Micro-Array, Golub et al. Data

Reampling-Based Testing

S. Stanley Young

Glaxo Wellcome Inc.

Michael Emptage

Glaxo Wellcome Inc.

Peter H. Westfall

Texas Tech University

Dmitri Zaykin

Glaxo Wellcome Inc.

Outline

1. Micro array data
2. Problem statement
3. Statistical methods
4. Results
5. Questions

Why micro array data?

Business and Science Drivers

- Drug target selection
- Bio network understanding
- Reduce drug development costs

Why micro array data?

Knowledge and technology converge

- Human Genome Project(s)
- Bio chip technology
- Informatics

Three Statistical Analysis Problems

1. Correlated genes
   (guilt by association).

2. General genetic structure.

3. Biology / gene associations.

Goal: Understand gene-phenotype relationships

- Level gene correlations
- Level k gene associations
- Level one gene/bio associations
- Level k gene/bio associations.

Method : Resampling-based testing!

What are the problems?

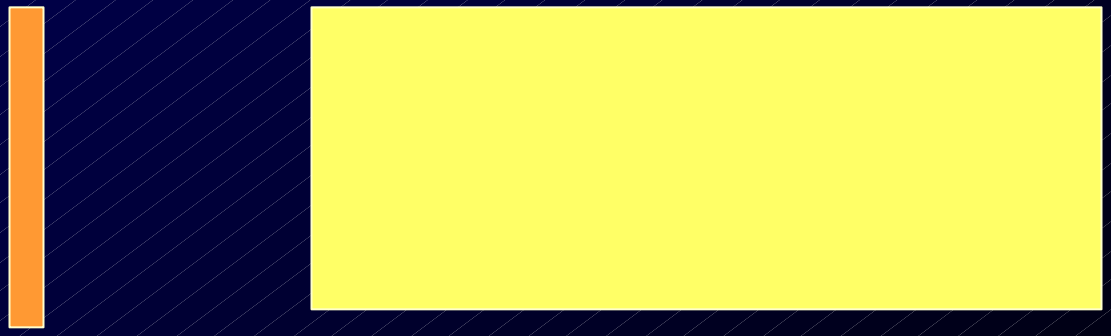
1. Few statistical experimental units
2. Very many genes
3. Non-normal distributions
4. Phenotype and data quality
5. Statistical methods

Data Formulation

Standard Formulation :

Phenotype = f(Genotype)

