I Just Received My Microarray Data, Now What?

Danny Park

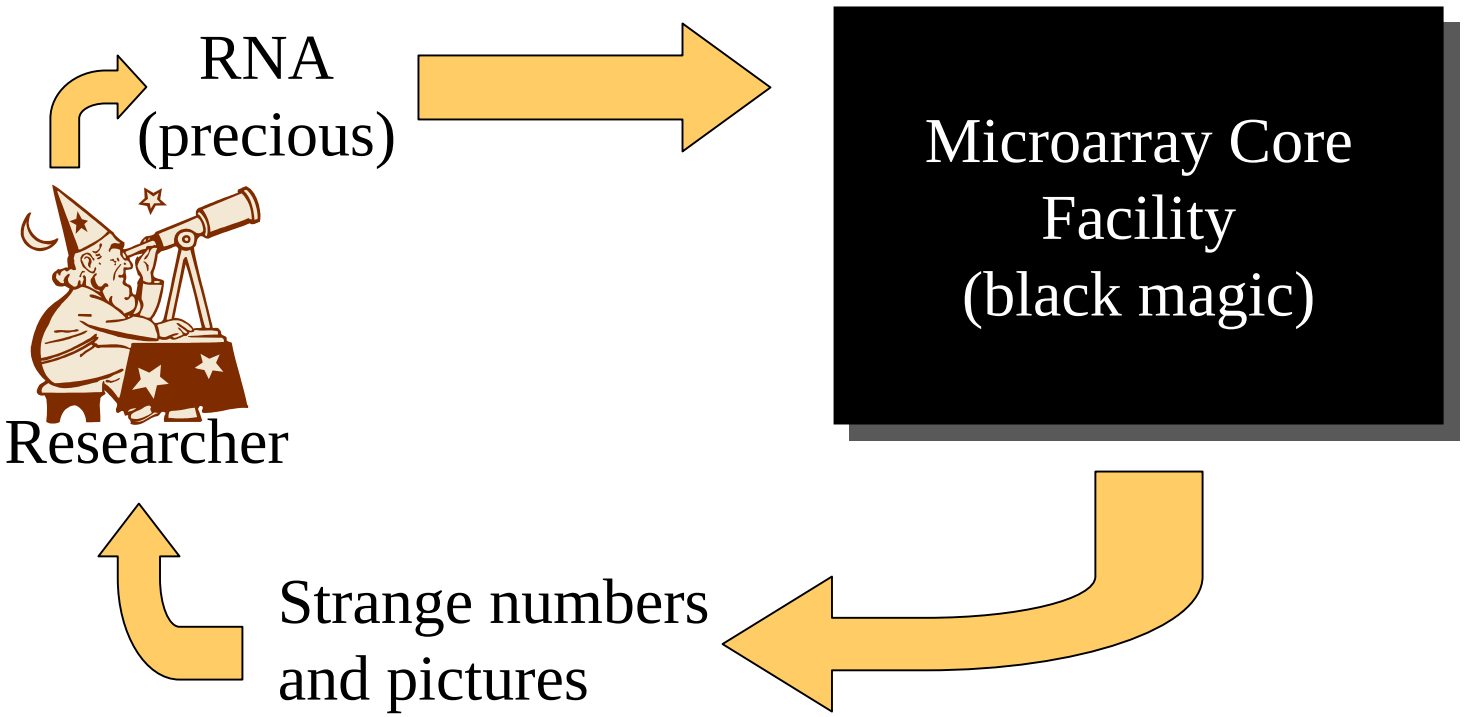
MGH-PGA (ParaBioSys)

Sat April 24, 2004

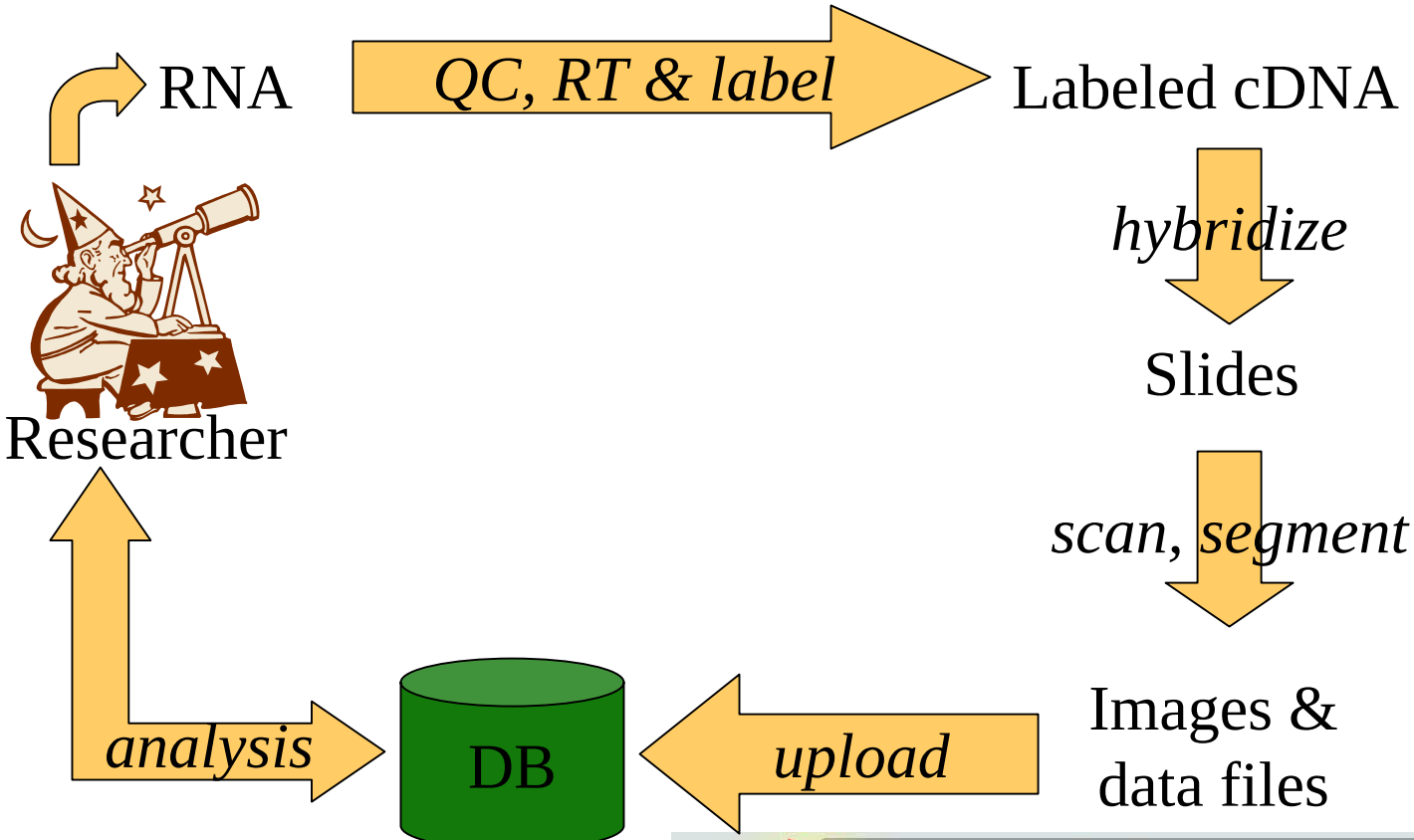
I Just Received My Microarray Data …

- Where did this come from?
  - A description of the process from RNA to raw data
- What do I do now?
  - A description of how to analyze your data

Demystifying the Core Facility

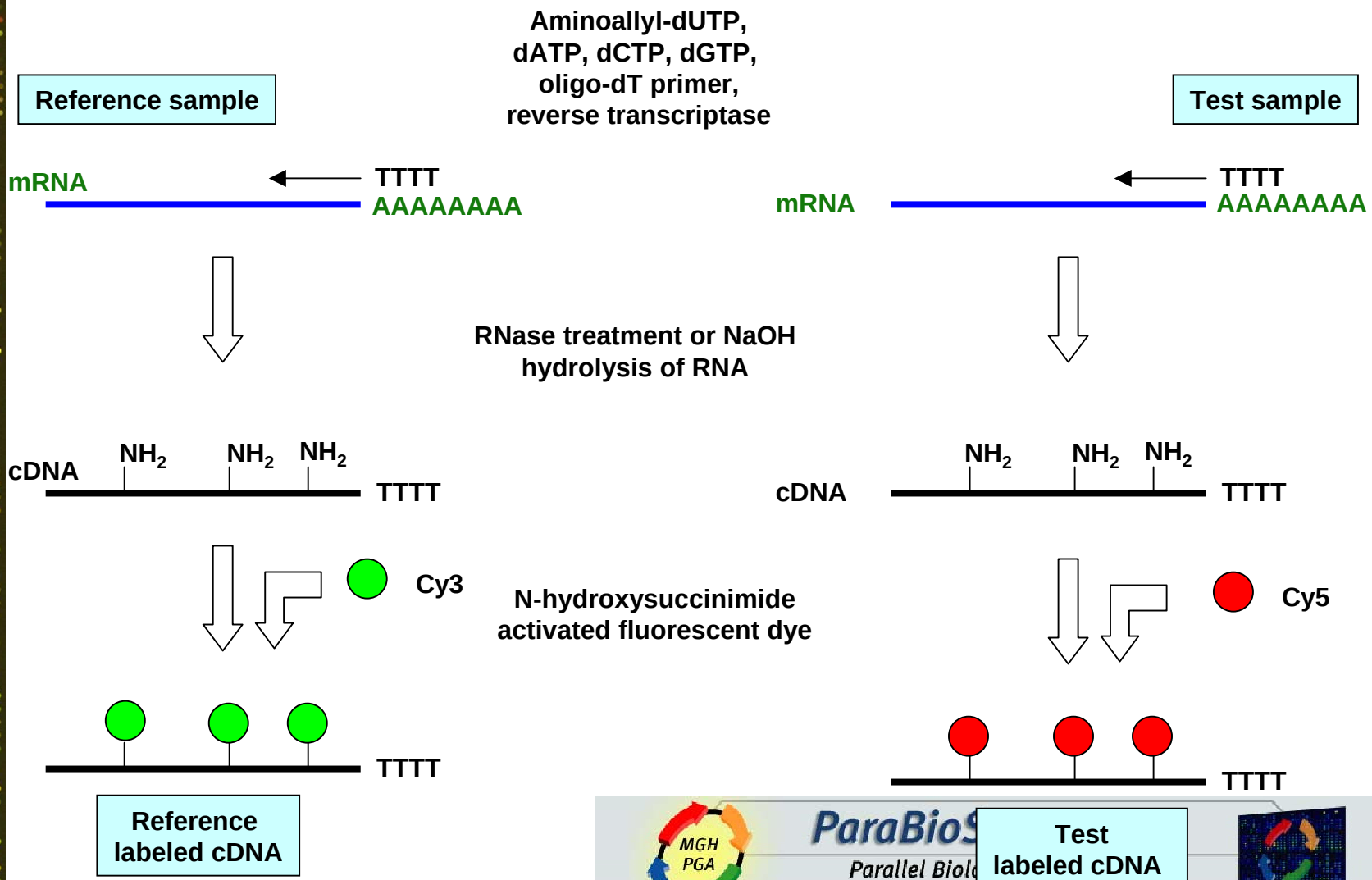


Demystifying the Core Facility



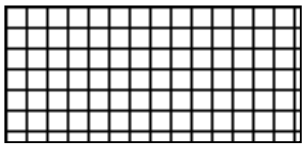
Approximately 20 μg total RNA required

RT & Labeling

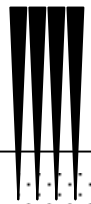


Hybridization

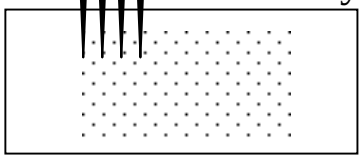
Synthesized oligonucleotides in 384 well plates

