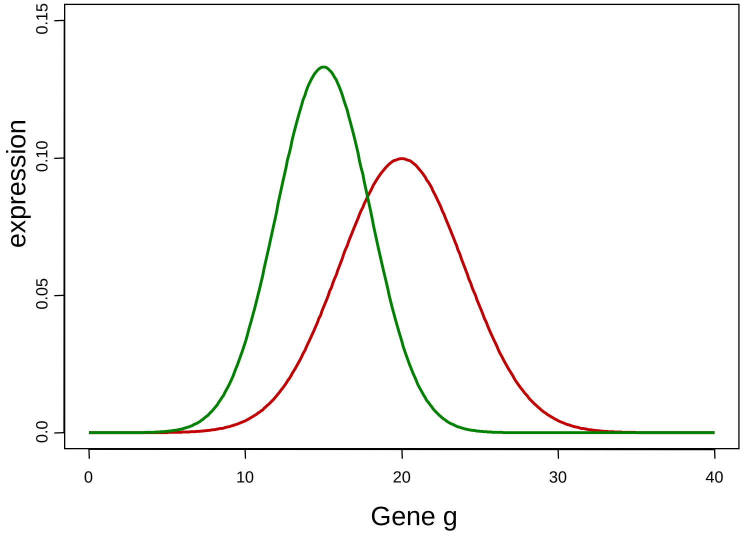
Multiple Testing

Test Hypothesis in Microarray Studies

Microarray studies

- aim to discover genes in biological samples that are differentially expressed under different experimental conditions
- aim at having high probability of declaring genes to be significantly expressed if they are truly expressed **(high power ~ low type II error risk)**, while keeping the probability of making false declarations of expression acceptably low **(controlling type I error risk)**



Lee & Whitmore (2002) Statistics in Medicine 21, 3543-3570

Multiple Testing

- Microarray studies typically involve the simultaneous study of thousands of genes, the probability of producing incorrect test conclusions (false positives and false negatives) must be controlled for the whole gene set.
- **for each gene there are two possible situations**
  - the gene is not differentially expressed, e.g. hypothesis **H_0** is true
  - the gene is differentially expressed at the level described by the alternative hypothesis **H_A**
- **test declaration (decision)** - the gene is differentially expressed (**H_0** rejected)
  - the gene is unexpressed (**H_0** not rejected)

	test declaration	
true hypothesis	**unexpressed** **(H_0 not rejected)**	**expressed** **(H_0 rejected)**
unexpressed (H_0)	true negative	**false positive** **(type I error α)**
expressed (H_A)	**false negative** **(type II error β)**	true positive

Lee & Whitmore (2002) Statistics in Medicine 21, 3543-3570

Multiple Testing

Testing simultaneously **G** hypothesis $H_1,\ldots, H_G$, G_0 of these hypothesis are true

	# not rejected hypothesis	# rejected hypothesis	
# true hypothesis (unexpressed genes)	U	V	G_0
# false hypothesis (expressed genes)	T	S	$G - G_0$
total	$G - R$	R	G

- counts ***U***, ***V***, ***S***, ***T*** are random variables in advance of the analysis of the study data
- **observed random variable *R* = number of rejected hypothesis**
- ***U***, ***V***, ***S***, ***T*** not observable random variables
- ***V* = number of type I errors (false positives)**

  ***T* = number of type II errors (false negatives)**

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Type I and II Error Rates

	# not rejected hypothesis	# rejected Hypothesis	
# true hypothesis (unexpressed genes)	U	V	G_0
# false hypothesis (expressed genes)	T	S	$G - G_0$
total	$G - R$	R	G

α_0 = probability of type I error for any gene = $E(V)/G_0$

β_1 = probability of type II error for any gene = $E(T)/(G-G_0)$

α_F = family-wise error rate (FWER) = $P(V > 0)$ (probability of at least one type I error)

False discovery rate (FDR) (Benjamini & Hochberg, 1995)

= expected proportion of false positives among the rejected hypothesis

$$FDR = E(Q), \qquad Q = \begin{cases} V/R & : R > 0 \\ 0 & : R = 0 \end{cases}$$

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Strong vs. weak control

- expectations and probabilities are conditional on which hypothesis are true

- **strong control:**

  control of the Type I error rate under any combination of true and false hypotheses, i.e., any value of $\mathbf{G_0}$

$$\bigcap\nolimits_{g \in \mathbf{G}_0} H_g, \quad \text{for all} \quad \mathbf{G}_0 \subseteq \{1, \ldots, G\}, \ |\mathbf{G}_0| = G_0$$

- **weak control:**

  control of the Type I error rate only when **all** hypothesis are true, i.e. under the complete null-hypothesis

$$H_0^C = \bigcap\nolimits_{g=1}^{G} H_g, \quad \text{with} \ G_0 = G$$

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Notations

For hypothesis H_g, $g = 1,\ldots, G$:

observed test statistics t_g

observed unadjusted p-values p_g

Ordered p-values and test statistics:

$$\{r_g\}_{g=1,\ldots,G}$$

$$p_{r_1} \le p_{r_2} \le \ldots \le p_{r_G}$$

$$|t_{r_1}| \ge |t_{r_2}| \ge \ldots \ge |t_{r_G}|$$

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Control of the family-wise error rate (FWER)

observed p-values	Bonferroni	Holm Step-down	Hochberg Step-up
p_{r_1}	α/G	α/G	α/G
p_{r_2}	α/G	$\alpha/(G\text{-}1)$	$\alpha/(G\text{-}1)$
:	:	:	:
p_{r_g}	α/G	$\alpha/(G\text{-}g+1)$	$\alpha/(G\text{-}g+1)$
:	:	:	:
$p_{r_{G-1}}$	α/G	$\alpha/2$	$\alpha/2$
p_{r_G}	α/G	α	α

Control of the family-wise error rate (FWER)

1. single-step Bonferroni procedure

reject H_g with $p_g \leq \alpha/G$, adjusted p-value $\tilde{p}_g = \min(G \cdot p_g, 1)$

2. Holm (1979) – step-down procedure

$g^* = \min\{g : p_{r_g} > \alpha/(G-g+1)\}$, reject H_g for $g = 1, \ldots, g^* - 1$,

adjusted p-value $\tilde{p}_{r_g} = \max_{k=1,\ldots,g}\{\min((m-k+1)p_{r_k}, 1)\}$

3. Hochberg (1988) – step-up procedure

$g^* = \max\{g : p_{r_g} \leq \alpha/(G-g+1)\}$, reject H_g for $g = 1, \ldots, g^*$,

adjusted p-value $\tilde{p}_{r_g} = \min_{k=g,\ldots,m}\{\min((G-k+1)p_{r_k}, 1)\}$

4. Single-step Šidák procedure

adjusted p-value $\tilde{p}_g = 1-(1-p_g)^G$

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Resampling

Estimate joint distribution of the test statistics $T_1,...,T_G$ under the complete null hypothesis H_0^C by permuting the columns of the Gene expression data matrix **X.**

Permutation algorithm for non-adjusted p-values

For the b-th permutation, $b = 1,...,B$

1. Permute the n columns of the data matrix **X.**
2. Compute test statistics $t_{1,b}$, ..., $t_{G,b}$ for each hypothesis.

The permutation distribution of the test statistic T_g for hypothesis $H_{g\,,}$ $g=1,...,G,$ is given by the empirical distribution of $t_{g,1}$, ... , $t_{g,B}$.

For two-sided alternative hypotheses, the permutation p-value for hypothesis H_g is

$$p_g^* = \tfrac{1}{B}\sum_{b=1}^{B} I(|\,t_{g,b}\,| \geq |\,t_j\,|)$$

where I(.) is the indicator function, equaling 1 if the condition in parenthesis is true, and 0 otherwise.

