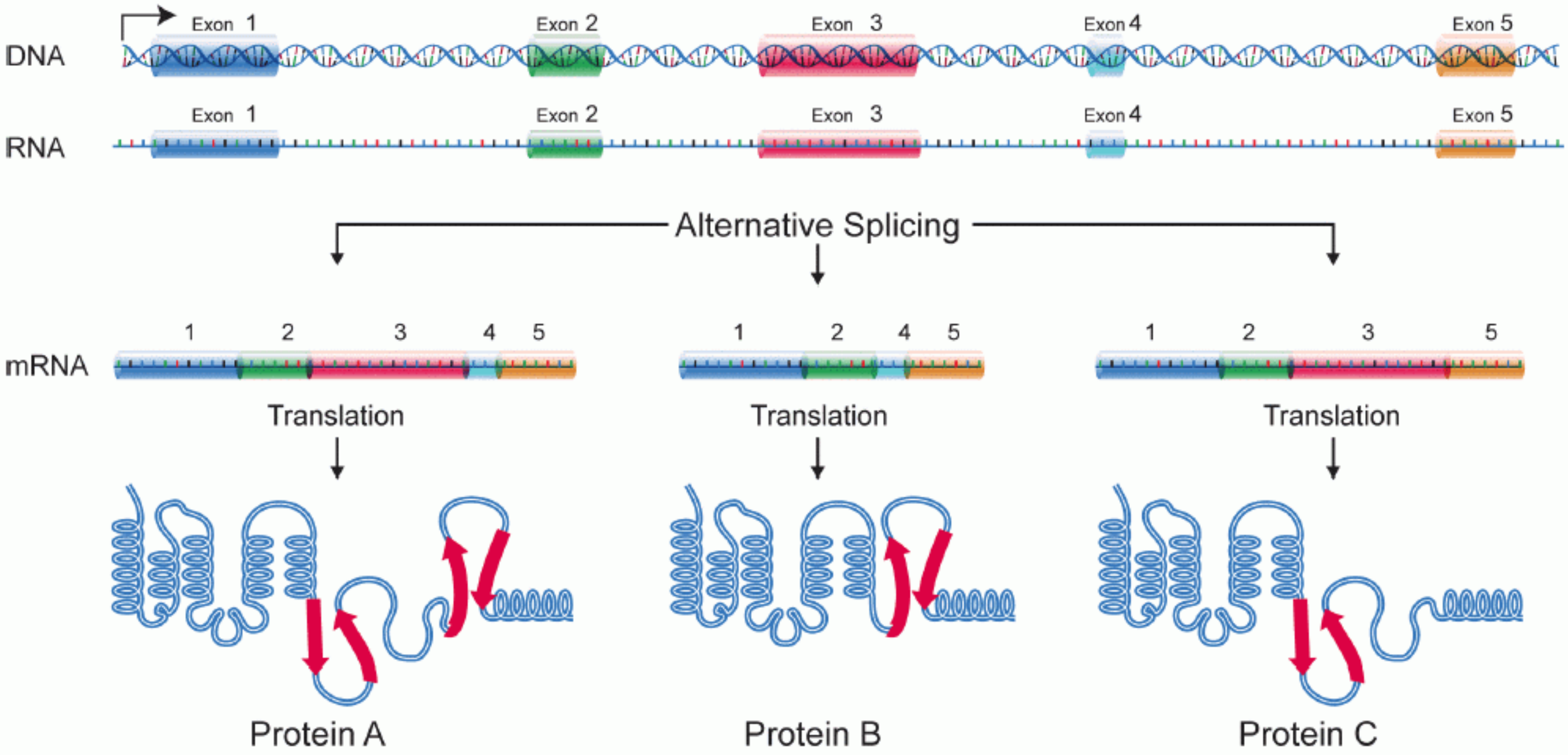
Differential expression analysis

Charlotte Soneson

Friedrich Miescher Institute for Biomedical Research &
SIB Swiss Institute of Bioinformatics

Hinxton, April 9, 2019


DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

Differential analysis types for RNA-seq

- Does the total output of a gene change between conditions? **DGE**
- Does the expression of individual transcripts change? **DTE**
- Does *any* isoform of a given gene change? **DTE+G**
- Does the isoform composition for a given gene change? **DTU/DIU/DEU**
- (Does *anything* change? GDE*)

- need **different** abundance quantification of transcriptomic features (genes, transcripts, exons)

*https://liorpachter.wordpress.com/2018/02/15/gde%C2%B2-dge%C2%B2-dtu%C2%B2-dte%E2%82%81%C2%B2-dte%E2%82%82%C2%B2/

Differential expression analysis

- Input: expression/abundance matrix
  (features x samples) + grouping/sample annotation

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	693	451	887	416	1148	1069	774	581
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	466	515	623	364	590	794	419	510
ENSG00000000457	326	274	372	223	356	450	308	297
ENSG00000000460	91	75	61	48	110	95	100	82
ENSG00000000938	0	0	2	0	1	0	0	0

