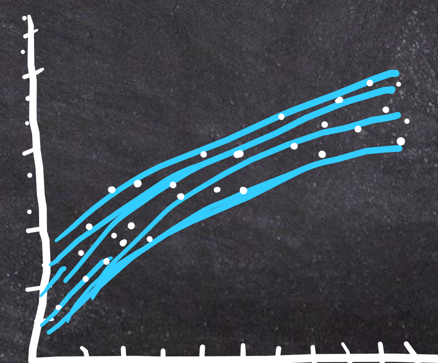
Data Analysis using R

SoSe 2022

Linear mixed models

23.06.2022



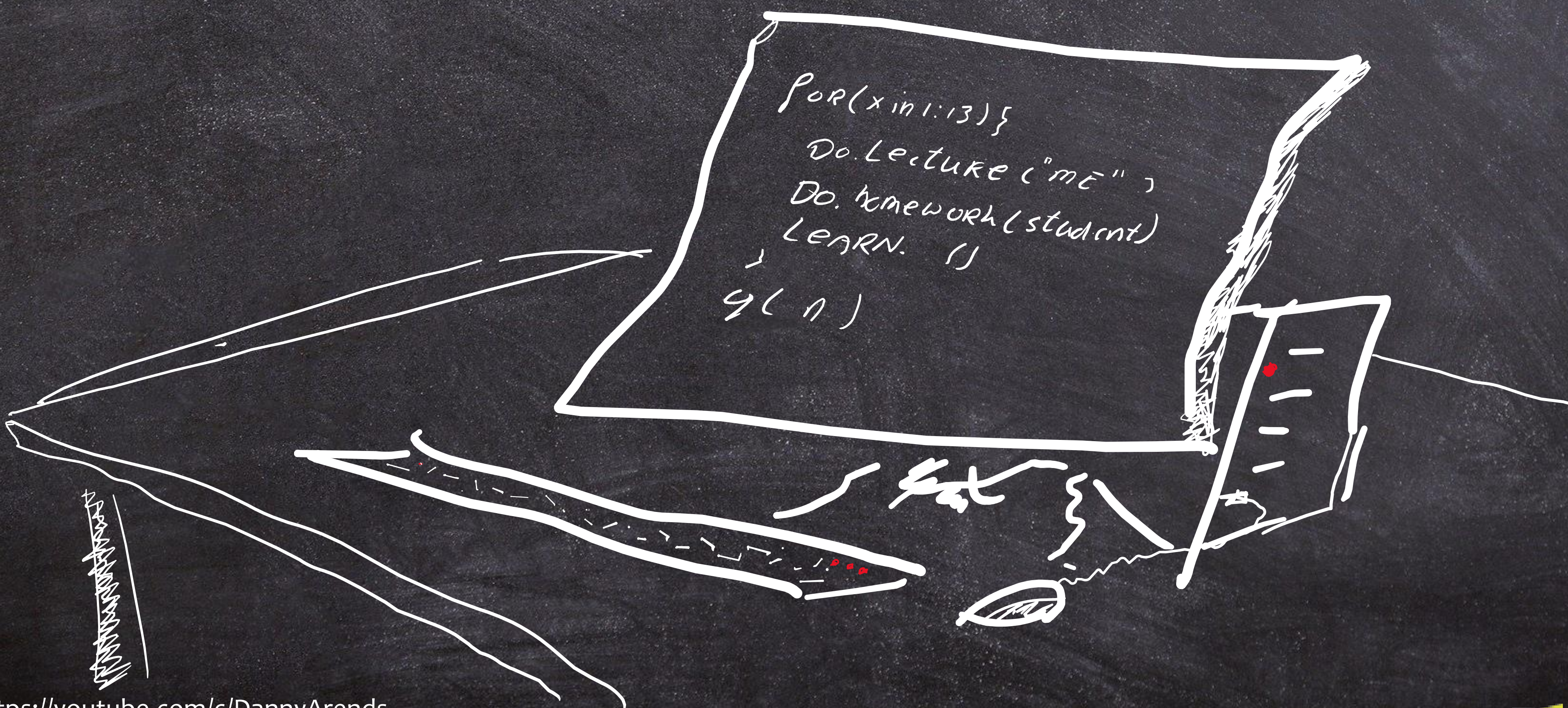
Danny Arends

Fachgebiet Züchtungsbiologie und molekulare Tierzüchtung

Humboldt-Universität zu Berlin

Assignments from last week

Let's take a look at my answers for lecture 7



Exam dates

* **Exam**
  * 28/07/2021 14:00 via Zoom/Moodle

* **Re-Exam**
  * 23/09/2021 14:00 via Zoom/Moodle

SIGN UP via AGNES

Best grade will count

Data Analysis using R

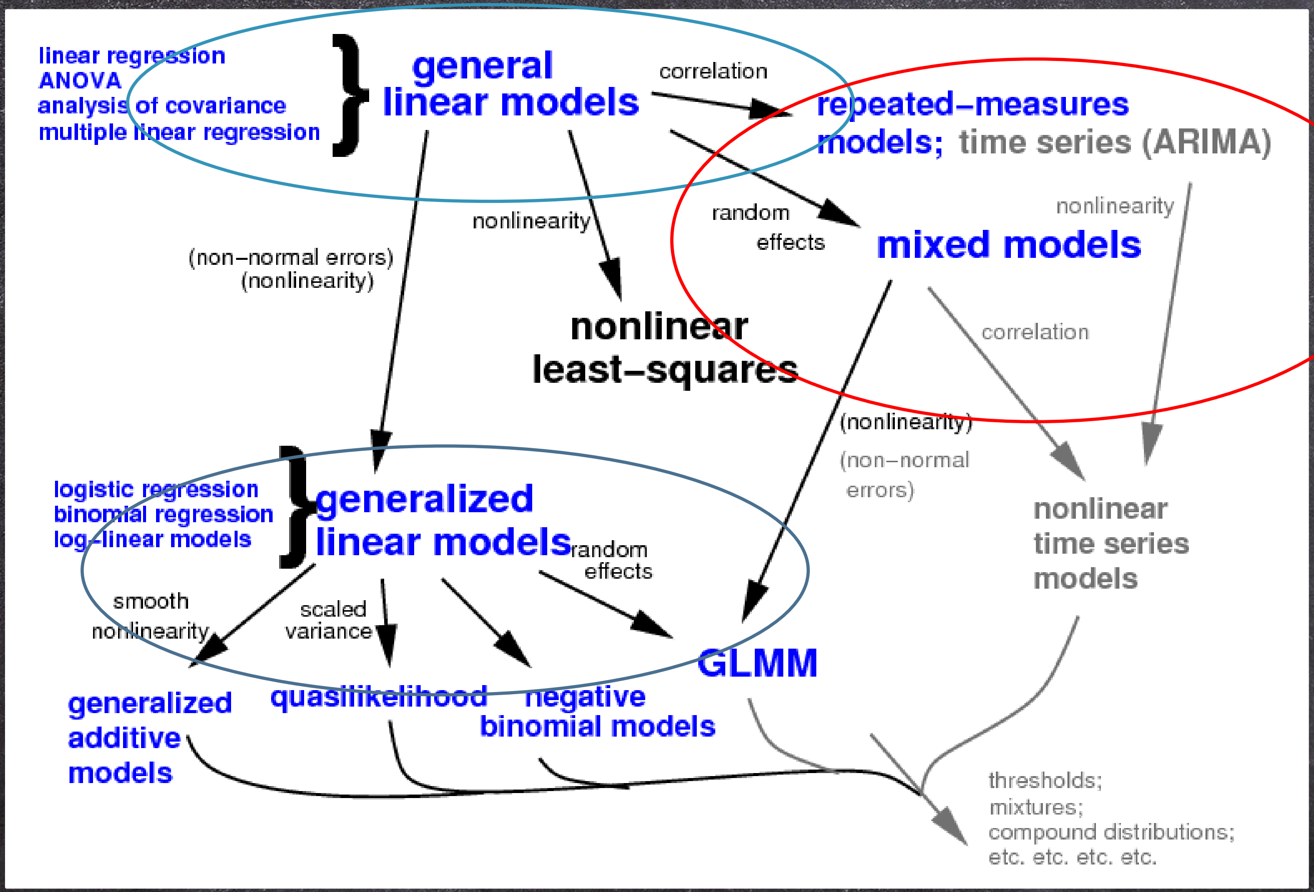
SoSe 2022

Linear mixed models

23.06.2022

Danny Arends
Fachgebiet Züchtungsbiologie und molekulare Tierzüchtung
Humboldt-Universität zu Berlin

Before we start



Today

* Lecture is an adaptation of
  * Introductory tutorial for performing linear mixed effects analyses (Tutorial 2) - Bodo Winter
  * http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf

  Also check out his tutorial on **Linear Models** (tutorial1)

* After the introduction I'll show an example from my current research

Short lecture

* The lecture is short, 22 pages of PDF
  * Compressed into 29 slides
* Read the PDF
  * I will ask questions about it on the exam
* Ask any questions you might have

Linear Mixed Effect Analyse

* Linear mixed effects analyse
  * Random Effects
* How to in R
  * Significance
* Random intercept model
* Random slope model

* Linear mixed model example on the Berlin Fat Mouse

Why ?

