Meta-analysis with a general genetic model:
ACTN3 & athletic performance

Damjan Vukcevic

Centre for Systems Genomics
University of Melbourne

25 May 2017
ViCBiostat Seminar

Systems Genomics

ViCBiostat

THE UNIVERSITY OF MELBOURNE

Murdoch Childrens Research Institute
Healthier Kids. Healthier Future.

Overview

Part 1

- Background
- Data
- Model
- Results

Part 2

- Simpler (misspecified) models
- Covariates
- Some properties of the model
- Questions for the audience

Part 1

Background
Data
Model
Results

ACTN3 and muscle fibres

The gene ***ACTN3***

Encodes the protein **alpha-actinin-3**

Expressed in **fast twitch muscle fibres**

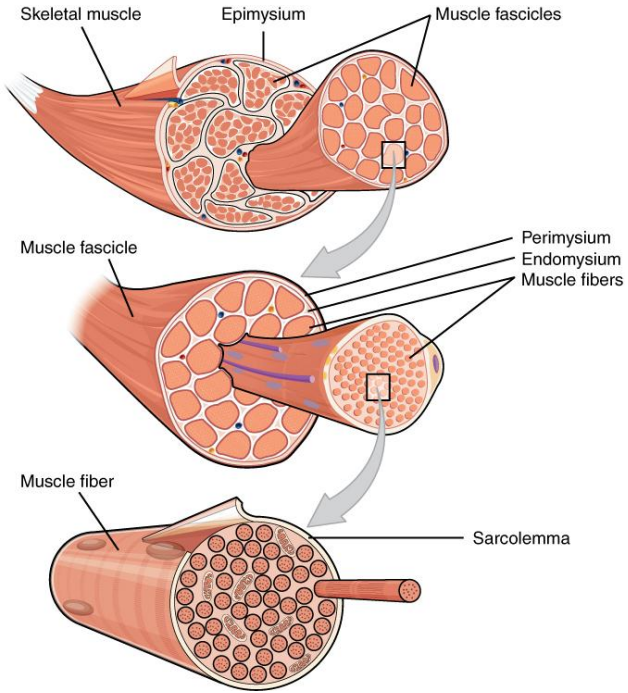
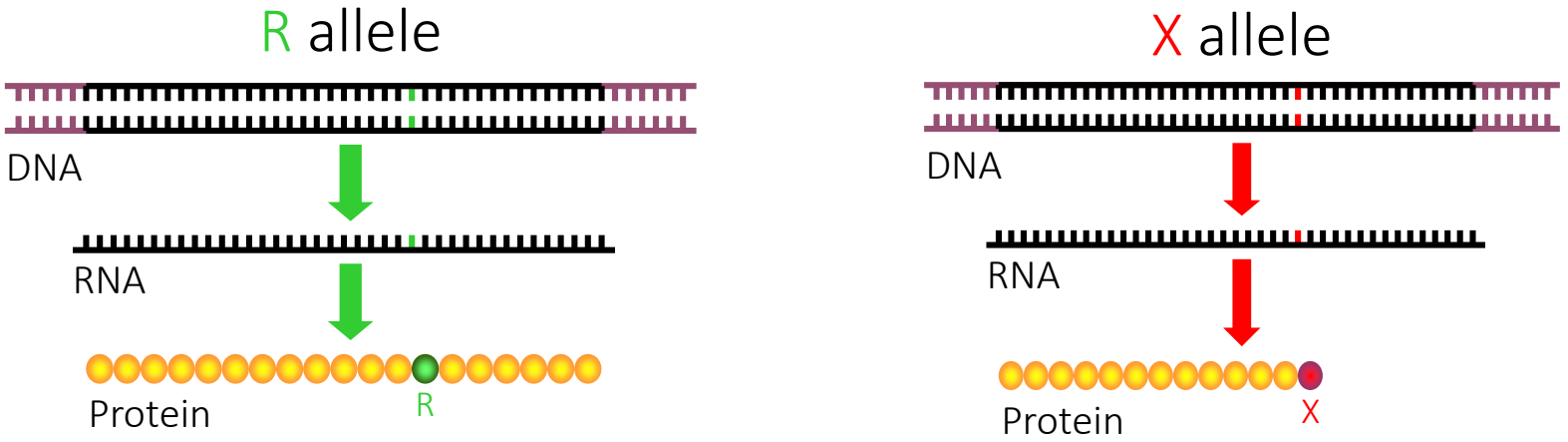


Image: Wikimedia Commons

R577X mutation in *ACTN3*



Am. J. Hum. Genet. 73:627–631, 2003

Report

ACTN3 Genotype Is Associated with Human Elite Athletic Performance

Nan Yang,[1] Daniel G. MacArthur,[1,2] Jason P. Gulbin,[3] Allan G. Hahn,[3] Alan H. Beggs,[5] Simon Easteal,[4] and Kathryn North[1,2]

[1]Institute for Neuromuscular Research, Children's Hospital at Westmead and [2]Discipline of Paediatrics and Child Health, Faculty of Medicine, University of Sydney, Sydney; [3]Australian Institute of Sport and [4]Human Genetics Group, John Curtin School of Medical Research, Australian National University, Canberra; and [5]Genetics Division, Children's Hospital, Boston

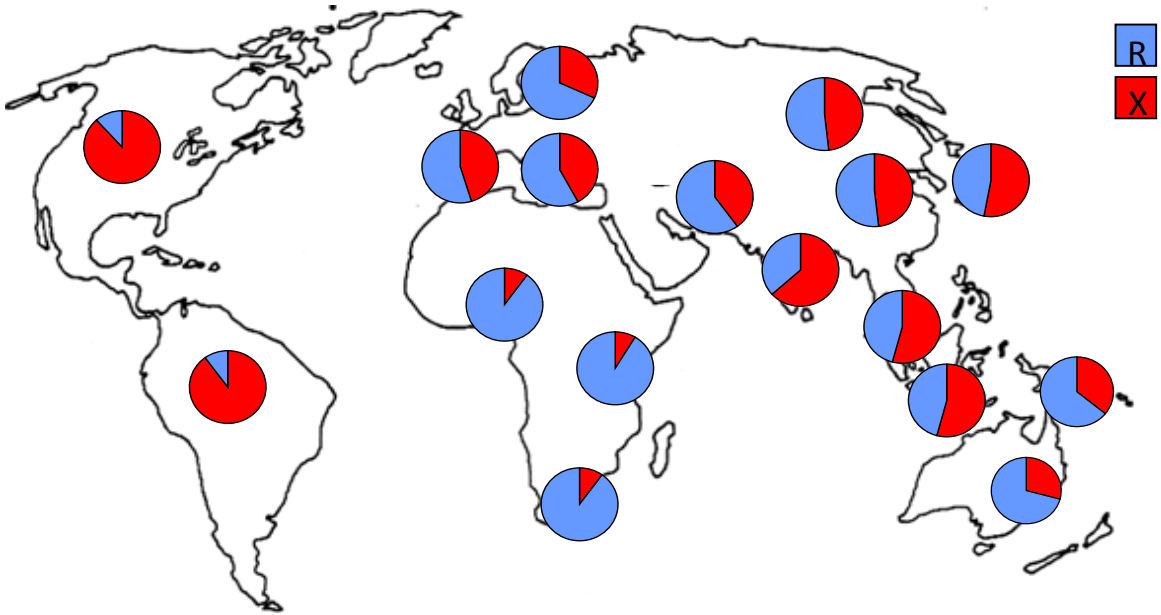
Image: Wikimedia Commons

'The **gene** for **speed**'

NATURE GENETICS VOLUME 39 | NUMBER 10 | OCTOBER 2007

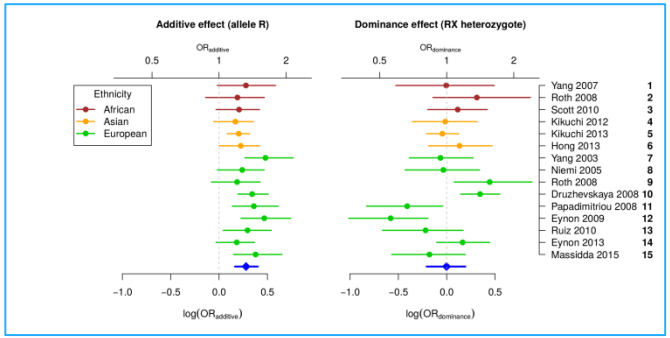
Loss of *ACTN3* gene function alters mouse muscle metabolism and shows evidence of positive selection in humans

Daniel G MacArthur[1,2], Jane T Seto[1,2], Joanna M Raftery[1], Kate G Quinlan[1,2], Gavin A Huttley[3], Jeff W Hook[4], Frances A Lemckert[4], Anthony J Kee[5], Michael R Edwards[6], Yemima Berman[1], Edna C Hardeman[5], Peter W Gunning[2,4], Simon Easteal[3], Nan Yang[1] & Kathryn N North[1,2]

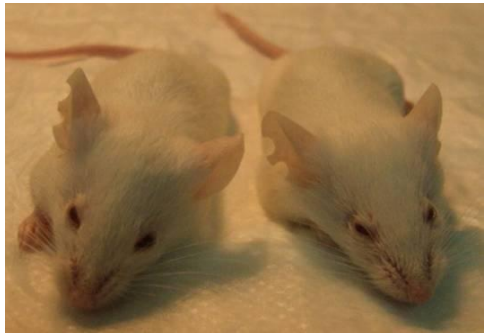


Aim

Study the effect of the **heterozygotes (RX)**



Meta-analysis



Novel experiments

Data

13 studies

Case-control design (**athletes vs controls**)

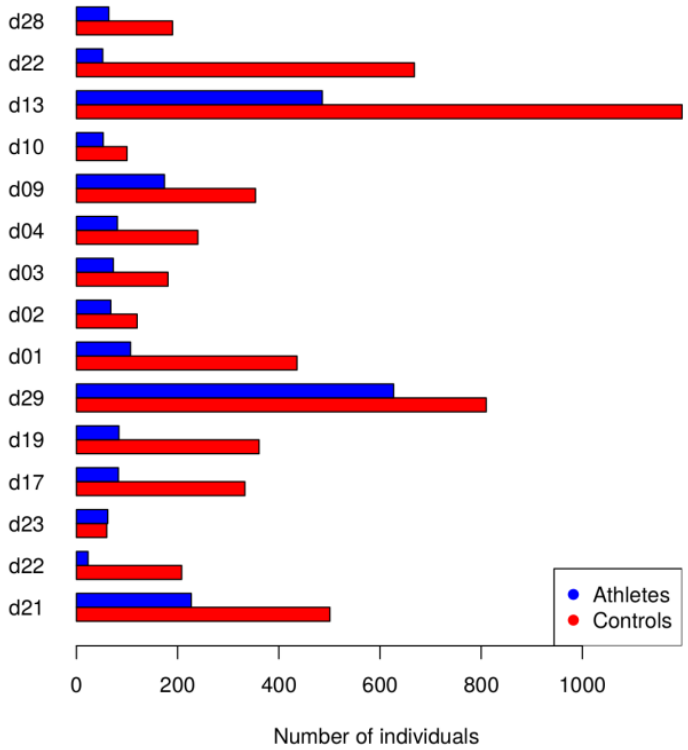
Phenotype: Elite athletic performance

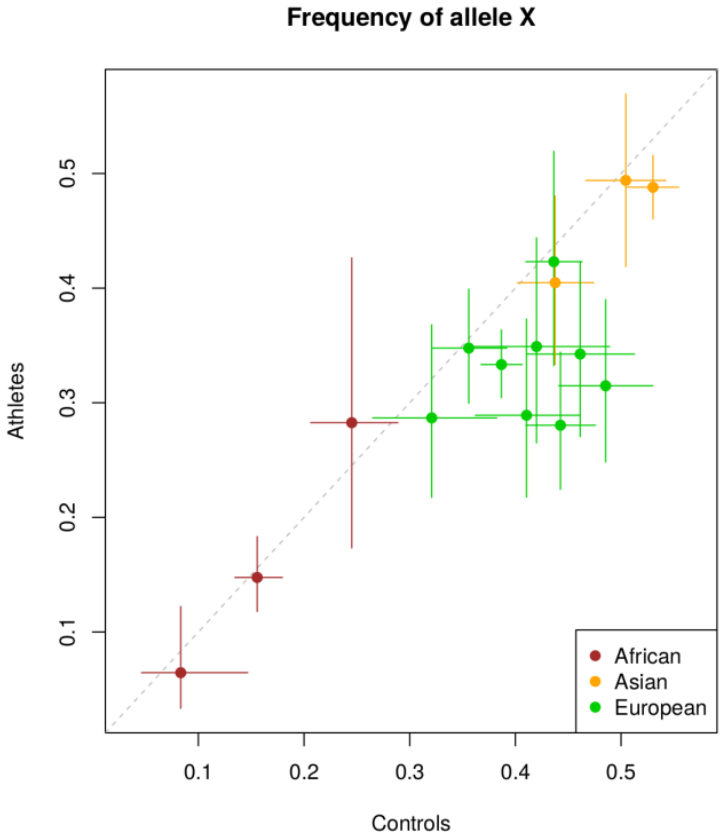
Genotypes: **rs1815739** (causes R → X)

Example:
(Papadimitriou 2008)

