Steps involved in microarray analysis after the experiments

- Scanning slides to create images
- Conversion of images to numerical data
- Processing of raw numerical data
- Further analysis
  - Clustering
  - Integration with genomic data

Steps involved in microarray analysis after the experiments

- Scanning slides to create images
- Conversion of images to numerical data
- Processing of raw numerical data
- Further analysis
  - Clustering
  - Integration with genomic data

Processing raw microarray data

- Main aims:
  - To identify and reduce the noise found in microarray data
  - To identify differentially expressed genes/fragments in an experiment

Methods used so far…

- Many have been *ad hoc*

Methods used so far…

- Many have been *ad hoc*
  - Visual identification (!)

Methods used so far…

- Many have been *ad hoc*
  - Visual identification (!)
  - 2-fold change in hybridisation signals

Methods used so far…

- Many have been *ad hoc*
  - Visual identification (!)
  - 2-fold change in hybridisation signals
  - Ranking ratios of hybridisation signals

Methods used so far…

- Many have been *ad hoc*
  - Visual identification (!)
  - 2-fold change in hybridisation signals
  - Ranking ratios of hybridisation signals
- Microarray expts are 

Methods used so far…

- Many have been *ad hoc*
  - Visual identification (!)
  - 2-fold change in hybridisation signals
  - Ranking ratios of hybridisation signals
- Microarray expts are 

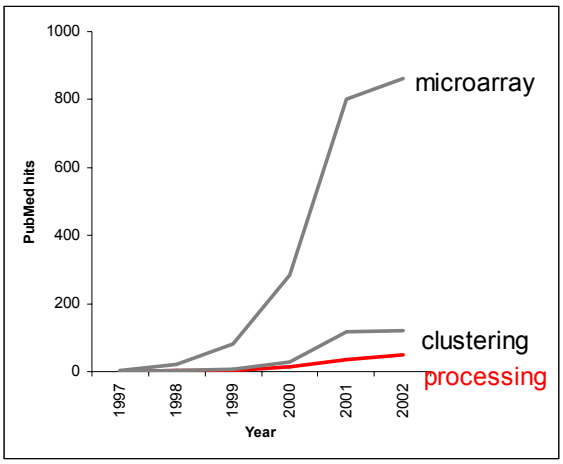
  - Easy to do, but very sensitive
  - Lot of noise, artifacts, errors
  - ~20,000 spots per slide

Methods used so far…

- Many have been *ad hoc*
  - Visual identification (!)
  - 2-fold change in hybridisation signals
  - Ranking ratios of hybridisation signals
- Microarray expts are hard
  - Easy to do, but very sensitive
  - Lot of noise, artifacts, errors
  - ~20,000 spots per slide

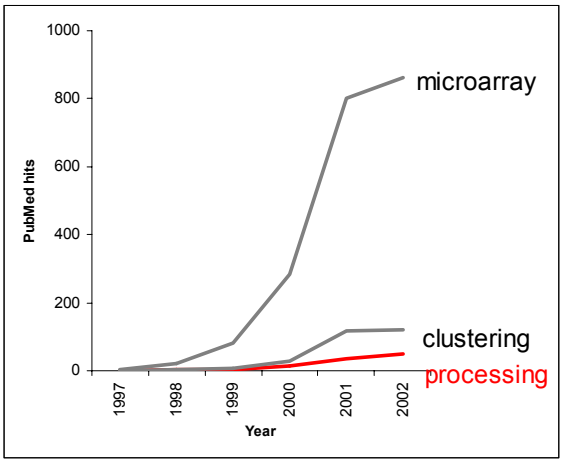
- **Without good processing the results can be completely wrong**

This is changing…..

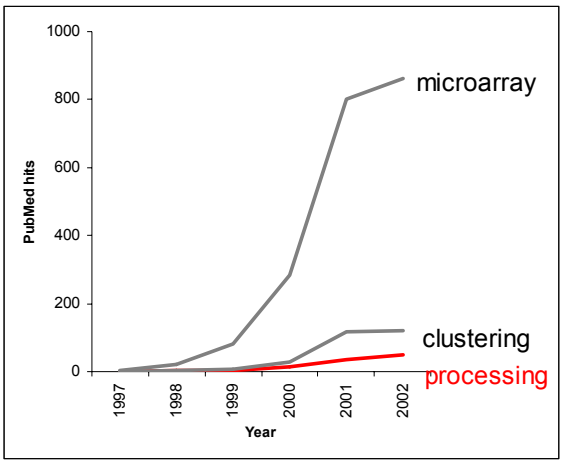


This is changing…..

- Opposing camps

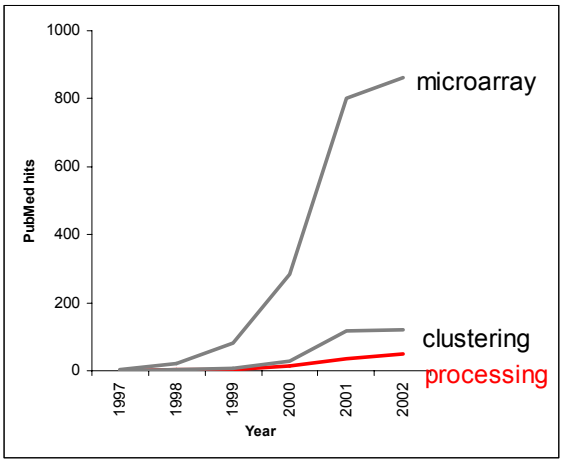


This is changing.....



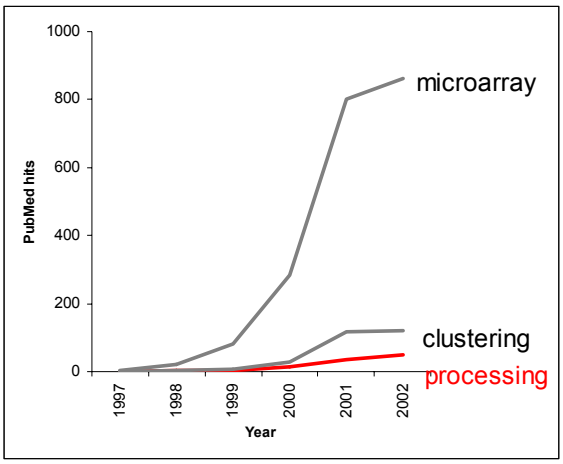
- Opposing camps
  - *Ad hoc* camp
    - Biologists
    - Too simple and may be wrong

This is changing.....



- Opposing camps
  - *Ad hoc* camp
    - Biologists
    - Too simple and may be wrong
  - Complicated maths
    - Biostatisticians
    - Incomprehensible
    - Idealised datasets

This is changing…..



