



3000788 Intro to Comp Molec Biol

Lecture 9: Transcriptomics technology

September 14, 2023



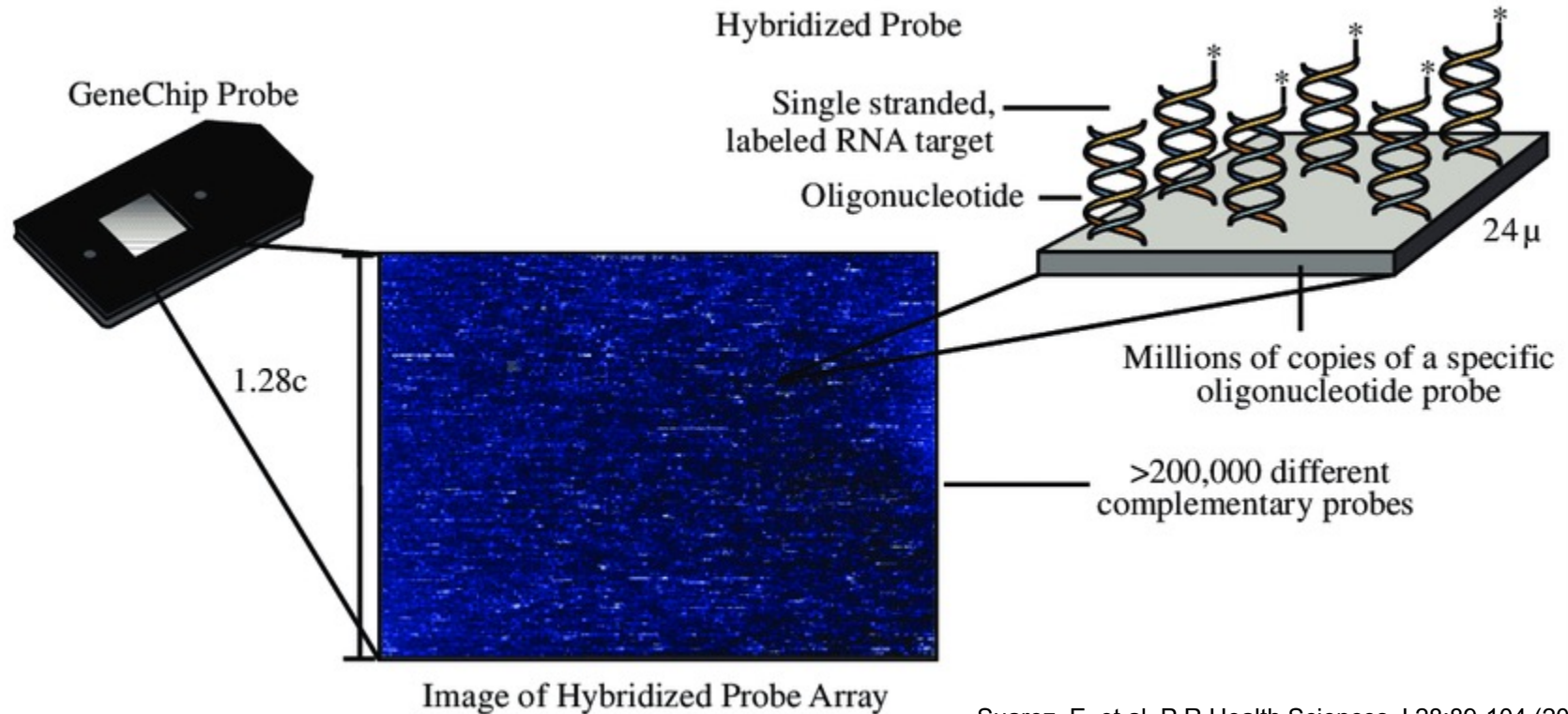
Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

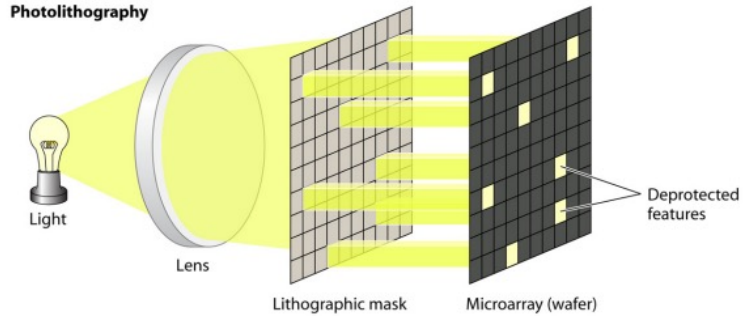


Oligonucleotide microarray

Microarray technology overview

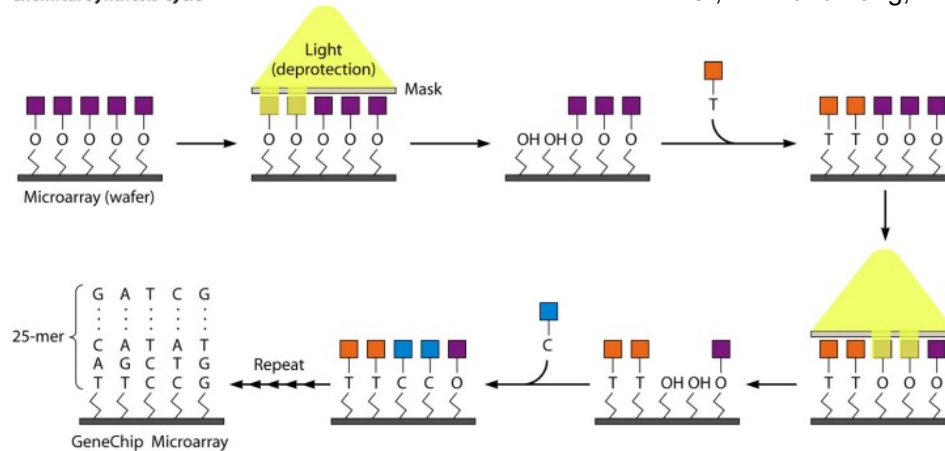


Microarray fabrication

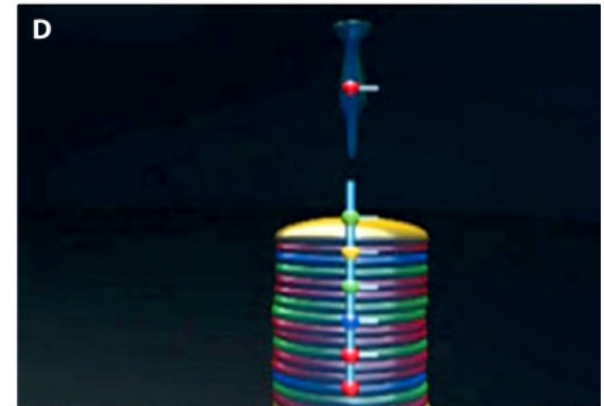
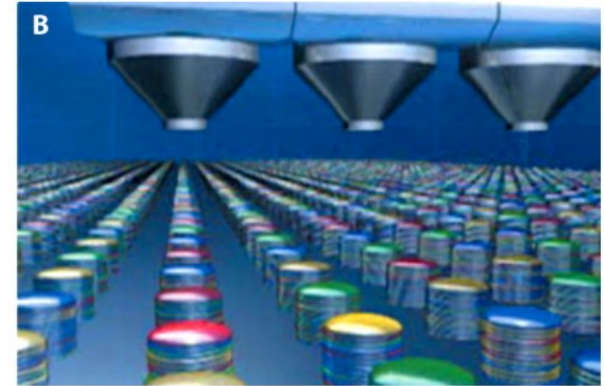
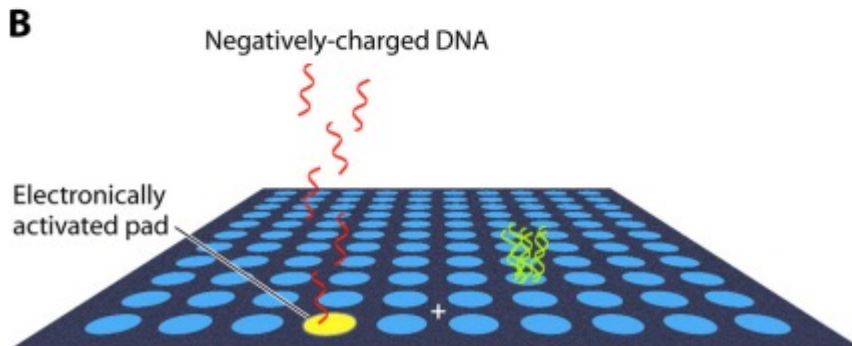
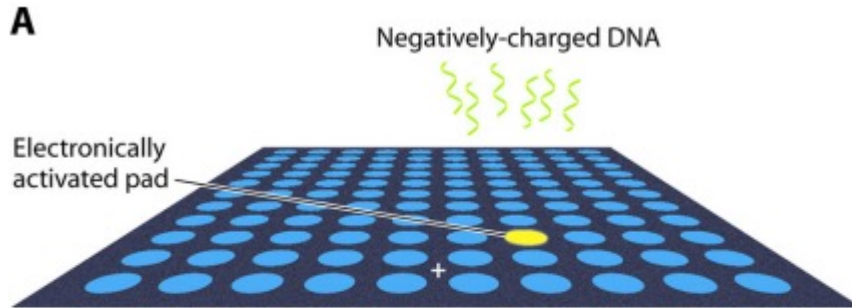


Chemical Synthesis Cycle

Miller, M.B. and Tang, Y.-W. Clin Microbiol Rev 611-633 (2009)



