Formatting, filtering, imputing genotype data

Aaron Lorenz
Data Bootcamp for Genomic Prediction in Plant Breeding
Ghent, Belgium
July 5-7, 2023

The data

- Data is from the Soybean Nested Association Mapping project
- 39 families of recombinant inbred lines (RILs)
- ~140 RILs per family
- 5487 total RILs recombinant inbred lines
- 4292 SNPs in dataset.
- Phenotypes, extracted from SoyNAM package, are de-regressed BLUPs from a multi-environment trial (18 environments).
- Traits: Yield, maturity, protein, oil, height, seed size

- Note: I subset this dataset to speed up computations for purposes of demonstration

The data

Genomic prediction in SoyNAM

Soybean Nested Association Mapping

Diers et al. (2018)

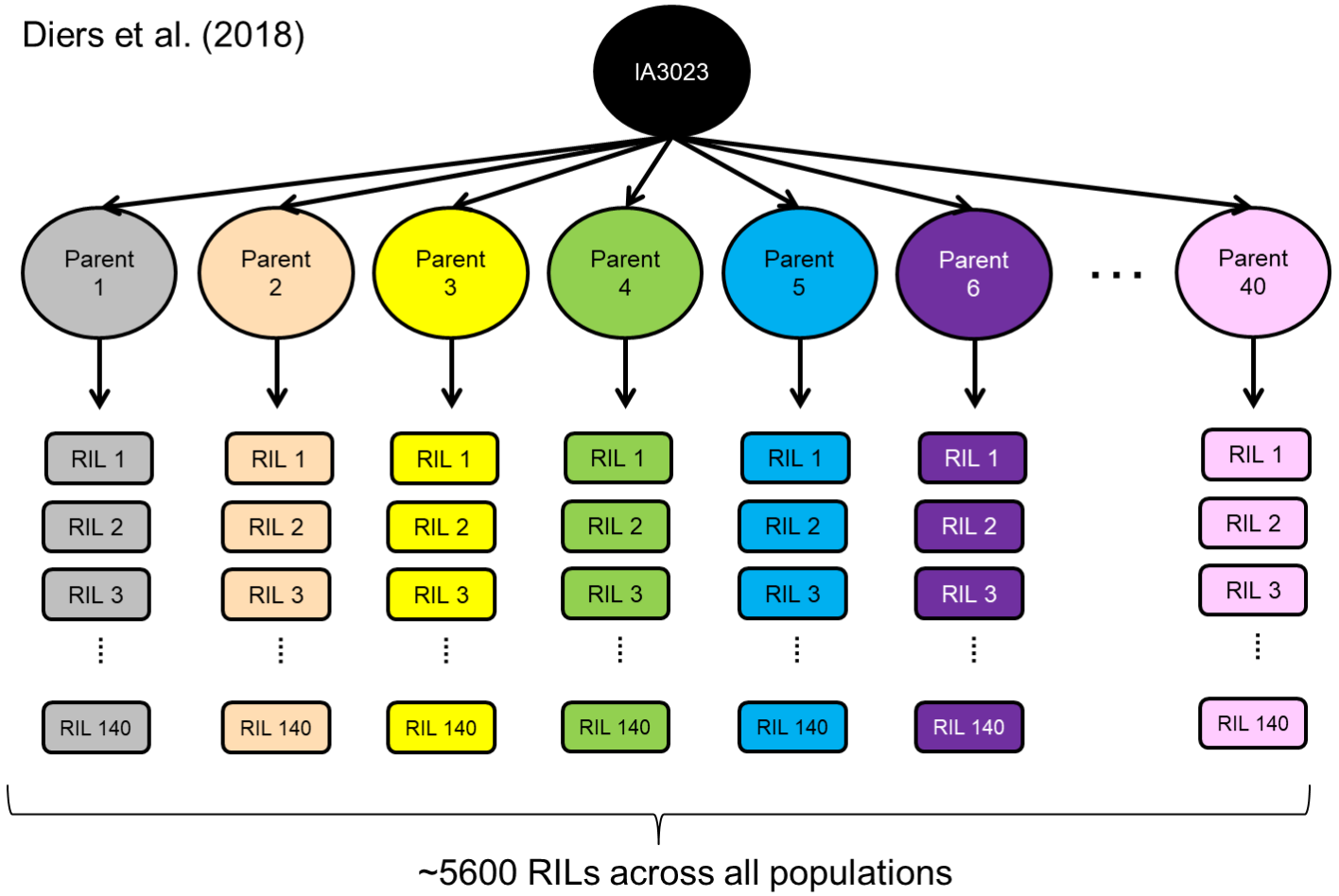
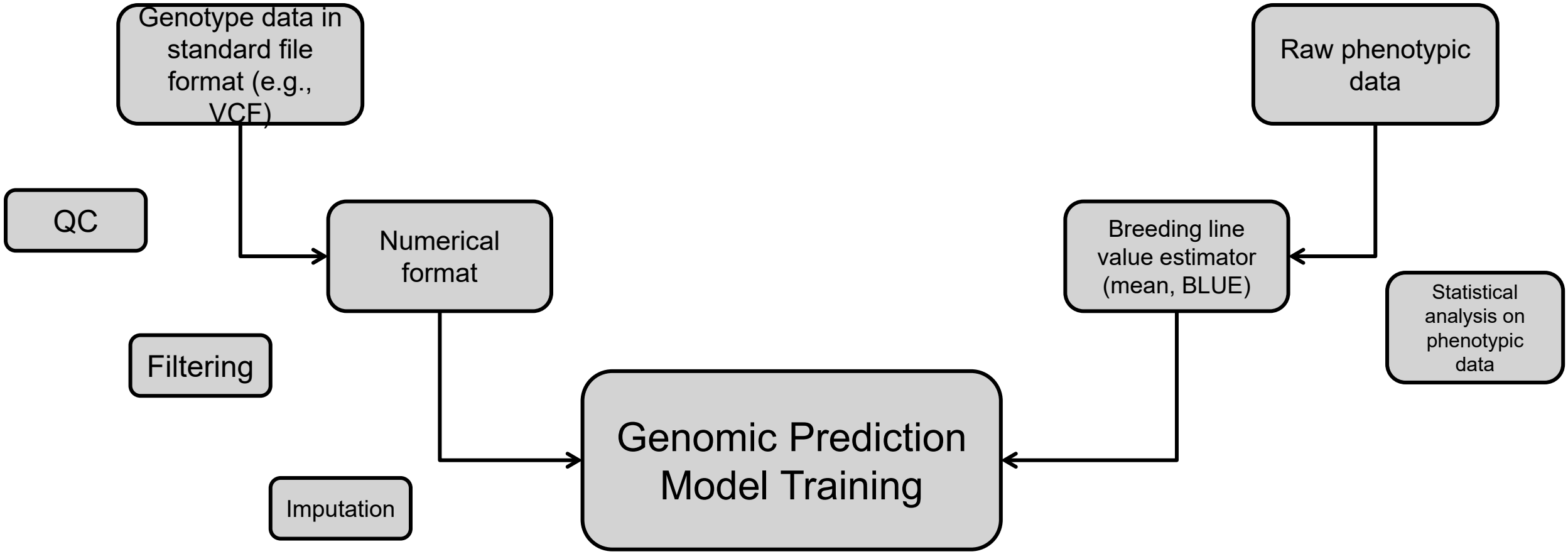