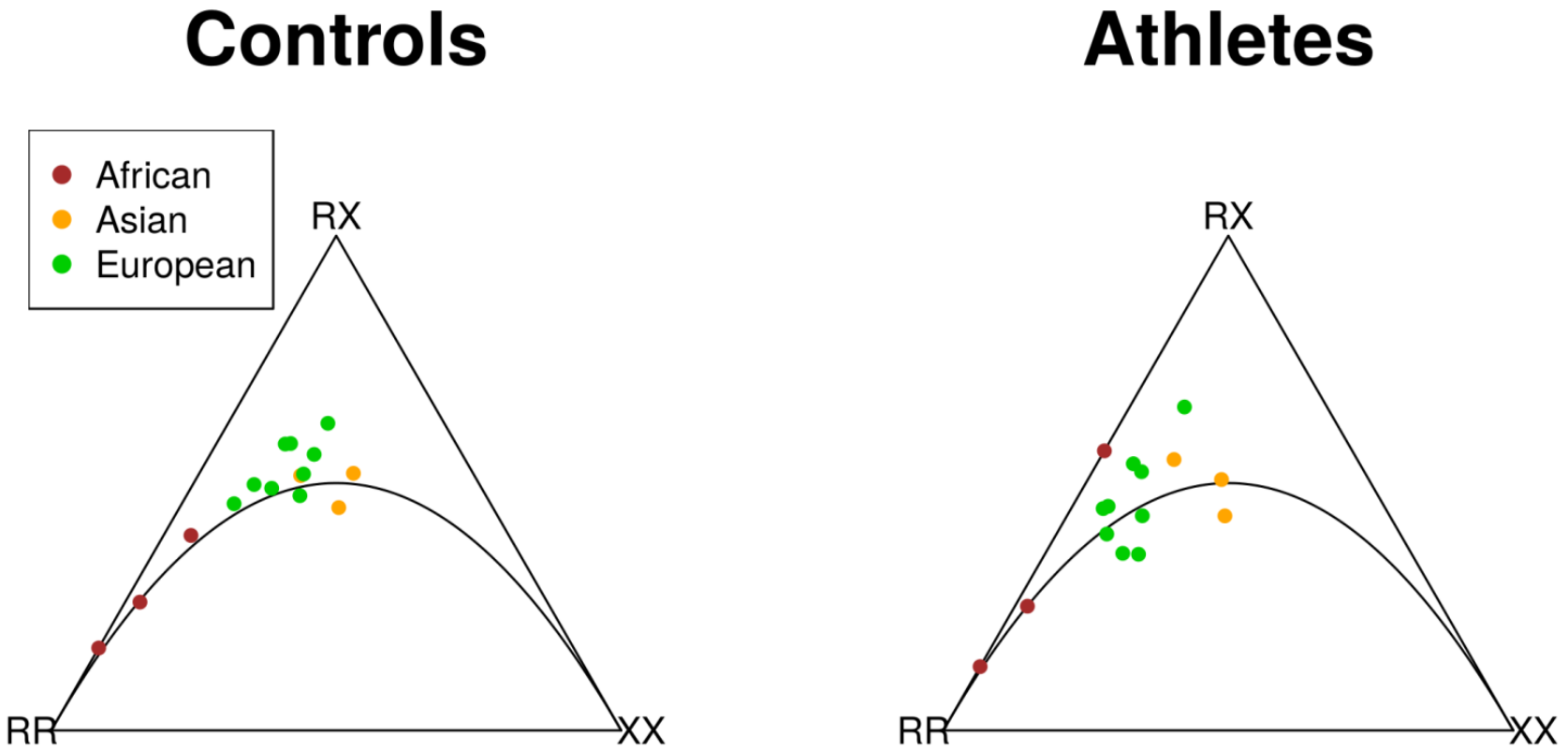
	RR	RX	XX
Athletes	35	26	12
Controls	47	101	33

Data

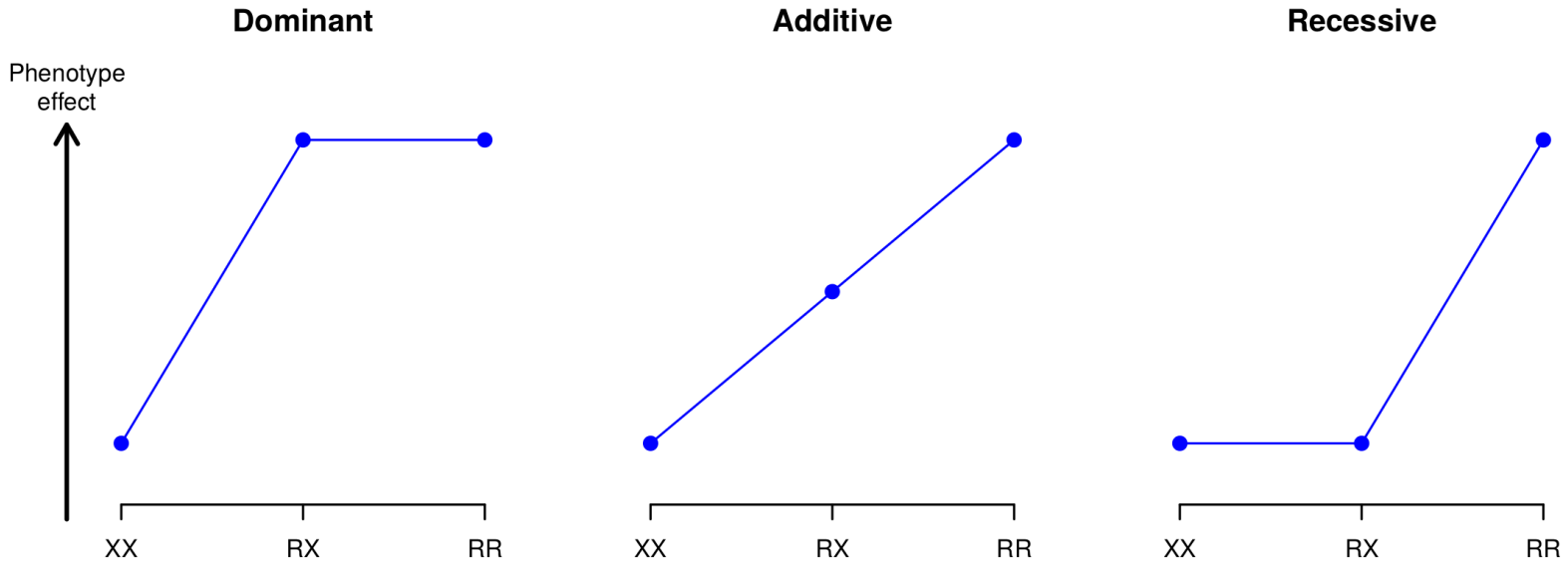

d28
d22
d13
d10
d09
d04
d03
d02
d01
d29
d19
d17
d23
d22
d21
Athletes
Controls
0
200
400
600
800
1000
Number of individuals


Frequency of allele X
Athletes
0.1
0.2
0.3
0.4
0.5
Controls
0.1
0.2
0.3
0.4
0.5
African
Asian
European

