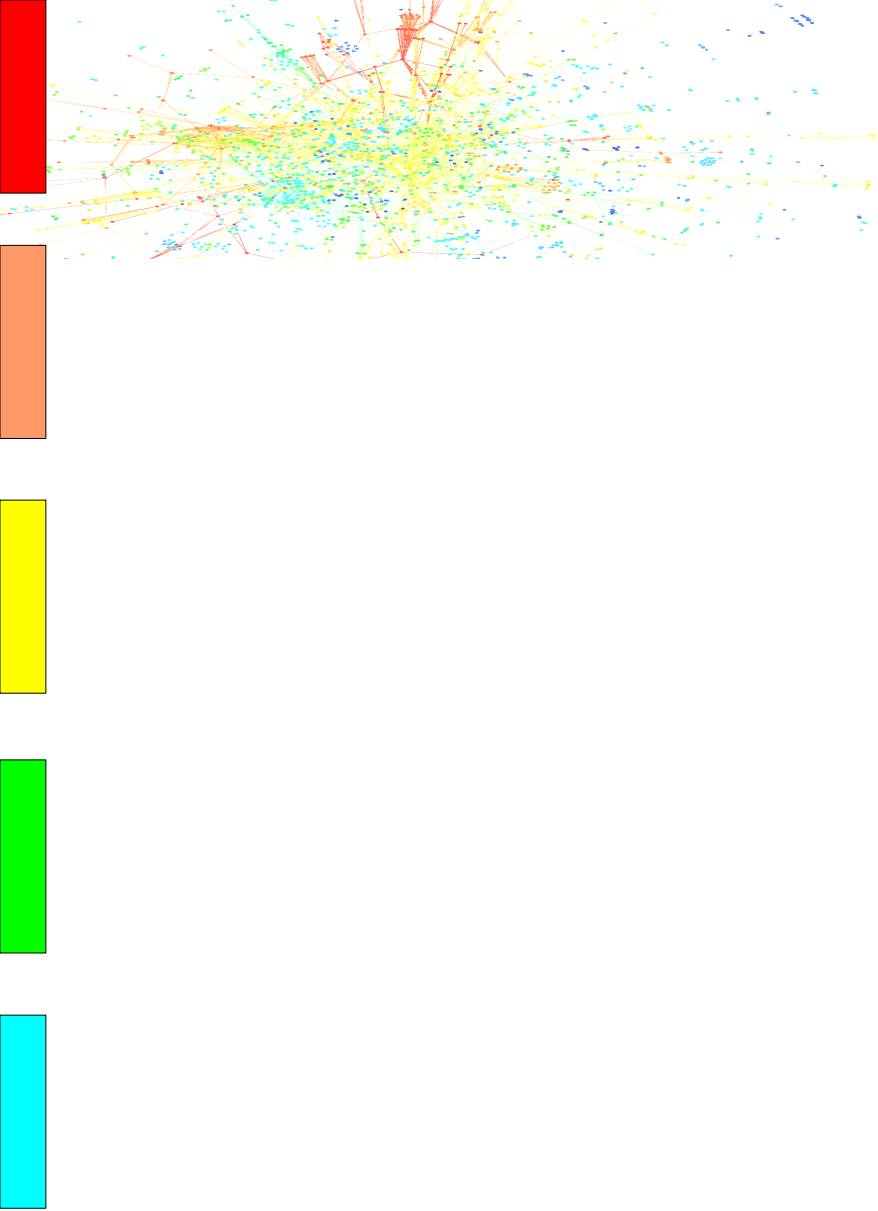


Reported Regulations

ID	C.I.	Values			Difference			Regu	Factor
		Red	Green	#	Lo	Normal	Hi		
R000008_01	[-1515.52:1812.48]	2502	81	1	-905.48	-2421	-4233.48	down	2.08
R000068_01	[-4894.72:5621.76]	18661	12833	2	-933.28	-5828	-11449.8	down	1.06
R000088_01	[-3194.88:2938.88]	7637	2963	2	-1479.12	-4674	-7612.88	down	1.32
R000137_01	[-307.2:307.2]	39	-294	2	-25.8	-333	-640.2	up	2.32
R000141_01	[-8273.92:7905.28]	24161	34097	2	18209.9	9936	2030.72	up	1.07
R000177_01	[-993.28:972.8]	1154	72	1	-88.72	-1082	-2054.8	down	1.16
R000186_01	[-204.8:215.04]	86	-155	1	-36.2	-241	-456.04	up	1.36
R000248_01	[-204.8:215.04]	115	-125	1	-35.2	-240	-455.04	up	1.04
R000293_01	[-3573.76:3737.6]	10913	7012	2	-327.24	-3901	-7638.6	down	1.04
R000310_01	[-3328:3665.92]	12767	2304	2	-7135	-10463	-14128.9	down	2.8
R000490_01	[-665.6:655.36]	826	149	2	-11.4	-677	-1332.36	down	1.02
R000504_01	[-10506.2:10537]	57663	69019	2	21862.2	11356	819.04	up	1.01
R000668_01	[-307.2:317.44]	340	-157	2	-189.8	-497	-814.44	down	1.58
R000711_01	[-665.6:655.36]	866	-206	2	-406.4	-1072	-1727.36	down	4.21

...