- Opposing camps
  - *Ad hoc* camp
    - Biologists
    - Too simple and may be wrong
  - Complicated maths
    - Biostatisticians
    - Incomprehensible
    - Idealised datasets
  - Must strike a balance

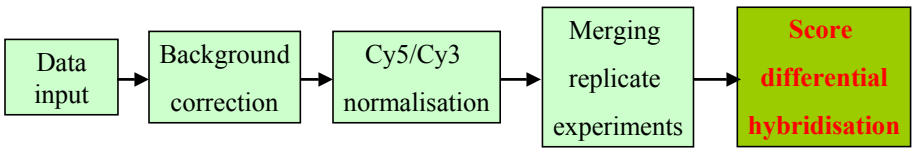
Getting the most out of your microarray

- Processing the raw data
- Cleaning and assessing the quality of your data
- Identifying differentially hybridised spots
- **How do you get the correct list of differentially expressed genes out of ~20000 data points?**

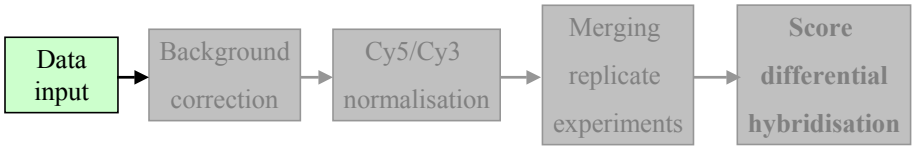
Getting the most out of your microarray

- **Processing the raw data**
- Cleaning and assessing the quality of your data
- Identifying differentially hybridised spots
- **How do you get the correct list of differentially expressed genes out of ~20000 data points?**

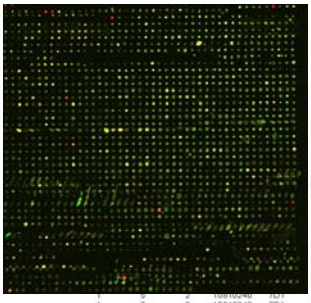
Processing flow chart



Processing flow chart



The data – a GenePix file

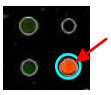


x	y	diameter	635_median	f635_mean	f635_sd	635_median	b635_mean	b635_sd	_gt_b635_r_gt_b635_r5_percent_532_media
2400	2760	140	65400	695	204	639	658	209	23	5	0	809
2650	2760	140	65400	674	194	634	647	190	24	5	0	870
2900	2760	140	613	642	200	645	660	217	13	3	0	825
3150	2760	140	609	627	173	624	735	1513	0	0	0	810
3400	2760	140	569	613	237	595	703	1509	0	0	0	705
3650	2760	140	573	603	201	595	607	181	20	5	0	637
3900	2760	120	932	947	213	582	802	261	62	23	0	762
4140	2790	130	985	987	261	588	808	243	70	33	0	741
4400	2760	140	656	685	280	601	616	181	26	8	0	669
4650	2760	140	587	623	192	591	601	170	23	9	0	675
4900	2760	140	560	570	169	555	573	167	20	2	0	609
5150	2760	140	571	588	191	563	582	168	20	6	0	643
5400	2760	140	575	592	179	588	581	169	16	6	0	677
5650	2760	140	601	600	182	588	578	171	21	3	0	675
5900	2760	140	529	544	180	556	579	181	14	2	0	614
6150	2760	140	609	949	3643	587	620	425	5	3	0	757
2400	2930	140	590	581	179	578	600	187	16	1	0	636
2650	2930	140	587	610	182	583	594	178	22	4	0	655
2900	2930	140	604	632	221	602	608	182	22	7	0	778
3150	2930	140	620	625	199	600	607	189	21	5	0	744
3400	2930	140	684	802	1217	595	600	183	33	10	0	792
1	6	2	10810240	7D1	3650	2930	140	664	664	201	590	694	663	0	0	0	748
1	7	2	10817028	7D7	3900	2930	140	558	584	183	589	689	664	0	0	0	648
1	8	2	10817028	7D7	4150	2930	140	554	588	213	577	591	174	19	5	0	600
1	9	2	10849817	7F1	4400	2930	140	573	575	177	576	596	177	14	1	0	630
1	10	2	10849817	7F1	4650	2930	140	588	593	291	570	585	180	15	5	0	610
1	11	2	10856406	7F7	4900	2930	140	577	585	198	559	587	260	8	1	0	584
1	12	2	10856406	7F7	5150	2930	140	600	593	179	566	598	261	10	0	0	618
1	13	2	10886046	7H1	5400	2930	140	566	597	177	578	586	179	19	4	0	696
1	14	2	10886046	7H1	5650	2930	140	598	603	194	581	589	174	19	5	0	667
1	15	2	10894096	7H7	5900	2930	140	567	588	161	566	585	160	17	3	0	636
1	16	2	10894096	7H7	6150	2930	140	608	620	193	582	593	182	21	5	0	832
1	1	3	11257735	11B1	2400	3110	140	590	591	163	567	641	1130	0	0	0	600
1	2	3	11257735	11B1	2650	3110	140	537	557	161	562	578	181	13	1	0	601
1	3	3	11262510	11B7	2900	3110	140	589	590	187	581	587	179	14	4	0	612
1	4	3	11262510	11B7	3150	3110	140	635	656	205	592	608	190	25	7	0	779
1	5	3	11262214	11D1	3400	3110	140	579	597	184	581	603	191	17	5	0	711
1	6	3	11262214	11D1	3650	3110	140	571	585	195	592	594	663	0	0	0	715
1	7	3	11287805	11D7	3900	3110	140	601	634	223	581	679	676	1	0	0	662
1	8	3	11287805	11D7	4150	3110	140	549	582	239	569	586	223	15	4	0	670
1	9	3	11313882	11F1	4400	3110	140	695	727	235	565	576	168	44	22	0	654

GenePix file – what do we look at?

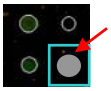
- Measure red and green intensities separately
- Signal intensity = foreground – background

GenePix file – what do we look at?



- Measure red and green intensities separately
- Signal intensity = foreground – background
- Foreground signal

GenePix file – what do we look at?

