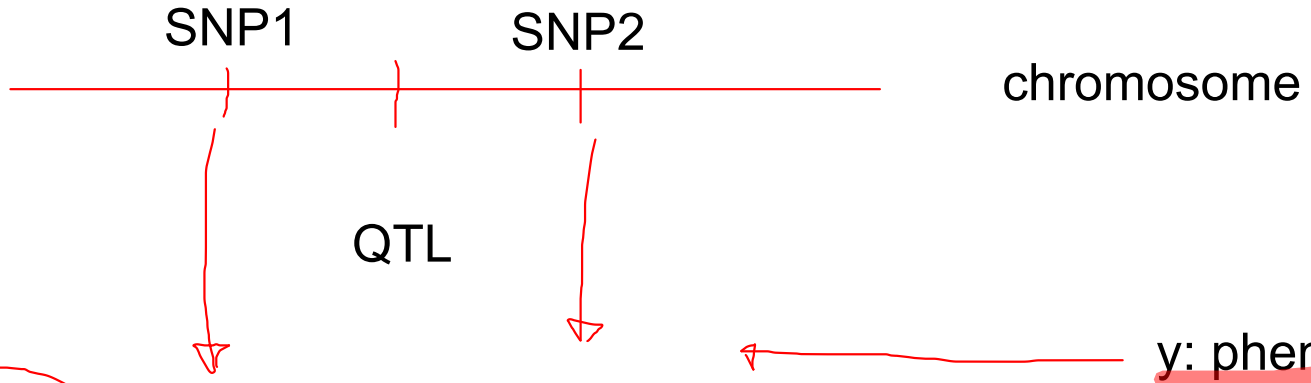
Genomic BLUP

Peter von Rohr

2021-03-08

So Far



y: phenotypes

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

N

a1: marker effect for SNP1