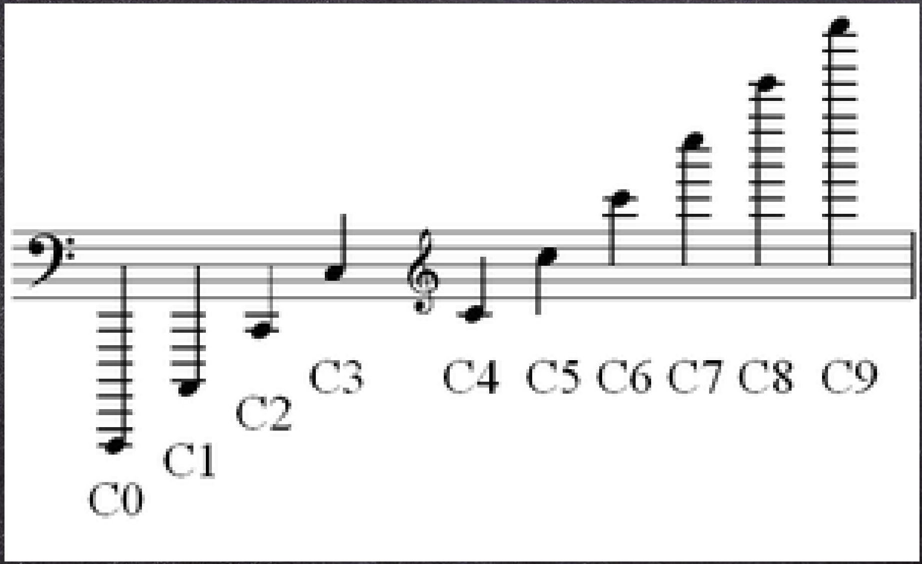
- Measurements are not independent
  - Same individual, over time (time series)
  - Related individuals
- We need to take into account the fact that the number of measurements we have != N
  - If we Don't:
    - Overestimate the statistical power
    - Significant results due to relatedness
      - *Spurious* relationships

Linear models

* Modeling a relationship
  * Response ~ Predictor

* In this tutorial we look at pitch

http://www.bodowinter.com/tutorial/politeness_data.csv



Data structure

- Subject (a person)
- Gender (sex of the person)
- Scenario (question)
- Attitude (polite versus informal)
- Frequency (aka Pitch)

politeness_data.csv

```
subject,gender,scenario,attitude,frequency
F1,F,1,pol,213.3
F1,F,1,inf,204.5
F1,F,2,pol,285.1
F1,F,2,inf,259.7
F1,F,3,pol,203.9
F1,F,3,inf,286.9
```

Most elemental linear model

* The hypothesis of Winter & Grawunder, 2012

frequency ~ attitude + ε

* Attitude - a two level categorical factor:
  * Formal & Informal

Image by Rinto F Rozi from Pixabay

Extending the linear model

* We include sex of the participant

frequency ~ attitude + gender + ε

* Now things get a little more complicated.
  * By design: Multiple measures per subject

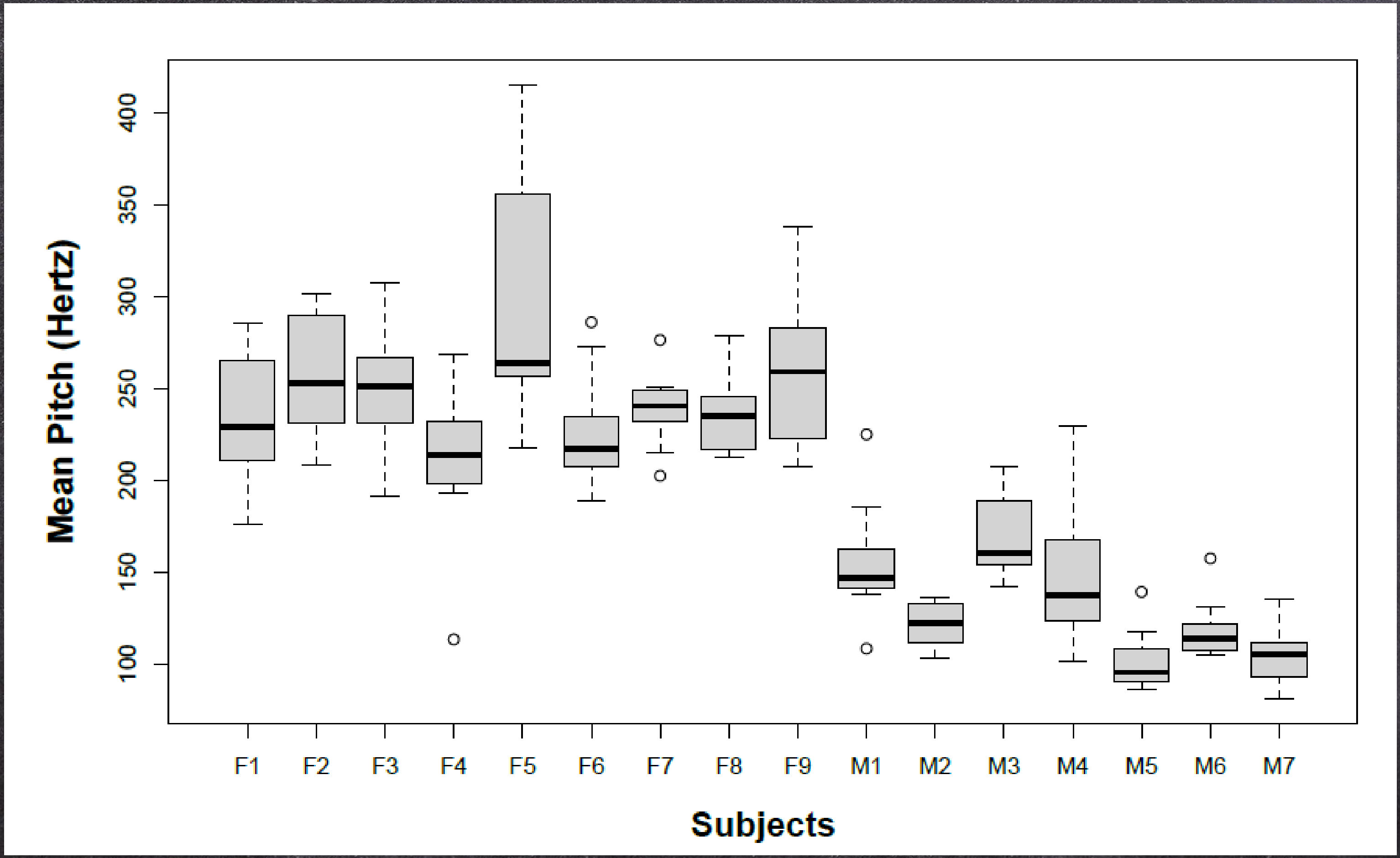
Random effects

Every subject has a slightly different voice pitch, and this is going to be a factor that affects all responses from the same subject, thus rendering these different responses inter-dependent rather than independent.

* So, subject 1 may have a mean voice pitch of 233 Hz across different utterances, and subject 2 may have a mean voice pitch of 210 Hz per subject

Image by photosforyou from Pixabay

Random effects



https://youtube.com/c/DannyArends

Extending the model

* Model individual differences by allowing different ***random intercepts*** for each individual (subject).

frequency ~ attitude + gender + **(1|subject)** + ε

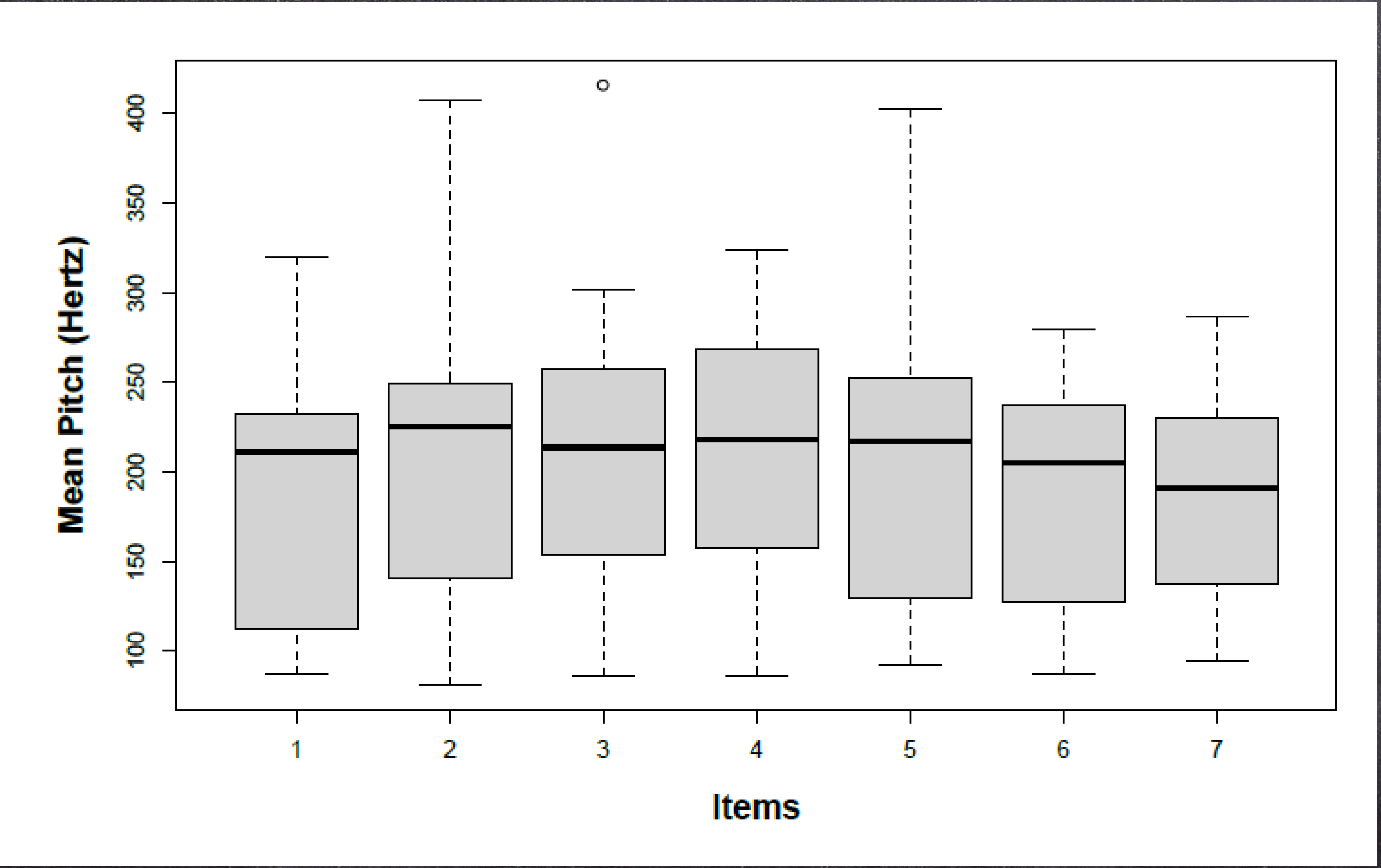
Different questions

* Similar to the case of by-subject variation, we also expect by-item variation.
* There might be something special about *"Excusing for coming too late"*
  * Leading to overall higher pitch compared to

"Asking for a favor"

  * Regardless of the influence of politeness

Different questions



Extended model

* Account for them in our model:

frequency ~ attitude + gender + (1|subject) + (1|item) + ε

In R

* No default support for linear mixed models
* lme4 package

* Provides the function **lmer()**
* Comparable to **lm()**

Some R - code

```
library(lme4)

url <- "http://www.bodowinter.com/tutorial/politeness_data.csv"
politeness = read.csv(url)
boxplot(frequency ~ attitude + gender,
        col = c("white","lightgray"), politeness)

lmer(frequency ~ attitude + gender, data = politeness)
Error in mkReTrms(findbars(RHSForm(formula)), fr) : No random effects
terms specified in formula

politeness.model = lmer(
    frequency ~ attitude + gender + (1|subject) + (1|scenario),
    data=politeness
)
summary(politeness.model)
```

summary(politeness.model)

```
Linear mixed model fit by REML ['lmerMod']
Formula: frequency ~ attitude + (1 | subject) + (1 | scenario)
   Data: politeness

REML criterion at convergence: 793.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2006 -0.5817 -0.0639  0.5625  3.4385

Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept)  219     14.80
 subject  (Intercept) 4015     63.36
 Residual              646     25.42
Number of obs: 83, groups: scenario, 7; subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  202.588     26.754   7.572
attitudepol  -19.695      5.585  -3.527

Correlation of Fixed Effects:
            (Intr)
attitudepol -0.103
```

Model significance

```
politeness.null = lmer(
  frequency ~ gender + (1|subject) + (1|scenario),
  data=politeness, REML = FALSE
)
```

* Include attitude into the model

```
politeness.model = lmer(
  frequency ~ attitude + gender + (1|subject) + (1|scenario),
  data=politeness, REML = FALSE
)
```

Comparison

* anova(politeness.null, politeness.model)

```
Data: politeness
Models:
politeness.null: frequency ~ gender + (1 | subject) + (1 | scenario)
politeness.model: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
                 Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
politeness.null   5 816.72 828.81 -403.36   806.72
politeness.model  6 807.10 821.61 -397.55   795.10 11.618      1  0.0006532 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Publication

"… politeness affected pitch ($\chi^2_{(1)}$=11.62, p=0.00065), lowering it by about 19.7 Hz ± 5.6 (standard errors) …"

Random slopes versus intercepts

- Random slopes versus random intercepts
- You see that each scenario and each subject is assigned a different intercept.
- That's what we would expect, given that we've told the model with "(1|subject)" and "(1|scenario)" to take by-subject and by-item variability into account.

