59th Annual Meeting of the European Association for Animal Production

Genome-wide Association Analysis (GWAS) in Livestock

Guilherme J. M. Rosa

University of Wisconsin - Madison

OUTLINE

- Introduction and Examples
- Descriptive Statistics and Data Cleaning
- Genetic Association Analysis
- Statistical Power and Multiple Testing
- Validation and Replication

GENE MAPPING

- ⇨ Linkage Analysis (QTL Analysis)
- ⇨ Fine Mapping Strategies (LDLA approach, Selective Genotyping, etc.)
- ⇨ Association Analysis, Candidate Gene Approach
- ⇨ Genome-wide Association Analysis (GWAS)

HIGH DENSITY SNP PANELS

⇨ Species: cattle, chicken, pigs

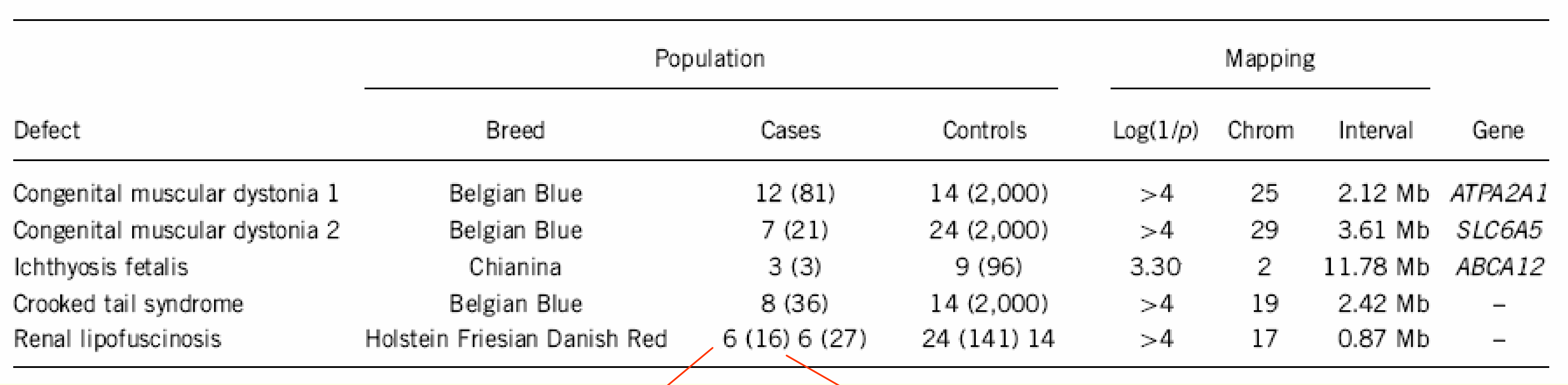
⇨ Technology (Affymetrix, Illumina, etc.)

⇨ Genome-wide Association Analysis (GWAS), Genome-wide Marker Assisted Selection (GWMAS), Population Structure, Selection Signature, etc.

EXAMPLE 1

(Charlier et al., 2008)

- Fine-scale mapping of recessive disorders in cattle
- Custom-made 60K iSelect panel and 25K Affymetrix array
- Case-control study
- Statistical analysis: detection of overlapping, unusually long, homozygous chromosome segments among affected animals



	Population			Mapping			
Defect	Breed	Cases	Controls	Log(1/*p*)	Chrom	Interval	Gene
Congenital muscular dystonia 1	Belgian Blue	12 (81)	14 (2,000)	>4	25	2.12 Mb	*ATPA2A1*
Congenital muscular dystonia 2	Belgian Blue	7 (21)	24 (2,000)	>4	29	3.61 Mb	*SLC6A5*
Ichthyosis fetalis	Chianina	3 (3)	9 (96)	3.30	2	11.78 Mb	*ABCA12*
Crooked tail syndrome	Belgian Blue	8 (36)	14 (2,000)	>4	19	2.42 Mb	–
Renal lipofuscinosis	Holstein Friesian Danish Red	6 (16) 6 (27)	24 (141) 14	>4	17	0.87 Mb	–

Number of animals genotyped and total available

EXAMPLE 2

(Kolbehdari et al., 2008)

- ⇨ 462 Canadian Holstein bulls
- ⇨ 1,536 SNPs
- ⇨ 17 conformation and functional traits
- ⇨ Trait-specific single locus LD regression model

$$\mathrm{EBV}_i = \mu + g_i\alpha + u_i + \varepsilon_i$$

$$\begin{cases} \mathbf{u} = [u_1, u_2, \ldots, u_q]' \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2) \\ \boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_q]' \sim N(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2) \end{cases} \qquad g_i = \begin{cases} 0 & \text{for } 1-1 \\ 1 & \text{for } 1-2 \\ 2 & \text{for } 2-2 \end{cases}$$

- ⇨ Genome- and chromosome-wise significance level
- ⇨ 45 and 151 SNPs found associated with at least 1 trait

EXAMPLE 3

(Daetwyler et al., 2008)

- 484 Holstein sires; 9,919 SNPs; 7 traits
- Selective genotyping within a granddaughter design
- HW, Heterozygosity (H), and PIC
- Variance component linkage analysis (VCLA)

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{v} + \mathbf{e} \quad \begin{cases} \mathbf{v} \sim \mathrm{N}(\mathbf{0}, \mathbf{G}\sigma^2_{\mathrm{QTL}}) \;\rightarrow\; \mathbf{G}\text{: IBD prob. matrix} \\ \mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_{\mathrm{e}}) \end{cases}$$

- Single locus LD regression model (LDRM)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} + \mathbf{e} \quad \rightarrow \quad \mathrm{y_i} = \mathrm{EBV_i}; \quad \mathrm{e_i} \overset{\mathrm{iid}}{\sim} \mathrm{N}(0, \sigma^2_{\mathrm{e}})$$

- 5% chromosome-wise FDR: 102 ‘potential’ (VCLA) and 144 significant (LDRM) QTL

EXAMPLE 4

(Barendse et al., 2007)

- Feed intake (RFI) in cattle
- Total of 1,472 animals from 7 breeds (Taurine and Zebu)
- Selective genotyping: 189 extreme animals within CG (sex, feed group, herd, and market destination)
- MegAllele Genotyping Bovine 10K SNP Panel on Affymetrix GeneChip
- Tests for genotypic frequency homogeneity across breeds, and HW (within?) breeds
- Single marker analysis using permutation test
- 161 SNPs with $P < 0.01$ (FDR 17.4%)
- Validation performed on 44 selected SNPs

ASSOCIATION ANALYSIS

- Data Cleaning: Data preprocessing
- Data Imputation: Missing genotypes (information from allelic frequencies, LD, recombination rates, phenotype, etc.)
- Statistical Analysis:
  - Significance analysis
  - ‘Large p, small n’ paradigm
  - Multiple testing

DESCRIPTIVE STATISTICS & DATA CLEANING

⇨ Measurement/recording error

⇨ Genotyping error; Mendelian inconsistencies

⇨ Redundancies

⇨ Heterozygosity (H)
Polymorphism Information Content (PIC)

⇨ Minor Allele Frequency (MAF)

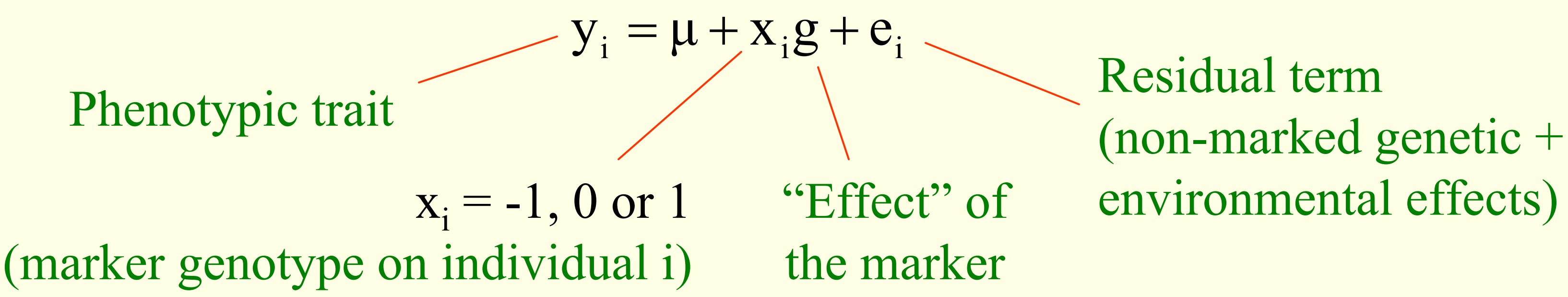
⇨ Hardy-Weinberg equilibrium

TYPOLOGY OF GENETIC ASSOCIATION TESTS

	Association	Association in the Presence of Linkage	
		Test Conditioned on Parental Genotypes (Directly or Indirectly)	Tests Based on Controlling for Background NLD
Residuals Unrelated	Ordinary Association Test	TDT’s	Structured Association Testing Genomic Control
Related Residuals	Ordinary Association Tests with Related Individuals	TDT’s with Multiple Offspring or Pedigrees	

SINGLE MARKER REGRESSION

- Diallelic marker (additive genetic effect only):



$$y_i = \mu + x_i g + e_i$$

Phenotypic trait

x_i = -1, 0 or 1
(marker genotype on individual i)

“Effect” of
the marker

Residual term
(non-marked genetic +
environmental effects)

- IBD and combined LD-LA approaches (Zhao et al. 2007)

- Dominance effect: $y_i = \mu + x_i\alpha + (1 - |x_i|)\delta + e_i$

- Multi-allelic marker (haplotype): $y_i = \mu + \sum_{k=1}^{m-1} x_{ik} g_k + e_i$ Calus et al. 2007
Hayes et al. 2007

- Population structure: $\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{g} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$

MULTIPLE MARKER REGRESSION

⇨ Diallelic markers (additive genetic effects only):

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j g_j + \mathbf{e}$$

- If the number of markers (p) is large, fitting such a model using standard regression approaches is not trivial.
- Various strategies have been proposed to overcome this difficulty, such as:

  - Stepwise selection methodology
  - Dimension reduction techniques, such as singular vale decomposition and partial least squares (Hastie et al. 2001)
  - Ridge regression (Whittaker et al. 2000, Muir 2007)
  - Shrinkage estimation (Meuwissen et al. 2001, Gianola et al. 2003, Xu 2003)