- Output: result table (one line per feature)

	logFC	logCPM	LR	PValue	FDR
ENSG00000109906	-5.882117	4.120149	924.1622	5.486794e-203	3.493826e-198
ENSG00000165995	-3.236681	4.603028	576.1025	2.641667e-127	8.410672e-123
ENSG00000189221	-3.316900	6.718559	562.9594	1.909251e-124	4.052512e-120
ENSG00000120129	-2.952536	7.255438	506.3838	3.881506e-112	6.179067e-108
ENSG00000196136	-3.225084	6.911908	463.2175	9.587512e-103	1.221008e-98
ENSG00000101347	-3.759902	9.290645	449.9697	7.323427e-100	7.772231e-96
ENSG00000211445	-3.755609	9.102440	433.4656	2.861624e-96	2.603138e-92
ENSG00000162692	3.616656	4.551120	402.0266	1.994189e-89	1.587300e-85
ENSG00000171819	-5.705289	3.474697	389.3431	1.150502e-86	8.140055e-83
ENSG00000152583	-4.364255	5.491013	376.1995	8.363745e-84	5.325782e-80

Differential expression analysis - input

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520	SRR1039521
ENSG00000000003	693	451	887	416	1148	1069	774	581
ENSG00000000005	0	0	0	0	0	0	0	0
ENSG00000000419	466	515	623	364	590	794	419	510
ENSG00000000457	326	274	372	223	356	450	308	297
ENSG00000000460	91	75	61	48	110	95	100	82
ENSG00000000938	0	0	2	0	1	0	0	0

- **Most** RNA-seq methods (e.g., edgeR, DESeq2, voom) need **raw counts** (or equivalent) as input
- **Don't** provide these methods with (e.g.) RPKMs, FPKMs, TPMs, CPMs, log-transformed counts, normalized counts, …
- Read documentation carefully!

Model formulas and design matrices

- Testing is done separately for each gene
- We must tell the packages **which model** to fit (e.g. which predictors to use)
- The design does *not* follow “automatically” from having the sample annotation table - many different designs are often possible
- Model formulas in R:

$$\text{response variable} \sim \text{predictors}$$

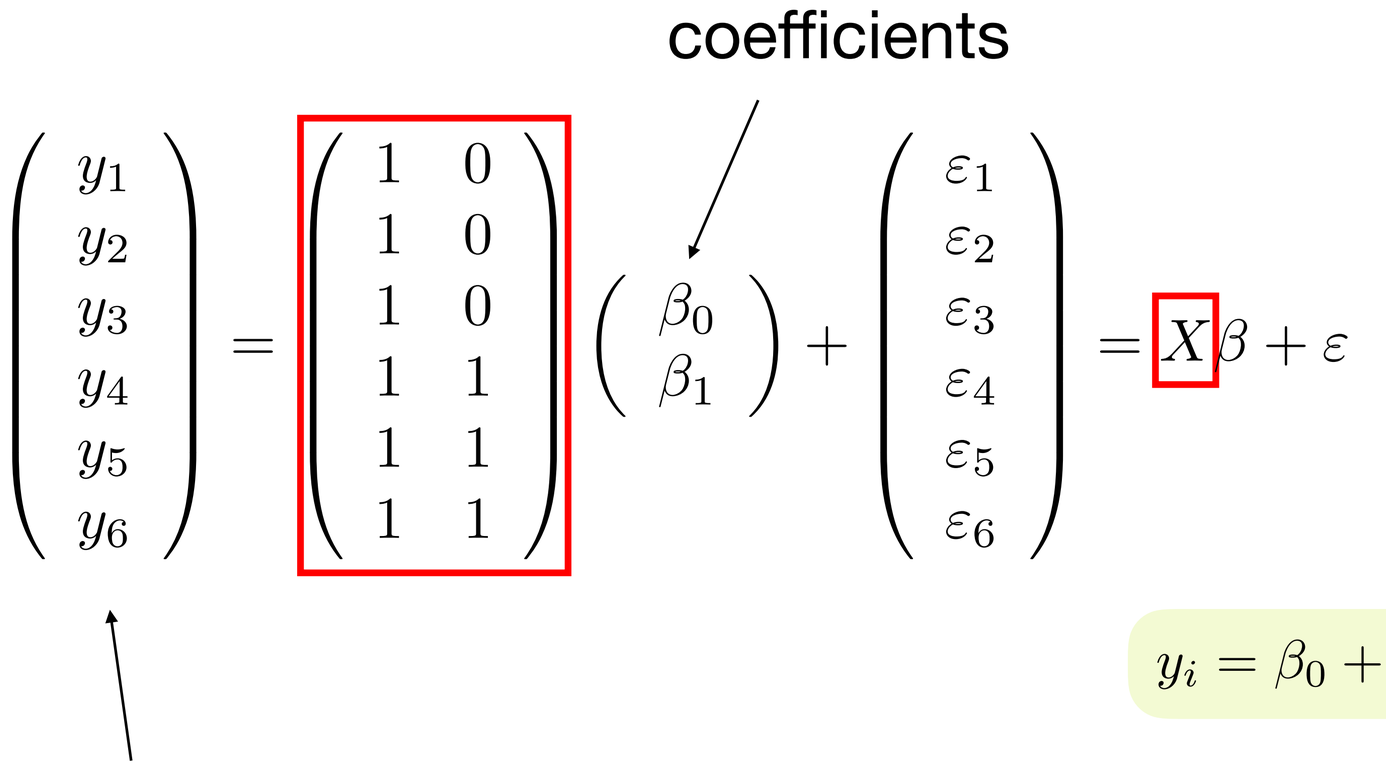
- Fit a separate model for each gene - response variable changes. Specify only predictors

