Introduction to GWAS

Linkage Disequilibrium and Linear Regression

Christian Werner

(Quantitative geneticist and biostatistician) **EiB, CIMMYT**, Texcoco (Mexico)

Filippo Biscarini HerrFalloppio

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

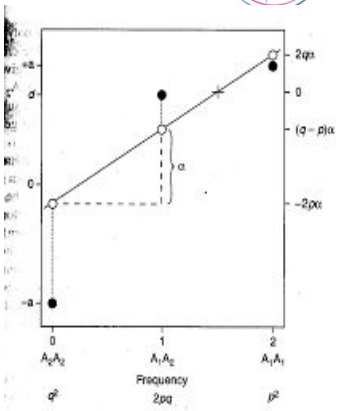
Oscar González-Recio OscarGenomics

(Computational biologist and quantitative geneticist) **INIA-UPM**, Madrid (Spain)

– What marker is associated with the phenotype?
– What individual is genetically more susceptible?

trait	id	snp1	snp2	snp3	snp4	snp5	snp6
9.616	1	Aa	Aa	AA	Aa	Aa	AA
10.140	2	aa	AA	AA	AA	AA	AA
10.687	3	AA	AA	aa	AA	aa	Aa
9.906	4	Aa	aa	Aa	AA	AA	aa

– What marker is associated with the phenotype?
– What individual is genetically more susceptible?



f(A)= mean(Aa)-mean(aa)

Allele substitution effect

The effect that the presence of a copy of an allele has on the phenotype (regarding the reference allele).

snp1	snp2	snp3	snp4	snp5	snp6
-0.4	0	0	0.2	-0.4	0.4

– What marker is associated with the phenotype?
– What individual is genetically more susceptible?

Allele substitution effect

P=G+E (may dominate over G)

snp1	snp2	snp3	snp4	snp5	snp6
-0.4	0	0	0.2	-0.4	0.4

GWAS aims to discover the loci with causal effect and their magnitude

– What marker is associated with the phenotype?
– What individual is genetically more susceptible?

id (pop. μ=10)	Genetic merit	trait	id
1	0.20	9.616	1
2	0.40	10.140	2
3	0.00	10.687	3
4	-0.80	9.906	4

Predicting the genetic response, the phenotype or the disease susceptibility is approached in Genome-Wide Prediction

Genome-wide association studies (GWAS)

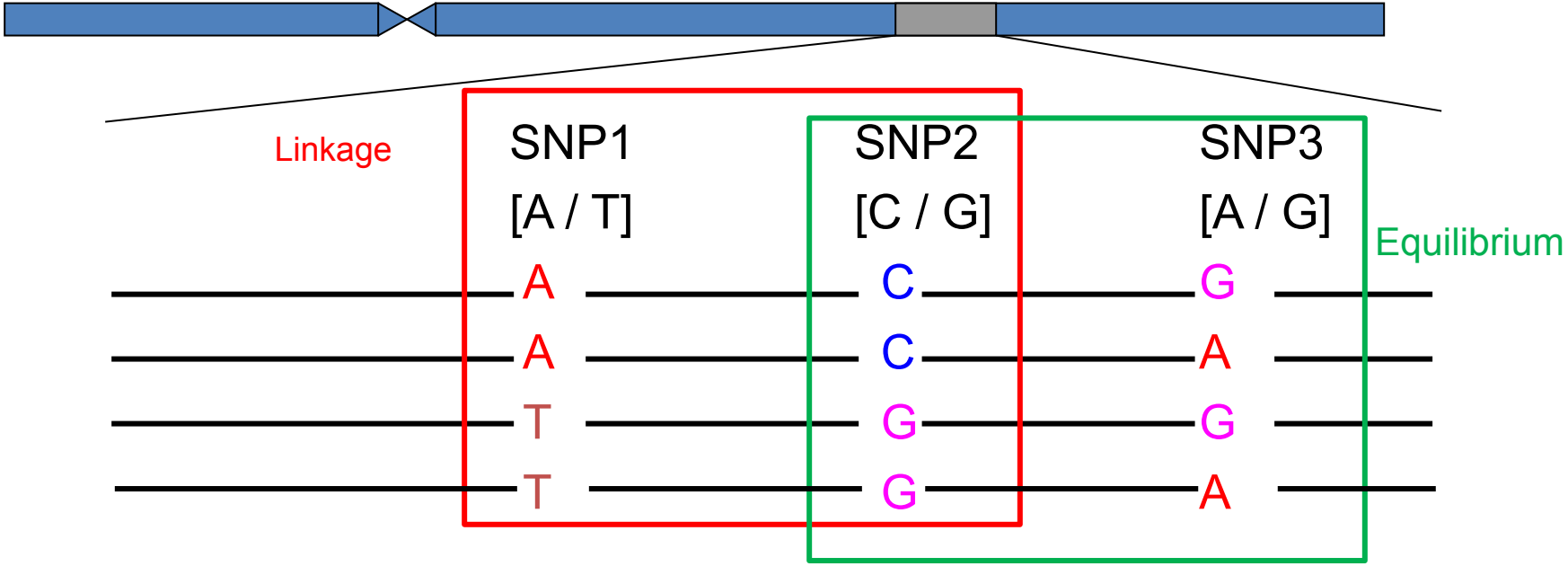
- Based on linear regression models (mathematics, linear algebra, statistics)
- Uses linkage disequilibrium between genomic markers and genes

Genome-wide association studies (GWAS)

- Linkage disequilibrium
- Linear regression

Linkage disequilibrium recap

Linkage disequilibrium (LD)



Haplotype: specific combination of alleles that appear in the same chromosome or segment (*in-cis*) and are inherited together from a single parental.