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Control of the family-wise error rate (FWER)

Permutation algorithm of Westfall & Young (1993)

- step-down procedure without assuming *t* distribution of the test statistics for each gene's differential expression
- adjusted p-values directly estimated by permutation
- strong control of FWER
- takes dependency structure of hypotheses into account

Control of the family-wise error rate (FWER)

Permutation algorithm of Westfall & Young (maxT)

- Order observed test statistics: $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_G}|$

- for the b-th permutation of the data (*b* = 1,...,*B*):

- divide data into artificial control and treatment group
- compute test statistics ***t***$\boldsymbol{t_{1b}}$**, ... ,** $\boldsymbol{t_{Gb}}$
- compute successive maxima of the test statistics

$$u_{G,b} = |t_{r_G,b}|$$

$$u_{g,b} = \max\{u_{g+1,b}, |t_{r_g,b}|\} \quad \text{für} \quad g = G-1, \ldots, 1$$

- compute adjusted p-values:

$$\tilde{p}^*_{r_g} = \tfrac{1}{B} \sum_{b=1}^{B} I(u_{g,b} \geq |t_{r_g}|)$$

Dudoit et al. (2002) Multiple Hypothesis Testing in Microarray Experiments, Technical Report

Control of the family-wise error rate (FWER)

Permutation algorithm of Westfall &Young – Example

gene	|t|	
1	0.1	t_{r_G}
4	0.2	$t_{r_{G-1}}$
5	2.8	:
2	3.4	t_{r_2}
3	7.1	t_{r_1}

sort observed values

gene	$\|t_b\|$	u_b	$I(u_b>\|t\|)$
1	1.3	1.3	1
4	0.8	1.3	1
5	3.0	3.0	1
2	2.1	3.0	0
3	1.8	3.0	0

B=1000 permutations

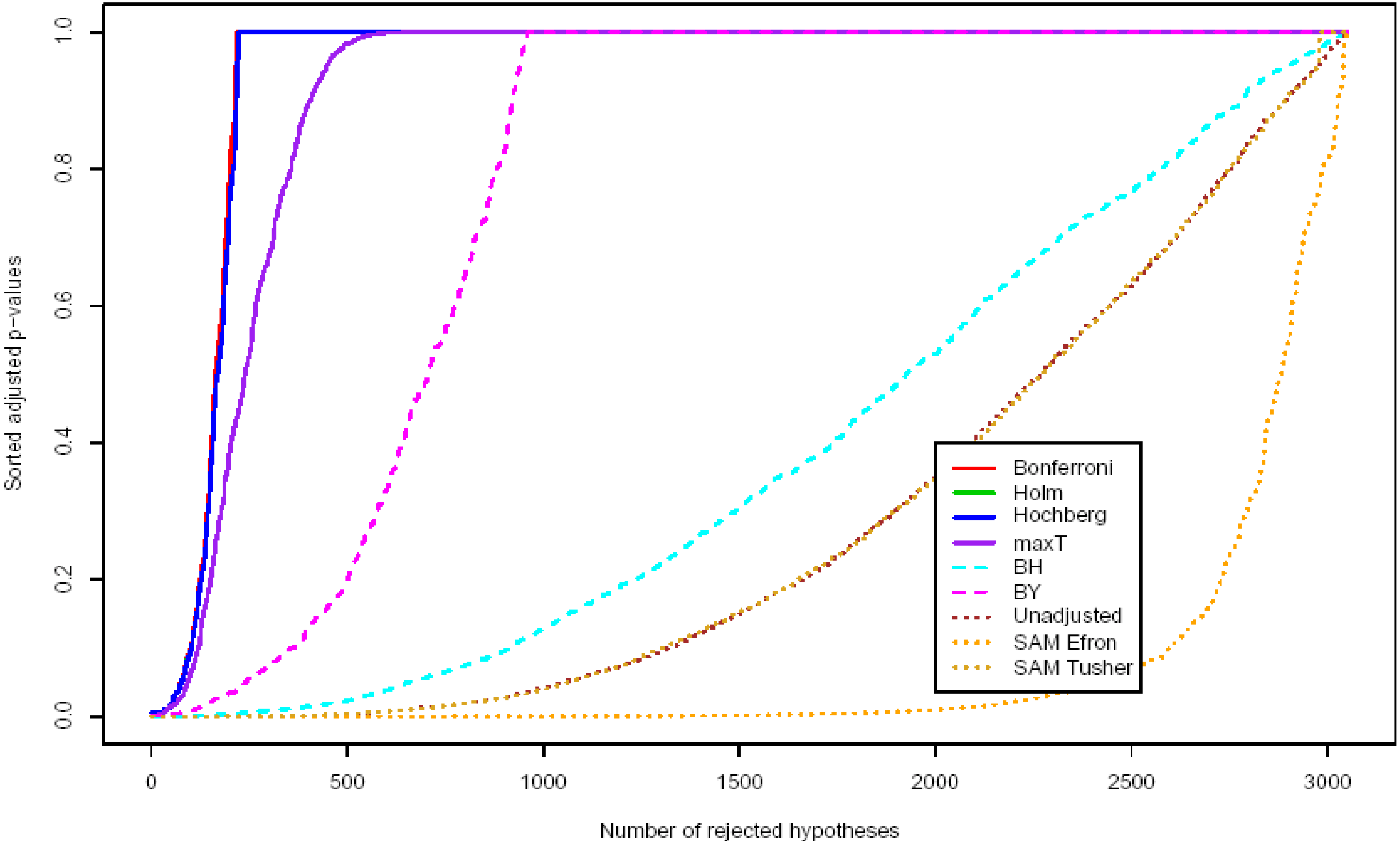
$\sum$	$\tilde{p} = \sum / B$
935	0.935
876	0.876
138	0.138
145	0.145
48	**0.048**

adjusted p-values

Example: Leukemia study, Golub et al. (1999)