Validation



Validation

- The Lo[w]ess set
 - 27648 containing 17411 NA&NA and 3359 NA|NA dots
=> 6878 remaining
 - 4007 pairs in agreement
- The Quantile Normalized C.I. set
 - 1422 dots reported
- Overlap
 - 311 dots
 - 1111 unique Q.C.I dots
 - 3696 unique lo[w]ess dots



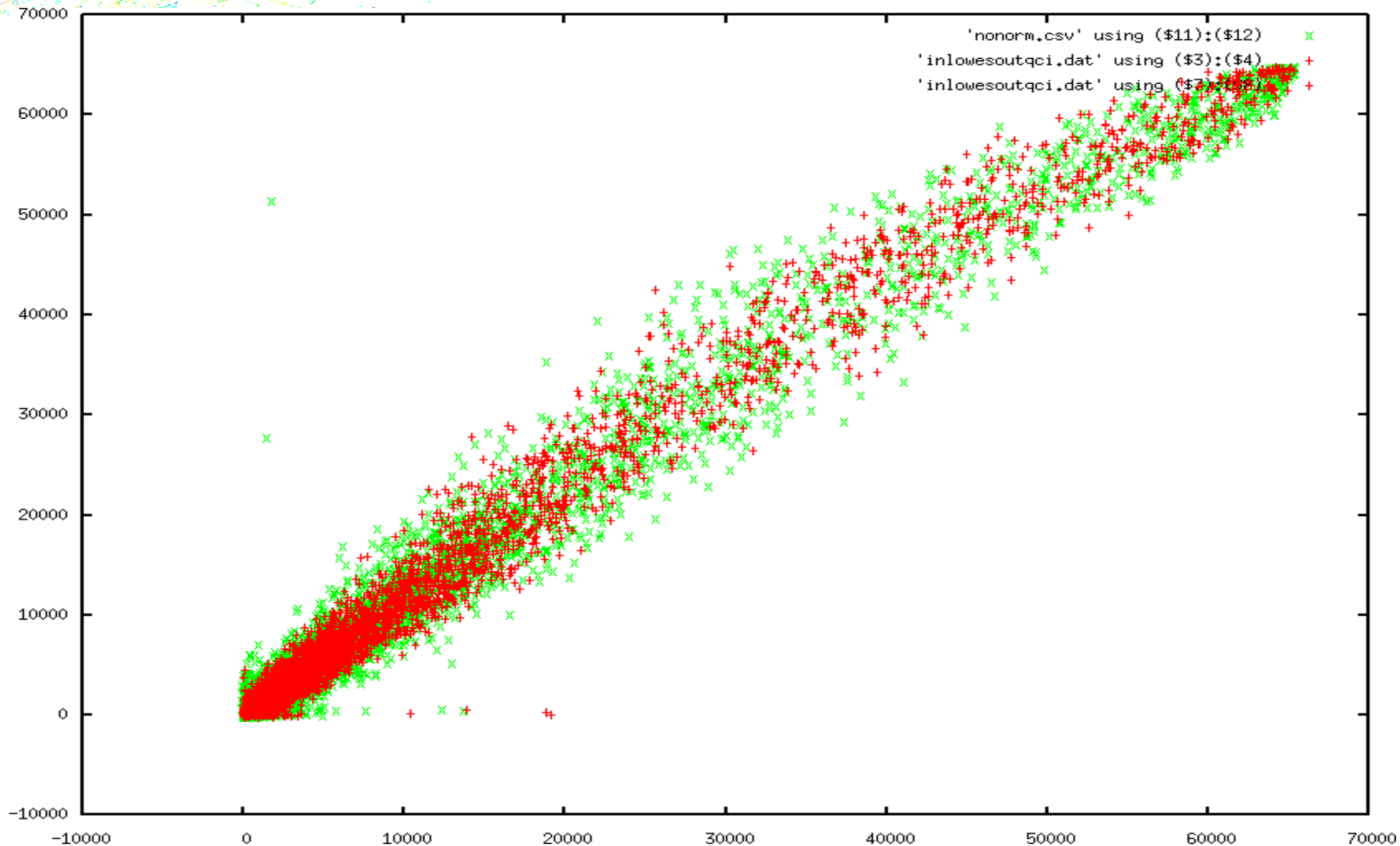
Questions

- Of those in the Lo[w]ess set but not in the Q.C.I. set

Q: Do we have a good argument why we did not take them into account ?