SHRINKAGE APPROACHES

⇨ Model: $\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j g_j + \mathbf{e}$

- Marker effects assumed normally distributed with a common variance, i.e.: $g_j \sim N(0, \sigma_0^2)$

- Estimates:

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X} \\ \mathbf{X}'\mathbf{1} & \mathbf{X}'\mathbf{X} + \mathbf{I}\gamma \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix}$$

where $\gamma = \sigma_e^2 / \sigma_0^2$

SHRINKAGE APPROACHES

(Meuwissen et al. 2001, Xu 2003)

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad \longrightarrow \quad \mathbf{y} \mid \mu, \mathbf{g}_j, \sigma_e^2 \sim N(\mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j, \mathbf{I}\sigma_e^2)$$

⇨ Prior distributions:

$$\begin{cases} g_j \mid \sigma_j^2 \sim N(0, \sigma_j^2) \\ \sigma_j^2 \sim \chi^{-2}(\nu, S) \\ \text{(scaled inverted chi-square distribution with scale parameter S and } \nu \text{ degrees of freedom)} \\ \sigma_e^2 \sim \chi^{-2}(-2,0) \end{cases}$$

SHRINKAGE APPROACHES

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \mathbf{e} \quad \longrightarrow \quad \mathbf{y} \mid \mu, \mathbf{g}_j, \sigma_e^2 \sim N(\mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j, \mathbf{I}\sigma_e^2)$$

⇨ Prior distributions:

$$\begin{cases} \begin{cases} g_j = 0 & \text{with probability } \pi \\ g_j \mid \sigma_j^2 \sim N(0, \sigma_j^2) & \text{with probability } (1 \text{ - } \pi) \end{cases} \\ \pi \sim \text{Beta}(\alpha, \beta) \\ \sigma_j^2 \sim \chi^{-2}(\nu, S) \\ \sigma_e^2 \sim \chi^{-2}(-2, 0) \end{cases}$$

⇨ Alternative distributions for $\mathbf{g}_j$: if instead of a Gaussian process, a double exponential distribution is adopted → Bayesian LASSO (Park and Casella 2008)

GWAS Including Non-Additive Genetic Effects

- Many studies that attempt to identify the genetic basis of complex traits ignore the possibility that loci interact, despite its known substantial contribution to genetic variation (Carborg and Haley 2005)

- Extensions of the GWAS model to accommodate dominance and some level of epistasis have been proposed (Yi et al. 2003, Huang et al. 2007, Xu 2007), which can be described as:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{X}_j \mathbf{g}_j + \sum_{j'>j}^{p} \mathbf{X}_{j'j} \mathbf{g}_{j'j} + \mathbf{e}$$

where the $\mathbf{g}_{j'j}$ refer to interaction terms relative to epistatic effects involving loci j and j', and $\mathbf{X}_{j'j}$ represent appropriate design matrices.

GWAS Including Non-Additive Genetic Effects

- ⇨ In the case of diallelic loci, each row of $\mathbf{X}_j\mathbf{g}_j$ can be factorize into additive and dominance effects as $\mathbf{x}_{ij}^{'}\mathbf{g}_j = x_{ij}\alpha_j + (1-|x_{ij}|)\delta_j$, where x_{ij} = -1, 0 or 1 for the three possible genotypes aa, Aa and AA, respectively, and α_j and δ_j represent the additive and dominance effects relative to loci j.

- ⇨ Similarly, the four degrees of freedom relative to each pairwise interaction between biallelic loci can be described as:

$$\mathbf{x}_{ij'j}^{'}\mathbf{g}_{j'j} = x_{ij'}x_{ij}\alpha\alpha_{j'j} + x_{ij'}(1-|x_{ij}|)\alpha\delta_{j'j}$$
$$+ x_{ij}(1-|x_{ij'}|)\delta\alpha_{jj'} + (1-|x_{ij'}|)(1-|x_{ij}|)\delta\delta_{j'j}$$

where $\alpha\alpha_{j'j}$, $\alpha\delta_{j'j}$, $\delta\alpha_{jj'}$, and $\delta\delta_{j'j}$ represent additive × additive, additive × dominance, dominance × additive, and dominance × dominance epistasis between loci j’ and j.

GWAS Including Non-Additive Genetic Effects

- ⇨ Similar statistical and computational strategies discussed previously can be used also for fitting the non-additive GWAS model, such as dimension reduction techniques and hierarchical modeling approaches.
- ⇨ The non-additive GWAS model presented, however, relies on strong assumptions, such as linearity, multivariate normality, and proportion of segregating loci (Gianola et al. 2006).
- ⇨ In addition, the genome seems to be much more highly interactive than what standard quantitative genetic models can accommodate. For example, the number of higher-order interactions (i.e., multi-loci epistatic effects) grows extremely quickly with the increase on the number of markers; moreover, the partition of genetic variance into orthogonal additive, dominance, additive x additive, additive x dominance, etc. components is possible only under highly idealized, unrealistic conditions (Cockerham 1954, Kempthorne 1954).

FEATURE SELECTION

⇨ Two-step approaches (e.g., Hoh et al. 2000): selection of a small number of influential markers (features), which are then used for more elaborate modeling of the relationship between markers and the target trait.

⇨ Two-step procedures require an efficient method for optimal selection of influential features. Long et al. (2007) developed a machine learning selection methodology for binary traits, which consisted of filtering (using information gain), and wrapping (using naïve Bayesian classification).

⇨ The **filter** is a preprocessing method, which reduces the large number of SNPs to a much smaller size, to facilitate the wrapper step.

⇨ The **wrapper** step then optimizes the performance of the top scoring SNPs selected by the filter. It consists of an iterative search-evaluate-search algorithm, using cross-validation accuracy to evaluate the selected feature subset's usefulness.

⇨ Long et al. (2007) found that the two-step method improved naïve Bayesian classification accuracy over the case without feature selection, from around 50 to above 90% without and with feature selection.

TESTING HYPOTHESES

HYPOTHESIS TESTING

	H_0 is not rejected	H_0 is rejected
H_0 is true	No error (1-α)	Type I error (α)
H_0 is false	Type II error (β)	No error (1-β)

Significance level

Power

- Standard approach:

① Specify an acceptable type I error rate (α)

② Seek tests that minimize the type II error rate (β), i.e., maximize power (1 - β)

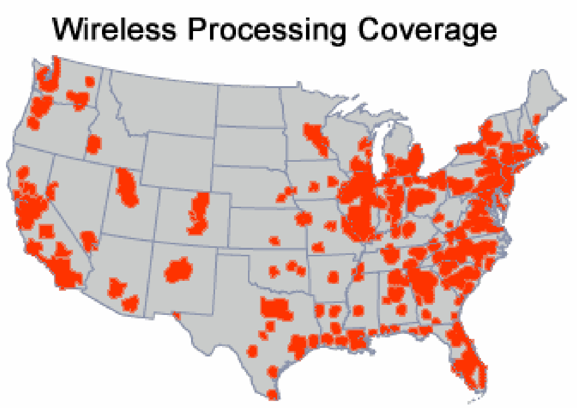
STATISTICAL POWER

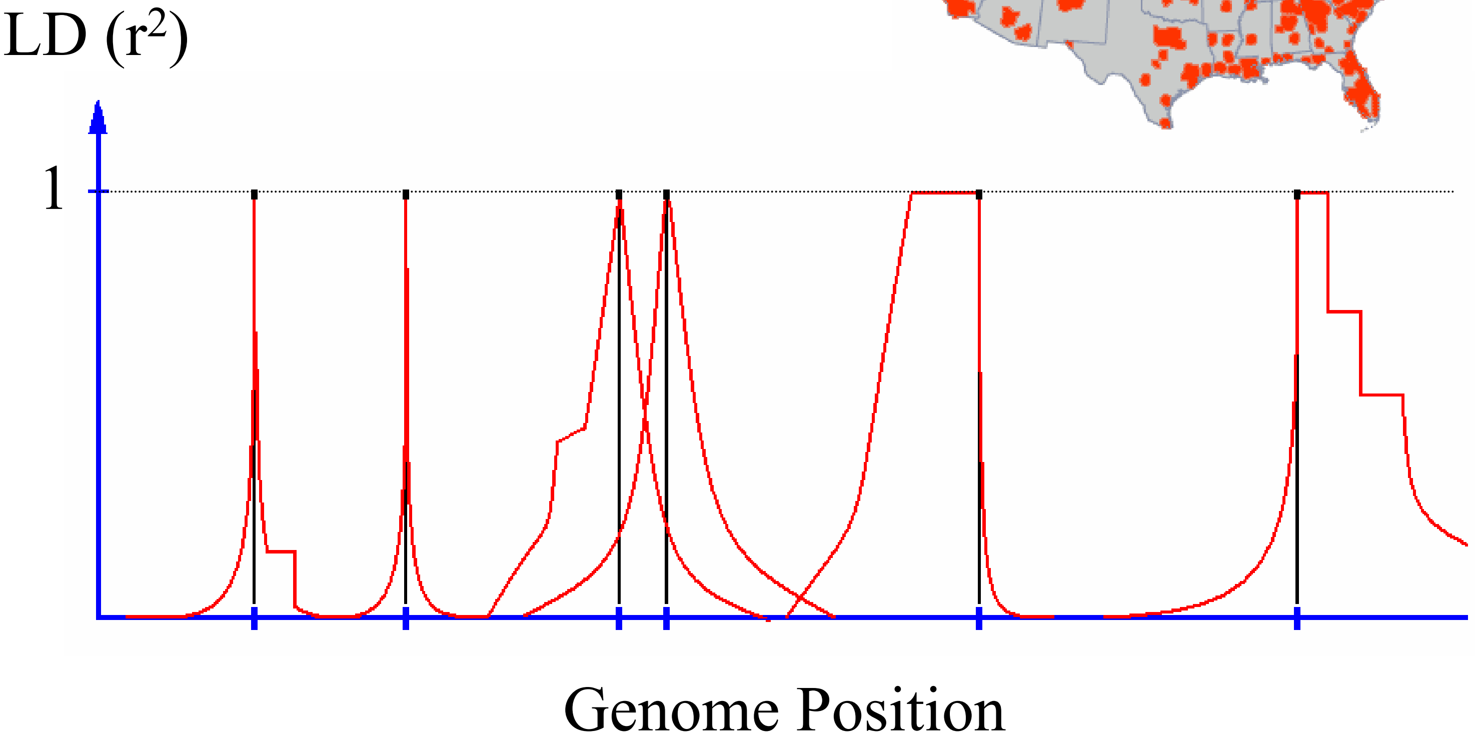
⇨ Power is a function of:

- Significance level (α)
- Sample size (n)
- Effect size (δ), expressed as a proportion of variance in measured phenotype, subsumes allele frequency, mode of inheritance, measurement reliability, degree of LD, and all other aspects of genetic model
- Test statistic (T)

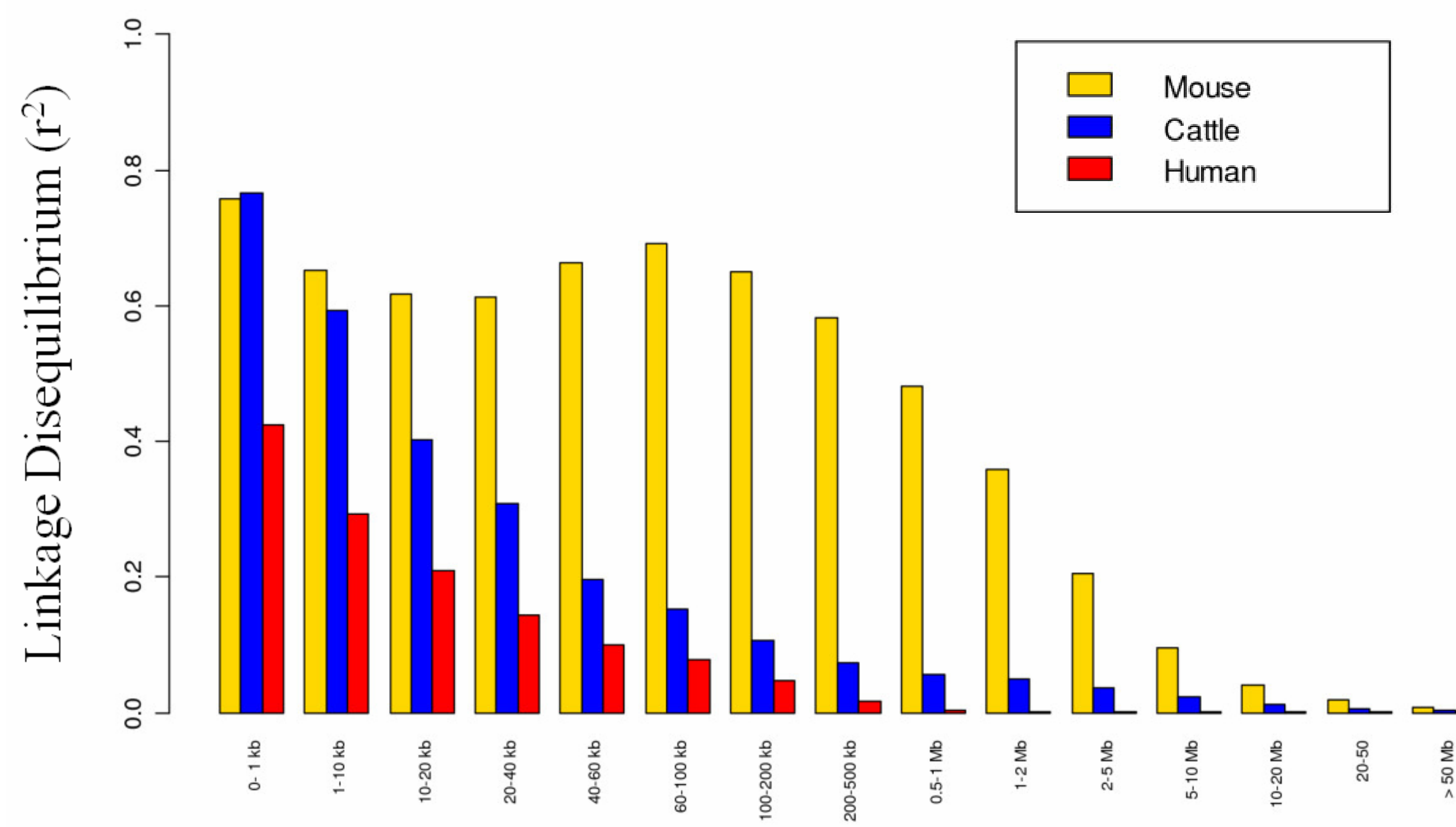
GENOME COVERAGE

- Lightning rod,
  or cellular coverage…





GENOME COVERAGE

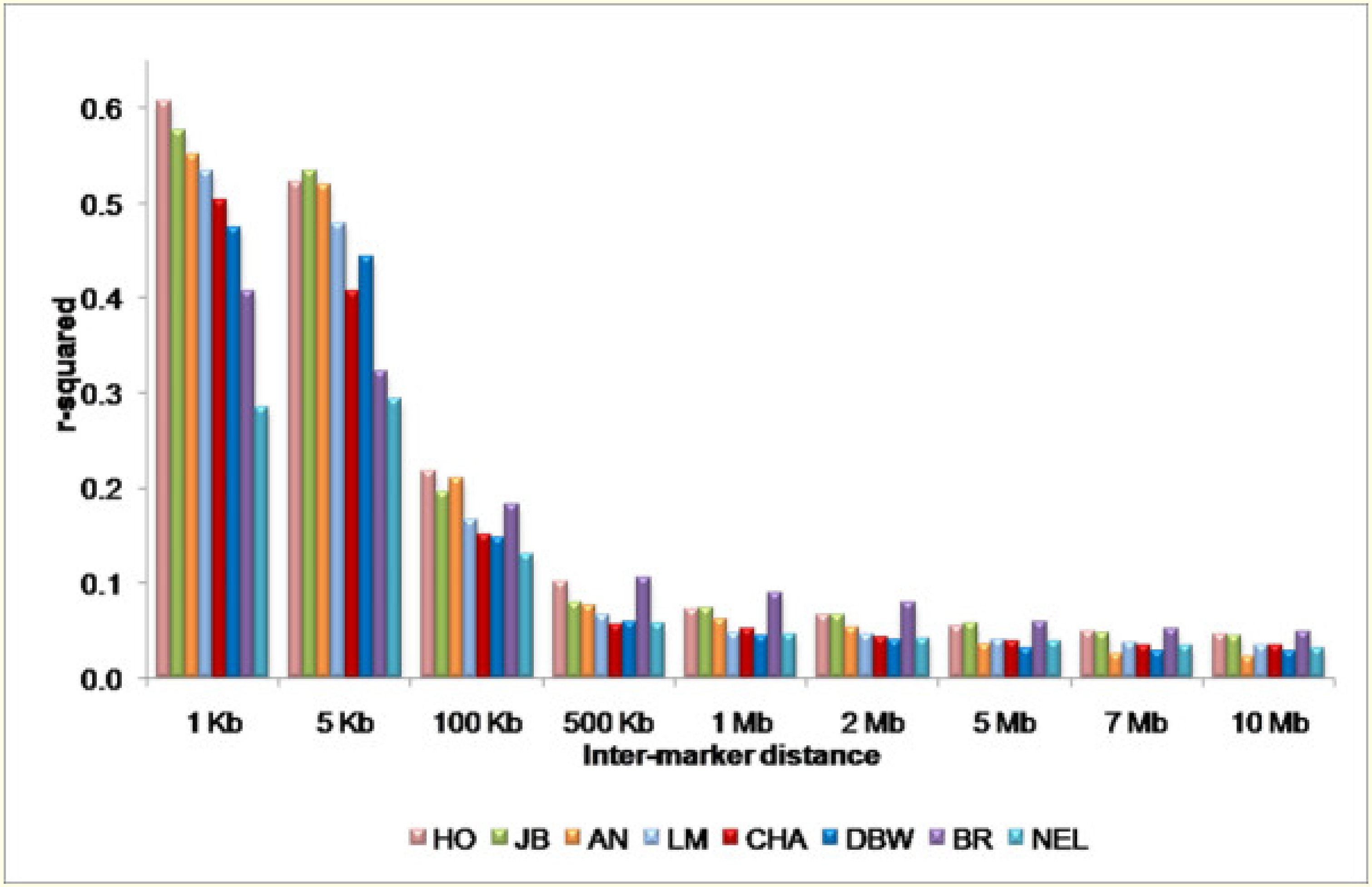


(Khatkar et al., 2008)

Pigs: $r^2 \approx 0.2$ at 1,000 kb (Du et al. 2007)

Chickens: $\chi^{2'} \geq 0.2$ 28-57% of marker pairs 5-10 cM apart (Heifetz et al. 2005)