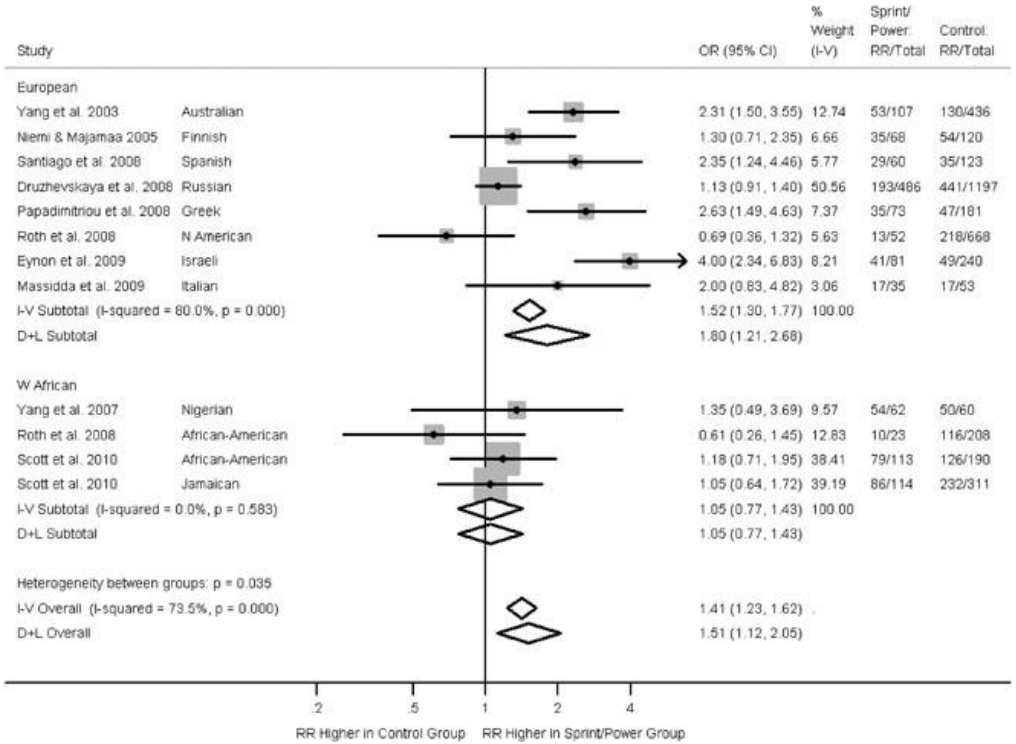
Data



Models


Dominant
Additive
Recessive
Phenotype effect
XX
RX
RR
XX
RX
RR
XX
RX
RR

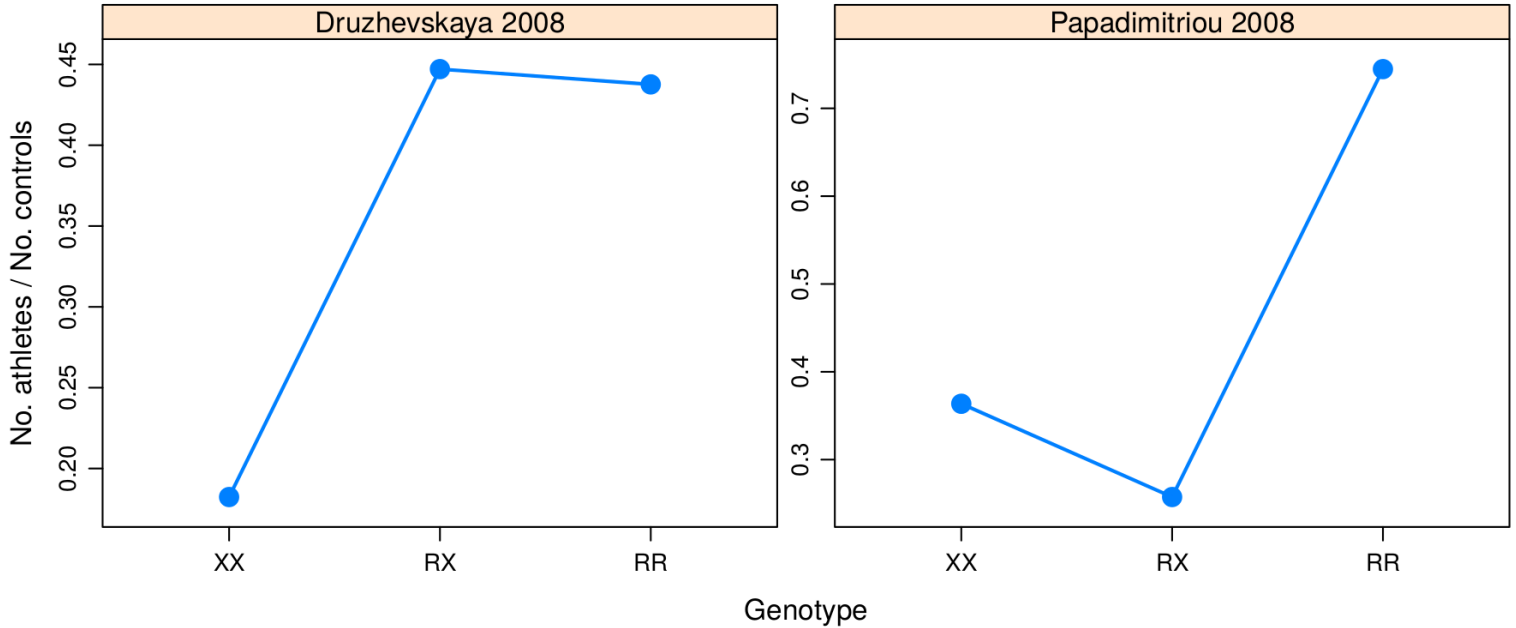
Previous meta-analysis



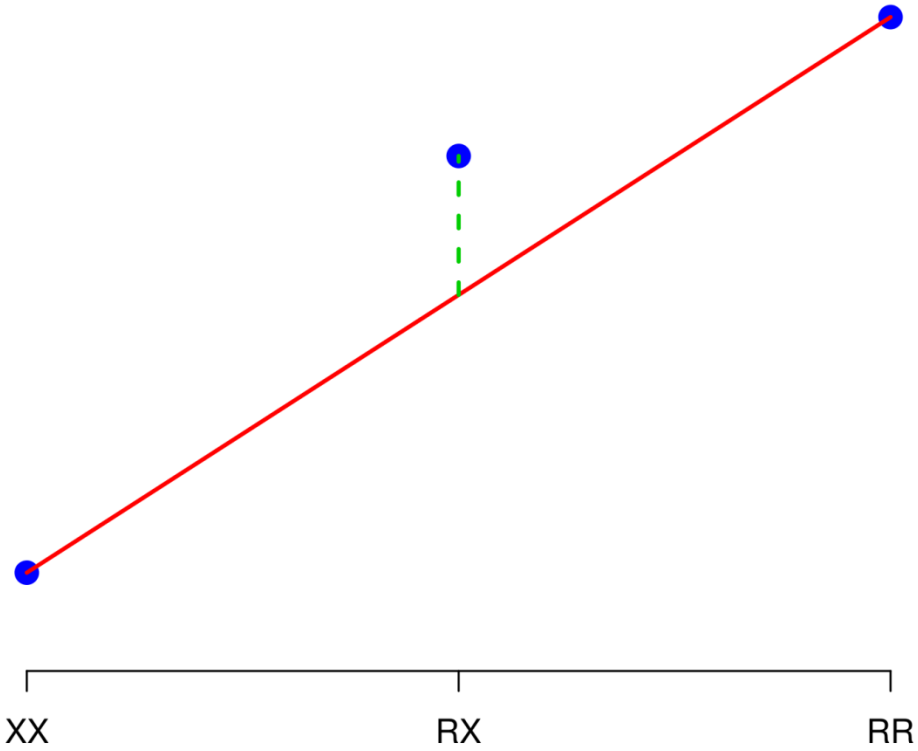
Assumed a **recessive** model

Alfred et al. 2011

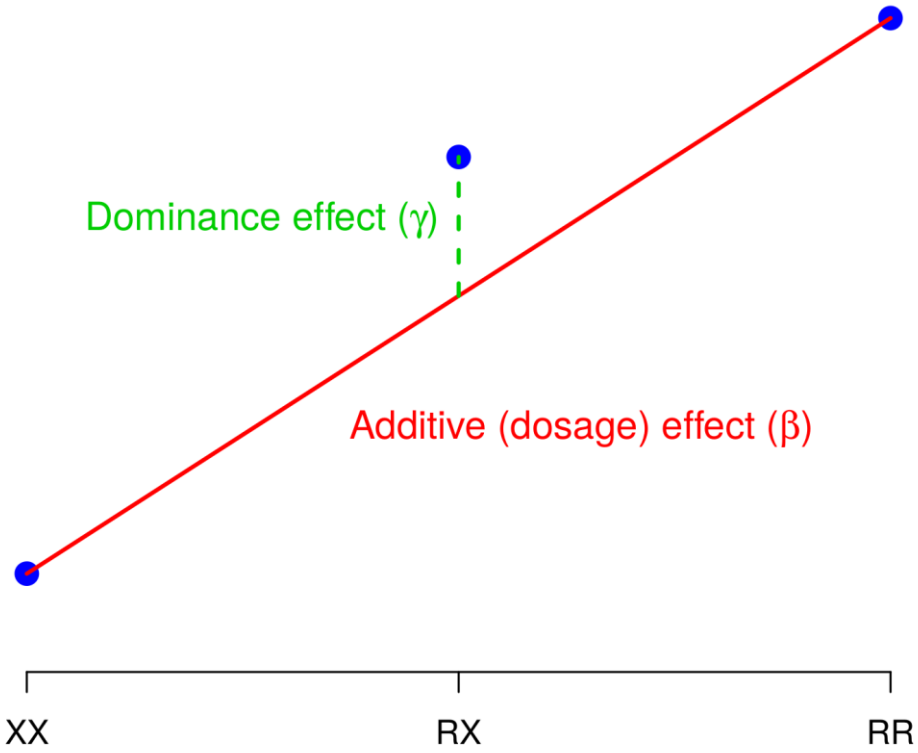
Diverse genetic effects



General model



General model



General model

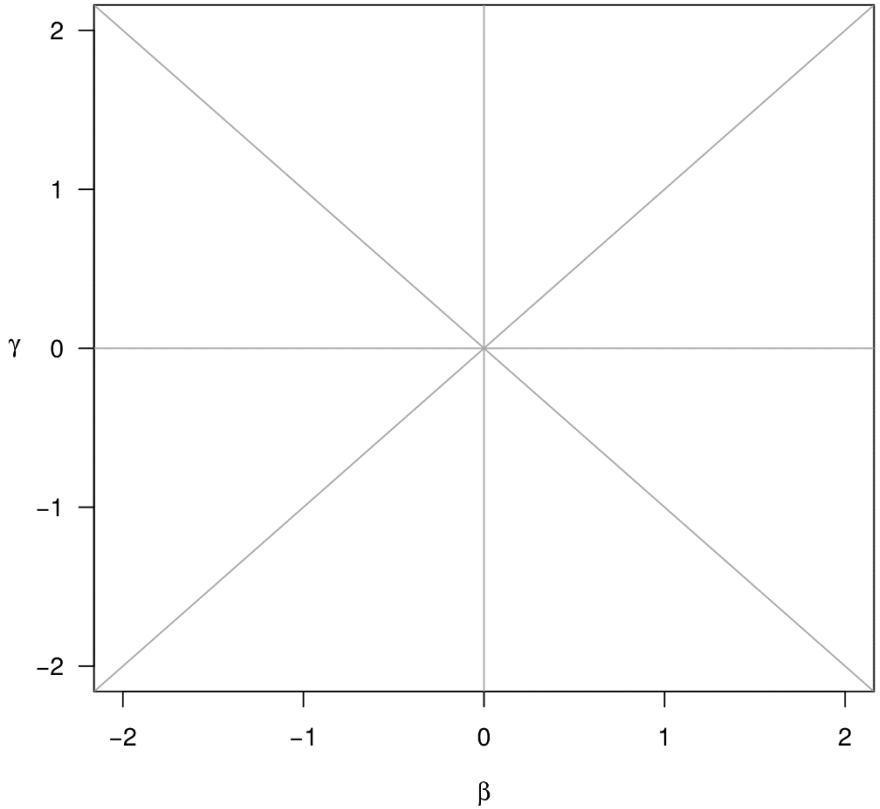
Study i, individual j, genotype G_{ij}

$$\log \frac{\Pr(\text{athlete}|G_{ij})}{\Pr(\text{control}|G_{ij})} = \mu_i + \beta_i G_{ij} + \gamma_i \, \mathrm{I}(G_{ij} = 1)$$

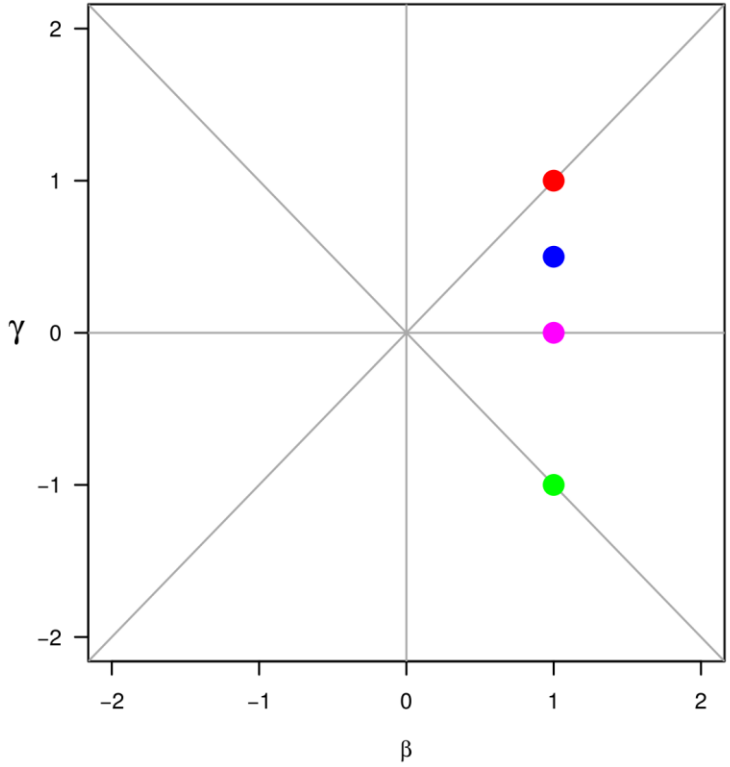
$$\begin{bmatrix} \beta_i \\ \gamma_i \end{bmatrix} \sim N\left(\begin{bmatrix} \beta \\ \gamma \end{bmatrix}, \begin{bmatrix} \tau_\beta^2 & \rho\tau_\beta\tau_\gamma \\ \rho\tau_\beta\tau_\gamma & \tau_\gamma^2 \end{bmatrix} \right)$$