```
$scenario
  (Intercept) attitudepol   genderM
1    243.4859   -19.72207 -108.5173
2    263.3592   -19.72207 -108.5173
3    268.1322   -19.72207 -108.5173
4    277.2546   -19.72207 -108.5173
5    254.9319   -19.72207 -108.5173
6    244.8015   -19.72207 -108.5173
7    245.9618   -19.72207 -108.5173

$subject
   (Intercept) attitudepol   genderM
F1    243.3684   -19.72207 -108.5173
F2    266.9443   -19.72207 -108.5173
F3    260.2276   -19.72207 -108.5173
M3    284.3536   -19.72207 -108.5173
M4    262.0575   -19.72207 -108.5173
M7    224.1292   -19.72207 -108.5173

attr(,"class")
[1] "coef.mer"
```

Random intercept model

Fixed effects (attitude and gender) are all the same for all subjects and items.

Our model is a ***random intercept model***.

In this model, we account for baseline-differences in pitch,
We assume that whatever the effect of politeness is, it's going to **be the same for all** subjects and items.

Random slope model

* For example, it might be expected that some people are more polite, others less.
  * If so, what we need is a ***random slope*** model, where subjects and items are not only allowed to have differing intercepts, but where they are also allowed to have different slopes for the effect of politeness.

Random slope model