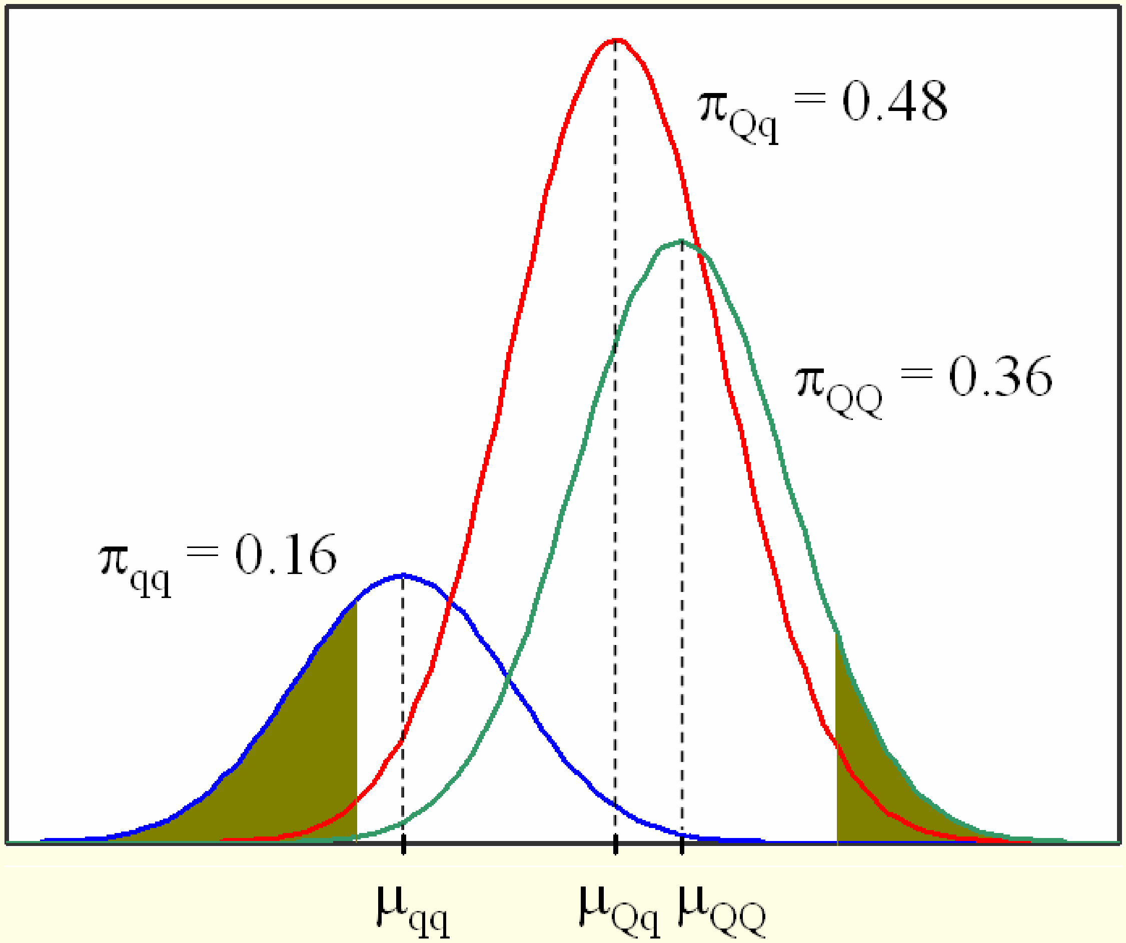
GENOME COVERAGE



(McKay et al. 2007)

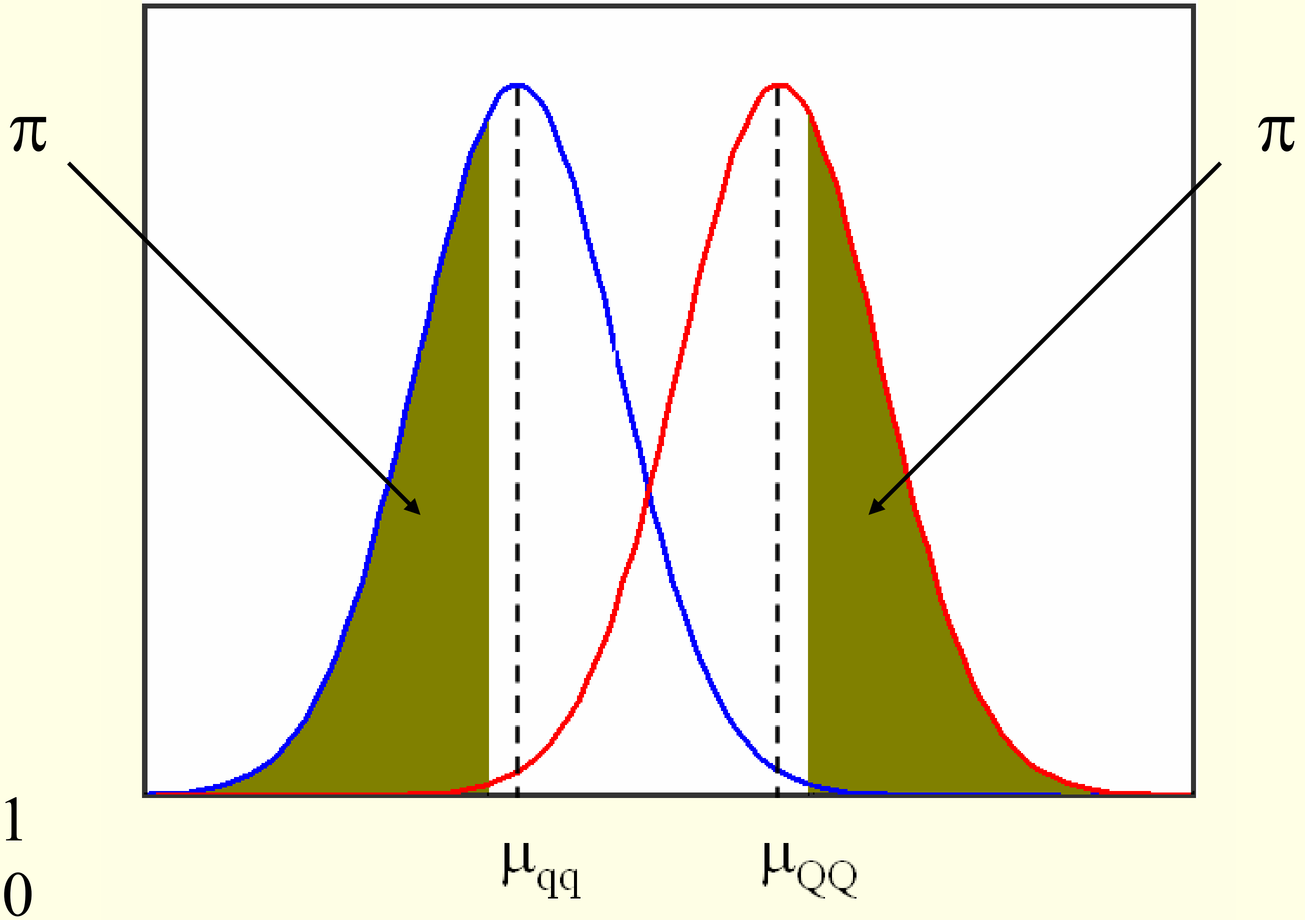
HO: Holstein, JB: Japonese Black, AN: Angus, LM: Limousin, CHA: Charolais, DBW: Dutch Black & White Dairy, BR: Brahman, NEL: Nelore

SELECTIVE GENOTYPING



$\alpha = 1.3$
$\delta = 0.6$
$\sigma^2 = 1.0$
$f(Q) = 0.6$
$f(q) = 0.4$

SIMULATION STUDY



$\Delta = 2$

$\sigma^2 = 81$

$n = 500$

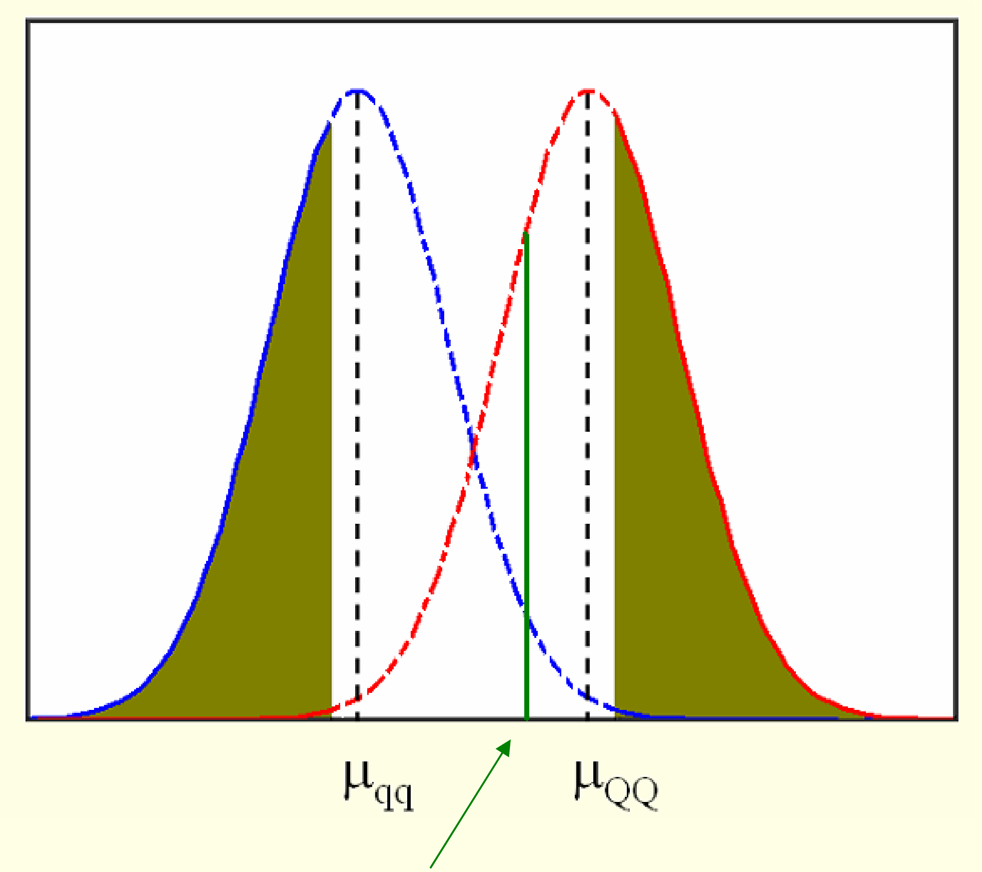
$\pi = 0.5,\ 0.25$ and 0.1

COMPARING GENOTYPIC FREQUENCIES

	Genotype		
Phenotype	A	B	Total
Low	LA	LB	L
High	HA	HB	H
Total	A	B	N

$$X^2 = \frac{N \times (LA \times HB - HA \times LB)^2}{A \times B \times L \times H} \sim \chi^2_{1df}$$

COMPARING MEANS WITH A MIXTURE MODEL



Genotype?

