Microarray

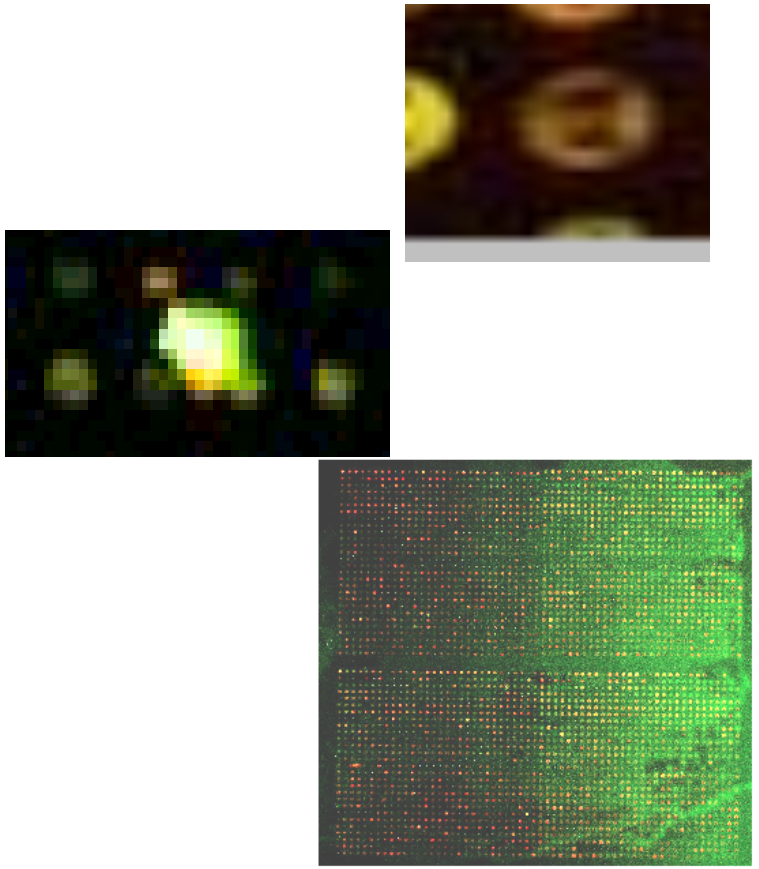
Lecture 4

Outline

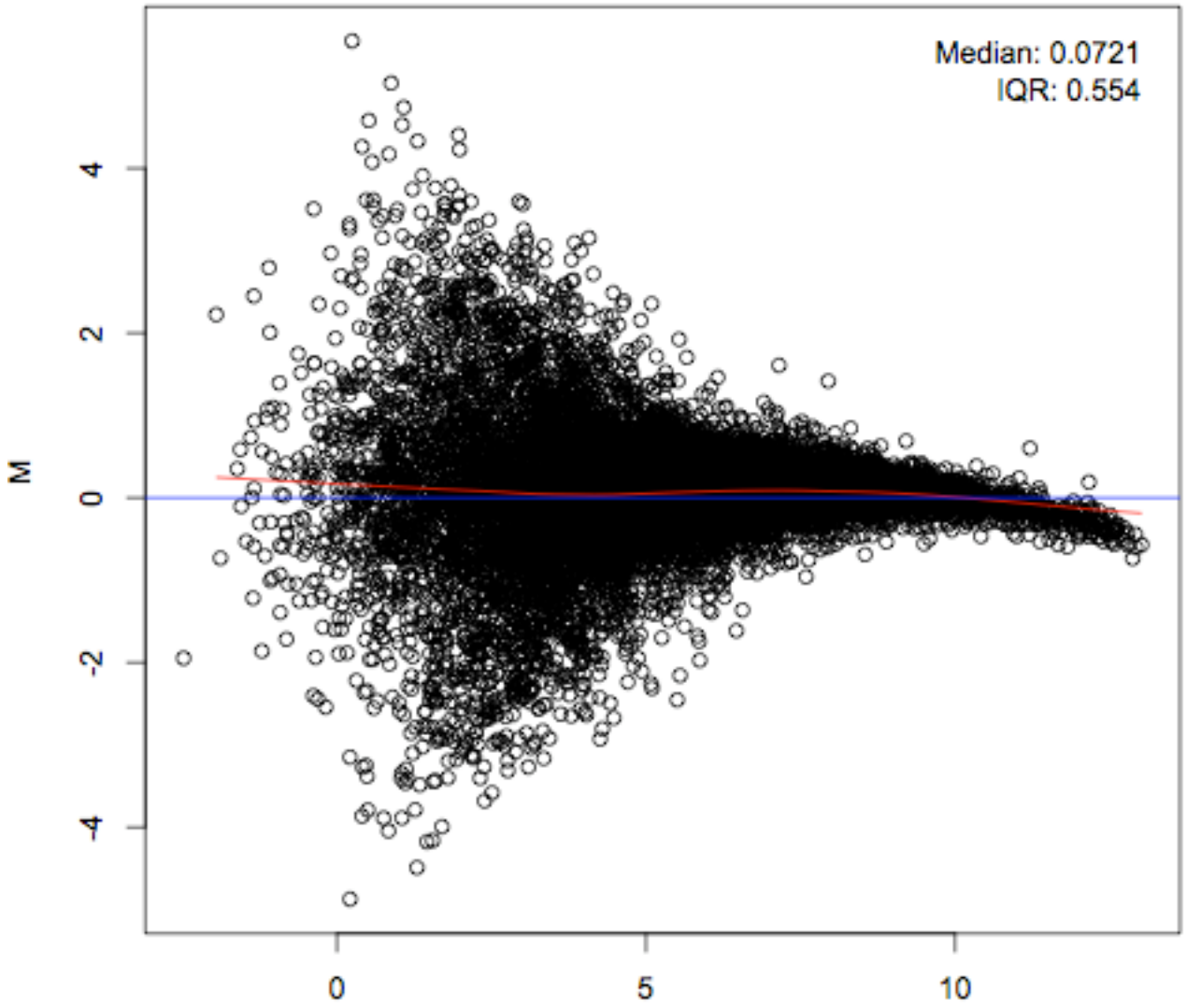
- Background
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

