

Introduction of Bioinformatics

Fall 2007

Part II: Microarray Data Analysis

Biological Principle Behind Microarray Experiments



- Base complementary $\begin{matrix} A & \longleftrightarrow & T, U \\ C & \longleftrightarrow & G \end{matrix}$

- Gene expression experiments measure the **amount of mRNA** to see which genes are being expressed in the cell.
- Measuring protein levels directly is also possible, but is currently harder.

Transcription



From DNA to mRNA

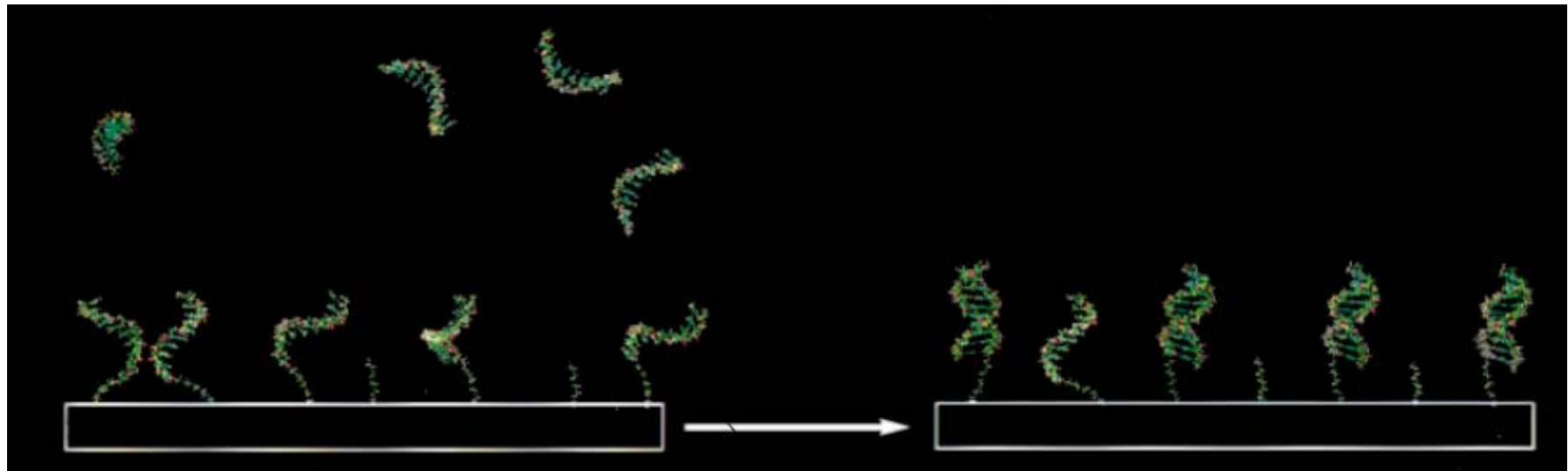
Reverse Transcription

Clone cDNA strands, complementary to the mRNA



Microarray Technology

- Microarray technology takes advantage of hybridization properties of nucleic acid
 - **probes**: complementary molecules (e.g. cDNAs) attached to a solid surface
 - **target**: the specific nucleic acid transcripts of interest



- A specific scanner is used to measure the amount of hybridized target at each probe, which is reported as an **intensity**.