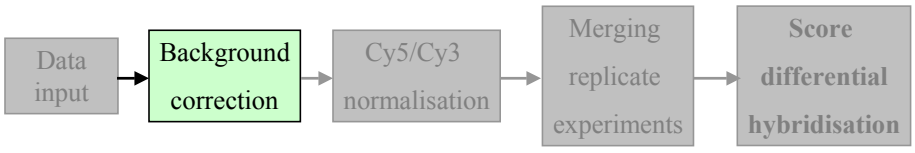


- Measure red and green intensities separately
- Signal intensity = foreground – background
- Foreground signal
- Background signal

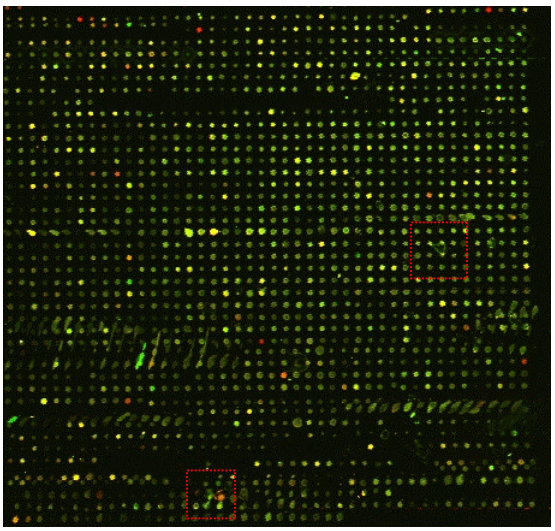
GenePix file – what do we look at?

- Measure red and green intensities separately
- Signal intensity = foreground – background
- Foreground signal
- Background signal
- Ratio = red/green or green/red

Processing flow chart

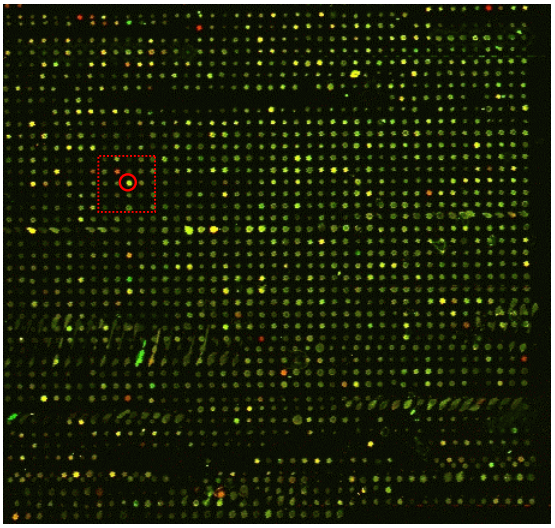


Background correction



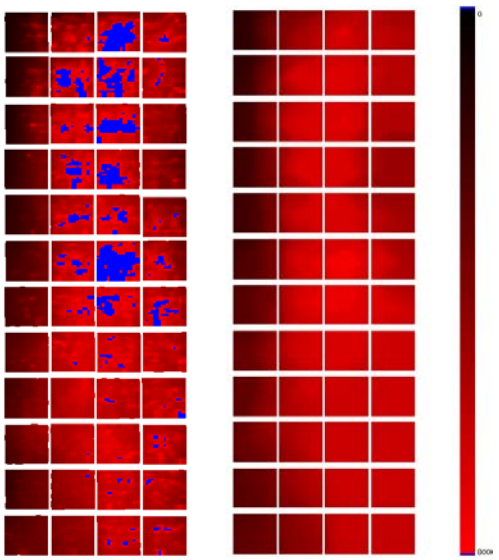
- Genepix background
  - Median intensity of immediate area surrounding each spot
- But
  - Very variable between individual spots
  - Artifactual background from smudges
- Therefore add to noise

Background correction



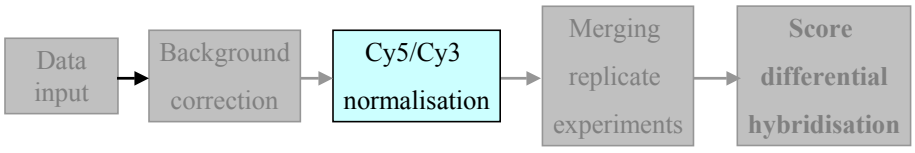
- Calculate the average background from surrounding area of spot
- Recommendation of 3x3 – 5x5 area
- Repeat for red and green separately

Background correction



- Still have variable distribution of intensity
- Much smoother distribution of background intensity
- Remove artifactual smudges

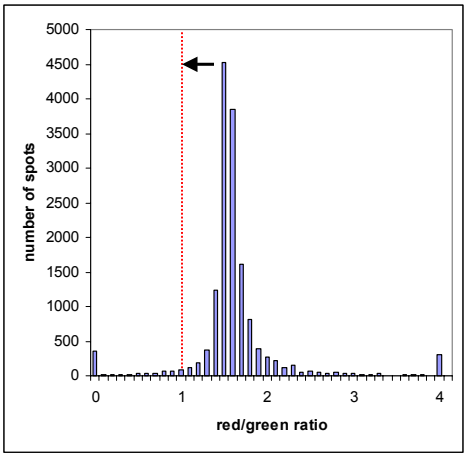
Processing flow chart



Red/green normalisation

- Normalise red and green intensities
  - spots with equal hybridisation should have similar intensities
  - *ie* ratio ~ 1 for similarly expressed genes
- Otherwise, will have wrong list of differentially expressed genes
- Multiply one set of intensities by a scale factor
- **But must obtain scale factor**

Majority method

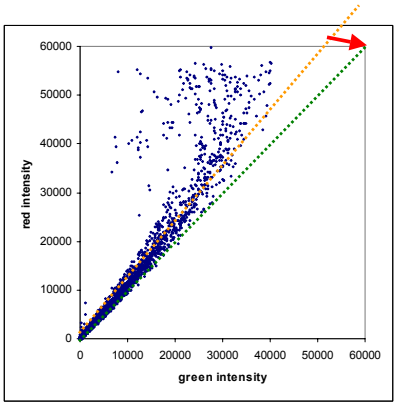


- Majority method:
  - Assume that most spots do not change expression level
  - Find average ratio (red/green intensity)
  - Scale factor is the amount by which need to multiply the ratio so it is 1.