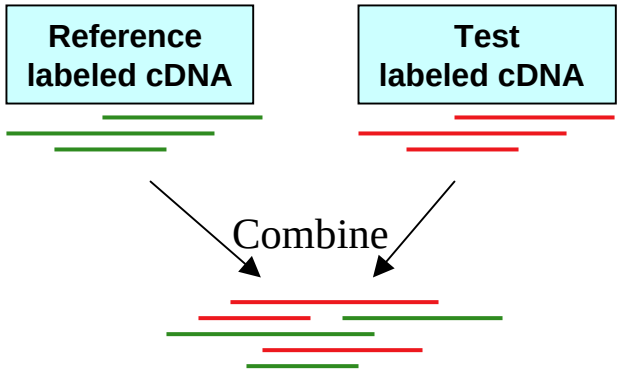


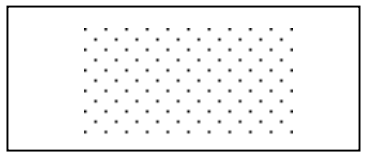
Robotic printing

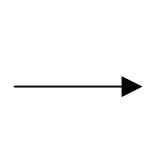


Microarray



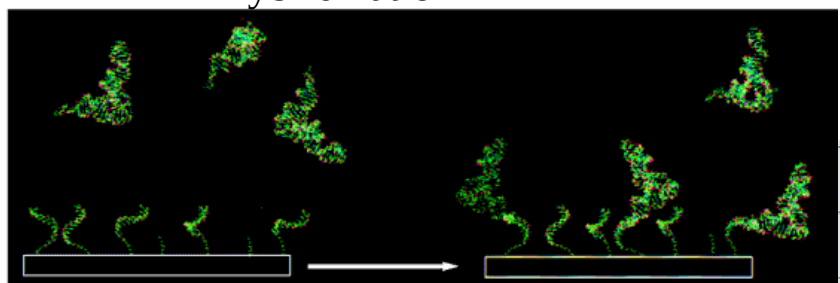






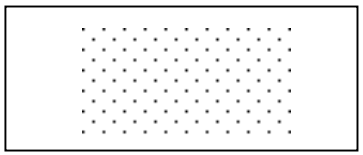
Genomic Solutions Hybridization Station (PerkinElmer)