A: They are too close to the expected measurement error

Too close to error





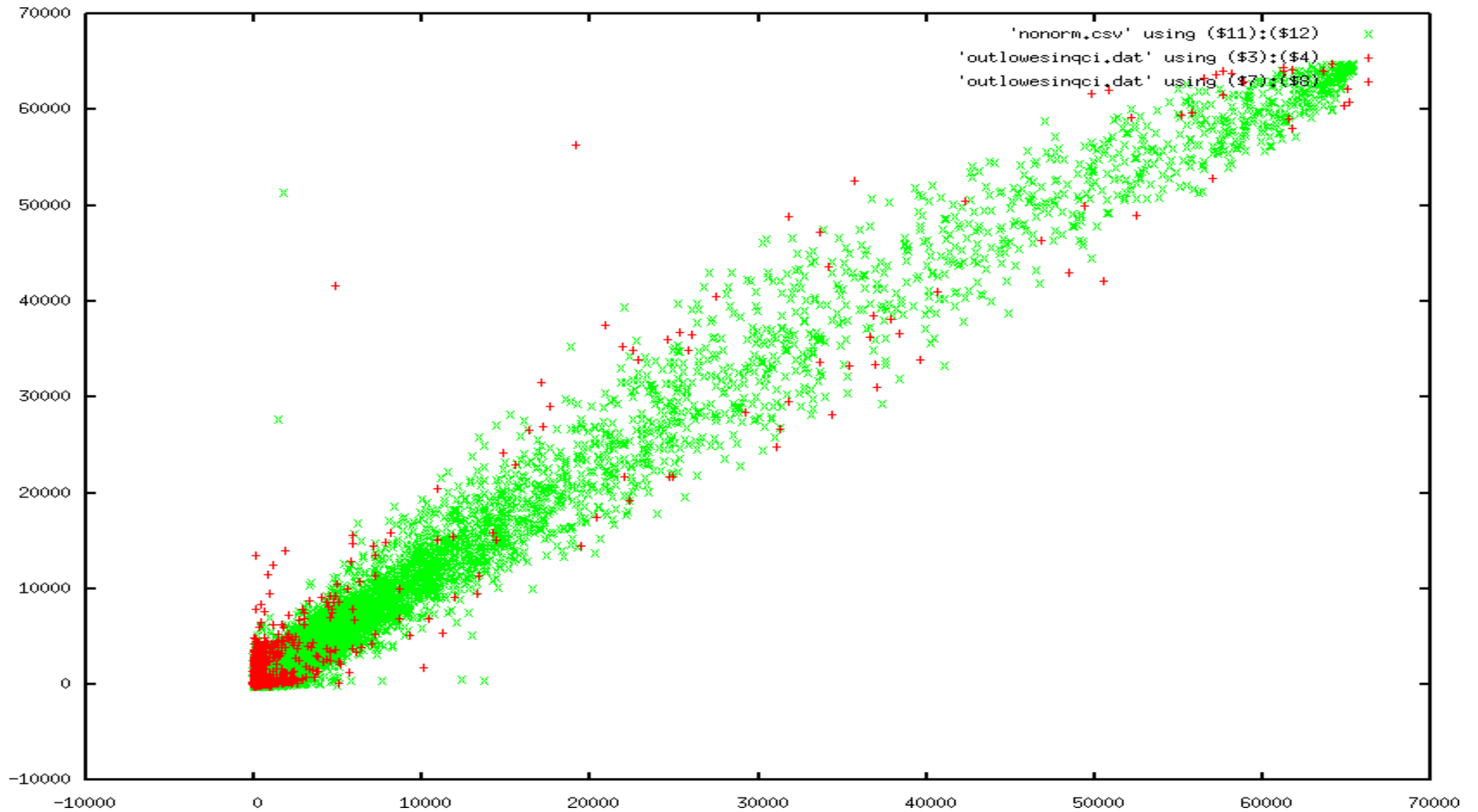
Questions

- Of those in the Q.C.I. set but not in the Lo[w]ess set ?

Q: Do we have a good argument why we did take them into account ?

A: The Q.C.I. Method takes into account multiple dots, which is information unavailable to the Lo[w]ess method.

Outside 95% C.I.





Questions

- Of those dots that are both in the Lowes and Q.C.I. set:

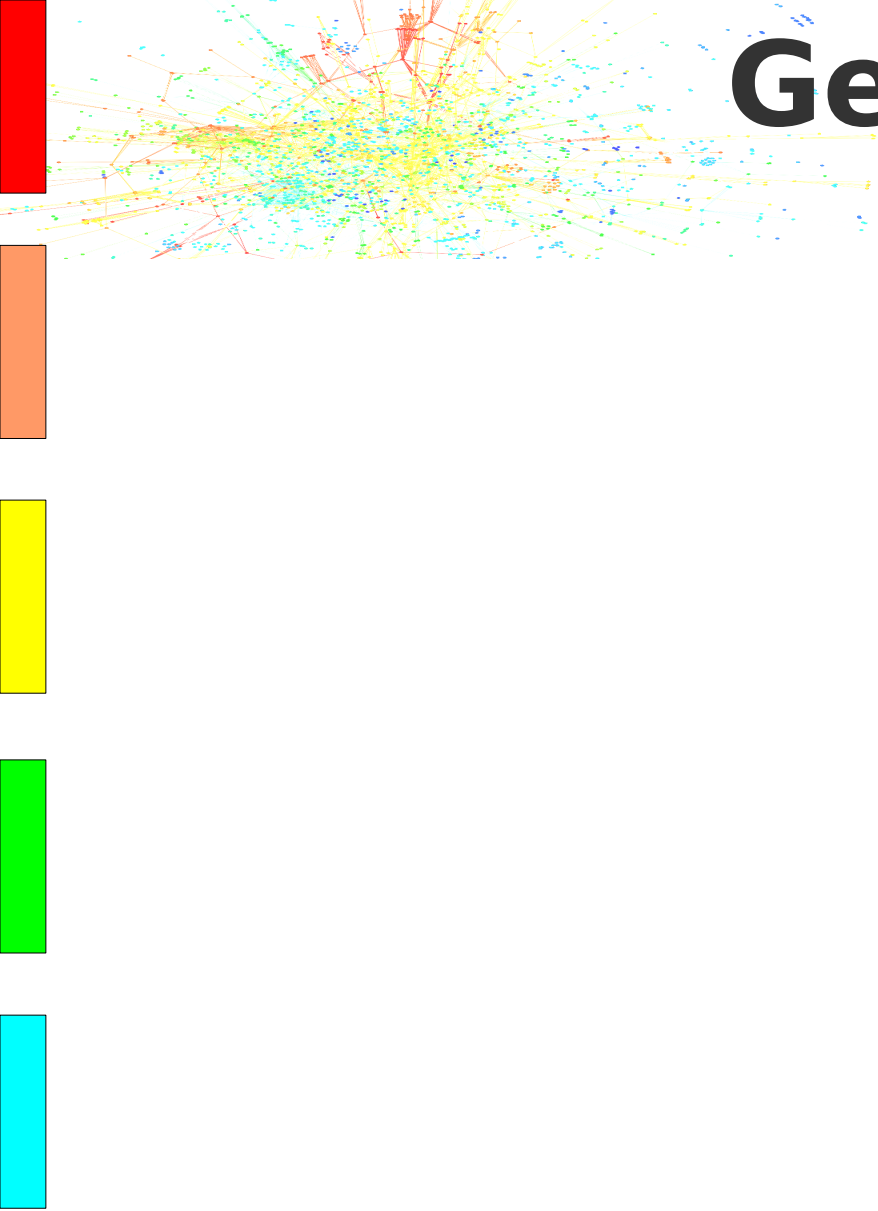
Q: Do they report the same qualitative regulation ?