Table 1 Founders of the 40 NAM families, their origin and group. For more information and photos, see: https://soybase.org/SoyNAM/imagebrowser.php

NAM Family	Parent	Origin	Group*
Hub	IA3023	Iowa State Univ.	Common parent
N02	TN05-3027	Univ. of Tenn.	EL
N03	4J105-3-4	Purdue Univ.	EL
N04	5M20-2-5-2	Purdue Univ.	EL
N05	CL0J095-4-6	Purdue Univ.	EL
N06	CL0J173-6-8	Purdue Univ.	EL
N08	HS6-3976	Ohio State Univ.	EL
N9	Prohio	USDA-ARS, Wooster, OH	EL
N10	LD00-3309	Univ. of Illinois	EL
N11	LD01-5907	Univ. of Illinois	EL
N12	LD02-4485	Univ. of Illinois	EL
N13	LD02-9050	Univ. of Illinois	EL
N14	Magellan	Univ. of Missouri	EL
N15	Maverick	Univ. of Missouri	EL
N17	S06-13640	Univ. of Missouri	EL
N18	NE3001	Univ. of Nebraska	EL
N22	Skylla	Mich. State Univ.	EL
N23	U03-100612	Univ. of Nebraska	EL
N24	LG03-2979	USDA-ARS, Urbana, IL	BX
N25	LG03-3191	USDA-ARS, Urbana, IL	BX
N26	LG04-4717	USDA-ARS, Urbana, IL	BX
N27	LG05-4292	USDA-ARS, Urbana, IL	BX
N28	LG05-4317	USDA-ARS, Urbana, IL	BX
N29	LG05-4464	USDA-ARS, Urbana, IL	BX
N30	LG05-4832	USDA-ARS, Urbana, IL	BX
N31	LG90-2550	USDA-ARS, Urbana, IL	BX
N32	LG92-1255	USDA-ARS, Urbana, IL	BX
N33	LG94-1128	USDA-ARS, Urbana, IL	BX
N34	LG94-1906	USDA-ARS, Urbana, IL	BX
N36	LG97-7012	USDA-ARS, Urbana, IL	BX
N37	LG98-1605	USDA-ARS, Urbana, IL	BX
N38	LG00-3372	USDA-ARS, Urbana, IL	BX
N39	LG04-6000	USDA-ARS, Urbana, IL	BX
N40	PI 398.881	South Korea	PI
N41	PI 427136	South Korea	PI
N42	PI 437169B	Russia	PI
N46	PI 507681B	Uzbekistan	PI
N48	PI 518751	Serbia	PI
N50	PI 561370	China	PI
N54	PI 404188A	China	PI
N64	PI 574486	China	PI

*Founder group designations are EL = Elite, BX = breeding lines with exotic ancestry, and PI = plant introduction.


Genotype data in standard file format (e.g., VCF)
QC
Numerical format
Filtering
Imputation
Genomic Prediction Model Training
Raw phenotypic data
Breeding line value estimator (mean, BLUE)
Statistical analysis on phenotypic data

Standardized genotype data file

- VCF (Variant Call Format)
  - Very common format for storing DNA sequence polymorphisms.
  - Developed as part of the “1000 Genomes Project” and maintained by the Global Alliance for Genomics & Health group

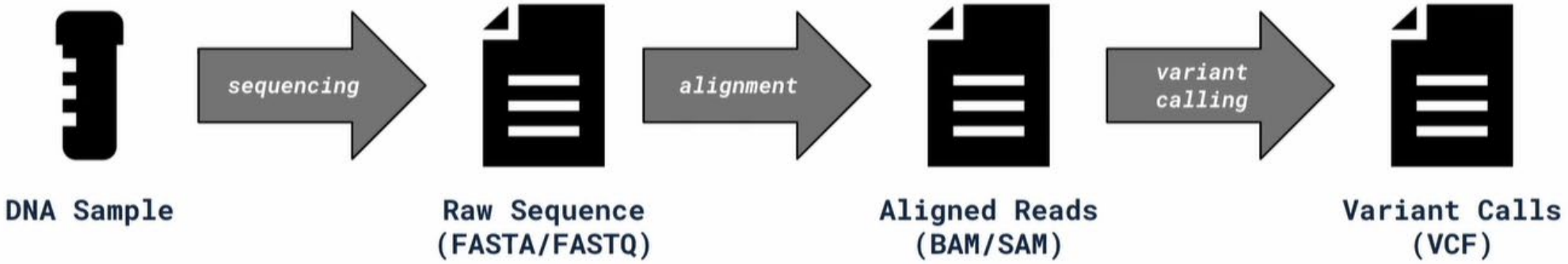
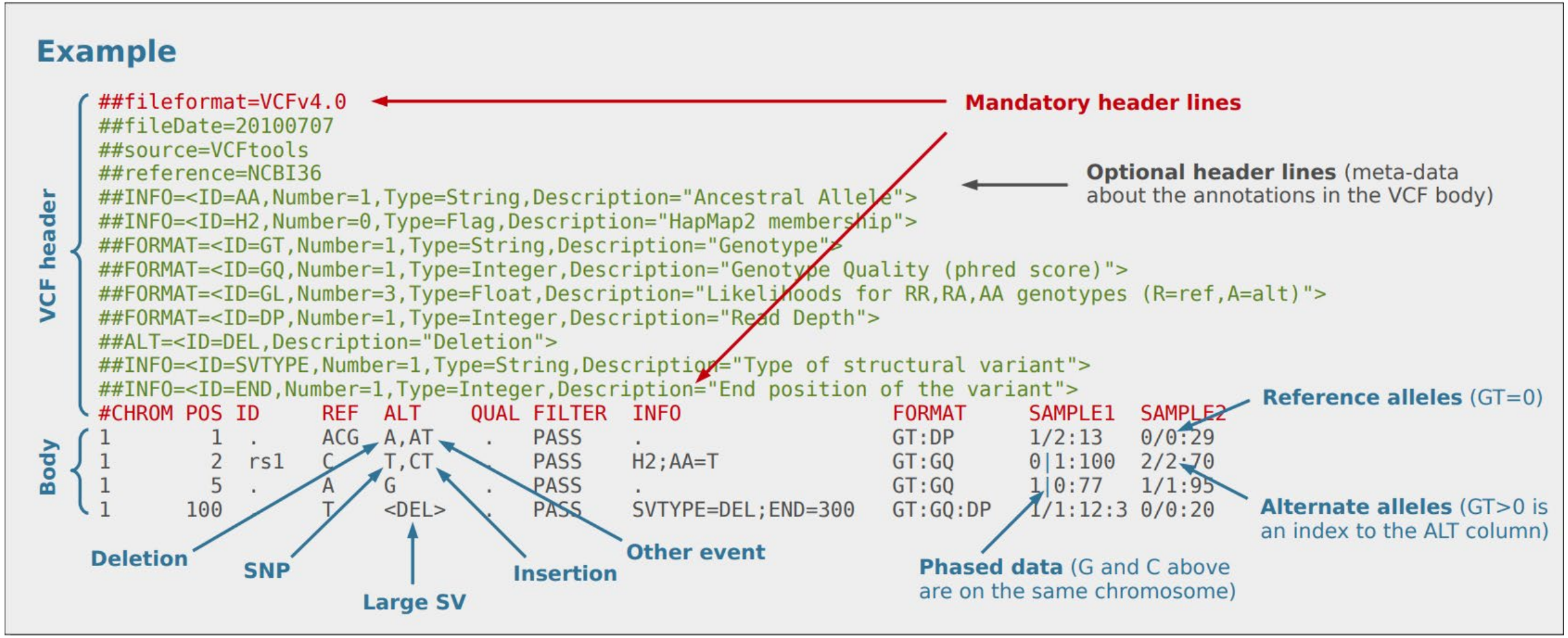