Testing and contrasts

- After fitting the model(s), we must decide *which* coefficient (or combination thereof) we want to apply a hypothesis test for.
- Combinations of coefficients are called *contrasts*.
- Design matrices can often be defined in many equivalent ways - important that the contrast is defined accordingly!

Model formulas and design matrices

- A design matrix contains the values of the predictor variables for each sample



Model formulas and design matrices - example 1
One predictor, two levels (without intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

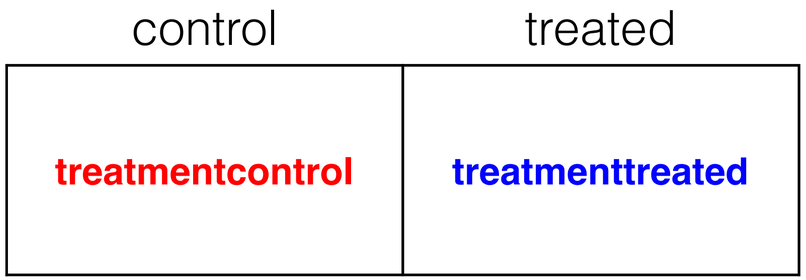
Design matrix:

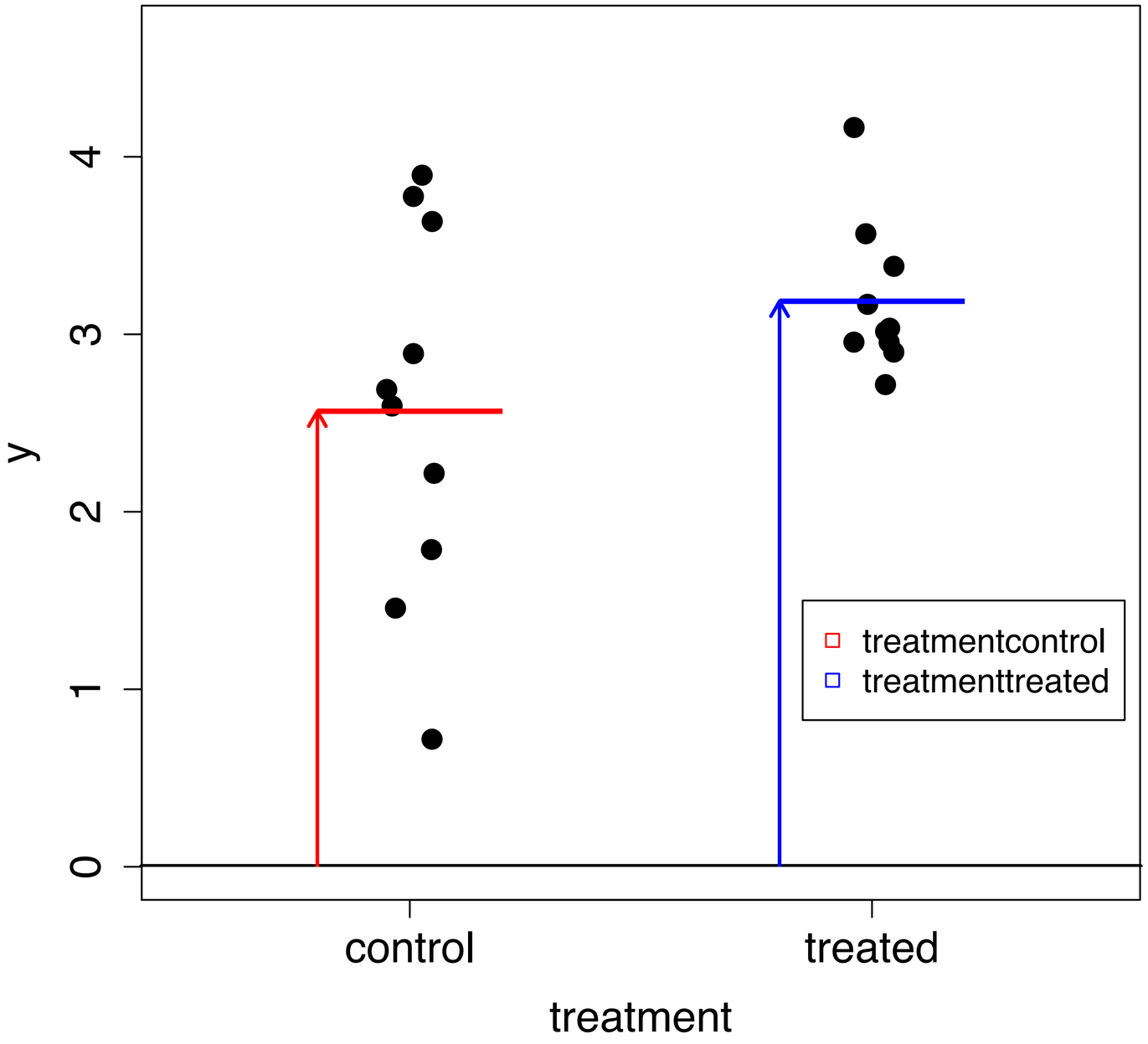
	treatmentcontrol	treatmenttreated
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1