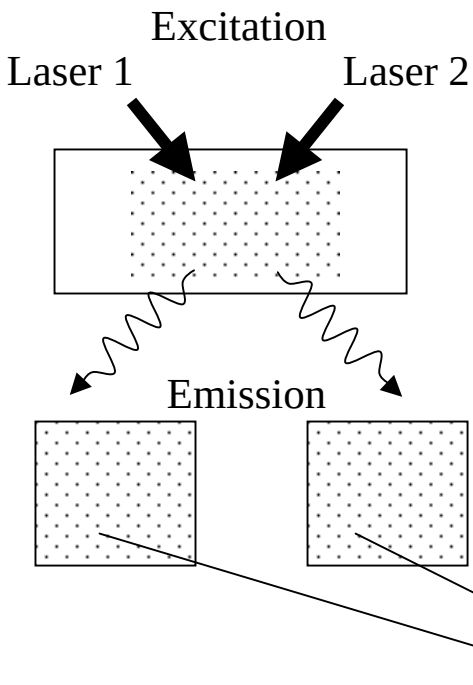
Hybridization

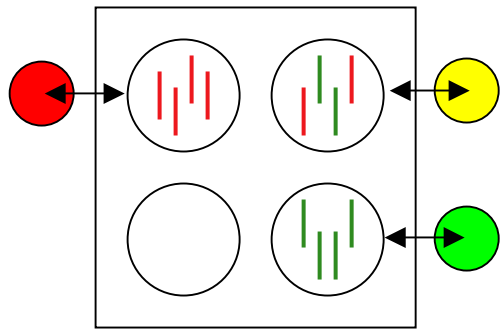


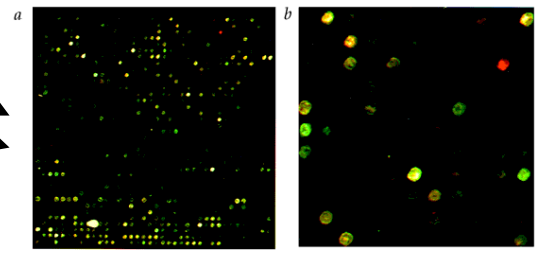
Scanning

Hybridized Microarray





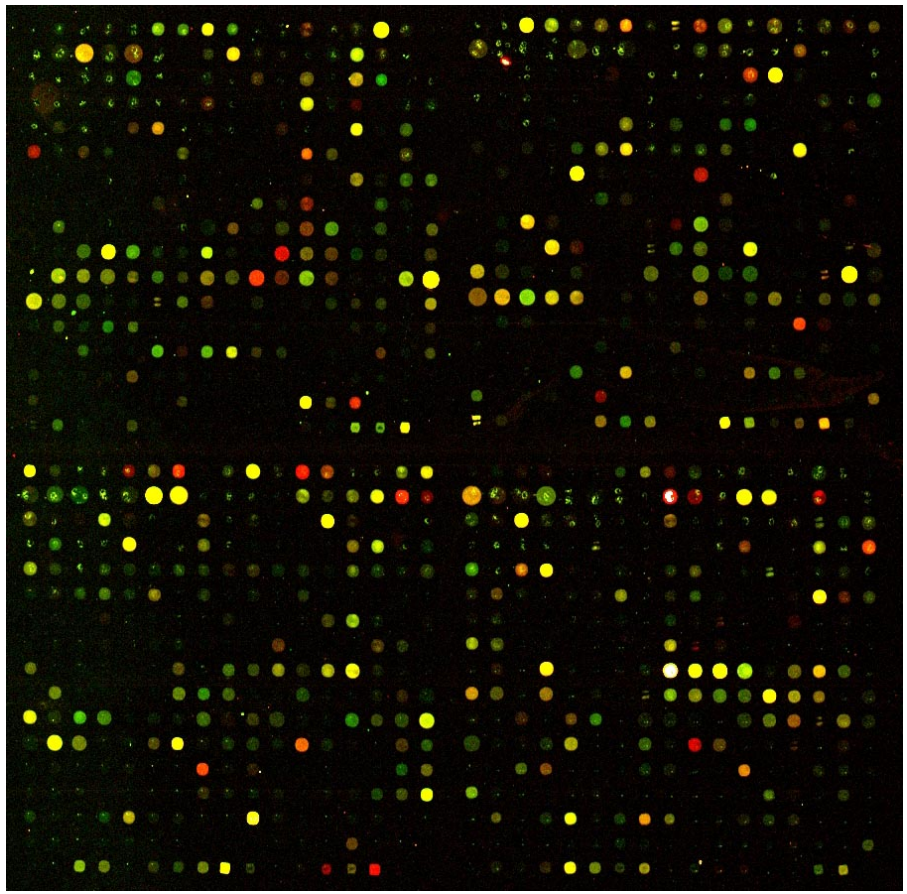


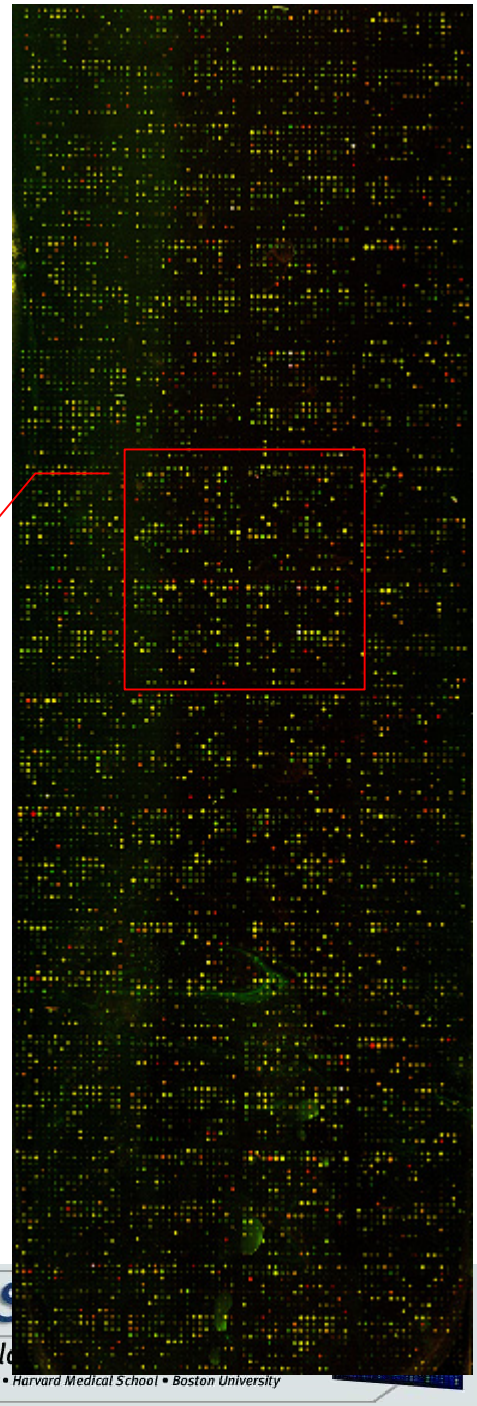


Monochrome pictures combined

Axon Instruments GenePix 4000B

Scanning





20 μg total RNA macrophage RAW

Cy5: 100 ng/mL LPS 2 h

Cy3: no treatment

MGH PGA

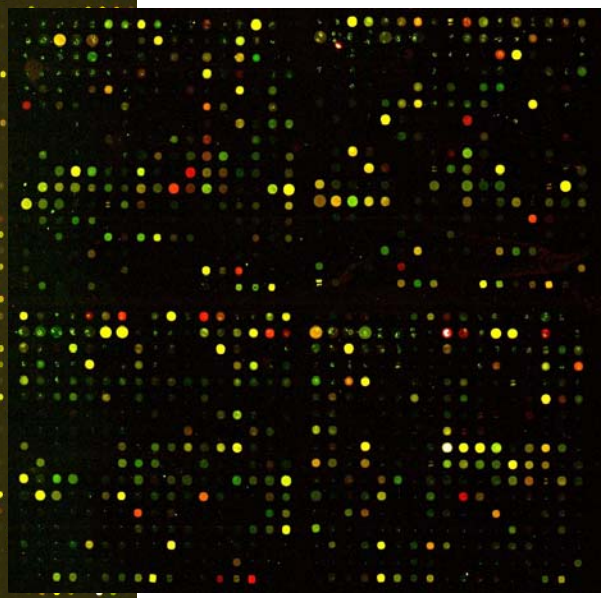
ParaBioS

Parallel Biol

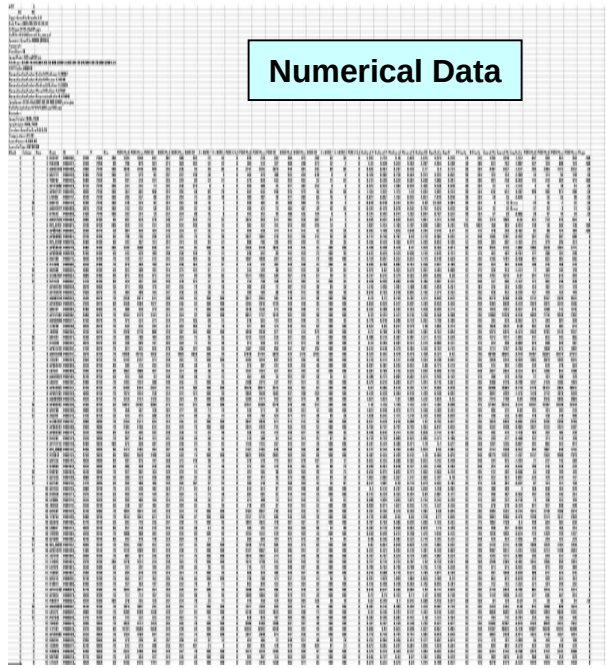
Mass. General Hospital • Harvard Medical School • Boston University