- EM algorithm and LRT

Phenotype	Genotype
y_1	A
y_2	A
y_3	?
y_4	B
y_5	B

RESULTS

$\pi = .10$

Statistic	Test	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Type I Error	χ^2	.014	.062	.086
	LRT	.042	.116	.186
Power	χ^2	.256	.536	.596
	LRT	.442	.678	.774

$\pi = .25$

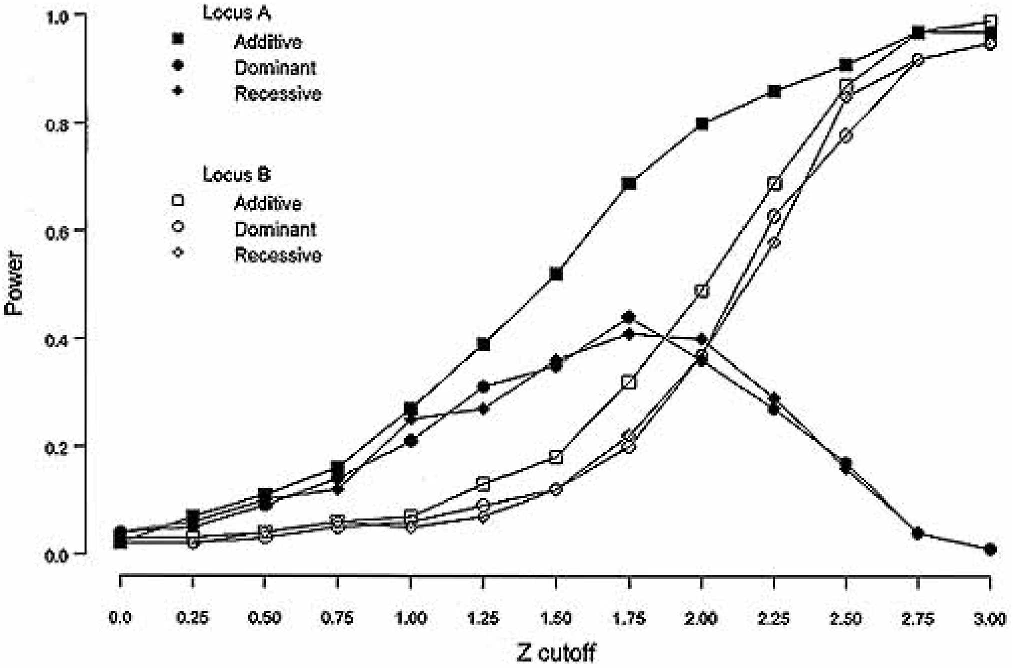
Statistic	Test	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Type I Error	χ^2	.008	.040	.072
	LRT	.010	.050	.094
Power	χ^2	.354	.644	.736
	LRT	.470	.718	.792

$\pi = .50$

Statistic	Test	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Type I Error	χ^2	.008	.046	.088
	LRT	.016	.042	.098
Power	χ^2	.254	.542	.642
	LRT	.464	.696	.796

SELECTIVE GENOTYPING

Locus	Parameter	Mode of inheritance		
		additive	dominant	recessive
A	μ_{AA} (σ_{wAA})	0.064 (1)	0.032 (1)	0.032 (1)
	μ_{Aa} (σ_{wAa})	0.564 (1)	1.032 (1)	0.032 (1)
	μ_{aa} (σ_{waa})	1.064 (1)	1.032 (1)	1.032 (1)
	P_a	0.500	0.077	0.385
B	μ_{BB} (σ_{wBB})	0.500 (1)	0.148 (1)	0.148 (1)
	μ_{Bb} (σ_{wBb})	2.500 (1)	4.148 (1)	0.148 (1)
	μ_{bb} (σ_{wbb})	4.500 (1)	4.148 (1)	4.148 (1)
	P_b	0.016	0.004	0.089



% in each side of the distribution: 50 40 30 20 10 5 1 .15

(Allison et al., 1998)

THE MULTIPLE TESTING ISSUE

Suppose you carry out 10 hypothesis tests at the 5% level

(assume independent tests)

The probability of declaring a particular test significant under its null hypothesis is 0.05

But the probability of declaring at least 1 of the 10 tests significant is $\boxed{0.401}$

$1 - 0.95^{10}$

If you perform 20 hypothesis tests, this probability increases to 0.642…

➡ Typically thousands of markers tested simultaneously

➡ Example: Suppose trait with $H^2 = 0$ and association analysis considering 100 markers and $\alpha = 5\%$ (for each test)

- Expected $100 \times 0.05 = 5$ false associations…

THE MULTIPLE TESTING ISSUE

	# H_0 not rejected	# H_0 rejected	
# true H_0	A	B	m_0
# false H_0	C	D	m_1
	m – R	R	m

Observable quantity (nº rejected H_0)

known quantity

- Family-wise error rate (FWER): $\mathrm{FWER} = \Pr(B \geq 1) = 1 - \Pr(B = 0)$
- False discovery rate (FDR): $$\mathrm{FDR} = \underbrace{E[B/R \mid R > 0]}_{\text{Positive FDR (pFDR); Storey (2002)}} \Pr(R > 0)$$

MULTIPLE TESTING CONTROL

① Controlling family-wise type I error rates (FWER) (Westfall and Young, 1993)

$$\mathrm{FWER} = \Pr(V \geq 1) = 1 - \Pr(V = 0)$$

$$\mathrm{FWER}_k = \Pr(V > k) = 1 - \Pr(V \leq k)$$ (Chen and Storey, 2006)

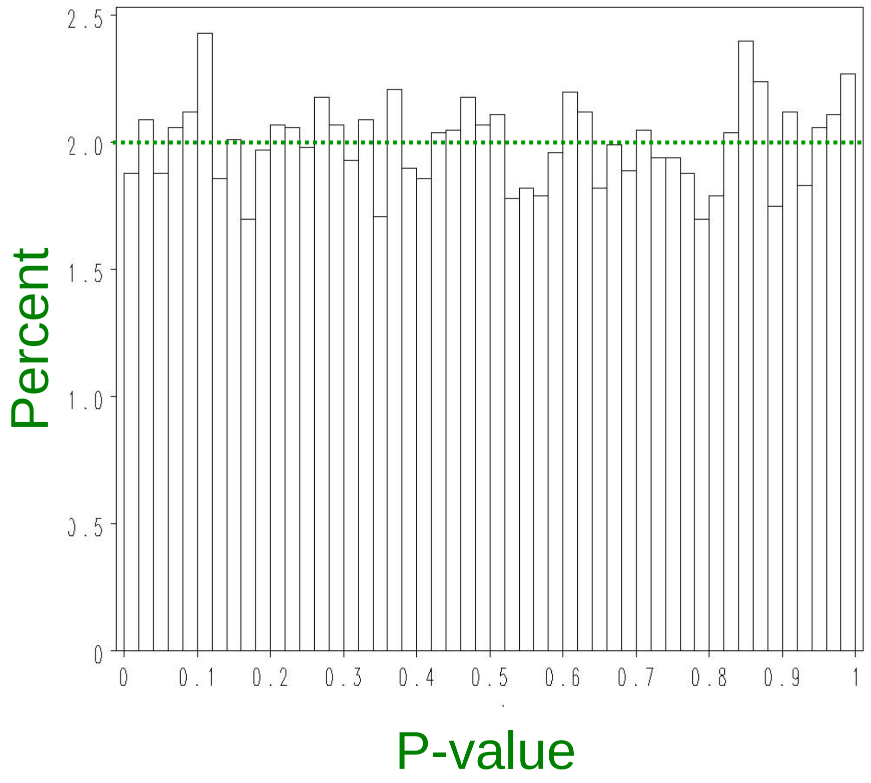
② False discovery rate (FDR) (Benjamini and Hochberg, 1995; Storey et al., 2002)

$$\mathrm{FDR} = \underbrace{E[V / R \mid R > 0]}_{} \Pr(R > 0)$$

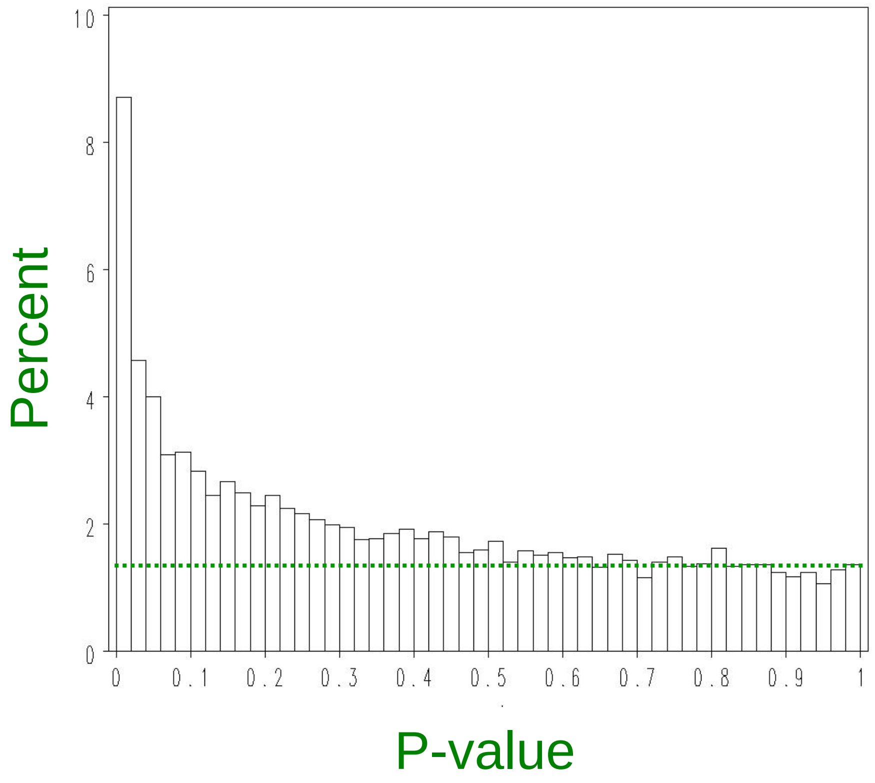
Positive FDR (pFDR); Storey (2002)

DISTRIBUTION OF P-VALUES (Histogram)