Image credit: GenomOncology

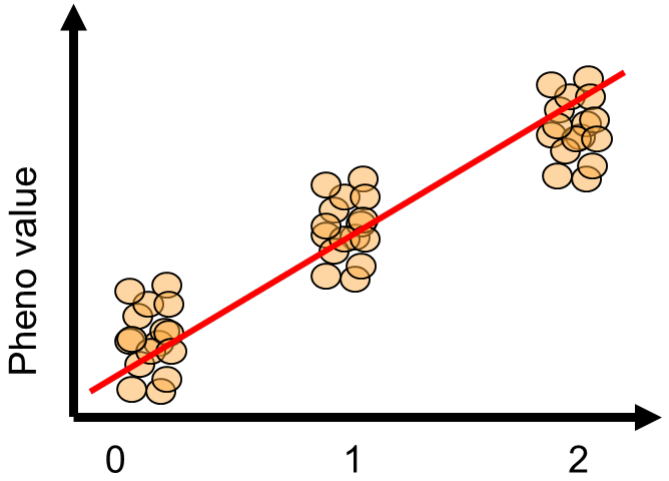
VCF files



http://vcftools.sourceforge.net/VCF-poster.pdf

Convert text strings in VCF file to numbers

- Genotypes need to be scored quantitatively so that they can be fit into statistical models and used to calculate genetic similarities.



$$y_i = u_i + \varepsilon_i$$

$$\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$$

Realized genomic relationship between individuals 2 and 1

$$\mathbf{G} = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nn} \end{bmatrix}$$

Filtering out markers and individuals based on missing data

- Markers with a high proportion of missing data are not informative
  - Can reduce prediction accuracy if missing data rate is very high (>40%).
  - Imputation (discussed later) can help but accuracy of imputation on markers with high rates of missing data is low.
  - Including lots of markers with high rates of missing data unnecessarily increases computational burden.

- Individuals with high rate of missing data will not be predicted well – not very much information – and will be compressed towards mean.
  - Also, indication of possible technical error in genotyping, producing inaccurate genotype scores.

Removing markers with low minor-allele frequency

Necessary?

Original idea of removing low-frequency variants was based on desire to minimize false positives, but findings have been mixed (see Tabangin et al., 2009).

The MAF cutoff depends on population size, and type of population.

For biparental populations, MAF cutoffs can be used to retain only polymorphic markers.

For broad-based populations or diversity panels, ensuring at least 20-30 individuals carry the allele helps ensure accurate allelic effect estimates.

If only a small number (1-5) individuals carry the allele, allelic effect estimates susceptible to outliers.

Situation not the same in genomic prediction as in GWAS because shrinkage is applied to effects.