Segmentation

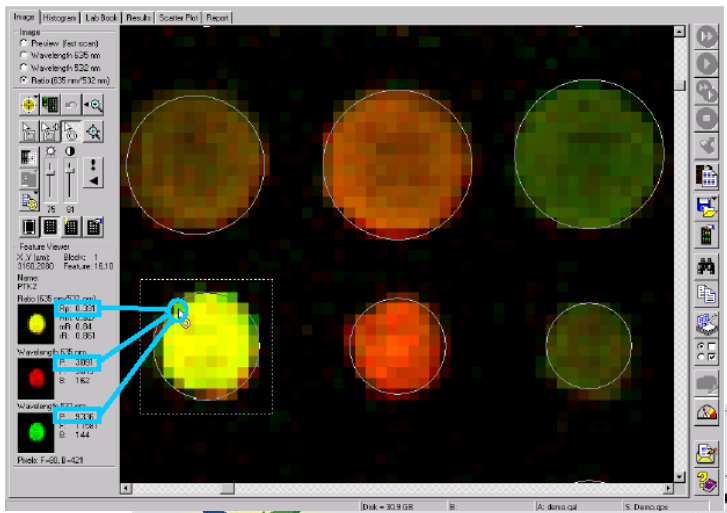
Scanned Image



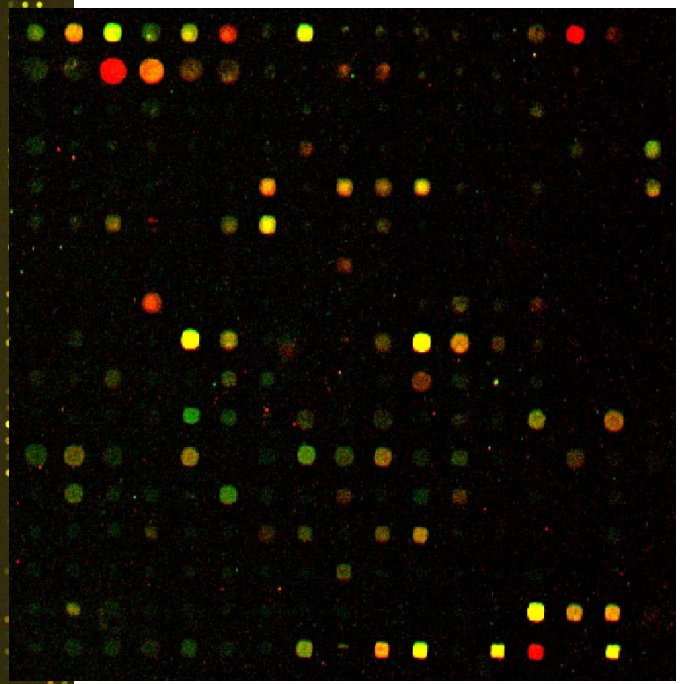
Numerical Data



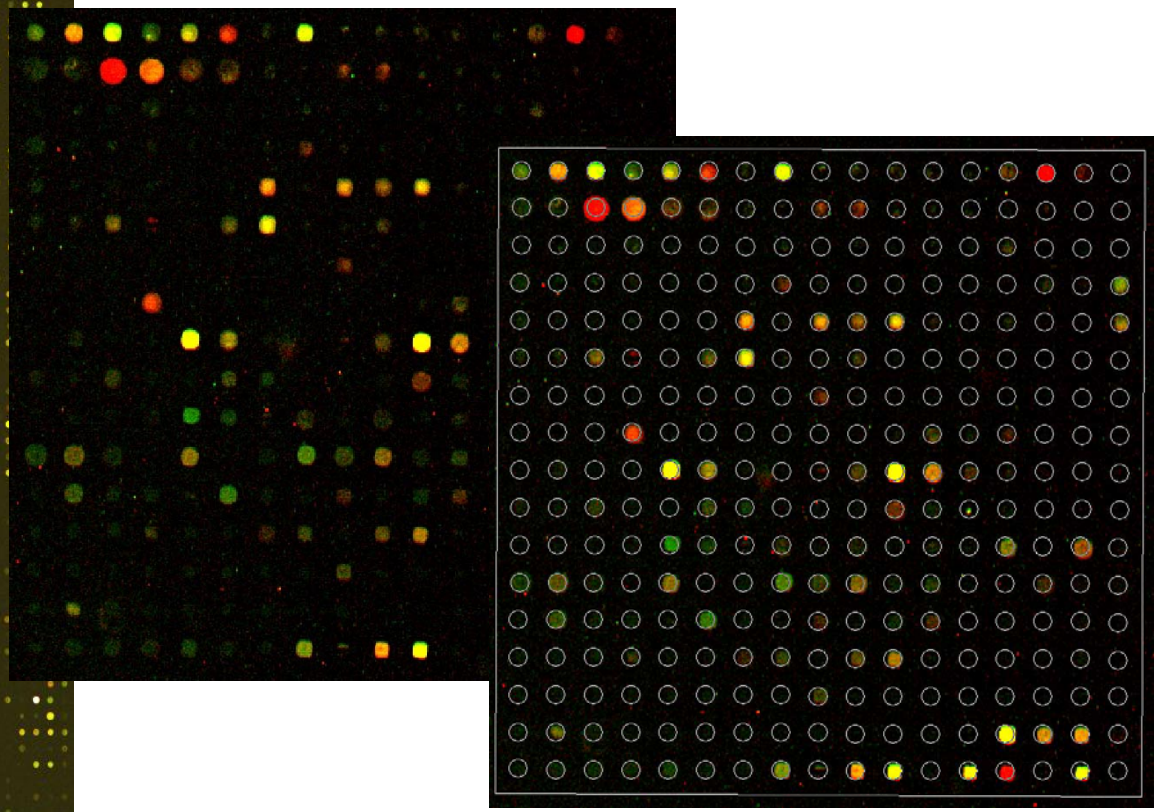
Segmentation Software



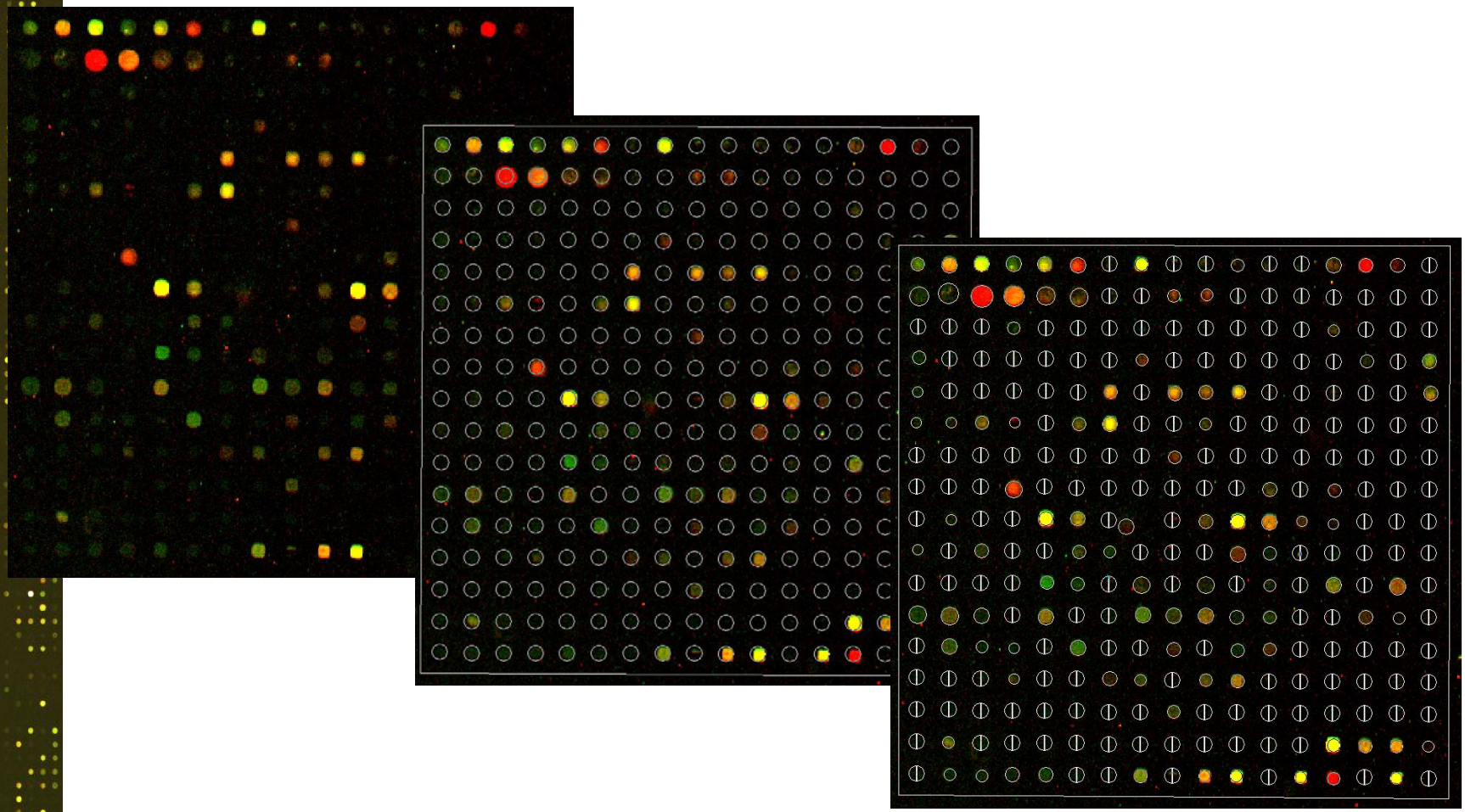
Segmentation



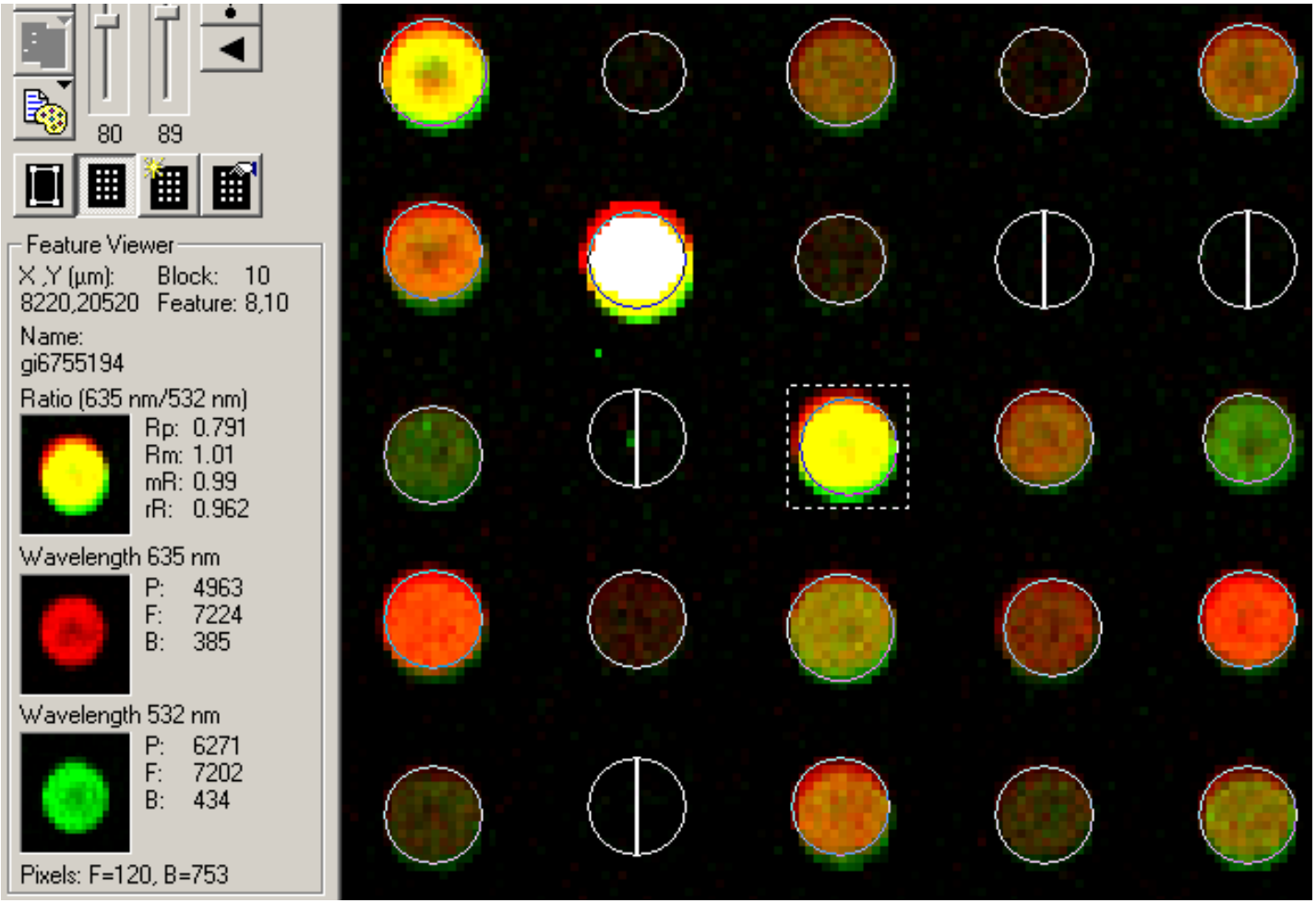
Segmentation



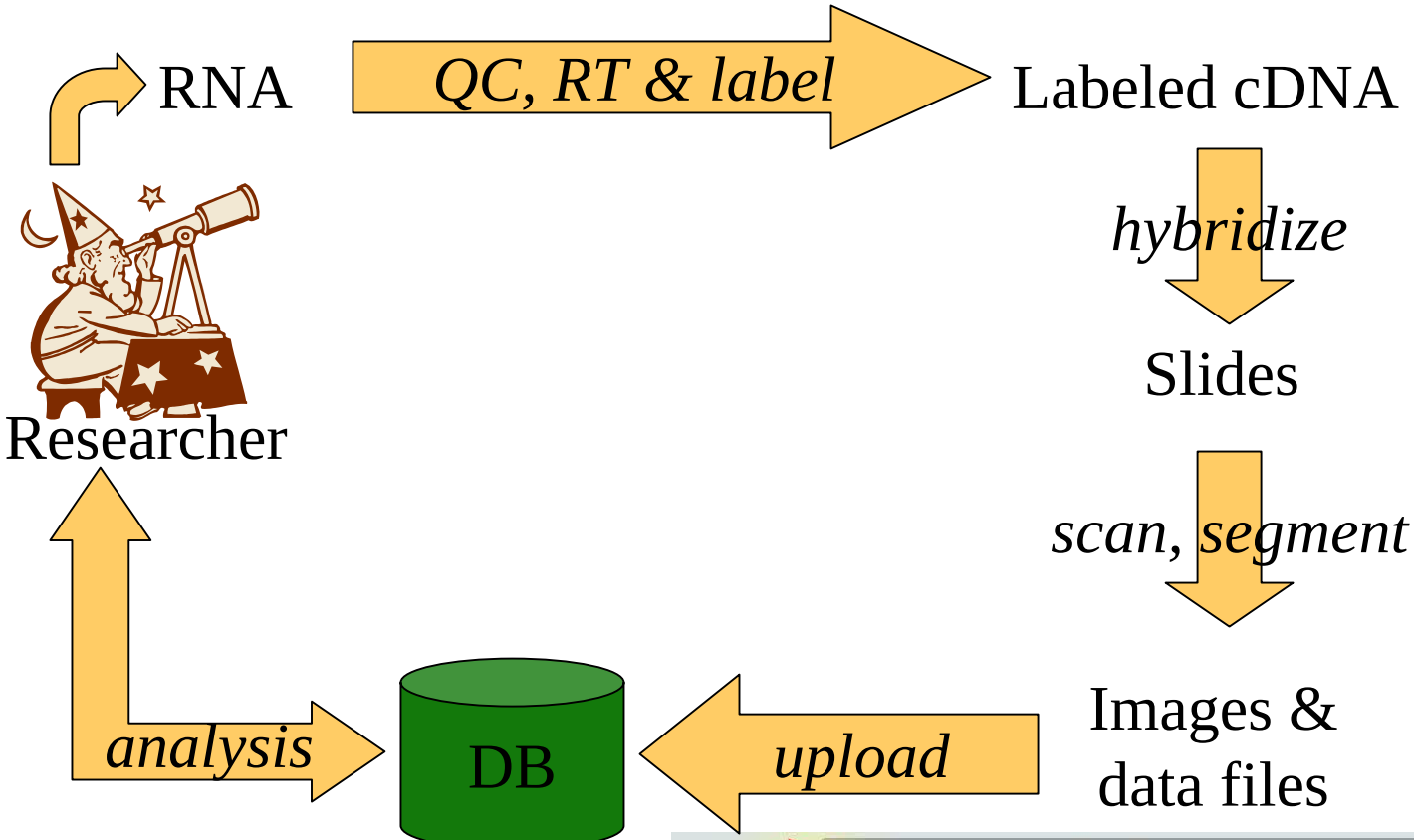
Segmentation



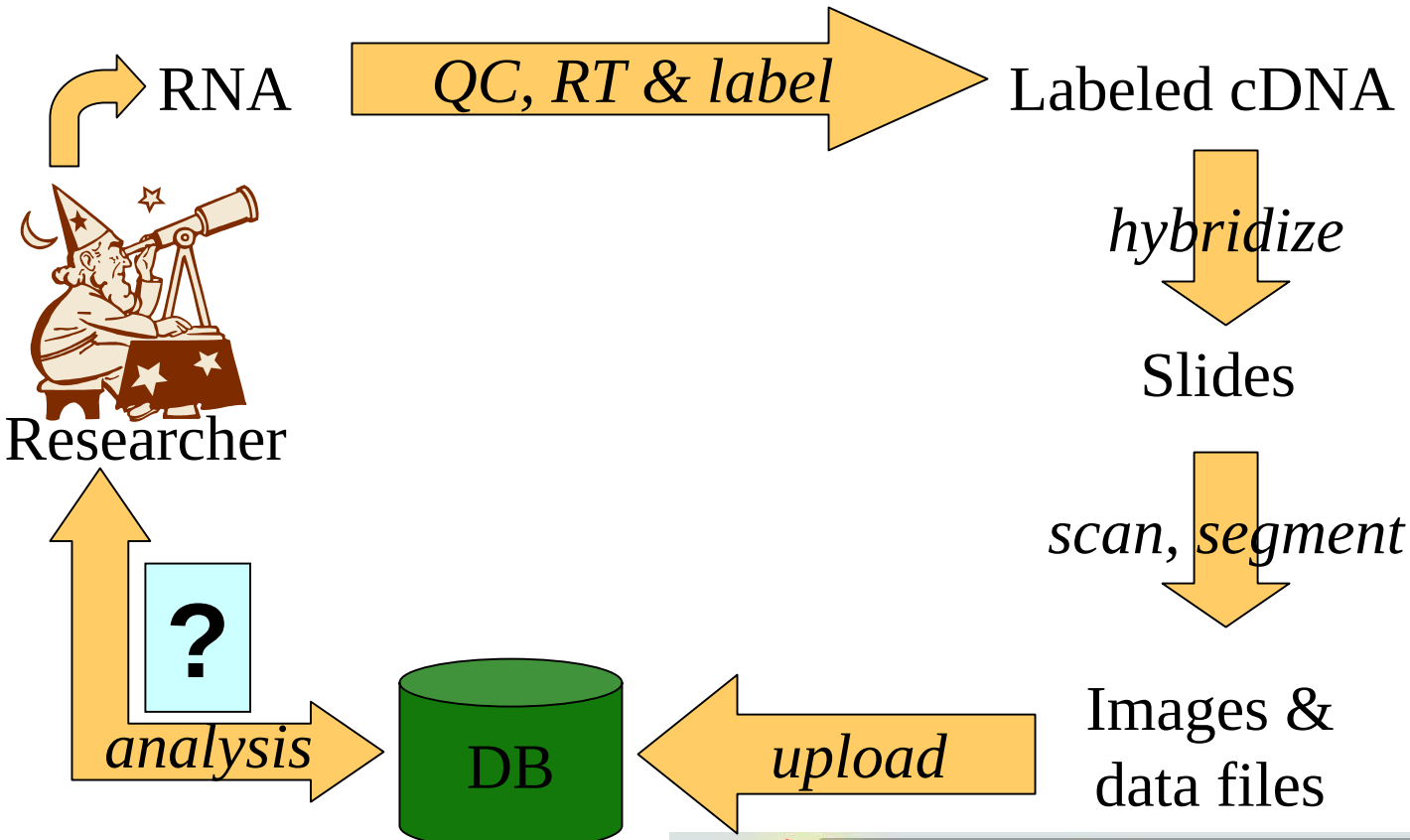
Segmentation



Core Facility – Demystified!