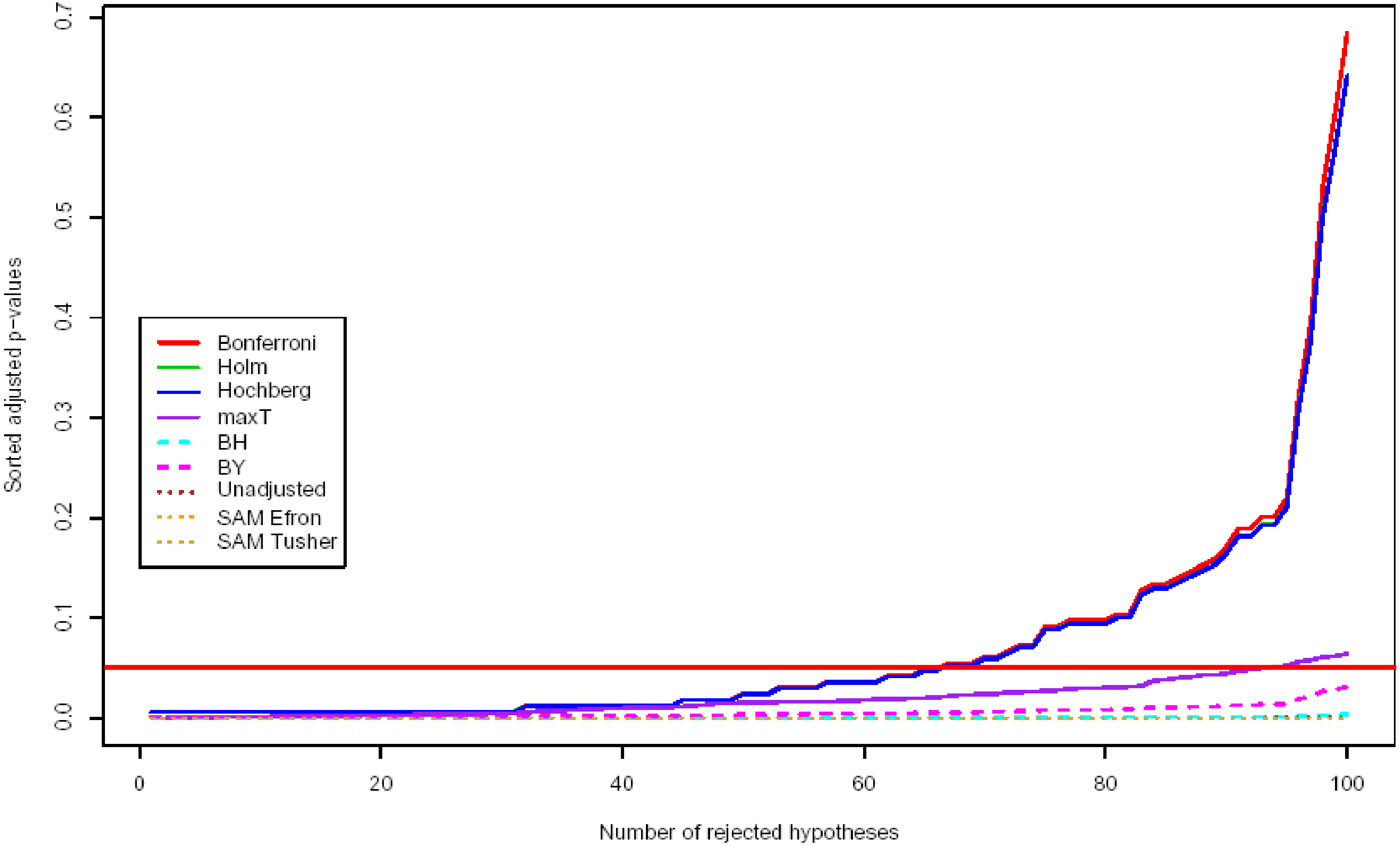
- patients with ALL (acute lymphoblastic leukemia) $n_1=27$
  AML (acute myeloid leukemia) $n_2=11$
- Affy-Chip: 6817 genes
- reduction to 3051 genes according to certain exclusion criteria for expression values

Example: Leukemia study, Golub et al. (1999)



Dudoit et al. (2002)

Example: Leukemia study, Golub et al. (1999)



Dudoit et al. (2002)

Control of the False Discovery Rate (FDR)

- While in some cases FWER control is needed, the multiplicity problem in microarray data does not require a protection against against even a single type I error, so that the serve loss of power involved in such protection is not justified.

- Instead, it may be more appropriate to emphasize the proportion of errors among the identified differentially expressed genes.
  The expectation of this proportion is the **False Discovery Rate (FDR)**.

$$FDR = E(Q), \qquad Q = \begin{cases} V/R & : R > 0 \\ 0 & : R = 0 \end{cases}$$

R **= number of rejected hypothesis**

V **= number of type I errors (false positives)**

Reiner, Yekutieli & Benjamini (2003) Bioinformatics 19, 368-375

Control of the False Discovery Rate (FDR)

1. Linear step-up procedure (Benjamini & Hochberg, 1995)

$$g^* = \max\{g : p_{r_g} \le \tfrac{g}{G} q\}, \qquad \text{reject } H_g \;\; \text{for } g = 1,\dots,g^*,$$

$$\text{adjusted p-value} \qquad \tilde{p}_{r_g} = \min_{k=g,\dots,G}\{\min(\tfrac{G}{k} p_{r_k}, 1)\}$$

- controls FDR at level *q* for independent test statistics $FDR \le q \cdot \frac{G_0}{G} \le q$

2. Benjamini & Yekutieli (2001)

- procedure 1 controls the FDR under certain dependency structures (positive regression dependency)
- step-up procedure for more general cases (replace *q* by $q / \sum_{i=1}^{G} 1/i$)

$$g^* = \max\left\{g : p_{r_g} \le q \cdot g / (G \sum_{i=1}^{G} 1/i)\right\}, \quad \text{reject } H_g \;\; \text{for g} = 1,\dots,g^*,$$

$$\text{adjusted p-value} \qquad \tilde{p}_{r_g} = \min_{k=g,\dots,G}\left\{\min(p_{r_k} \tfrac{G}{k} \sum_{i=1}^{G} 1/i, 1)\right\}$$

- this modification may be to conservative for the microarray problem

Reiner, Yekutieli & Benjamini (2003) Bioinformatics 19, 368-375

Control of the False Discovery Rate (FDR)

3. Adaptive procedures (Benjamini & Hochberg, 2000)

- try to estimate G_0 and use $q^{*}=q\ G_0/G$ instead of q in procedure 1 to gain more power

- Storey (2001) suggests a similar version to estimate G_0, which are implemented in SAM (Storey & Tibshirani, 2003)

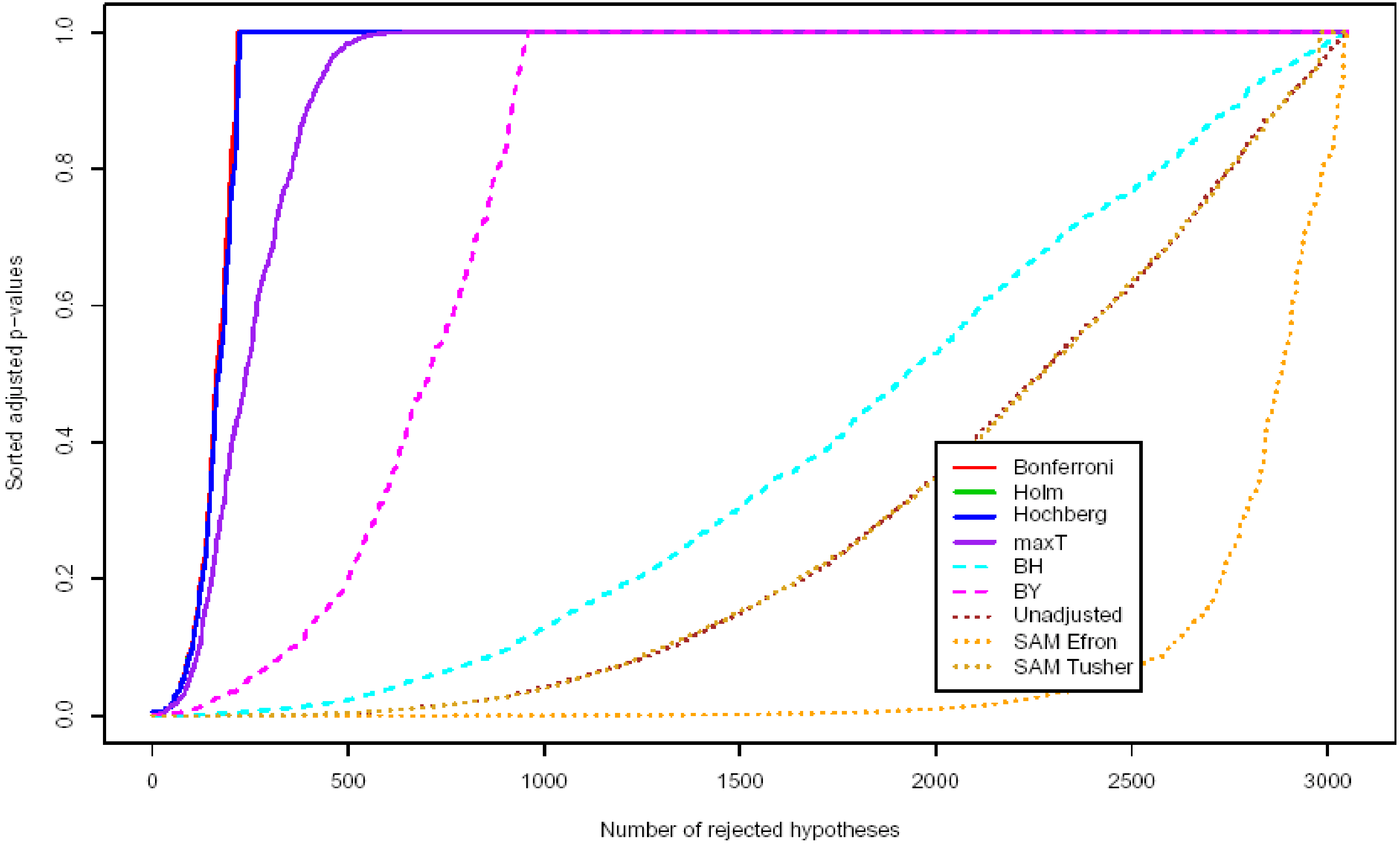
- adaptive methods offer better performance only by utilizing the difference between G_0/G and 1, if the difference is small, i.e. when the potential proportion of differentially expressed genes is small, they offer little advantage in power while their properties are not well established.

