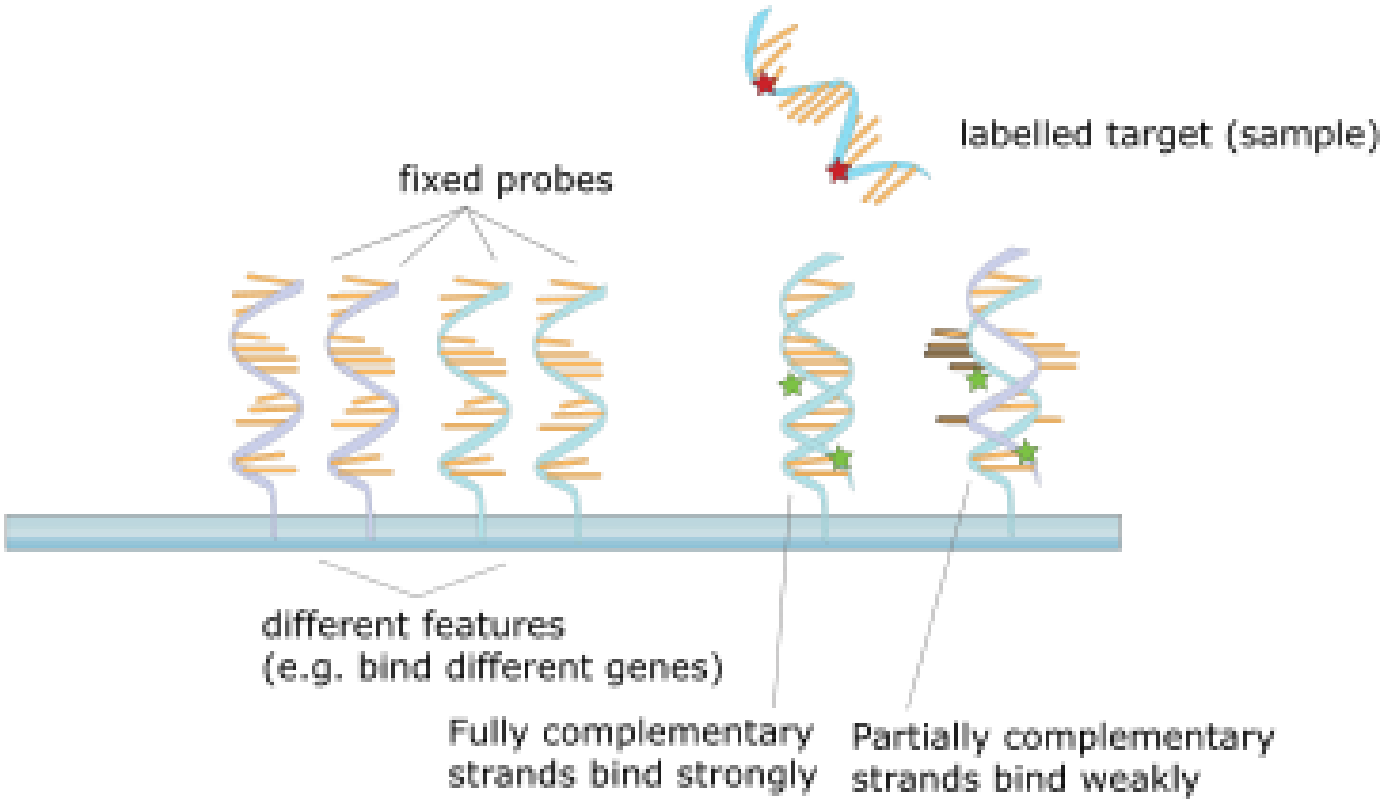