Majority method

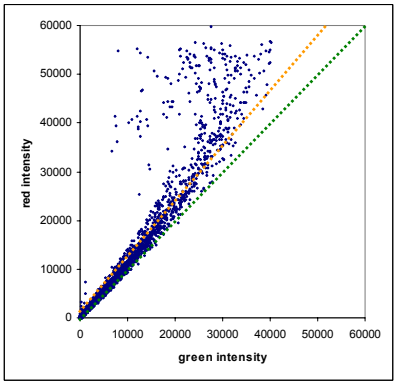
- But several issues
  - Hybridisation levels differ according to location on slide
  - Scanning properties for different colours differ at different intensities
- Scale factor must take these into account

Intensity considerations



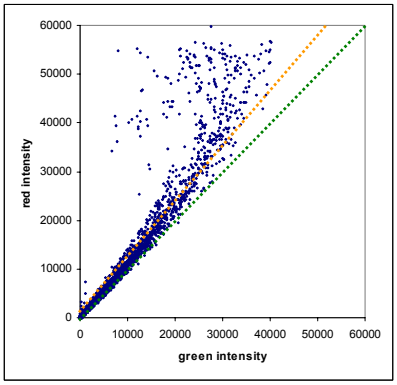
- Simple scale factor fits a straight line

Intensity considerations



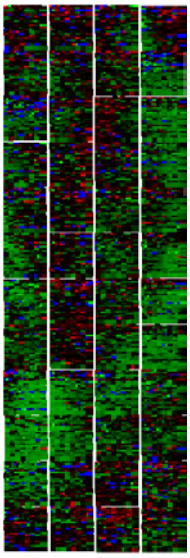
- Simple scale factor fits a straight line
- Distribution is curved
  - Difference in ratio can be 10-fold different depending on intensity

Intensity considerations



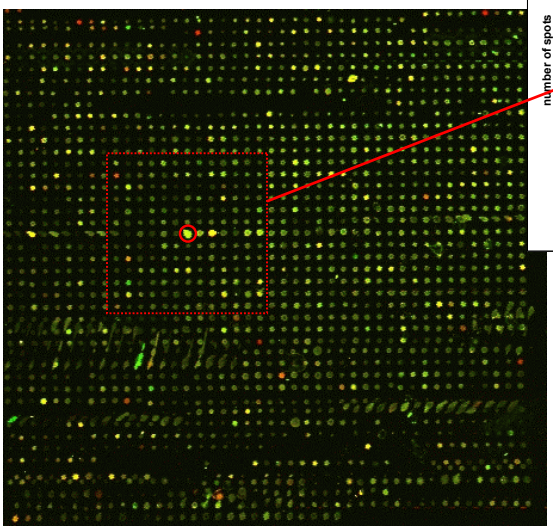
- Simple scale factor fits a straight line
- Distribution is curved
  - Difference in ratio can be 10-fold different depending on intensity
- Different scale factors should be used for different intensities

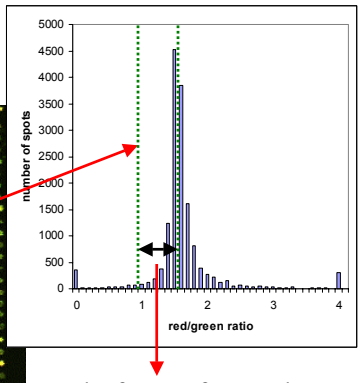
Positional considerations



- Different regions of the slide have different levels of hybridization
- Difference in average ratio can be ~10 fold
- Different scale factors needed for each region of slide

Positional considerations

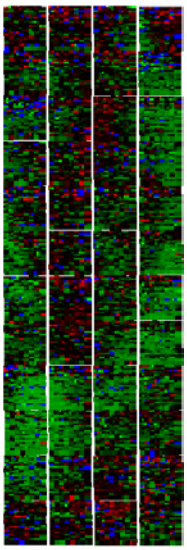




Scale factor for each spot

- Calculate the scale factor using surrounding area of spot
- Recommendation of 12x12 – 20x20 area

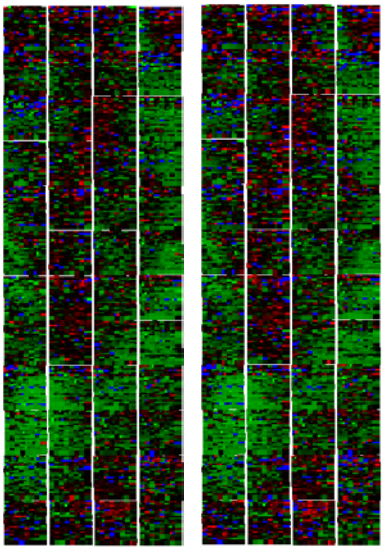
Positional considerations



Before

- Raw data has large positional dependence

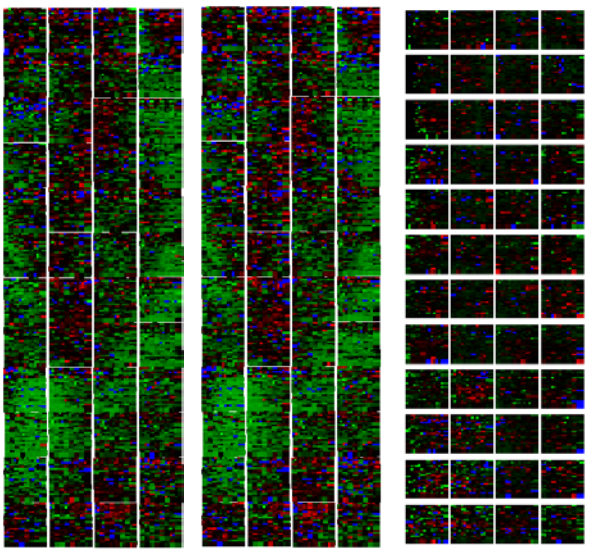
Positional considerations



Before No positional data

- Raw data has large positional dependence
- Normalisation without pos. shifted ratios towards red intensity, but does not remove artifact

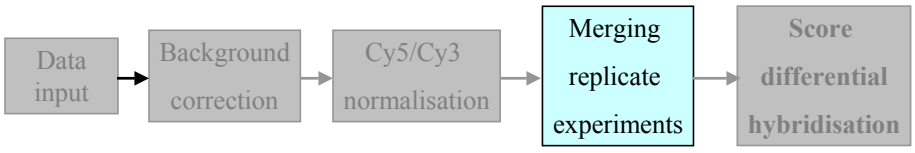
Positional considerations



Before No positional data Positional data

- Raw data has large positional dependence
- Normalisation without pos. shifted ratios towards red intensity, but does not remove artifact
- Positional normalisation removes most of the artifact

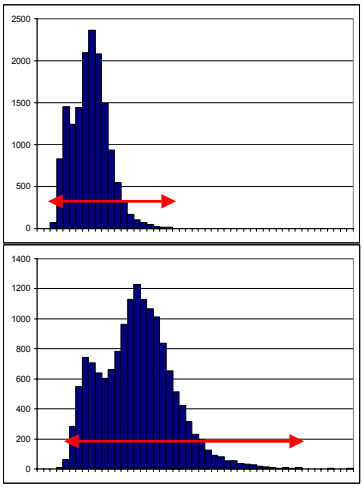
Processing flow chart



Combining multiple experiments

- Often have replicates of the same experiment
- Do you have to scale between them?
- How do you combine the data from them?