Spot QA for cDNA Spotted Arrays

- Spot Measures
  - Uniformity
  - Spot Area
- Inspect images for artifacts
- Global Measures
  - Qualitative assessments
  - Averages of spot measures

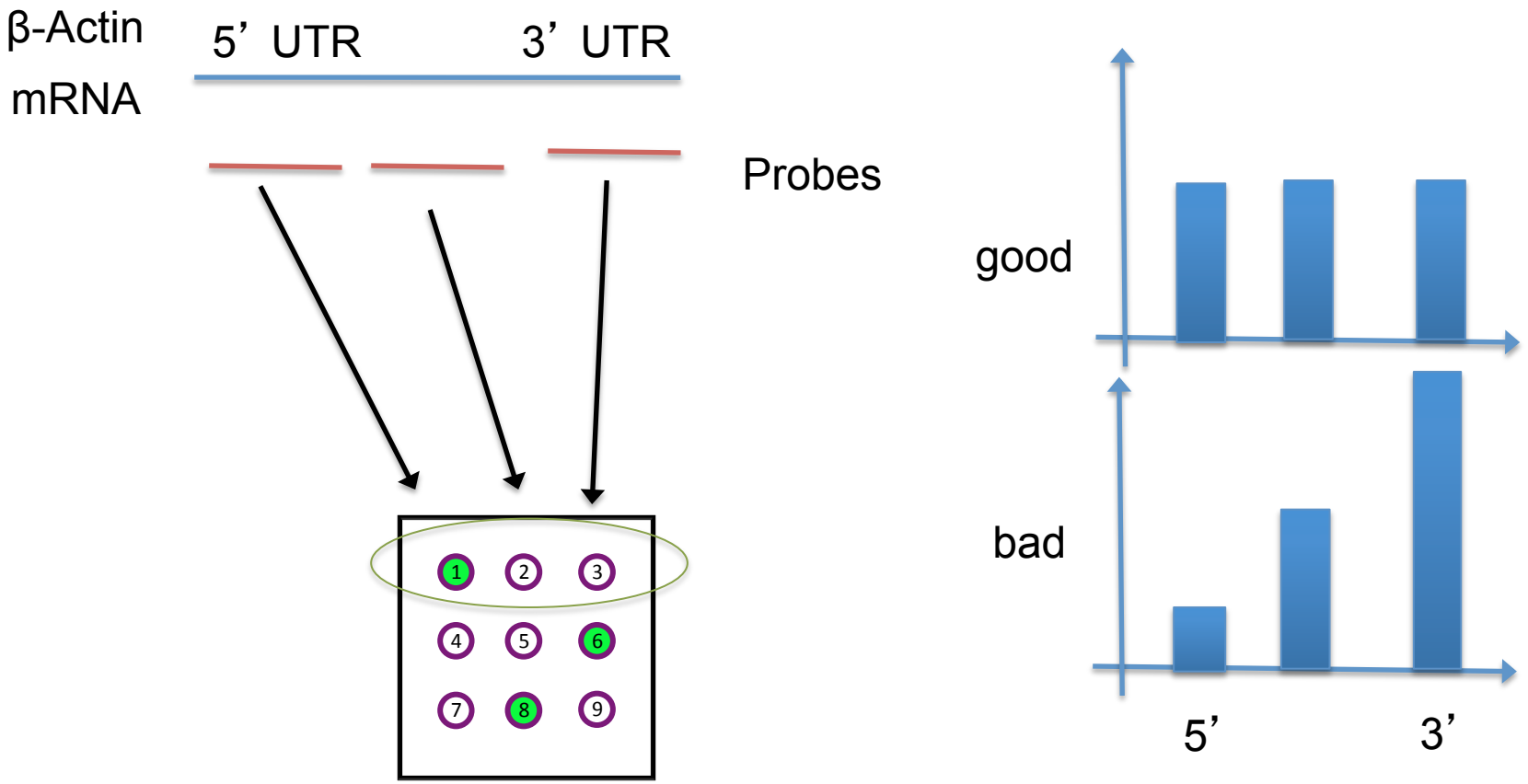


MA plot



The general assumption is that most of the genes would not see any change in their expression.

