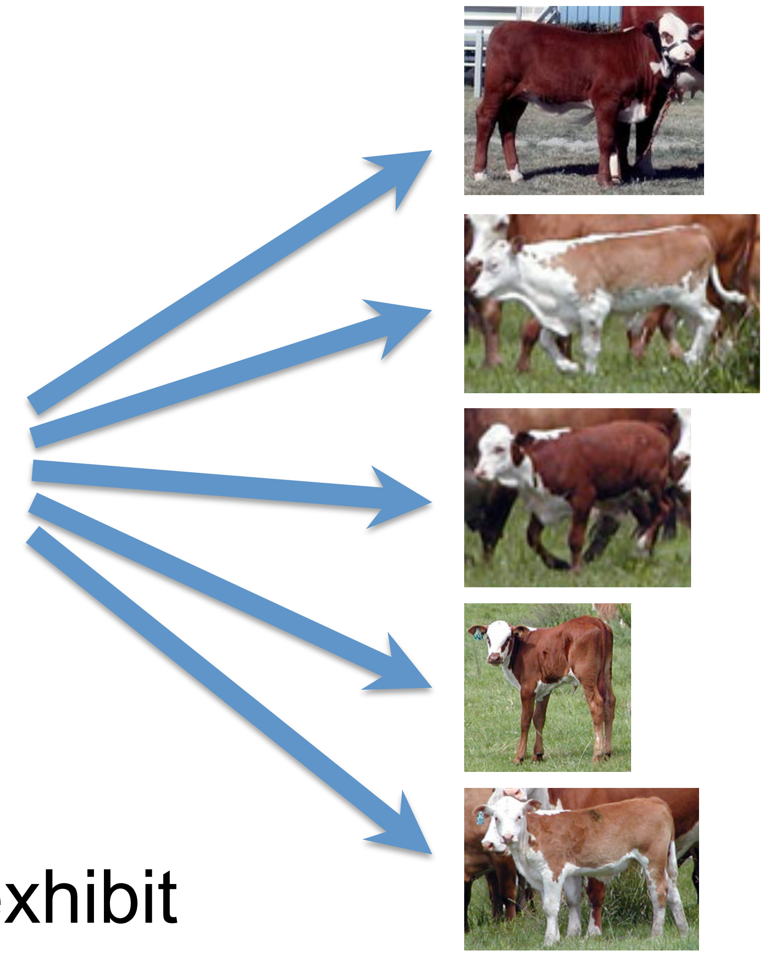
Implementation of genomic prediction in routine genetic evaluations: state of the art in different species, pitfalls, future developments

dorian@iastate.edu

Performance of the Progeny

Sire



+30 kg

+15 kg

-10 kg

+ 5 kg

+10 kg

Progeny +10 kg

Offspring of one sire exhibit more than ¾ diversity of the entire population

We learn about parents from progeny

Sire

+30 kg

+15 kg

-10 kg

+ 5 kg

+10 kg

(EBV is "shrunk")