Replicate scaling



- Different slides may have different spreads in intensity and ratios
- Adjust spread of distributions by measuring standard deviation
- Estimate of variance by quartiles, fitting distribution, or bootstrapping

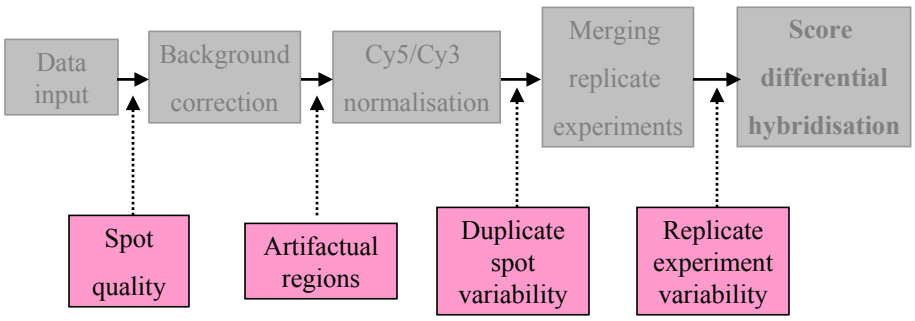
Combining replicate data

- Take medians of ratios?
- Take means of intensity values?
- Take weighted means?
- Treat each experiment individually and see which spots are consistently differntially expressed?

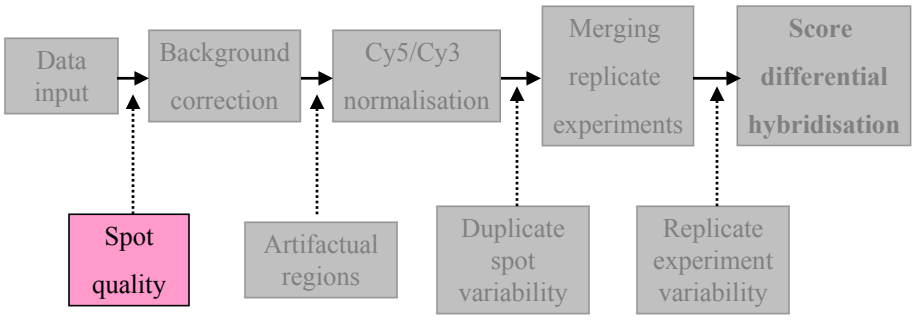
Getting the most out of your microarray

- Processing the raw data
- **Cleaning and assessing the quality of your data**
- Identifying differentially hybridised spots
- **How do you get the correct list of differentially expressed genes out of ~20000 data points?**

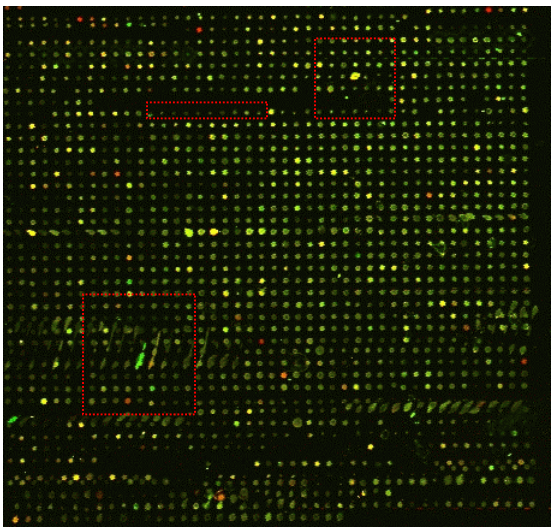
Processing flow chart



Processing flow chart



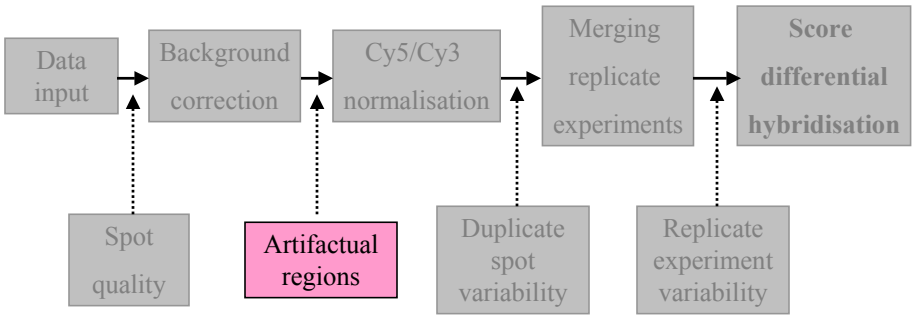
Filter bad spots



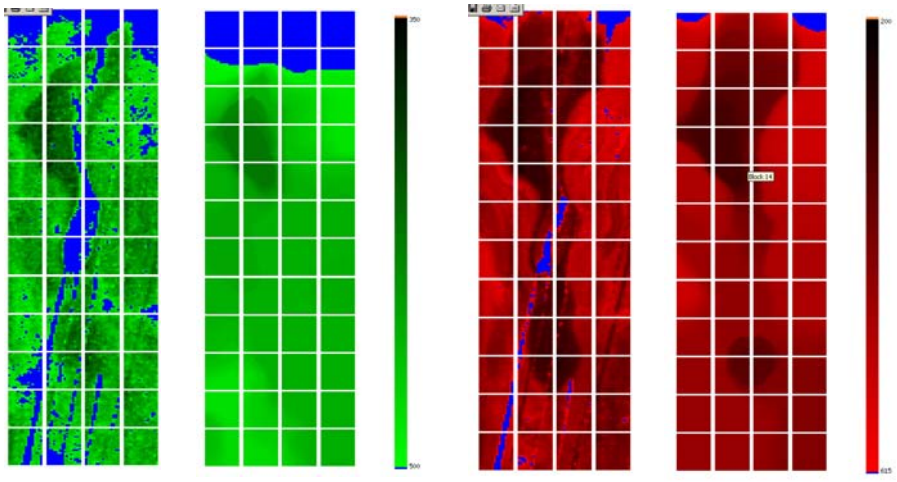
- Small spots
- Smudgy spots
- Non-round spots