Under H_0

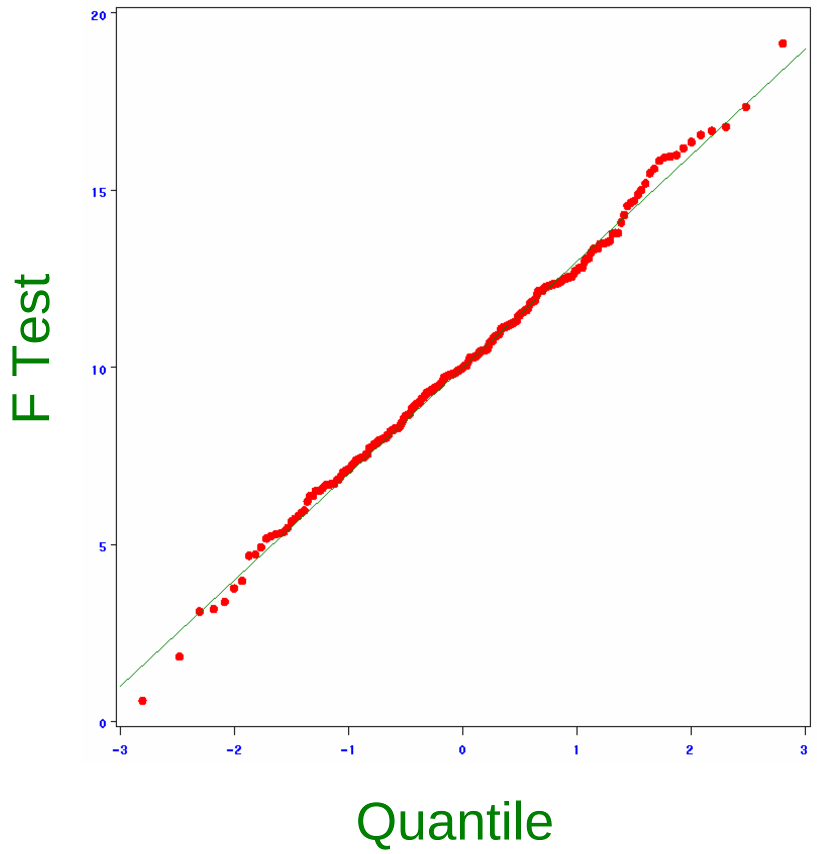


Mixture of H_0 and H_a

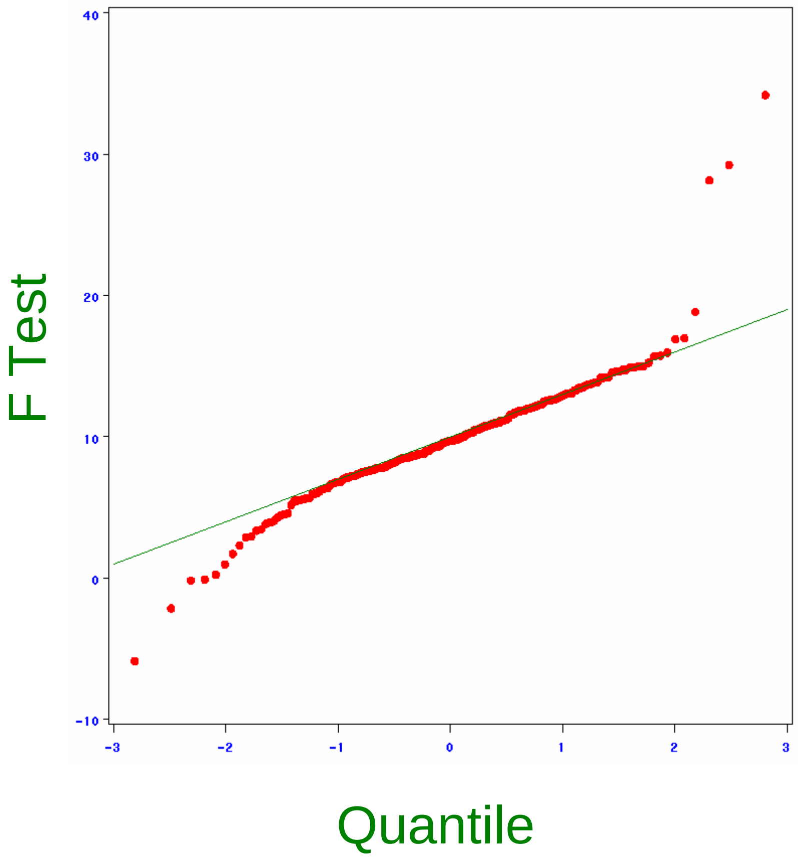


DISTRIBUTION OF P-VALUES (Q-Q Plot)

Under H_0



Mixture of H_0 and H_a



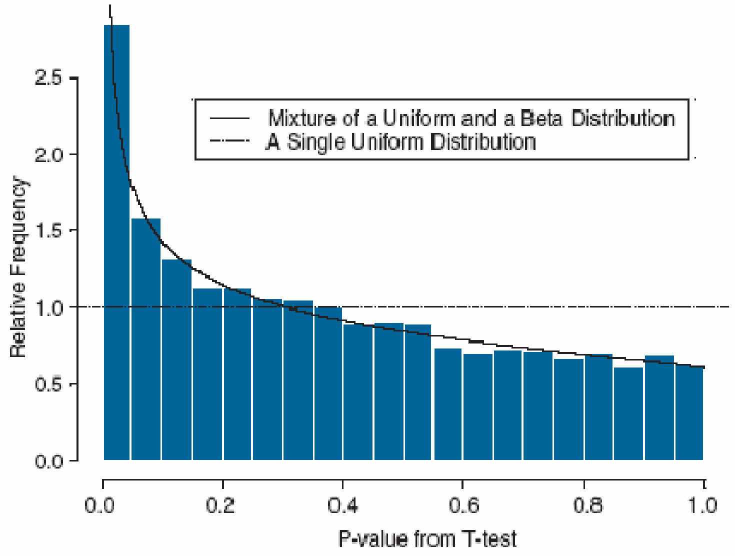
HOW MANY SAMPLES SHOULD I USE?

* In the context of multiple testing:

Gadbury et al. (2004)

$$\mathrm{TP} = \frac{\mathrm{D}}{\mathrm{C}+\mathrm{D}}, \quad \mathrm{TN} = \frac{\mathrm{A}}{\mathrm{A}+\mathrm{B}},$$

$$\mathrm{EDR} = \frac{\mathrm{D}}{\mathrm{B}+\mathrm{D}}$$

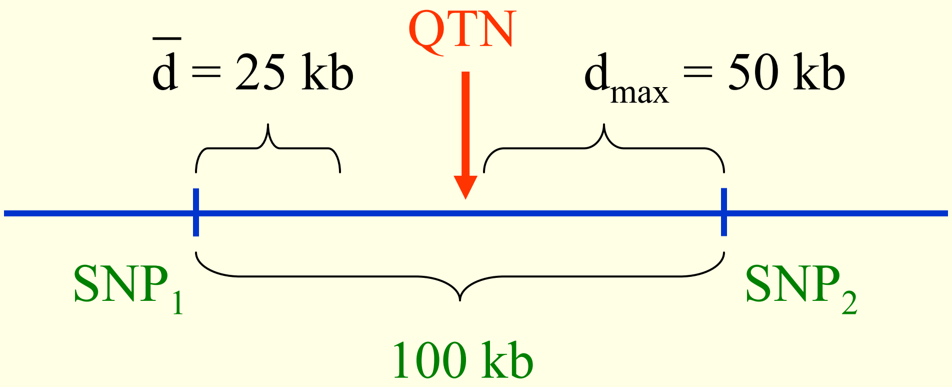


- p-value $\xrightarrow{n}$ t $\xrightarrow{n^*}$ t^* $\rightarrow$ p-value* $\xrightarrow{\tau}$ $\begin{cases} \text{TP} \\ \text{TN} \\ \text{EDR} \end{cases}$

Other methods (FDR-based): Muller et al. (2004), Hu et al. (2005) and Jung (2005)

EXAMPLE

- GWAS in dairy cattle with the 50K SNP bovine chip
- Fertilization and embryo survival rates: $y \sim \mathrm{Bin}(m\ ,\ p)$
- Even if only 40-50% of SNPs are polymorphic and with MAF > 0.10 → about 10 SNPs/cM, i.e. an average spacing of 100 kb between SNPs



$\bar{d} = 25$ kb

QTN

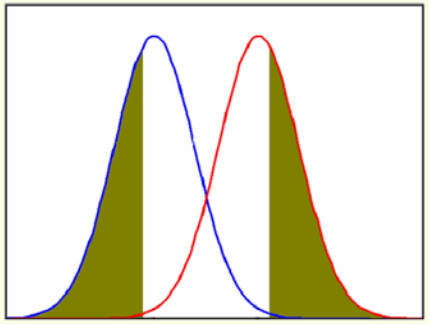
$d_{max} = 50$ kb

SNP_1

SNP_2

100 kb

- Selective genotyping:



$$X^2 = \frac{N\times(LA\times HB - HA\times LB)^2}{A\times B\times L\times H} \sim \chi^2_{1df}$$

$$T = \frac{\bar{y}_1 - \bar{y}_2}{s.e.} \approx t_{\varphi df}$$

EXAMPLE

⇨ $$\left.\begin{cases} \text{Group 1:} \quad x_1, x_2, \ldots, x_{n_x} \sim \text{Bin}(m_i, p_x) \\ \text{Group 2:} \quad y_1, y_2, \ldots, y_{n_y} \sim \text{Bin}(m_i, p_y) \end{cases}\right\} \; H_0 : p_x = p_y$$