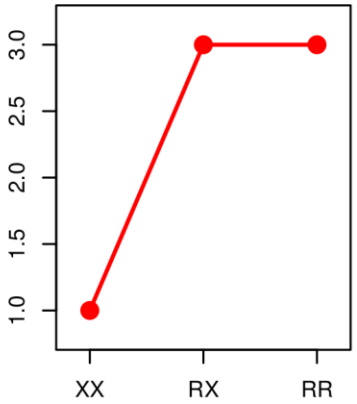
Use ‘default’ weakly informative priors

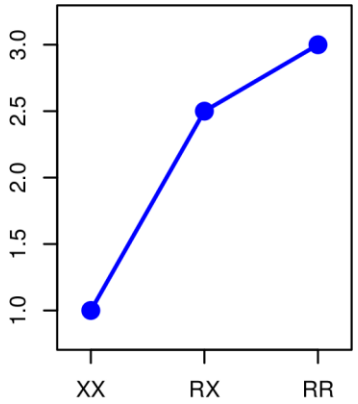
Model space plot

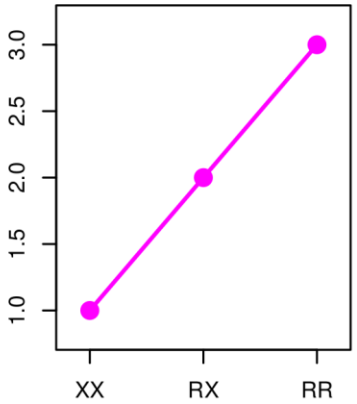


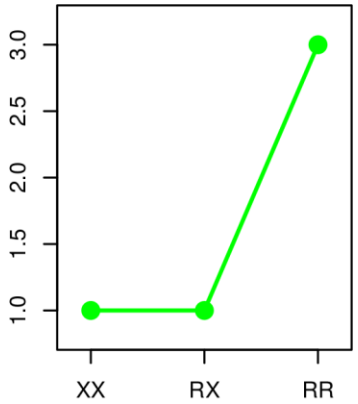
Model space plot



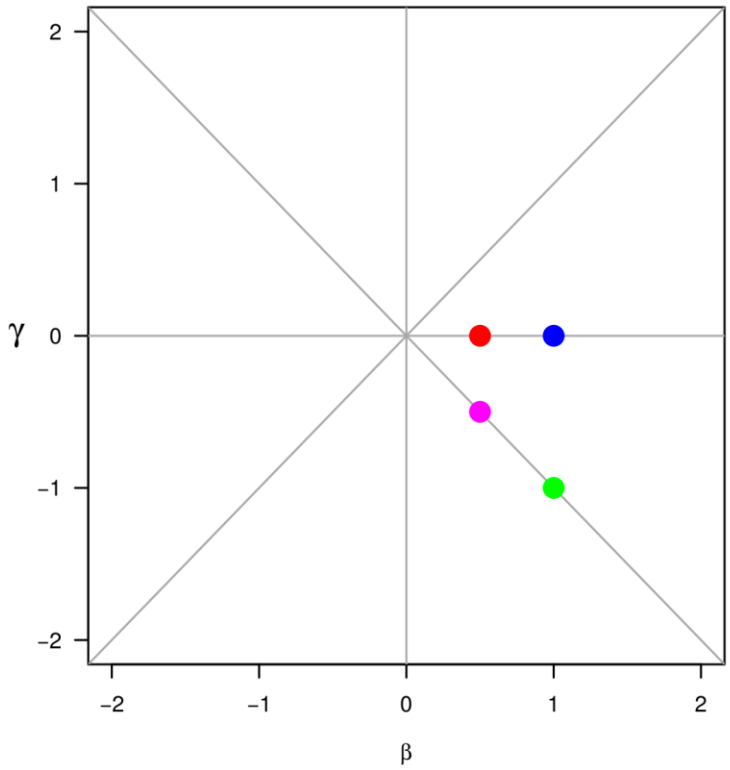


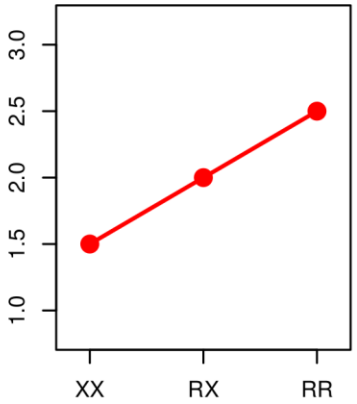


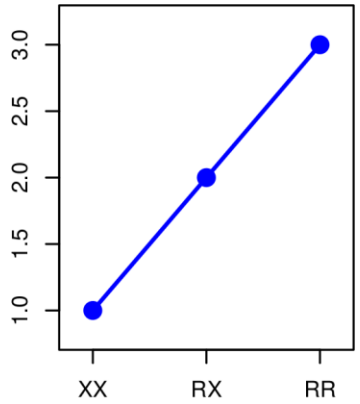


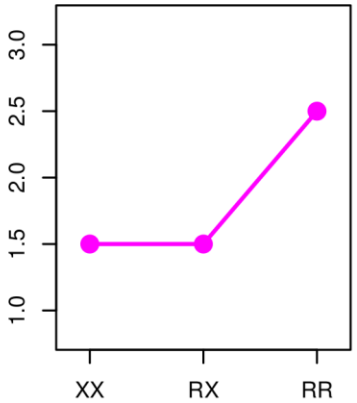


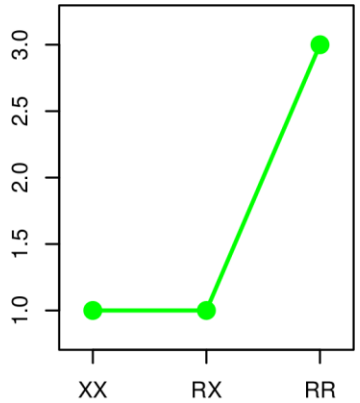
Model space plot



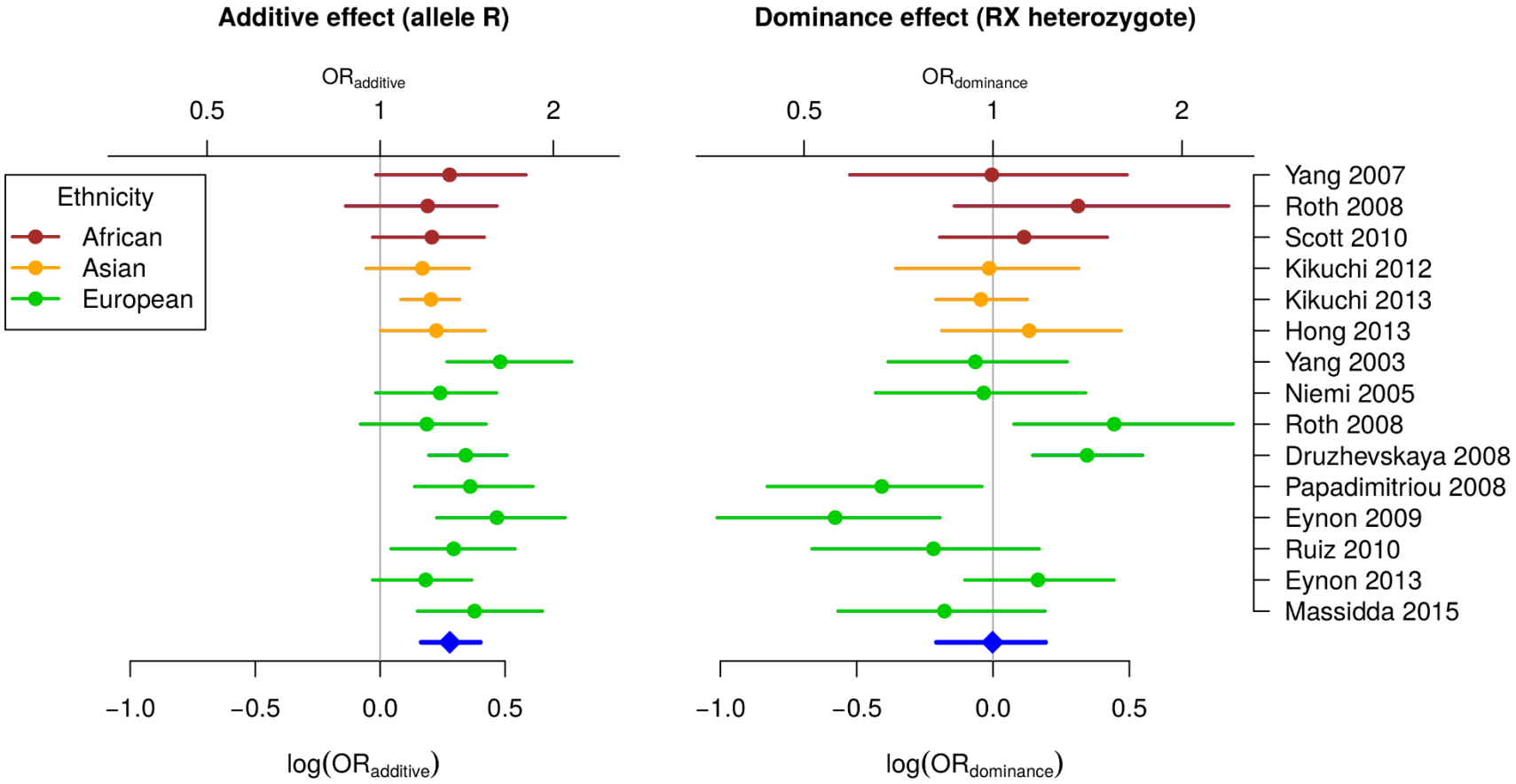




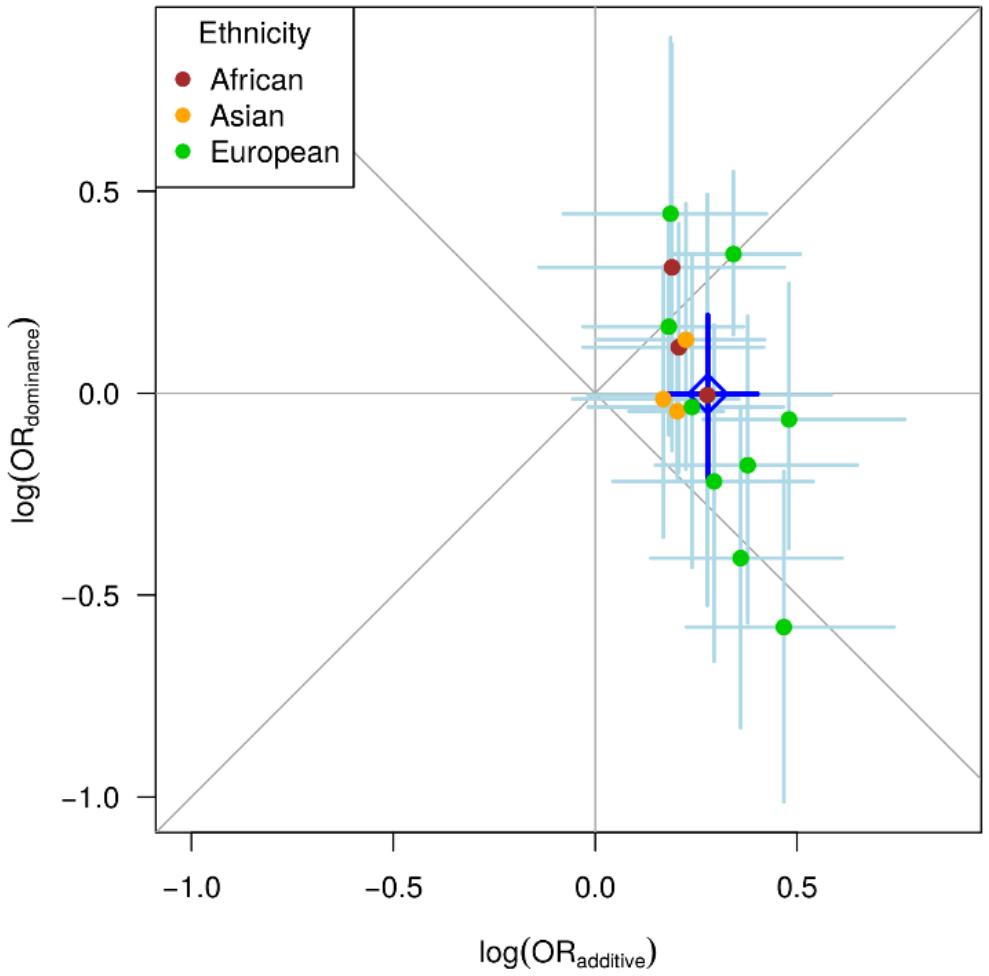




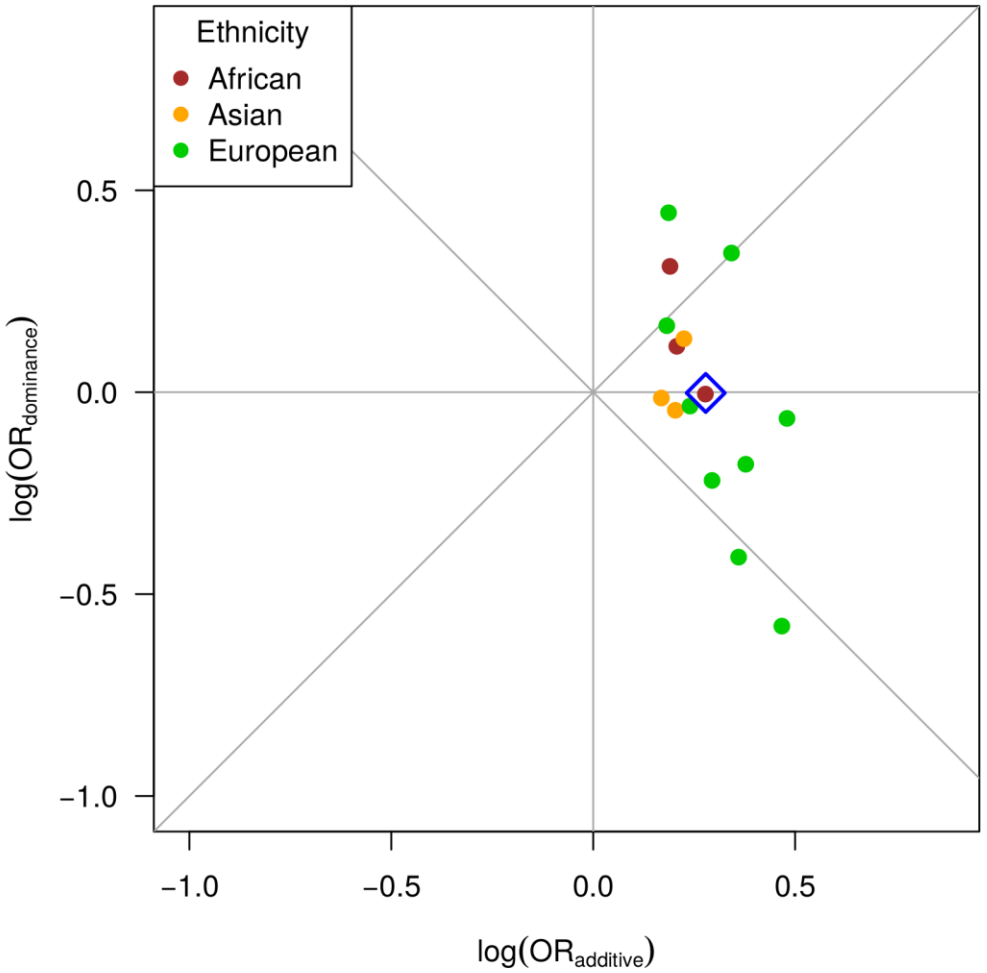
Results


Additive effect (allele R)
Dominance effect (RX heterozygote)
OR additive
OR dominance
0.5
1
2
Ethnicity
African
Asian
European
Yang 2007
Roth 2008
Scott 2010
Kikuchi 2012
Kikuchi 2013
Hong 2013
Yang 2003
Niemi 2005
Roth 2008
Druzhevskaya 2008
Papadimitriou 2008
Eynon 2009
Ruiz 2010
Eynon 2013
Massidda 2015
−1.0
−0.5
0.0
0.5
log(OR additive)
log(OR dominance)

Results


Ethnicity
African
Asian
European
0.5
0.0
−0.5
−1.0
log(OR_dominance)
−1.0
−0.5
0.0
0.5
log(OR_additive)

Results


Ethnicity
African
Asian
European
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
log(OR_dominance)
log(OR_additive)

Results

Overall mean genetic effect

$$OR_{add} = e^{\hat{\beta}} = \mathbf{1.3}\ (1.2–1.6)$$

$$OR_{dom} = e^{\hat{\gamma}} = \mathbf{1.0}\ (0.76–1.3)$$

Heterogeneity of effects

$$\hat{\tau}_{\beta} = \mathbf{0.17}\ (0.02–0.36)$$

$$\hat{\tau}_{\gamma} = \mathbf{0.44}\ (0.21–0.77)$$

Summary (Part 1)

- Clear evidence of an association (recapitulates main conclusion from past studies)
- Large heterogeneity of effects, no **simple** genetic model fits the data

- **Additive** component **relatively consistent** across studies
- **Dominance** component (heterozygote effect) **highly heterogeneous**, especially for **Europeans**

- Why the heterogeneity?
- Are the covariates useful?