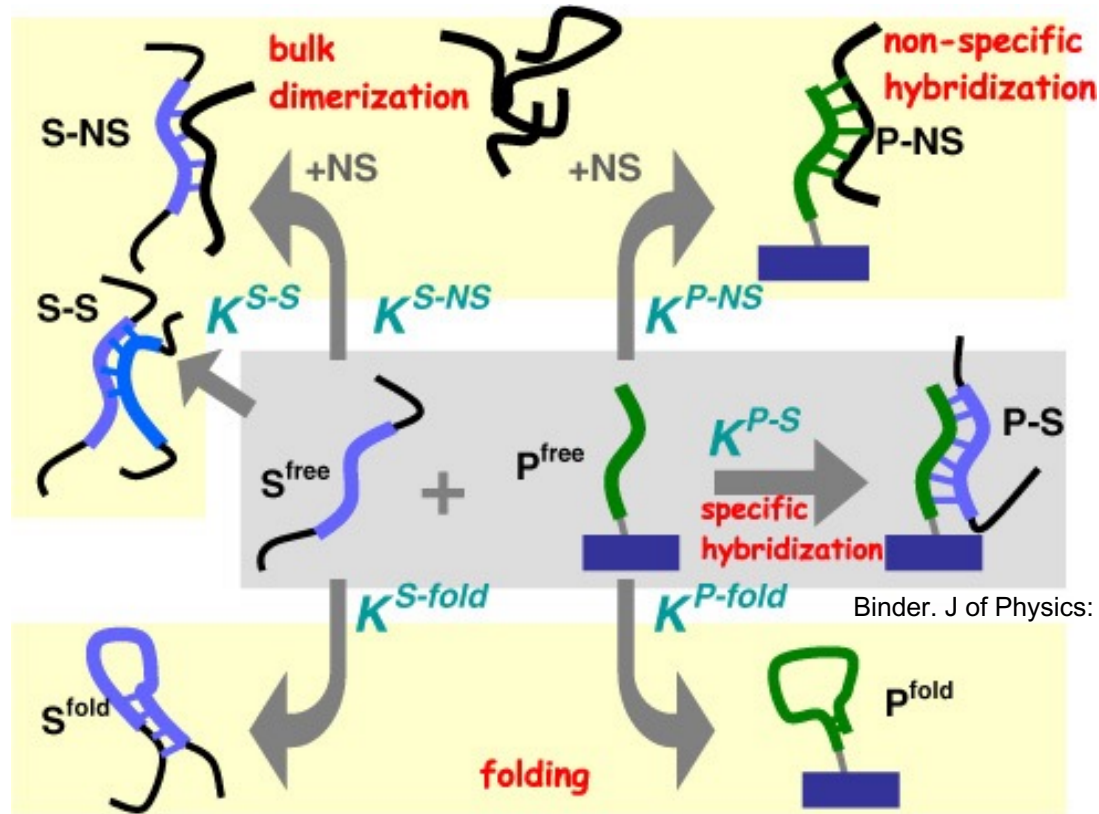
Microarray fabrication





Probe design for microarray

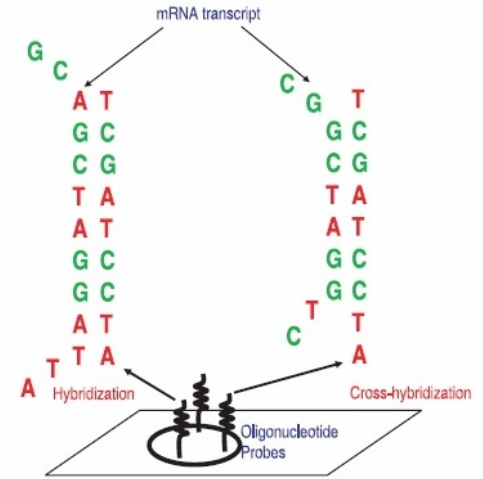
Unwanted probe interactions



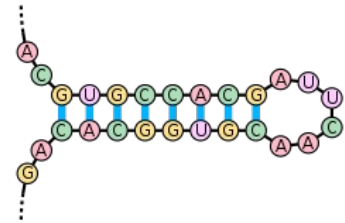
Binder. J of Physics: Condensed Matter 18. (2006)

Probe design principles

- **Sensitivity**
 - Complementary to each target genomic region
 - Multiple probes for each target
- **Specificity**
 - Cross-hybridization: similarity to other targets
 - BLAST
 - Negative control probes
- **Technical issues**
 - Secondary structure: self-complementary
 - Hybridization energy
 - Position-specific bias

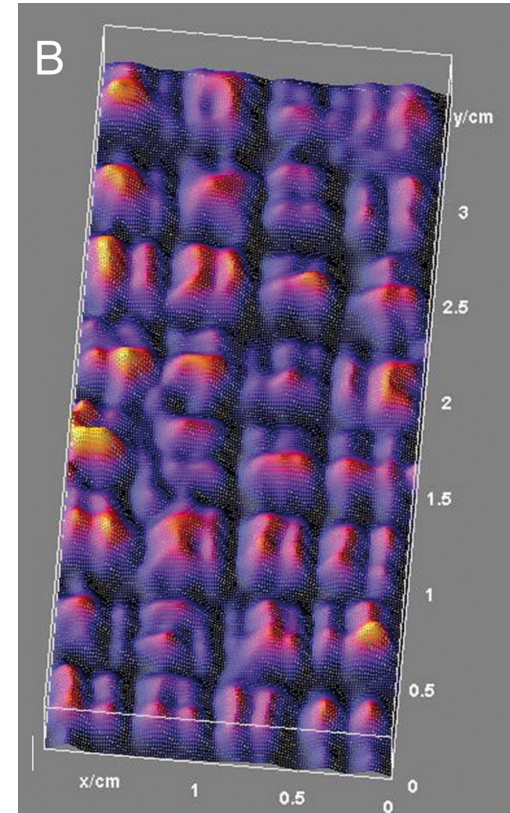
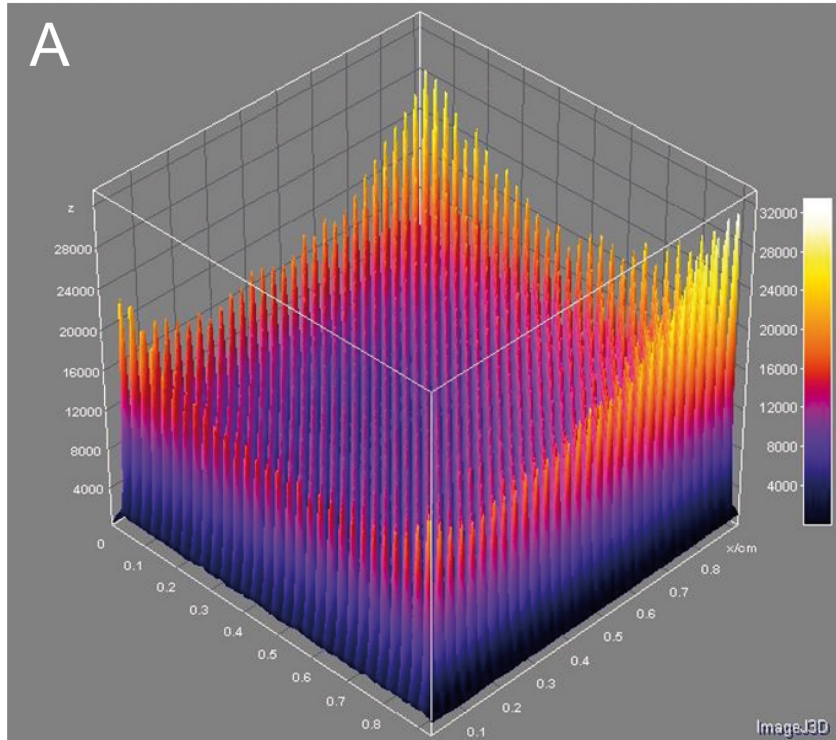


Malutan, R. and Vilda, P.G. Thermodynamics of Microarray Hybridization (2011)

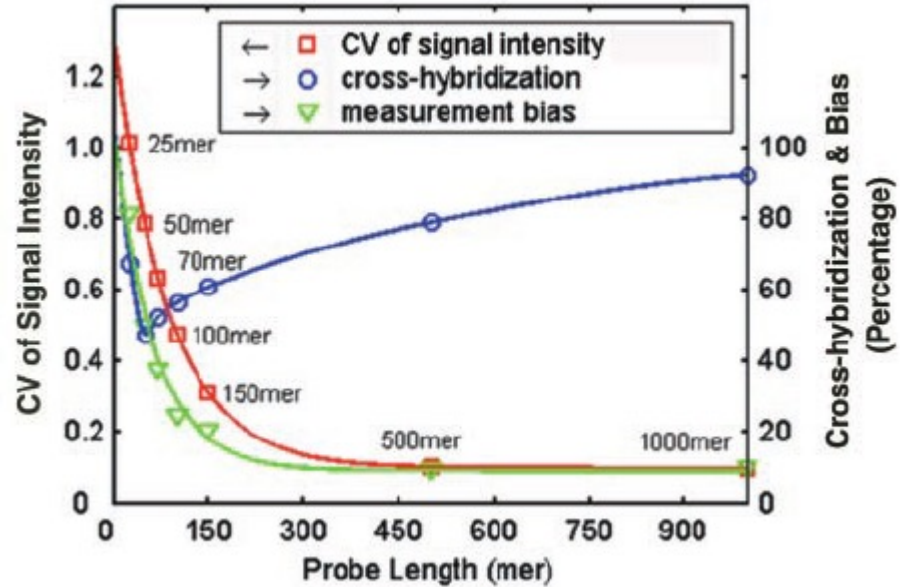
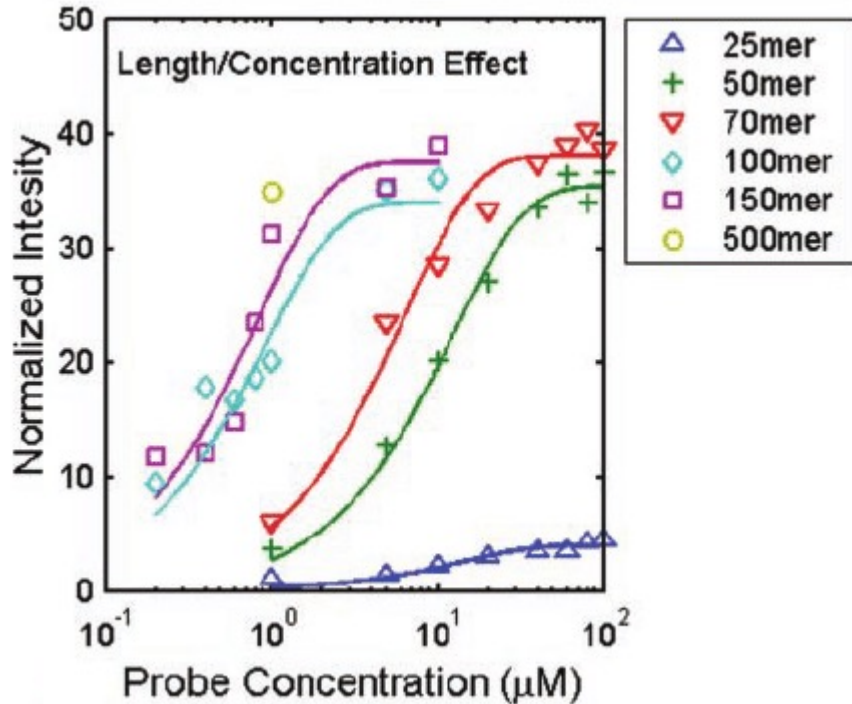