Phenotype Genes



Problems with Standard Formultions

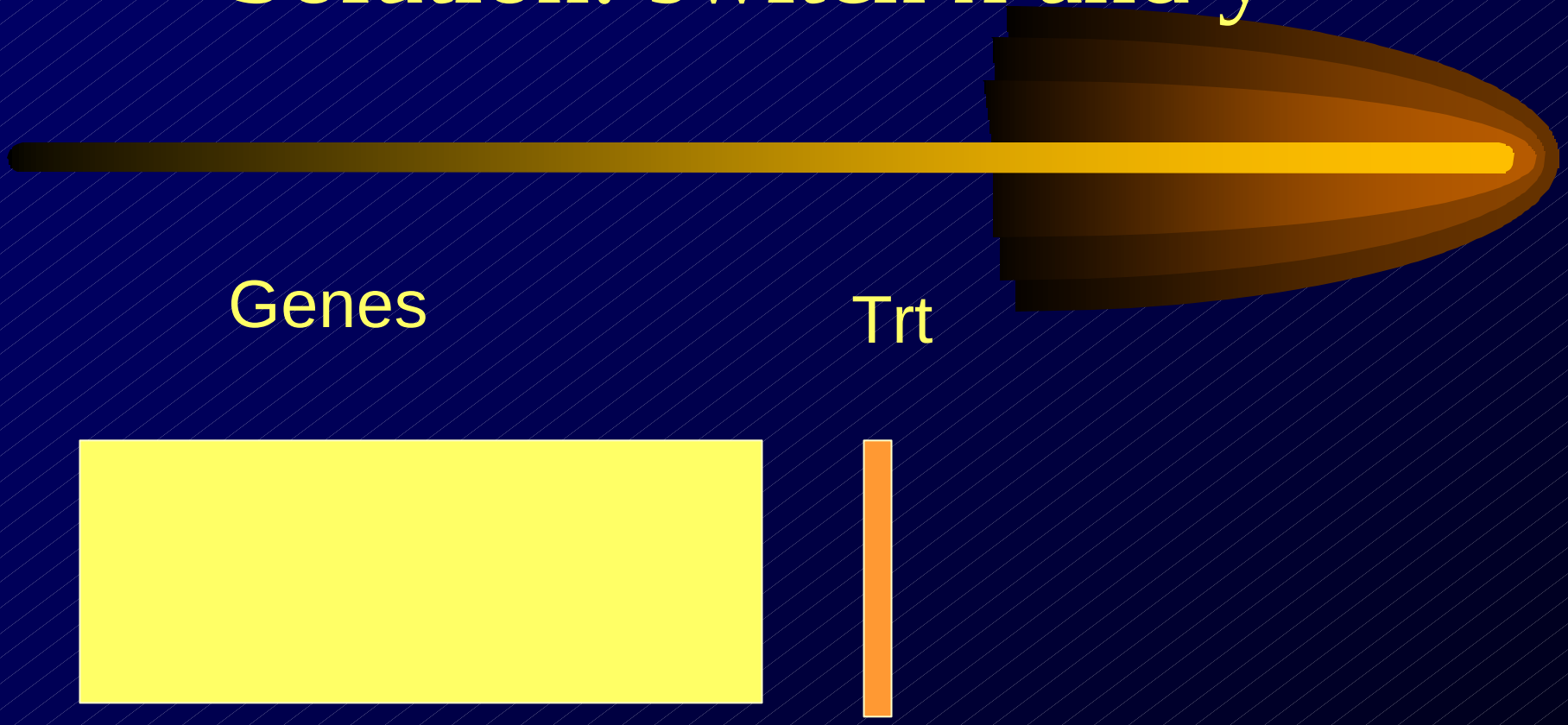
Standard Formulation : Phenotype = f(Genotype)

1. Gene expression measured with error.

2. Genotype relatively error free.

3. Enormous number of genes.

Solution: switch x and y



Statistical Plan : permute Trt at random, and compute Max t over all genes.

Statistical Testing Strategy

1. Treat micro array data as Y vector.

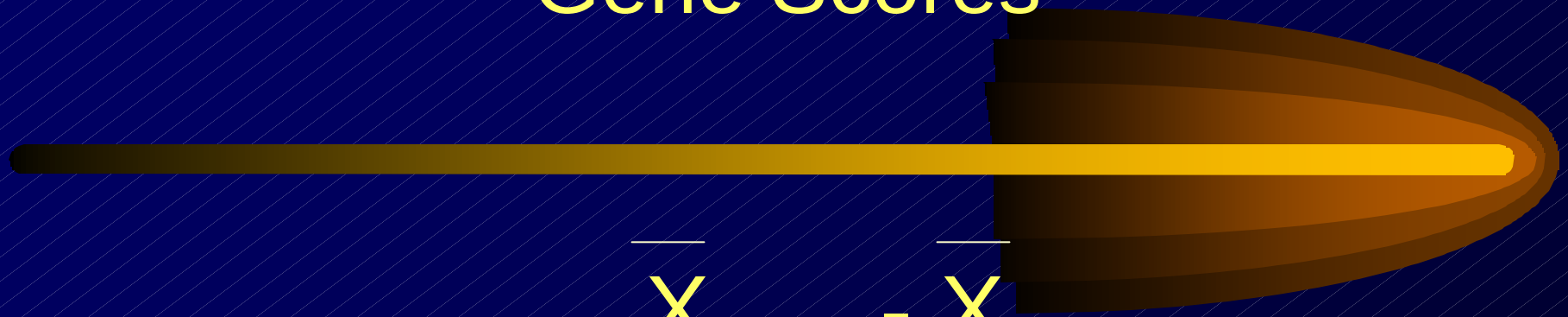
2. Use t-test as score for *each* gene.