Formula:

$$\sim 0 + \text{treatment}$$

Modeled values:




y
0
1
2
3
4
control
treated
treatment
treatmentcontrol
treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
1 * Intercept + 0 * treatmenttreated	1 * Intercept + 1 * treatmenttreated

Model formulas and design matrices - example 1
One predictor, two levels (with intercept)

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	control
4	s4	treated
5	s5	treated
6	s6	treated

Design matrix:

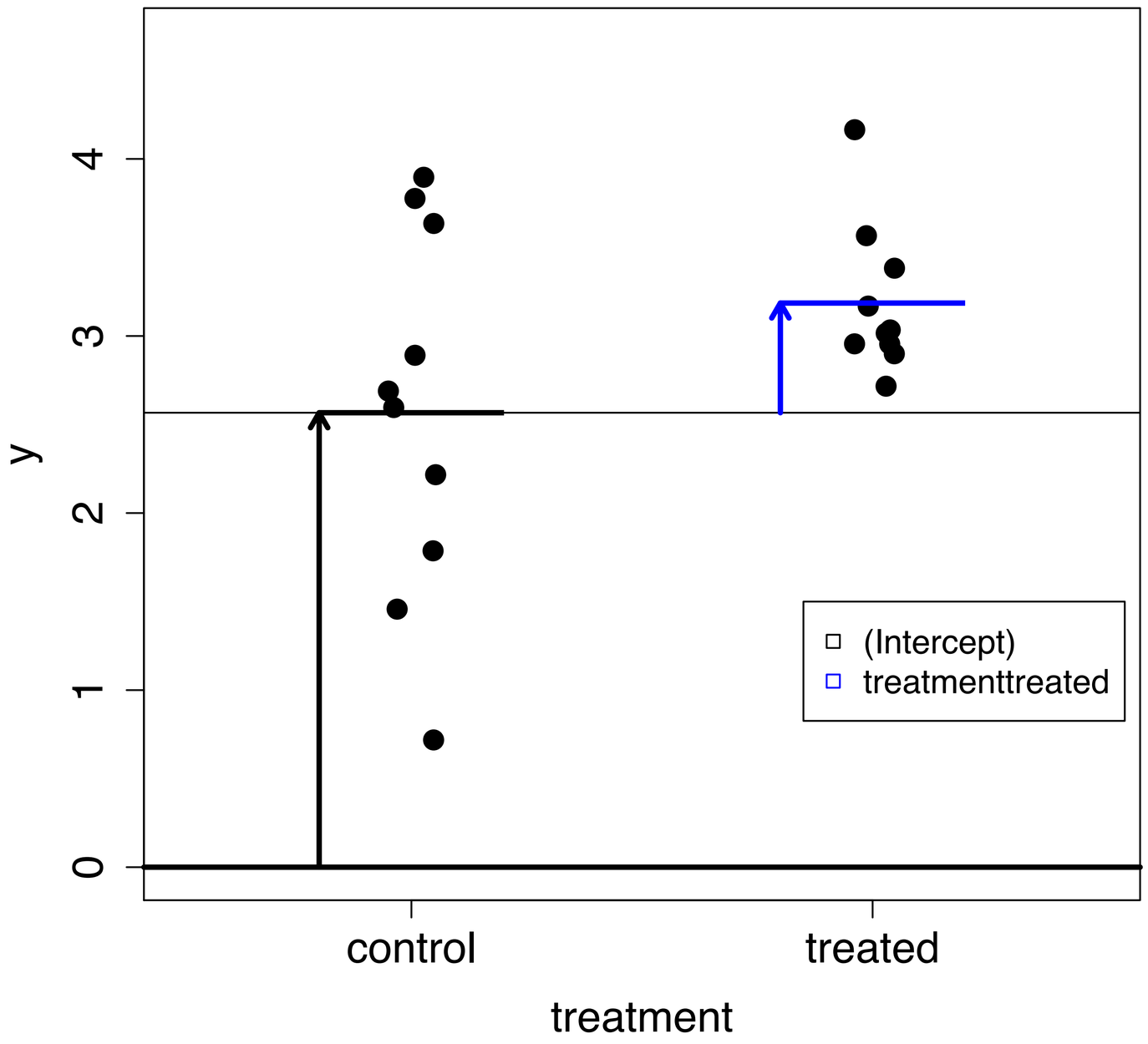
	(Intercept)	treatmenttreated
1	1	0
2	1	0
3	1	0
4	1	1
5	1	1
6	1	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treated
Intercept	**Intercept + treatmenttreated**