Core Facility – Demystified?



What Do I Do Now? (data analysis)

- What was I asking?
  - Remember your experimental design
- How do I analyze the data?
  - Learn some typical filters, transformations, and statistics
  - Learn the necessary software tools
  - Consult biostatistician

What Was I Asking?

- Typically: “which genes changed expression patterns when I did ____”
- Common ____’s:
  - Binary conditions: knock out, treatment, etc
  - Unordered discrete scales: multiple types of treatment or mutations
  - Continuous scales: time courses, levels of treatment, etc
- **My focus: binary conditions (aka “diagnostic experiments”)**

Diagnostic Experiments

- Two-sample comparison w/N replicates
  - KO vs. WT
  - Treated vs. untreated
  - Diseased vs. normal
  - Etc
- Question of interest: which genes or groups are (most) differentially expressed?

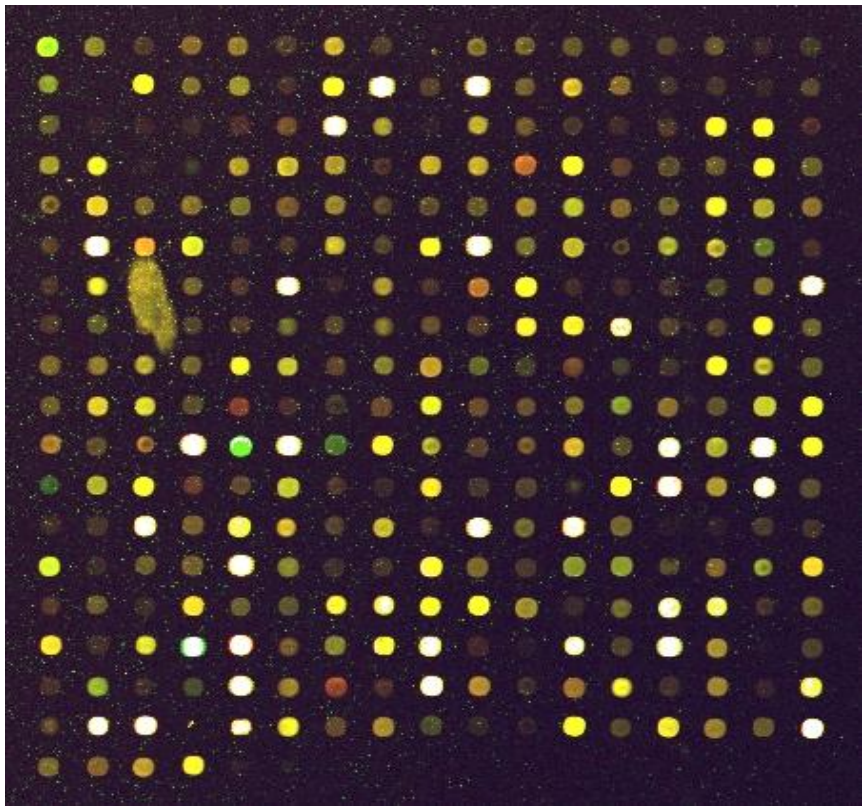
Software Tools?

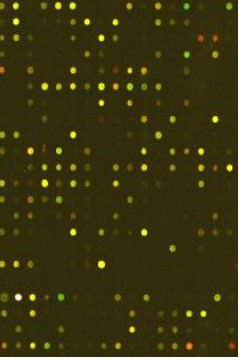
- BASE – BioArray Software Environment
  - Data storage and distribution
  - Simple filtering, normalization, averaging, and statistics
  - Export/Download results to other tools
- R, Bioconductor (complex statistics)
- MS Excel (general)
- TIGR Multi Experiment Viewer (clustering)
- GenMAPP (ontologies)

Analyzing a Diagnostic Experiment

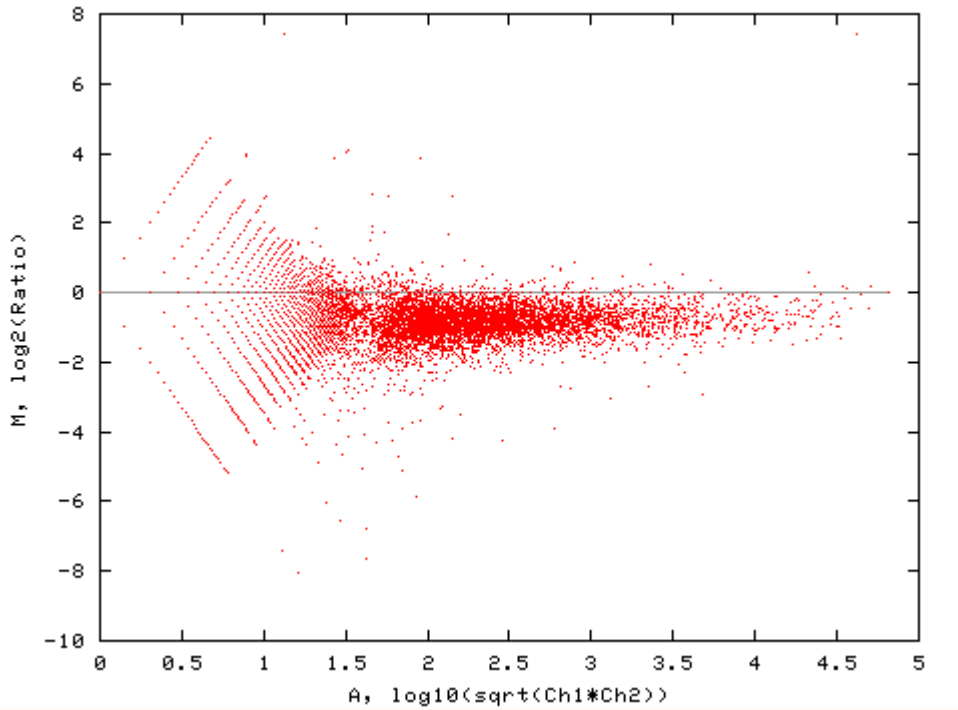
- Filter out bad spots
- Adjust low intensities
- Normalize – correct for non-linearities and dye inconsistencies
- Calculate average ratios and significance values per gene
- Rank, sort, filter, squint, sift data
- Validate results

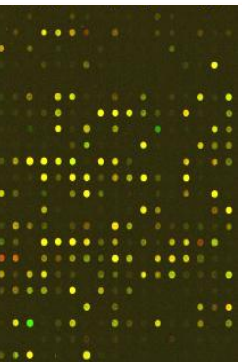
Filtering bad spots – Why?



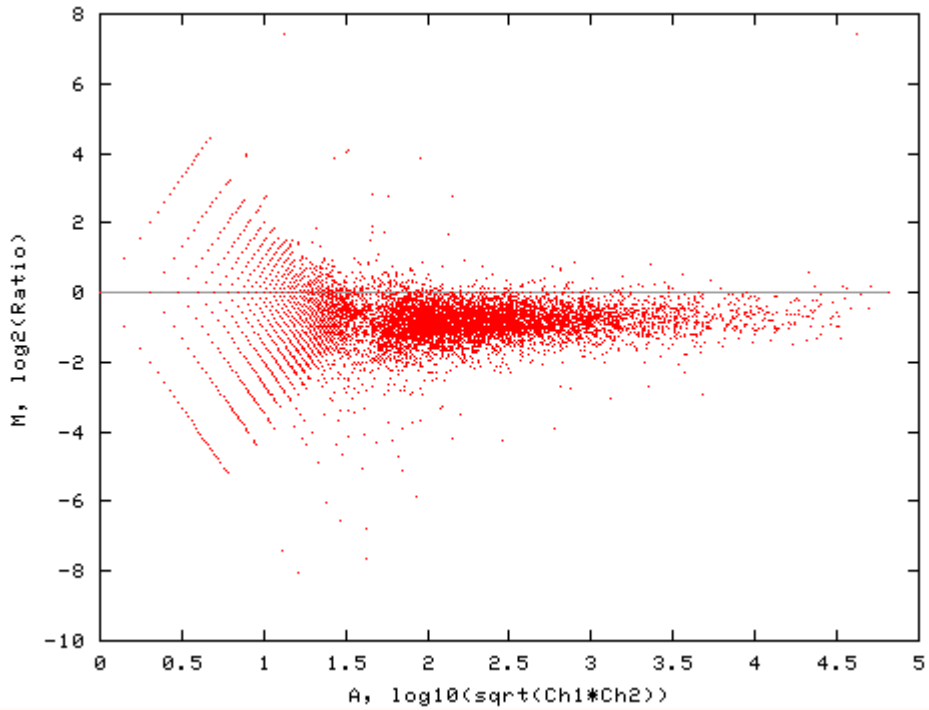


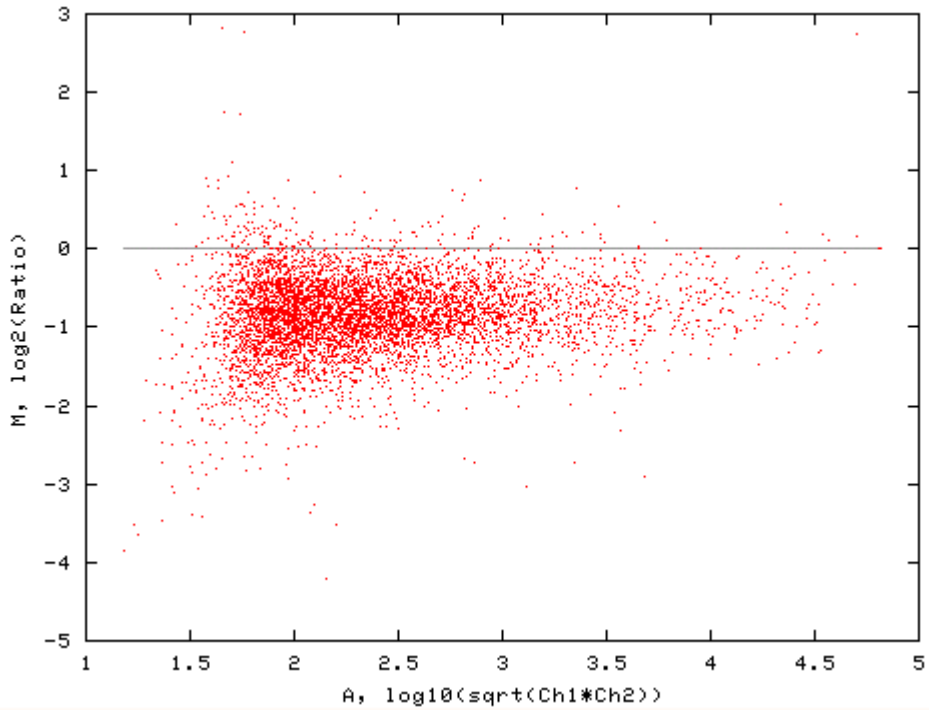
Filtering bad spots





Filtering bad spots





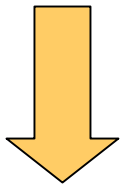
Filter

Adjusting low intensities – Why?