Part 2

Simpler (misspecified) models
Covariates
Some properties of the model
Questions for the audience

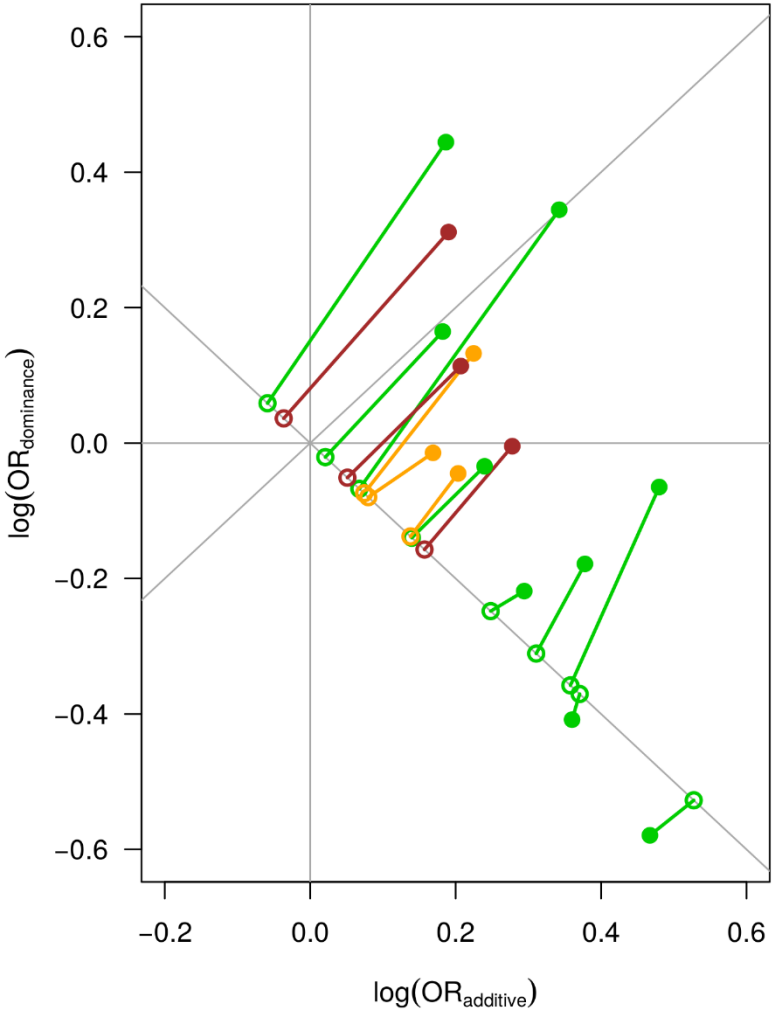
Assume a recessive model

Overall mean genetic effect

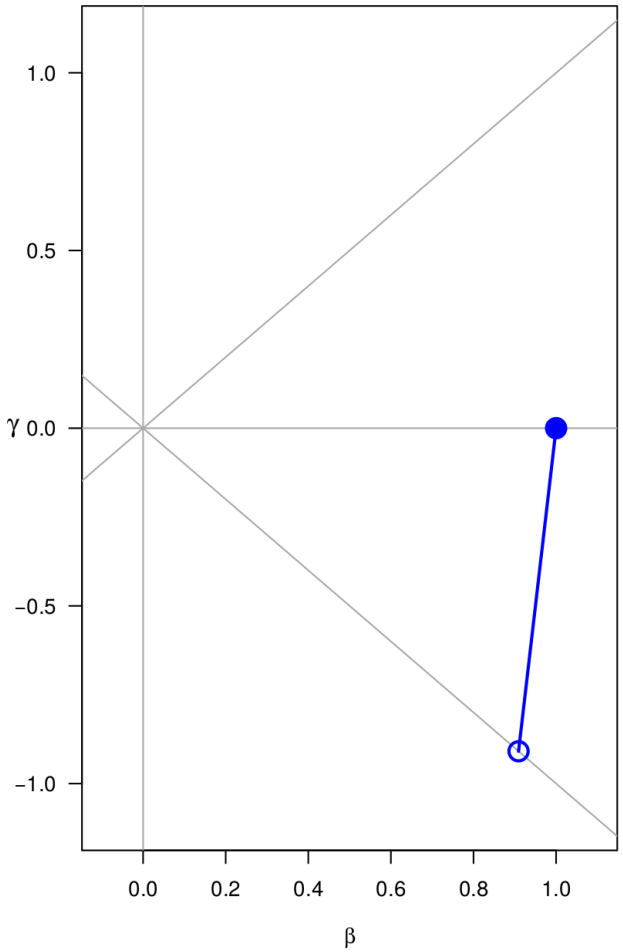
$\sqrt{OR_{RR}}$ = **1.2** (1.0–1.4)

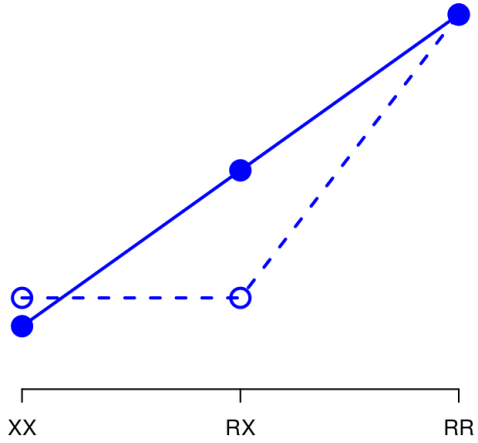
Heterogeneity of effects

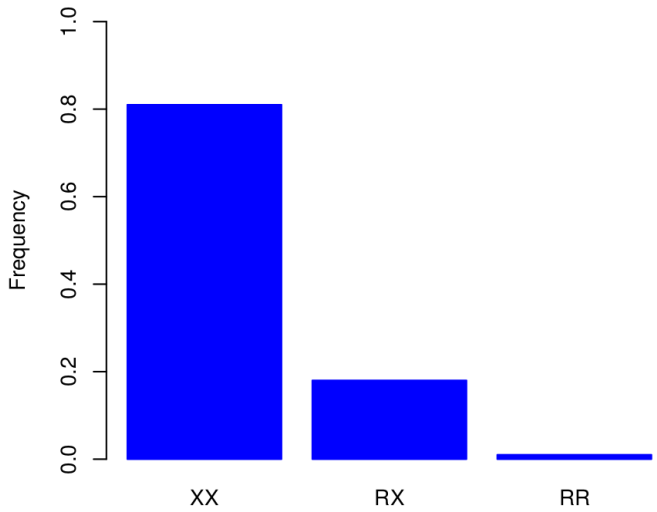
$\hat{\tau}$ = **0.23** (0.11–0.39)

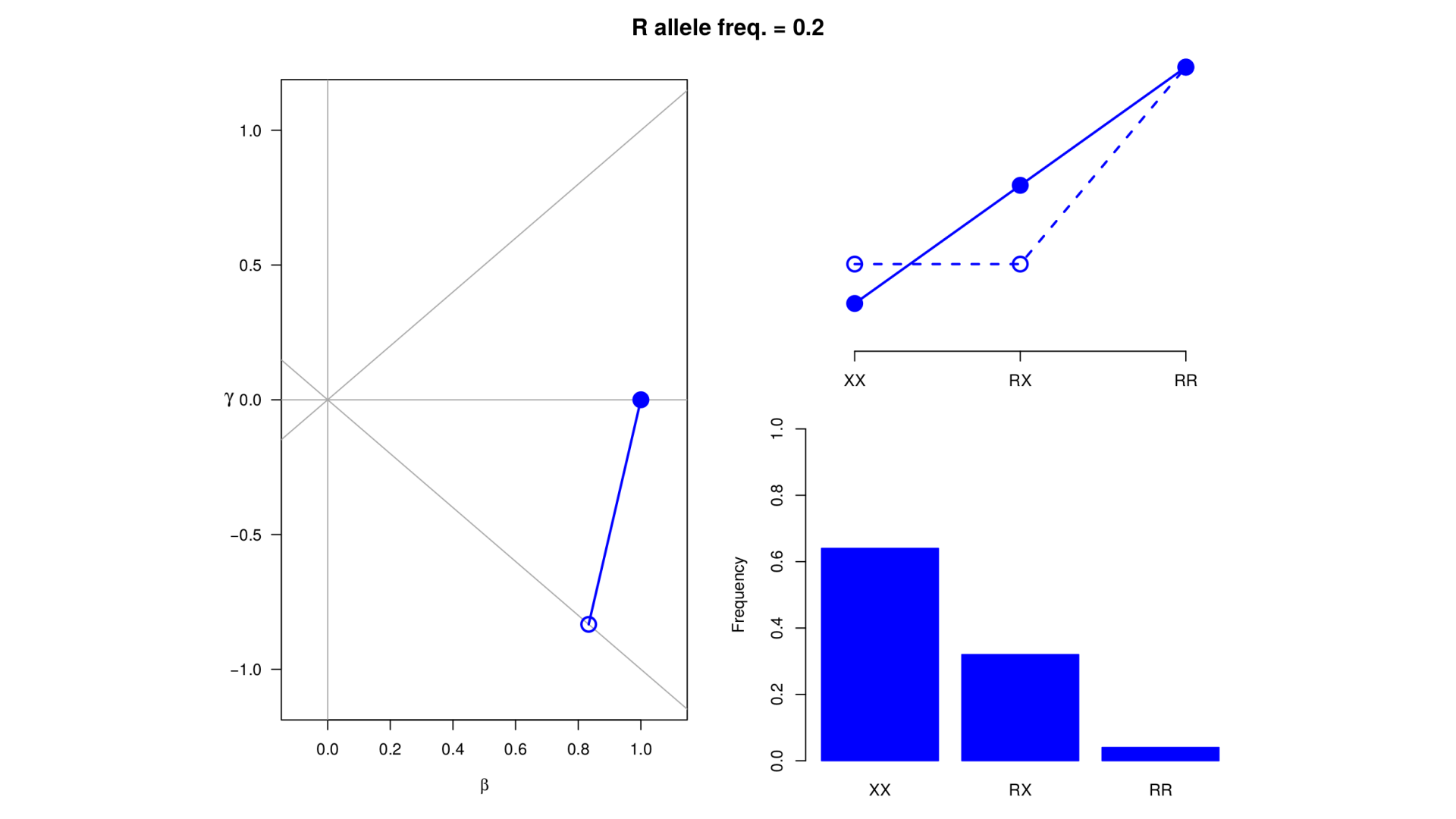


R allele freq. = 0.1

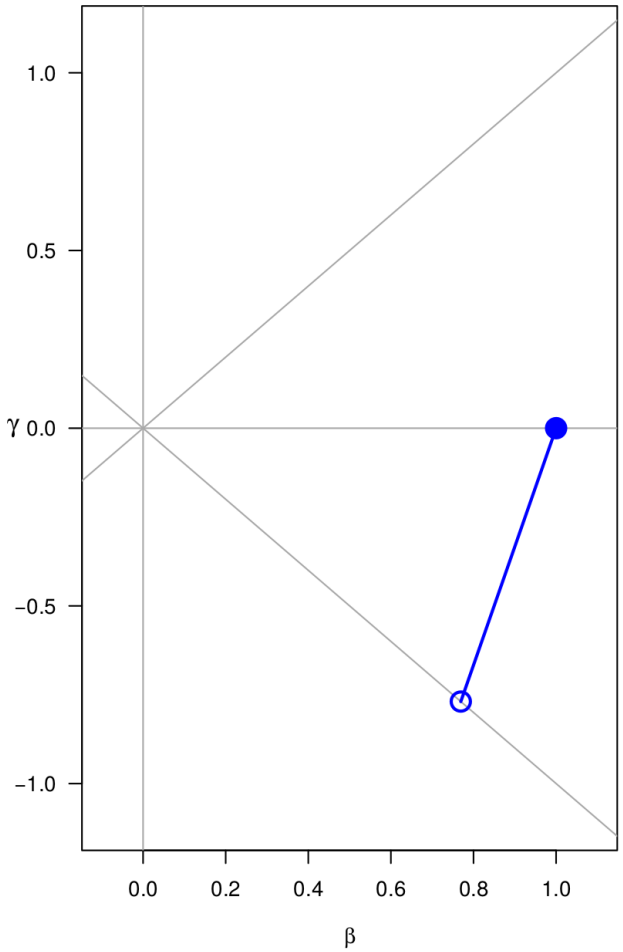


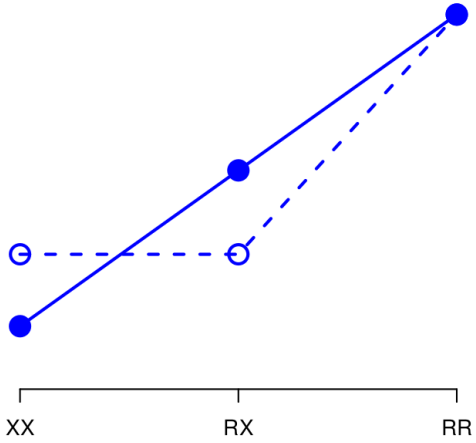


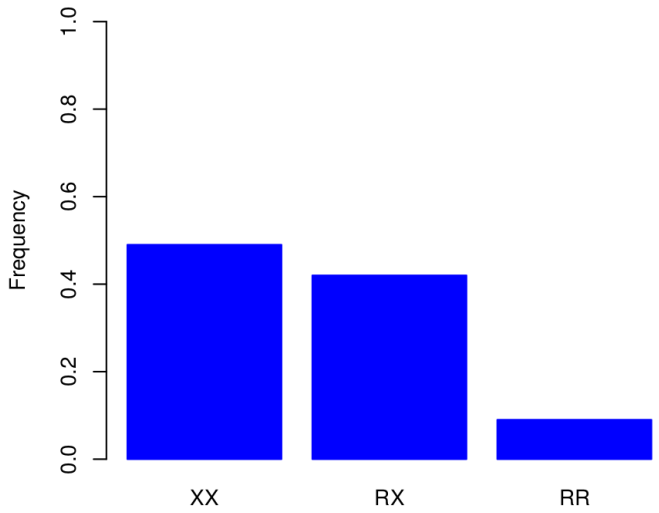



R allele freq. = 0.2
1.0
0.5
γ 0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
β
XX
RX
RR
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
XX
RX
RR