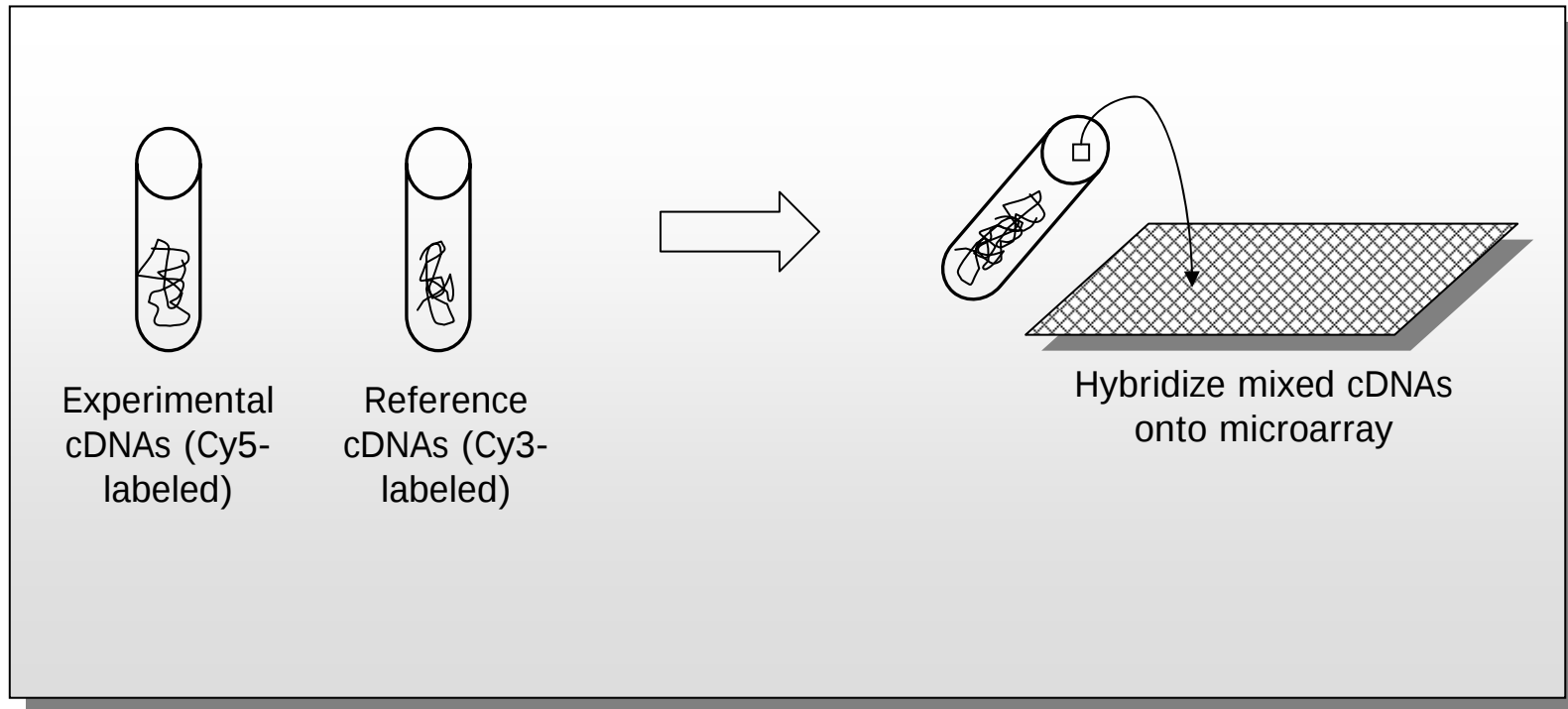
Microarray Platforms

- Two main classes of platforms:
 - High-density oligonucleotide array
 - Two-color spotted array