y
0
1
2
3
4
control
treated
treatment
(Intercept)
treatmenttreated

Model formulas and design matrices - example 2
One continuous predictor

Sample table:

	sample	age
1	s1	21
2	s2	12
3	s3	64
4	s4	44
5	s5	19
6	s6	26

Design matrix:

	(Intercept)	age
1	1	21
2	1	12
3	1	64
4	1	44
5	1	19
6	1	26

Formula:

$$\sim \text{age}$$

Modeled values:

s1	s2	s3	s4	s5	s6
Intercept + 21 * age	Intercept + 12 * age	Intercept + 64 * age	Intercept + 44 * age	Intercept + 19 * age	Intercept + 26 * age

Model formulas and design matrices - example 3
One predictor, three levels

Sample table:

	sample	treatment
1	s1	control
2	s2	control
3	s3	treatA
4	s4	treatA
5	s5	treatB
6	s6	treatB

Design matrix:

	(Intercept)	treatmenttreatA	treatmenttreatB
1	1	0	0
2	1	0	0
3	1	1	0
4	1	1	0
5	1	0	1
6	1	0	1

Formula:

$$\sim \text{treatment}$$

Modeled values:

control	treatA	treatB
Intercept	**Intercept + treatmenttreatA**	**Intercept + treatmenttreatB**

Model formulas and design matrices - example 4
One predictor, paired data (or two predictors)

Sample table:

	sample	treatment
1	s1	control
2	s1	treated
3	s2	control
4	s2	treated
5	s3	control
6	s3	treated

Design matrix:

	(Intercept)	samples2	samples3	treatmenttreated
1	1	0	0	0
2	1	0	0	1
3	1	1	0	0
4	1	1	0	1
5	1	0	1	0
6	1	0	1	1

Formula:

$$\sim \text{sample} + \text{treatment}$$

Modeled values:

	s1	s2	s3
control	**Intercept**	**Intercept + samples2**	**Intercept + samples3**
treated	**Intercept + treatmenttreated**	**Intercept + samples2 + treatmenttreated**	**Intercept + samples3 + treatmenttreated**

Model formulas and design matrices - example 4
One predictor, paired data (or two predictors)

Sample table:

	genotype	treatment
1	A	control
2	A	control
3	A	treated
4	A	treated
5	B	control
6	B	control
7	B	treated
8	B	treated

Design matrix:

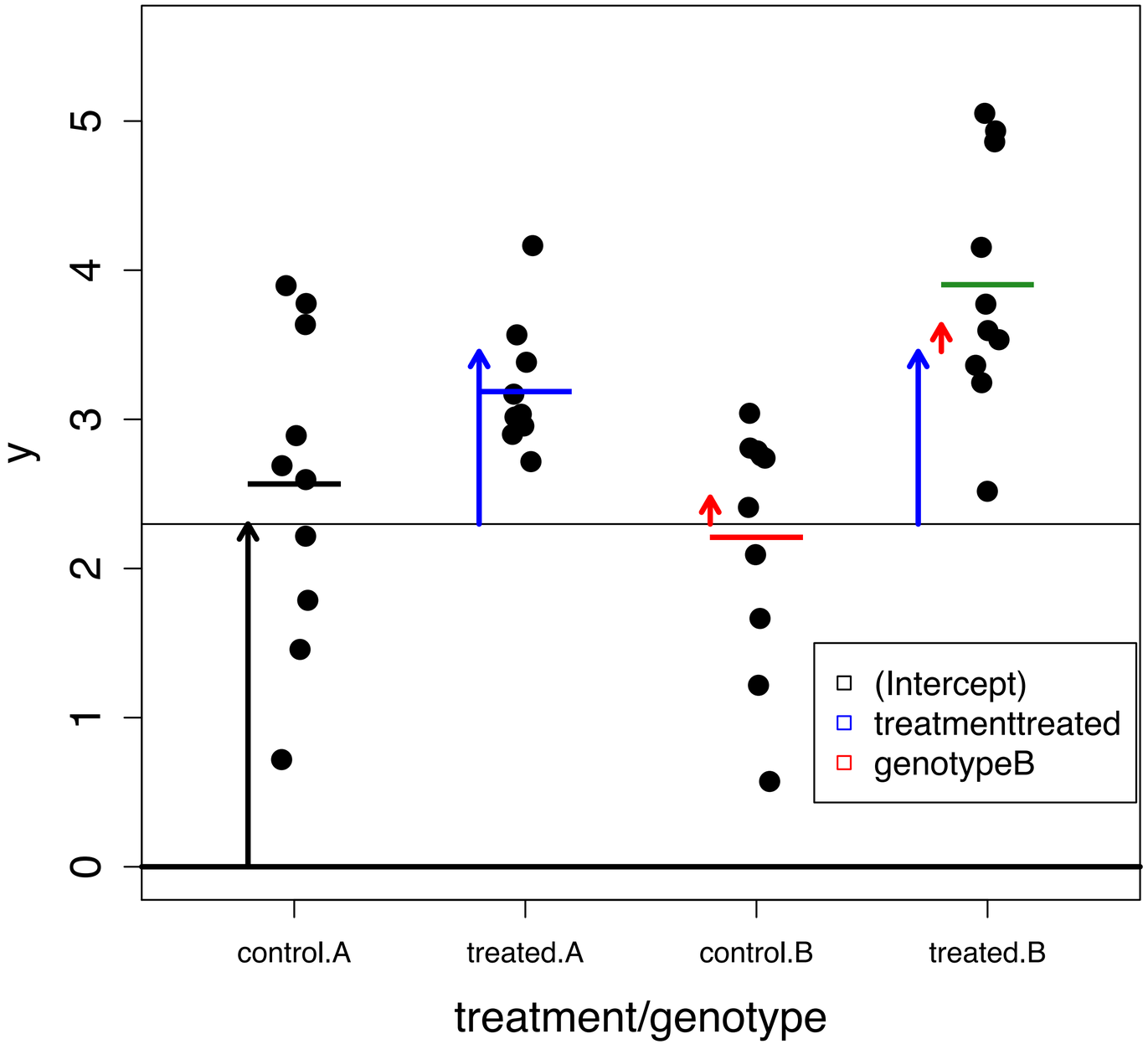
	(Intercept)	genotypeB	treatmenttreated
1	1	0	0
2	1	0	0
3	1	0	1
4	1	0	1
5	1	1	0
6	1	1	0
7	1	1	1
8	1	1	1