```
politeness.model = lmer(
  frequency ~ attitude + gender + (1 + attitude | subject)
                              + (1 + attitude | scenario),
  data=politeness, REML=FALSE
)
```

* The notation **“(1+attitude|subject)”** means that you tell the model to expect differing baseline-levels of frequency (the intercept, represented by 1) as well as differing responses to the main factor in question (attitude).

Summary

* Random effects
* Mixed Models
  * Random intercept model
  * Random slope model

* For the Assignment
  * 1) Read the tutorial

  http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf
  * 2) More practice exercises

An example:
Linear mixed model multiple QTL time series mapping

Overview

* Short introduction
  * Linear mixed models (LMM)
  * Multiple QTL mapping (MQM)
  * The Berlin fat mouse advanced inbred line (AIL)
* Model selection
  * Akaike information criterion (AIC)
  * Litter size + Litter Number = Litter type
  * Growth curves
* Results of LMM MQM time series mapping
* Conclusions / Discussion

Short introduction

Linear mixed models

- An extension to linear models, allowing for a combination of fixed and random effects
  - Fixed effects
    - Model parameters are fixed (non-random) quantities
    - Advantage: Non-biased estimates for parameters
  - Random effects
    - Model parameters are considered as random variables
    - Hierarchy between variables
    - Advantage: efficient, as such random/mixed effect models are good at dealing with repeated measurements

An example

Wikipedia

- m large elementary schools from a single country
- n pupils are chosen randomly at each school
- Y_{ij} is the score of the j^{th} pupil at the i^{th} school

Random effects example

Wikipedia

$$Y_{ij} = \mu + U_i + W_{ij}$$

* μ: Average test score for the entire population
* U_i: School-specific random effect