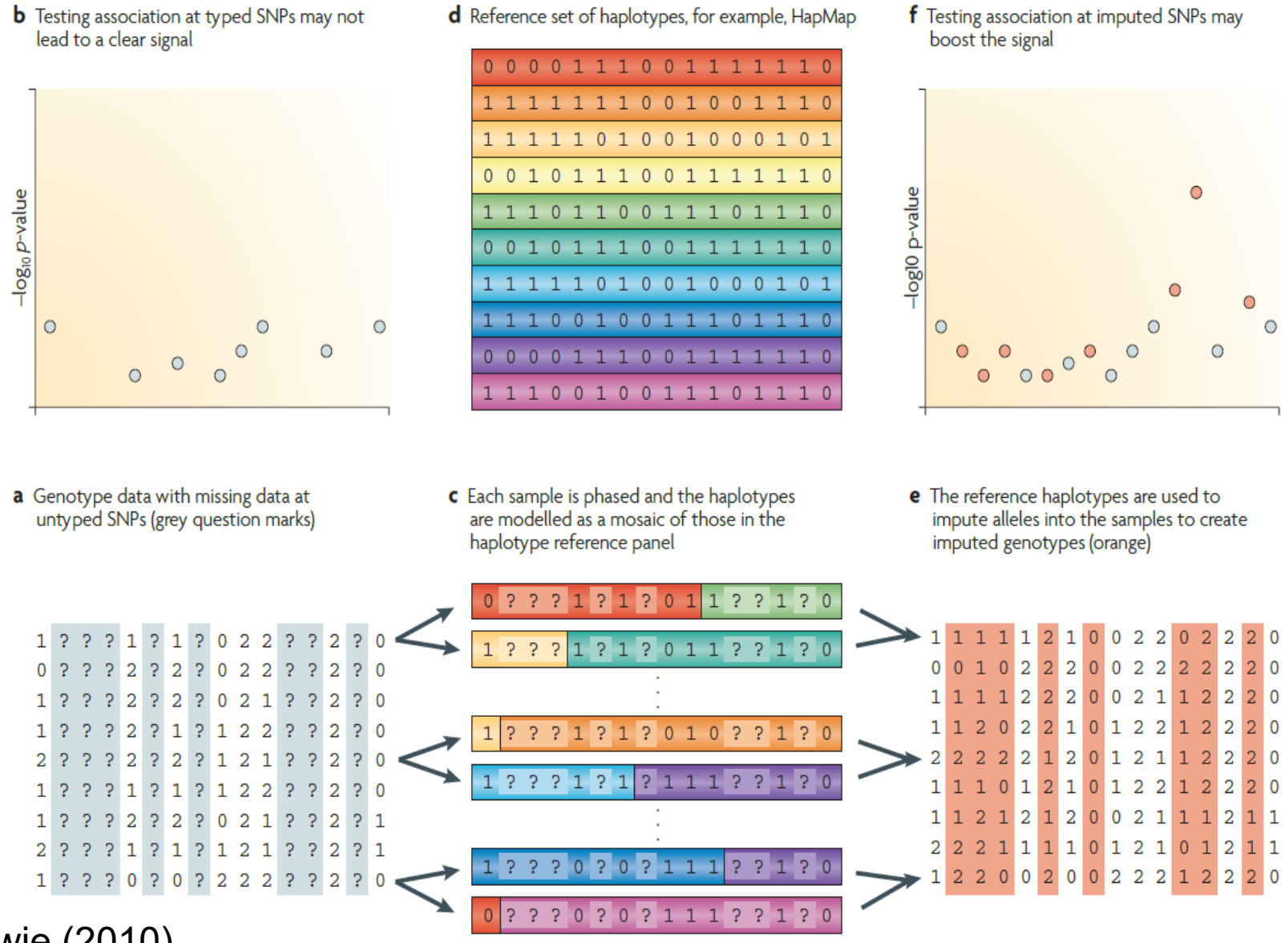
Genotype imputation

- Defn of imputation (as used in statistics): Replace missing data with substituted values.
- For purposes of genomic prediction and QTL mapping, imputation can either…
  - Add previously missing information to increase statistical power
  - Provide numerical values in data matrix so it can be used in calculations

“Naïve” imputation

- Imputation based on allele frequencies in the unimputed data.
- Extreme example: If freq(1) = 1 and freq(0) = 0 in the unimputed data, all missing marker scores will be imputed as “1”.
  - If freq(1) = freq(0) = 0.5, imputed values will be 0.5.
- If genotypes coded as X = {0, 0.5, 1}, mean(X_j) = mean(X_1, X_2, X_3, …, X_n) = freq(1)
- Essentially, by imputing missing genotype scores to the mean genotype score, we’re taking our best guess based on only the data at that locus, not considering genotype scores at surrounding loci.

Imputation through leveraging a reference panel

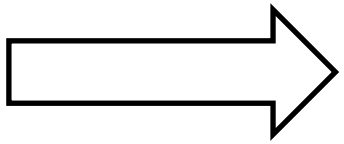


Marchini and Howie (2010)

Imputation through haplotypes in unimputed data

2	0	0	2	2	2
2	?	2	0	?	0
2	?	2	0	2	2
0	2	2	?	0	0
2	?	0	0	0	0
2	2	2	?	?	2

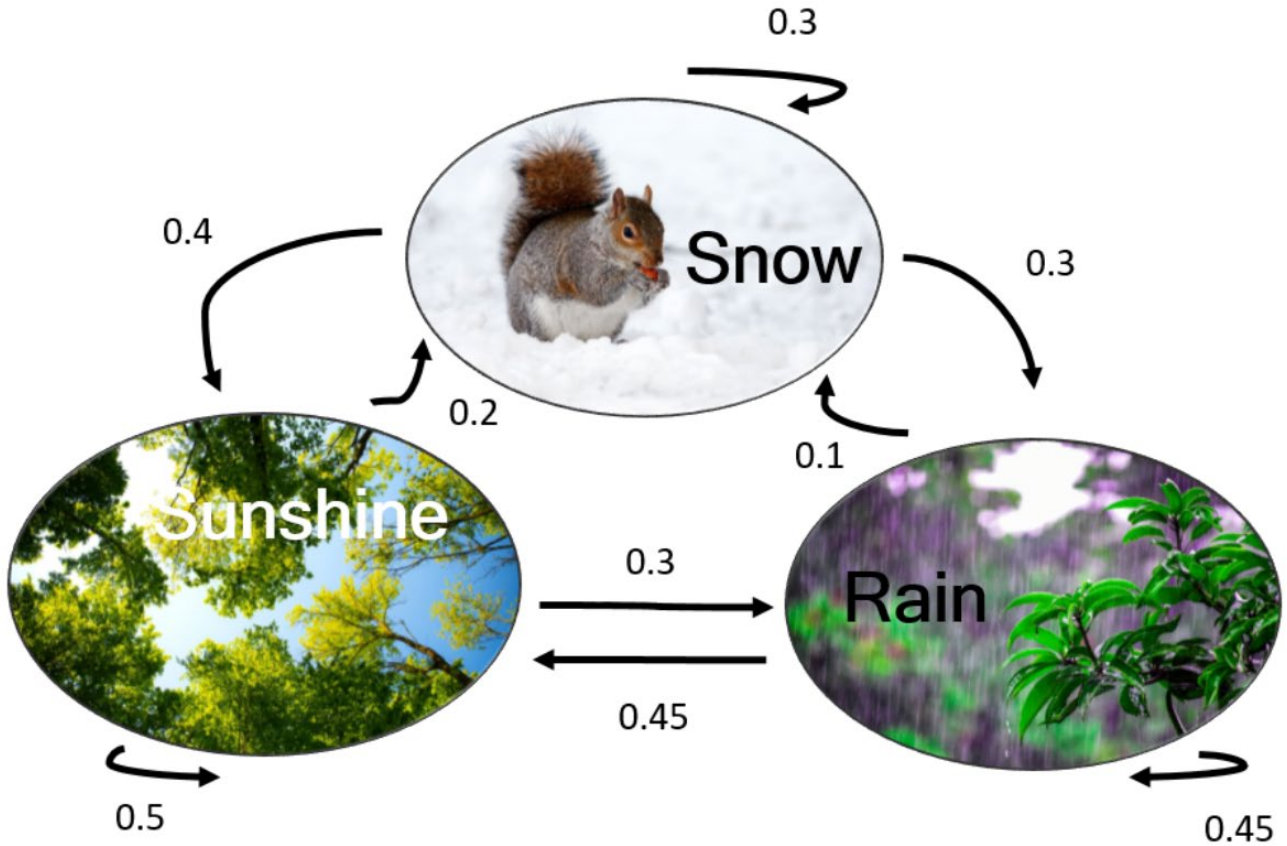
Hidden Markov model



2	0	0	2	2	2
2	2	2	0	0	0
2	2	2	0	2	2
0	2	2	0	0	0
2	0	0	0	0	0
2	2	2	0	2	2

Hidden Markov model (very high-level overview)