3. Use resampling to evaluate Max t.

Characteristics of method?

1. Identifies individual genes.
2. Adjusts for multiple testing.
3. Preserves correlation structure.
4. Exact p-values, modulo simulation.

Gene Scores



$$T = \frac{\bar{X}_{ALL} - \bar{X}_{AML}}{S_p \sqrt{1/11 + 1/27}}$$

$$S_{Golub} = \frac{\bar{X}_{ALL} - \bar{X}_{AML}}{SD_{ALL} + SD_{AML}}$$

SAS proc multtest code

```
proc multtest data=gene.espress
          out=adjp stepperm holm
          n=10000 noprint;

     classes disease;

     test mean(gene1-gene7129);

     contrast “AML vs ALL” -1 1;
run;
```

SAS code (2)

```
proc sort data=adjp
          (where=(stppermp le .05));
    by raw_p;

proc print data=adjp
           (where=(stppermp le .05))
            noobs label;
    var _var_ raw_p  stpbon_p stppermp;
run;
```

Results (1)

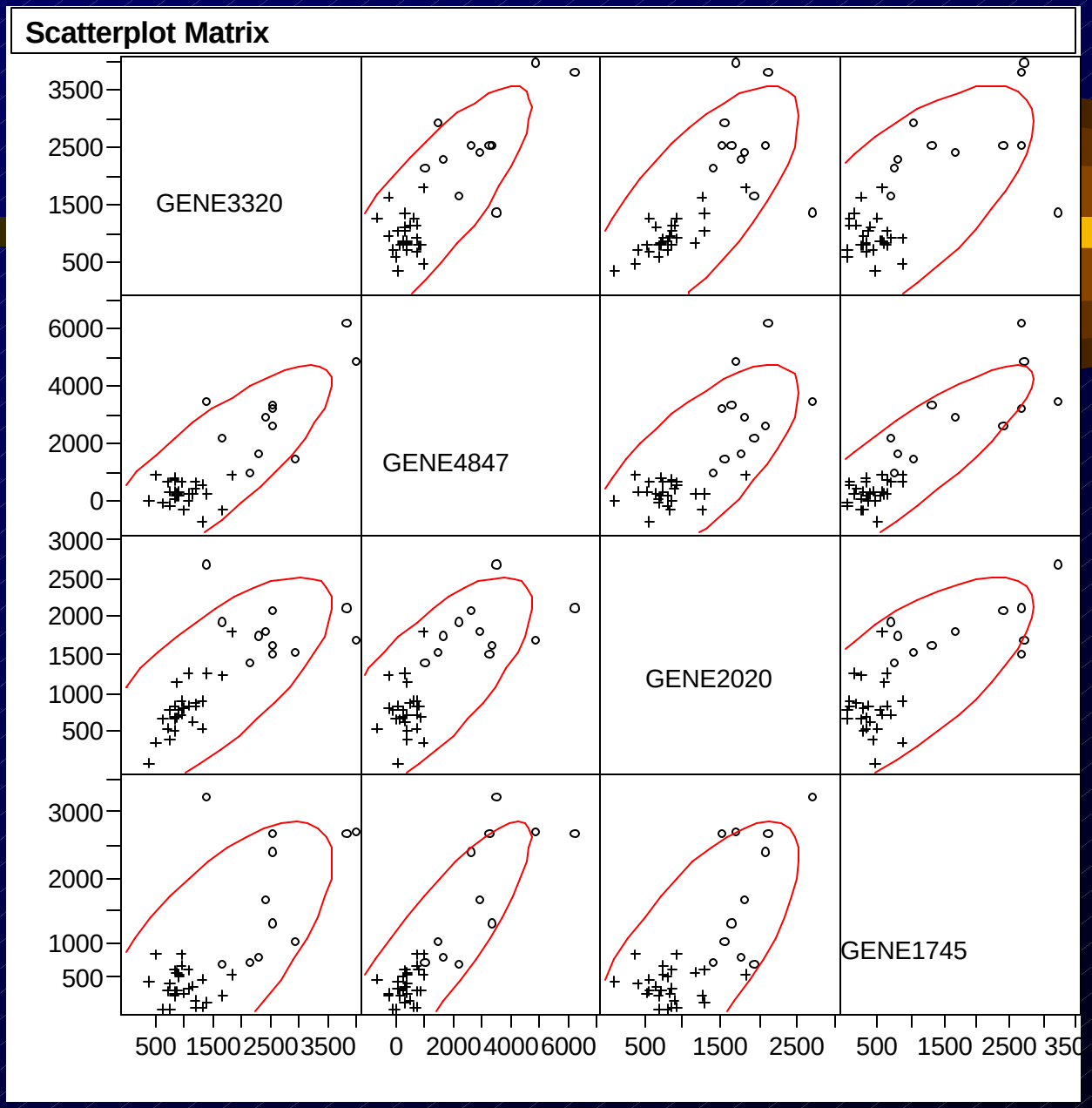
Gene	RawP	Holm	CMinP
GENE3320	1.38e-10	0.000001	0.0001
GENE4847	2.44e-10	0.000002	0.0001
GENE2020	6.58e-10	0.000005	0.0001
GENE1745	1 e- 8	0.000070	0.0004
GENE5039	1 e- 8	0.000072	0.0004
GENE1834	1.5 e- 8	0.000108	0.0005
GENE 461	3.6 e- 8	0.000257	0.0005
GENE4196	6.2 e- 8	0.000438	0.0009
GENE3847	7.2 e- 8	0.000510	0.0010