  * The difference between the average score at school *i* and the average score in the entire country
* W_{ij}: Individual-specific random effect
  * The difference of the *j*-th pupil's score from the average for the *i*-th school

Fixed effects example

Wikipedia

* Fixed effects can capture differences in scores among different groups across different schools
  * For example:
    * Sex of the individual (Male, Female)
    * Race (White, Black, Chinese)
    * Parent education level

$$Y_{ij} = \mu + \beta_1 \mathrm{Sex}_{ij} + \beta_2 \mathrm{Race}_{ij} + \beta_3 \mathrm{ParentsEduc}_{ij} + U_i + W_{ij}$$

Short introduction

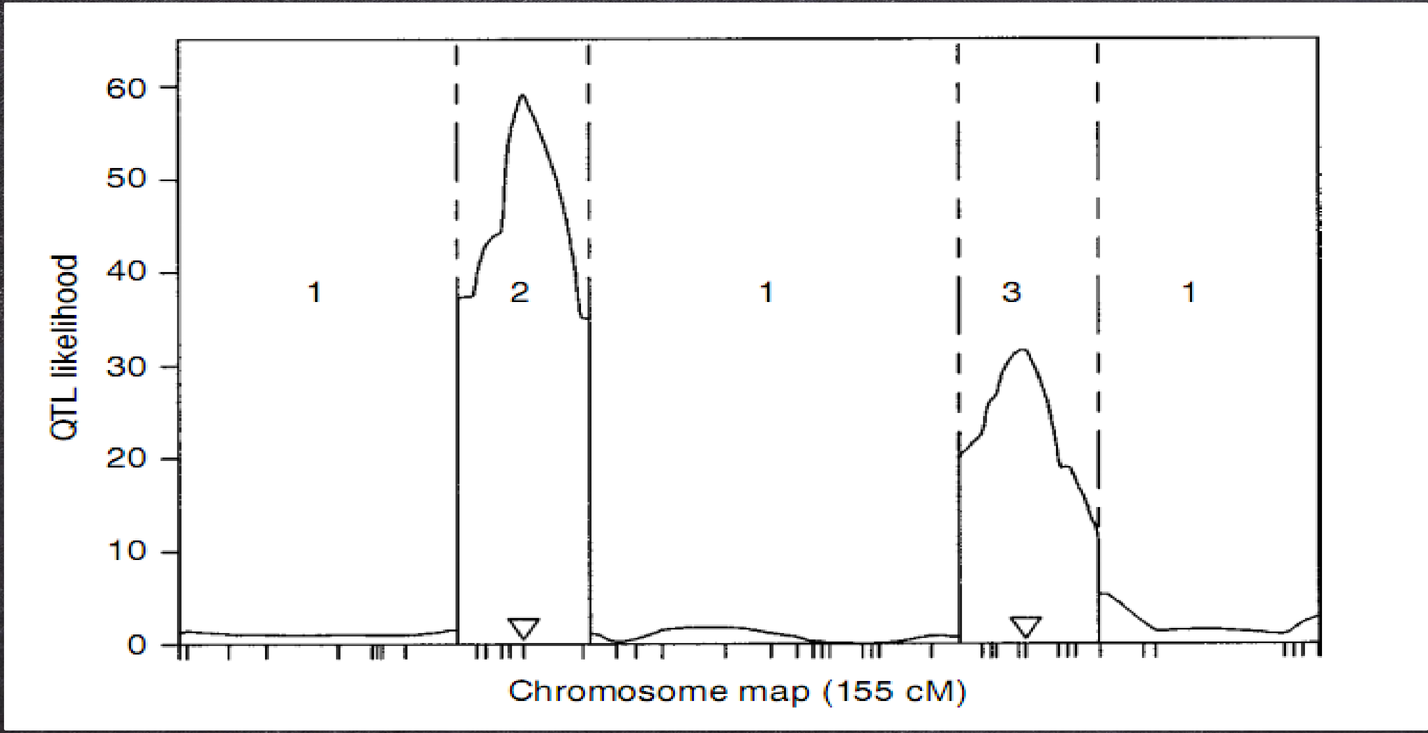
Multiple QTL mapping

* Genetic markers as fixed effects into the model
  * Account for known genetic effects
  * More power to detect other effects
  * Disentangle QTLs in close proximity (LD)
  * QTLs with opposite direction of effects

* Model selection
* QTL detection using the best model

Short introduction

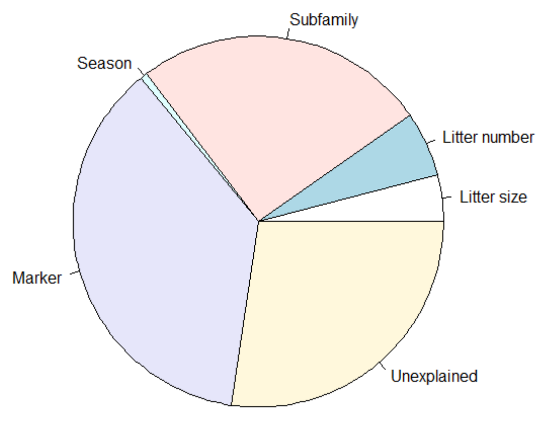
Multiple QTL mapping

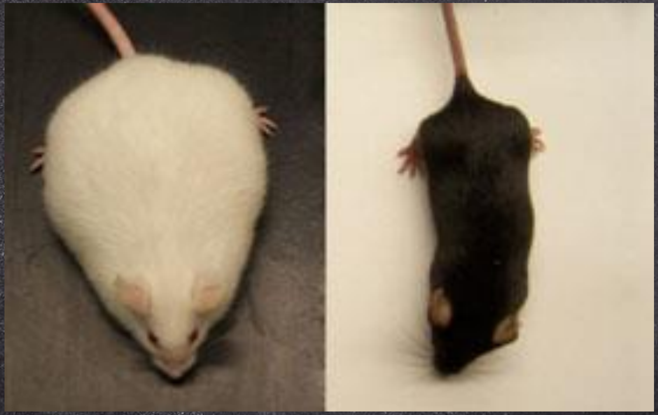


Short introduction

The Berlin fat mouse advanced inbred line

* Model organism for polygenic obesity
* Five fold increased fat percentage (compared to B6)
* Long-term selected for high fatness
* Several features of the metabolic syndrome





Materials & Methods

* 344 individuals in generation 28
* 17.971 genetic markers (after QC)

* Time series data on body weight
  * Days: 21, 28, 35, 42, 49, 56, 63, 70

Model selection

Akaike information criterion (AIC)

* Model selection is the task **of selecting a statistical model** from a set of **candidate models**
* Akaike information criterion (AIC) is an estimator of the **relative quality** of statistical models
* Lower = Better

Model selection

Litter size + Litter Number = Litter type