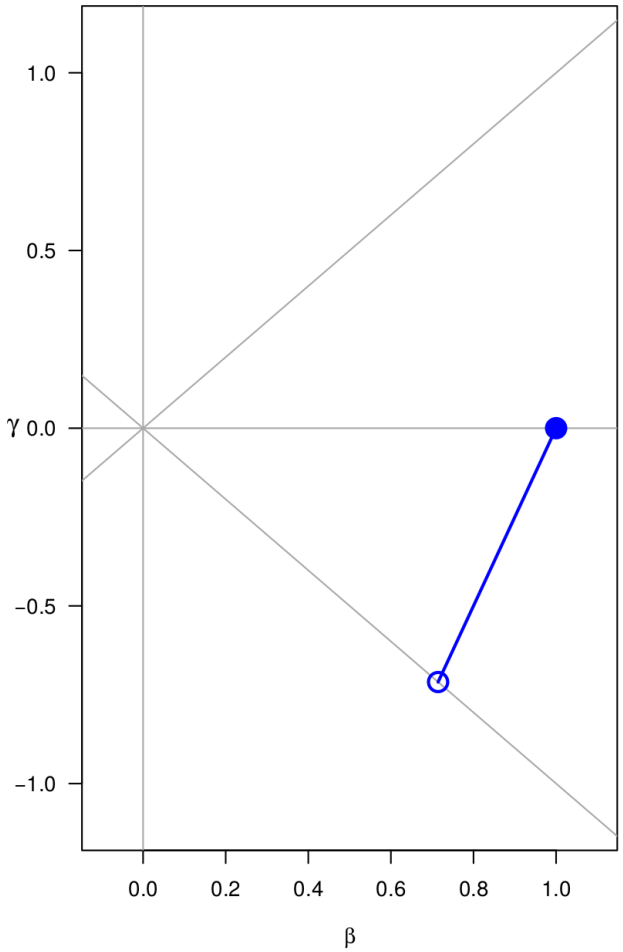
R allele freq. = 0.3

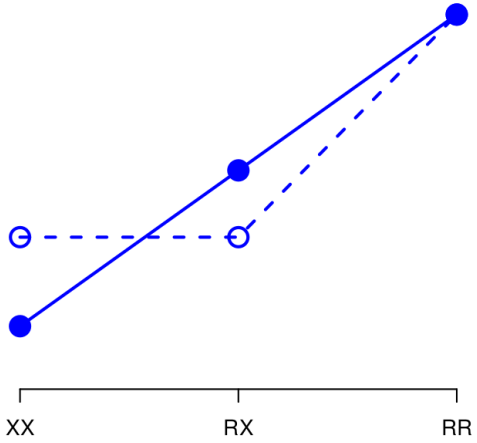


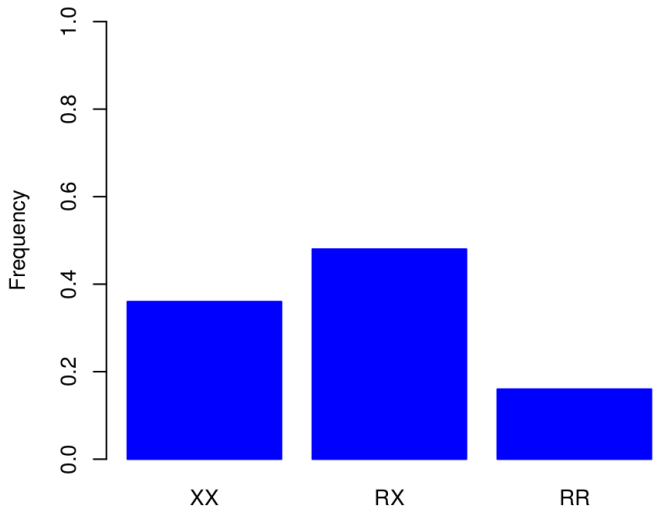




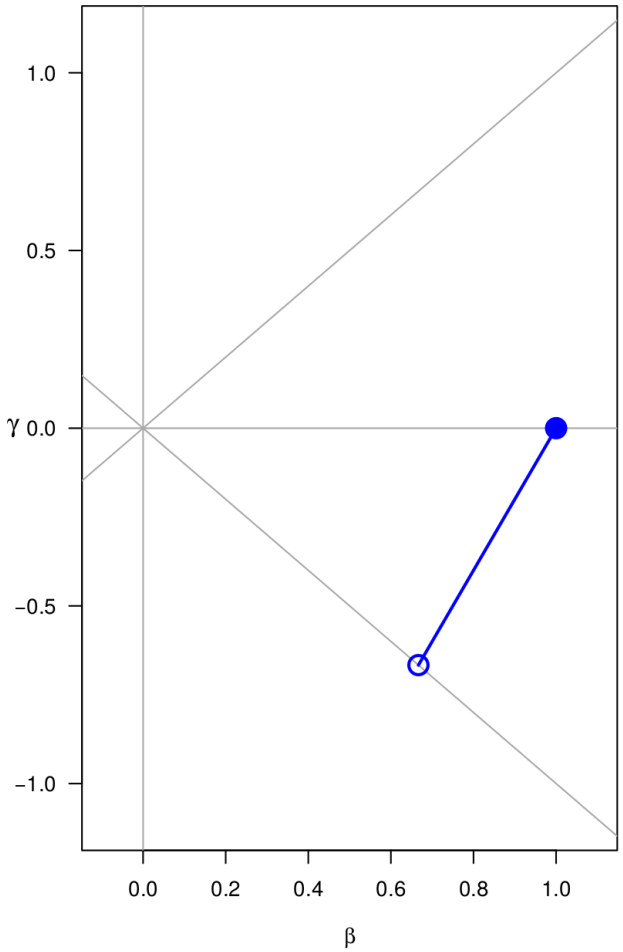
R allele freq. = 0.4

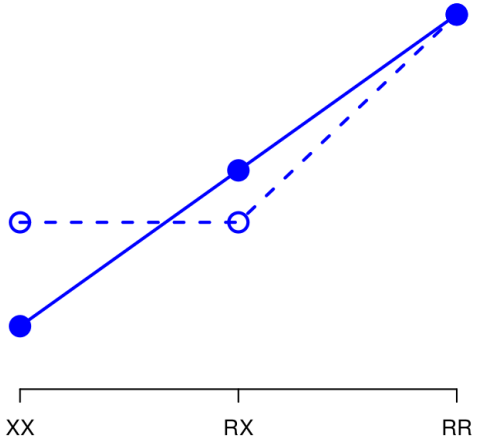


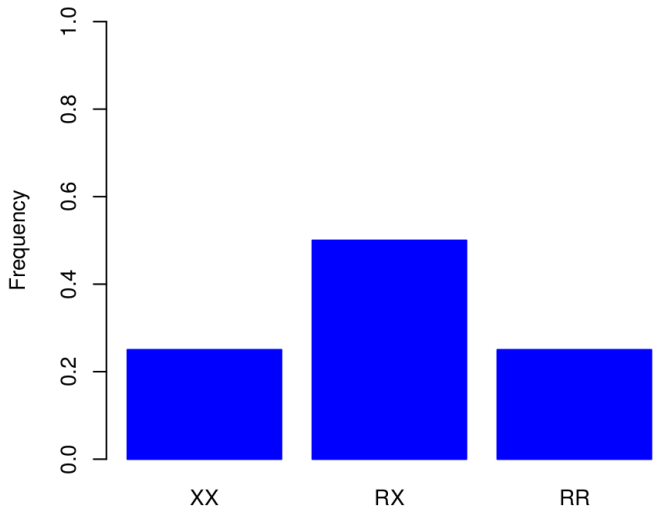




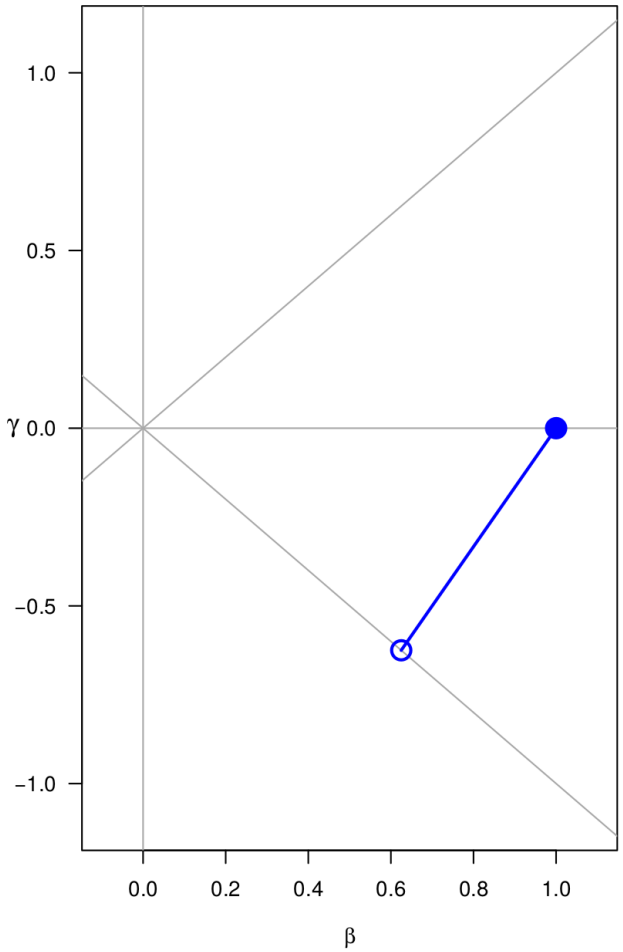
R allele freq. = 0.5

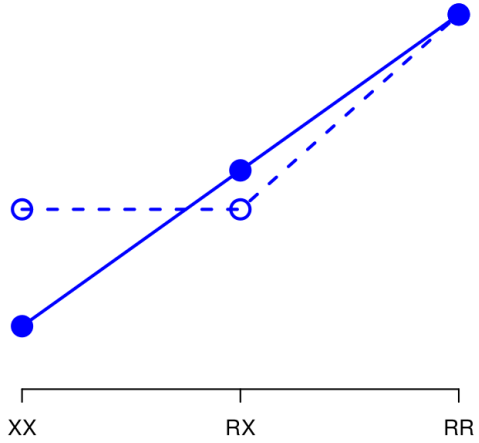


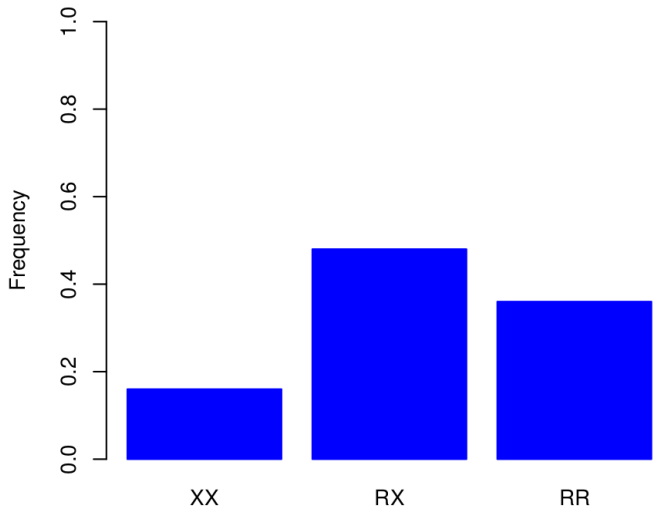




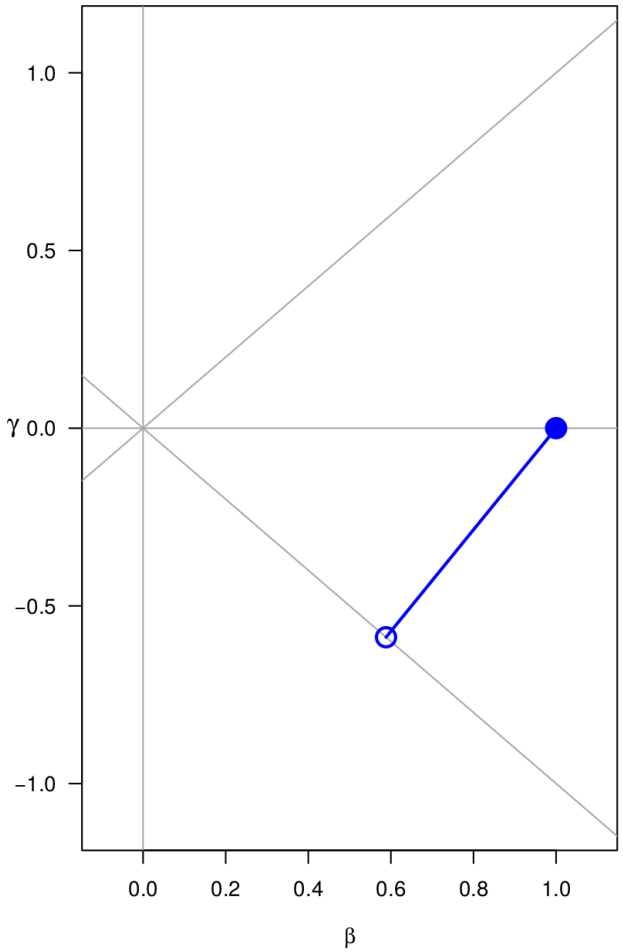
R allele freq. = 0.6

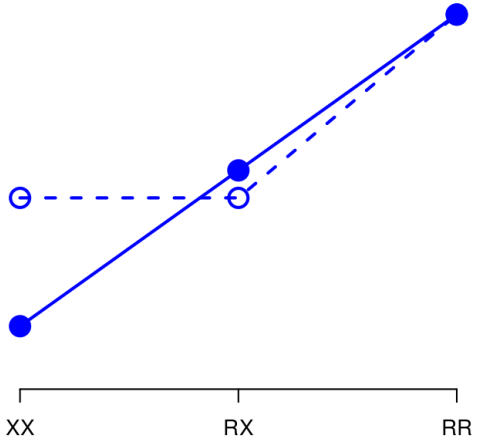


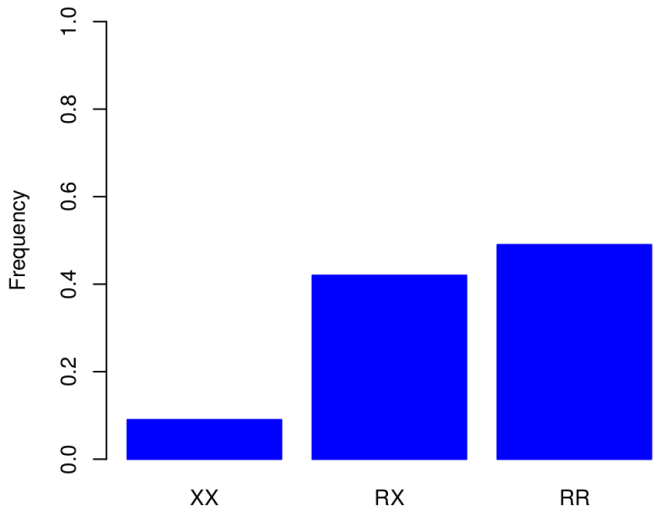




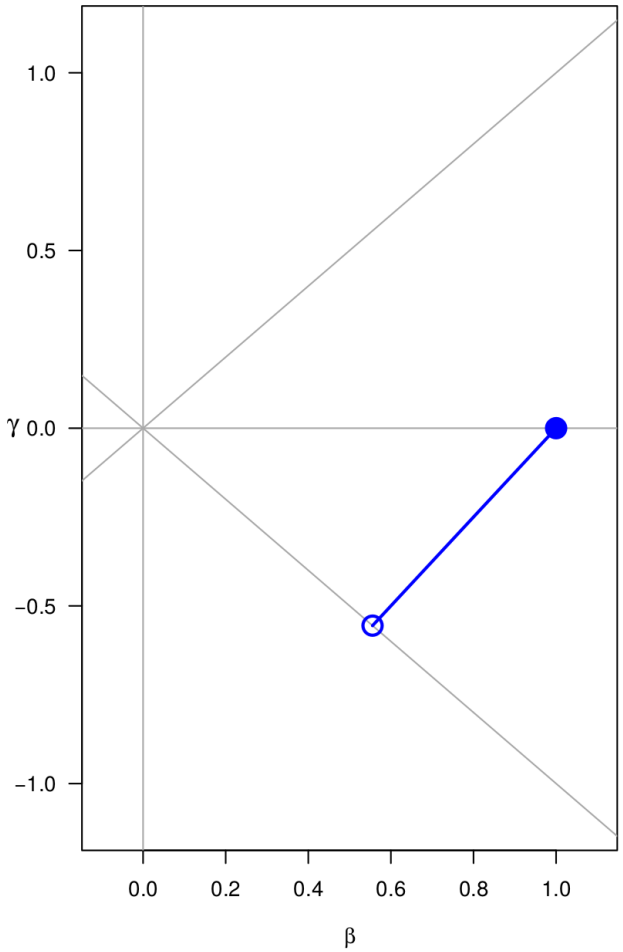
R allele freq. = 0.7

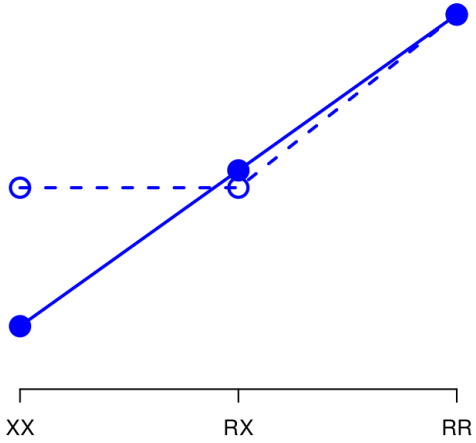


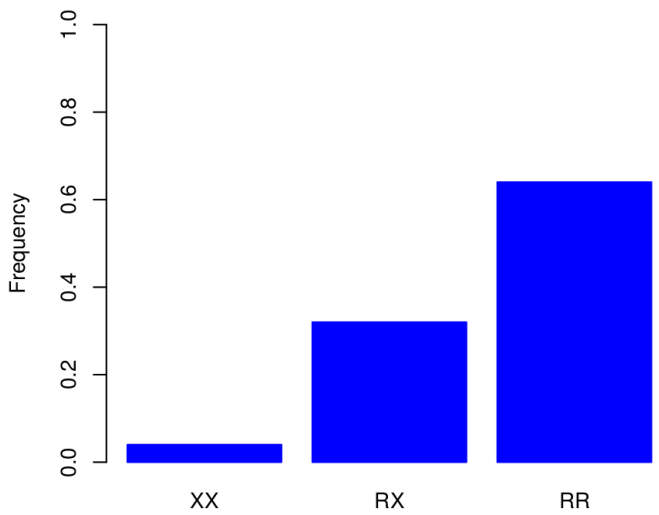




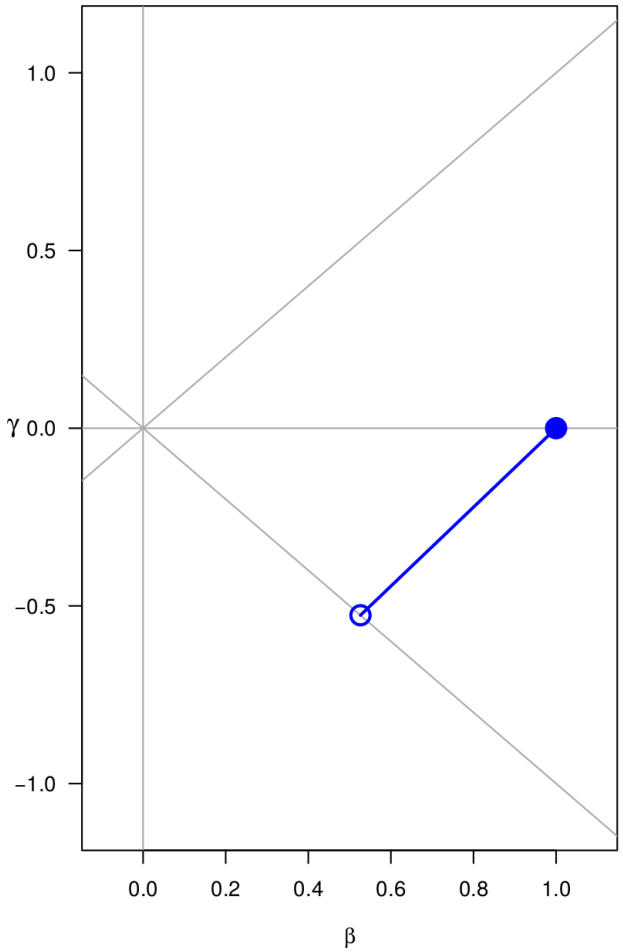
R allele freq. = 0.8

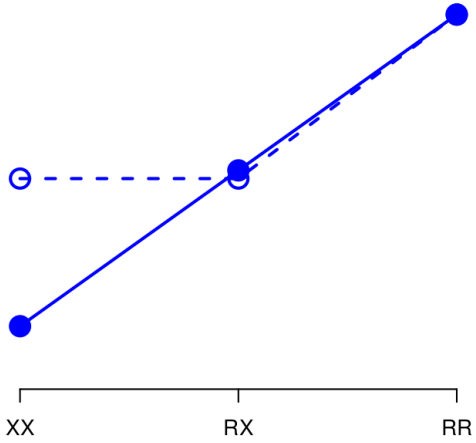


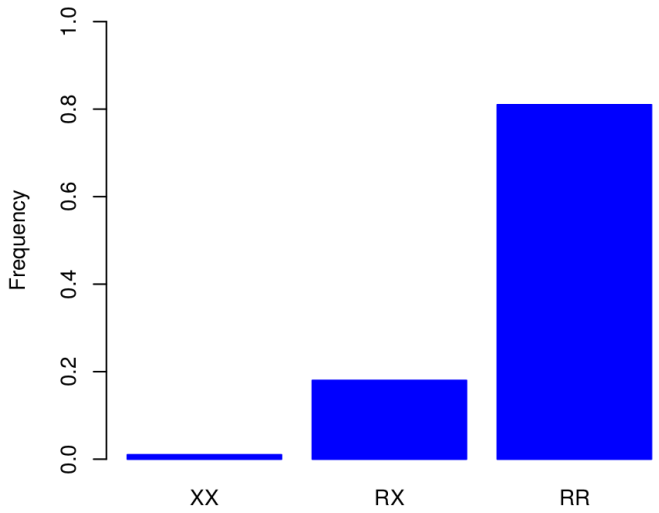




R allele freq. = 0.9







Using covariates

Covariates

1. Ethnicity
2. Sex
3. Competition level (international/national)
4. Sport (i.e. mix of sports)

Mostly only have per-study summaries

Some data are missing (esp. 2)

Some covariates only defined for athletes (3 & 4)

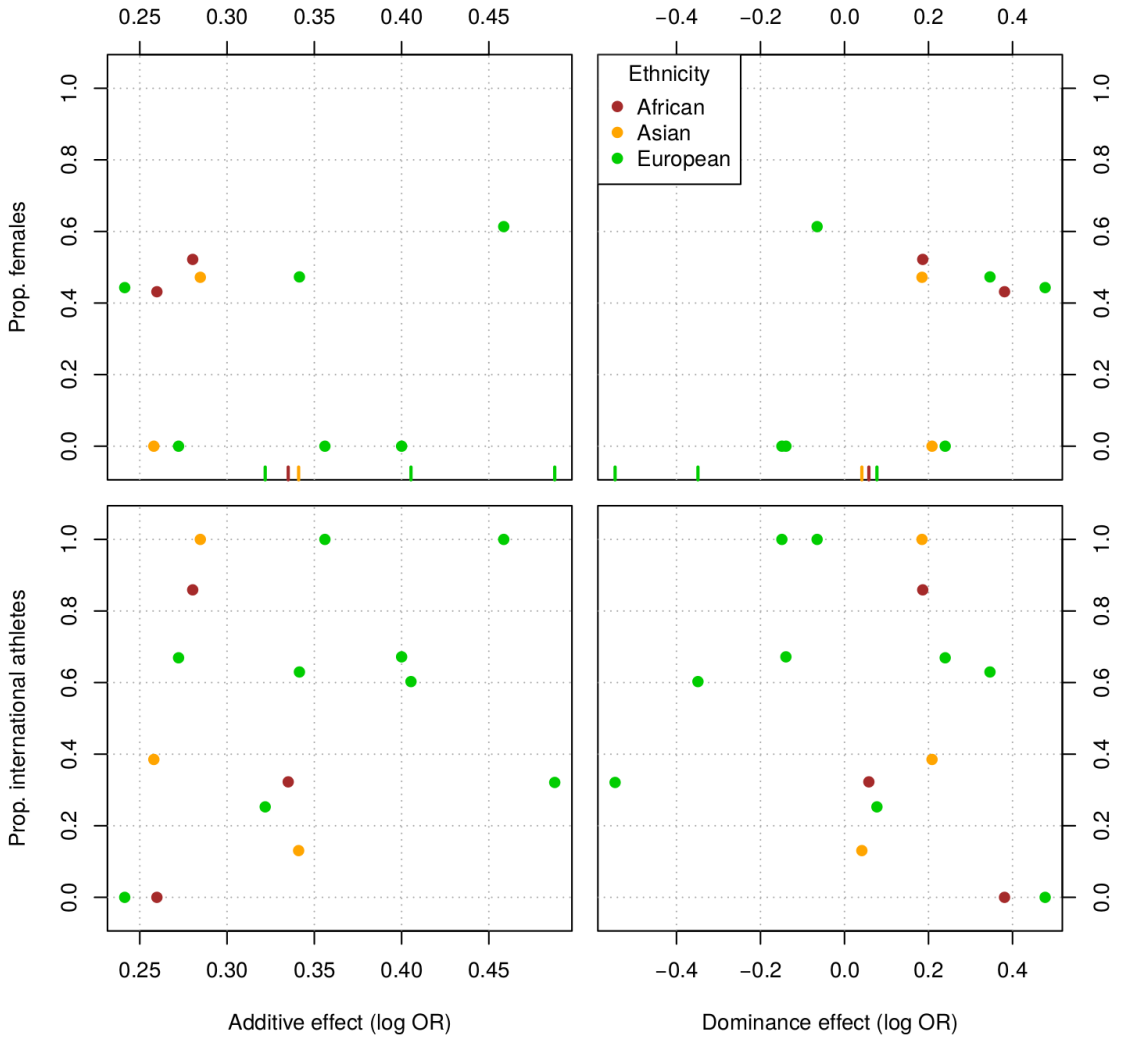
Questions

- Stratify the data?
- Should male & female controls be pooled?
- How to cope with athlete-specific covariates?
- Perhaps multinomial logistic regression? (Seems messy…)
- Need to shift to a retrospective likelihood?
- Currently, I do something hacky…

Comparison against covariates

An ‘informal assessment’ of the impact of covariates

Haven’t yet looked at sport (covariate 4)



Sport (covariate 4) is messy…

Study reference	Country of origin	Sex	Athletes (number, % international)