etc…

Processing flow chart



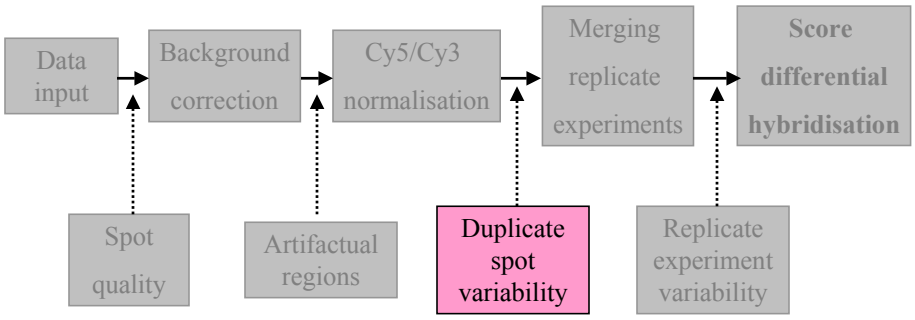
Artifactual array regions



Remove regions that have artifactual background after correction

Background artifacts usually vary from slide to slide – ie not consistent

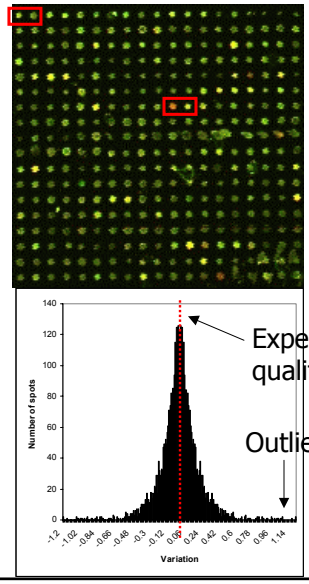
Processing flow chart



Measurement of chip quality

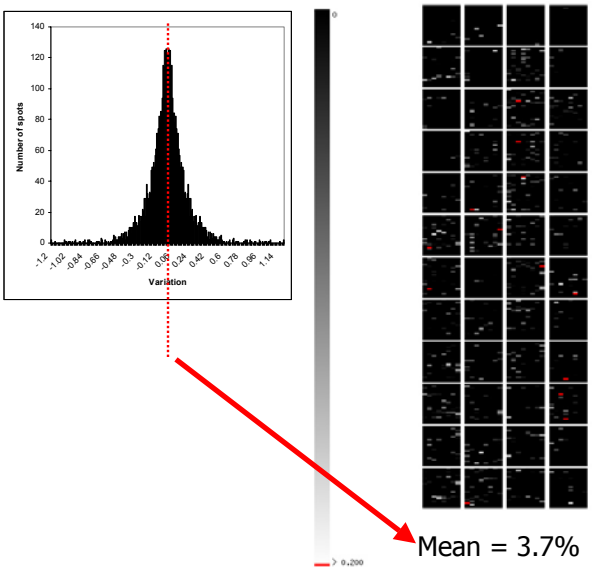
- How successful is an experiment?
  - How consistent are the hybridisations within experiments

Measurement of **intra**chip quality

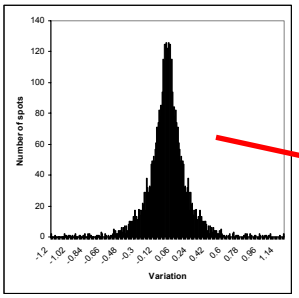


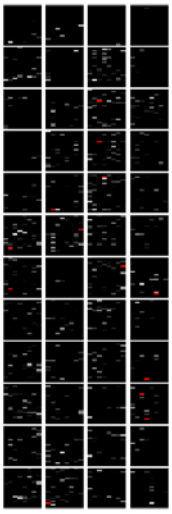
- Genes are often placed as neighbouring pairs
- Variation between them provides a measure of variation within an experiment
  - (x1 – x2)/(x1+x2)
- Mean gives overall variation for expt.
- Remove outliers
- Are the same spots always inconsistent across replicate expts?