* Litter size - Number of individuals in a litter
* Litter number - The nth litter of a female
  * Encoded in two different ways
    * Litter A (1st), Litter B (2nd), etc (5 levels)
    * F (1st) versus N (not the 1st)

* Litter type - Combination of Litter size and number
  * Lt5 = A8, B10, C12
  * Lt2 = F8, N10, F10, N12

Model selection

Litter size + Litter Number = Litter type

* Null-model (M0)

$$P = F + (1|\text{individual})$$

Fixed effect

Random effect

  * P = Body weight
  * F = ID of the Father

ID	Model	Random effect	Degrees of freedom
m0	P = F	1|individual	30
m1_L2	$P = F + L_{n2}$	1|individual	30 + 1
m1_L5	$P = F + L_{n5}$	1|individual	30 + 4
m2_L2	$P = F + L_{n2} + L_s$	1|individual	30 + 1 + 4
m2_L5	$P = F + L_{n5} + L_s$	1|individual	30 + 4 + 4
m2_Lt2	$P = F + L_{t2}$	1|individual	30 + 7
m2_Lt5	$P = F + L_{t5}$	1|individual	30 + 13

	m1_L2	m1_L5	m2_L2	m2_L5	m2_Lt2	m2_Lt5
m0	-18.452	-20.114	-21.706	-19.988	-23.245	-19.254
m1_L2		-1.662	-3.254	-1.537	-4.794	-0.802
m1_L5	1.662		-1.592	0.126	-3.131	0.860
m2_L2	3.254	1.592		1.718	-1.540	2.452
m2_L5	1.537	-0.126	-1.718		-3.257	0.734
m2_Lt2	4.794	3.131	1.540	3.257		3.992
m2_Lt5	0.802	-0.860	-2.152	-0.734	-3.992	
	-6.403	-18.039	-28.881	-17.158	-39.960	-12.017
Rank	6	3	2	4	1	5

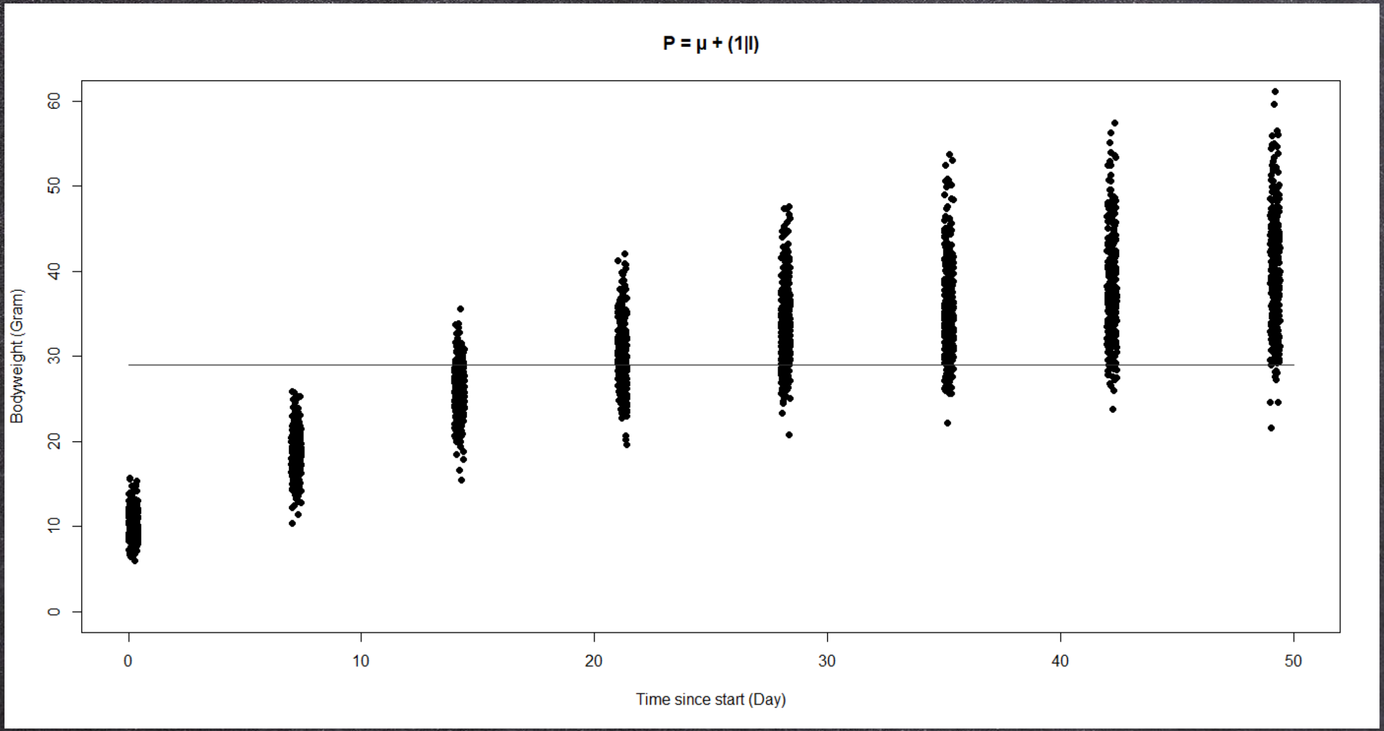
Model selection

Growth curves

- Stepwise model selection
  - AIC drop of > 10 is considered model improvement

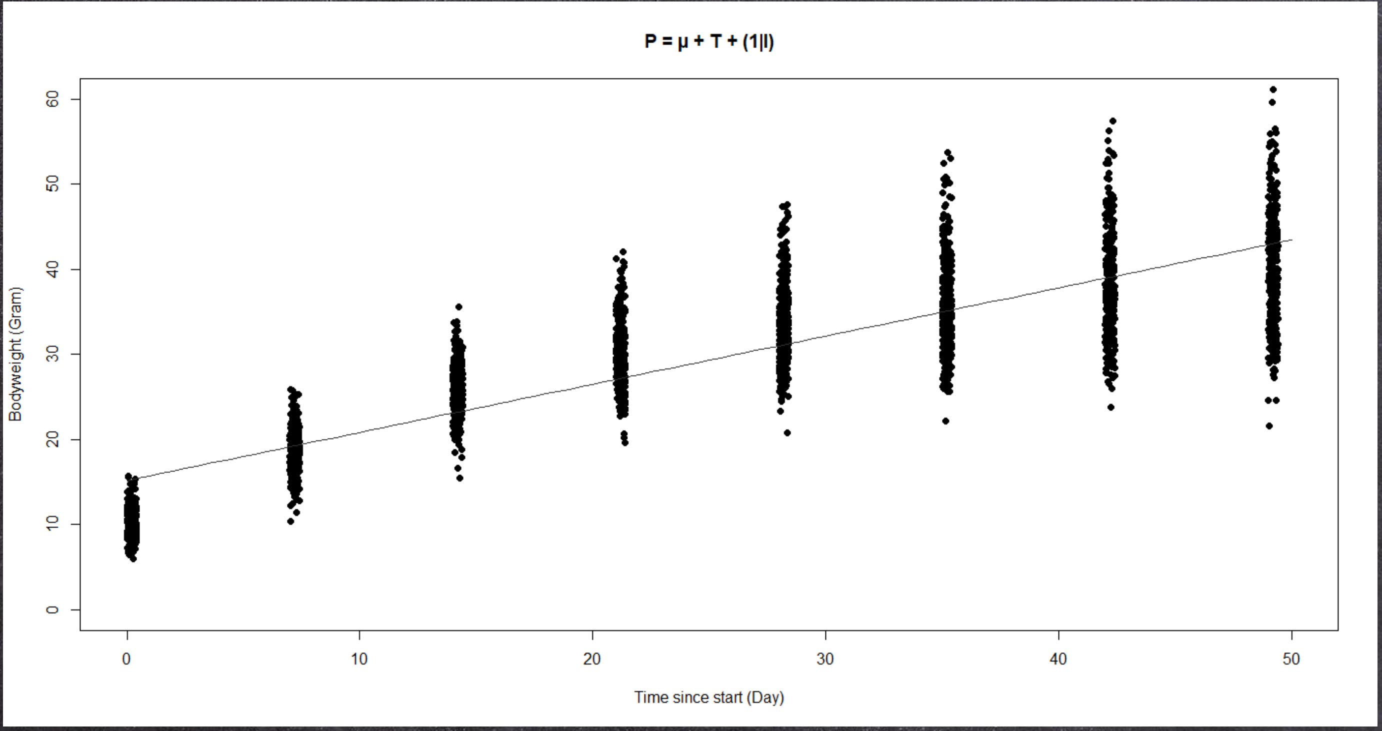
ID	Model	Random effect	Δ AIC	Model comparison result
m3	$P = F + L_{t2}$	1|individual		
m4	$P = F + L_{t2} + S$	1|individual	1.7	season should NOT be a fixed effect
m5	$P = F + L_{t2} + T$	1|individual	-4700.2	time should be a fixed effect
m6	$P = F + L_{t2} + T$	time|individual	-770.2	time should be a random slope effect
m7	$P = F + L_{t2} + T + T^2$	time|individual	-3556.0	$time^2$ should be a fixed effect