Yang et al. 2003	Australia	M&F	Track and field athletes (≤800m) (n=46), swimmers (≤200m) (n=42), judo athletes (n=9), short-distance track cyclists (n=7), and speed skaters (n=3). (n= 107, 100%)
Niemi & Majamaa 2005	Finland	M&F	Sprinters (100-400m) & field athletes (n= 23, international, n=68 national level^)
Papadimitriou et al. 2008	Greece	M&F	Sprinters (100- 400m), jumpers, throwers and decathletes (international n=44, n=29 national)
Eynon et al. 2009	Israel	M&F	Sprinters (100 to 200m) (n= 26, international, n=55 national)
Massidda et al. 2015	Italy	M	Sprinters (n=16), swimmer (n=1), wrestlers (n=17), power lifters (n=11), artistic gymnasts (n=19) (n=64, 67%)
…	…	…	…

Prospective vs retrospective

Prospective likelihood:

$$\log \frac{\Pr(\text{athlete}|G)}{\Pr(\text{control}|G)} = \mu + \beta G + \gamma\, \mathrm{I}(G = 1)$$

Retrospective likelihood:

	$G = 0$	$G = 1$	$G = 2$
$\Pr(G\|\text{control})$	g_0	g_1	g_2
$\Pr(G\|\text{athlete})$	$\frac{g_0}{Z}$	$\frac{g_1 r_1}{Z}$	$\frac{g_2 r_2}{Z}$

- The g_i describe the **genotype distribution for controls** (2 free parameters), replacing μ.
- The r_i are **odds ratios**, naturally parameterised by (β, γ), same as before.
- Z is just a normalisation parameter
- Overall, there is **1 extra parameter**
- Prospective likelihood **implicitly requires pairing** of cases & controls

Retrospective: potential benefits

Would allow the **control cohorts to partially pool**
(via the genotype distribution)

Would allow the **athlete cohorts to be stratified more elegantly**
(the odds ratios refer only to an athlete cohort,
rather than to an athlete/control pair of cohorts)

Is this the best approach?