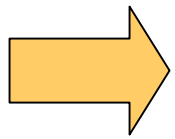
Gene	T	C	Ratio	log2	Sp	Reporter name
2210417C17Rik	-48	1942	NaN	NaN		homolog to OPIOID GROWTH FACTOR [...]
B230114J08Rik	-91	1141	NaN	NaN		homolog to MEMBRANE PROTEIN C21ORF4 - [...]
Tgtp	-49	3020	NaN	NaN		T-cell specific GTPase

Adjusting low intensities – Why?

Gene	T	C	Ratio	log2	Sp	Reporter name
2210417C17Rik	-48	1942	NaN	NaN		homolog to OPIOID GROWTH FACTOR [...]
B230114J08Rik	-91	1141	NaN	NaN		homolog to MEMBRANE PROTEIN C21ORF4 - [...]
Tgtp	-49	3020	NaN	NaN		T-cell specific GTPase

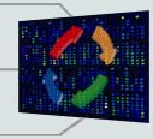


LOWESS Normalization



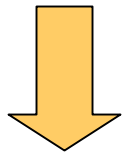
Data is gone!

log(T), log(C)



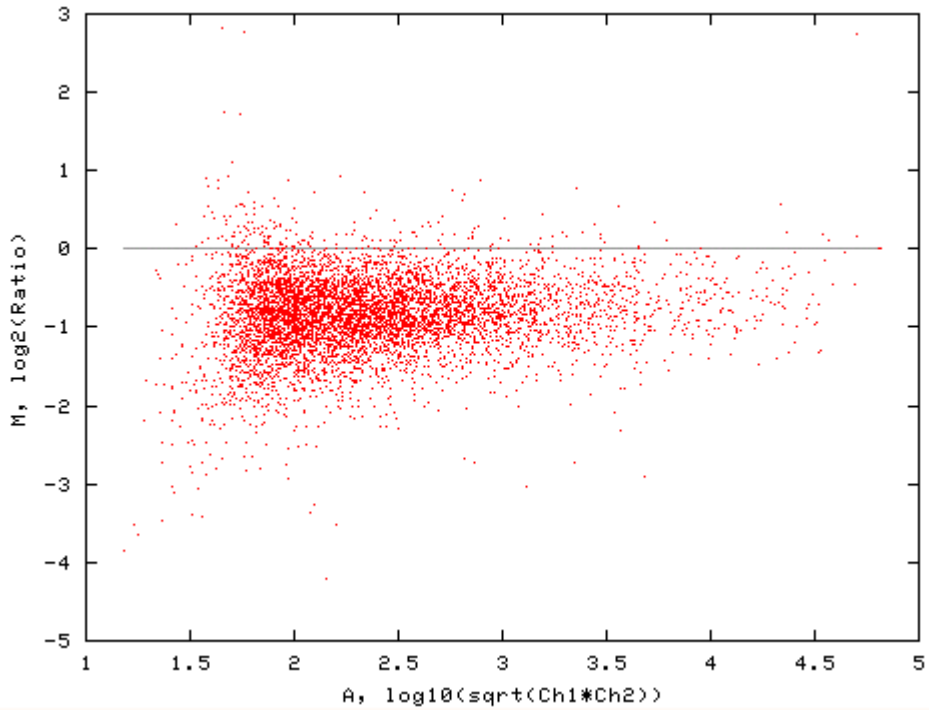
Adjusting low intensities

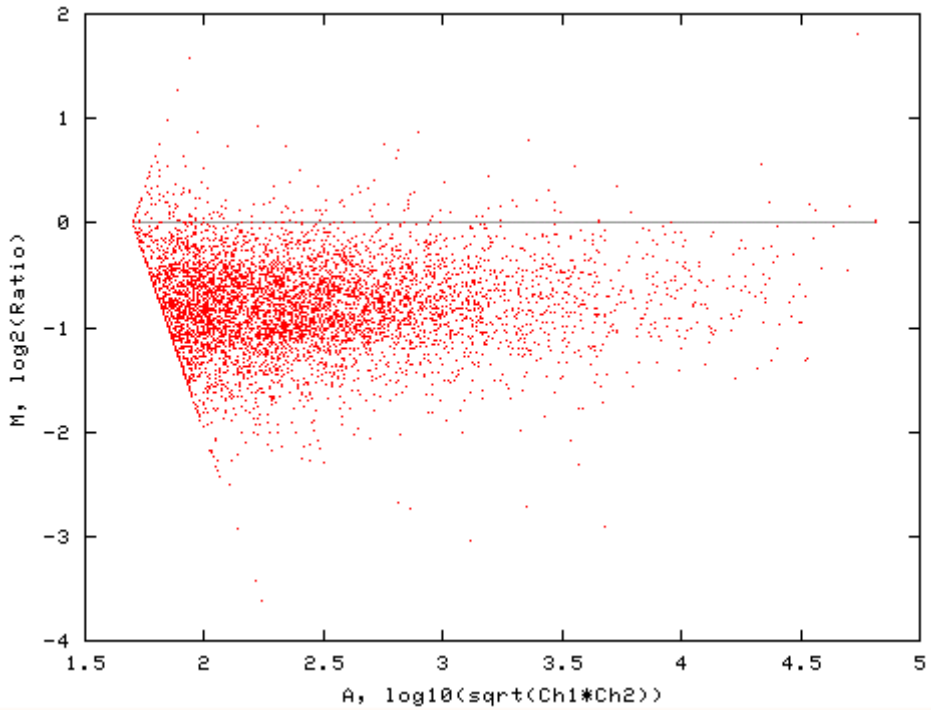
Gene	T	C	Ratio	log2	Sp	Reporter name
2210417C17Rik	-48	1942	NaN	NaN		homolog to OPIOID GROWTH FACTOR [...]
B230114J08Rik	-91	1141	NaN	NaN		homolog to MEMBRANE PROTEIN C21ORF4 - [...]
Tgtp	-49	3020	NaN	NaN		T-cell specific GTPase



Gene	T	C	Ratio	log2	Sp	Reporter name
2210417C17Rik	50	1942	0.03	-5.28		homolog to OPIOID GROWTH FACTOR [...]
B230114J08Rik	50	1141	0.04	-4.51		homolog to MEMBRANE PROTEIN C21ORF4 - [...]
Tgtp	50	3020	0.02	-5.92		T-cell specific GTPase

Adjusting low intensities

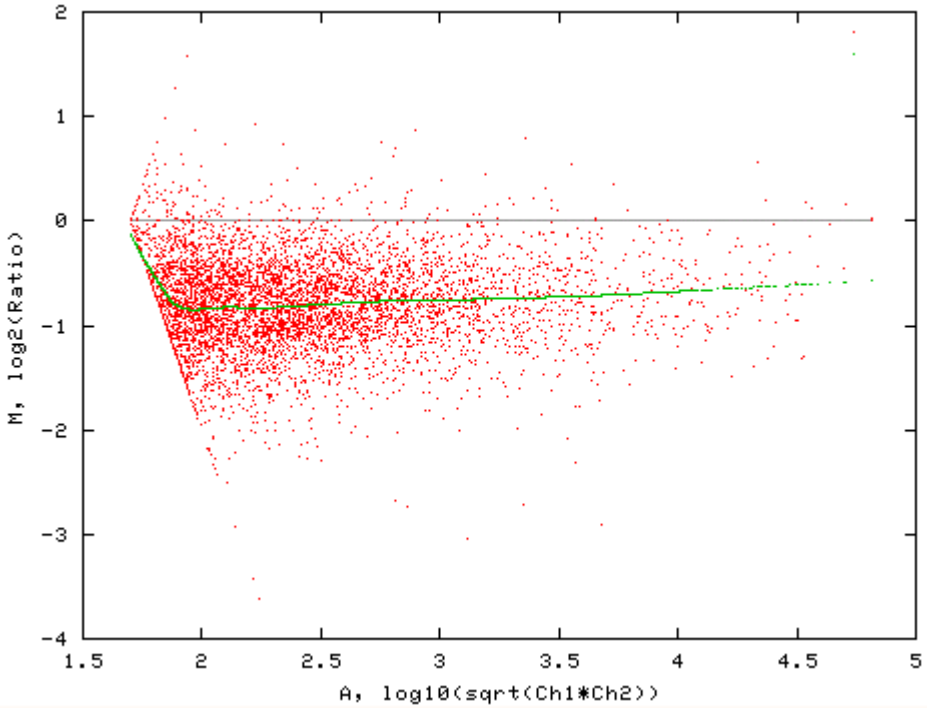




Int Limit

Normalization – Why?

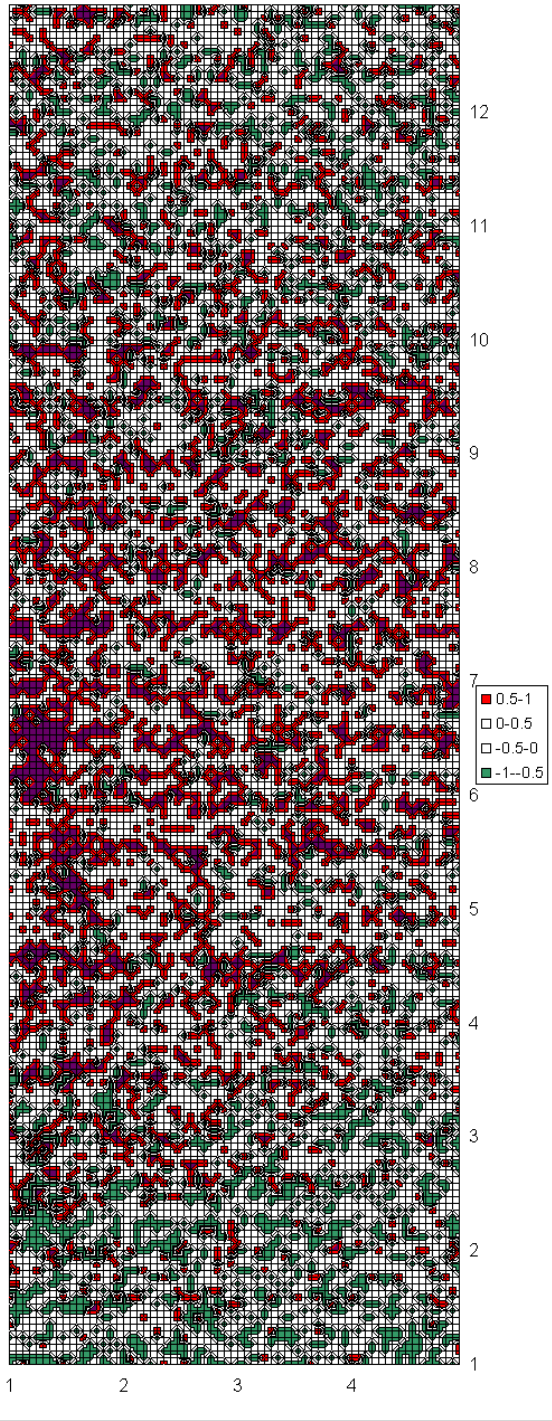
- Not perfectly centered around zero
- Implies that nearly all genes down regulated?
- There are dye effects