m8	$P = F + L_{t2} + T + T^2 + T^3$	time|individual	-962.5	$time^3$ should be a fixed effect
m9	$P = F + L_{t2} + T + T^2 + T^3 + T^4$	time|individual	-6.6	$time^4$ should NOT be a fixed effect
m10	$P = F + L_{t2} + T + T^2 + T^3 + M_{(jObes1)} + (M_{(jObes1)}:T)$	time|individual	-225.2	jObes1 top marker and interaction with time should be included

Visualizing the LMM models
(random effects not shows)



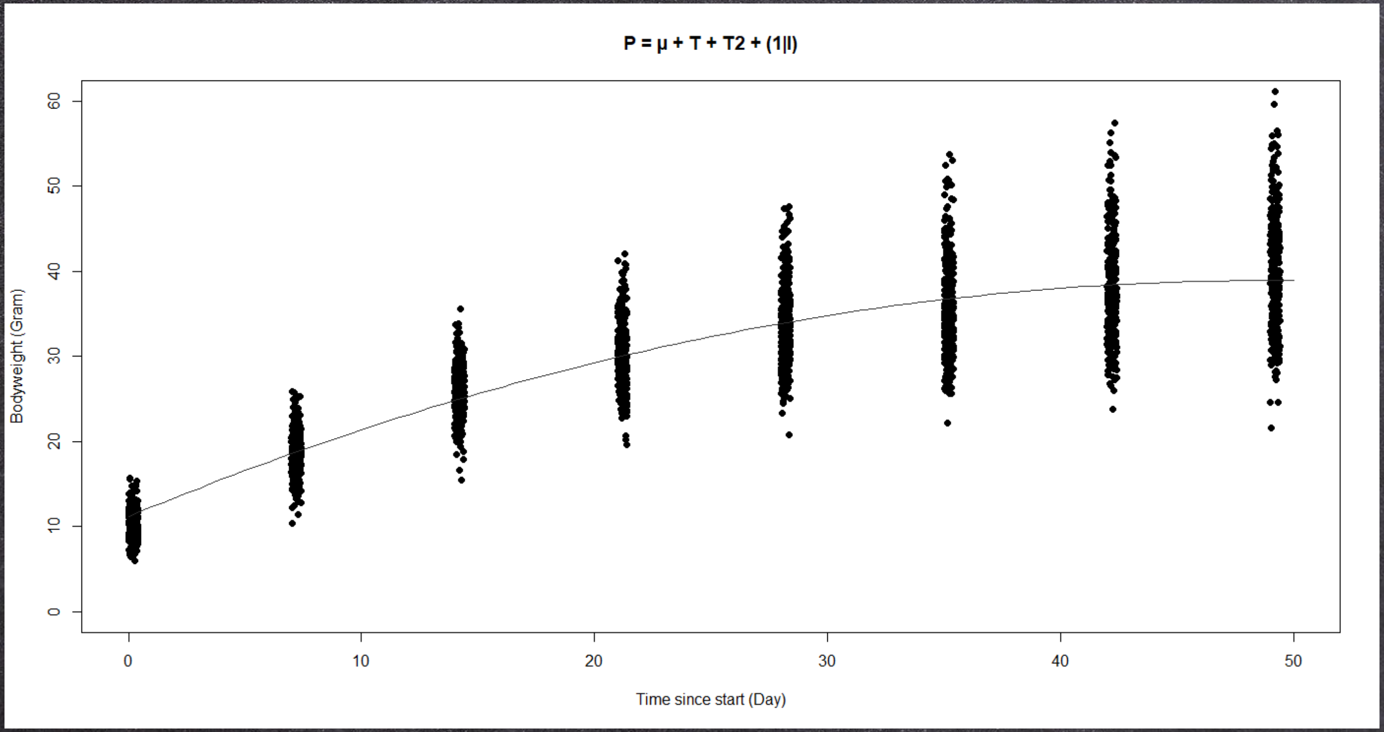
Estimate the global mean

Visualizing the LMM models
(random effects not shows)



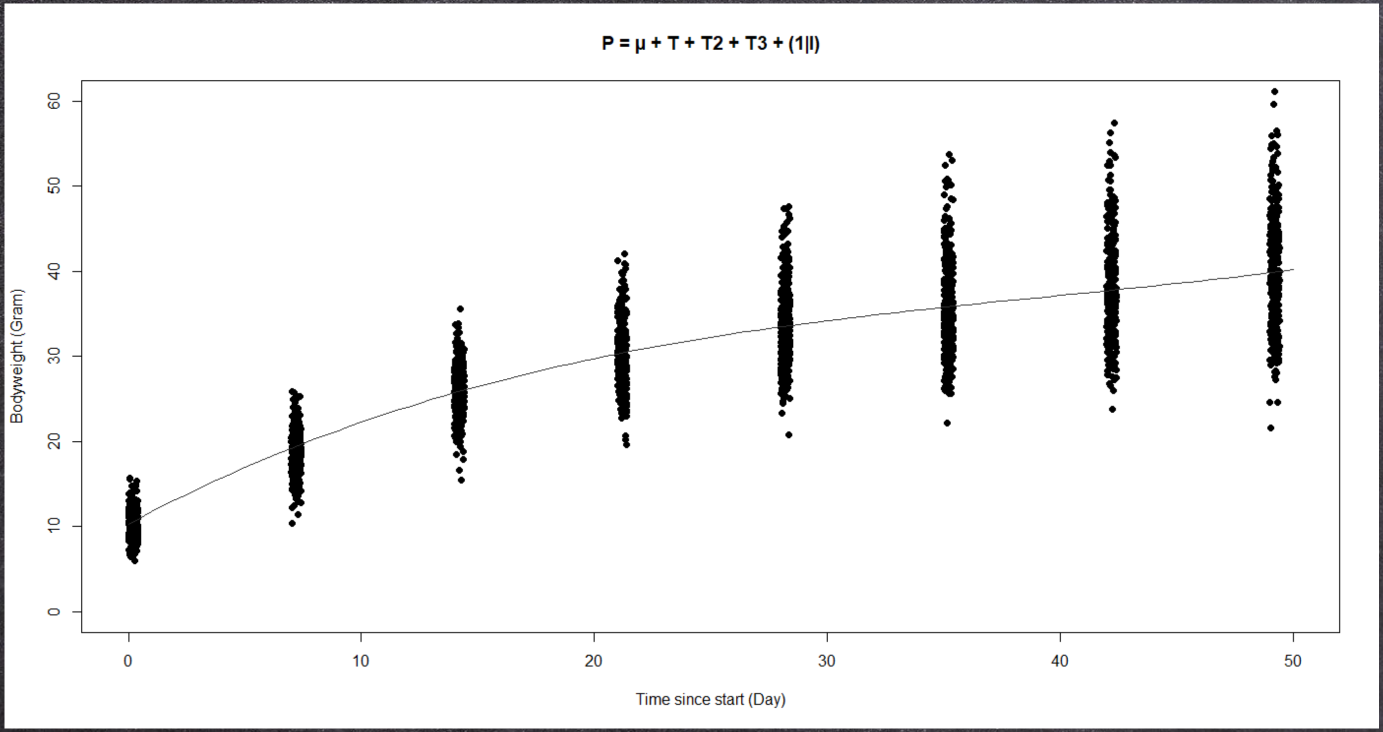
Mean + Time

Visualizing the LMM models
(random effects not shows)



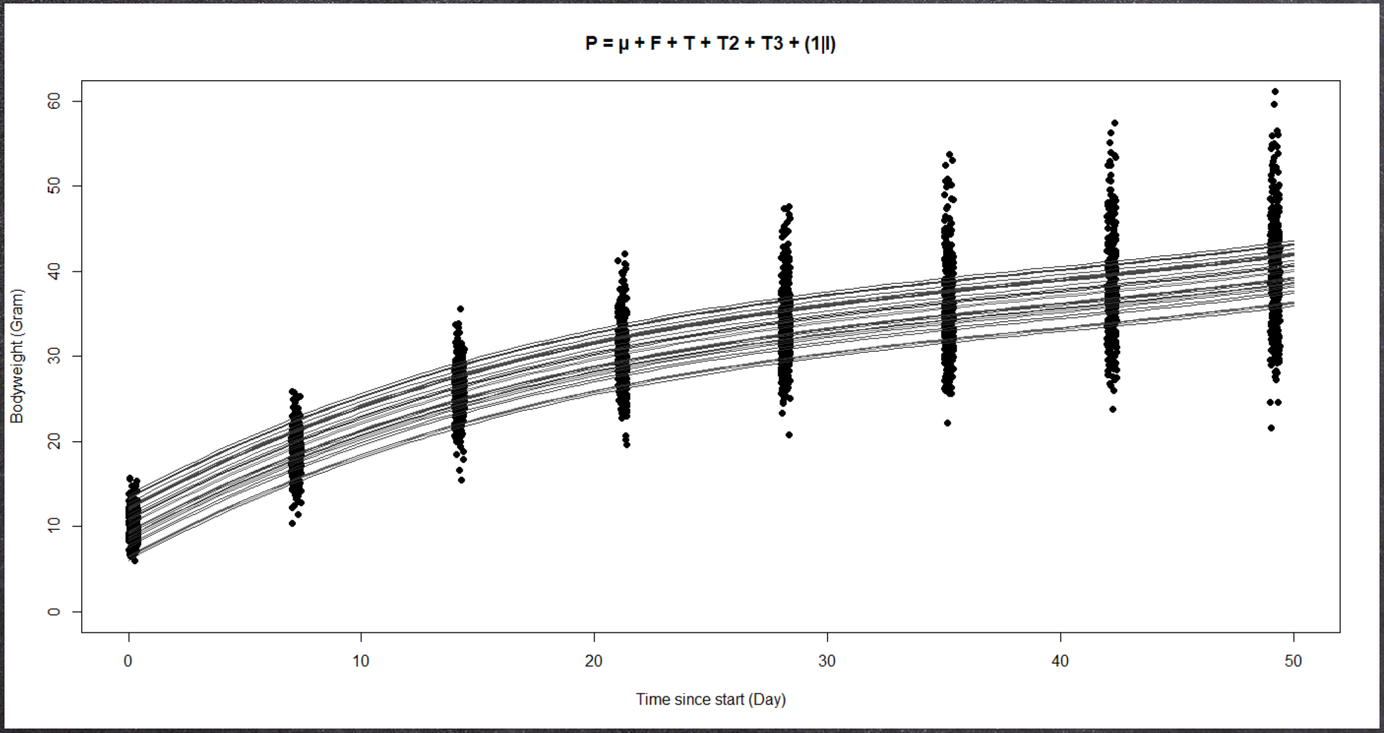
Mean + Time + $Time^2$

Visualizing the LMM models
(random effects not shows)



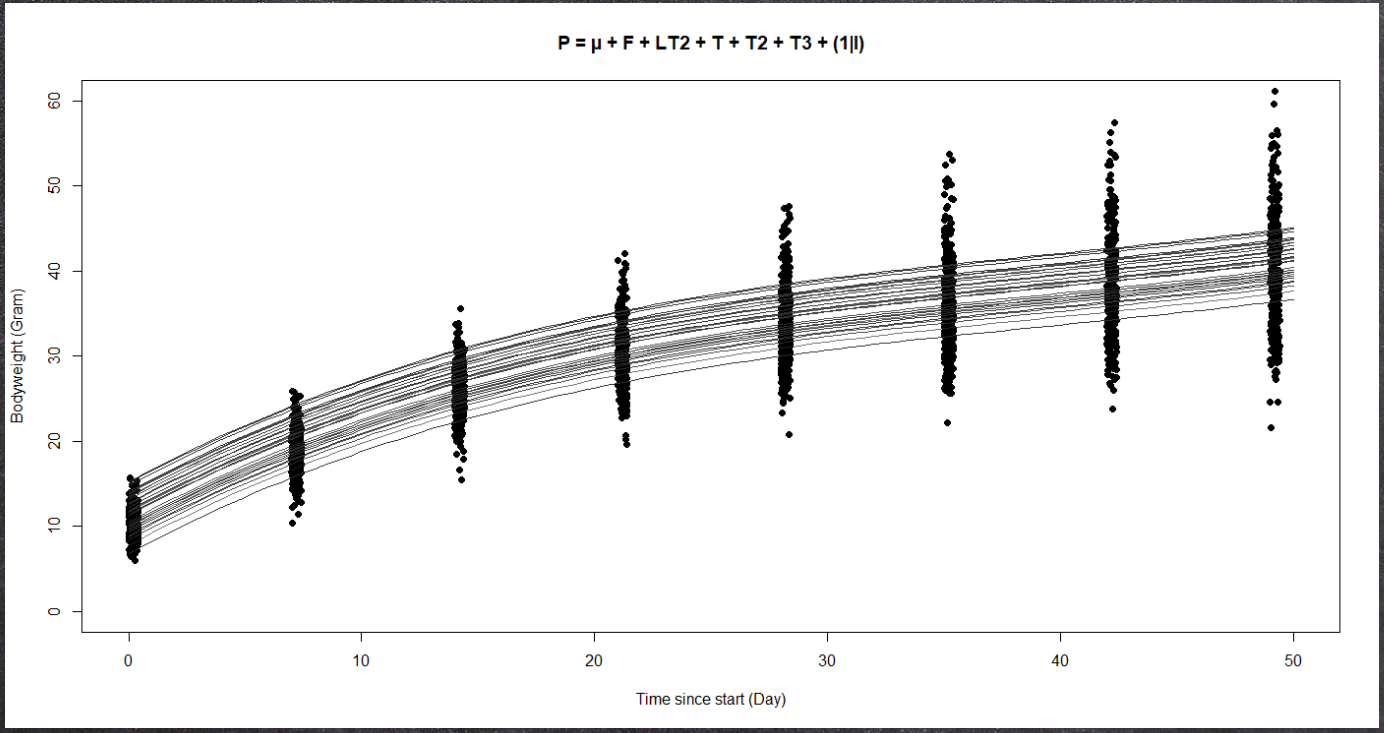
Mean + Time + Time^2+ Time^3

Visualizing the LMM models
(random effects not shown)



Mean + Family + Time + Time^2+ Time^3

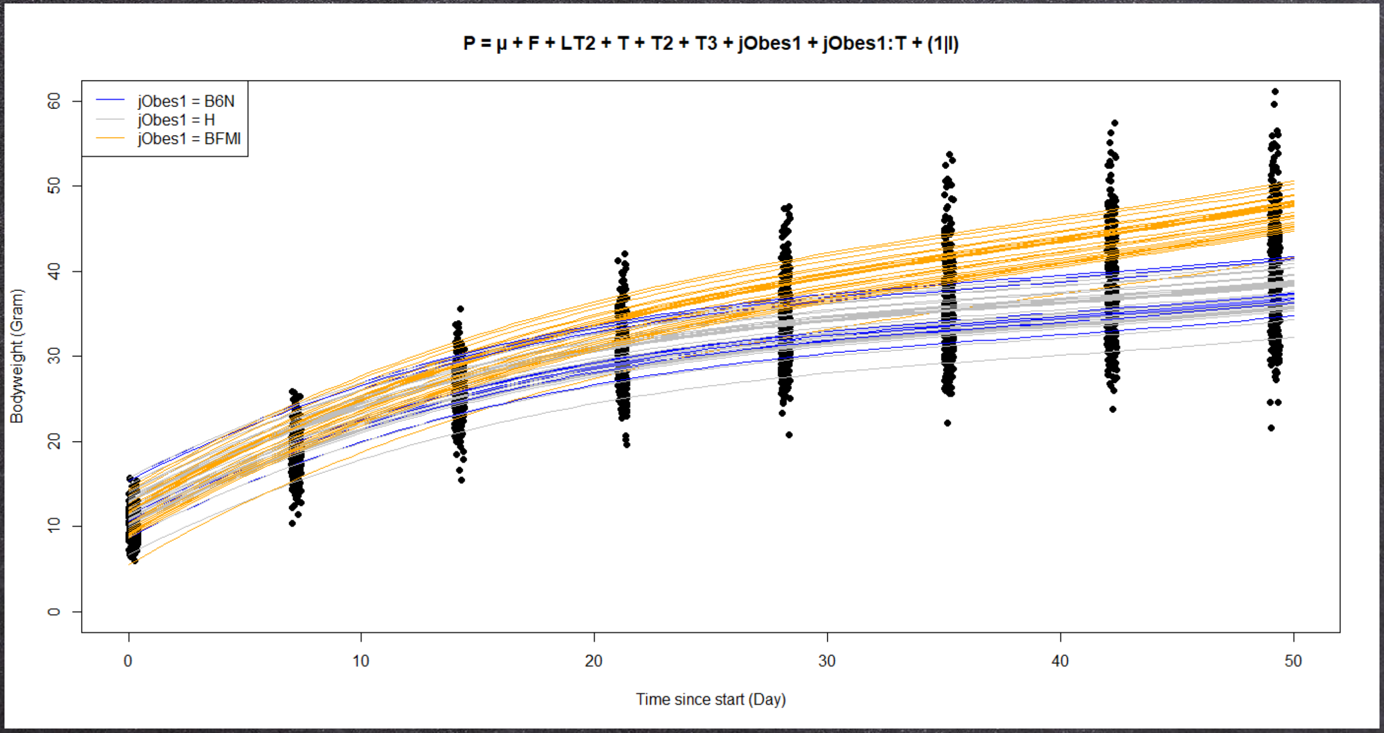
Visualizing the LMM models
(random effects not shows)



Mean + Family + Litter type$_{(2)}$ + Time + Time^2+ Time^3

Visualizing the LMM models

(random effects not shows)

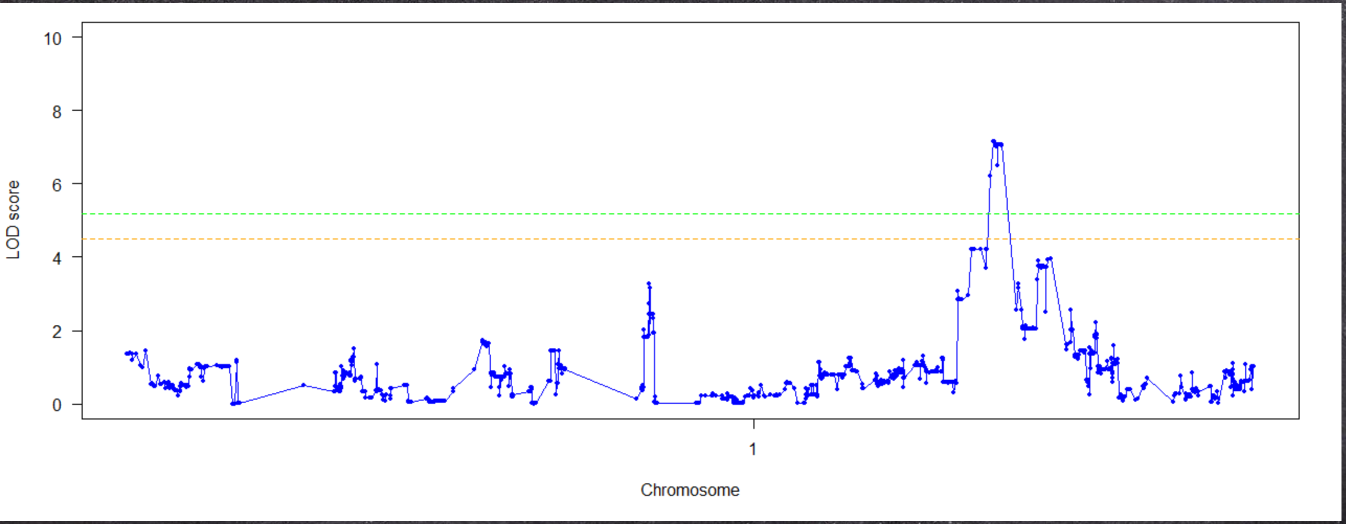


Mean + Family + Litter type$_{(2)}$ + Time + Time^2+ Time^3+ jObes1 + jObes1:Time

QTL mapping

- Scan to the genome
  - Add the marker under consideration to the model
  - For example: chromosome 1

$+ M_x + M_x{:}T$



Mean + Family + Litter $\text{type}_{(2)}$ + T+ T^2+ T^3+ jObes1 + jObes1:Time+ M_x + $M_x{:}T$

Results

LMM MQM time series mapping

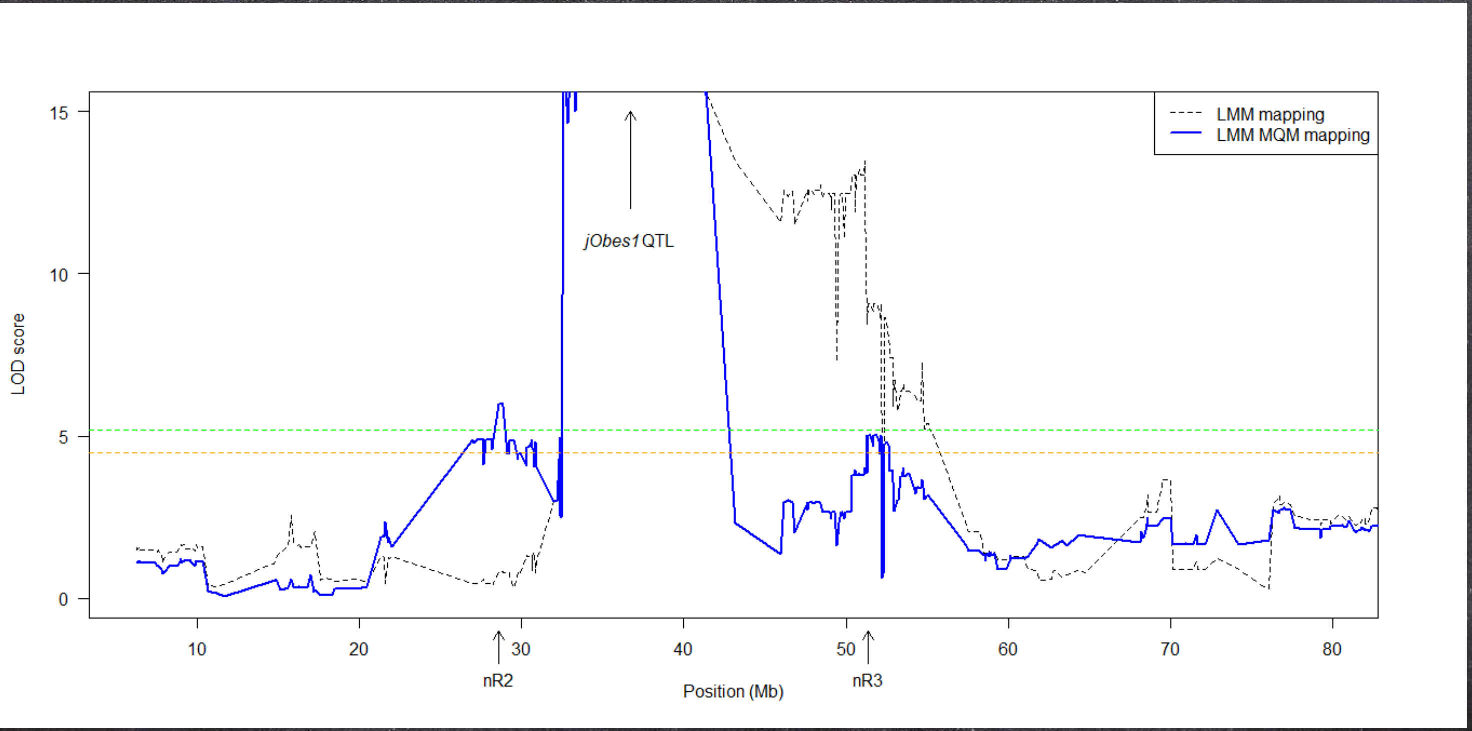
* After scanning all chromosomes
  * 5 new QTL detected

Name	Chr	Start	Top marker	Stop	LOD	Number of alleles			Effect relative to B6N				
						B6N	H	BFMI	H	BFMI		H/Day	BFMI/Day
nR1	1	149,553,681	UNC1938399	154,868,088	**7.14**	83	145	116	0.80	0.41		-0.078	-0.067