Results (1)

Gene	RawP	Holm	CMinP
GENE2288	8.90e-8	0.000635	0.0011
GENE1249	1.74e-7	0.001239	0.0017
GENE6201	1.76e-7	0.001250	0.0017
GENE2242	1.95e-7	0.001386	0.0020
GENE3258	2.11e-7	0.001500	0.0021
GENE1882	3.19e-7	0.002267	0.0024
GENE2111	3.66e-7	0.002606	0.0027
GENE2121	5.78e-7	0.004115	0.0041
GENE6200	6.23e-7	0.004428	0.0042
GENE6373	8.19e-7	0.005823	0.0058

Results (3)

Gene	RawP	Holm	CMinP
GENE6677	0.000003	0.024412	0.0196
GENE4052	0.000004	0.026268	0.0220
GENE1394	0.000005	0.034948	0.0282
GENE6405	0.000005	0.037980	0.0300
GENE248	0.000006	0.045267	0.0346
GENE2267	0.000006	0.046019	0.0352
GENE6041	0.000008	0.055335	0.0421
GENE6005	0.000008	0.056861	0.0428
GENE5772	0.000009	0.063771	0.0471
GENE6378	0.000010	0.067993	0.0500


Scatterplot Matrix
GENE3320
GENE4847
GENE2020
GENE1745
3500
2500
1500
500
6000
4000
2000
0
3000
2500
2000
1500
1000
500
3000
2000
1000
500
500 1500 2500 3500
0 2000 4000 6000
500 1500 2500
500 1500 2500 350

Current Research

Try some sort of linear combination of genes

connonical correlation-like? PLS? RP?

Q: Which Ys differeniate cancer type?
Q: How many real cancer types?
Find single gene then correlates to that gene.
Then find second orthogonal gene
that helps the prediction.