https://en.wikipedia.org/wiki/Complementarity_%28molecular_biology%29

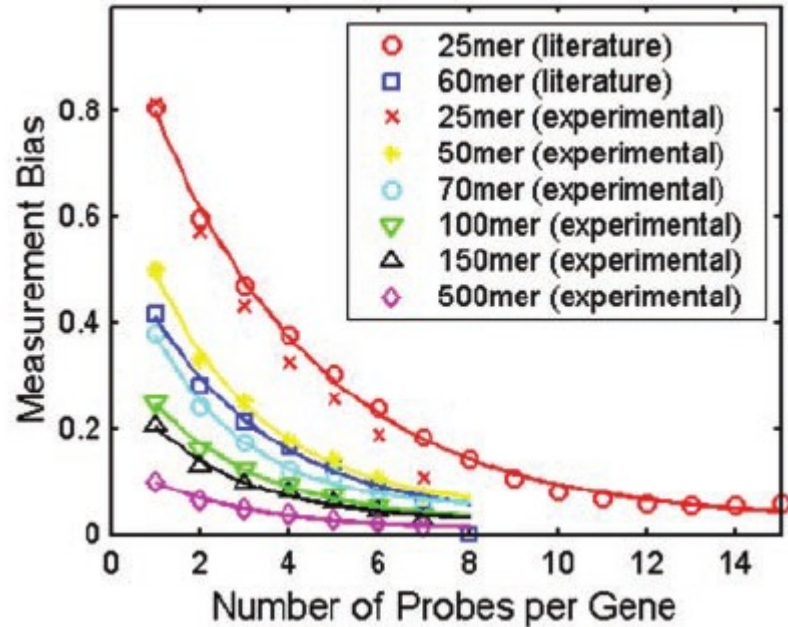
Position-specific intensity bias



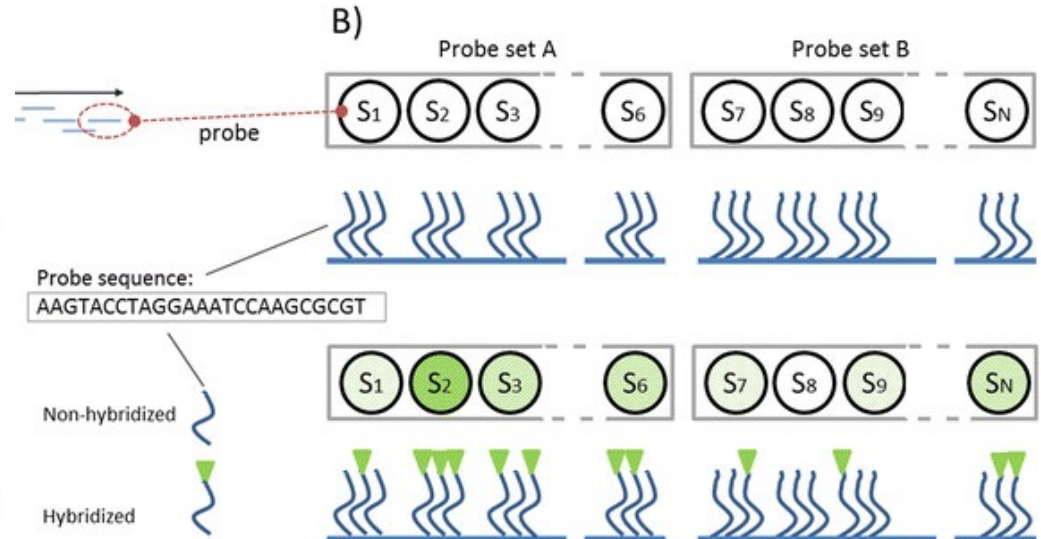
Impact of probe length



Probe set = multiple probes per gene

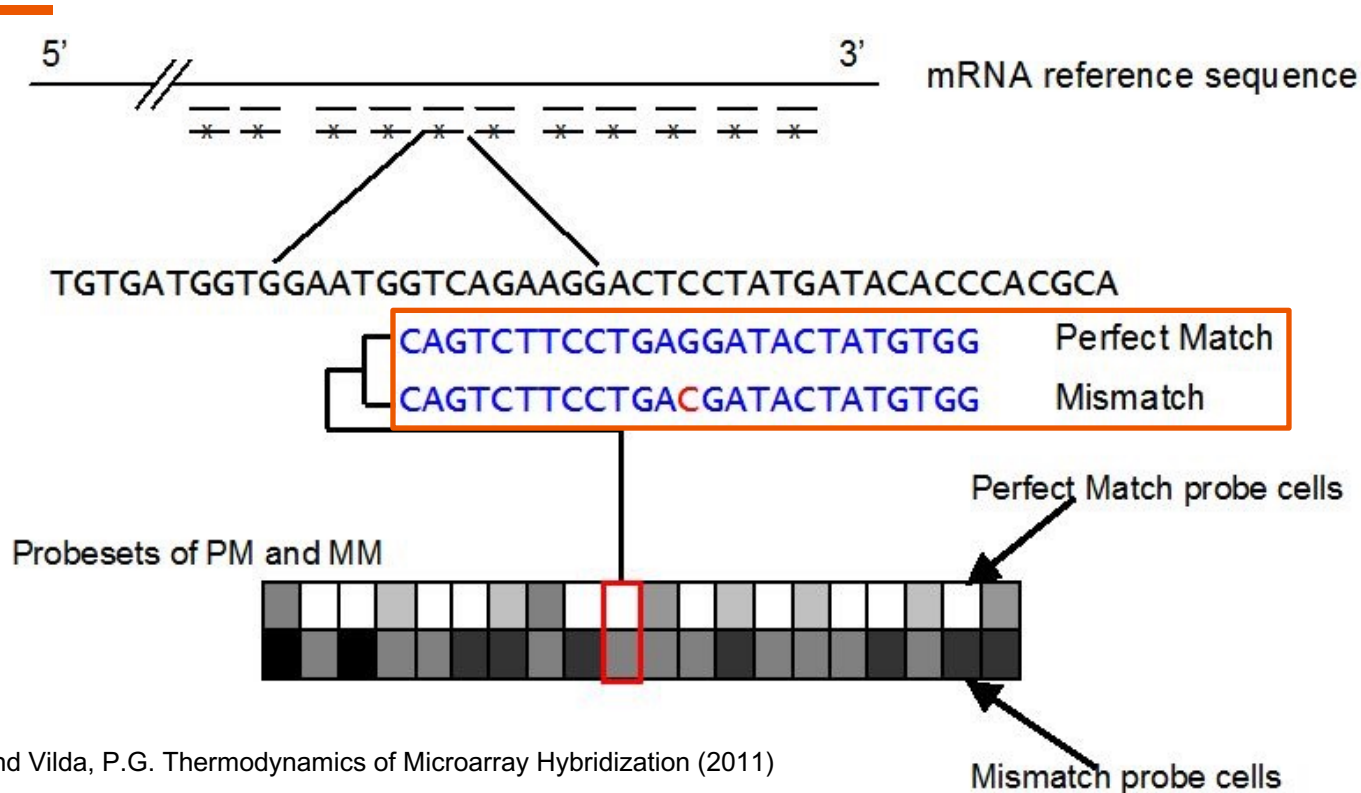


Chou, C.-C. NAR 32:e99 (2004)






Jaksik, R. et al. Biology Direct 10:46 (2015)

Perfect match (PM) and mismatch (MM)