3’/5’ ratio: RNA quality



The assumption is that RNA degradation, or problems during labeling, can lead to under intensity representation at the 5′ end of RNA

Outline

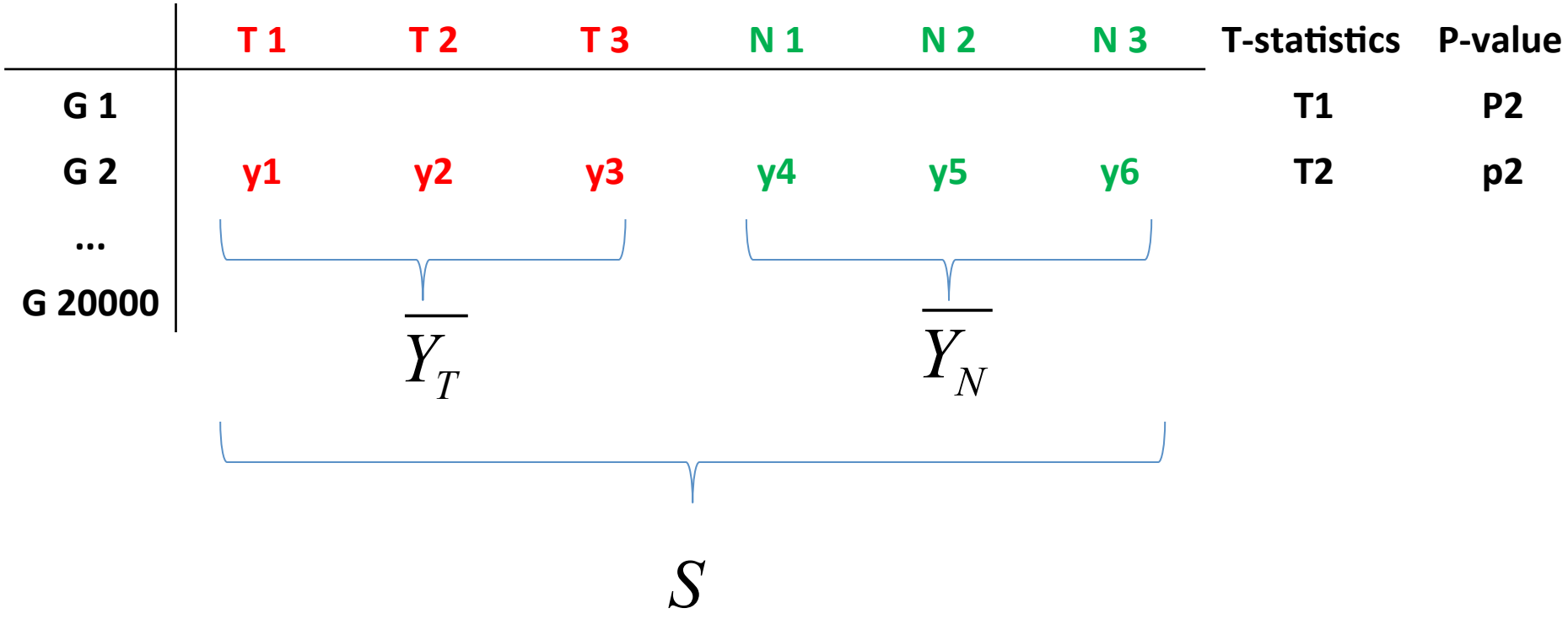
- Background
- Preprocessing of oligonucleotide microarray
- Quality Assessment for oligonucleotide Microarray
- Differential Expression Testing

