

# Genomic BLUP

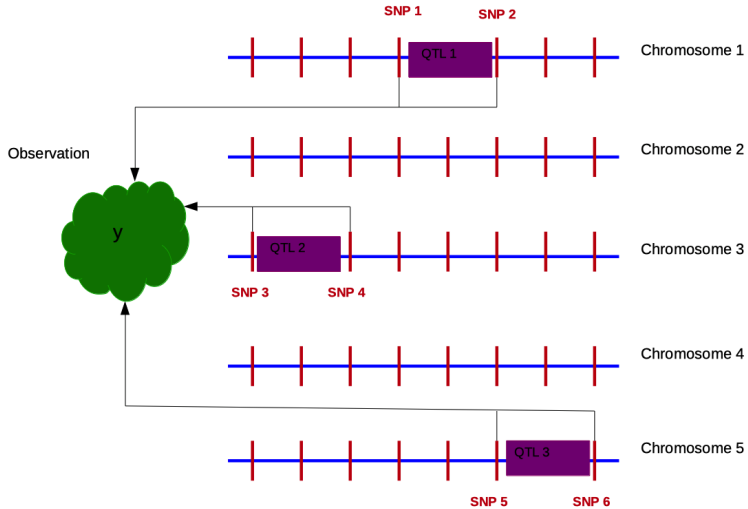
Peter von Rohr

2021-03-08

## So Far

- ▶ Estimate effect of few SNP loci linked to QTL
- ▶ Use parameter estimates to predict genomic breeding values
- ▶ **New:** Many SNP, find the important ones

# Situation



Goal: Find SNP 1 – SNP 6 out of the many SNPs

# Approaches in Fixed Linear Model Framework

## Two Approaches

1. Forward selection: Start with empty model, include predictors that improve model
2. Backward elimination: Start with full model, remove predictors as long as model does not get worse

## Forward Selection

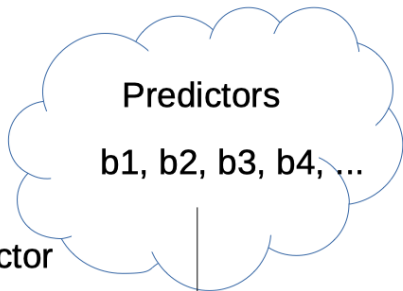
Start with empty model

$$y = b_0 + e$$



Include additional predictor

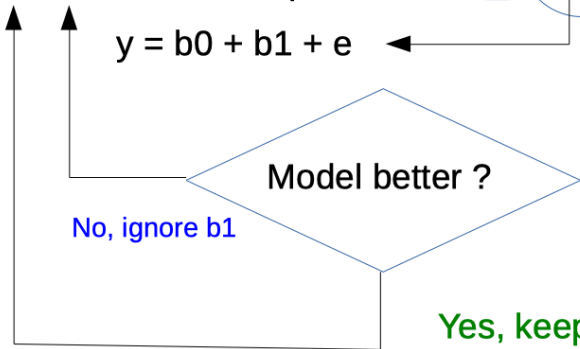
$$y = b_0 + b_1 + e$$



Model better ?

No, ignore b1

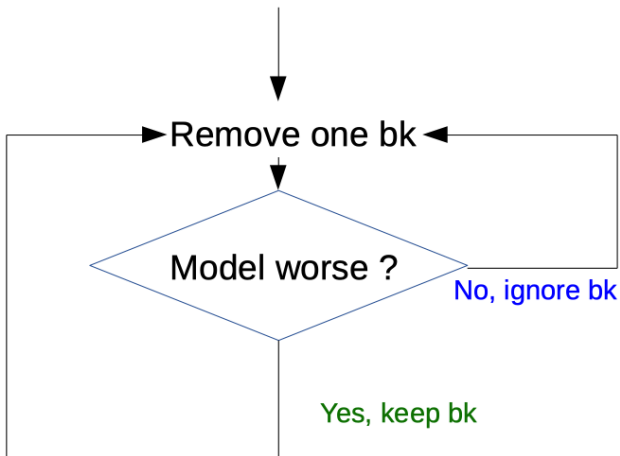
Yes, keep b1



# Backward Elimination

Start with full model

$$Y = b_0 + b_1 + b_2 + b_3 + \dots + e$$



# Model Selection With Genomic Data

- ▶ Only backward elimination really works in practical problems
- ▶ Large number of predictors ( $1.5 * 10^5$ )
- ▶ How to determine sequence of predictors to eliminate
- ▶ Fitting the full model is problematic

## Mixed Linear Effect Model

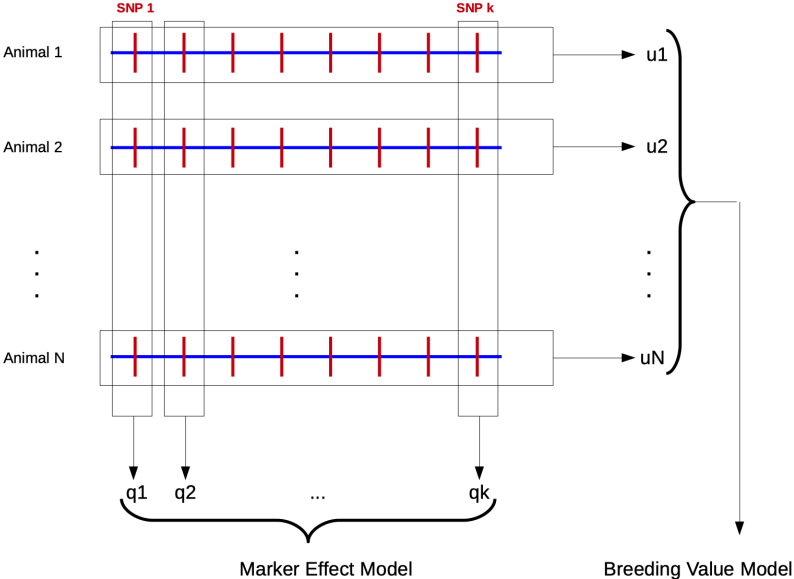
- ▶ One solution: replace fixed linear effect model by **mixed** linear effect model (mle)
- ▶ MLE: additional random effect besides error term
- ▶ Random effects are specified by expected value and variance
- ▶ In livestock breeding MLE have a good reputation from BLUP animal model



# MLE In Genomics

- ▶ Two different parametrizations
  1. Marker Effect Model (MEM)
  2. Breeding Value Model (BVM)

# Overview



## Marker Effect Model

In MEM random effects of markers are directly included in the model. For an idealized data set we can write

$$y = 1_n\mu + Wq + e$$

where

- $y$  vector of length  $n$  with observations
- $\mu$  general mean denoting fixed effects
- $1_n$  vector of length  $n$  of all ones
- $q$  vector of length  $m$  of random SNP effects
- $W$  design matrix relating SNP-genotypes to observations
- $e$  vector of length  $n$  of random error terms

# Breeding Value Model

$$y = Xb + Zg + e$$

where

- $y$  vector of length  $n$  with observations
- $b$  vector of length  $r$  with fixed effects
- $X$  incidence matrix linking elements in  $b$  to observations
- $g$  vector of length  $t$  with random genomic breeding values
- $Z$  incidence matrix linking elements in  $g$  to observations
- $e$  vector of length  $n$  of random error terms