Example microarray metadata

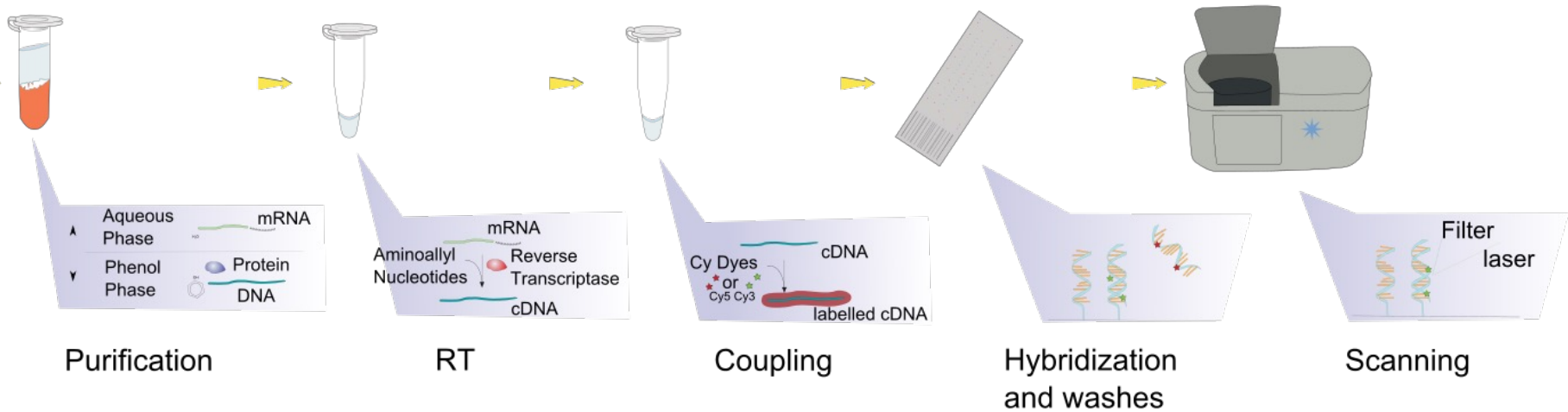


ControlType	ProbeName	SystematicName	PositionX	PositionY
1	GE_BrightCorner	GE_BrightCorner	584.922	4464.27
1	DarkCorner	DarkCorner	606.433	4464.3
1	DarkCorner	DarkCorner	626.841	4464.18
0	A_23_P326296	NM_144987	648.069	4464.19
0	A_24_P287941	NM_013290	669.667	4464.39
0	A_24_P325046	BC022434	691	4464.5
0	A_23_P200404	NM_001625	712	4464.5
0	A_19_P00800513	lincRNA:chr7:226042-232442_R	733.224	4464.48
0	A_23_P15619	NM_032391	754.4	4464.41
0	A_33_P3402354	L40403	775.5	4464.32
0	A_33_P3338798	NM_001145251	798.041	4464.16
0	A_32_P98683	NM_005937	817.068	4464.27
0	A_23_P137543	NM_152493	838.533	4464.4
0	A_19_P00803040	lincRNA:chr8:104254399-104295074_F	859.965	4464.37
0	A_23_P117852	NM_014736	881	4464.3
0	A_33_P3285585	AK127191	902.5	4464.5
0	A_24_P328231	NM_017871	923.214	4464.57
0	A_33_P3415668	NR_028328	944.776	4464.52
0	A_23_P73609	NM_000266	966	4464.5
0	A_24_P186124	NM_182501	986.871	4464.53



Microarray data pre-processing

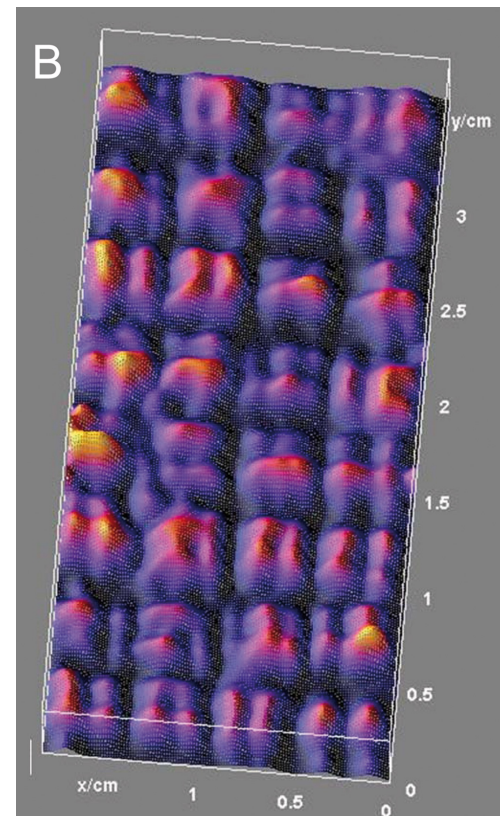
Microarray data collection steps



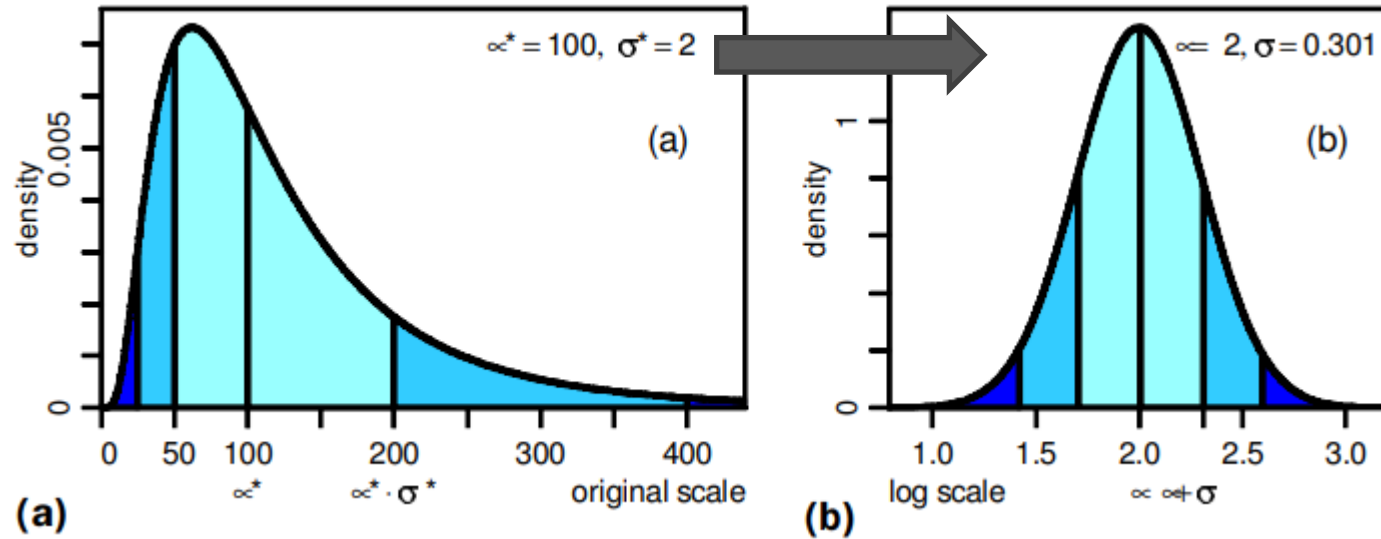
https://en.wikipedia.org/wiki/Microarray_analysis_techniques

Key processing steps

- Redefining probe set
 - BLAST to latest genome annotation
- Intensity correction
 - Model background using probe location & sequence
 - Perfect match (PM) vs mismatch (MM)
 - Global & local correction
- Outlier removal
- Probe set aggregation
- Log transform

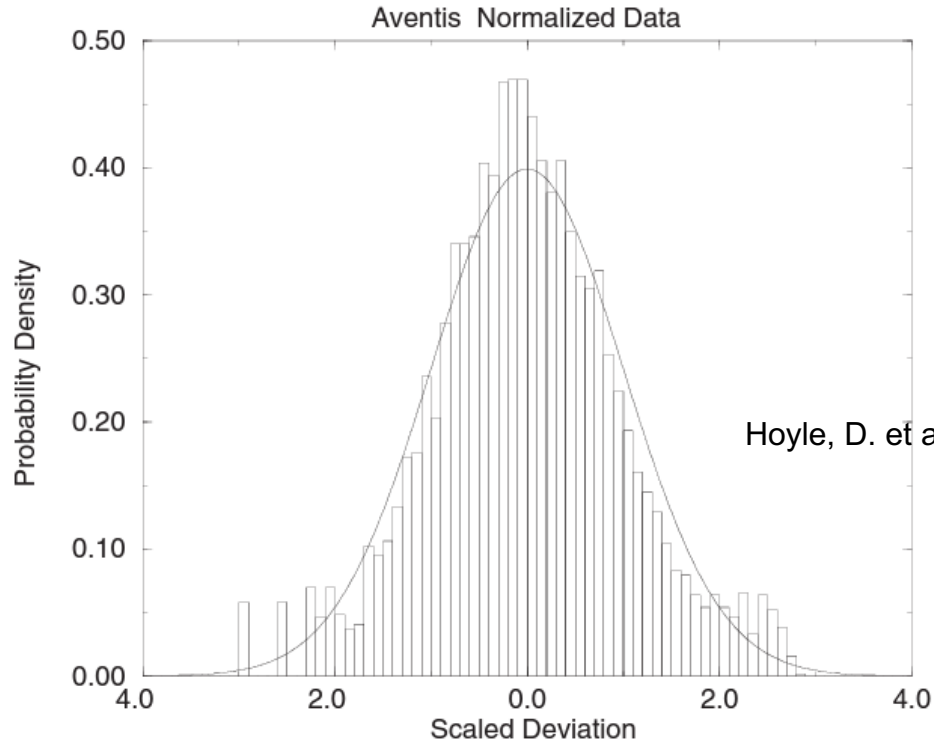


Log-normal distribution



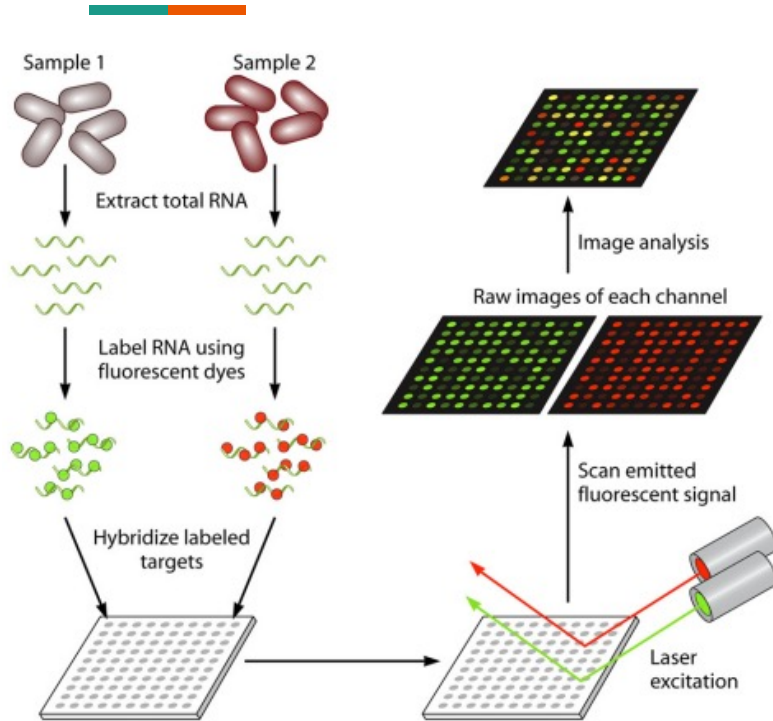
Limpert, Stahel, and Abbt. BioScience 2001.