4. Resampling FDR adjustments

- Yekutieli & Benjamini (1999) *J. Statist. Plan. Inference 82, 171-196*
- Reiner, Yekutieli & Benjamini (2003) *Bioinformatics 19, 368-375*

Reiner, Yekutieli & Benjamini (2003) Bioinformatics 19, 368-375

Example: Leukemia study, Golub et al. (1999)



Dudoit et al. (2002)

Example: Apo AI Exp., Callow et al. (2000)

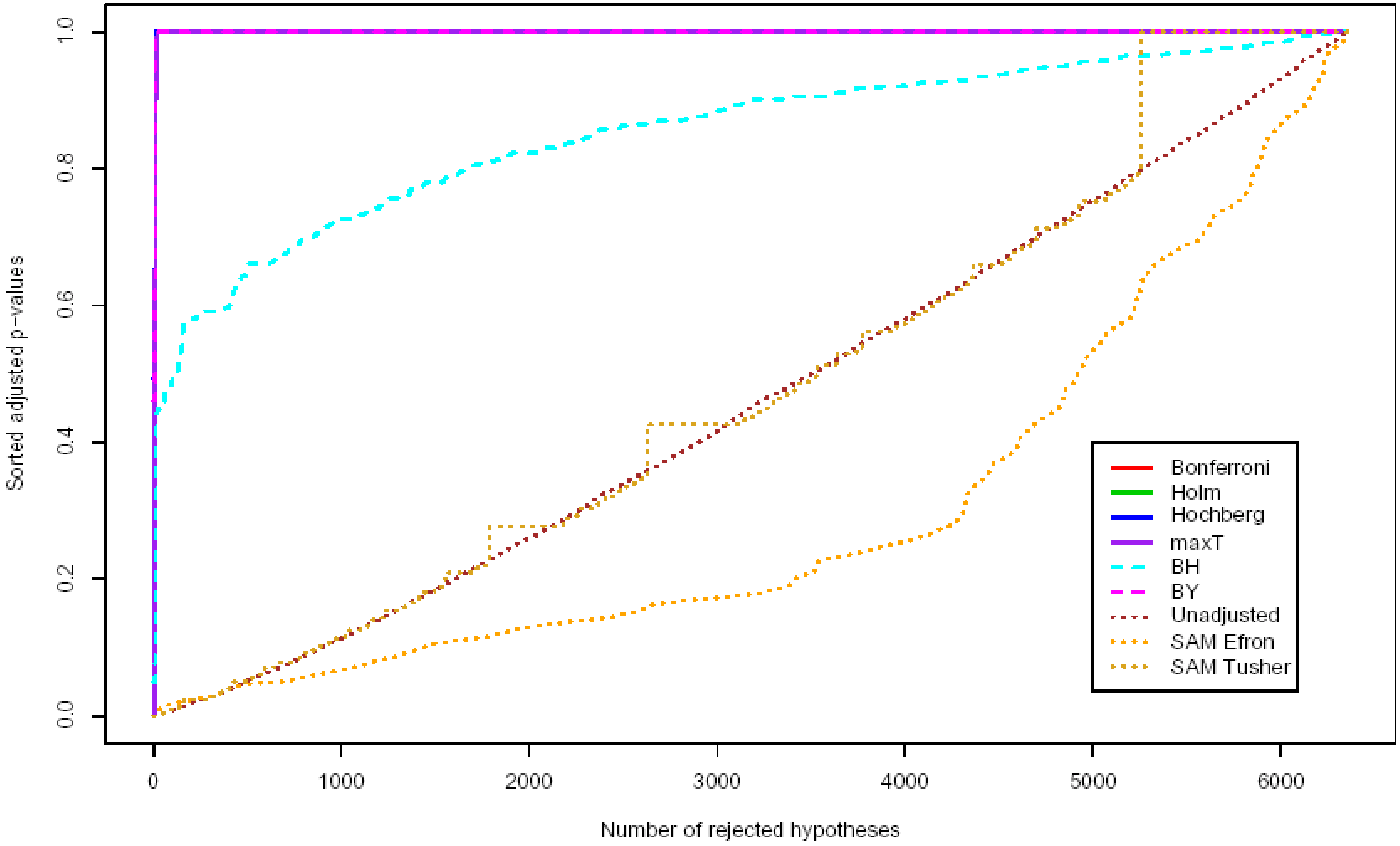
Apolipoprotein A1 (Apo A1) experiment in mice

- aim: identification of differentially expressed genes in liver tissues
- **experimental group:** 8 mice with apo A1-gene knocked out (apo A1 KO)
- **control group:** 8 C57B1/6 mice

- **experimental sample:** cDNA for each of the 16 mice ⇒ labeled with **red (Cy5)**
- **reference-sample:** pooled cDNA of the 8 control mice ⇒ labeled with **green (Cy3)**