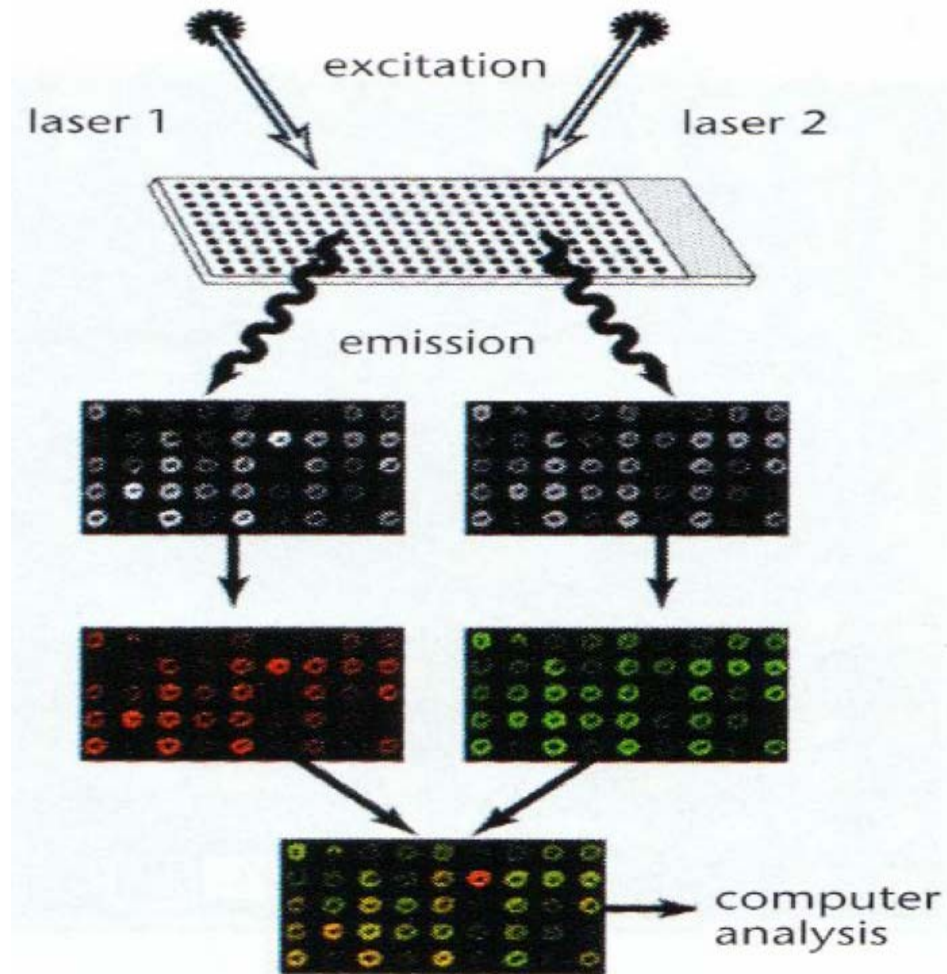
Two-color Spotted Arrays

- Two-color (two-channel) spotted platforms:
 - Two colors represent the two samples (experiment and reference) competitively hybridized
 - Each spot has “red” and “green” measurements associated with it.

Two-color Spotted Arrays



Two-color Spotted Arrays



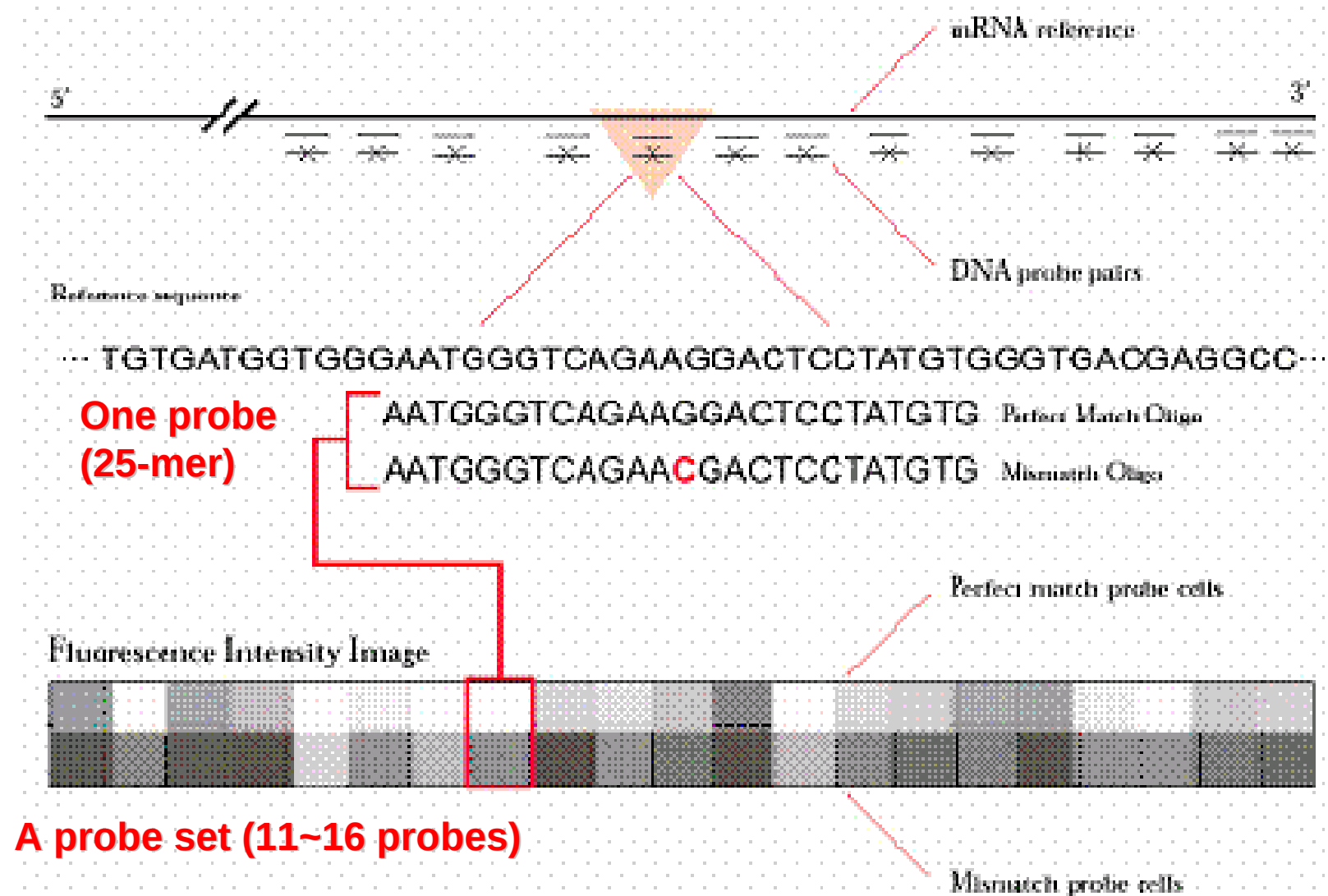
A	B	C	D	E	F	G	H	I	J	K
NAME	TYPE	ACC	CLID	SPOT	svcc77 CH1D	svcc77 CH2D	svcc77 FLAG	svcc78 CH1D	svcc78 CH2D	svcc78 FLAG
zinc finger	cDNA	AA406467	753234	1	1846	2088	0	6660	10328	
small prolin	cDNA	AA447835	813614	577	352	492	0	527	275	
zinc finger	cDNA	T57959	71626	1153	2601	1883	0	2677	1686	
486544	cDNA	AA043334	486544	1729	3527	2745	0	2069	1431	
zinc finger	cDNA	H17047	50794	2	3444	3039	0	6830	9446	
small induc	cDNA	H52985	205633	578	842	1292	0	1330	1514	
Human PC	cDNA	AA425602	768644	1154	548	655	0	1275	860	
small induc	cDNA	AA425102	768551	1730	4951	1797	0	2714	1055	
ESTs, High	cDNA	VW16724	302190	3	2170	1897	0	6442	13371	