Formula:

$$\sim \text{genotype} + \text{treatment}$$

Modeled values:

	genotype A	genotype B
control	**Intercept**	**Intercept + genotypeB**
treated	**Intercept + treatmenttreated**	**Intercept + genotypeB + treatmenttreated**


y
0
1
2
3
4
5
control.A
treated.A
control.B
treated.B
treatment/genotype
(Intercept)
treatmenttreated
genotypeB

Model formulas and design matrices - example 5
Two predictors, with interaction

Sample table:

	genotype	treatment
1	A	control
2	A	control
3	A	treated
4	A	treated
5	B	control
6	B	control
7	B	treated
8	B	treated

Design matrix:

	(Intercept)	genotypeB	treatmenttreated	genotypeB:treatmenttreated
1	1	0	0	0
2	1	0	0	0
3	1	0	1	0
4	1	0	1	0
5	1	1	0	0
6	1	1	0	0
7	1	1	1	1
8	1	1	1	1

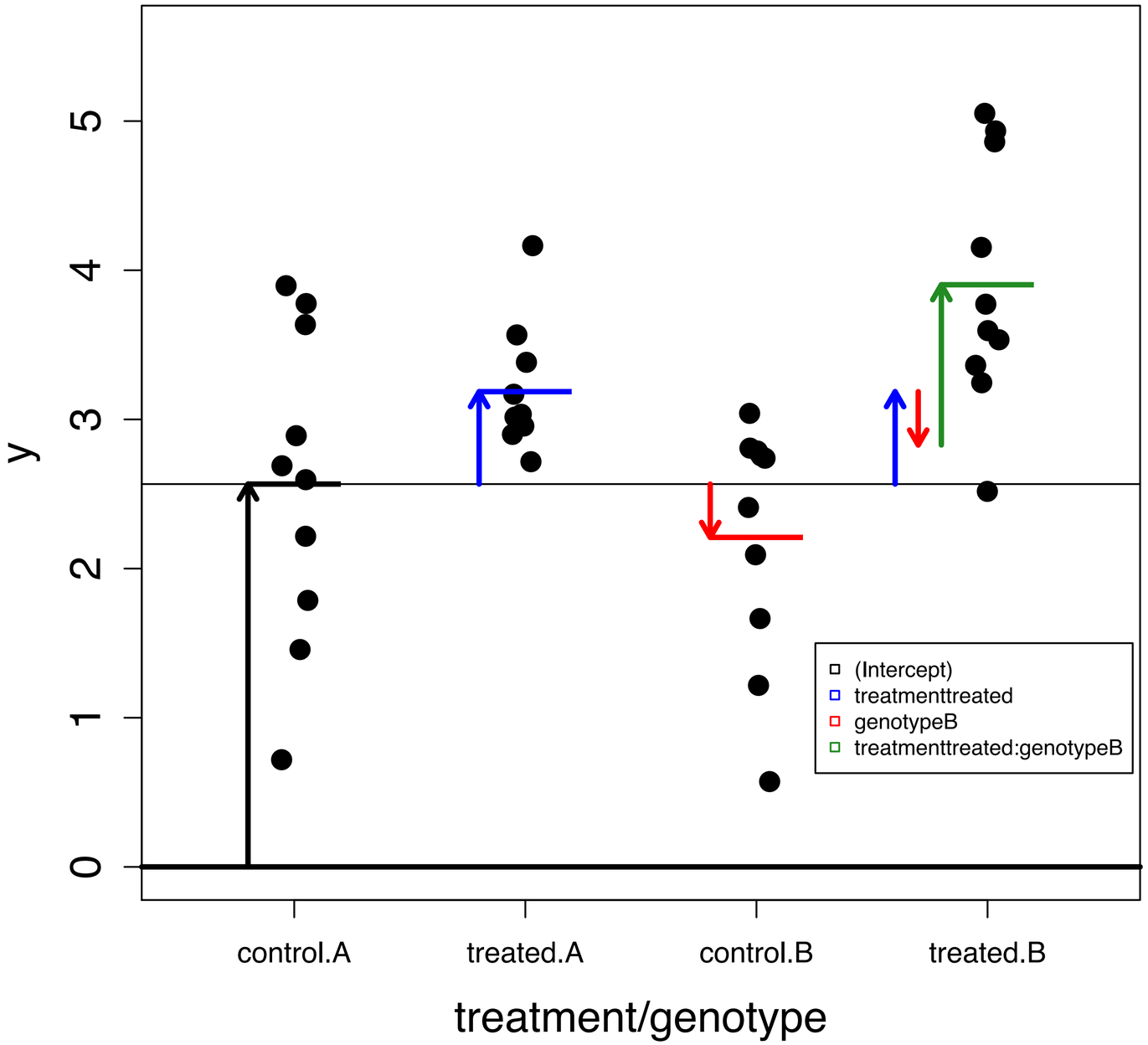
Formula:

$\sim$ genotype * treatment