Microarray data are log-normal distributed



Hoyle, D. et al. Bioinformatics 18:576-584 (2001)

Two-channel microarray

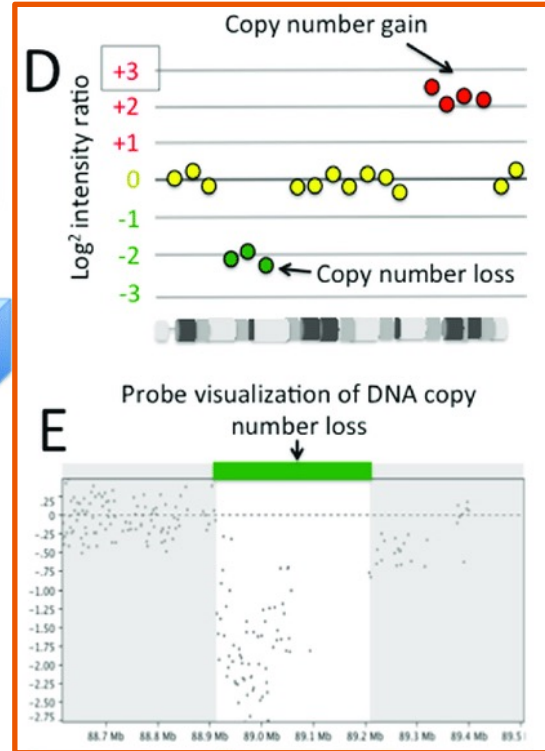
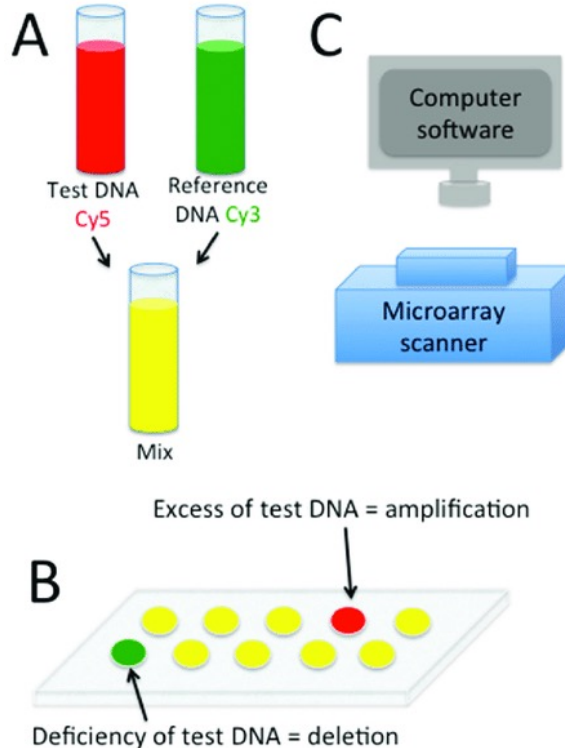


- Two samples are labeled with different dyes
- Mix and hybridize to microarray
- Relative fluorescence signal (ratio) directly indicates fold difference in gene expression
- **Minimize technical variance**



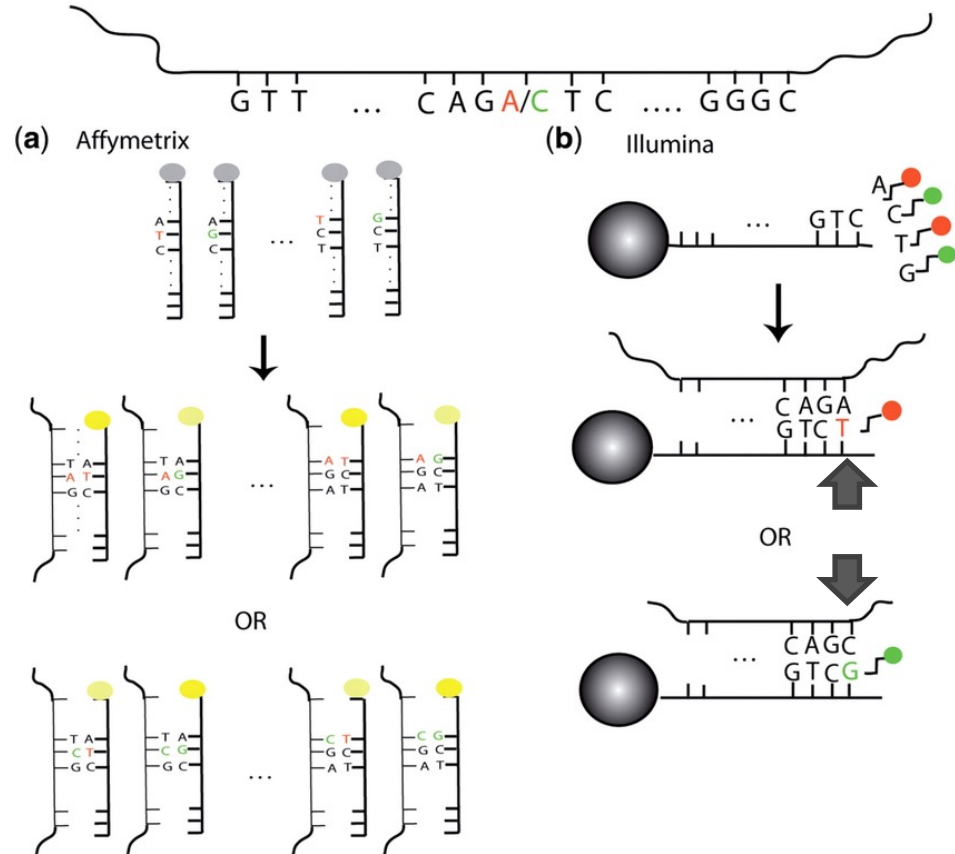
Beyond transcriptomics

Comparative genome hybridization (CGH)



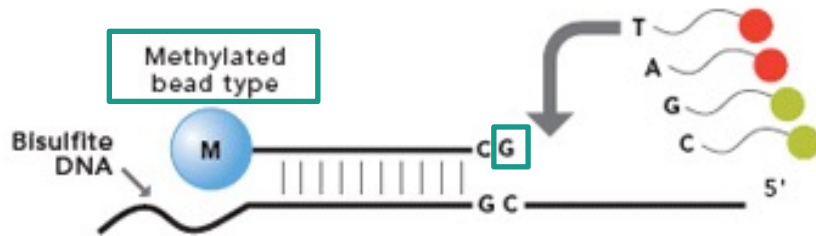
SNP genotyping array

- Design probes for alternative SNPs at each position
 - Relative hybridization
- Single-nucleotide sequencing
 - Probe acts as primer
 - Match to the position up until right before the SNP
 - Incorporation of the next nucleotide determine the genotype

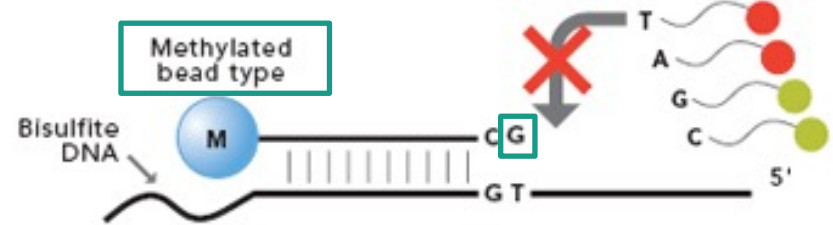
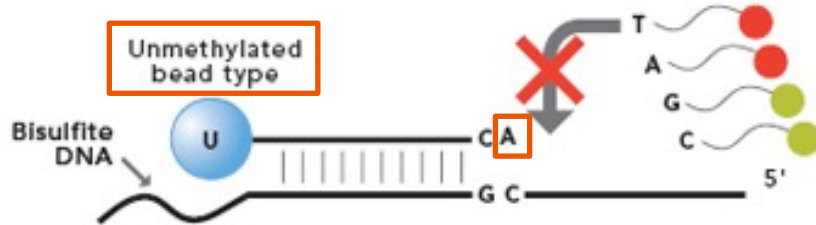
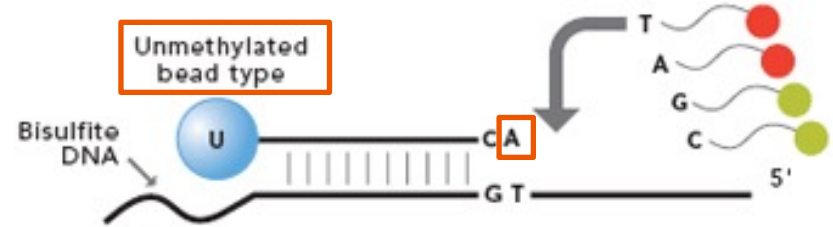