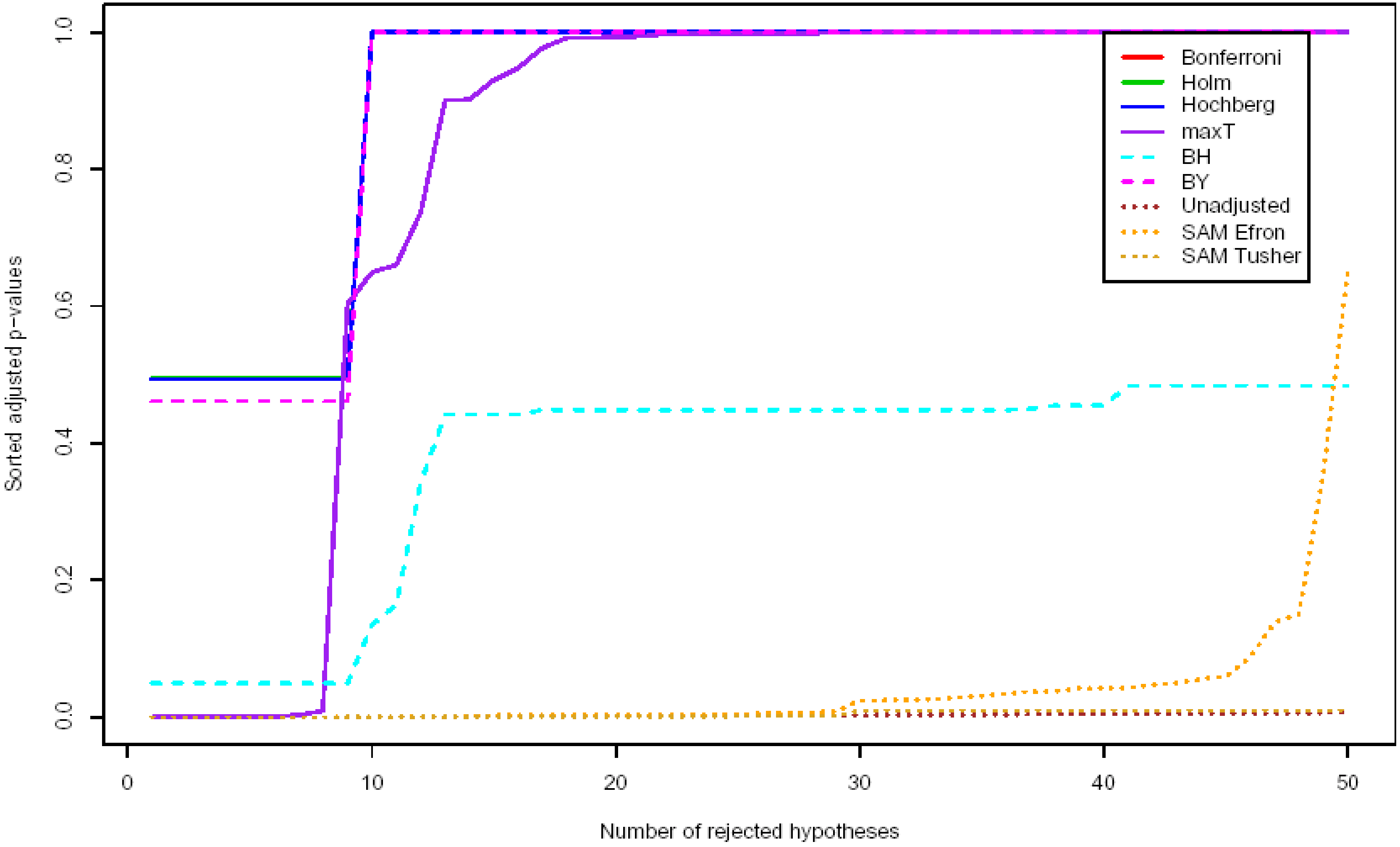
- cDNA Arrays with 6384 cDNA probes, 200 related to lipid-metabolism
- 16 hybridizations overall

Example: Apo AI Exp., Callow et al. (2000)



Dudoit et al. (2002)

Beispiel 2: Apo AI Exp., Callow et al. (2000)



Dudoit et al. (2002)

Multiple Testing - Summary

- For multiple testing problems there are several methods to control the family-wise error rate (FWER).

- FDR controlling procedures are promising alternatives to more conservative FWER controlling procedures.

- Strong control of the type one error rate is essential in the microarray context.

- Adjusted p-values provide flexible summaries of the results from a multiple testing procedure and allow for a comparison of different methods.

- Substantial gain in power can be obtained by taking into account the joint distribution of the test statistics
  (e.g. Westfall & Young, 1993; Reiner, Yekutieli & Benjamini 2003).

- **Recommended software: Bioconductor R *multtest* package (http://www.bioconductor.org/)**

Adapted from S. Dudoit, Bioconductor short course 2002

Multiple Testing - Literature

- **Benjamini, Y. & Hochberg, Y. (1995).** Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. B* **57**: 289-300.

- **Benjamini,Y. and Hochberg,Y. (2000)** On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.*, **25**, 60–83.

- **Benjamini,Y. and Yekutieli,D. (2001b)** The control of the false discovery rate under dependency. *Ann Stat.* **29**, 1165–1188.

- **Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P. & Rubin, E. M. (2000).** Microarray expression profiling identifies genes with altered expression in HDL deficient mice, *Genome Research* **10**(12): 2022-2029.

- **S. Dudoit, J. P. Shaffer, and J. C. Boldrick (Submitted).** Multiple hypothesis testing in microarray experiments, Technical Report #110 (http://stat-www.berkeley.edu/users/sandrine/publications.html)

- **Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek,M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomeld, C. D. & Lander, E. S. (1999).** Molecular classication of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**: 531-537.

Multiple Testing - Literature

- **Hochberg, Y. (1988).** A sharper bonferroni procedure for multiple tests of significance, *Biometrika* **75**: 800- 802.
- **Holm, S. (1979).** A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* **6**: 65-70.
- **M.-L. T. Lee & G.A. Whitmore (2002)** Power and sample size for DNA microarray studies. *Statistics in Medicine* 21, 3543-3570.
- **A. Reiner, D. Yekutieli & Y. Benjamini (2003)** Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19, 368-375
- **Westfall, P. H. & Young, S. S. (1993).** *Resampling-based multiple testing: Examples and methods for p-value adjustment*, John Wiley & Sons.
- **Yekutieli,D. and Benjamini,Y. (1999)** Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan Infer.*, **82**, 171–196.

2 x 2 Factorial Experiments

Two experimental factors, e.g. **treatment** (untreated T -, treated T +)

strain (knock out KN, wild-type WT)

Linear model

$$\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)}$$

$$x_1 = \begin{cases} 0 & : strain = KN \\ 1 & : strain = WT \end{cases}$$

$$x_2 = \begin{cases} 0 & : treatment = T - \\ 1 & : treatment = T + \end{cases}$$