Type of problem

- To compare two groups
  - Treatment group vs. control group
- To compare multiple groups
  - Treatment A, Treatment B, Control group
- To consider multiple variables (factors) simultaneously
  - Treatment variable (Treatment vs. Control), age variables (>50 vs. <50), …

Two-group comparisons: Student's T-test

- Then, for gene2, calculated the difference between group means, divided by global standard error; obtain T2 and P2



Statistics methods for two-group comparisons

- T-test
  - Student's t-test: assumes normally distributed data in each group, equal variance within groups
  - Welch t-test: as above, but allows unequal vairance
- Univariate linear model
- Nonparametric test
  - Wilcoxon, or rank-sums test: non-parametric, rank-based
  - Permutation test: estimate the distribution of the test statistics under the null hypothesis by permutations of the sample labels

Univariate Linear Model

- The expression of gene x is modeled as a baseline expression level (from the normal group) plus the group effect (i.e., tumor vs. normal)

$$Y = Y_N + \beta Z + \varepsilon$$

Expression level of gene x

Baseline expression level of gene x

Group effect

Group vector
- ❑ 0 for normal group
- ❑ 1 for tumor group

- β represents the group effect. It is P-value (through F-test) can be used to test whether the expression in tumor is different from normal (i.e., showing group effect).

Univariate Linear Model

- Under R: > glm()
- Example output for a single gene:

Variable	Effect estimate	St.error	t-statistic	p-value
Intercept	0.4015	0.4334	0.926	0.381329
Group	-3.43	0.6129	-5.597	0.000512
Multiple R-square	0.7966		F-statistic	31.32
Adjusted R-square	0.7711		df	(1,8)
			p-value	0.000512

Type of problem

- To compare two groups
  - Treatment group vs. control group
- To compare multiple groups
  - Treatment A, Treatment B, Control group
  - Solutions: ANOVA (F test)
- To consider multiple variables (factors) simultaneously
  - Treatment variable (Treatment vs. Control), age variables (>50 vs. <50), …

multiple group comparison

g1	*g2*	*control*
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

F-test

- **Step 1:** Calculate the mean within each group.
- **Step 2:** Calculate the overall mean.
- **Step 3:** Calculate the "between-group" sum of squares $\sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y})^2/(K-1)$
- **Step 4:** Calculate the "within-group" sum of squares $\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2/(N-K),$
- **Step 5:** The F-ratio

$$F = \frac{MS_B}{MS_W}$$

F-test

g1	*g2*	*control*
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Step 1: Calculate the mean within each group.

Y1=5

Y2=9

Y3=10

F-test

g1	*g2*	*control*
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

- **Step 2:** Calculate the overall mean.

Y1=5

Y2=9

Y3=10

Y=(5+9+20)/3=8

F-test

g1	*g2*	*control*
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

- **Step 3:** Calculate the "between-group" sum of squares

$Y1=5, Y2=9, Y3=10$

$Y=8$

$S=6x(5-8)^2+6x(9-8)^2+6x(10-8)^2=84$

F-test

g1	*g2*	*control*
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

- **Step 3:** Calculate the "between-group" sum of squares

Y1=5, Y2=9, Y3=10

Y=8

$S=6x(5-8)^2+6x(9-8)^2+6x(10-8)^2=84$

Freedom=3group-1=2

F-test

g1	*g2*	*control*
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

- **Step 3:** Calculate the "between-group" sum of squares

Y1=5, Y2=9, Y3=10

Y=8

$S=6x(5-8)^2+6x(9-8)^2+6x(10-8)^2=84$

Freedom=3group-1=2

MS=84/2=42

F-test

g1	*g2*	*control*
6 – 5 = 1	8 – 9 = -1	13 – 10 = 3
8 – 5 = 3	12 – 9 = 3	9 – 10 = -1
4 – 5 = -1	9 – 9 = 0	11 – 10 = 1
5 – 5 = 0	11 – 9 = 2	8 – 10 = -2
3 – 5 = -2	6 – 9 = -3	7 – 10 = -3
4 – 5 = -1	8 – 9 = -1	12 – 10 = 2