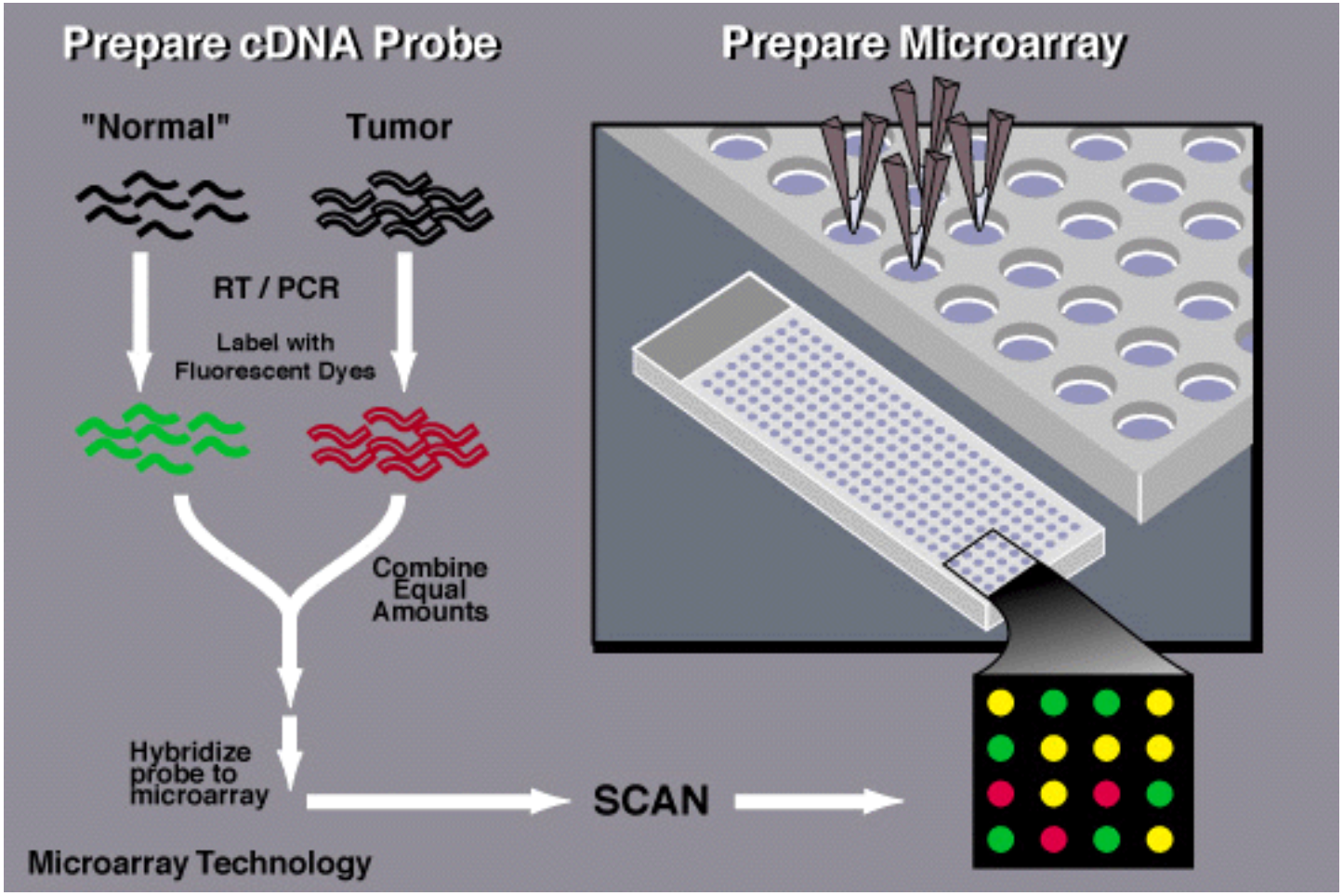
Microarrays