Sjogren syn	cDNA	H29484	49970	579	3710	2356	0	10049	2749	
zinc finger	cDNA	AA082564	511814	1155	3106	1890	0	3429	2074	
signal rec	cDNA	AA411407	754958	1731	3538	3169	0	3523	1513	
Down synd	cDNA	H19439	51408	4	1694	1256	0	6505	6343	
sex hormo	cDNA	T69346	82871	580	387	676	0	502	406	
alpha thal	cDNA	AA410435	753430	1156	94	185	0	929	798	
selectin L	cDNA	H00756	149910	1732	1338	949	0	900	764	
wingless-t	cDNA	VW49672	324901	5	6690	3050	0	8633	5541	
selectin E	cDNA	H39560	186132	581	207	517	0	254	954	
wingless-t	cDNA	T99653	122762	1157	272	2075	0	365	1404	
sarcoglyc	cDNA	AA234982	868829	1733	476	586	0	654	529	
von Hippel	cDNA	H73054	234856	6	2517	2255	0	3252	2177	
steroid sul	cDNA	H15215	49591	582	5075	5328	0	7069	6892	
visinin-like	cDNA	H65066	210575	1158	1098	666	0	1046	448	
SR Y (sex-	cDNA	AA400739	753184	1734	5026	7070	0	2180	3533	
ESTs, Mod	cDNA	H69834	213280	2305	112	374	0	420	274	
phosphoin	cDNA	AA464765	810372	2881	879	409	1	867	87	