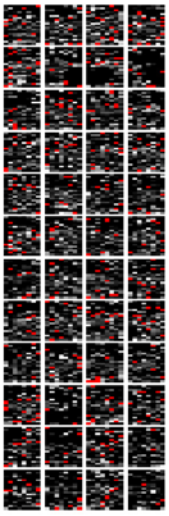
Intrachip variability



Intrachip variability – single experiment quality



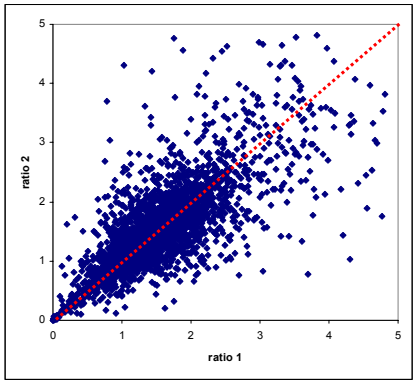




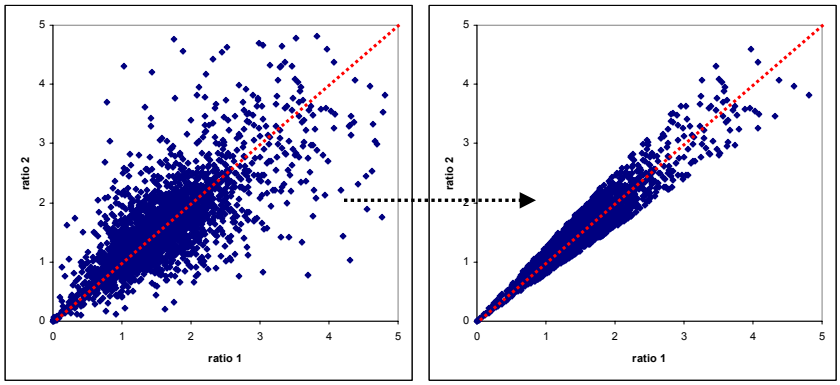
Mean = 3.7%

Mean = 10.7%

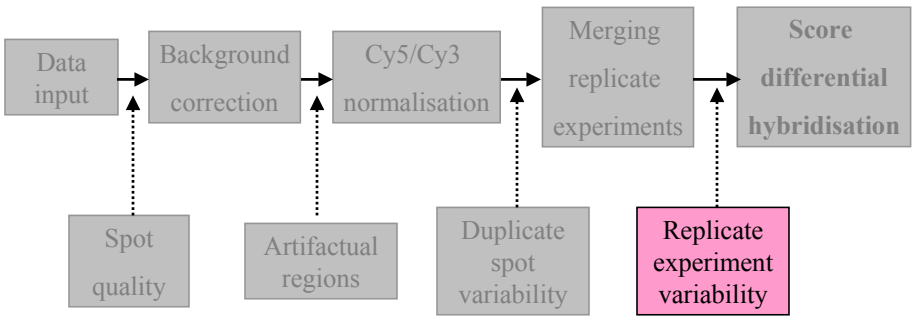
Filtering poor duplicates



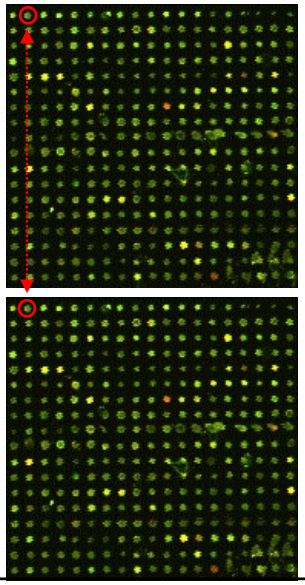
Filtering poor duplicates



Processing flow chart

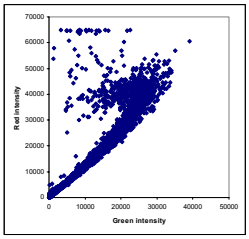


Measurement of **inter**chip quality

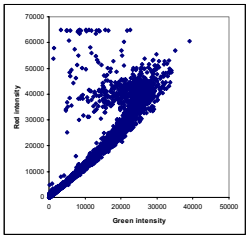


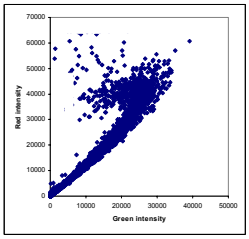
- How consistent are replicate experiments?
- Use same measure for equivalent spots :
  (x1 – x2)/(x1+x2)
- Mean gives overall variation for expt. With respect to other replicates
- Remove outlier experiments