More target in sample = more binding to probe = brighter signal

Two channel


Prepare cDNA Probe
Prepare Microarray
"Normal"
Tumor
RT / PCR
Label with
Fluorescent Dyes
Combine
Equal
Amounts
Hybridize
probe to
microarray
SCAN
Microarray Technology

Q1) GDS3716

- Diseased (ER+/ER-) vs normal in breast cancer

probe ID gene name

samples

ID_REF	IDENTIFIER	GSM512539	GSM512540	GSM512541	GSM512542	GSM512543	GSM512544
1007_s_at	DDR1	2461.400	3435.700	1932.500	2377.700	3055.300	2978.100
1053_at	RFC2	26.700	159.000	31.200	140.700	69.900	98.500
117_at	HSPA6	82.600	243.400	150.200	95.100	209.300	103.400
121_at	PAX8	942.300	897.500	840.800	870.900	685.400	791.800
1255_g_at	GUCA1A	71.800	87.900	75.400	58.100	31.800	40.300
1294_at	UBA7	630.200	571.400	346.300	679.900	1289.300	421.100
1316_at	THRA	186.000	208.700	141.000	135.000	167.500	48.700
1320_at	PTPN21	15.800	18.000	10.600	17.900	51.800	95.400
1405_i_at	CCL5	71.300	26.100	26.300	23.800	100.100	10.900
1431_at	CYP2E1	58.700	45.000	98.300	64.900	56.900	53.000
1438_at	EPHB3	120.800	23.200	113.400	121.000	150.200	107.600
1487_at	ESRRA	343.100	396.500	190.300	260.000	439.600	200.100
1494_f_at	CYP2A6	327.700	438.400	310.000	267.700	248.800	279.600
1598_g_at	GAS6	949.300	1143.800	940.200	1539.900	2623.800	853.500
160020_at	MMP14	540.000	467.900	430.500	424.300	527.000	214.100
1729_at	TRADD	374.300	539.400	301.800	387.100	542.500	241.800
1773_at	FNTB	91.500	52.900	83.300	41.500	30.300	80.600
177_at	PLD1	142.100	170.200	180.300	222.400	135.200	139.700
179_at	PMS2L11	285.400	444.700	254.900	352.600	378.300	387.600
1861_at	BAD	5.300	52.000	5.200	68.600	77.000	43.200

ID_REF	GSM512544	GSM512545	GSM512546	GSM512547	GSM512548	GSM512549	GSM512550
1007_s_at	2978.100	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1053_at	86.700	10[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
117_at	162.700	20[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
121_at	843.100	77[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1255_g_at	51.600	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1294_at	778.10[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1316_at	176.300	24[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1320_at	20[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1405_i_at	12.700	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1431_at	84[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1438_at	96[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1487_at	35[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1494_f_at	242.50[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1598_g_at	15[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
160020_at	504.70[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1729_at	33[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1773_at	60[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
177_at	21[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
179_at	39[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1861_at	13.200	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Microarray Analysis

- Normalisation
  - transformation: compare cancerous to healthy
  - ratio $_{gene}$= cancerous $_{gene}$ / healthy $_{gene}$
    - up-regulation = ratio > 1
    - down-regulation = 0-1
    - `ge3716.getRatio()`

Code

```
from genome import *
ge3716 = readGEOFile('GDS3716.soft')

#empty GeneExpression class
ratio = GeneExpression('GDS3716_ratio')

#example for 1 sample
ratio.addSamples('S1_ER+/Healthy', ge3716.getRatio(33,0))
```

Healthy controls			Cancer patients		
Index	Name	Description	Index	Name	Description
0	#GSM512539	control (RM) sample 1	33	#GSM512557	cancer: ER+ (breast) sample 1
1	#GSM512540	control (RM) sample 2	34	#GSM512558	cancer: ER+ (breast) sample 2
2	#GSM512541	control (RM) sample 3	35	#GSM512559	cancer: ER+ (breast) sample 3
3	#GSM512542	control (RM) sample 4	36	#GSM512560	cancer: ER+ (breast) sample 4
4	#GSM512543	control (RM) sample 5	37	#GSM512561	cancer: ER+ (breast) sample 5
5	#GSM512544	control (RM) sample 6	38	#GSM512562	cancer: ER+ (breast) sample 6
6	#GSM512545	control (RM) sample 7	39	#GSM512563	cancer: ER+ (breast) sample 7
7	#GSM512546	control (RM) sample 8	40	#GSM512564	cancer: ER+ (breast) sample 8
8	#GSM512547	control (RM) sample 9	41	#GSM512565	cancer: ER+ (breast) sample 9
9	#GSM512548	control (RM) sample 10	24	#GSM512566	cancer: ER- (breast) sample 1
10	#GSM512549	control (RM) sample 11	25	#GSM512567	cancer: ER- (breast) sample 2
11	#GSM512550	control (RM) sample 12	26	#GSM512568	cancer: ER- (breast) sample 3
12	#GSM512551	control (RM) sample 13	27	#GSM512569	cancer: ER- (breast) sample 4
13	#GSM512552	control (RM) sample 14	28	#GSM512570	cancer: ER- (breast) sample 5
14	#GSM512553	control (RM) sample 15	29	#GSM512571	cancer: ER- (breast) sample 6
15	#GSM512554	control (RM) sample 16	30	#GSM512572	cancer: ER- (breast) sample 7
16	#GSM512555	control (RM) sample 17	31	#GSM512573	cancer: ER- (breast) sample 8
17	#GSM512556	control (RM) sample 18	32	#GSM512574	cancer: ER- (breast) sample 9