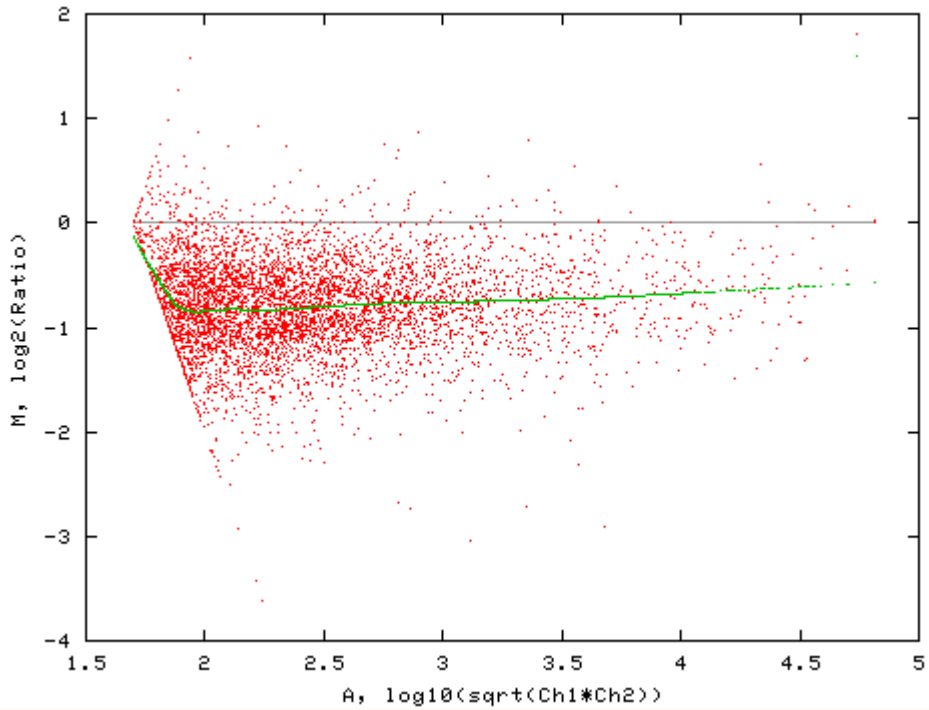
Normalization – Why?

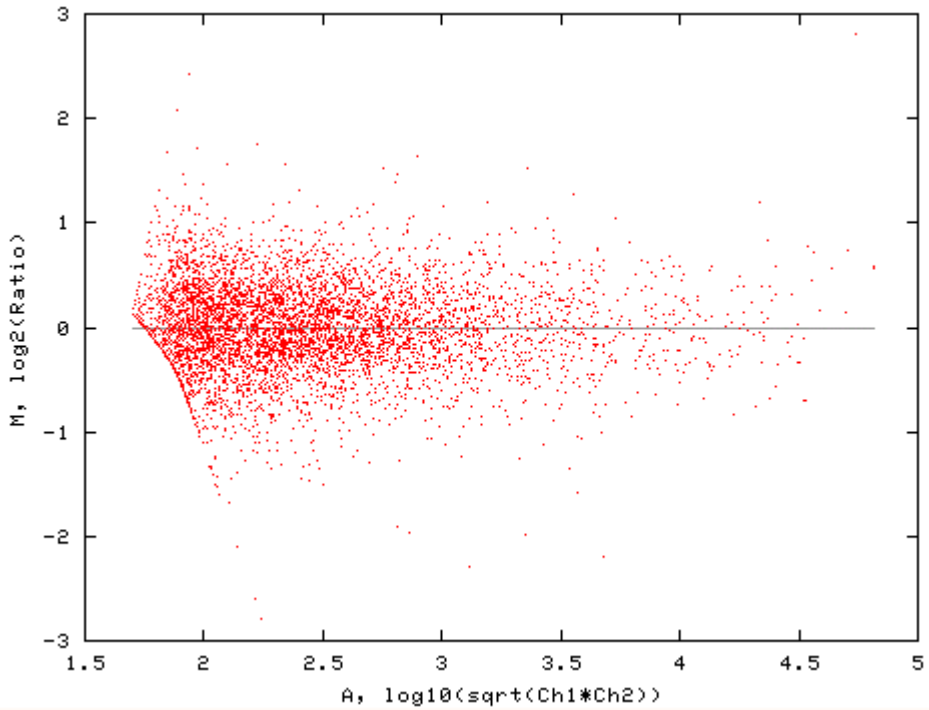
- Regional variations
- Up (red) and down (green) regulated genes should be randomly distributed across the slide (but they're not)

Green corner!



Normalization





LOWESS

Normalization – Thoughts

- There are many different ways to normalize data
  - Global median, LOWESS, LOESS, etc
  - By print tip, spatial, etc
- Choose one wisely
- BUT: don't expect it to fix bad data!
  - Won't make up for lack of replicates
  - Won't make up for horrible slides

Average Fold Ratios – Why?

- You don't care about *spots* up or down regulated
- You care about *genes* up or down regulated
- Data is highly variable, so do a lot of replicates

Fibroblast growth factor 9 (20 repl)

T	C	Ratio ▽	log2
2485.3	7729.6	0.32	-1.64
1468.7	3989.8	0.37	-1.44
1134.0	2986.4	0.38	-1.40
335.38	837.71	0.40	-1.32
3037.6	7517.7	0.40	-1.31
762.22	1718.4	0.44	-1.17
1322.1	2649.0	0.50	-1.00
400.72	718.72	0.56	-0.84
1637.1	2732.0	0.60	-0.74
2689.5	4316.8	0.62	-0.68
365.45	573.39	0.64	-0.65
1536.3	2366.7	0.65	-0.62
2520.0	3838.5	0.66	-0.61
2749.4	3721.3	0.74	-0.44
1047.7	1304.4	0.80	-0.32
1524.7	1868.4	0.82	-0.29
3207.2	3454.1	0.93	-0.11
5443.6	5753.6	0.95	-0.08
1376.3	1321.3	1.04	0.06
942.51	772.20	1.22	0.29

Average Fold Ratios

Per-spot ratios

Ch1 int	Ch2 int	Ratio ▽	log2
2485.3	7729.6	0.32	-1.64
1468.7	3989.8	0.37	-1.44
1134.0	2986.4	0.38	-1.40
335.38	837.71	0.40	-1.32
3037.6	7517.7	0.40	-1.31
762.22	1718.4	0.44	-1.17
1322.1	2649.0	0.50	-1.00
400.72	718.72	0.56	-0.84
1637.1	2732.0	0.60	-0.74
2689.5	4316.8	0.62	-0.68
365.45	573.39	0.64	-0.65
1536.3	2366.7	0.65	-0.62
2520.0	3838.5	0.66	-0.61
2749.4	3721.3	0.74	-0.44
1047.7	1304.4	0.80	-0.32
1524.7	1868.4	0.82	-0.29
3207.2	3454.1	0.93	-0.11
5443.6	5753.6	0.95	-0.08
1376.3	1321.3	1.04	0.06
942.51	772.20	1.22	0.29

Per-gene ratios

myd88 fold ratio △	#	Reporter ID	Cluster	Gene	Reporter name
6.32166	1	M011911_01	Mm.157750	2400003M24Rik	pro-platelet basic protein
5.03461	2	M001974_01	Mm.5034	Pde7a	phosphodiesterase 7A
4.80875	3	M003900_01	Mm.25045	Vav2	Vav2 oncogene
4.70454	4	M009551_01	Mm.35830	Hbb-y	hemoglobin Y, beta-like embryonic chain
4.63545	5	M012310_01	Mm.159023	1700070G05Rik	RIKEN cDNA 1700070G05
4.43047	6	M013108_01	Mm.178707	1700011L22Rik	
4.28467	7	M009631_01	Mm.37922	4933405A16Rik	RIKEN cDNA 4933405A16
4.05637	8	M006174_01	Mm.116788	Gdf2	growth/differentiation factor-2
4.02236	9	M011938_01	Mm.158116	2310020N23Rik	RIKEN cDNA 2310020N23
3.8972	10	M011359_01	Mm.121197	5730402C15Rik	
3.88726	11	M003809_01	Mm.23895	Tekt2	tektin 2
3.86665	12	M005522_01	Mm.57212	Adra2a	adrenergic receptor, alpha 2a
3.4735	13	M012406_01	Mm.159221	5430405H02Rik	
3.47212	14	M001926_01	Mm.4956	Fgf4	fibroblast growth factor 4
3.36209	15	M000603_01	Mm.1641	Car4	carbonic anhydrase 4

Fold ratio average

Statistical Significance – Why?

- Which gene is more likely to be down regulated?
  - Fibroblast growth factor 9 – ratio: 0.6
  - ETS-related transcription factor – ratio: 0.6

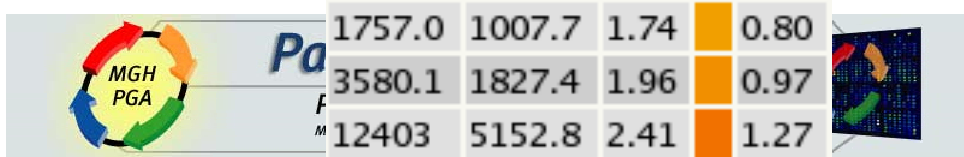
Statistical Significance – Why?