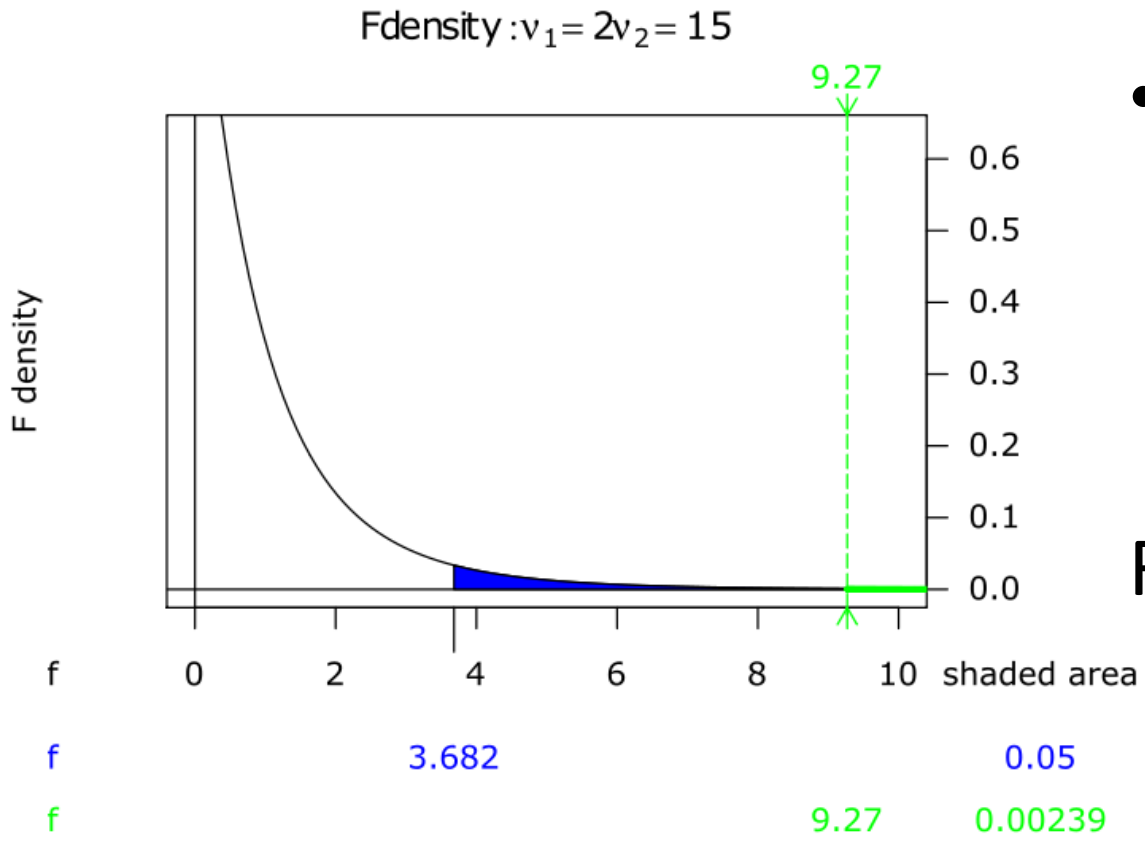
- **Step4:** Calculate the "within-group" sum of squares

S=1+9+1+1+0+4+1+1+9+0+4+9+1+9+1+1+4+9+4=68

Freedom=3x(6-1)=15

MS=68/15=4.5

F-test



- **Step 5:** The F-ratio

F=42/4.5=9.3

The critical value is the number that the test statistic must exceed to reject the test. In this case, $F_{crit}(2,15) = 3.68$ at $\alpha = 0.05$. Since $F = 9.3 > 3.68$, the results are significant at the 5% significance level. One would reject the null hypothesis, concluding that there is strong evidence that the expected values in the three groups differ. The p-value for this test is 0.002.