Methylation array

Methylated DNA Locus



Unmethylated DNA Locus



Microarray vs DNA sequencing

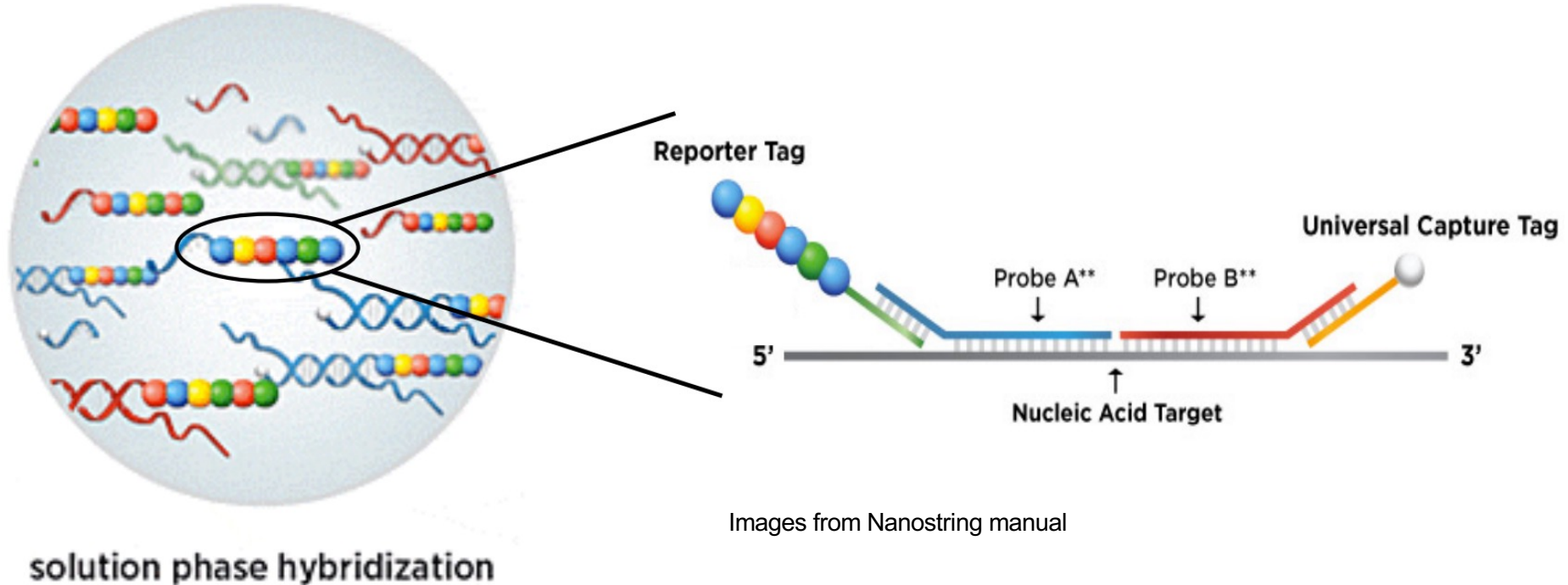


- Microarray and DNA sequencing are **interchangeable in many applications**
 - Genome tiling array
 - Fusion gene
 - ChIP-chip
 - Pathogen-specific probes
- Designed once for each task and can be reused
- Cheaper than DNA sequencing
 - But lack the ability to detect novel molecules

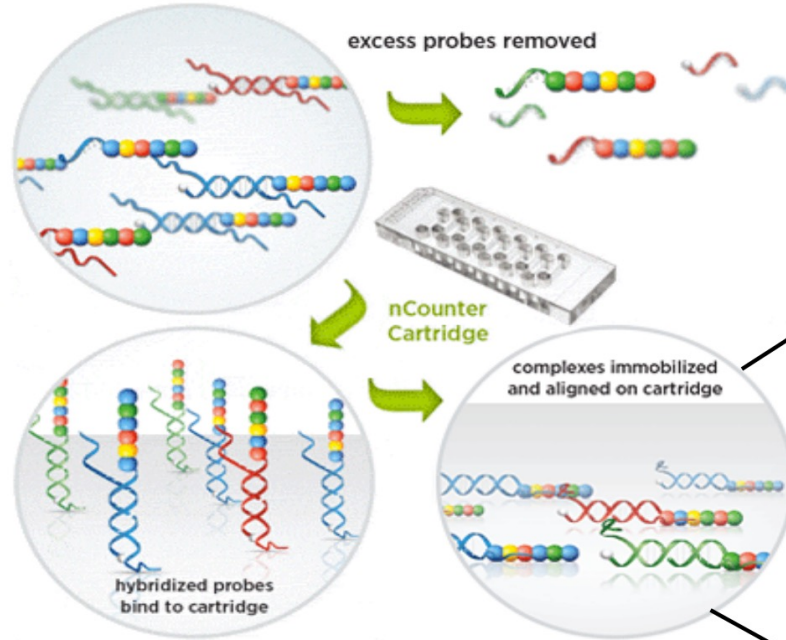


Nanostring

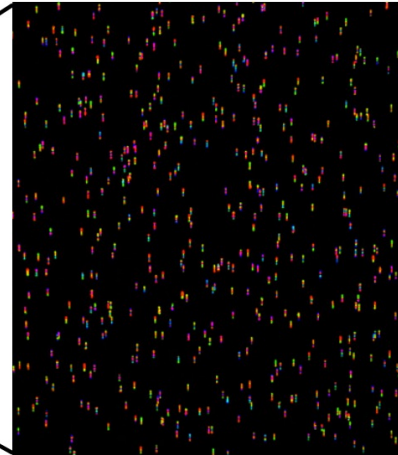
Transcript-specific probes & fluorescence barcodes



Counting number of molecules

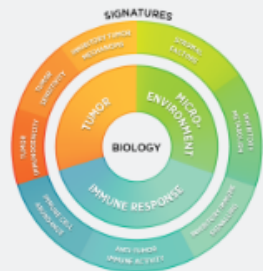


Barcode	Counts	Identity
	3	XLSA
	2	FOX5
	1	INSULIN



Images from Nanostring manual

Prebuilt barcode set (up to 800 targets)



PanCancer IO 360

Human  Mouse 

750 cancer-related genes involved in the complex interplay between the tumor, microenvironment and immune response including 20 internal reference controls.

Application:

Oncology

Species:

Human, Mouse

Genes in panel:

770, 770

% Match:

100%, 100%

Panel type:

Inventoried

Platform:

nCounter Analysis System



Canine IO

Canine 

The nCounter® Canine IO Panel includes 780 genes covering 47 annotated pathways involved in canine immune response to IO treatments, and 20 internal reference genes for [show more](#)

Application:

Oncology

Species:

Canine

Genes in panel:

800

% Match:

100%

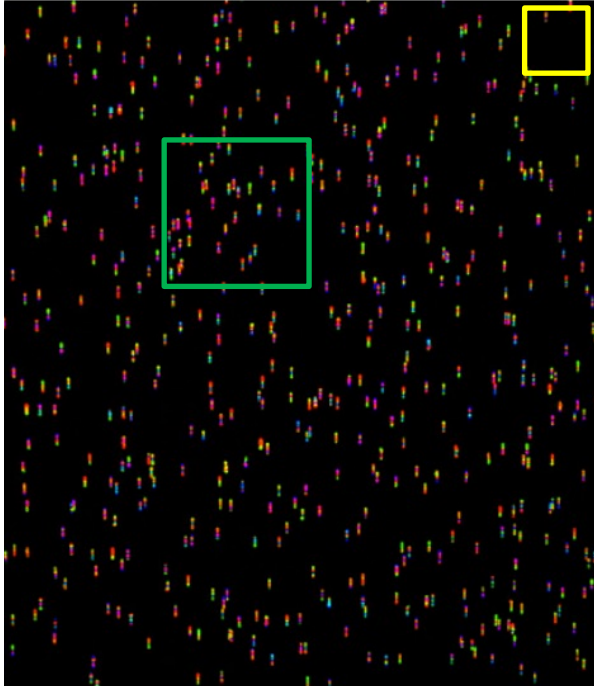
Panel type:

Inventoried

Platform:

nCounter Analysis System

Nanostring quality control



- Imaging QC
 - % of successful imaging field of view > 75%
- Binding QC
 - 0.1-2 molecules per square micron
- Positive control
 - Six synthetic DNA ranging from 0.125-128 fM
- Negative control
 - Eight synthetic DNA that do not bind to probe



Nanostring data preprocessing

Negative and positive control

☒ Background Subtraction/ Thresholding

☐ Background Subtraction ☒ Background Thresholding

☒ Negative control count

Class	Name	Avg. Count	Selected
Negative	NEG_A	14.5	<input checked="" type="checkbox"/>
Negative	NEG_B	15.583	<input checked="" type="checkbox"/>
Negative	NEG_C	24.416	<input checked="" type="checkbox"/>
Negative	NEG_D	15.166	<input checked="" type="checkbox"/>
Negative	NEG_E	15.083	<input checked="" type="checkbox"/>
Negative	NEG_F	14.5	<input checked="" type="checkbox"/>
Negative	NEG_G	18.916	<input checked="" type="checkbox"/>
Negative	NEG_H	21.083	<input checked="" type="checkbox"/>

Threshold to of Negative Controls
+ standard deviations

Raw Data					
			Sample 1	Sample 2	Sample 3
Positive	POS_A	ERCC_00117.1	24573	21007	21856
Positive	POS_B	ERCC_00112.1	6948	6414	6589
Positive	POS_C	ERCC_00002.1	2123	1826	1932
Positive	POS_D	ERCC_00092.1	432	363	425
Positive	POS_E	ERCC_00035.1	52	68	53
Positive	POS_F	ERCC_00034.1	49	38	52
		Geomean of POS:	858.01	783.19	829.55
Arithmetic mean of geomeans:			823.58		
POS control normalization factors:			0.96	1.05	0.99

Housekeeping control

☒ 2. CodeSet Content (Reference or Housekeeping) Normalization

☒ Standard ☐ Other

Set normalization genes as default for subsequent experiments.