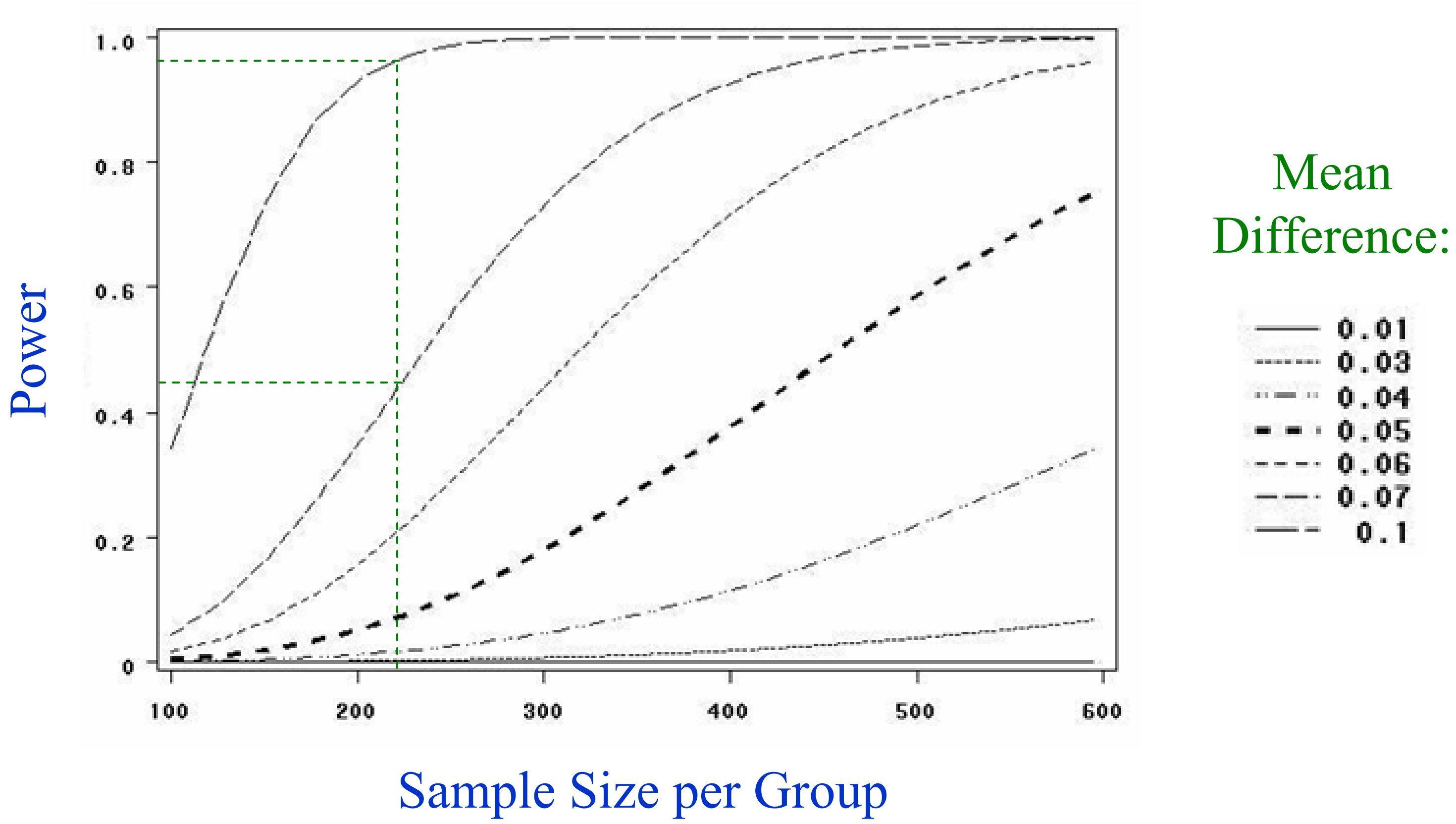
⇨ $$x_i \sim \text{Bin}(m, p) \;\rightarrow\; \bar{x} = \frac{1}{n}\sum x_i \overset{n\rightarrow\infty}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

$$\text{Upper limit} = \frac{0.5 \times 0.5}{n} = \frac{1}{4n}$$

⇨ Multiple testing: Assuming an equivalent to 25,000 independent tests:

$$\alpha^* = 0.05 / 25{,}000 = 0.000002 \quad \text{(Bonferroni)}$$

EXAMPLE



⇨ Previous studies with STAT5A: Differences of 7.7% in fertilization rates and 12.8% in survival rates (Khatib et al. 2008)

EXAMPLE

⇨ However, LD level should be taken into account

⇨ Example: Genetic effect of 12.8%

$$\begin{cases} r^2 = 1 \rightarrow \text{Power} \approx 90\% \\ r^2 = 0.5 \rightarrow \text{Power} \approx 35\% \end{cases}$$

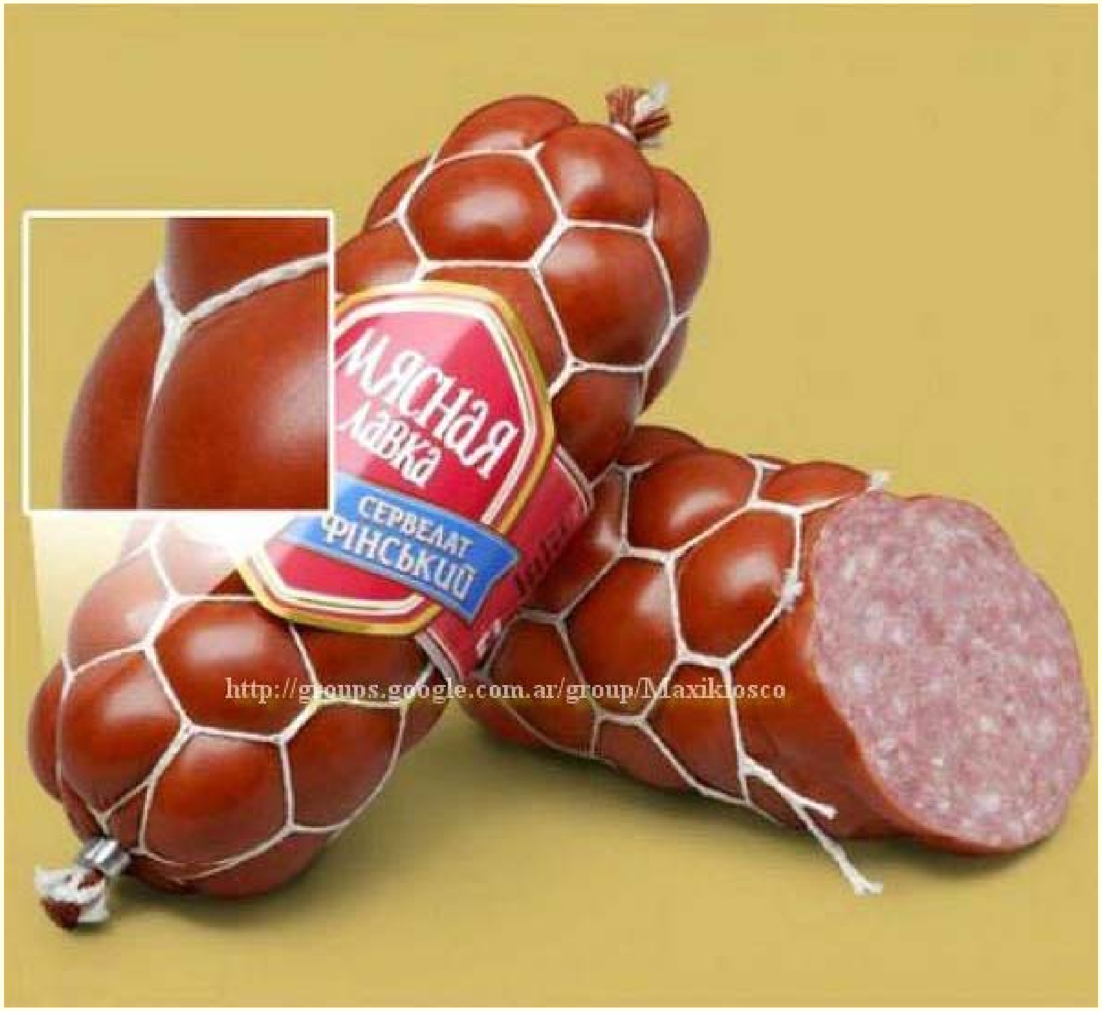
But still approx. 1/3 chance of detecting QTL of such size

⇨ Selective genotyping can improve power:

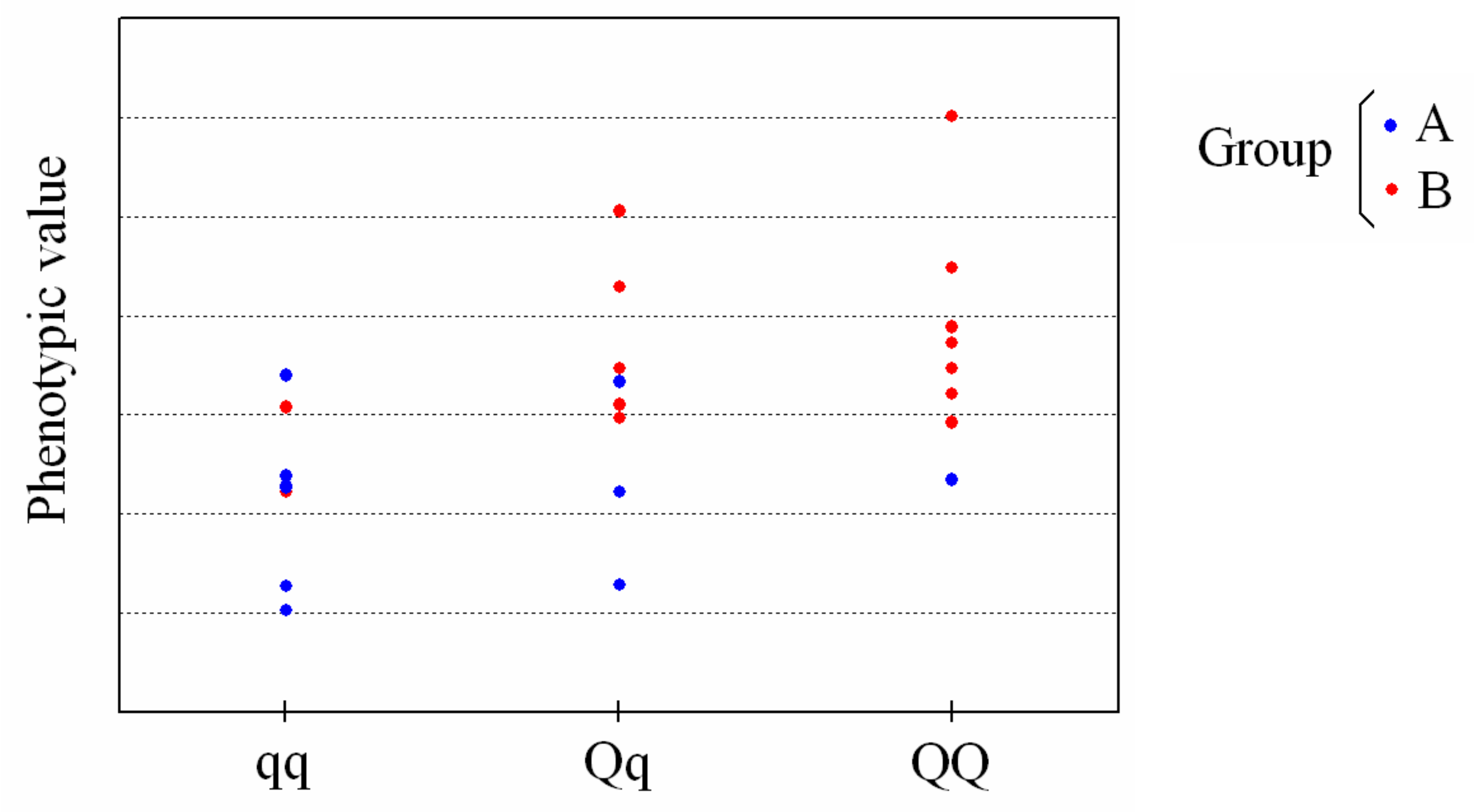
Kathib et al. (2008) estimated survival rates of 52.7 and 25.9% for CC and GG cows, respectively.