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

150'000 - 20 Mio

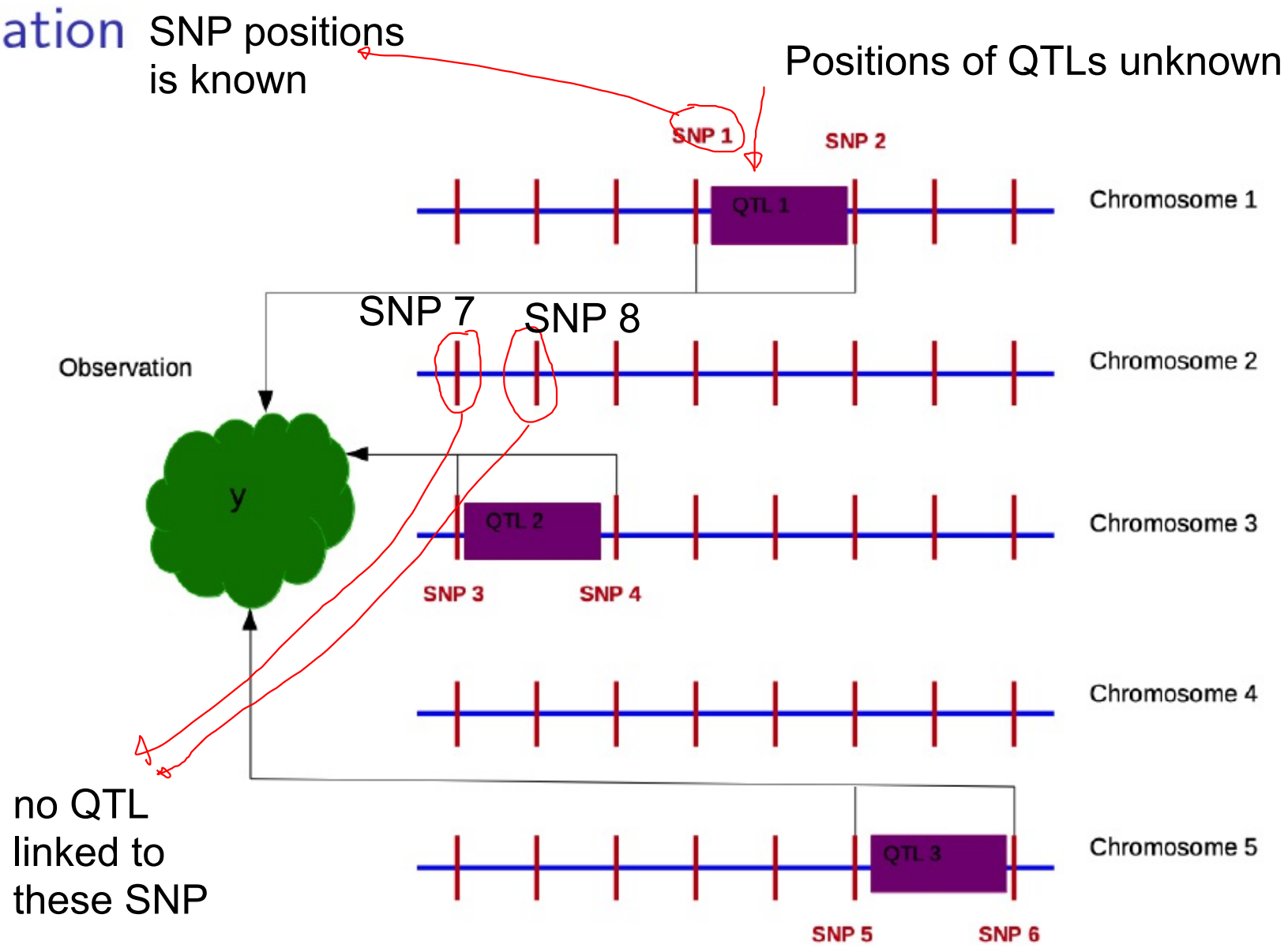
denoted by the variable p

$$p \gg N$$

combine estimation of marker effects together with availability of many SNP in the dataset