Measurement of replicate chip quality

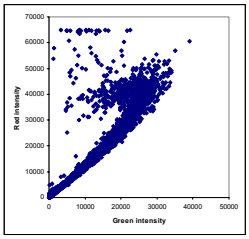


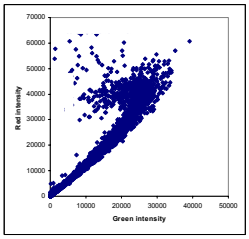
Measurement of replicate chip quality

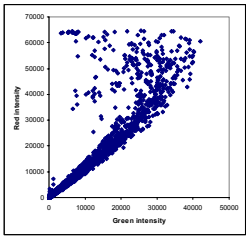




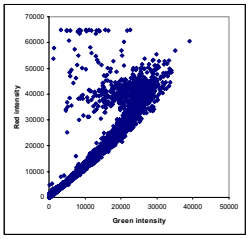
Measurement of replicate chip quality

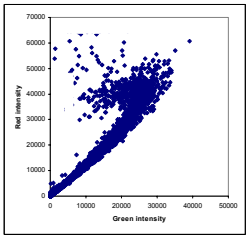


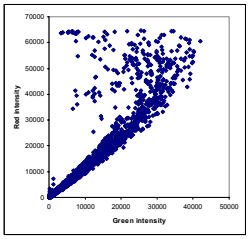


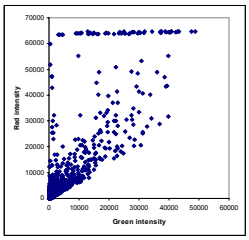


Measurement of replicate chip quality

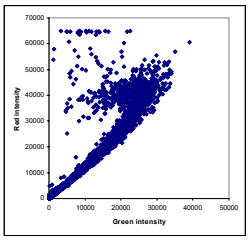


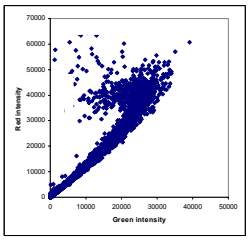


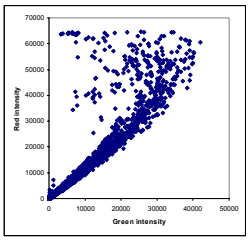


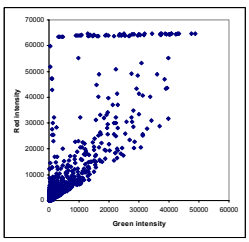


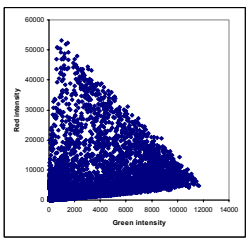
Measurement of replicate chip quality



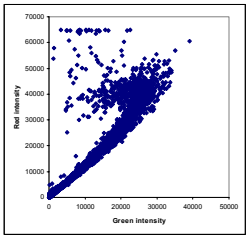




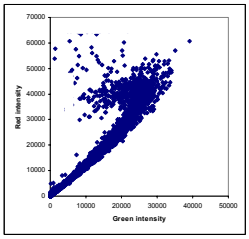




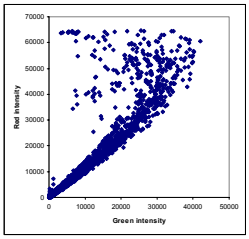
Measurement of replicate chip quality



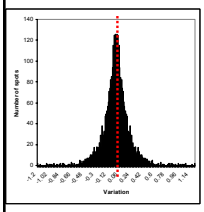
Mean var. = 7.3%

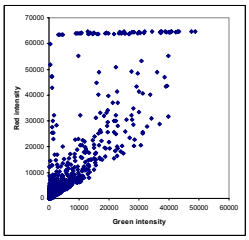


Mean var. = 8.2%

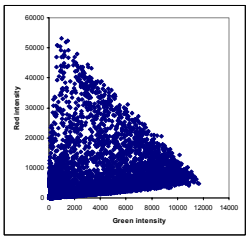


Mean var. = 8.4%



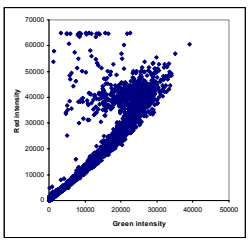


Mean var. = 15.2%

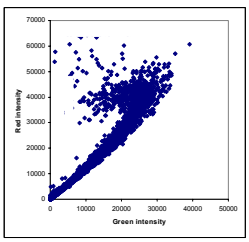


Mean var. = 32.3%

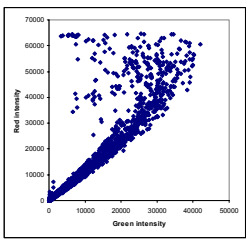
Measurement of replicate chip quality



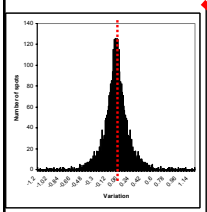
Mean var. = 7.3%

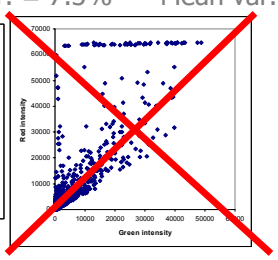


Mean var. = 8.2%

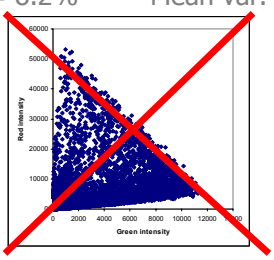


Mean var. = 8.4%





Mean var. = 15.2%



Mean var. = 32.3%

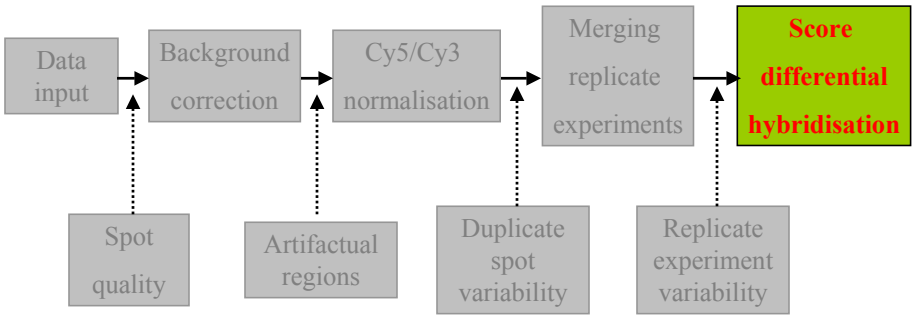
Measurement of **inter**chip quality

- Quite easy to see by eye, but provides a systematic and objective method for determining consistency
- Use to measure overall consistency of replicates
- Identify and remove bad replicate expts
- Also identify regions of the slide that are consistently error prone

Getting the most out of your microarray

- Processing the raw data
- Cleaning and assessing the quality of your data
- **Identifying differentially hybridised spots**
- **How do you get the correct list of differentially expressed genes out of ~20000 data points?**

Processing flow chart



Scoring differentially hybridized probes

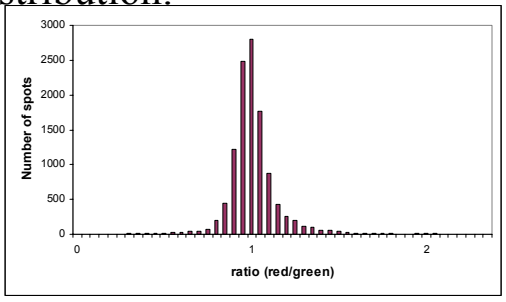
- Identify spots that have differential hybridisation on red and green channels
- Many different methods being published

Scoring differentially: ratio-based method

- Calculate red/green ratio for each spot

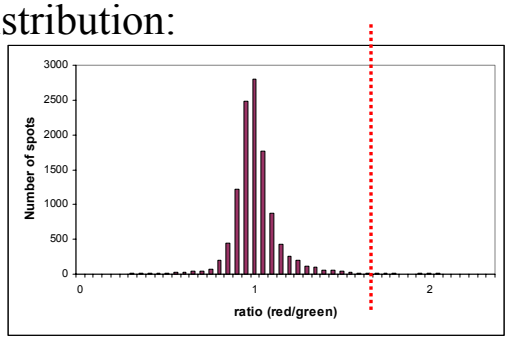
Scoring differentially: ratio-based method

- Calculate red/green ratio for each spot
- Plot distribution:



Scoring differentially: ratio-based method

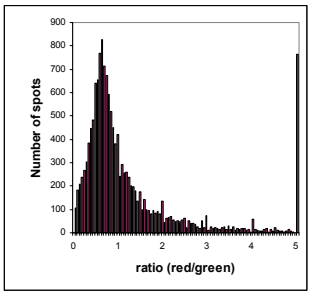
- Calculate red/green ratio for each spot
- Plot distribution:



- Define cut-off based on normal distribution (or use 2-fold cut-off)

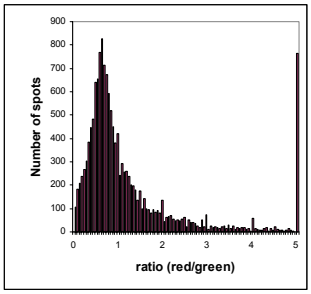
Problems

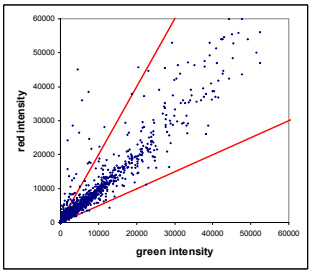
Problems



- Many experiments don't give normal distribution

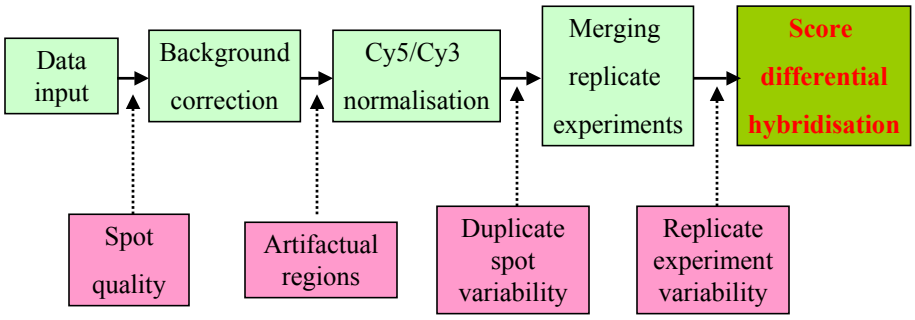
Problems



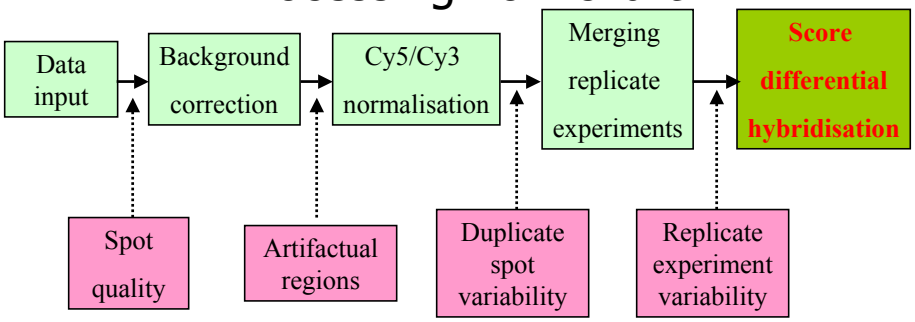


- Many experiments don't give normal distribution
- Ratios ignore the signal intensity
  - More stringent for high intensity spots

Processing flow chart



Processing flow chart



- Raw microarray requires a lot of initial processing before being useful
- Very important as can completely change the answers you get
- The issues are beginning to emerge
  - Different people have different ideas of how to resolve them
  - There is no standard method yet – each has problems
- Very labour intensive, but can be computed relatively easily