A: 301 do, 10 don't => 3% mismatch

Differences

ID	C.I.				Difference		Values		Regulation	Factor	
	Low	Norm	Hi	Green	Red	Count	Lo	Hi			
D4D1	D6D3										
Green	Red	Green	Red								
Rn30026543	[-3983.36:3993.6]	-277.64	-4261	-8254.6	10661	6400	2 down	1.03	2.87	Q.C.I	
					0.49	0.55	up			Lowes Log	
	4336	2218	7262	3661						Non normalized	
Rn30009746	[-1904.64:1812.48]	-41.36	-1946	-3758.48	2743	797	2 down	1.02	33.41	Q.C.I	
					0.12	0.02	up			Lowes Log	
	911	683	2001	113						Non normalized	
Rn30025831	[-2918.4:3246.08]	-545.6	-3464	-6710.08	8138	4674	2 down	1.09	3.2	Q.C.I	
					0.21	0.41	up			Lowes Log	
	2274	1383	6508	2910						Non normalized	
Rn30026511	[-8212.48:8407.04]	-1460.52	-9673	-18080	41854	32181	2 down	1.04	1.65	Q.C.I	
					0.43	0.06	up			Lowes Log	
	10631	8385	34489	20727						Non normalized	
Rn30023124	[-5539.84:5683.2]	11256.8	5717	33.8	14262	19979	2 up	1	1.98	Q.C.I	
					-0.13	-0.11	down			Lowes Log	
	7556	8168	8065	10364						Non normalized	
Rn30026938	[-2959.36:2826.24]	-580.64	-3540	-6366.24	5297	1757	2 down	1.18	19.51	Q.C.I	
					0.02	0.02	up			Lowes Log	
	2104	827	3590	880						Non normalized	
Rn30026618	[-7618.56:8785.92]	17364.6	9746	960.08	109415	119161	2 up	1.01	1.16	Q.C.I	
					-0.01	-0.13	down			Lowes Log	
	53944	57496	57493	60704						Non normalized	
Rn30026891	[-2938.88:3481.6]	-444.12	-3383	-6864.6	7455	4072	2 down	1.08	3.94	Q.C.I	
					0.01	0.1	up			Lowes Log	
	2347	1004	5737	2757						Non normalized	
Rn30000378	[-6338.56:7075.84]	13860.6	7522	446.16	114279	121801	2 up	1	1.12	Q.C.I	
					-0.26	-0.12	down			Lowes Log	
	63294	64294	53346	56126						Non normalized	
Rn30018614	[-1904.64:1853.44]	3883.64	1979	125.56	851	2830	2 up	1.07	37.33	Q.C.I	
					-0.1	0	down			Lowes Log	
	528	958	311	1711						Non normalized	