β_1 - strain effect

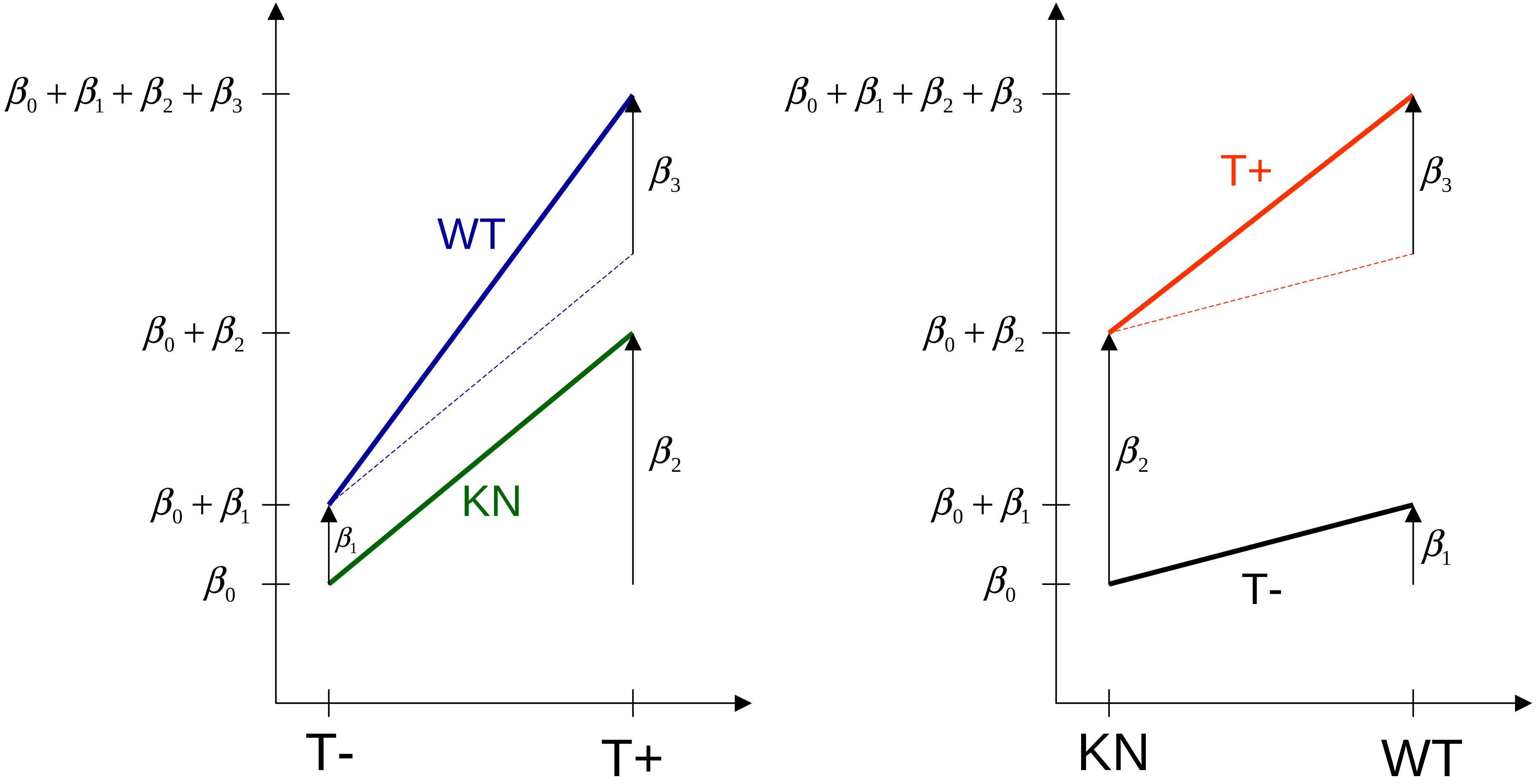
β_2 - treatment effect

β_3 - interaction effect of strain and treatment

		strain	
		KN	**WT**
treatment	**T -**	β_0	$\beta_0 + \beta_1$
	T+	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_3$

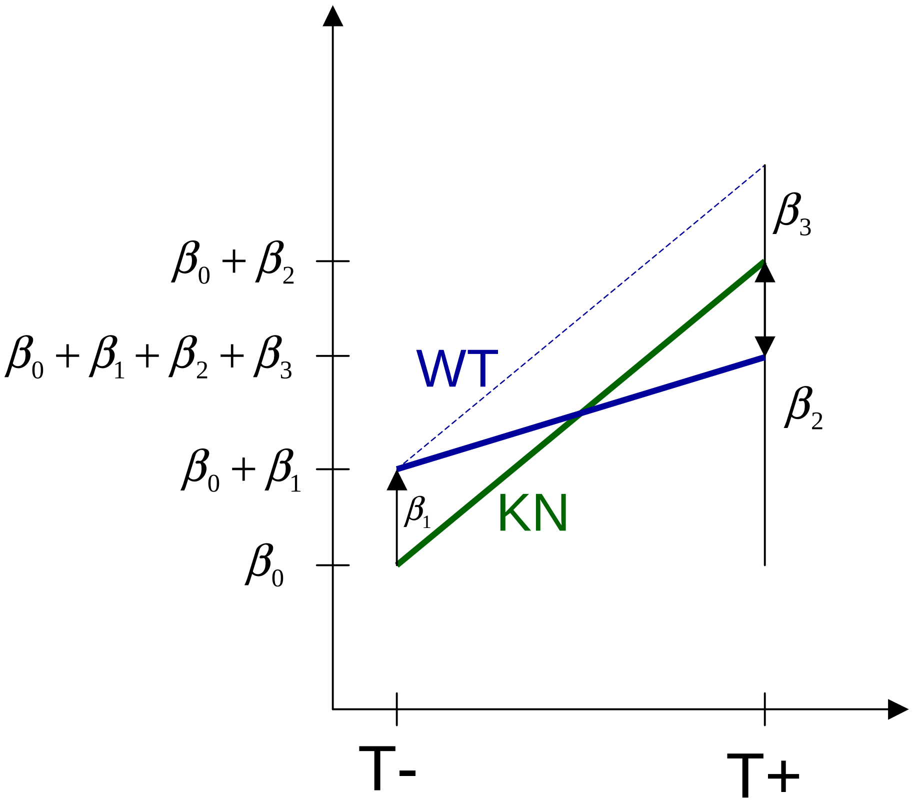
2 x 2 Factorial Experiments

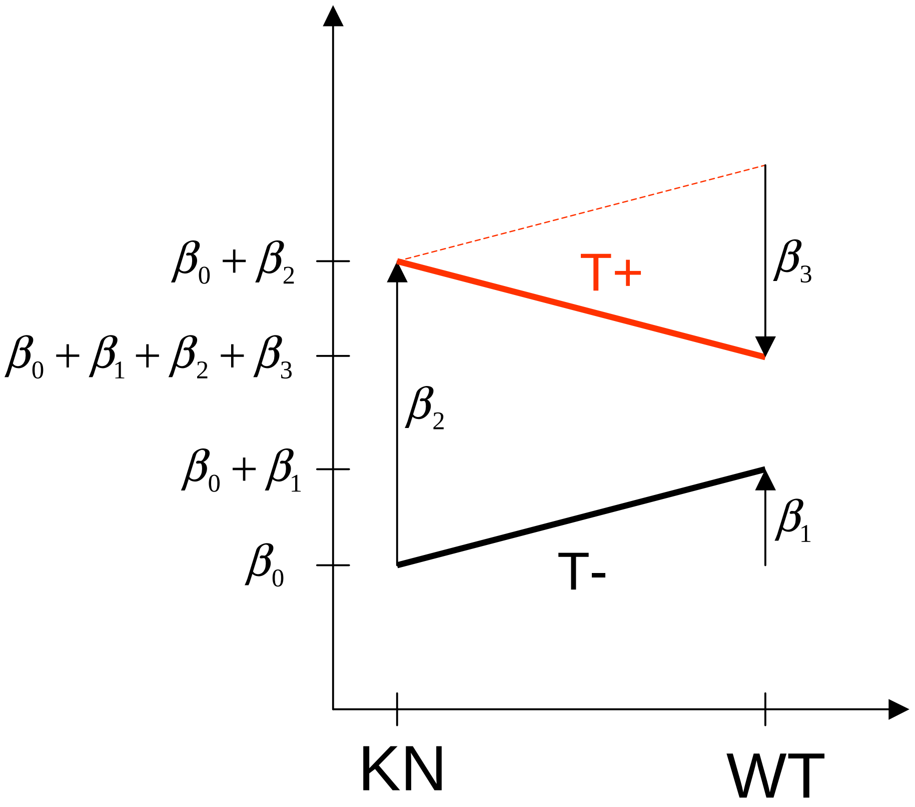
$\boxed{\beta_3 > 0}$



2 x 2 Factorial Experiments

$$\boxed{\beta_3 < 0}$$





2 x 2 Factorial Experiments

H_0: $\beta_3 = 0$

- effect of strain is independent of treatment or
- effect of treatment is independent of strain or
- strain and treatment are additive

H_A: $\beta_3 \neq 0$

- treatment interacts with strain
- treatment modifies effect of strain
- strain modifies effect of treatment
- treatment and strain are nonadditive

H_0: $\beta_1 = \beta_3 = 0$

- strain is not associated with expression *Y*

H_A: $\beta_1 \neq 0$ or $\beta_3 \neq 0$

- strain is associated with expression *Y*
- strain is associated with expression *Y* for either T- or T+