Fibroblast growth factor 9 (20 repl)

T	C	Ratio ▽	log2
2485.3	7729.6	0.32	-1.64
1468.7	3989.8	0.37	-1.44
1134.0	2986.4	0.38	-1.40
335.38	837.71	0.40	-1.32
3037.6	7517.7	0.40	-1.31
762.22	1718.4	0.44	-1.17
1322.1	2649.0	0.50	-1.00
400.72	718.72	0.56	-0.84
1637.1	2732.0	0.60	-0.74
2689.5	4316.8	0.62	-0.68
365.45	573.39	0.64	-0.65
1536.3	2366.7	0.65	-0.62
2520.0	3838.5	0.66	-0.61
2749.4	3721.3	0.74	-0.44
1047.7	1304.4	0.80	-0.32
1524.7	1868.4	0.82	-0.29
3207.2	3454.1	0.93	-0.11
5443.6	5753.6	0.95	-0.08
1376.3	1321.3	1.04	0.06
942.51	772.20	1.22	0.29

ETS-related transcription factor

T	C	Ratio ▽	log2
414.09	2934.2	0.14	-2.82
2062.2	10543	0.20	-2.35
1504.6	3871	0.39	-1.36
741.63	1831.2	0.40	-1.30
5254.1	12294	0.43	-1.23
4781.5	10782	0.44	-1.17
3086.2	6862.7	0.45	-1.15
3820.4	7397.7	0.52	-0.95
3572.1	6085.5	0.59	-0.77
2774.5	4388.4	0.63	-0.66
3935.4	6027.9	0.65	-0.62
1110.5	1177	0.94	-0.08
1717.0	1177.6	1.46	0.54
1475.6	1010.3	1.46	0.55
1539.5	1020.3	1.51	0.59
9997.8	6037.2	1.66	0.73
4927.9	2942.6	1.67	0.74
1757.0	1007.7	1.74	0.80
3580.1	1827.4	1.96	0.97
12403	5152.8	2.41	1.27



Statistical Significance – Why?

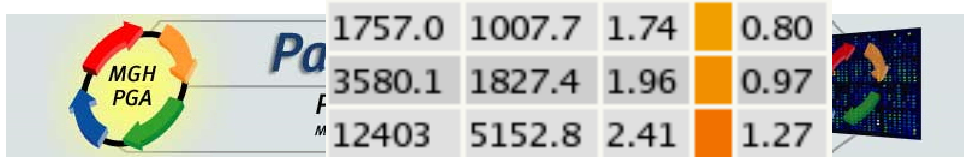
Fibroblast growth factor 9 (20 repl)

T	C	Ratio ▽	log2
2485.3	7729.6	0.32	-1.64
1468.7	3989.8	0.37	-1.44
1134.0	2986.4	0.38	-1.40
335.38	837.71	0.40	-1.32
3037.6	7517.7	0.40	
762.22	1718.4	0.44	
1322.1	2649.0	0.50	
400.72	718.72	0.56	
1637.1	2732.0	0.60	
2689.5	4316.8	0.62	
365.45	573.39	0.64	
1536.3	2366.7	0.65	-0.62
2520.0	3838.5	0.66	-0.61
2749.4	3721.3	0.74	-0.44
1047.7	1304.4	0.80	-0.32
1524.7	1868.4	0.82	-0.29
3207.2	3454.1	0.93	-0.11
5443.6	5753.6	0.95	-0.08
1376.3	1321.3	1.04	0.06
942.51	772.20	1.22	0.29

ETS-related transcription factor

T	C	Ratio ▽	log2
414.09	2934.2	0.14	-2.82
2062.2	10543	0.20	-2.35
1504.6	3871	0.39	-1.36
741.63	1831.2	0.40	-1.30
	2294	0.43	-1.23
	0782	0.44	-1.17
	862.7	0.45	-1.15
	397.7	0.52	-0.95
	085.5	0.59	-0.77
	388.4	0.63	-0.66
	027.9	0.65	-0.62
1110.5	1177	0.94	-0.08
1717.0	1177.6	1.46	0.54
1475.6	1010.3	1.46	0.55
1539.5	1020.3	1.51	0.59
9997.8	6037.2	1.66	0.73
4927.9	2942.6	1.67	0.74
1757.0	1007.7	1.74	0.80
3580.1	1827.4	1.96	0.97
12403	5152.8	2.41	1.27

Same average fold ratio, but the gene on the right has almost as many replicates going up as it does down!



Statistical Significance – Why?

Fibroblast growth factor 9 (20 repl)

T	C	Ratio	log2
2485.3	7729.6	0.32	-1.64
1468.7	3989.8	0.37	-1.44
1134.0	2986.4	0.38	-1.40
335.38	837.71	0.40	-1.32
3037.6	7517.7	0.40	
762.22	1718.4	0.44	
1322.1	2649.0	0.50	
400.72	718.72	0.56	
1637.1	2732.0	0.60	-0.74
2689.5	4316.8	0.62	-0.68
365.45	573.39	0.64	-0.65
			-0.62
			-0.61
			-0.44
			-0.32
			-0.29
3207.2	3454.1	0.93	-0.11
5443.6	5753.6	0.95	-0.08
1376.3	1321.3	1.04	0.06
942.51	772.20	1.22	0.29

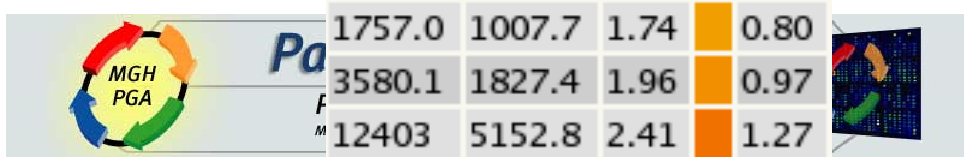
ETS-related transcription factor

T	C	Ratio	log2
414.09	2934.2	0.14	-2.82
2062.2	10543	0.20	-2.35
1504.6	3871	0.39	-1.36
741.63	1831.2	0.40	-1.30
	2294	0.43	-1.23
	0782	0.44	-1.17
	862.7	0.45	-1.15
	397.7	0.52	-0.95
3572.1	6085.5	0.59	-0.77
2774.5	4388.4	0.63	-0.66
3935.4	6027.9	0.65	-0.62
			-0.08
			0.54
			0.55
			0.59
			0.73
4927.9	2942.6	1.67	0.74
1757.0	1007.7	1.74	0.80
3580.1	1827.4	1.96	0.97
12403	5152.8	2.41	1.27

Same average fold ratio, but different P-values!

Probability of null hypothesis via t-test: 0.0012%

Probability of null hypothesis via t-test: 13%



Statistical Significance

Variability data

Ch1 int	Ch2 int	Ratio	log2
2485.3	7729.6	0.32	-1.64
1468.7	3989.8	0.37	-1.44
1134.0	2986.4	0.38	-1.40
335.38	837.71	0.40	-1.32
3037.6	7517.7	0.40	-1.31
762.22	1718.4	0.44	-1.17
1322.1	2649.0	0.50	-1.00
400.72	718.72	0.56	-0.84
1637.1	2732.0	0.60	-0.74
2689.5	4316.8	0.62	-0.68
365.45	573.39	0.64	-0.65
1536.3	2366.7	0.65	-0.62
2520.0	3838.5	0.66	-0.61
2749.4	3721.3	0.74	-0.44
1047.7	1304.4	0.80	-0.32
1524.7	1868.4	0.82	-0.29
3207.2	3454.1	0.93	-0.11
5443.6	5753.6	0.95	-0.08
1376.3	1321.3	1.04	0.06
942.51	772.20	1.22	0.29

P-values

myd88 fold ratio	myd88 p-value	#	Reporter ID	Cluster	Gene	Reporter name
6.32166	0.0283105	1	M011911_01	Mm.157750	2400003M24Rik	pro-platelet basic protein
5.03461	0.0957318	2	M001974_01	Mm.5034	Pde7a	phosphodiesterase 7A
4.80875	0.0141359	3	M003900_01	Mm.25045	Vav2	Vav2 oncogene
4.70454	0.00544169	4	M009551_01	Mm.35830	Hbb-y	hemoglobin Y, beta-like embryonic chain
4.63545	0.0703521	5	M012310_01	Mm.159023	1700070G05Rik	RIKEN cDNA 1700070G05
4.43047	0.123722	6	M013108_01	Mm.178707	1700011L22Rik	
4.28467	0.00264129	7	M009631_01	Mm.37922	4933405A16Rik	RIKEN cDNA 4933405A16
4.05637	0.00189037	8	M006174_01	Mm.116788	Gdf2	growth/differentiation factor-2
4.02236	0.0778869	9	M011938_01	Mm.158116	2310020N23Rik	RIKEN cDNA 2310020N23
3.8972	0.00547626	10	M011359_01	Mm.121197	5730402C15Rik	
3.88726	0.275014	11	M003809_01	Mm.23895	Tekt2	tektin 2
3.86665	0.0200593	12	M005522_01	Mm.57212	Adra2a	adrenergic receptor, alpha 2a
3.4735	0.18172	13	M012406_01	Mm.159221	5430405H02Rik	
3.47212	0.0356278	14	M001926_01	Mm.4956	Fgf4	fibroblast growth factor 4
3.36209	0.0210486	15	M000603_01	Mm.1641	Car4	carbonic anhydrase 4

Statistical Significance – Thoughts

- There are many different statistical significance metrics
  - T-test (P values), SAM (T values), Wilcoxon RST, ANOVA (F-statistics), many more
- Choose one (or more!) wisely
- BUT: don't let it make decisions for you!
  - There will *always* be false pos/neg hits
  - Ultimately, *biological* significance matters

Statistical Significance #'s – How Should We Use Them?

- To sort and rank data
- To reduce data set of 1000s genes to 10s or 100s
- With annotations and biological insight
- As a guide in selecting which genes to validate more precisely (qPCR)

Analysis Pipeline: Summary

- Filter out bad spots
- Adjust low intensities
- Normalize – correct for non-linearities and dye inconsistencies
- Calculate average ratios and significance values per gene
- Rank, sort, filter, squint, sift data
- Validate results

Analysis Pipeline: Summary

- Filter out bad spots
- Adjust low intensities
- Normalize – and dye inc
- Calculate a significance values per gene
- Rank, sort, filter, squint, sift data
- Validate results

There's no one way to do it—just many variations on a theme!

The Biologist's Creed (adapted from the US Marine Corps)

- This is my microarray analysis pipeline. There are many like it, but this one is mine. It is my life. I must master it as I must master my life. Without me my pipeline is useless. Without my pipeline, I am useless.
- My pipeline and I know that what counts in research is not the P-values we choose, the normalization parameters we pick, or the pretty plots we generate. We know that it is the validations that count. We will validate.

The Biologist's Creed (adapted from the US Marine Corps)

- My pipeline is human, even as I am human, because it is my life. Thus, I will learn it as a brother. I will learn its weaknesses, its strengths, its parts, its statistical assumptions, its usage, its variations, and their effects on my conclusions.
- I will keep my pipeline clean and ready, even as I am clean and ready. We will become part of each other.

The End! – What Have We Covered?

- The path from RNA samples to numeric data
- Typical steps & concerns in “data scrubbing”
- Typical analysis of diagnostic experiments

The End! – What Have We *Not* Covered?

- Different flavors of filters, normalizations and stat. significance metrics (10:45a)
- Analysis of time course & multiple treatment experiments (1:00p)
- Clustering, visualization methods (1:00p)
- Step by step tutorial of software (1:45p)

Acknowledgements

MGH Lipid Metabolism Unit
Mason Freeman
Harry Björkbacka

MGH Microarray Core
Glenn Short
Jocelyn Burke
Najib El Messadi
Jason Frietas
Zhiyong Ren

MGH Molecular Biology Bioinformatics Group
Chuck Cooper
Xiaowei Wang

Harvard School of Public Health Biostatistics
Xiaoman Li

BU BioMolecular Engineering Research Center
Temple Smith
Gabriel Eichler
Sean Quinlan
Prashanth Vishwanath