- Probabilistic way of imputing genotypes based on surrounding marker states.
- The probability of being in a certain state depends on the prior state.
- Applied to markers, the probability a missing value is truly a certain allele depends on the surrounding marker genotypes.



https://towardsdatascience.com/introduction-to-hidden-markov-models-cd2c93e6b781

What determines your accuracy of imputation? (non-exhaustive list)

- **Allele frequency**
  - It is much harder to accurately impute low frequency alleles (esp. rare alleles) compared to common alleles
- **Effective population size**
  - Function of the diversity of the population
  - Diverse populations have faster LD decay (shorter shared haplotype blocks), thus making it more difficult to impute using surrounding marker information
- **Reference panel size and its relationship to target population**
  - A larger reference panel will capture more low frequency or rare alleles

Fit an RR-BLUP model using rrBLUP package

Published November, 2011

ORIGINAL RESEARCH

Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP

Jeffrey B. Endelman*

Link to article

Ridge regression BLUP (RR-BLUP): Convenient way of estimating genome-wide marker effects

Interested in an estimate of the overall genetic value: $\hat{a} = \sum_{i=1}^{n} \hat{\beta}_i X_i$

Least-squares estimators

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Number of predictors (*p*) cannot exceed number of observations (*n*)

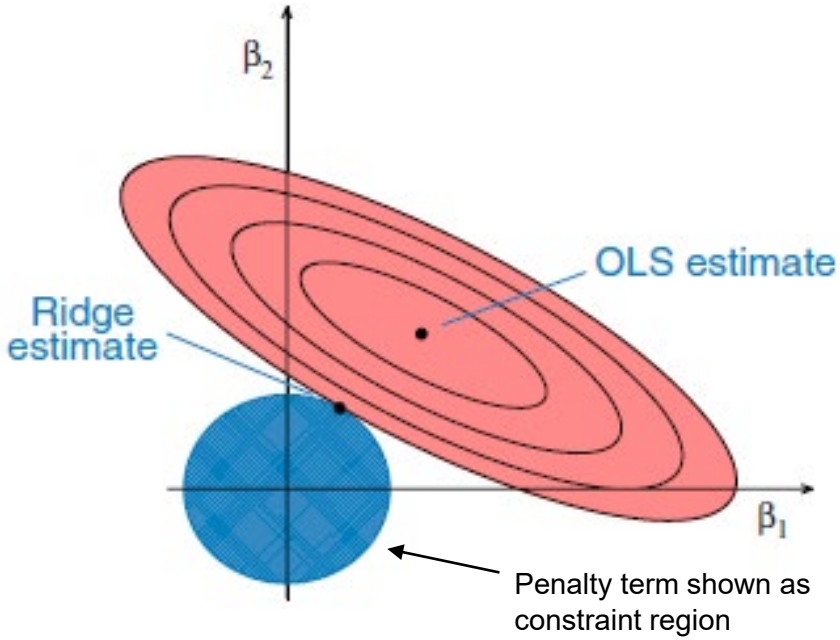
- $X^T X$ becomes singular. Cannot solve equation
- Even if p < n, but p approaches n, variance of $\hat{\beta}$ high.

RR-BLUP estimators

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$$

Where $\lambda = \sigma_e^2 / \sigma_\beta^2$. Addition of λI term reduces collinearity and prevents $X^T X$ from becoming singular.

- Originally, ridge regression used grid search to find optimal λ.
- When $\lambda = \sigma_e^2 / \sigma_\beta^2$ is used, it has become known as ridge regression BLUP (RR-BLUP).



Ridge regression best linear unbiased prediction
RR-BLUP

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

$\lambda = \sigma_e^2 / \sigma_\beta^2$

Where σ_e^2 is the residual variance and σ_β^2 is the variance accounted for by the markers

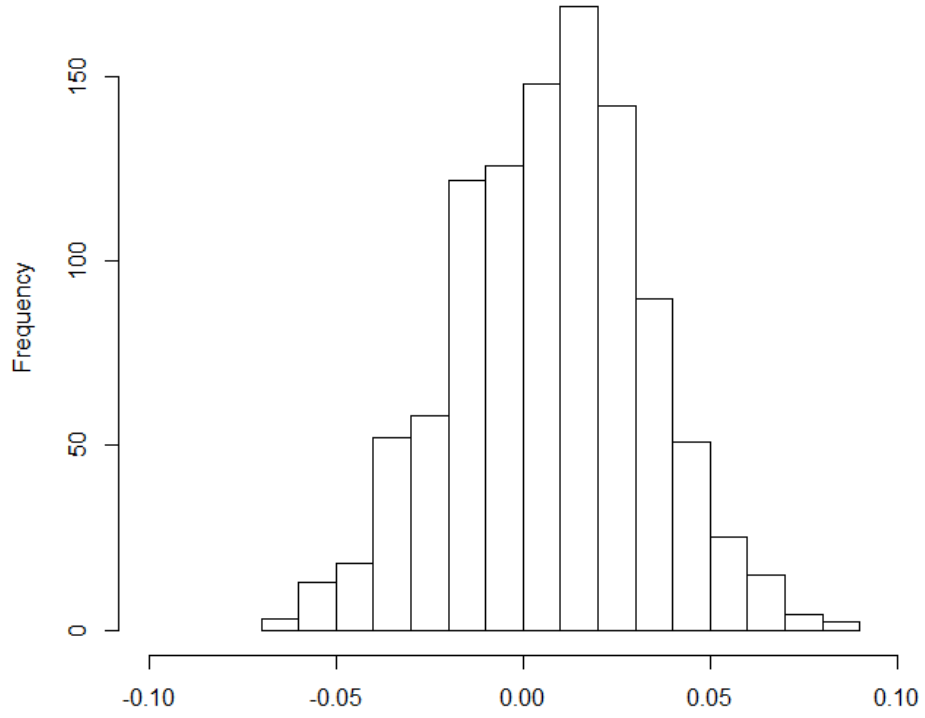
As $\lambda = \sigma_e^2 / \sigma_\beta^2$ increases, marker effect predictions ($\hat{\beta}$) will become more shrunken (regressed) towards 0.