High-density Oligonucleotide Array

- 寡聚核苷酸晶片 $\sim 1.28 \text{ cm}^2$
- Contain one set of probe-level data per microarray; some probes for specific finding and others for nonspecific finding



High-density Oligonucleotide Array



*<http://keck.med.yale.edu/affymetrix/technology.htm>

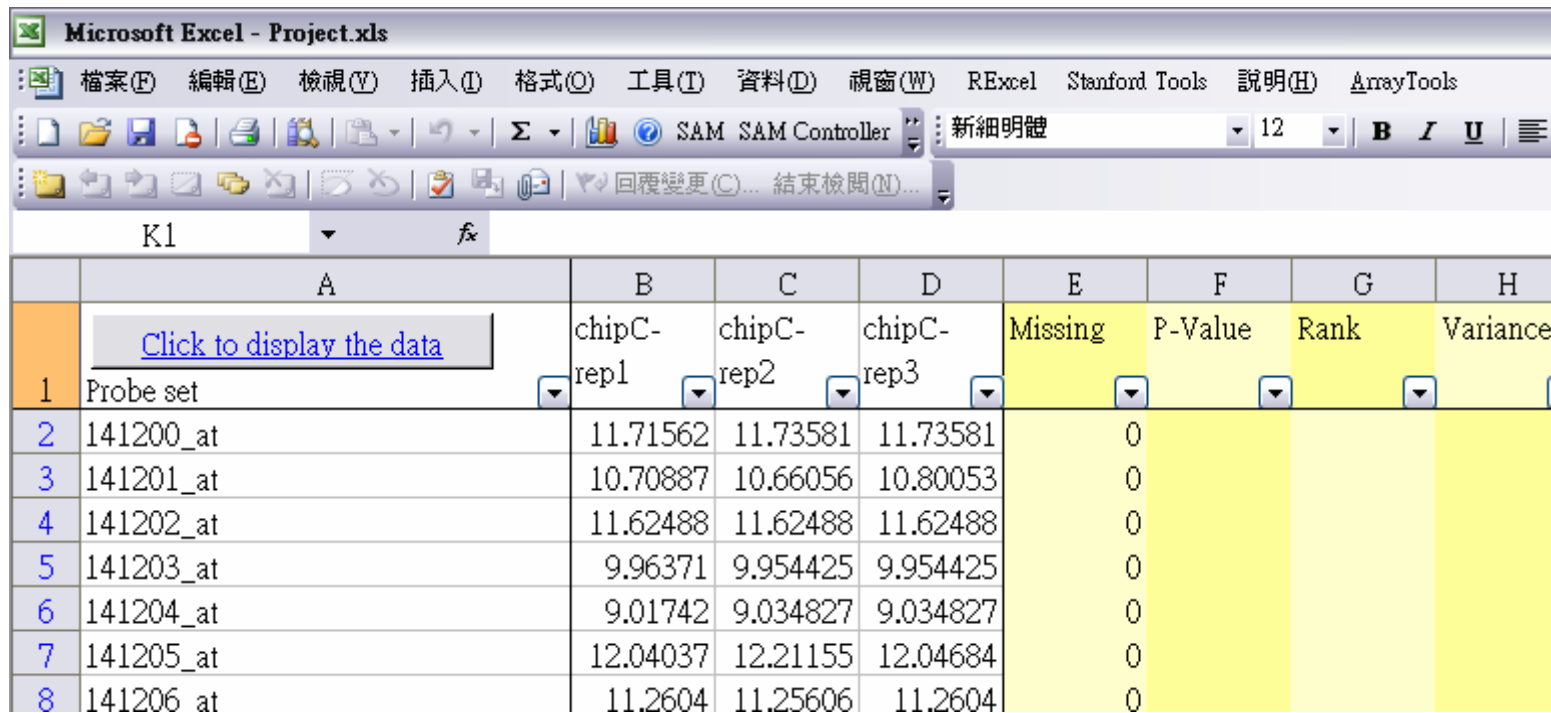
High-density Oligonucleotide Array

- **PM (Perfect Match):** The perfect match probe has a sequence exactly complementary to the particular gene.
- **MM (Mismatch):** The mismatch probe differs from the perfect match probe by a single base substitution at the center base position.
- $\sum w_k(\text{PM}_k - \text{MM}_k)^+$ quantifies the expression level of a gene (MAS 5.0).

Affymetrix GeneGhips

.dat file: a huge image file

.cel file: cell intensity file



Microsoft Excel - Project.xls

檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O) 工具(T) 資料(D) 視窗(W) RExcel Stanford Tools 說明(H) ArrayTools

新細明體 12 B I U

回覆變更(C)... 結束檢閱(N)...

	A	B	C	D	E	F	G	H
	Click to display the data	chipC-rep1	chipC-rep2	chipC-rep3	Missing	P-Value	Rank	Variance
1	Probe set							
2	141200_at	11.71562	11.73581	11.73581	0			
3	141201_at	10.70887	10.66056	10.80053	0			
4	141202_at	11.62488	11.62488	11.62488	0			
5	141203_at	9.96371	9.954425	9.954425	0			
6	141204_at	9.01742	9.034827	9.034827	0			
7	141205_at	12.04037	12.21155	12.04684	0			
8	141206 at	11.2604	11.25606	11.2604	0			

Microarray Result

- **Probe-level data:** the intensities read for each of the components.
- **Genomic-level data:** the measures being used in real research.

Preprocessing

- Preprocessing is the process to account for various sources of variation and make the probe-level data to genomic-level data

probe-level data $\xrightarrow{\text{preprocessing}}$ genomic-level data

- Statistical analyses will be applied to genomic-level data.

Outline for Part II

1. Introduction of Microarray

2 . Statistical Analyses:

- Distance Measures in DNA Microarray Data Analysis
- Cluster Analysis of Genomic Data
- Analysis of Differential Gene Expression Studies
- Multiple Testing Procedures and Applications to Genomics
- Machine Learning Concepts and Tools for Statistical Genomics

3. Data Visualization

4. Preprocessing

- Preprocessing High-density Oligonucleotide and Two-Color Spotted Arrays
- Preprocessing SELDI-TOF Mass Spectrometry Protein Data