Linkage disequilibrium (LD)

Major allele frequency:	p(A)	p(B)
Minor allele frequency:	p(a)	p(b)

<u>Marker segregation:</u>

Haplotype frequency $p(ab) = p(a) \times p(b)$ Linkage equilibrium

Haplotype frequency $p(ab) \neq p(a) \times p(b)$ Linkage disequilibrium

Linkage disequilibrium (LD)

SNP2 Allele

SNP1 Allele	B	b	
A	AB p(A)p(B)	p(A)p(b) Ab	p(A)
a	aB p(a)p(B)	p(a)p(b) ab	p(a)
	p(B)	p(b)	

Example:

p(ab)=p(a)p(b) **Expected Haplotype Frequencies**

p(A)p(B)+p(a)p(B)=p(B){ p(A)+p(a) } = p(B)

Linkage disequilibrium (LD)

SNP2 Allele

SNP1 Allele	B	b	
A	AB p(A)p(B)+D	p(A)p(b)-D Ab	p(A)
a	aB p(a)p(B)-D	p(a)p(b) +D ab	p(a)
	p(B)	p(b)	

Example:

$p(ab) \neq p(a)p(b) \rightarrow p(ab)=p(a)p(b)+D$

Non-Expected Haplotype Frequencies (*D is LD degree*)

$\{p(A)p(B)+D\}+\{p(a)p(B)-D\}=p(B)\{ p(A)+p(a) \} = p(B)$

Linkage disequilibrium (LD)

Question: What markers are in LD?

- A frequency=0.6; B frequency=0.3; haplotype **ab** frequency =0.40
- a frequency=0.4; C frequency=0.3; haplotype AC frequency=0.18
- haplotype **b**C frequency=0.49

Linkage disequilibrium (LD)

Most common measurements

-1 ← D′ (Lewontin, 1964) → 1

Recombination

0 ← r^2 → 1

Correlation

Linkage disequilibrium (LD)

D’ (Lewontin, 1964)

$$D' = D / D_{max}$$

$$D_{max} = \min(p(A)p(B), p(a)p(b)) \quad D < 0$$

$$D_{max} = \min(p(A)p(b), p(a)p(B)) \quad D > 0$$

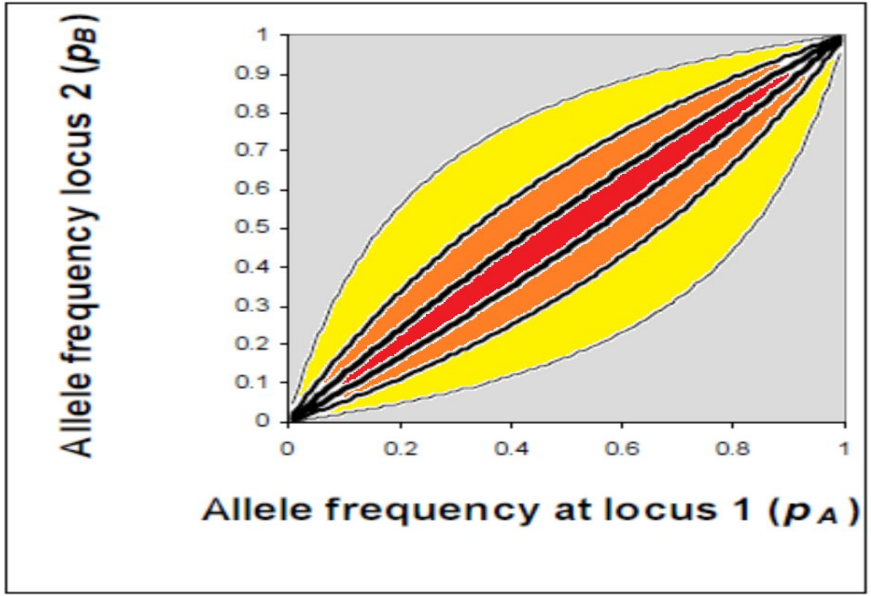
D’ sign depends on the reference allele

Linkage disequilibrium (LD)

r^2 (Hill & Robertson, 1968)

$$r^2 = \frac{D^2}{p(a)p(b)p(A)p(B)}$$

- Between 0 - 1
- Depends on allele frequency
- Most common metric



r^2>0.80
r^2>0.50
r^2>0.20
r^2<0.20

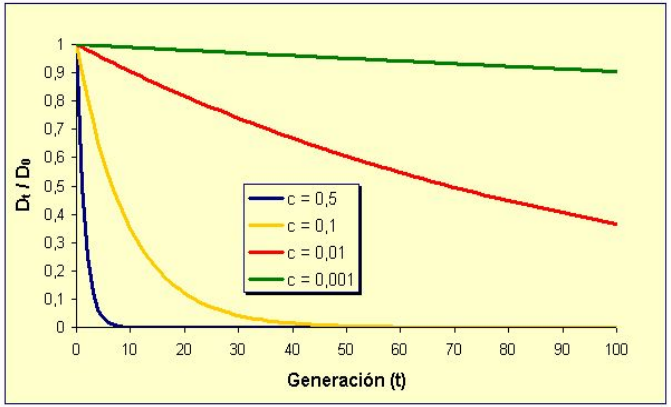
Linkage disequilibrium (LD)