Comparison to least-squares, MAS approach

Least-squares

Marker	Effect estimate (fixed)
umc1346	1.68
ufg17	1.65
umc1650	-1.90
umc1147	-1.29
umc1658	2.09
mmp102	1.86
asg27a	0.99
umc1926	1.43
bnlg1018	1.31
csu48	1.54
umc1608	1.27
umc1028	-1.92
csu324b	-0.96

RR-BLUP



Level of shrinkage of effects applied proportional to error/signal ratio

I.e., $\lambda = \sigma_e^2 / \sigma_\beta^2$

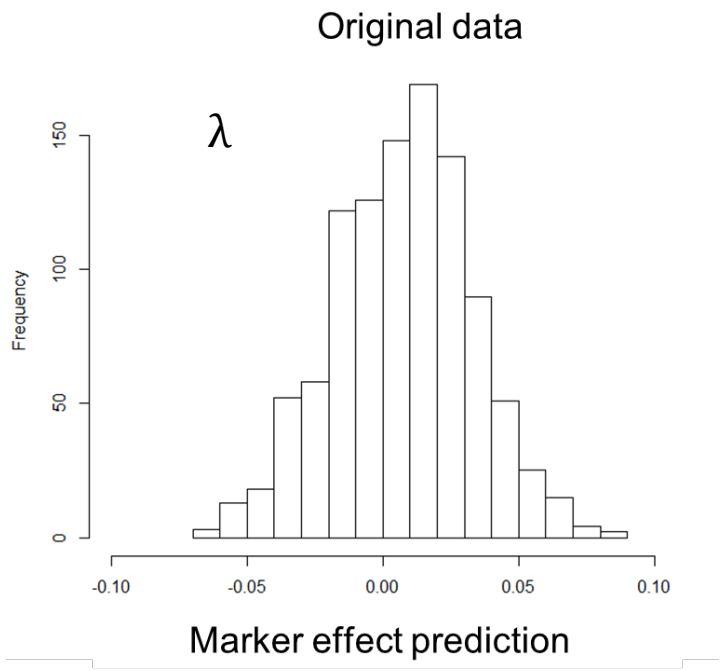
The foregoing parameterization essentially fits *β* as a random effect.

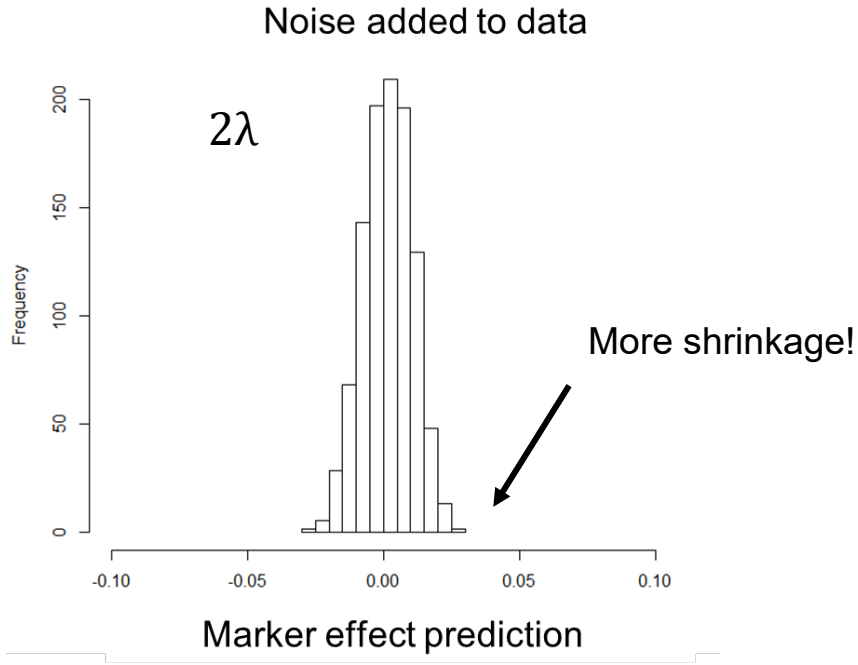
Introduces simplifying assumption that all marker effects are sampled from common distribution of equal variance.

Only one degree of freedom required to estimate variance of this distribution.

This does not mean that all marker effects are the same, only that they have a common variance scaled by λ.

Degree of scaling is called “shrinkage”