Codeset Content

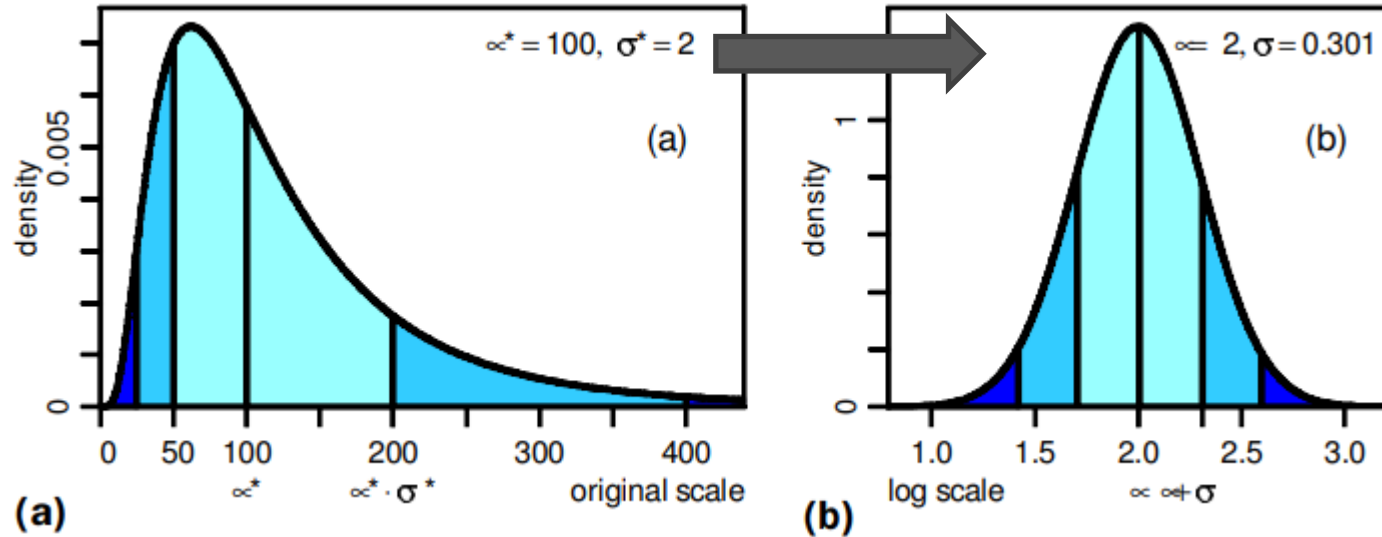
Gene Name	Class Name	Avg Count	%CV
ABCC4	Endogenous	54.231	71.243
ADM	Endogenous	259	82.216
AMD1	Endogenous	1,211.692	58.834
APC	Endogenous	107.769	58.735
ASPA	Endogenous	8.538	127.636
BTBD15	Endogenous	312.615	62.163
C11orf58	Endogenous	1,375.385	55.009
C13orf23	Endogenous	291.308	66.131
CCNA2	Endogenous	953.154	84.552
CDH1	Endogenous	1,487.385	150.15
CHGB	Endogenous	21.154	75.706
CYR61	Endogenous	1,766.385	94.931

Normalization Codes

Gene Name	Class Name	Avg Count	%CV
ACTB	Endogenous	23,095.23	82.706
POLR1B	Endogenous	213.846	58.665
LDHA	Endogenous	11,240.385	74.918

Use to compute normalization factor

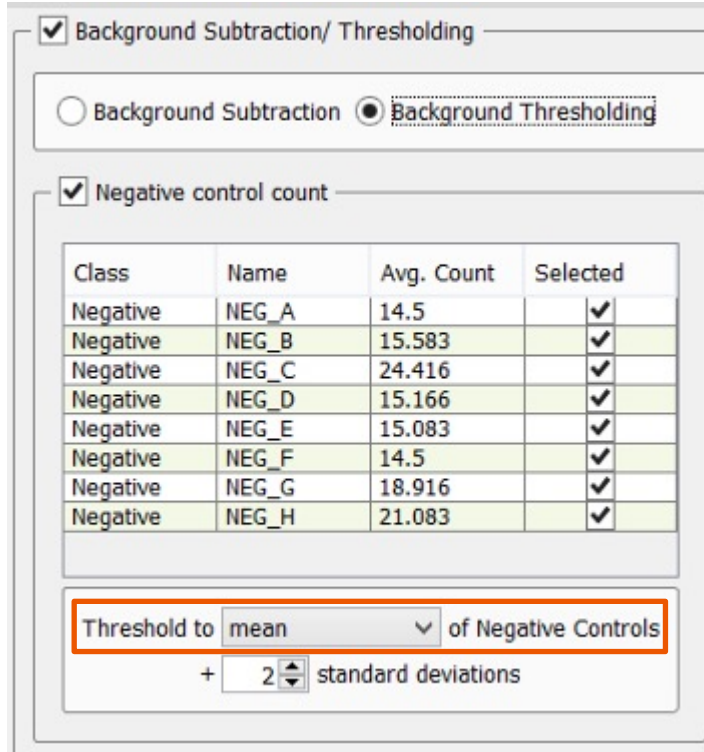
Arithmetic Mean vs Geometric Mean



Limpert, Stahel, and Abbt. BioScience 2001.

$$- \frac{\log(x) + \log(y)}{2} = \log(\sqrt{xy}) \rightarrow \text{AM of log-transformed} = \text{GM of original data}$$

Arithmetic mean of background noises



☒ Background Subtraction/ Thresholding

☐ Background Subtraction ☒ Background Thresholding

☒ Negative control count

Class	Name	Avg. Count	Selected
Negative	NEG_A	14.5	<input checked="" type="checkbox"/>
Negative	NEG_B	15.583	<input checked="" type="checkbox"/>
Negative	NEG_C	24.416	<input checked="" type="checkbox"/>
Negative	NEG_D	15.166	<input checked="" type="checkbox"/>
Negative	NEG_E	15.083	<input checked="" type="checkbox"/>
Negative	NEG_F	14.5	<input checked="" type="checkbox"/>
Negative	NEG_G	18.916	<input checked="" type="checkbox"/>
Negative	NEG_H	21.083	<input checked="" type="checkbox"/>

Threshold to mean of Negative Controls

+ 2 standard deviations

- Background noises are Normal
- Arithmetic Mean is ok

Geometric mean of molecule counts

Raw Data					
			Sample 1	Sample 2	Sample 3
Positive	POS_A	ERCC_00117.1	24573	21007	21856
Positive	POS_B	ERCC_00112.1	6948	6414	6589
Positive	POS_C	ERCC_00002.1	2123	1826	1932
Positive	POS_D	ERCC_00092.1	432	363	425
Positive	POS_E	ERCC_00035.1	52	68	53
Positive	POS_F	ERCC_00034.1	49	38	52
		Geomean of POS:	858.01	783.19	829.55
Arithmetic mean of geomeans:			823.58		
POS control normalization factors:			0.96	1.05	0.99

- Real expression data are closer to Log-Normal
- Use GM to represent AM of log-transformed data

Geometric mean of molecule counts

☒ 2. CodeSet Content (Reference or Housekeeping) Normalization

☒ Standard ☐ Other

Set normalization genes as default for subsequent experiments.

Gene Name	Class Name	Avg Count	%CV
ABCC4	Endogenous	54.231	71.243
ADM	Endogenous	259	82.216
AMD1	Endogenous	1,211.692	58.834
APC	Endogenous	107.769	58.735
ASPA	Endogenous	8.538	127.636
BTBD15	Endogenous	312.615	62.163
C11orf58	Endogenous	1,375.385	55.009
C13orf23	Endogenous	291.308	66.131
CCNA2	Endogenous	953.154	84.552
CDH1	Endogenous	1,487.385	150.15
CHGB	Endogenous	21.154	75.706
CYR61	Endogenous	1,766.385	94.931

Gene Name	Class Name	Avg Count	%CV
ACTB	Endogenous	23,095.23	82.706
POLR1B	Endogenous	213.846	58.665
LDHA	Endogenous	11,240.385	74.918

Use to compute normalization factor

- Real expression data are closer to Log-Normal
- Use GM to represent AM of log-transformed data



Simple transcriptomics analysis

Transformed data can be analyzed with *t*-test

Control

Treatment

J15								
	A	B	C	D	E	F	G	H
1	Acc ID	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	
2	NM_007818	67540.89	70924.09	80243.76	3501.2	5697.47	2426.72	
3	NM_001105160	811.93	801.36	740.71	128.67	104.42	101.33	
4	NM_028089	190.41	211.06	236.19	9.05	23.33	8.44	
5	NM_016696	66.77	57.56	101.09	750.9	659.84	491.89	
6	NM_013459	3.3	11.29	1.89	735.82	816.46	118.22	
7	NM_007809	45.34	36.12	51.02	245.27	372.13	335.67	
8	NM_009999	103.04	370.21	200.29	17.09	13.33	8.44	
9	NM_133960	7708.78	6976.38	6569.04	1731	1641.81	1853.55	
10	NM_027881	31.32	10.16	24.56	268.39	186.62	135.11	
11	NM_054053	31.32	24.83	19.84	323.68	428.78	116.11	
12	NM_007377	47.81	89.17	70.86	370.93	378.79	279.72	
13	NM_028064	703.95	689.62	662.29	214.11	168.85	144.61	
14	NM_008182	222.56	339.73	226.75	30.16	63.32	26.39	
15	NM_013661	12.36	11.29	8.5	97.51	77.76	71.78	
16	NM_007815	20613.09	25218.13	31540.46	5209.07	7680.3	6312.2	

- Log transform
- Perform *t*-test on each gene
- Correct the p-values for multiple testing

Choosing the right t -test

=t.test(A2:A13,B2:B13,

T.TEST(array1, array2, tails, type)