- Non-random allele segregation between 2 or more loci

- Loci on the same chromosome with LD have larger probability of being transmitted together (less likely recombination)

- Physically close loci tend to be in higher LD, as recombination rates are low

Linkage disequilibrium (LD)



Possible causes of LD

- Close location
- Low recombination rates
- Recent populations with low founder size
- Selective breeding (e.g. Holstein Long range LD)
- New mutations
- Inbreeding
- Population bottlenecks
- Population stratification
- Asexual reproduction

Linkage disequilibrium (LD)

Applications

- We can infer the genotype of one locus given we know the genotype of another locus with LD (Imputation)
- We can predict the effect of a gene, given we know a marker genotype in LD (GWAS – GWP)
- Caution
    - LD may disappear in future generations due to new recombinations. Consider this cautiously when applying results to further apart generations or low LD markers
    - Differentiate between statistical linkage and physical linkage

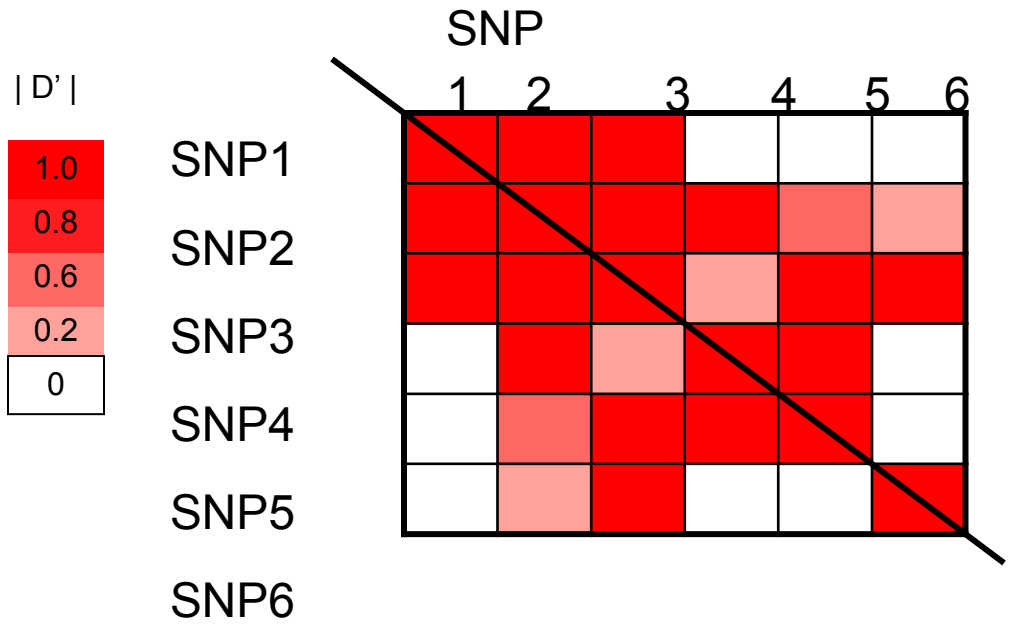
Linkage disequilibrium (LD)

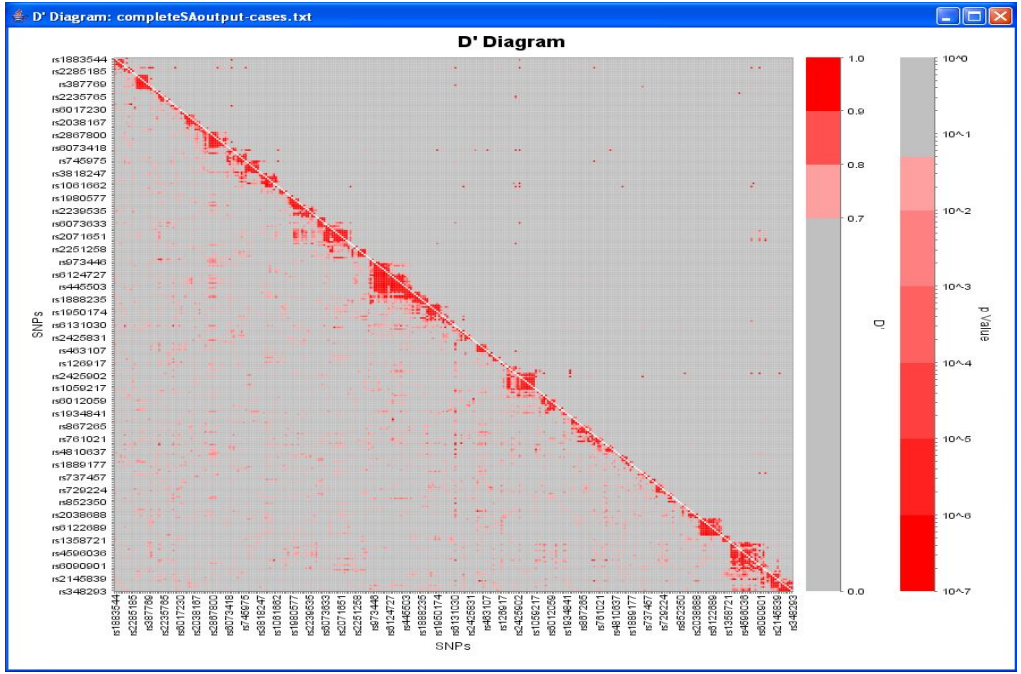
Why do we care about LD?