Type of problem

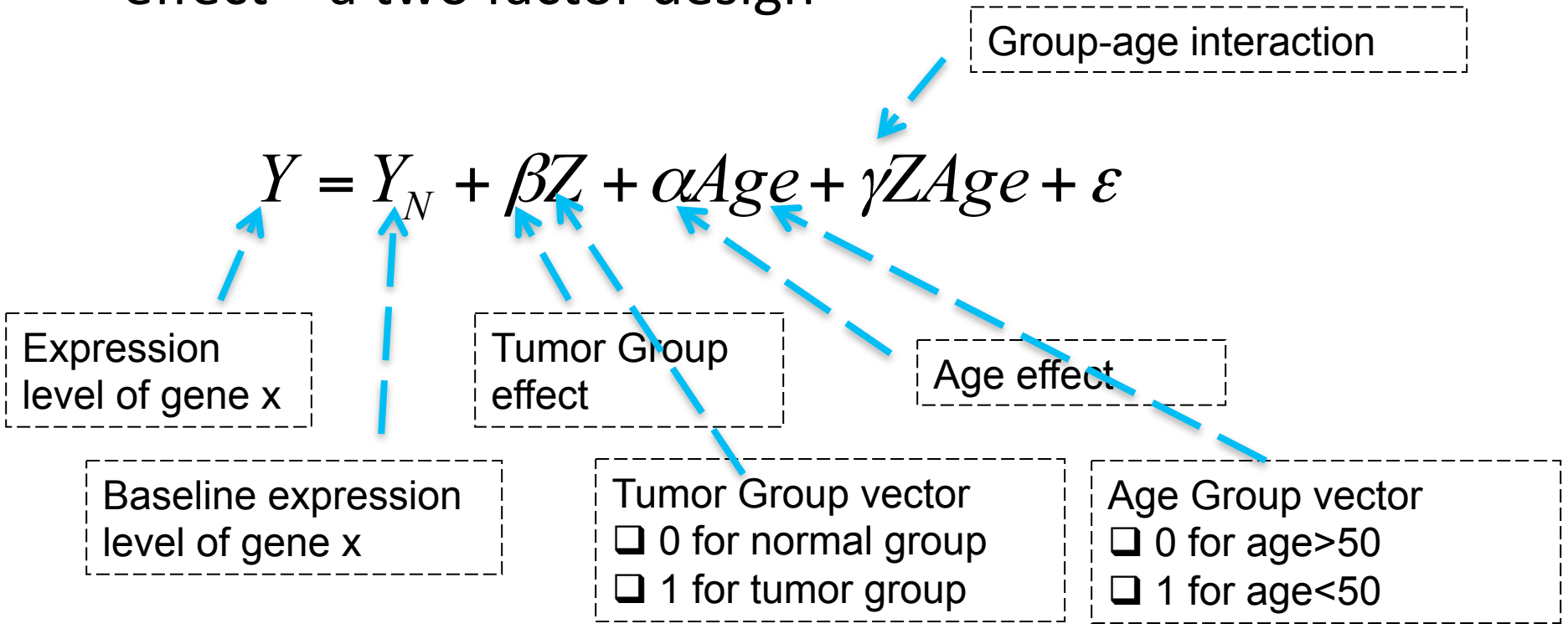
- To compare two groups
  - Treatment group vs. control group
- To compare multiple groups
  - Treatment A, Treatment B, Control group
  - Solutions: ANOVA (F test)
- To consider multiple variables (factors) simultaneously
  - Treatment variable (Treatment vs. Control), age variables (>50 vs. <50), …
  - Solutions: multivariate linear model

The two-factor experiment

	tumor	age
Sample 1	**0**	**0**
Sample 2	**0**	**0**
Sample 3	**1**	**0**
Sample 4	**1**	**0**
Sample 5	**0**	**1**
Sample 6	**0**	**1**
Sample 7	**1**	**1**
Sample 8	**1**	**1**

Multivariate Linear Model

- The expression of gene x is modeled as a baseline expression level (from the normal group) plus the group effect, age effect, and group-age interaction effect – a two factor design



Multivariate Linear Model

- Under R: > glm()
- Example output for a single gene:

	d.f.	Sum Sq	Mean Sq	F statistic	p-value
Treatment	1	20.6848	20.6848	25.9737	0.000263
Age	2	27.2838	13.6419	17.13	0.000305
Treatment:Age	2	0.5526	0.2763	0.3469	0.713707
Residuals	12	9.5565	0.7964		

- Both factors - treatment (tumor vs. normal) and age, has effect on gene's differential expression. However, no evidence for their interaction effect .

LIMMA Package

- The LIMMA package (one of the Bioconductor package) is for differential expression analysis of data arising from microarray experiments
  - The central idea is to fit a linear model to the expression data for each gene, with the experiment design information contained within the design matrix.
  - Also, empirical Bayes are used to borrow information across genes to estimate the standard deviation.