Code

```
from genome import *
ge3716 = readGEOFile('GDS3716.soft')

#empty GeneExpression class
ratio = GeneExpression('GDS3716_ratio')

#example for 1 sample
ratio.addSamples('S1_ER+/Healthy', ge3716.getRatio(33,0))


#set up age-matched sample indices (0: healthy, 1: cancer)
paired_ERpos = [[0,1,2,3,4,5,6,7,8],[33,34,35,36,37,38,39,40,41]]
paired_ERneg = #fill in yourself

#Fill class with ER+ matched samples
i = 0  #sample counter
while i < len(paired_ERpos[0]):
    name = 'S'+str(i+1)+'_ER+/Healthy' #meaningful name for column header
    ratio.addSamples(#fill this in yourself)
    i+=1

#Fill class with ER- matched samples
```

Microarray Analysis

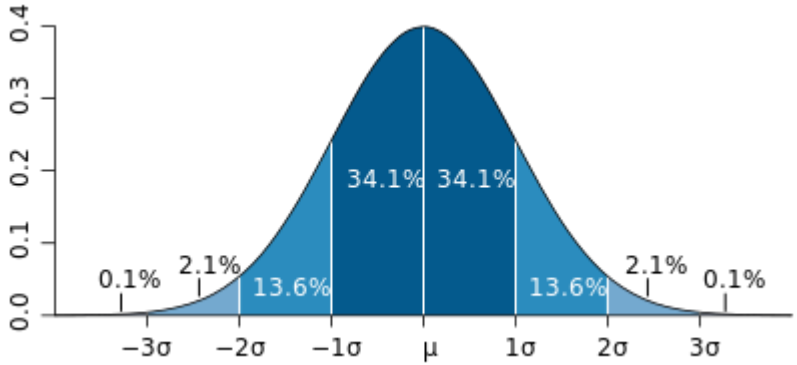
- Normalisation
  - $\log_2$ transformation: easy to compare
    - $\log_2$ (cancerous $_{gene}$ / healthy $_{gene}$)
    - up-regulation = log(ratio) > 0
    - down-regulation = log(ratio) < 0
    - `ge3716.getLogRatio()`

Q1a)

- Why are log2 transformations “useful”?
- Pick 3 probes, e.g.
  - 204531_s_at (BRCA1)
  - 209969_s_at (STAT1)
  - 211300_s_at (TP53)
- For 1st ER+/healthy-matched sample:
  - raw expression values
  - ratio value
  - log2 ratio value

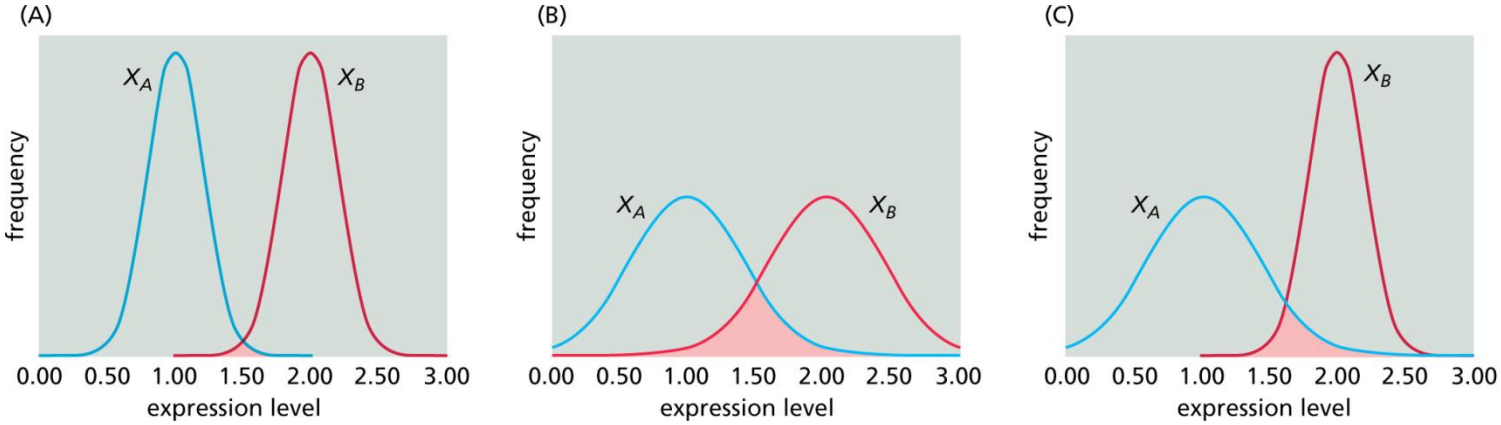
Q1b) & c)

- Histogram for ratio values (all probes, all samples)
- Histogram for log2 values (all probes, all samples)
  - compare to ratio histogram
  - does log2 follow a normal distribution?