The power of an indirect association study depends on 4 parameters:

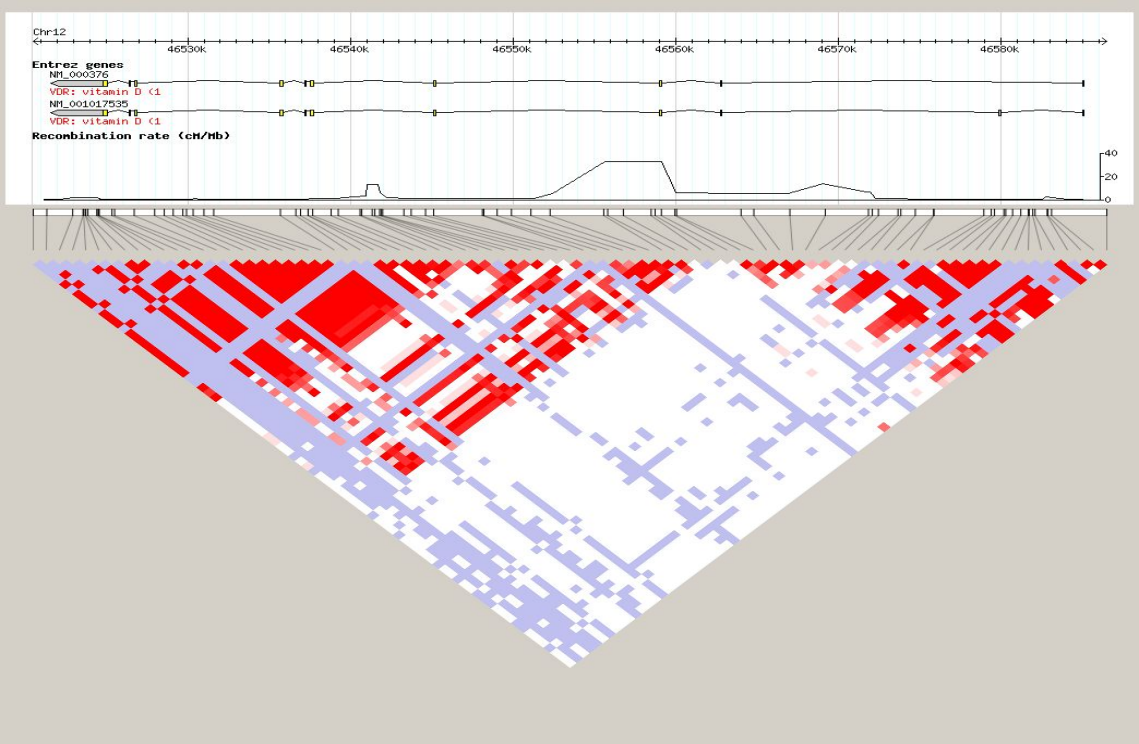
- Disease-allele effect size
- Disease-allele frequency
- Marker-allele frequency
- Extent of LD between marker and disease locus

Visualizing LD


SNP
1
2
3
4
5
6
| D' |
1.0
0.8
0.6
0.2
0
SNP1
SNP2
SNP3
SNP4
SNP5
SNP6


D' Diagram: completeSAoutput-cases.txt
D' Diagram
SNPs
rs1883544
rs2285185
rs387769
rs2235765
rs6017230
rs2038167
rs2867800
rs6073418
rs745975
rs3818247
rs1061862
rs1980577
rs2239535
rs6073633
rs2071651
rs2251258
rs973446
rs6124727
rs445503
rs1888235
rs1950174
rs8131030
rs2425831
rs463107
rs126917
rs2425902
rs1059217
rs6012059
rs1934841
rs867265
rs761021
rs4810637
rs1889177
rs737457
rs729224
rs852350
rs2038688
rs6122689
rs1358721
rs4596036
rs6090901
rs2145839
rs346293
D'
1.0
0.9
0.8
0.7
0.0
p value
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7

Haploview: TCN2 (r^2)