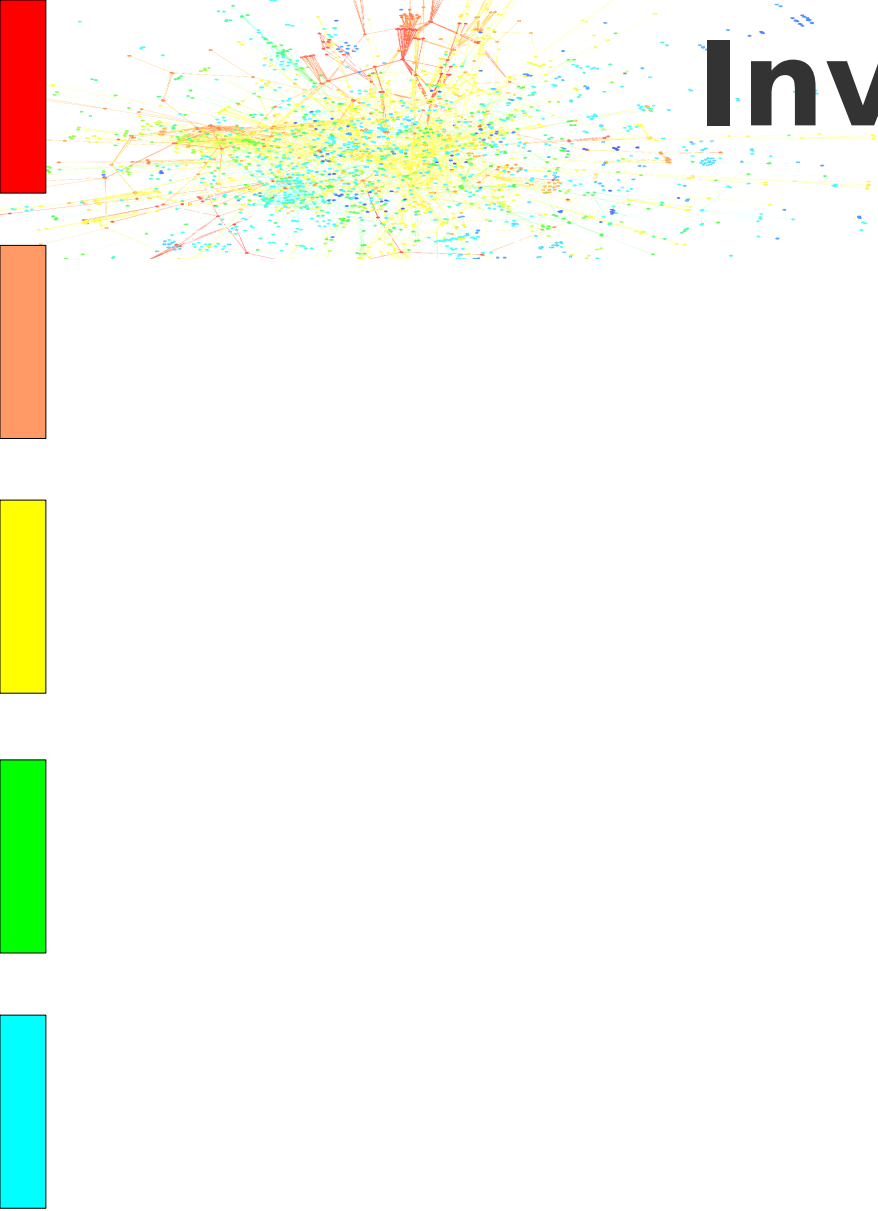
Gene Expression



Gene Expression

Regulation Factor				GO		Cell Location	Description RefSeq
At least	At most	Biological Process	Molecular Function				
up 3.86	4.87	inf	-	metal ion binding	-	zinc finger, FYVE domain containing 26 (predicted)	
down 3.21	4.62	inf	-	-	integral to membrane	CD37 antigen	
down 2.62	4.69	inf	metabolism	N-acyltransferase activity	-	N-Acetyltransferase-2 (arylamine N-acetyltransferase)	
up 2.45	3.9	inf	cell communication	-	integral to membrane	kringle containing transmembrane protein	
up 2.34	4.04	inf	macromolecule metabolism	catalytic activity	-	protein tyrosine phosphatase, receptor type, E	
down 2.14	2.45	3.49	protein metabolism	structural molecule activity	small ribosomal subunit	similar to 40S ribosomal protein S2	
up 1.78	4.41	inf	metabolism	catalytic activity	-	guanosine monophosphate reductase	
up 1.74	2.21	4.73	cell communication	binding	-	Yamaguchi sarcoma viral (v-yes) oncogene homolog 1	
down 1.7	4.41	inf	transcription, DNA-dependent	binding	intracellular	X-box binding protein 1	
up 1.66	4.68	inf	monovalent inorganic cation transport	binding	voltage-gated potassium channel complex	potassium voltage-gated channel, shaker-related subfamily, member 3	
up 1.65	2.55	41.77	cell communication	binding	integral to membrane	neuroligin 2	
down 1.65	3.72	inf	signal transduction	protein binding	integrin complex	integrin alpha 1	
up 1.61	2.7	inf	-	nucleic acid binding	extracellular	similar to Cc2-27 (predicted)	
down 1.6	2.01	inf	metabolism	N-acyltransferase activity	-	N-acetyltransferase 1 (arylamine N-acetyltransferase)	

Involved Proteins





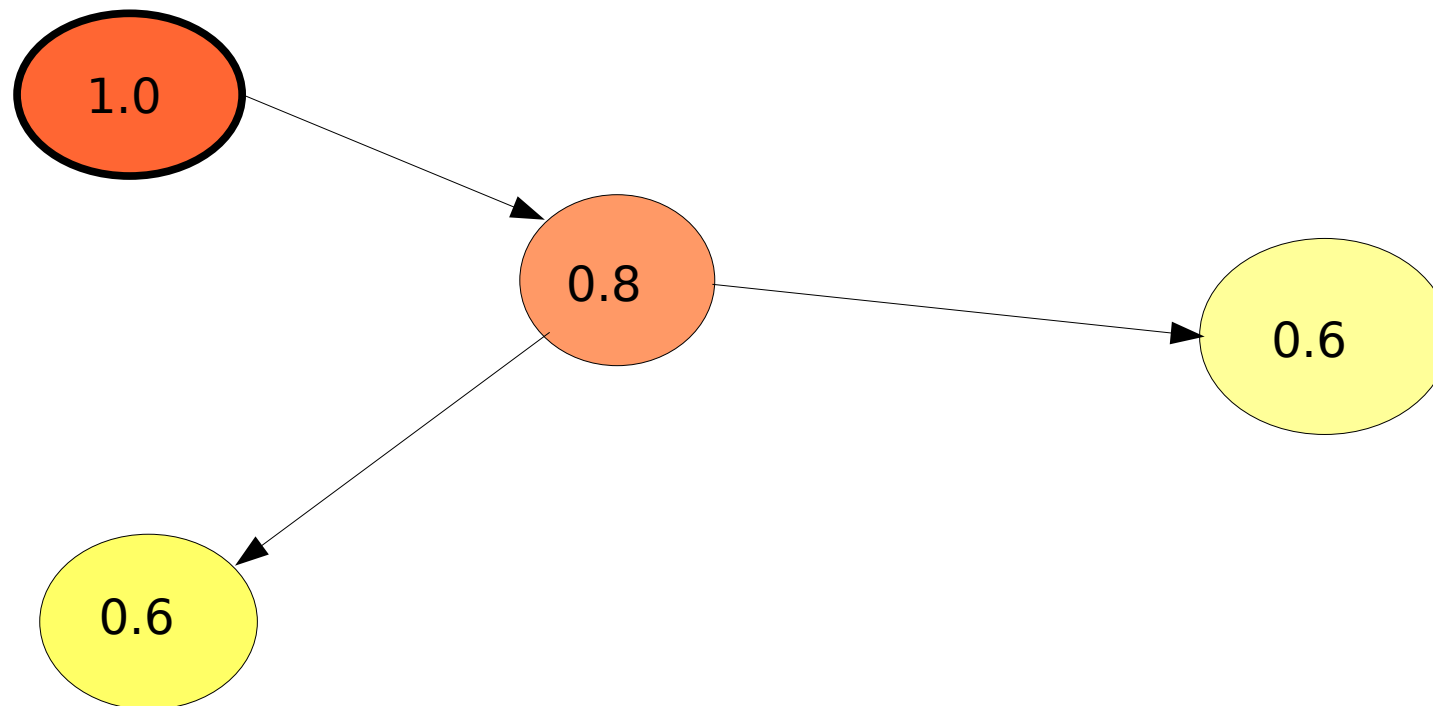
Influenced by/Influences

- MK5 -> Multiple changes in gene expression
- 27000 gene expressions measured
- Those that change will very likely influence other proteins

Which proteins are likely influenced by our measured up/down regulations ?