VALIDATION

VALIDATION


Мясная лавка
Сервелат
Фінський
http://groups.google.com.ar/group/Maxikiosco

CONFOUNDING



⇨ True model: $y_{ij} = \mu + \text{Group}_i + e_{ij}$

REPLICATION

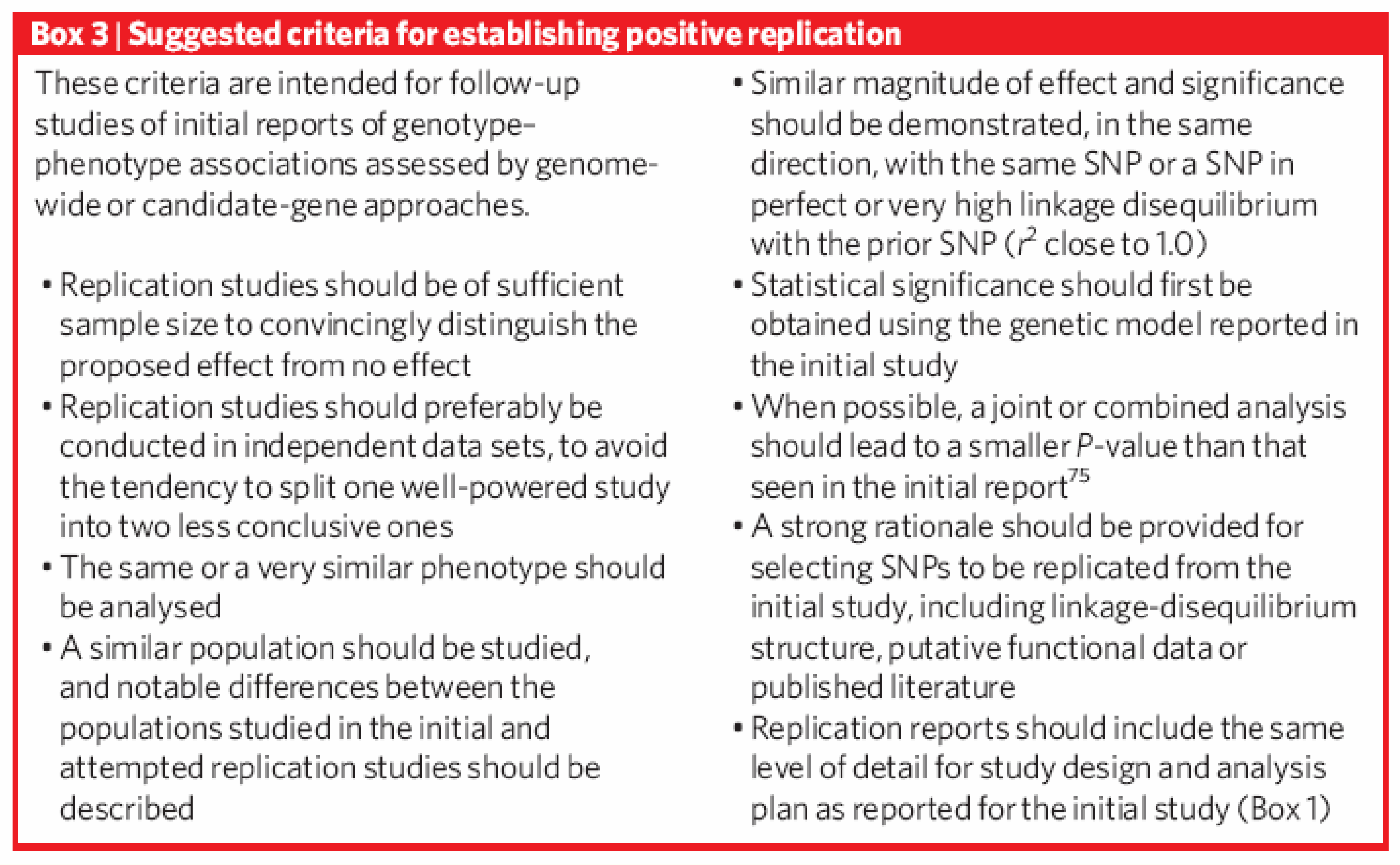
- ⇨ Confounding factors, population structure and stratification, Type I error, etc.
- ⇨ Biased estimates of gene effects due to significance threshold
- ⇨ Multiple genes, with modest individual effects
- ⇨ Gene × gene and gene × environment interactions
- ⇨ Inter population heterogeneity
- ⇨ Low statistical power
- ⇨ Validation of association findings
- ⇨ But what constitutes a replication?

REPLICATION

(Chanock et al. 2007)

- ⇨ Comprehensive reviews of the literature demonstrate a plethora of questionable genotype-phenotype associations, replication of which has often failed in independent studies
- ⇨ “Replication is essential for establishing the credibility of a genotype-phenotype association, whether derived from candidate-gene or genome-wide association studies”
- ⇨ But what consists a replication? How should validation study be performed? ‘Independent’ samples, independent labs, different statistical analysis approach, etc.?
- ⇨ Jiont analysis is more efficient than replication-based analysis for two-stage GWAS (Skol et al. 2006)

REPLICATION



Box 3 | Suggested criteria for establishing positive replication

These criteria are intended for follow-up studies of initial reports of genotype–phenotype associations assessed by genome-wide or candidate-gene approaches.

- Replication studies should be of sufficient sample size to convincingly distinguish the proposed effect from no effect
- Replication studies should preferably be conducted in independent data sets, to avoid the tendency to split one well-powered study into two less conclusive ones
- The same or a very similar phenotype should be analysed
- A similar population should be studied, and notable differences between the populations studied in the initial and attempted replication studies should be described
- Similar magnitude of effect and significance should be demonstrated, in the same direction, with the same SNP or a SNP in perfect or very high linkage disequilibrium with the prior SNP (r^2 close to 1.0)
- Statistical significance should first be obtained using the genetic model reported in the initial study
- When possible, a joint or combined analysis should lead to a smaller *P*-value than that seen in the initial report[75]
- A strong rationale should be provided for selecting SNPs to be replicated from the initial study, including linkage-disequilibrium structure, putative functional data or published literature
- Replication reports should include the same level of detail for study design and analysis plan as reported for the initial study (Box 1)

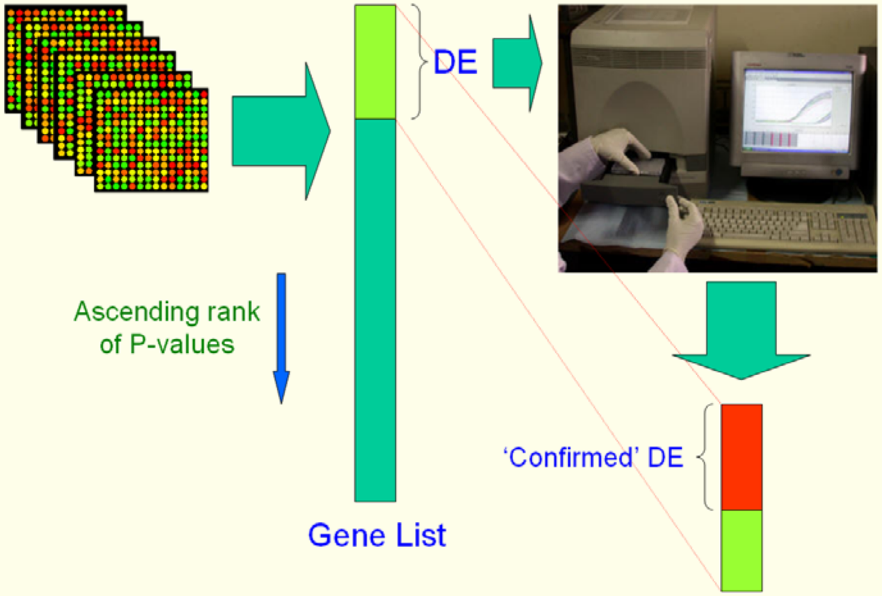
TWO-STAGE DESIGNS

⇨ GWAS (Satagopan et al. 2003, Skol et al. 2007)

- 1st stage: All markers available
- 2nd stage: Selected markers

⇨ Transcriptional Profiling (Steibel et al. 2008)

- 1st stage: Microarray chips
- 2nd stage: qRT-PCR



CONCLUDING REMARKS

- Current (or oncoming) 50-60 K SNP chips provide reasonable genome coverage in cattle, pig and chicken
- Sample sizes still limited for reasonable power, except for ‘major’ QTNs
- Two-stage studies with selective genotyping may reduce costs and improve results
- Appropriate design and statistical analysis of GWAS
  - High dimensionality
  - Multiple testing
  - G × G and G × E interactions

ACKNOWLEDGMENTS

THE UNIVERSITY of WISCONSIN MADISON

Dan Gianola

Kent Weigel

Hasan Khatib

Nick Wu

Oscar González-Recio

Nanye Long

Gustavo de los Campos

Bill Muir PURDUE UNIVERSITY

Hans Cheng USDA ars

Morris Soller The Hebrew University of Jerusalem

Juan Pedro Steibel

Rob Tempelman

Cathy Ernst

Ron Bates

MICHIGAN STATE UNIVERSITY

Santiago Avendaño

Aviagen

CSREES EDUCATION USDA RESEARCH EXTENSION

USDA 2004-35604-14580

USDA 2004-33120-18646