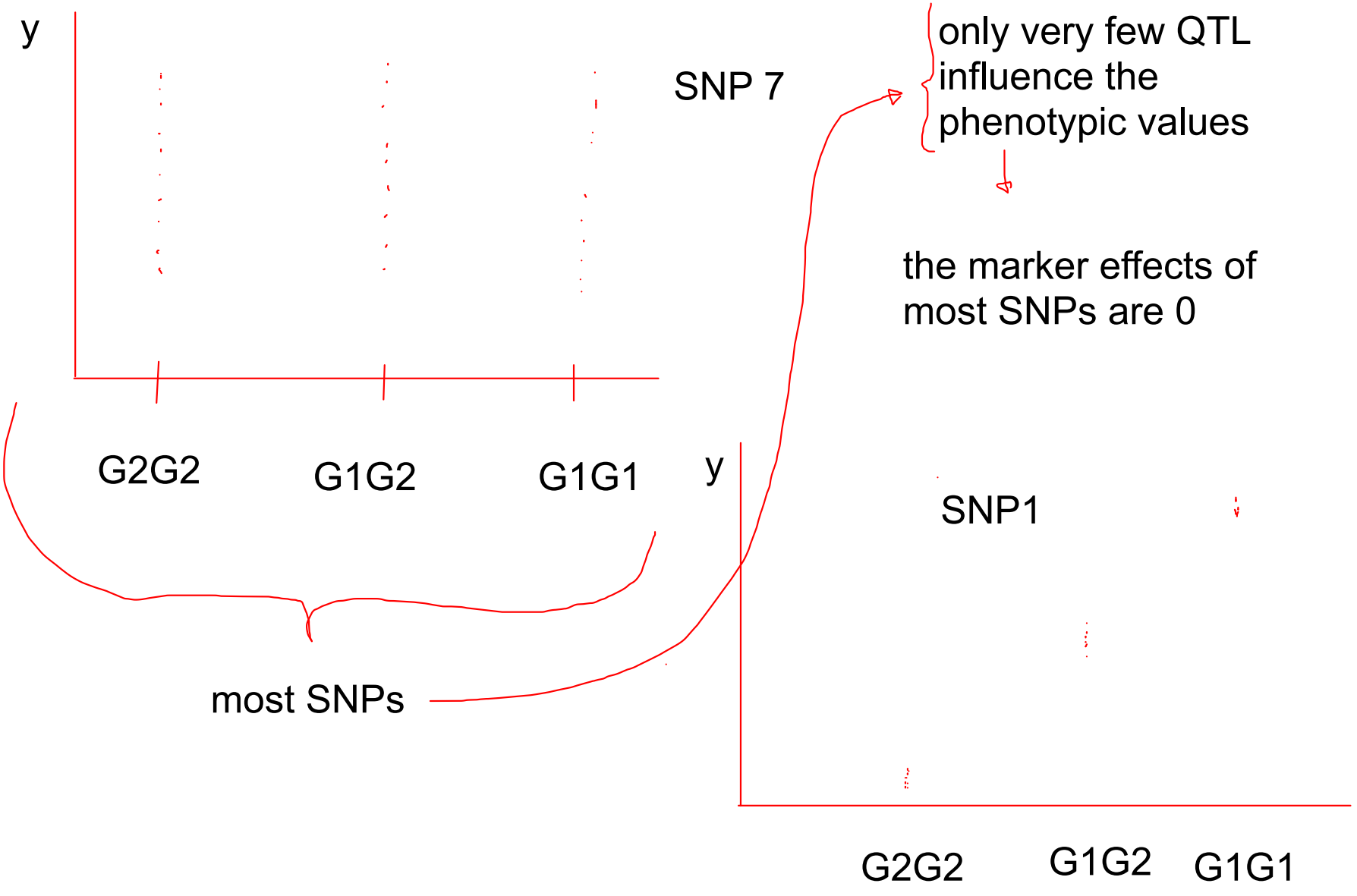
least squares cannot be used to estimate marker effects

Situation



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

For the two SNPs 7 and 8 that are not linked to a QTL



Summary: 2 Problems

1. if we consider all SNP in our data set, then p>>N
   ==> least squares cannot be used
2. from genetic model: only few QTL for a given trait,
   ==> most SNP have marker effects (a) = 0

because the position of the QTL is unknown, we do not know which SNP have marker effects = 0

Approaches in Fixed Linear Model Framework

Possible solution for problem 2: Model selection to determine which SNP have marker effect that are not 0

no SNP

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

all SNP, not possible due to p>>N

Forward Selection

no SNP

Start with empty model

M_0: y = b0 + e

intercept

$SSR = r_1^2 + r_2^2 + \dots s_N^2$

Predictors

SNP

b1, b2, b3, b4, ...

marker effects

Include additional predictor

M_1: y = b0 + b1 + e

M_2: $y = b_0 + b_1 + b_2 + e$

M_p:

choose b1 such that new SSR is reduced the most

Model better ?