H_0: $\beta_2 = \beta_3 = 0$

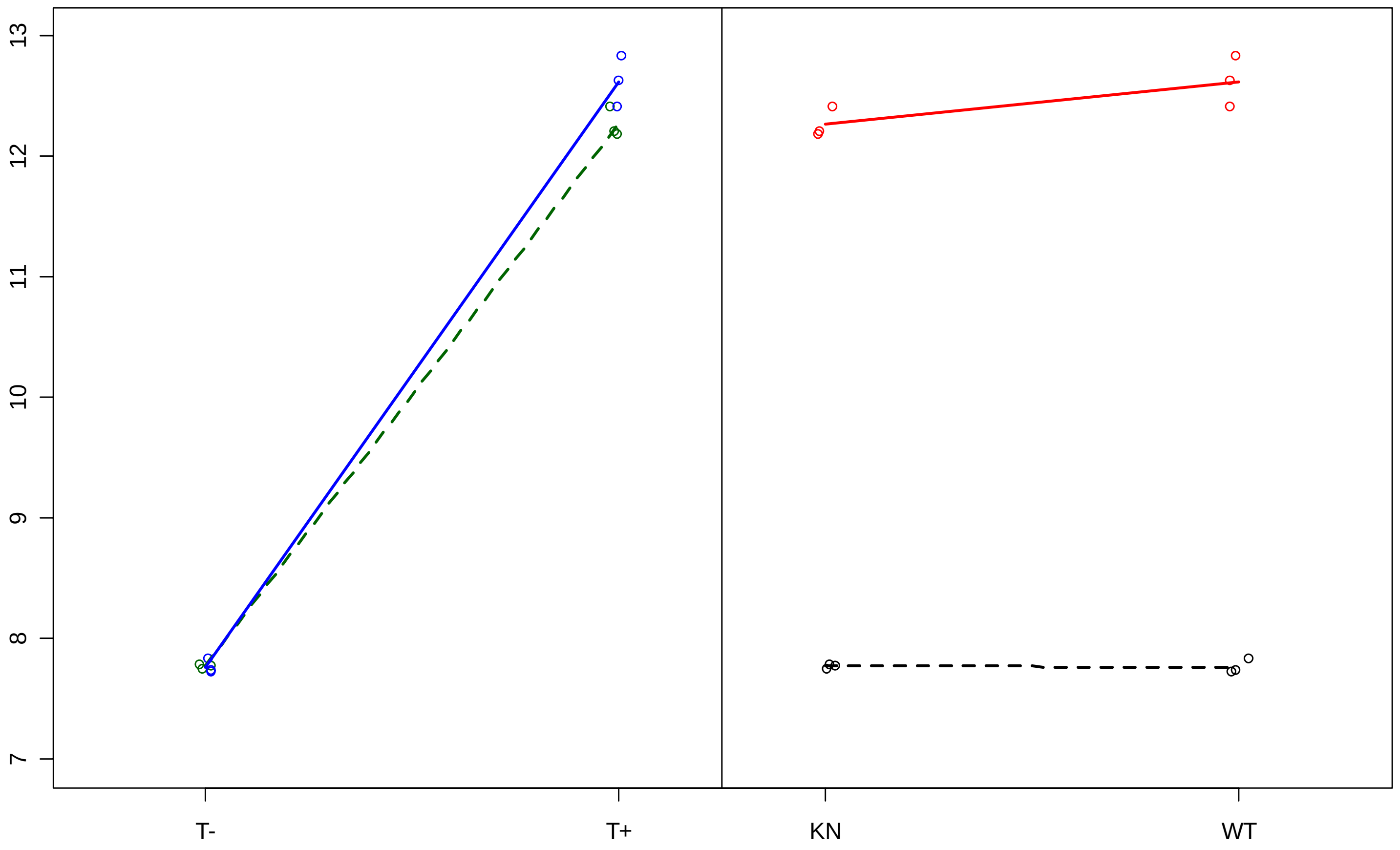
- treatment is not associated with expression *Y*

H_A: $\beta_2 \neq 0$ or $\beta_3 \neq 0$

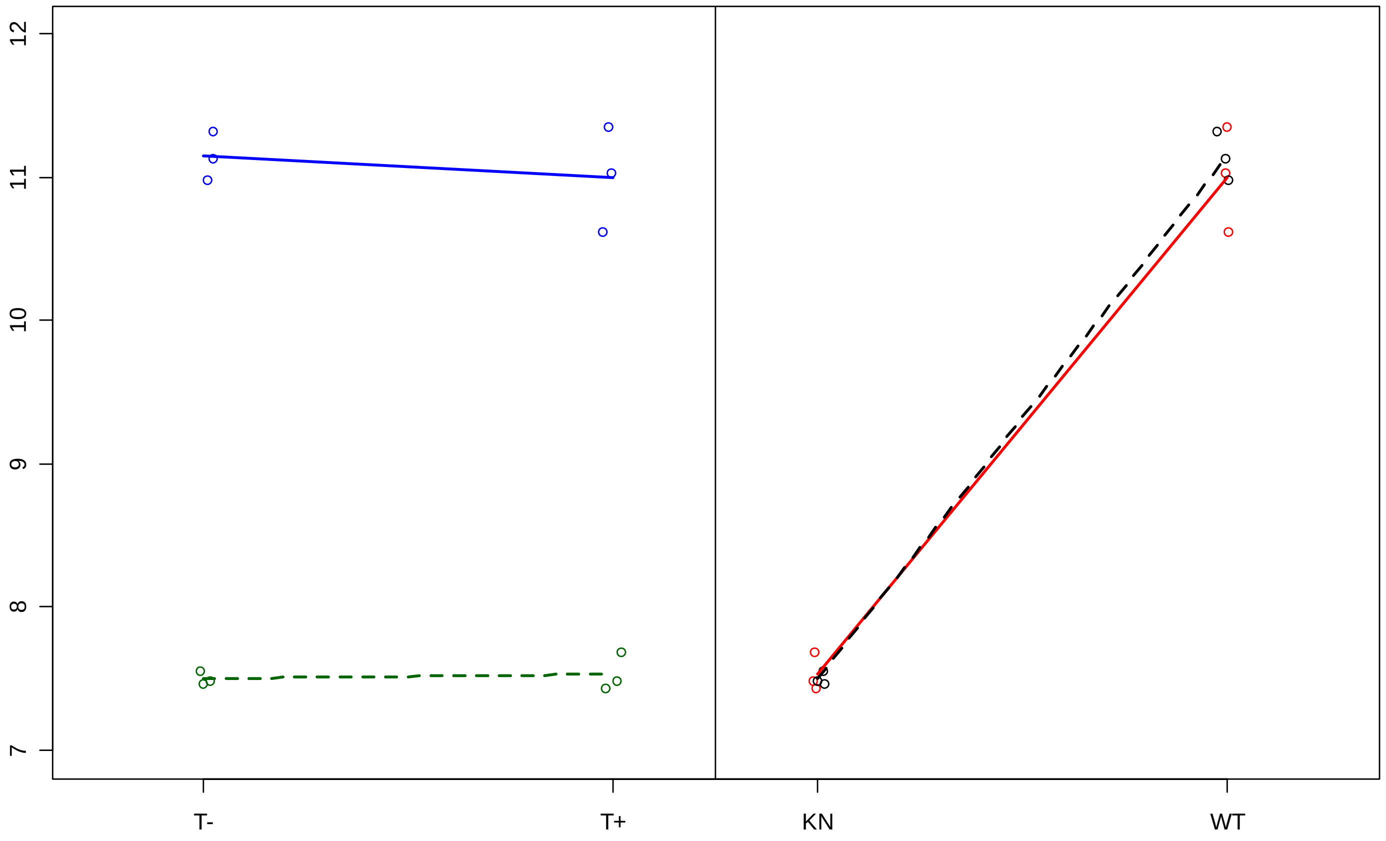
- treatment is associated with expression *Y*
- treatment is associated with expression *Y* for either KN or WT