$\sim$ genotype + treatment + genotype:treatment

Modeled values:

	genotype A	genotype B
control	**Intercept**	**Intercept + genotypeB**
treated	**Intercept + treatmenttreated**	**Intercept + genotypeB + treatmenttreated + genotypeB:treatmenttreated**


y
0
1
2
3
4
5
control.A
treated.A
control.B
treated.B
treatment/genotype
(Intercept)
treatmenttreated
genotypeB
treatmenttreated:genotypeB

Model formulas and design matrices - example 6
Two predictors, with interaction

Sample table:

	treat.gt
1	control.A
2	control.A
3	treated.A
4	treated.A
5	control.B
6	control.B
7	treated.B
8	treated.B

Design matrix:

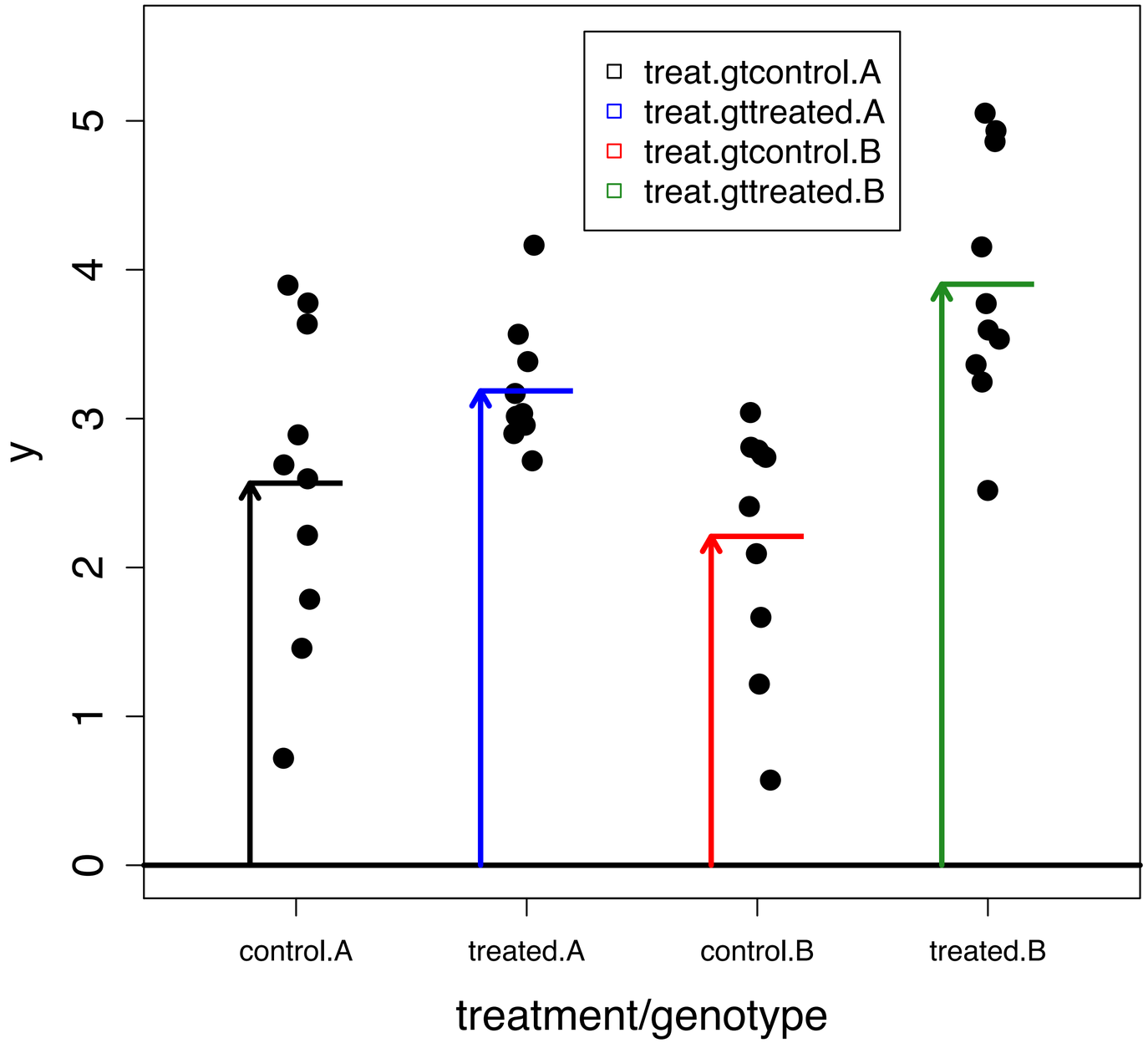
	treat.gtcontrol.A	treat.gttreated.A	treat.gtcontrol.B	treat.gttreated.B
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	1	0	0
5	0	0	1	0
6	0	0	1	0
7	0	0	0	1
8	0	0	0	1