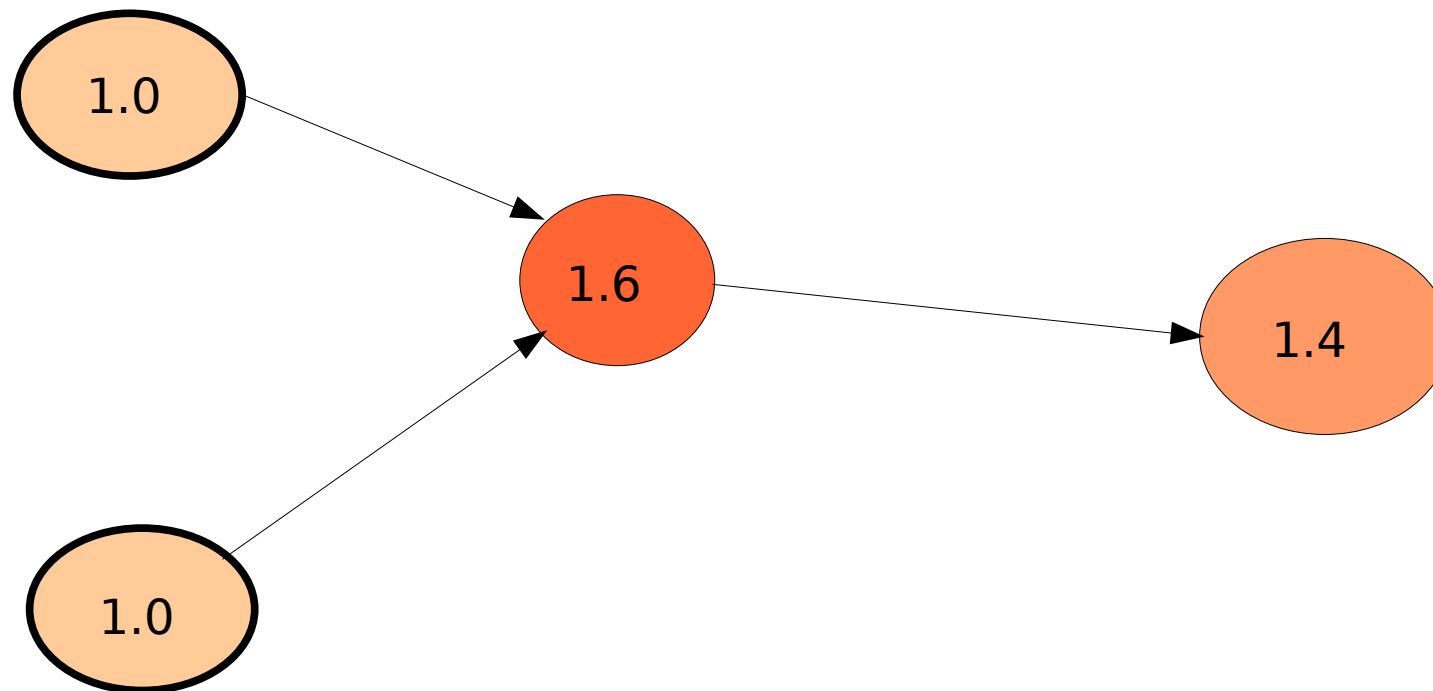
The 'Involved' Game

- Protein change will influence nearby proteins, which in turn ...



The 'Involved' Game

- Multiple proteins changes will all influence their neighbors as well.





The 'Involved' Game

- This network is iterated a number of times to expand the sphere of influence of all the altered gene expressions.
 - affected proteins will have higher numbers
- Protein Interaction key mechanism for signal transduction
 - Protein Interaction Network as published by

Jean François Rual *et al*- Towards a Proteome Scale Map of the Human Protein Protein Interaction Network – Nature 2005 – vol 437, p. 1173-1178



Involved Proteins by Rank

PROTEIN CGI-126 (PROTEIN HSPC155)

RAD50-INTERACTING PROTEIN 1. [Source:RefSeq;Acc:NM_021930]

RHO-RELATED BTB DOMAIN-CONTAINING PROTEIN 2 (DELETED IN BREAST CANCER 2 GENE PROTEIN) (P83).

NADH-UBIQUINONE OXIDOREDUCTASE 18 KDA SUBUNIT, MITOCHONDRIAL PRECURSOR (EC 1.6.5.3) (EC 1.6.99.3) (COMPLEX I-18 KDA) (CI-18 KDA) (COMPLEX I- AQDQ) (CI-AQDQ).

CHROMATIN ACCESSIBILITY COMPLEX PROTEIN 1 (CHRAC-1) (CHRAC-15) (HUCHRAC15) (DNA POLYMERASE EPSILON SUBUNIT P15).

ADIPONECTIN RECEPTOR 2. [Source:RefSeq;Acc:NM_024551]

ODD-SKIPPED RELATED 1; ODZ (ODD OZ/TEN-M) RELATED 1.

DNA POLYMERASE EPSILON P12 SUBUNIT (DNA POLYMERASE EPSILON SUBUNIT 4)

PROTEIN X 0004. [Source:RefSeq;Acc:NM_016301]

XPA BINDING PROTEIN 1; MBD2 INTERACTOR PROTEIN; PUTATIVE ATP(GTP)-BINDING PROTEIN HBS1-LIKE. [Source:RefSeq;Acc:NM_006620]

HOMEBOX PROTEIN HLX1 (HOMEBOX PROTEIN HB24).



Involved Proteins by Rank

NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT BETA (NF-Y PROTEIN CHAIN B) (NF-YB) (CCAAT-BINDING TRANSCRIPTION FACTOR SUBUNIT A) (CBF-A) (CAAT- BOX DNA BINDING PROTEIN SUBUNIT B).

GROWTH FACTOR RECEPTOR-BOUND PROTEIN 2 (GRB2 ADAPTER PROTEIN) (SH2/SH3 ADAPTER GRB2) (ASH PROTEIN).

SERINE/THREONINE-PROTEIN KINASE NEK2 (EC 2.7.1.37) (NIMA-RELATED PROTEIN KINASE 2) (NIMA-LIKE PROTEIN KINASE 1) (HSPK 21).