nR2	3	26,989,539	UNC030576333	35,953,921	**5.99**	46	173	125	0.10	0.42		0.039	-0.025
nR3	3	49,901,885	JAX00522656	52,973,026	5.03	60	158	126	0.21	-0.08		0.046	0.081
nR4	9	86,816,288	UNC090485124	99,363,348	**6.34**	171	133	40	-0.05	0.49		-0.027	-0.105
nR5	19	37,825,545	UNC30294194	40,410,259	4.83	81	179	81	0.11	-0.37		0.012	0.069
jObes1	3	36,481,201	UNC5048297	36,854,743	**43.23**	39	165	140	-0.07	-1.44		-0.011	0.201

Results

LMM MQM time series mapping

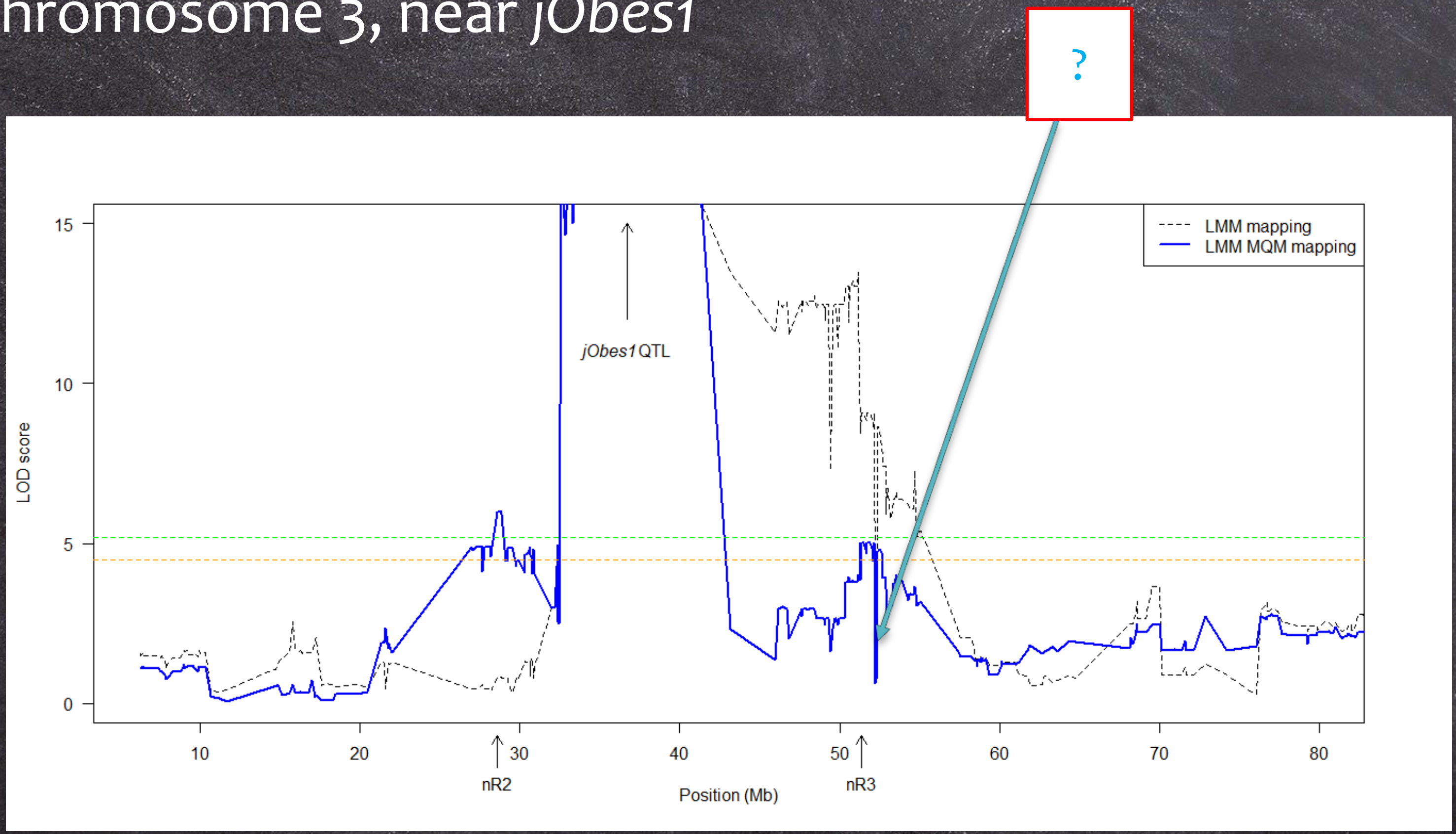
* Chromosome 3, near *jObes1*



Results

LMM MQM time series mapping

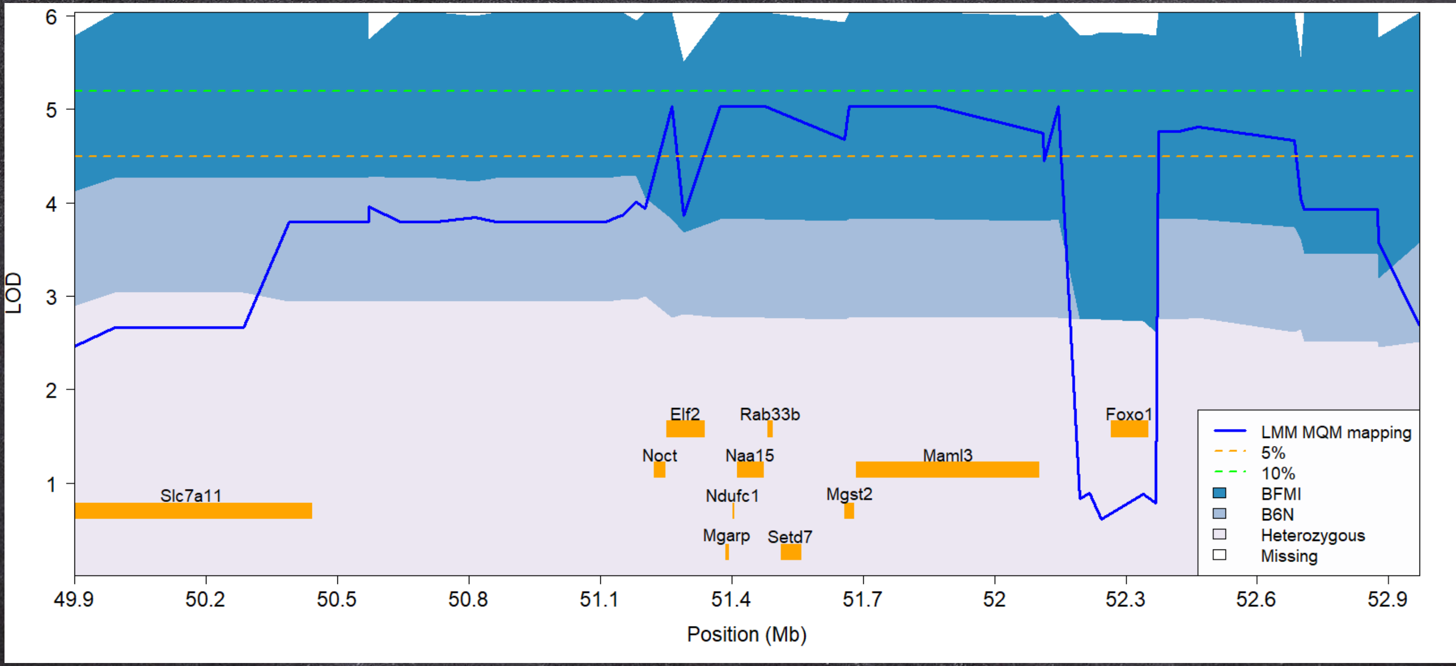
* Chromosome 3, near *jObes1*



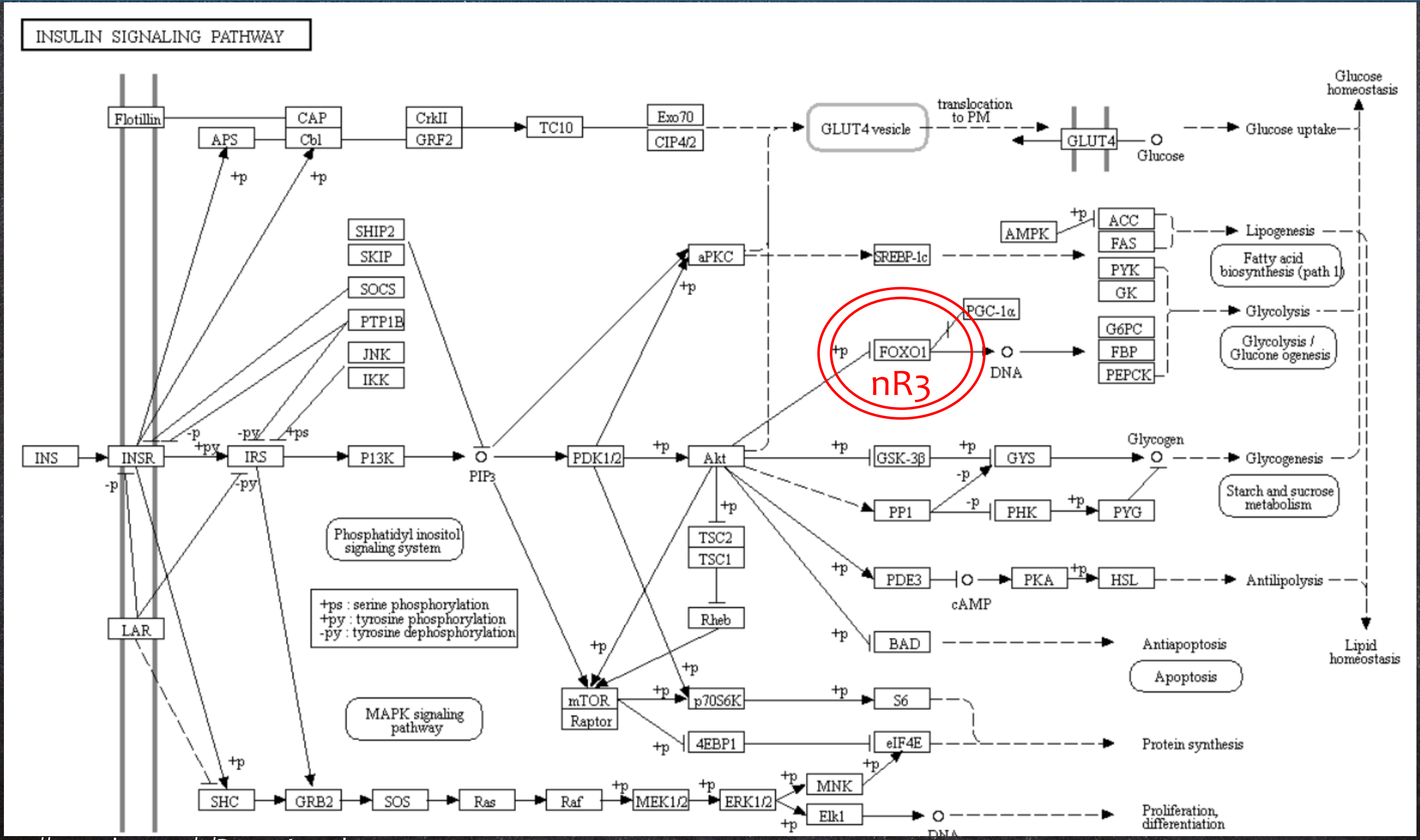
Results

LMM MQM time series mapping

- Segregation distortion at nR3
  - *Foxo1* is a well known regulator of insulin

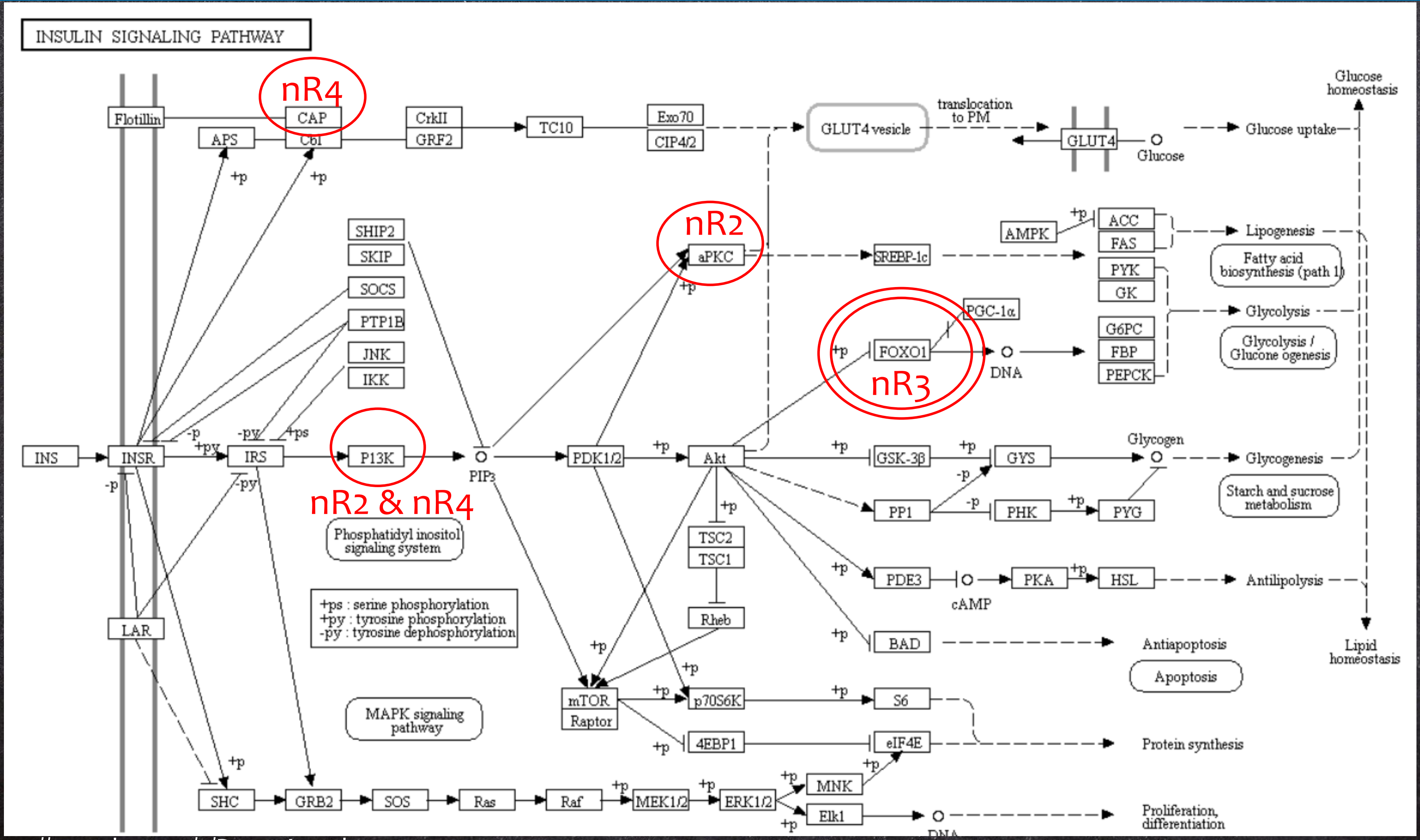


Insulin Pathway



https://youtube.com/c/DannyArends

Insulin Pathway



https://youtube.com/c/DannyArends

Conclusions / Discussion

* LMM MQM time series mapping is more sensitive
  * It uses all available data
  * Corrects for known genetic effects
* 5 novel QTL are detected
* Within the nR3 QTL segregation distortion is observed
  * *Foxo1* is the only gene in this region
* Many genes from the insulin pathway located underneath the newly identified regions

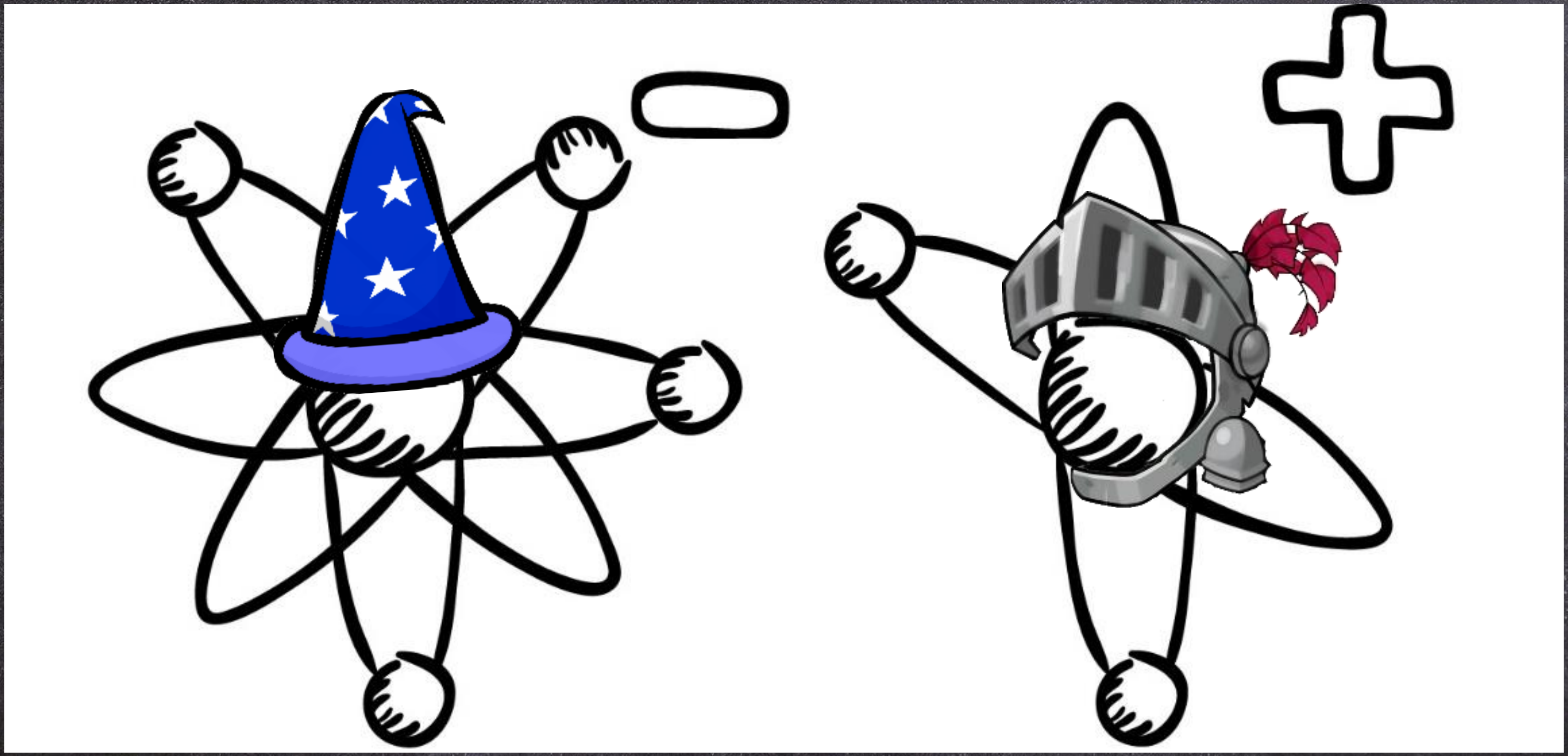
Summary

* Random effects
* Mixed Models
  * Random intercept model
  * Random slope model

* For the Assignment
  * 1) Read the tutorial
  http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf
  * 2) More practice exercises

* An example on how linear mixed models can improve QTL detection

Quest ions ?



Ions by **Iluvia ramos https://prezi.com**
Wizard Hat - New Horizon & Interactive Studios - http://clubpenguin.wikia.com
Knight helm - Plants vs Zombies 2 - PopCap Games (Juli 2013)