Limma for two-group case

	20A	20B	10A	10B
G 1				
G 2				
…				
G 12625				

```
> d.exp=exprs(Dilution)
> design <- cbind(WT=c(1,1,0,0),MU=c(0,0,1,1))
> fit <- lmFit(d.exp, design)
> cont.matrix <- makeContrasts(MUvsWT=WT-MU, levels=design)
> fit2 <- contrasts.fit(fit, cont.matrix)
> fit2 <- eBayes(fit2)
> results=topTable(fit2, number=20, adjust="fdr", lfc=1)
```

Results

> results

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
156253	156253	861.50	1687.750	14.351607	0.0001304622	0.885185	-4.595113
236914	236914	585.75	1252.375	12.166323	0.0002507076	0.885185	-4.595113
11614	11614	450.05	620.025	9.628034	0.0006270167	0.885185	-4.595113
21225	21225	358.50	587.250	8.843570	0.0008718059	0.885185	-4.595114
209366	209366	396.50	875.250	7.855832	0.0013746967	0.885185	-4.595114
212569	212569	431.10	718.950	7.086558	0.0020342135	0.885185	-4.595114
48215	48215	693.75	1243.525	6.988311	0.0021443453	0.885185	-4.595114
90347	90347	1571.35	2912.325	6.890651	0.0022611877	0.885185	-4.595114
28257	28257	545.75	1226.525	6.703351	0.0025080364	0.885185	-4.595114
47650	47650	543.10	890.450	6.609500	0.0026441801	0.885185	-4.595114
62342	62342	348.35	887.825	6.504775	0.0028069954	0.885185	-4.595114
52062	52062	320.75	641.525	6.494663	0.0028233597	0.885185	-4.595114
14171	14171	317.60	379.350	6.486309	0.0028369685	0.885185	-4.595114
53342	53342	262.75	383.125	6.293804	0.0031741131	0.885185	-4.595114

Limma Package

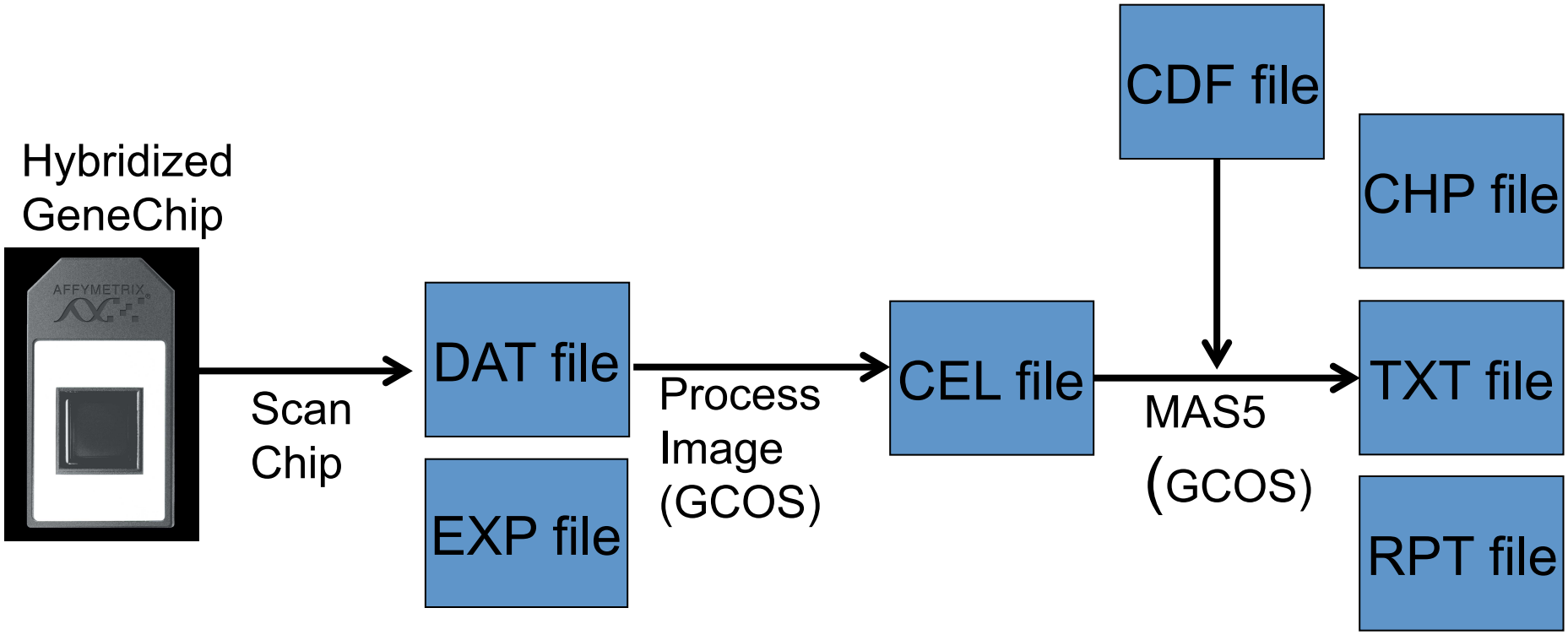
- http://bioconductor.org/packages/release/bioc/html/limma.html