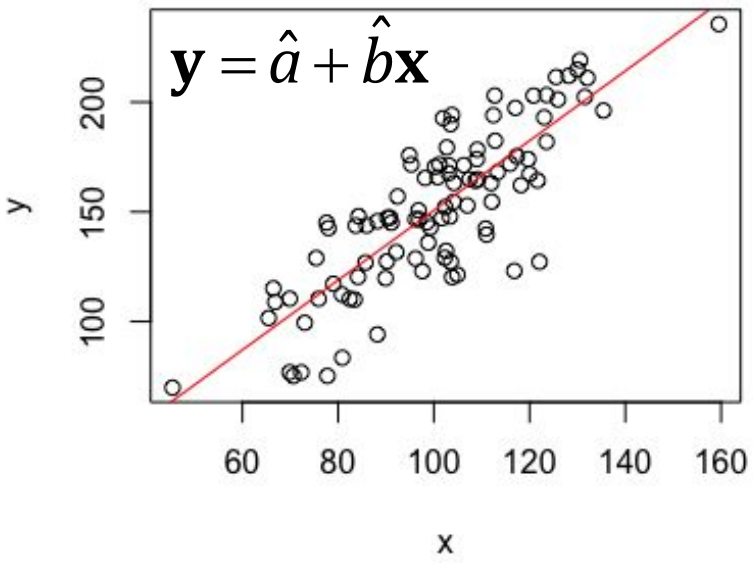


Linear regression recap

Linear regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Estimate the line that minimizes the MSE Sum(y-ÿ)^2.



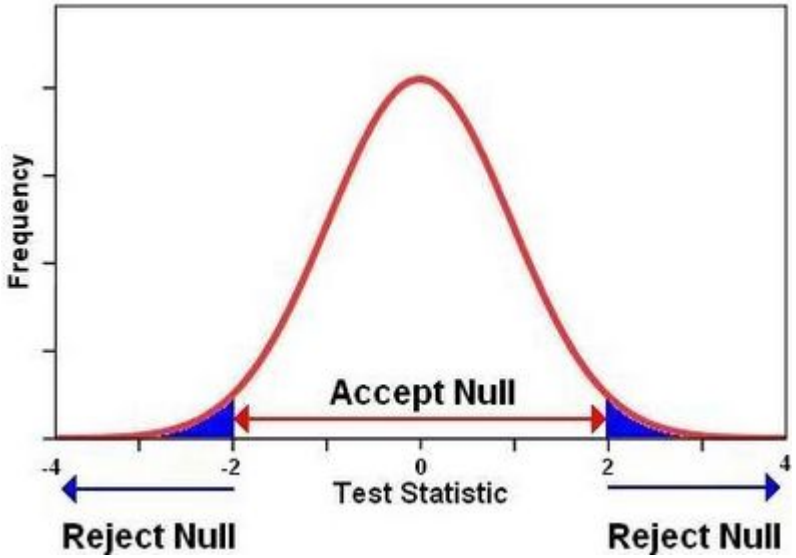
$$\hat{b} = \mathrm{cov}(\mathbf{x},\mathbf{y}) / Var(\mathbf{x})$$

$$\hat{a} = \overline{\mathbf{y}} - \hat{b}\overline{\mathbf{x}}$$

Is the *b* estimate statistically significant? → Real effect (association)

Linear regression – Hypothesis test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$



- Continuous variable → t-Student
- Binary variable → Chi-squared, log-ratio test

<u>Significance level α (e.g. 0.05, 0.01,…)</u>

Linear regression – Hypothesis test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

	One-tail (left)	Two-tails	One-tail (right)
Contrast	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 < b_1$	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 \neq b_1$	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 > b_1$
Statistic (t-Student)	$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\hat{s}_R^{\,2} / SS_{xx}}}$ con $\hat{s}_R^{\,2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}$		
Reject Region	$t < t_{a,n-2}$	$\lvert t \rvert > t_{1-a/2,n-2}$	$t > t_{1-a,n-2}$

$$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\hat{s}_R^{\,2} / SS_{xx}}}$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y},$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = n\sigma_x^2$$

Does my *t-statistic* value belong to the null hypothesis distribution

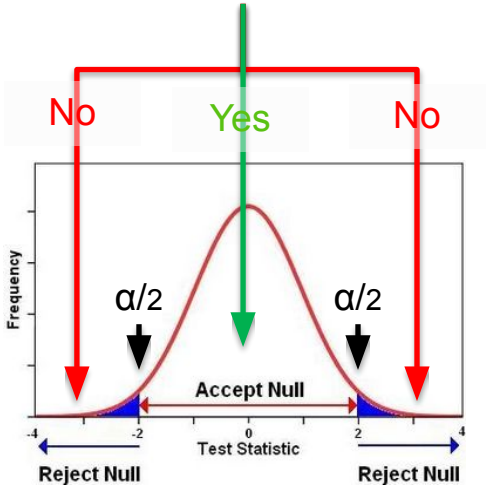
$$\hat{s}_R^{\,2} = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} \qquad SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i)^2$$

Linear regression – Hypothesis test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{\hat{s}_R^{\,2} / SS_{xx}}}$$