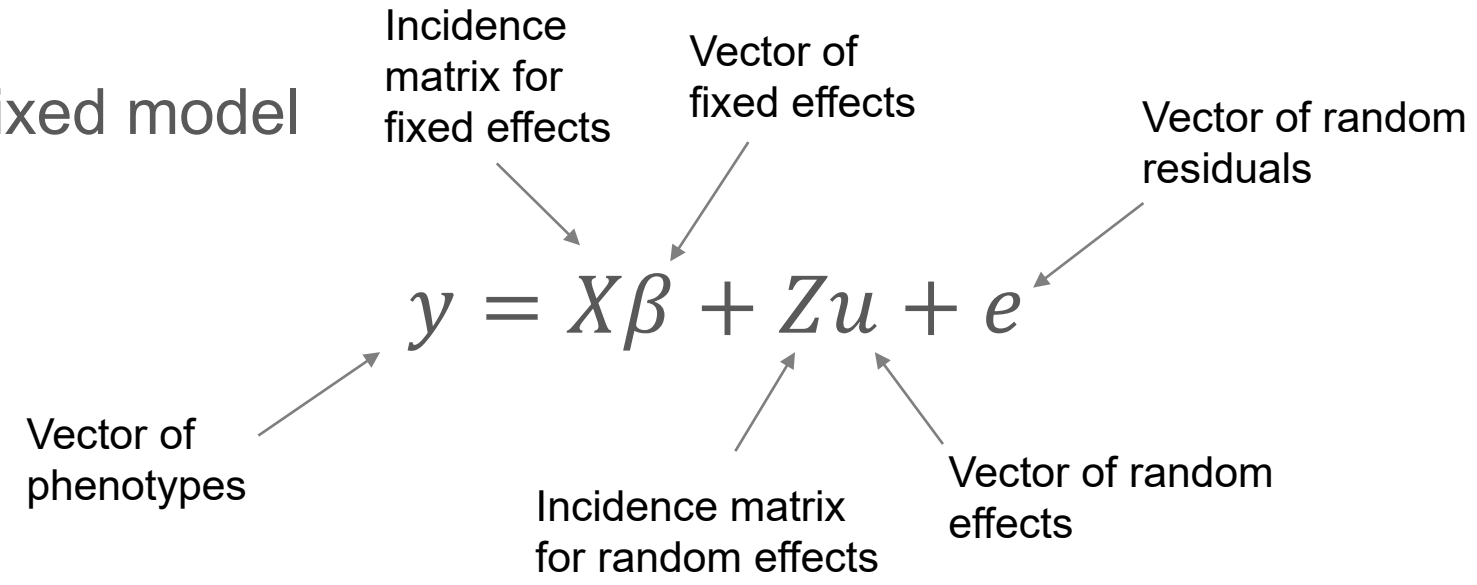


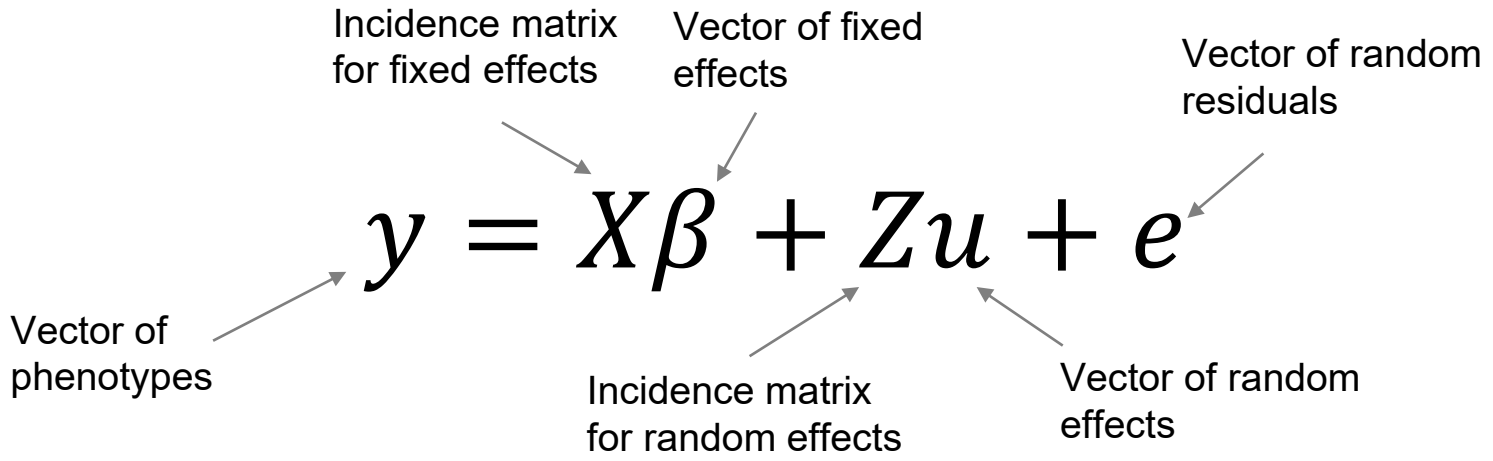


Genomic best linear unbiased prediction (G-BLUP)

- Similar to traditional BLUP with pedigrees in a mixed model
- Pedigree relationship matrix is substituted with genomic relationship matrix
- Use genomic relationships in mixed-linear model to predict breeding value of relatives

- General mixed model





Random effects are assumed to be drawn from some underlying probability distribution and thus can be assigned a covariance structure.

Here, it is normally assumed that $u \sim MVN(0, G)$ where G describes the covariances among random effects.

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Information between relatives being shared through G matrix

G-BLUP leverages sharing of information between relatives according to their realized relationships as estimated by markers

$$y_i = u_i + \varepsilon_i$$

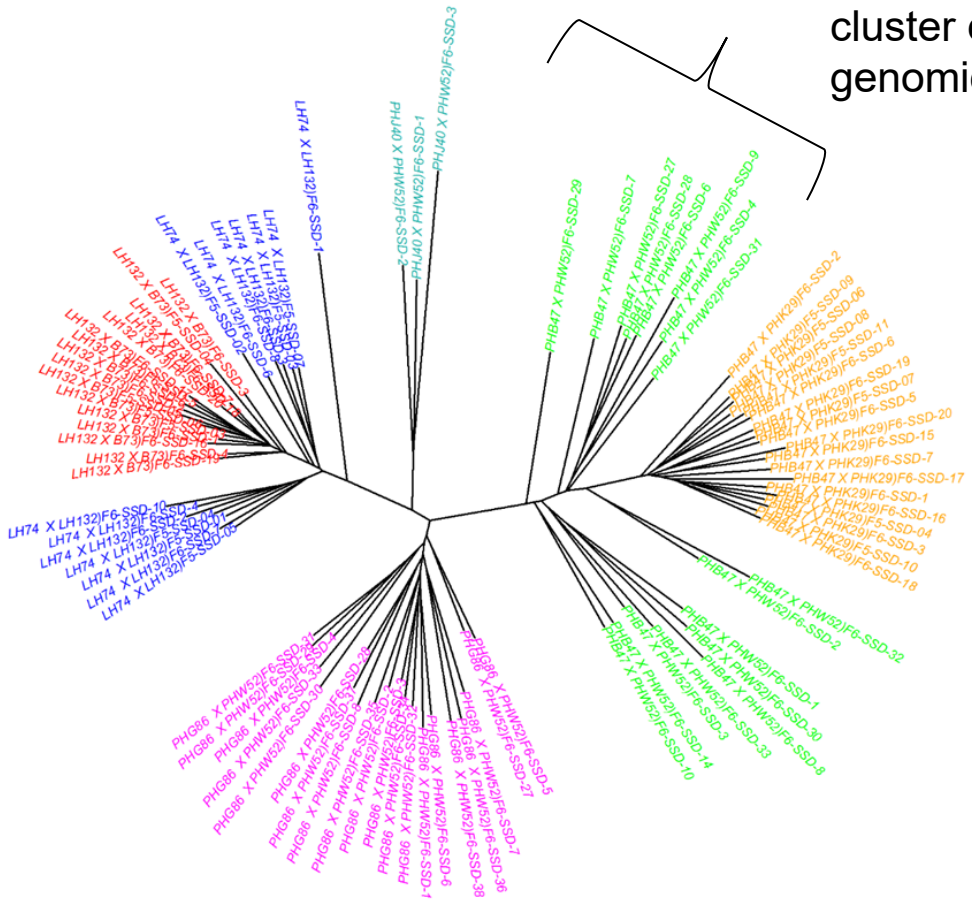
$$\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$$

Realized genomic relationship between individuals 2 and 1

$$\mathbf{G} = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1n} \\ G_{21} & G_{22} & \cdots & G_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nn} \end{bmatrix}$$

Sharing of information between relatives through the genomic relationship matrix to predict values for quantitative traits

Individuals from same family (i.e., same pedigree) cluster differentially because of different realized genomic relationships