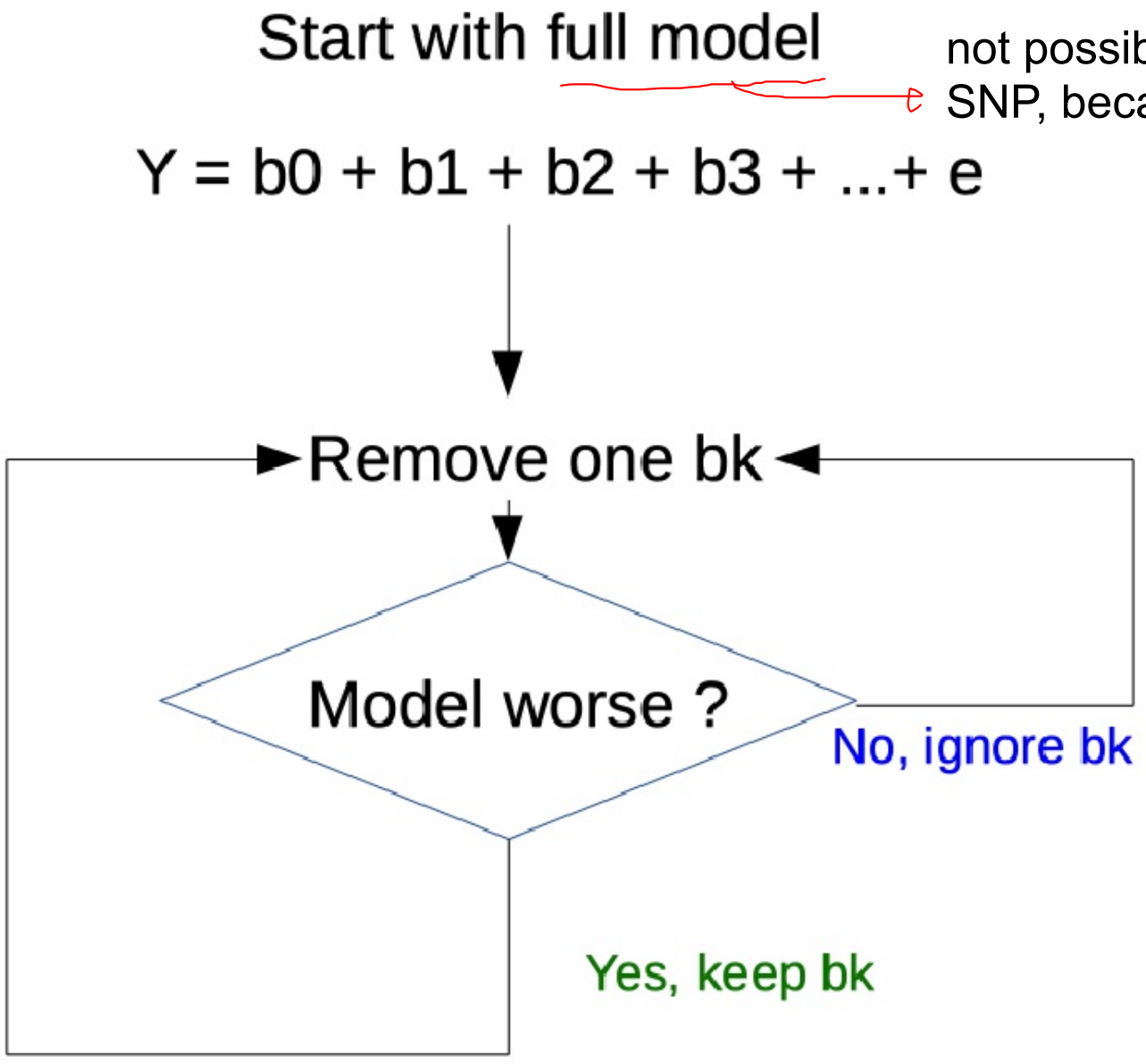
Criterions: Mallow Cp, Akaike

No, ignore b1

Yes, keep b1

Backward Elimination

except for SNP data, this is the preferred way



Model Selection With Genomic Data

cannot use backward elimination with genomic data, because parameter estimation in the full model cannot be