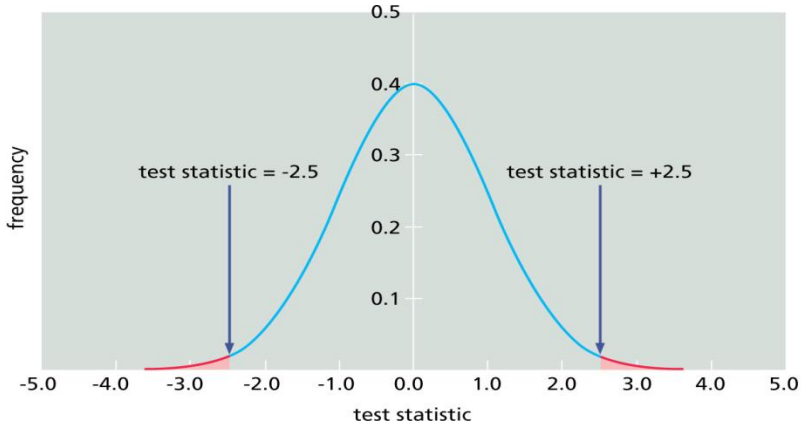
When are two measurements significantly different?

- An expression ratio is significant only if it is big enough or small enough
- A two-fold ratio (for example) is only significant if the variances of the underlying measurements are sufficiently small
- The significance is related to the area of the overlap of the underlying distributions



The *Z*-test

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}.$$



- If the data is approximately normal, convert it to a *Z*-score
  - The *Z*-test is a <u>one</u> location/observation test
  - *X* can be the log expression ratio; μ is then 0
  - σ is the sample standard deviation; *n* is the number of repeats
- The *Z*-score is distributed *N*(0,1) (standard normal).
- The significance level is the area in the tail(s) of the standard normal distribution (*P*-value)

Z-test

```
sdict={}  #dictionary for gene: no. of significant samples

#set up gene dictionary
genes = logratio.getGenes()
for gene in genes: #fill sdict with 0 so can add counter to them
    sdict[gene] = 0

#Z-score calculation for each sample
for #loop through samples to get zscore
    zscores=logratio.getZScore(sample)
    #iterate the dictionary to identify probes
    for key in zscores.iterkeys():
        if #Key condition(Zscore greater than 2 or less than -2)
            sdict[key] +=1 #add to sample count

#display significant genes
for key in sdict:
    if #Key condition (at least 11 sample pairs)
        #print something
```

Q2)

- List of significantly up/down regulated genes
  - Code used to get list
- Explain how Z-score is calculated
  - Use a specific probe (e.g. 214493_s_at) as example

Co-expressed genes: correlation

- GDS38: yeast cell cycle gene expression
  - two channel
  - already transformed
- Focus on gene CLN2
  - what other genes have similar gene expression patterns?

Pearson’s correlation co-efficient

- Cluster analysis method: correlation between expression values of CLN2 and each other gene

-1	0	+1
Negative correlation	No correlation	Positive correlation

Gene correlation with CLN2

```
ge38 = readGEOFile('GDS38.soft', id_column=1)
cln2 = ge38.getGenes('CLN2') #expression for CLN2

#get Pearson co-efficient for CLN2
cln2R = ge38.getPearson(#do something)

#need to get top 5 genes - but cln2R is a dictionary, no order
#function sort() in GeneExpression class can handle this

#convert cln2R to GeneExpression class
gecln2R = GeneExpression(#'name', [#'column header'], cln2R)
gecln2Rsorted = gecln2R.sort(#do something)

#print out top 5 genes
```

Q3a)

- List the top 5 genes with correlated CLN2
  - Include code

Q3b)

- Get GO terms for top 5 genes
  - Include code + output
- Summarise findings (1 statement per gene)
- Are they involved in cell cycle? Justify
- Useful code:

```
#import webservice module
from webservice import *
#get uniprot id
rows=search('taxonomy:4932+AND+gene:CLN2,'uniprot',format='list')
#get GO terms for first search term
terms=getGOTerms(rows[0])
#get definition of GO terms
geneinfo = getGODef(term)
```