F.E. Harrell, Jr. (2001) Regression Modeling Strategies, Springer

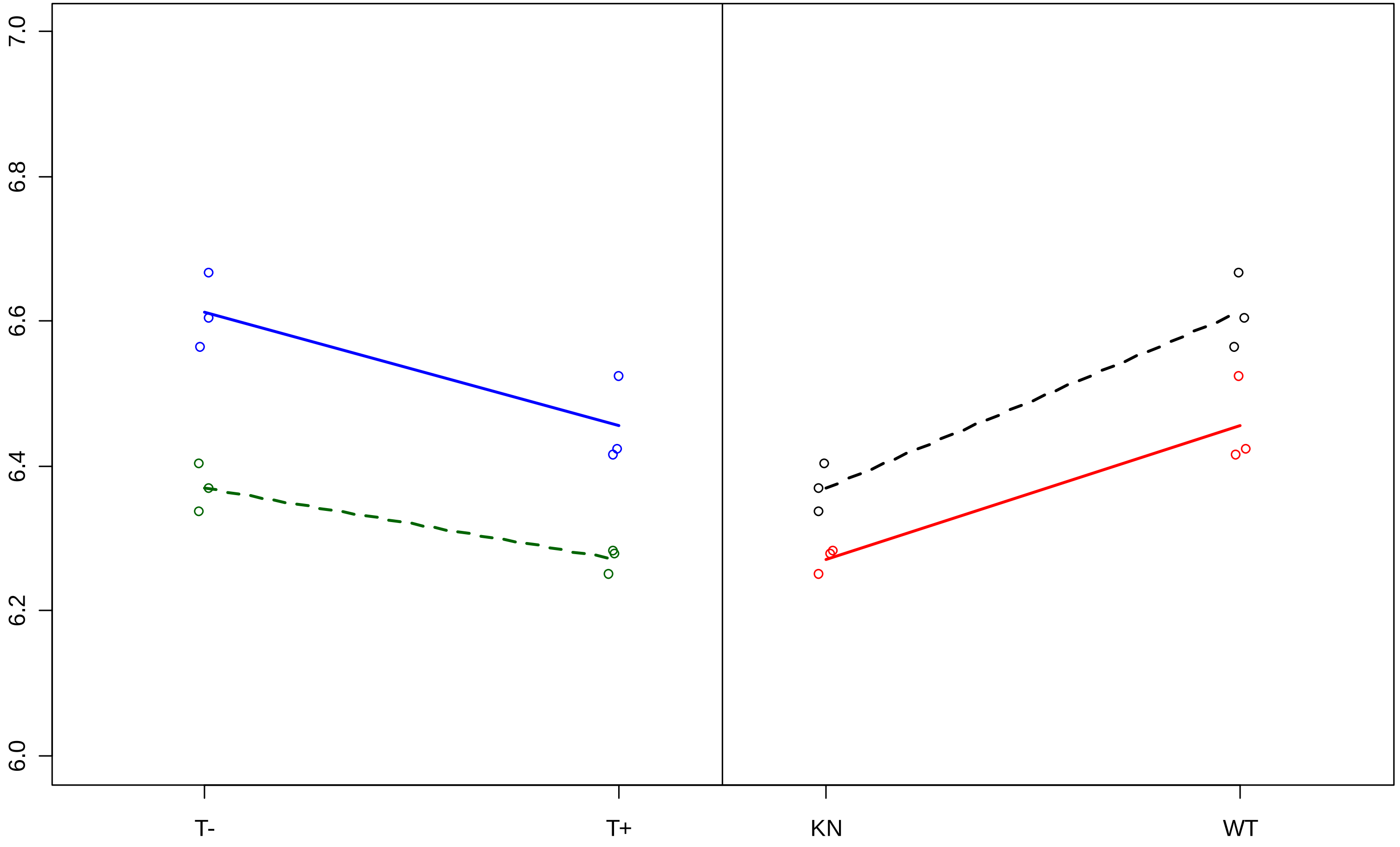
2 x 2 Factorial Experiments - Treatment effect



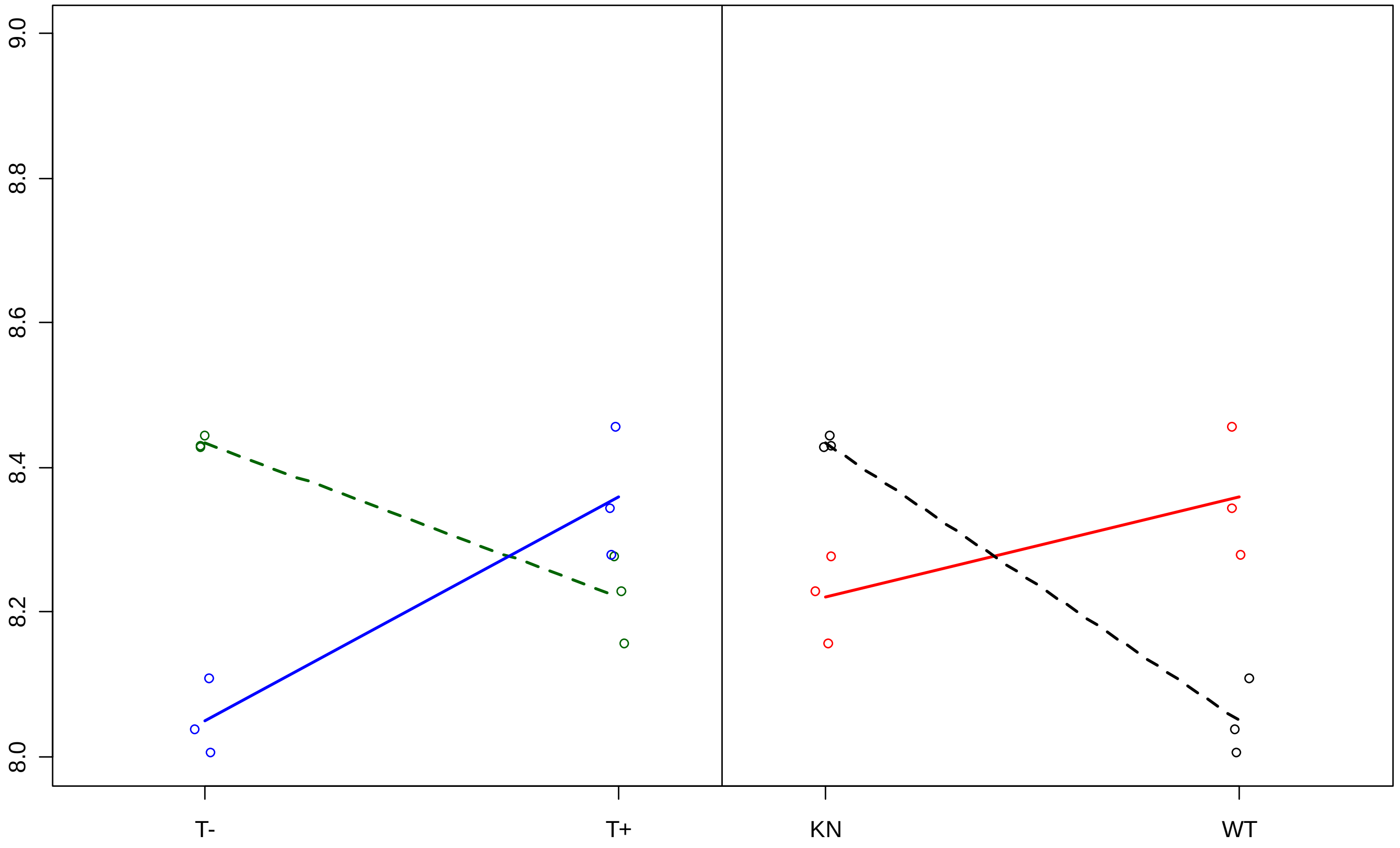
2 x 2 Factorial Experiments - Strain effect



2 x 2 Factorial Experiments - Strain effect



2 x 2 Factorial Experiments - Interaction effect


T-
T+
KN
WT