E2A-PBX1-ASSOCIATED PROTEIN; PUTATIVE 47 KDA PROTEIN.

NEURON NAVIGATOR 1; NEURON NAVIGATOR-1; PORE MEMBRANE AND/OR FILAMENT INTERACTING LIKE PROTEIN 3.

NEURON NAVIGATOR 3; PORE MEMBRANE AND/OR FILAMENT INTERACTING LIKE PROTEIN 1; STEERIN 3.

NEURON NAVIGATOR 2 ISOFORM L; RETINOIC ACID INDUCIBLE IN NEUROBLASTOMA; PORE MEMBRANE AND/OR FILAMENT INTERACTING LIKE PROTEIN 2; HELICASE HELAD1.

GRB2-RELATED ADAPTOR PROTEIN.

DNA-REPAIR PROTEIN COMPLEMENTING XP-G CELLS (XERODERMA PIGMENTOSUM GROUP G COMPLEMENTING PROTEIN) (DNA EXCISION REPAIR PROTEIN ERCC-5).

DOCKING PROTEIN 2 (P56(DOK-2)) (DOWNSTREAM OF TYROSINE KINASE 2).



Involved Proteins by Rank

TYROSINE-PROTEIN KINASE FRK (EC 2.7.1.112) (NUCLEAR TYROSINE PROTEIN KINASE RAK).

[Source:SWISSPROT;Acc:P42685]

DOK-LIKE PROTEIN. [Source:RefSeq;Acc:NM_024872]

DOCKING PROTEIN 1 (P62(DOK)) (DOWNSTREAM OF TYROSINE KINASE 1) (PP62).

[Source:SWISSPROT;Acc:Q99704]

TATA-BINDING PROTEIN-ASSOCIATED PHOSPHOPROTEIN (DOWN-REGULATOR OF TRANSCRIPTION 1) (DR1 PROTEIN)

WISKOTT-ALDRICH SYNDROME PROTEIN (WASP).

PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC (EC 2.7.1.112) (P60-SRC) (C-SRC).

TYROSINE-PROTEIN KINASE HCK (EC 2.7.1.112) (P59-HCK/P60-HCK) (HEMOPOIETIC CELL KINASE). [Source:SWISSPROT;Acc:P08631]

PROTO-ONCOGENE C-CRK (P38) (ADAPTER MOLECULE CRK). [Source:SWISSPROT;Acc:P46108]

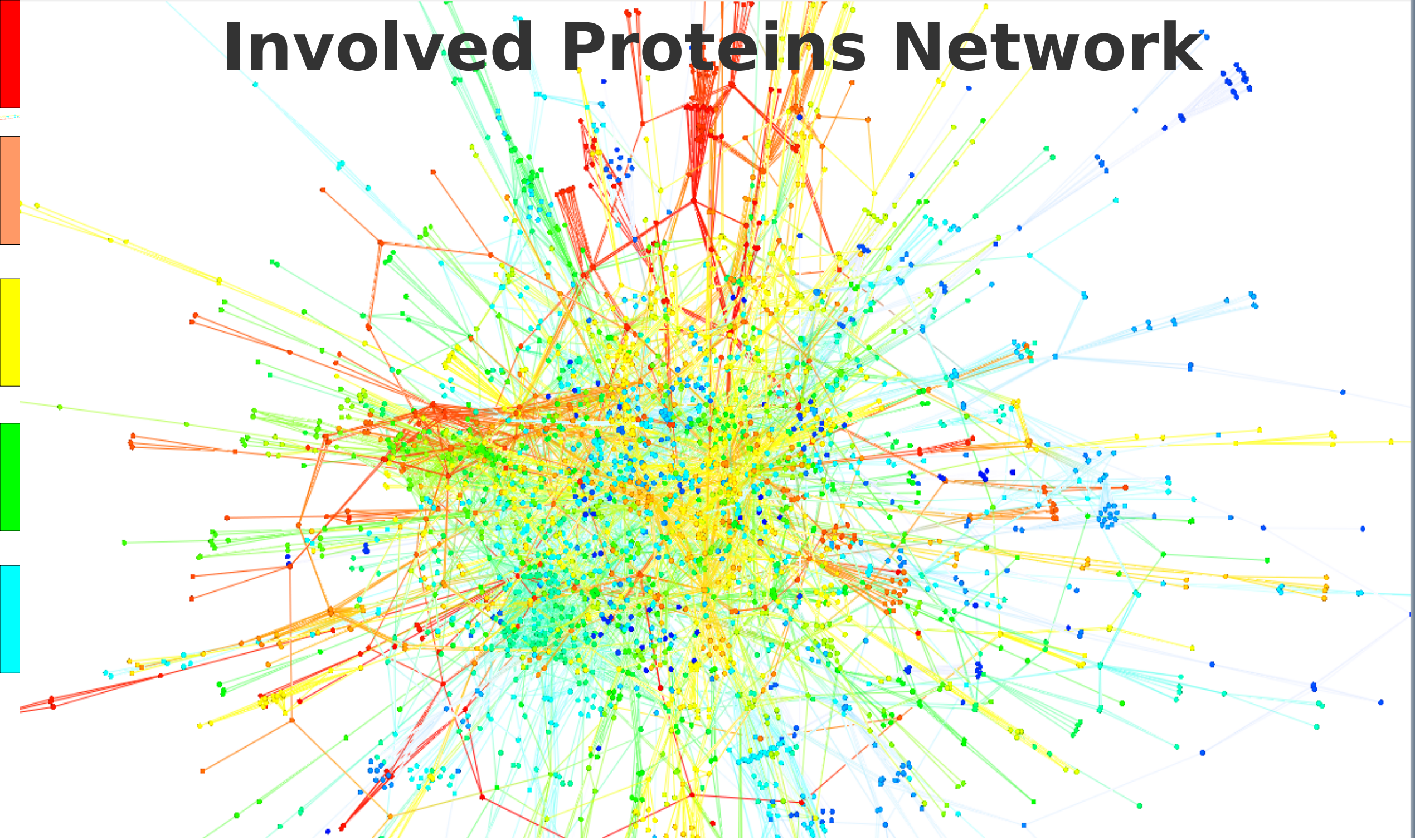
CRK-LIKE PROTEIN. [Source:SWISSPROT;Acc:P46109]

PROTO-ONCOGENE TYROSINE-PROTEIN KINASE ABL1 (EC 2.7.1.112) (P150) (C-ABL).