Does my *t-statistic* value belong to the null hypothesis distribution



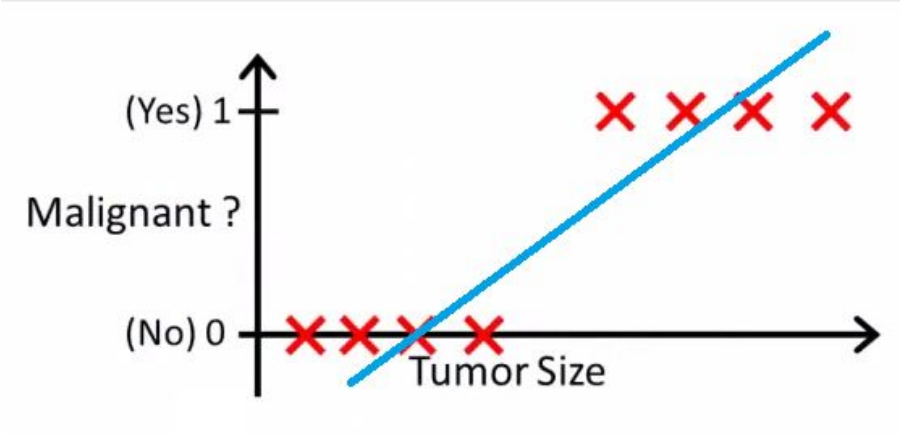
Calculate the ***p-value***: probability of obtaining a value of the statistic at least as extreme as the one obtained in the experiment, if we repeat the experiment an infinite number of times

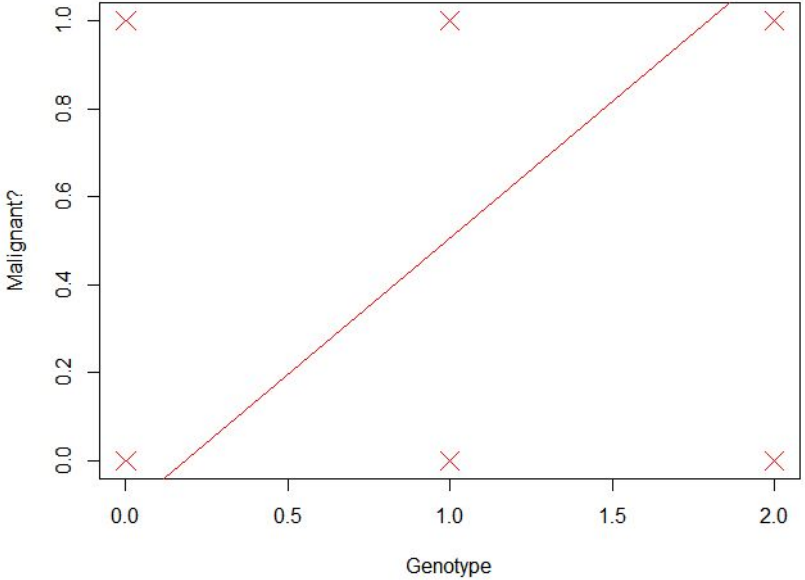
i.e. two-tails test

$$2 \cdot (1 - p(z \leq t))$$

If *p-value*< *α* → reject null-hypothesis

Logistic regression





Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Uses the sigmoid function

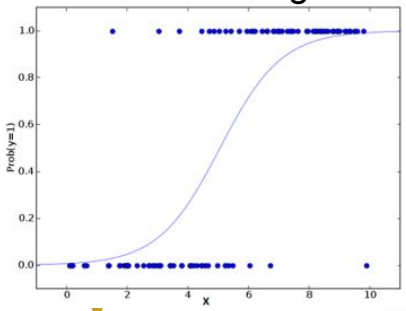
$$\log_{base} \frac{P(Y=1)}{1-P(Y=1)} = a_0 + a_1 x$$

$$\frac{P(Y=1)}{1-P(Y=1)} = base^{(a_0+b_1 x)} \Rightarrow P(Y=1) = base^{(a_0+b_1 x)} - P(Y=1)\, base^{(a_0+b_1 x)} \Rightarrow$$

$$\left(1 + base^{(a_0+b_1 x)}\right) P(Y=1) = base^{(a_0+b_1 x)} \Rightarrow$$

$$P(Y=1) = \frac{base^{(a_0+b_1 x)}}{1+base^{(a_0+b_1 x)}} = \frac{1}{1+base^{-(a_0+b_1 x)}}$$

Sigmoid function



Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Estimation of coefficients *a* and *b*

	0(control)	1(case)
AA	AA_0	AA_1
Aa	Aa_0	Aa_1

The intercept *a* is the log of the odds for the control group (AA)

$$a = \log \frac{AA_1 / AA_{totales}}{AA_0 / AA_{totales}} = \log \frac{AA_1}{AA_0}$$

The regression coefficient *b* is the log of the odds ratio between the Aa group and the AA group:

$$b = \log \frac{\dfrac{Aa_1 / Aa_0}{AA_1 / AA_0}}{\dfrac{Aa_1 / AA_0}{Aa_0 / AA_1}} = \frac{Aa_1 / AA_0}{Aa_0 / AA_1}$$

Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- How to interpret the Odd Ratio
  - It is a measurement of the degree of association between 2 variables, and is estimated with respect to a given base
  - Ranges between 0 and ∞
  - OR=1→ lack of association;OR<1 → negative association; OR>1→ positive association
  - If OR<1, estimate the inverse relationship is preferred for easiness of interpretation

	0(control)	1(case)
AA	AA_0	AA_1
Aa	Aa_0	Aa_1

Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Two examples
  - Colorectal cancer and meat consumption.
    USA male (30-55yr) and male (40-75yr)
    High vs low levels of processed meat intake

    Population prevalence: 1%
    OR=1.22

Meta-Analysis > Int J Cancer. 2006 Dec 1;119(11):2657-64. doi: 10.1002/ijc.22170.

Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies

Susanna C Larsson [1], Alicja Wolk

Affiliations + expand

PMID: 16991129 DOI: 10.1002/ijc.22170 Sign in

Free article

Abstract

Accumulating epidemiologic evidence indicates that high consumption of red meat and of processed meat may increase the risk of colorectal cancer. We quantitatively assessed the association between red meat and processed meat consumption and the risk of colorectal cancer in a meta-analysis of prospective studies published through March 2006. Random-effects models were used to pool study results and to assess dose-response relationships. We identified 15 prospective studies on red meat (involving 7,367 cases) and 14 prospective studies on processed meat consumption (7,903 cases). The summary relative risks (RRs) of colorectal cancer for the highest vs. the lowest intake categories were 1.28 (95% confidence interval (CI) = 1.15-1.42) for red meat and 1.20 (95% CI = 1.11-1.31) for processed meat. The estimated summary RRs were 1.28 (95% CI = 1.18-1.39) for an increase of 120 g/day of red meat and 1.09 (95% CI = 1.05-1.13) for an increase of 30 g/day of processed meat. Consumption of red meat and processed meat was positively associated with risk of both colon and rectal cancer, although the association with red meat appeared to be stronger for rectal cancer. In 3 studies that reported results for subsites in the colon, high consumption of processed meat was associated with an increased risk of distal colon cancer but not of proximal colon cancer. The results of this meta-analysis of prospective studies support the hypothesis that high consumption of red meat and of processed meat is associated with an increased risk of colorectal cancer.

Prevalence in the high intake levels of meat ?

Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Two examples
  - Colorectal cancer and meat consumption.
    USA male (30-55yr) and male (40-75yr)
    High vs low levels of processed meat intake

    Population prevalence: 1%
    OR=1.22

Meta-Analysis > Int J Cancer. 2006 Dec 1;119(11):2657-64. doi: 10.1002/ijc.22170.

Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies

Susanna C Larsson [1], Alicja Wolk

Affiliations + expand

PMID: 16991129 DOI: 10.1002/ijc.22170 Sign in

Free article

Abstract

Accumulating epidemiologic evidence indicates that high consumption of red meat and of processed meat may increase the risk of colorectal cancer. We quantitatively assessed the association between red meat and processed meat consumption and the risk of colorectal cancer in a meta-analysis of prospective studies published through March 2006. Random-effects models were used to pool study results and to assess dose-response relationships. We identified 15 prospective studies on red meat (involving 7,367 cases) and 14 prospective studies on processed meat consumption (7,903 cases). The summary relative risks (RRs) of colorectal cancer for the highest vs. the lowest intake categories were 1.28 (95% confidence interval (CI) = 1.15-1.42) for red meat and 1.20 (95% CI = 1.11-1.31) for processed meat. The estimated summary RRs were 1.28 (95% CI = 1.18-1.39) for an increase of 120 g/day of red meat and 1.09 (95% CI = 1.05-1.13) for an increase of 30 g/day of processed meat. Consumption of red meat and processed meat was positively associated with risk of both colon and rectal cancer, although the association with red meat appeared to be stronger for rectal cancer. In 3 studies that reported results for subsites in the colon, high consumption of processed meat was associated with an increased risk of distal colon cancer but not of proximal colon cancer. The results of this meta-analysis of prospective studies support the hypothesis that high consumption of red meat and of processed meat is associated with an increased risk of colorectal cancer.

Prevalence in the high intake levels of meat ?

answer: 1.22%

Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Two examples
  - COVID-19 and Neardenthal gene (Chr3).
    UK Biobank (679531 controls), severe COVID (1128 cases)

Table 1 | Lead variants from independent genome-wide significant regions

SNP	Chr.: pos.	Risk	Alt.	RAF_{gcc}	RAF_{ukb}	OR	CI	$P_{gcc.ukb}$	$P_{gcc.gs}$	$P_{gcc.100k}$	Locus
rs73064425	3: 45,901,089	T	C	0.15	0.07	2.1	1.88–2.45	4.8×10^{-30}	2.9×10^{-27}	3.6×10^{-32}	*LZTFL1*
rs9380142	6: 29,798,794	A	G	0.74	0.69	1.3	1.18–1.43	3.2×10^{-8}	0.00091	1.8×10^{-8}	*HLA-G*
rs143334143	6: 31,121,426	A	G	0.12	0.07	1.8	1.61–2.13	8.8×10^{-18}	2.6×10^{-24}	5.8×10^{-18}	*CCHCR1*
rs3131294	6: 32,180,146	G	A	0.9	0.86	1.5	1.28–1.66	2.8×10^{-8}	1.3×10^{-10}	2.3×10^{-8}	*NOTCH4*
rs10735079	12: 113,380,008	A	G	0.68	0.63	1.3	1.18–1.42	1.6×10^{-8}	2.8×10^{-9}	4.7×10^{-6}	*OAS1–OAS3*
rs2109069	19: 4,719,443	A	G	0.38	0.32	1.4	1.25–1.48	4×10^{-12}	4.5×10^{-7}	2.4×10^{-8}	*DPP9*
rs74956615	19: 10,427,721	A	T	0.079	0.05	1.6	1.35–1.87	2.3×10^{-8}	2.2×10^{-13}	3.9×10^{-6}	*TYK2*
rs2236757	21: 34,624,917	A	G	0.34	0.28	1.3	1.17–1.41	5×10^{-8}	8.9×10^{-5}	8.3×10^{-7}	*IFNAR2*