- 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

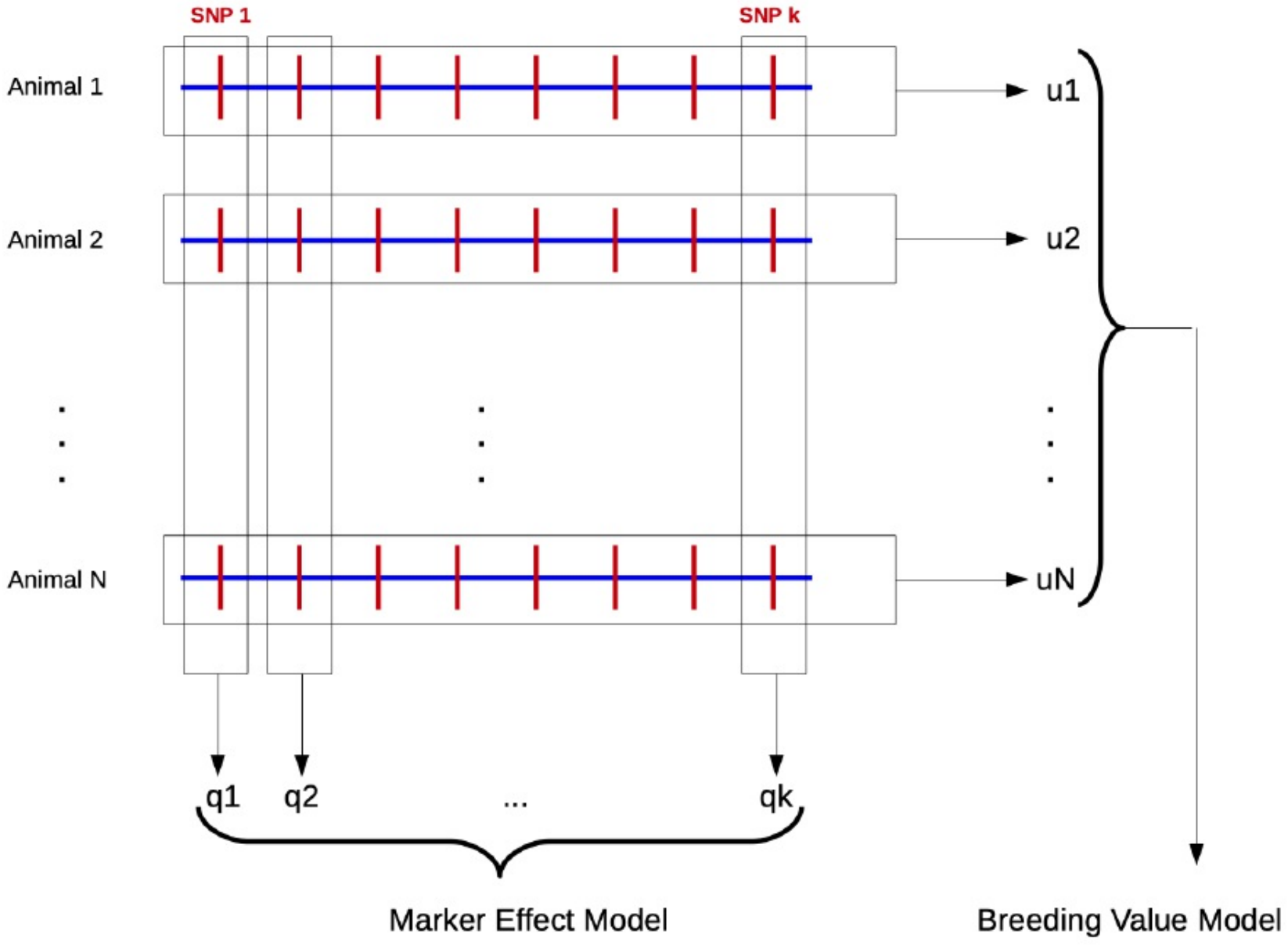
- One solution: replace fixed linear effect model by **mixed** linear effect model (mle)
- MLE: additional random effect besides error term
- Random effects are specified by expected value and variance
- In livestock breeding MLE have a good reputation from BLUP animal model

MLE In Genomics

- Two different parametrizations

1. Marker Effect Model (MEM)
2. Breeding Value Model (BVM)

Overview


SNP 1
SNP k
Animal 1
Animal 2
Animal N
u1
u2
uN
q1
q2
...
qk
Marker Effect Model
Breeding Value Model

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
μ	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