PROTO-ONCOGENE TYROSINE-PROTEIN KINASE FGR (EC 2.7.1.112) (P55-FGR) (C-FGR).

TYROSINE-PROTEIN KINASE ABL2 (EC 2.7.1.112) (TYROSINE KINASE ARG).

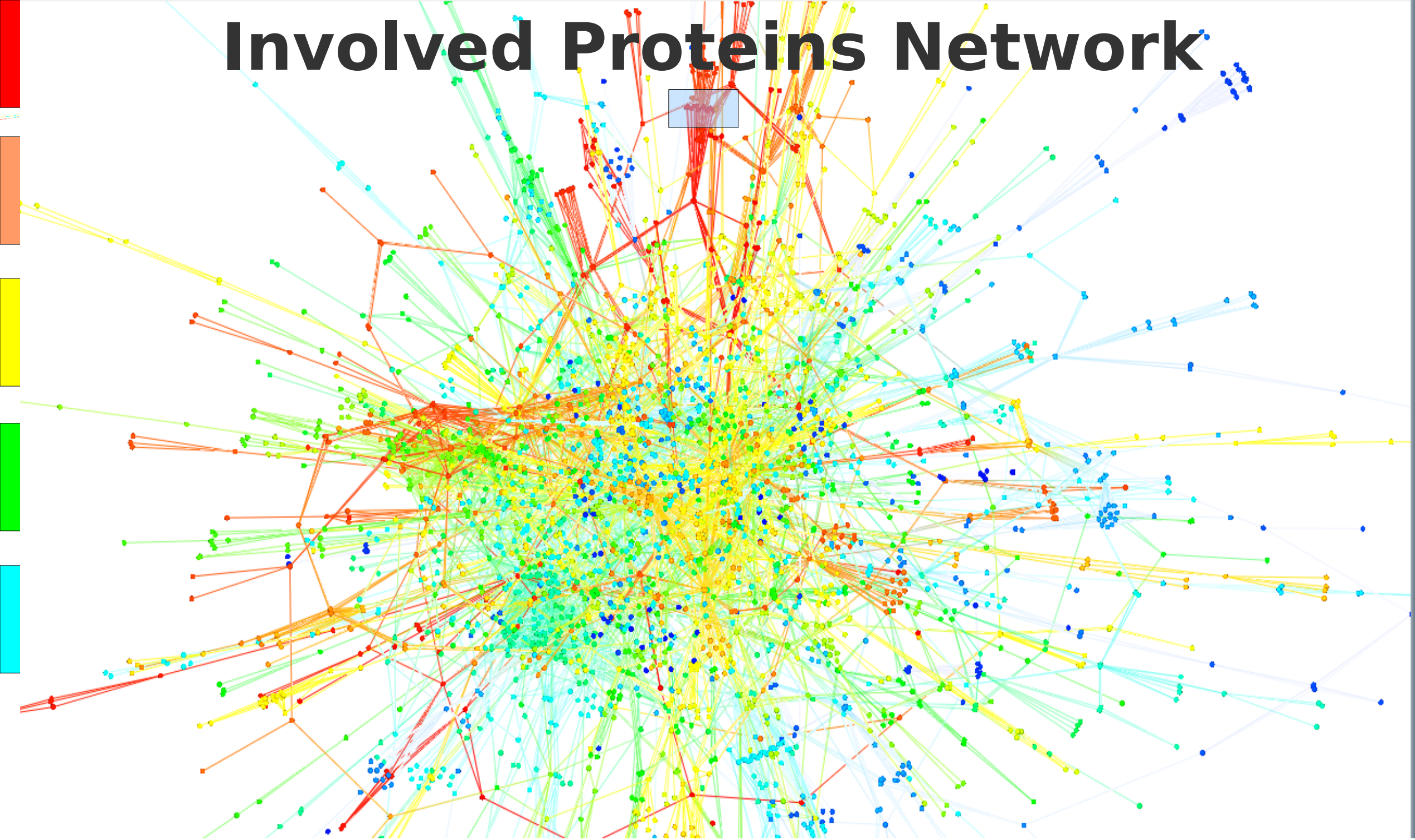
Involved Proteins Network



Involved Proteins Network

- 
- Red = Highest involvement; Blue = Lowest Involvement
 - Based on our lowest estimates for up/down regulation
 - Based on the high confidence set of protein interactions
 - Measured gene expressions are not listed

Involved Proteins Network



Involved Protein Network

11 precursor {er-associated dnaJ P 3} {erj3} {er-associated hsp40 co-chaperone} {hdj9} {pwp1- Interacting P 4}

101371 proto-oncogene tyrosine-P kinase src {ec 2.7.1.112} {p60-src} {c-src}

97007 proto-oncogene tyrosine-P kinase abl1 {ec 2.7.1.112} {p150} {c-abl}

167193 proto-oncogene c-crk {p38} {adapter molecule crk}

101336 tyrosine-P kinase hck {ec 2.7.1.112} {p59-hck/p60-hck} {hemopoietic cell kinase}

143322 tyrosine-P kinase abl2 {ec 2.7.1.112} {tyrosine kinase arg}

Involved Protein Network





Conclusions

- Novel micro-array analysis approach to measure quantitative regulation
 - Confidence intervals based on control sample
 - Confidence becomes greater with more dots
- Network of Involved Proteins
 - Prediction based on network simulation
 - Visualization & Clustering