As this is a meta-analysis of all available data, external replication cannot be attempted, so SNPs are included if they meet a more stringent *P*-value threshold of $P < 10^{-8}$. SNP, the strongest SNP in the locus. Chr.: pos., chromosome and position of the top SNP (build 37); Risk, risk allele; Alt., alternative allele; RAF, risk allele frequency; OR, effect size (odds ratio) of the risk allele in the GenOMICC European analysis; CI, 95% confidence interval for the odds ratio in the GenOMICC European cohort; *P*, *P* value; Locus, gene nearest to the top SNP. Subscript identifiers indicate the cohorts used for cases (gcc, GenOMICC European cohort) and controls (ukb, UK Biobank; gs, Generation Scotland; 100k, 100,000 Genomes Project).

Population prevalence: 0.16%
OR=2.1

Prevalence in the Risk Allele group ?

Logistic regression

$$\mathbf{y} = a + b\mathbf{x} + \mathbf{e}$$

- Two examples
  - COVID-19 and Neardenthal gene (Chr3).
    UK Biobank (679531 controls), severe COVID (1128 cases)

Table 1 | Lead variants from independent genome-wide significant regions

SNP	Chr.: pos.	Risk	Alt.	RAF_{gcc}	RAF_{ukb}	OR	CI	$P_{gcc,ukb}$	$P_{gcc,gs}$	$P_{gcc,100k}$	Locus
rs73064425	3: 45,901,089	T	C	0.15	0.07	2.1	1.88–2.45	4.8×10^{-30}	2.9×10^{-27}	3.6×10^{-32}	*LZTFL1*
rs9380142	6: 29,798,794	A	G	0.74	0.69	1.3	1.18–1.43	3.2×10^{-8}	0.00091	1.8×10^{-8}	*HLA-G*
rs143334143	6: 31,121,426	A	G	0.12	0.07	1.8	1.61–2.13	8.8×10^{-18}	2.6×10^{-24}	5.8×10^{-18}	*CCHCR1*
rs3131294	6: 32,180,146	G	A	0.9	0.86	1.5	1.28–1.66	2.8×10^{-8}	1.3×10^{-10}	2.3×10^{-8}	*NOTCH4*
rs10735079	12: 113,380,008	A	G	0.68	0.63	1.3	1.18–1.42	1.6×10^{-8}	2.8×10^{-9}	4.7×10^{-6}	*OAS1–OAS3*
rs2109069	19: 4,719,443	A	G	0.38	0.32	1.4	1.25–1.48	4×10^{-12}	4.5×10^{-7}	2.4×10^{-8}	*DPP9*
rs74956615	19: 10,427,721	A	T	0.079	0.05	1.6	1.35–1.87	2.3×10^{-8}	2.2×10^{-13}	3.9×10^{-6}	*TYK2*
rs2236757	21: 34,624,917	A	G	0.34	0.28	1.3	1.17–1.41	5×10^{-8}	8.9×10^{-5}	8.3×10^{-7}	*IFNAR2*

As this is a meta-analysis of all available data, external replication cannot be attempted, so SNPs are included if they meet a more stringent *P*-value threshold of $P < 10^{-8}$. SNP, the strongest SNP in the locus. Chr.: pos., chromosome and position of the top SNP (build 37); Risk, risk allele; Alt., alternative allele; RAF, risk allele frequency; OR, effect size (odds ratio) of the risk allele in the GenOMICC European analysis; CI, 95% confidence interval for the odds ratio in the GenOMICC European cohort; *P*, *P* value; Locus, gene nearest to the top SNP. Subscript identifiers indicate the cohorts used for cases (gcc, GenOMICC European cohort) and controls (ukb, UK Biobank; gs, Generation Scotland; 100k, 100,000 Genomes Project).

nature

Explore content | Journal information | Publish with us

nature > articles > article

Article | Published: 11 December 2020

Genetic mechanisms of critical illness in COVID-19

Erola Pairo-Castineira, Sara Clohisey, [...] J. Kenneth Baillie

Nature **591**, 92–98(2021) | Cite this article

11k Accesses | 55 Citations | 2734 Altmetric | Metrics

Population prevalence: 0.16%

OR=2.1

Prevalence in the Risk Allele group ?

answer: 0.34%

(needs to be considered carefully, because control group did not report COVID19; we don't know what would have been the progress of the disease)

Linear regression – Hypothesis test

Now you!

Compute the t-student statistic given some data, and accept or reject the null-hypothesis

1. Basis_of_linear_regression.R
2. Basis_of_logistic_regression.Rmd
3. Homework: Analyze data in Exercise1.exampleData.xlsx,
   a) Determine SNPs associated to the phenotype, and calculate the Odds-Ratio
   b) If you got 2 significant SNPs, that's wrong. Figure out why.