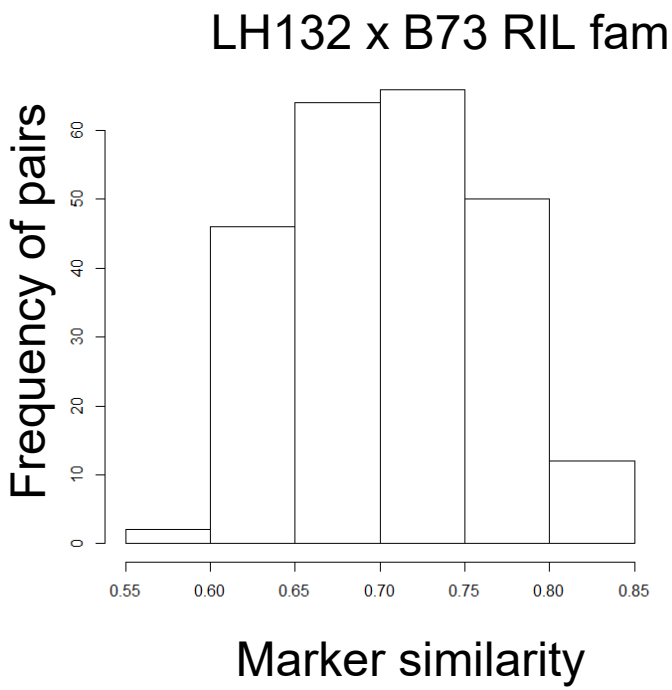


Neighbor joining tree representation of realized genomic relationships

Spectrum of resemblance among relatives for polygenic traits

Mendelian sampling term causes deviations from expected resemblance

- Markers can capture realized relationships which pedigree cannot



Equivalency between RR-BLUP and G-BLUP

If….

- The number of QTL is large
- No major effect QTL exist
- QTL effects are evenly distributed through genome

Then….RR-BLUP and G-BLUP are equivalent

Arises from properties of multivariate normal distribution, which is the distribution underlying G-BLUP.

$$\mathbf{y} = \mu + \sum_k \mathbf{x}_k \beta_k + \mathbf{e} \qquad \beta_k \sim N(0, \sigma^2_\beta)$$

$$\mathbf{u} = \sum_k \mathbf{x}_k \beta_k = \mathbf{X}\boldsymbol{\beta}$$

From MVN distribution properties:

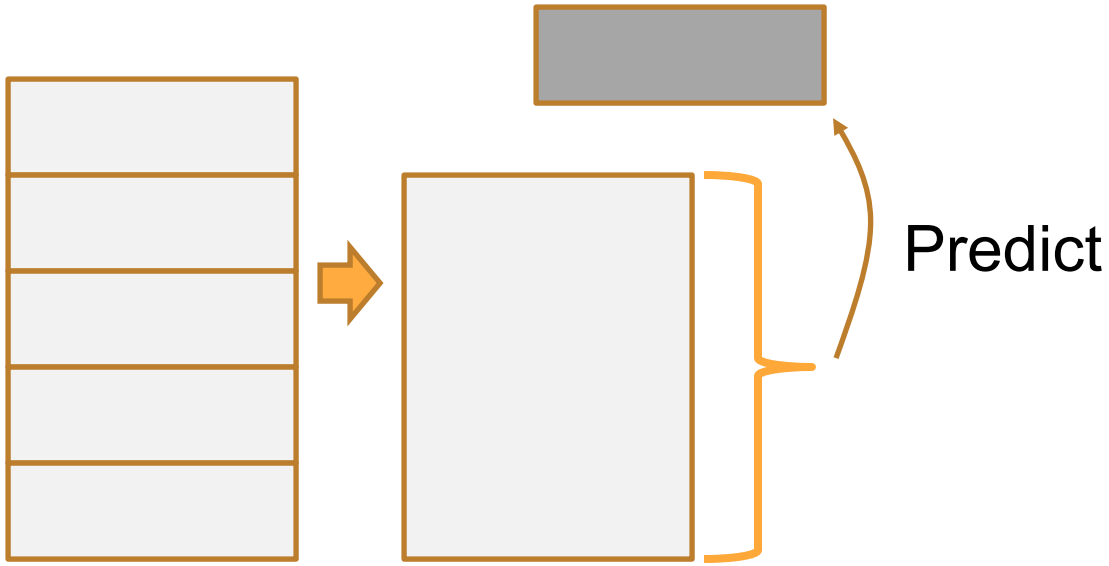
$$\mathrm{var}(\mathbf{u}) = \mathbf{X}\mathbf{X}^T\sigma^2_\beta = \mathbf{G}\sigma^2_u$$

$$\mathbf{G} \propto \mathbf{X}\mathbf{X}^T$$

Only valid with the normal prior!

Assessing success

- Calculate prediction accuracy using independent validation

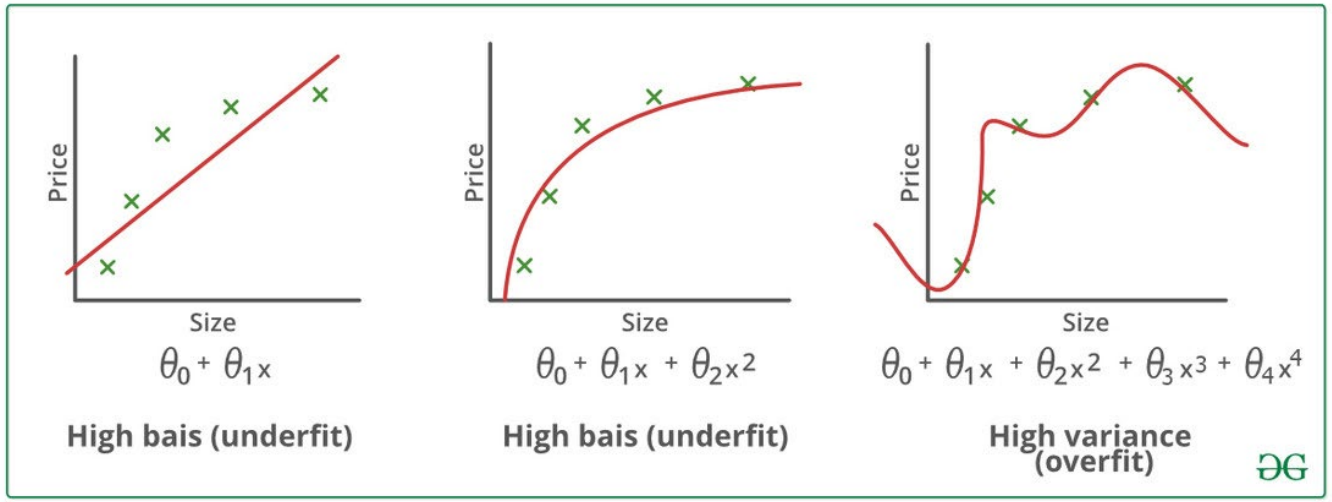


Why do we do this?

Model overfitting

- A model that is overfit includes more parameters than can be justified by the data. Basically, you run the risk of fitting “noise”, or residual variation unknowingly as if that variation can be represented by the model structure.

- Results in a model that is very good at explaining variation in the current (training) dataset, but poor at predicting future observations.



Source: https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Assessing success

- Calculate prediction accuracy using independent validation