How to deal with raw affy data

Affymetrix File Types

- DAT file:
  - Raw (TIFF) optical image of the hybridized chip
- CDF File (Chip Description File):
  - Provided by Affy, describes layout of chip. Each chip has a corresponding **CDF** which describes probe locations and probeset groupings on the chip.
- CEL File:
  - Processed DAT file (intensity/position values)
- CHP File:
  - Experiment results created from CEL and CDF files
- TXT File:
  - Probeset expression values with annotation (CHP file in text format)
- EXP File
  - Small text file of Experiment details (time, name, etc)
- RPT File
  - Generated by Affy software, report of QC info

Affymetrix Data Flow



Read CEL files

-Note we can read multiple CEL at the same time.

(1) need edit a plain text file saving all CEL names.

(2) read those affy files.

Read CEL files

-Edit a plain text file with the following format and save it as “targets.txt”.

```
FileName
Control_filename1.CEL
Control_filename2.CEL
Control_filename3.CEL
Treatment_filename1.CEL
Treatment_filename2.CEL
Treatment_filename3.CEL
```

Read CEL files

-(2) read those affy files.

```
> #red target filetargets
> Targets <- readTargets("targets.txt")
> #read microarry
> dataab <- ReadAffy(filenames=targets$FileName)
> preprocessed_dataab=rma(dataab)
```

Limma for two-group case

```
> design <- cbind(CT=c(1,1,1,0,0,0),TR=c(0,0,0,1,1,1))
> fit <- lmFit(preprocessed_dataab, design)
> cont.matrix <- makeContrasts(CT-TR, levels=design)
> fit2 <- contrasts.fit(fit, cont.matrix)
> fit2 <- eBayes(fit2)
> results=topTable(fit2, number=20, adjust="fdr", lfc=1)
```

Limma for two-group case

```
> design
    CT TR
[1,]  1  0
[2,]  1  0
[3,]  1  0
[4,]  0  1
[5,]  0  1
[6,]  0  1
```

```
> cont.matrix
       Contrasts
Levels CT - TR
    CT       1
    TR      -1
```

Log(fold change)=log(Control/Treatmetn)

Limma for two-group case

```
> results=topTable(fit2, number=20, adjust="fdr", lfc=1)

> results=topTable(fit2,  adjust="fdr", lfc=1)
```

Filter in R

> results

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
13611	Zm.7576.1.A1_at	-7.294902	9.368417	-48.18529	2.227504e-10	3.950256e-06	13.440325
181	Zm.100.1.A1_at	-6.798488	9.147950	-35.79768	1.917041e-09	1.699841e-05	12.069663
3684	Zm.14489.1.S1_at	-5.545995	8.421467	-31.88446	4.427037e-09	2.520650e-05	11.443865
2002	Zm.12635.1.S1_s_at	-4.640604	10.588414	-29.71106	7.371049e-09	2.520650e-05	11.039271
9216	Zm.3633.1.A1_at	-4.731900	8.121015	-28.43470	1.011871e-08	2.520650e-05	10.779434
12870	Zm.6721.4.A1_s_at	-4.271442	8.528172	-27.17756	1.401924e-08	2.520650e-05	10.505617
12918	Zm.6757.1.S1_at	-5.111377	8.398503	-27.10415	1.429522e-08	2.520650e-05	10.489045
5250	Zm.16735.1.A1_at	-4.019595	8.592958	-27.02468	1.460095e-08	2.520650e-05	10.471031

> results[,6]<0.000001

```
 [1] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

>results[results[,6]<0.000001,] ←what is the result?