- (...) 1 - One-tailed distribution
- (...) 2 - Two-tailed distribution

T.TEST uses the one-tailed distribution

- Two-tailed tests whether the expression is higher or lower

=t.test(A2:A13,B2:B13,2,

T.TEST(array1, array2, tails, type)

- (...) 1 - Paired
- (...) 2 - Two-sample equal variance (homoscedastic)
- (...) 3 - Two-sample unequal variance (heteroscedastic)

T.TEST performs a paired t-Test

- Use paired only for before & after treatment data of the same sample
- Otherwise, assuming unequal variance (Welch) is safer

Correction for multiple testings



- P-value cutoff of 0.05 means that under the null hypothesis, there is only 5% chance of observing the same or more extreme result
- Applying similar test 1,000 times will result in 50 tests on average with smaller p-value than 0.05 just by chance
 - Differential expression analysis tests thousands of gene at once
- This is unacceptable if a conclusion relies on multiple tests
 - Biological interpretation assumes that selected genes are truly differentially expressed

Bonferroni method



- Divide the p-value cutoff by the number of test
- Adjusted p-value cutoff = $0.05 / 1000 = 0.00005$
- Applying similar test 1,000 times will result in 0.05 tests on average with smaller p-value than 0.00005 just by chance
- Easy to calculate but lose power

False discovery rate (FDR)



- P-value operates under the **null hypothesis**
- But in practice, we want to control the number of errors in the output
 - The number of DEGs that were incorrectly proposed
- **False Discovery Rate (FDR)** = Probability of getting a false positive
 - Probability that a DEG is not truly differentially expressed
- But FDR involves **alternative hypothesis**
- There are ways to control FDR through p-value!

Benjamini-Hochberg procedure



- Valid under **broad assumptions** (independent tests, positively correlated tests, etc.)
- Given a series of tests with p-values, p_1, p_2, \dots, p_n
- To control FDR to be within 0.05
 - Sort p-values from low to high, p'_1, p'_2, \dots, p'_n
 - Find largest k such that $p'_k \leq 0.05 \times k / n$
 - For the smallest p-value, this is equivalent to **Bonferroni**
 - For other p-values, this technique gradually loosens the cutoff
 - Reject null hypothesis for tests corresponding to p'_1, p'_2, \dots, p'_k

Benjamini-Yekutieli procedure



- Valid under **broader assumption** (some dependence between tests)
- Given a series of tests with p-values, p_1, p_2, \dots, p_n
- To control FDR to be within 0.05
 - Sort p-values from low to high, p'_1, p'_2, \dots, p'_n
 - Find largest k such that $p'_k \leq (0.05 \times k) / (n \times c(k))$
 - $c(k) = \sum_{i=1}^k \frac{1}{i}$
 - For the smallest p-value, this is equivalent to **Benjamini-Hochberg** & **Bonferroni**
 - For other p-values, this technique gradually loosens the cutoff – but not as much as **Benjamini-Hochberg**
 - Reject null hypothesis for tests corresponding to p'_1, p'_2, \dots, p'_k

Correction method comparison



P-value	Bonferroni	B-H	B-Y
Smallest	0.0005	0.0005	0.0005
2 nd smallest	0.0005	0.001	0.000667
3 rd smallest	0.0005	0.0015	0.000818
4 th smallest	0.0005	0.002	0.00096
5 th smallest	0.0005	0.0025	0.001095

- There are 100 tests
- Target p-value or FDR cutoff = 0.05

Effect of correction



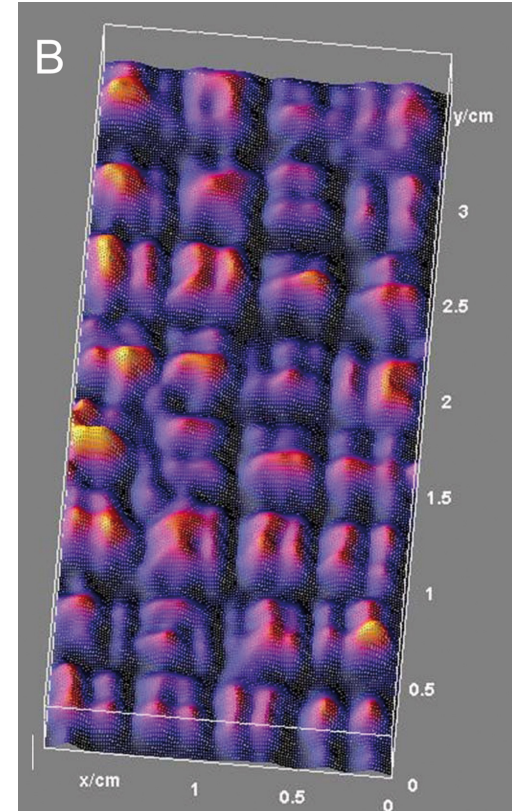
Gene	Sorted p-value	Rank	Benjamini-Hochberg	Result	c(rank)	Benjamini-Yekutieli	Result
Gene M	0.0000001	1	0.0005	Pass	1	0.0005	Pass
Gene S	0.0000035	2	0.001	Pass	1.5	0.00067	Pass
Gene A	0.00028	3	0.0015	Pass	1.83	0.00082	Pass
Gene C	0.0011	4	0.002	Pass	2.08	0.00096	Fail
Gene P	0.06	5	0.0025	Fail	2.28	0.0011	



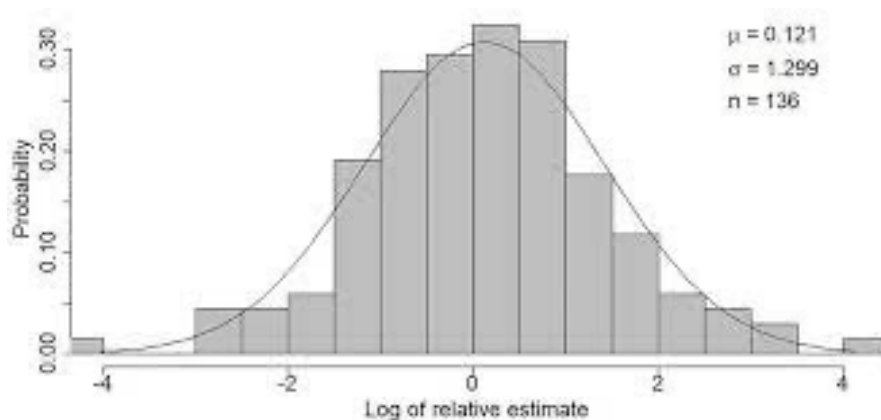
Linear effect model

Background noise correction models

- Null hypothesis
 - Background noise is normally distributed and is the same over the entire array
- Linear effect model
 - Background noise is normally distributed with mean depending on (x, y) positions and a fixed variance



Fitting normally distributed data



- Negative control probe intensities: n_1, n_2, \dots, n_k
 - Fitted mean and variance: $\mu = \frac{\sum_i n_i}{k}$ and $\sigma^2 = \frac{1}{k-1} \sum_i (n_i - \mu)^2$
 - Likelihood: $\prod_i P(n_i | \mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^k e^{-\frac{1}{2} \sum_i \left(\frac{n_i - \mu}{\sigma} \right)^2}$

Linear effect model



- Position of probe i with intensity n_i is (x_i, y_i)
- Fitted mean: $\mu(x_i, y_i) = ax_i + by_i + c$
 - Solve for a, b, c that minimize MSE: $\sum_i (n_i - (ax_i + by_i + c))^2$
- Recall calculus:
 - $\frac{\delta MSE}{\delta a} = \sum_i 2(n_i - (ax_i + by_i + c))(-x_i)$
 - $\frac{\delta MSE}{\delta b} = \sum_i 2(n_i - (ax_i + by_i + c))(-y_i)$
 - $\frac{\delta MSE}{\delta c} = \sum_i 2(n_i - (ax_i + by_i + c))(-1)$

Finding optimal solution by setting derivative to zero



- Boundary: a , b , or c approaches infinity
- Local optima: setting partial derivatives to zero
 - $0 = \sum_i 2(n_i - (ax_i + by_i + c))(-x_i)$
 - $0 = \sum_i 2(n_i - (ax_i + by_i + c))(-y_i)$
 - $0 = \sum_i 2(n_i - (ax_i + by_i + c))(-1)$

Some algebra exercises



- Setting partial derivatives to zero
 - $0 = \sum_i (n_i - (ax_i + by_i + c))(-x_i)$
 - $0 = \sum_i (n_i - (ax_i + by_i + c))(-y_i)$
 - $0 = \sum_i (n_i - (ax_i + by_i + c))$
- Or equivalently
 - $a \sum_i x_i^2 + b \sum_i x_i y_i + c \sum_i x_i = \sum_i n_i x_i$
 - $a \sum_i x_i y_i + b \sum_i y_i^2 + c \sum_i y_i = \sum_i n_i y_i$
 - $a \sum_i x_i + b \sum_i y_i + ck = \sum_i n_i$
- Three linear equations with three variables 😊

Hypothesis testing



- Fitted mean: $\mu(x_i, y_i) = ax_i + by_i + c$
 - Solve for a, b, c that minimize MSE: $\sum_i (n_i - (ax_i + by_i + c))^2$
- Fitted variance: $\sigma^2 = \frac{1}{k-1} \sum_i (n_i - \mu(x_i, y_i))^2$
- Likelihood: $\prod_i P(n_i | \mu(x_i, y_i), \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^k e^{-\frac{1}{2} \sum_i \left(\frac{n_i - \mu(x_i, y_i)}{\sigma} \right)^2}$
 - Compare to likelihood from Null hypothesis
 - **Likelihood ratio test** or **nested model testing** (degree of freedom = 2)

Incorporating confounding variable



- Design matrix

Sample	Condition	Batch	Patient's Age
S1	Control	1	35
S2	Control	2	21
S3	Control	3	45
S4	Treatment	1	18
S5	Treatment	2	37
S6	Treatment	3	52

Any question?



- See you on September 18