Formula:

$$\sim 0 + \text{treat.gt}$$

Modeled values:

	genotype A	genotype B
control	**treat.gtcontrol.A**	**treat.gtcontrol.B**
treated	**treat.gttreated.A**	**treat.gttreated.B**


treat.gtcontrol.A
treat.gttreated.A
treat.gtcontrol.B
treat.gttreated.B
y
0
1
2
3
4
5
control.A
treated.A
control.B
treated.B
treatment/genotype

Using contrasts vs subsetting data set

- Fitting model to full data set and using contrasts gives more samples to estimate parameters (generally recommended)
- Also assumes that dispersion is similar in all groups (estimates one dispersion parameter per gene)
- In some situations, subsetting to only groups of interest is advantageous:

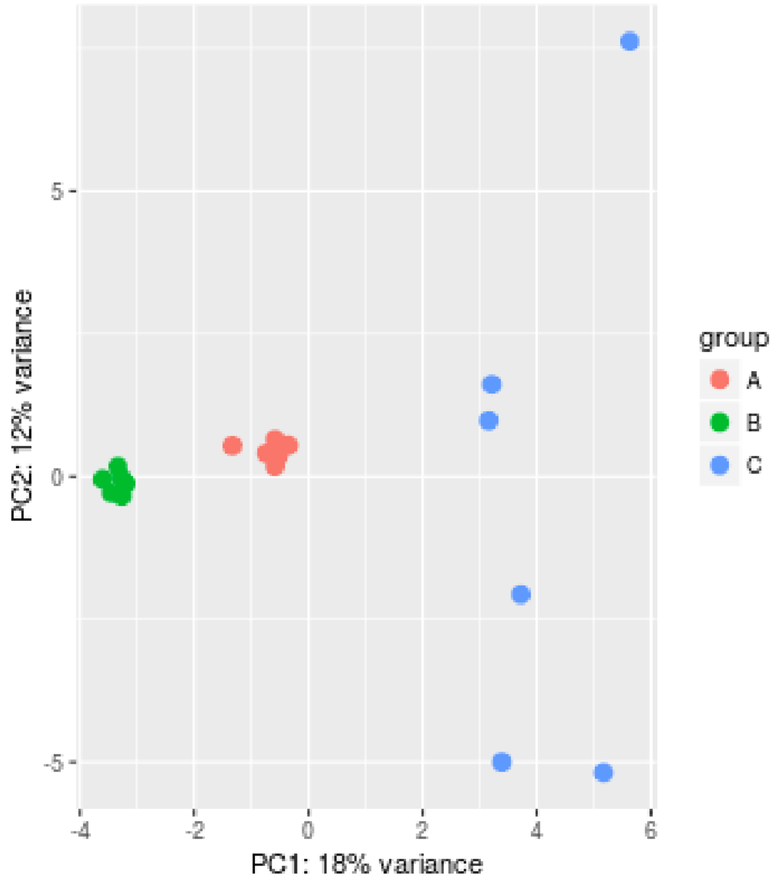


Illustration from DESeq2 vignette