Can these be achieved with a prospective likelihood?

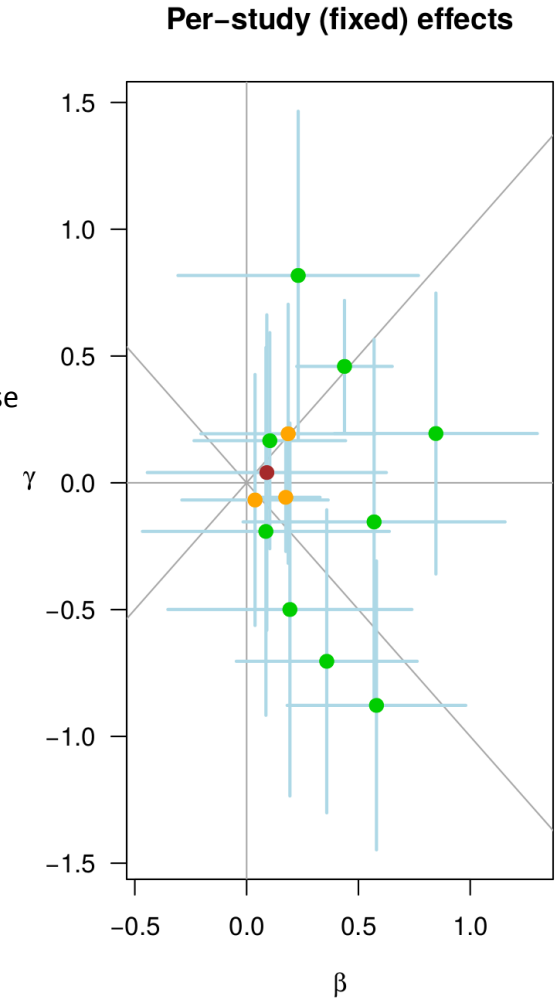
Presentation of results

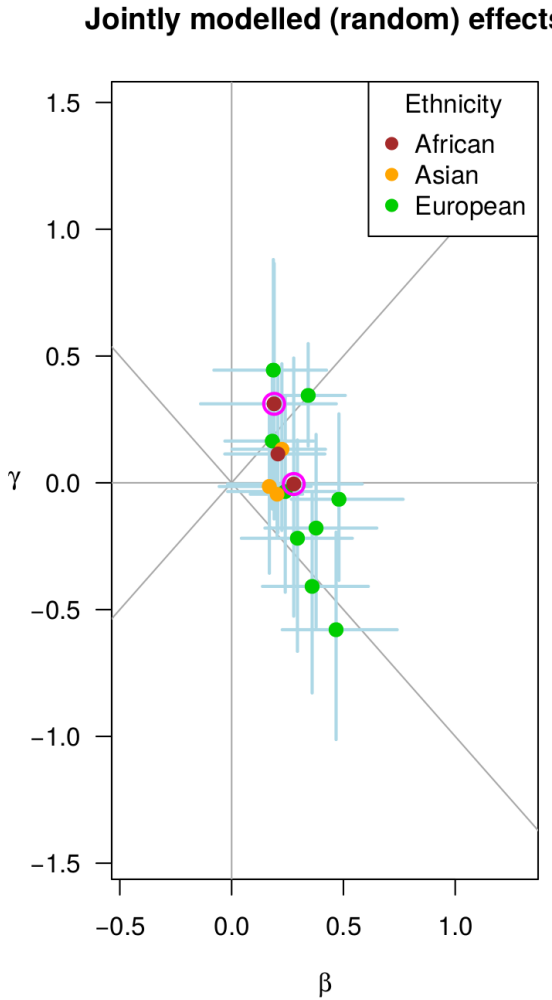
- Main figure is **not** analogous to a forest plot
- Shows the estimates from the **joint model**, rather than per-study models
- Therefore, **shrinkage**!

Shrinkage illustration

Points circled in magenta don't appear in the per-study plot

A general model cannot be fitted for those studies, due to the presence of zero genotype counts

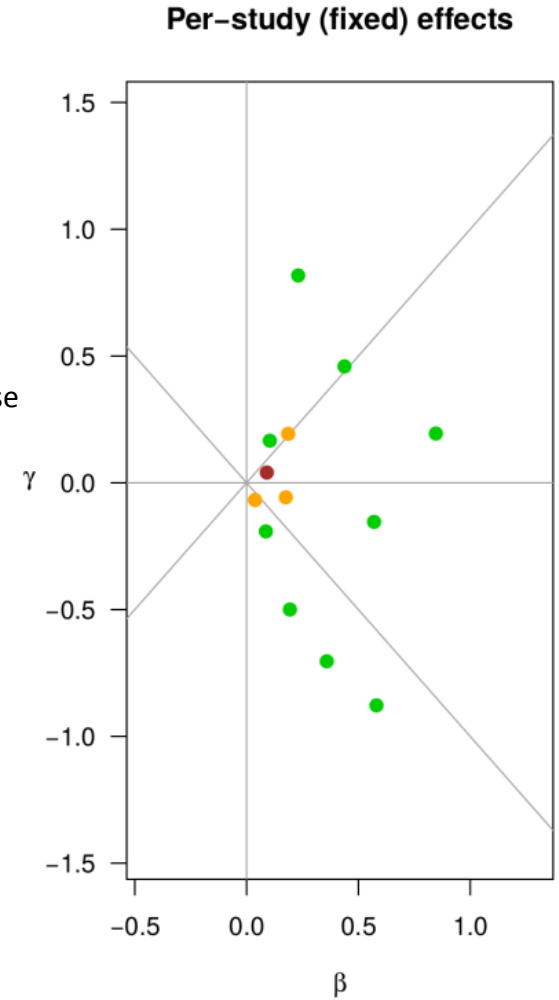


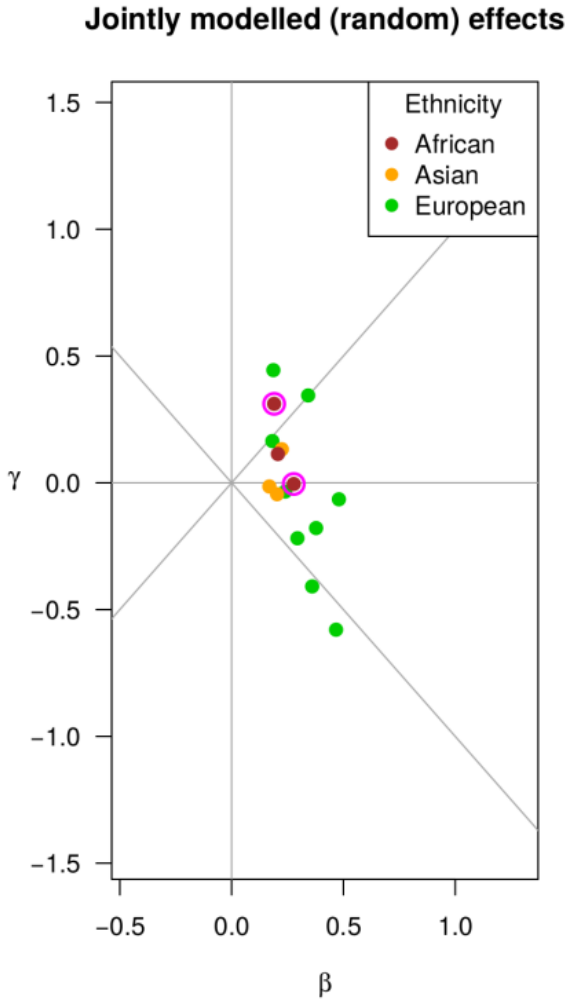


Shrinkage illustration

Points circled in magenta don't appear in the per-study plot

A general model cannot be fitted for those studies, due to the presence of zero genotype counts

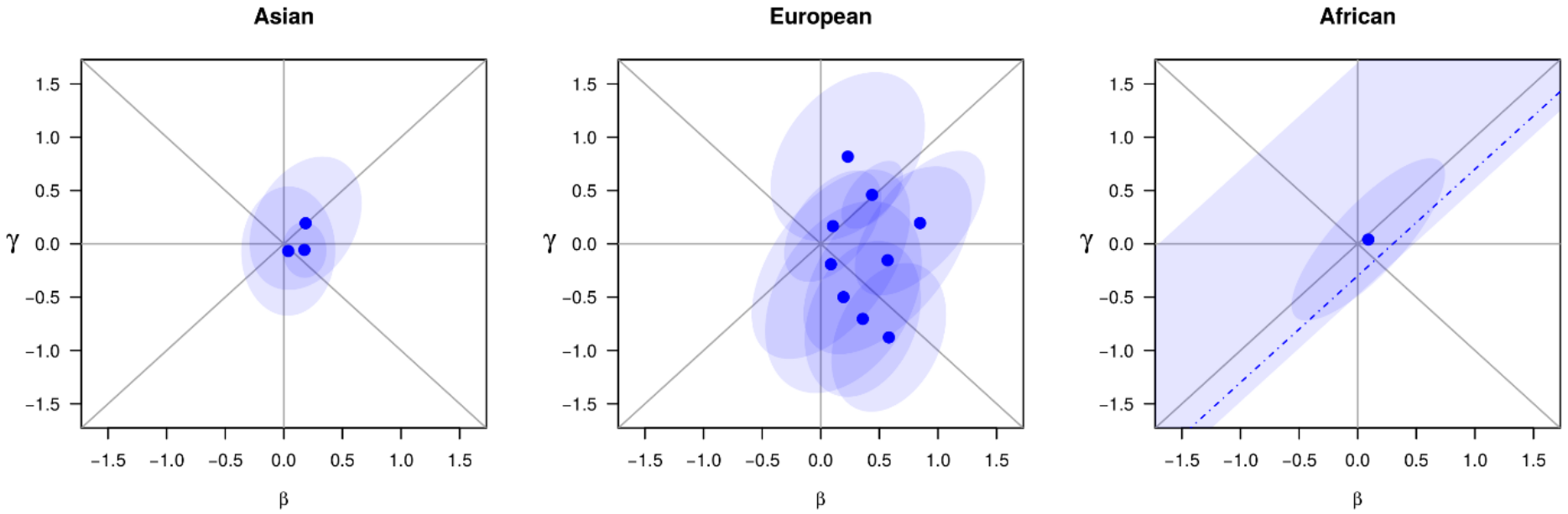




Correlation of effect estimates

- The per-study **estimates are correlated**
- Correlation depends on the **allele frequency**
- Should I depict this? With ellipses? With rotated crosses?

Per-study model fits



Interpretation of results

Any ideas beyond just saying
“there’s substantial heterogeneity in the heterozygote effect”?

Heterogeneity

How should we summarise and represent heterogeneity?

Some ideas:

- Estimate the **variance components**?
  (I did this, but it feels too obscure…)
- Work out a **2D analogue** of the **usual heterogeneity measures** used in standard meta-analyses? (Also seems obscure…)
- Calculate a **posterior distribution** over the **three canonical genetic models** (additive, recessive, dominant)?

Summary (Part 2)

- Use of a **general model** led to **clearer insights** and conclusions about the nature of the evidence in the data
- **Cause of heterogeneity still unclear**, but some ideas still to explore
- Assuming a more **restricted model** can give rise to **spurious heterogeneity**

- Still exploring to best ways to:
  - Visualise and present the results
  - Interpret or investigate the heterogeneity
  - Allow partial pooling beyond the case-control pairing

Not discussed today

- Details of the prior distributions
- Stan programming issues
- Previous work on this or similar problems

Some further work

- Investigate if the type of **athletic events** can explain heterogeneity
- Investigate how to evaluate possible biases (e.g. **funnel plots**)
- **Sensitivity analysis** (to choice of prior)
- Apply to other data: esp. known **GWAS loci** with **highly variable allele frequency** across populations

Acknowledgements

Centre for Systems Genomics

Stephen Leslie

Clinical Epidemiology & Biostatistics

Diana Zannino

Susan Donath

Neuromuscular Research

Fleur Garton (→ Uni. Qld)

Kathryn North

Questions?

…answers??