Sire EBV +16-18 kg ← Progeny +10 kg

<2x progeny difference

Pedigree Prediction

$$y = Xb + Zu + e$$

Single trait mixed effects linear model

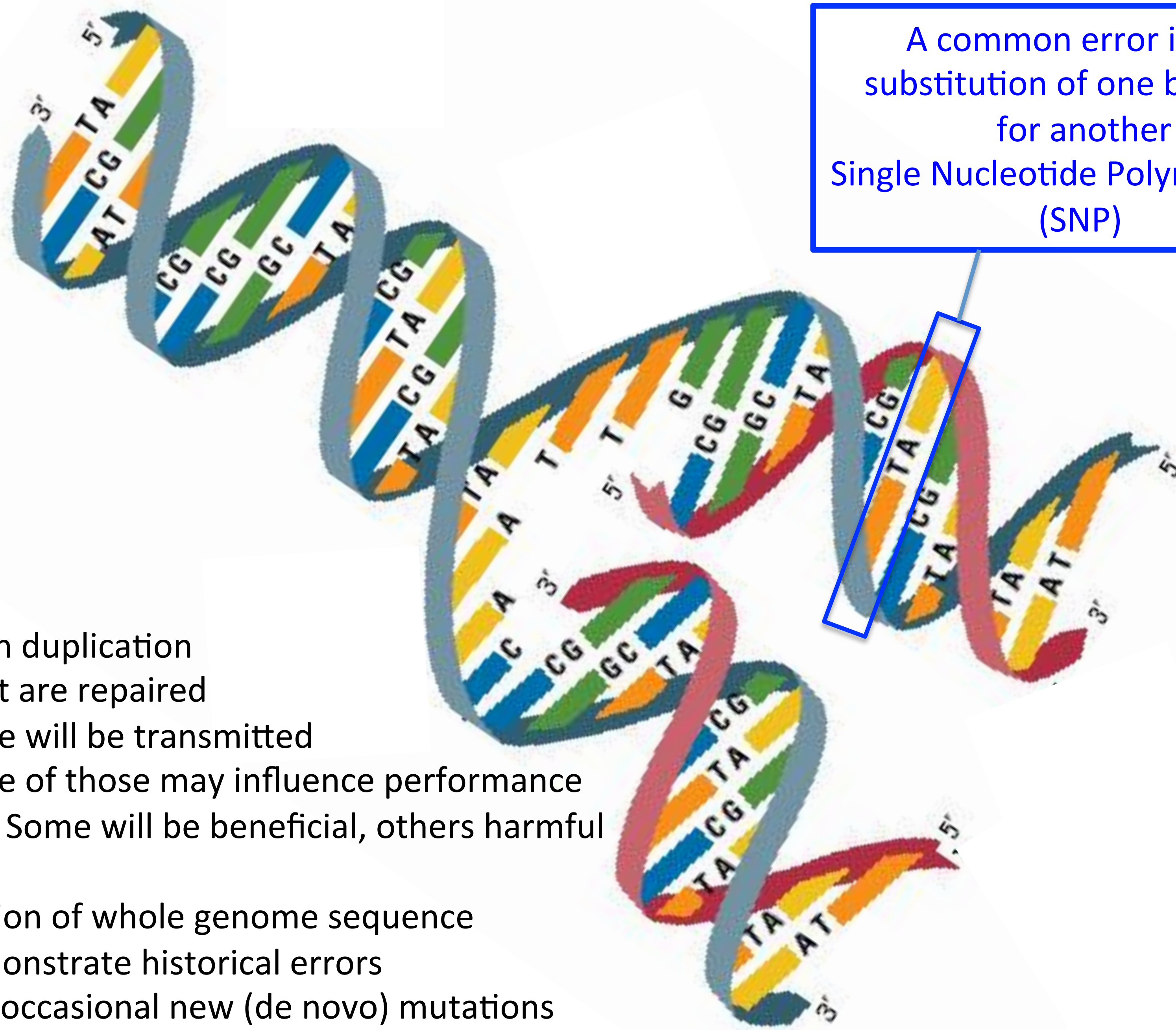
$$var(u) = G = A\sigma_g^2 \qquad var(e) = R = I\sigma_e^2$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

A = pedigree based numerator relationship matrix

$$\lambda = \sigma_e^2 / \sigma_g^2$$

Henderson 1949 (Phd), Henderson et al, 1959 Biometrics 15:192

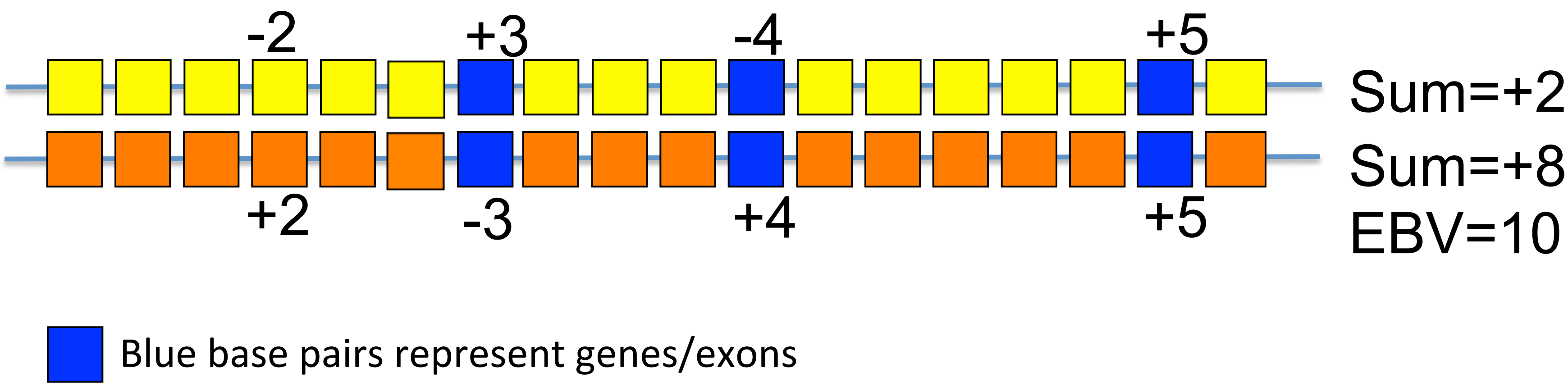


A common error is the substitution of one base pair for another
Single Nucleotide Polymorphism (SNP)

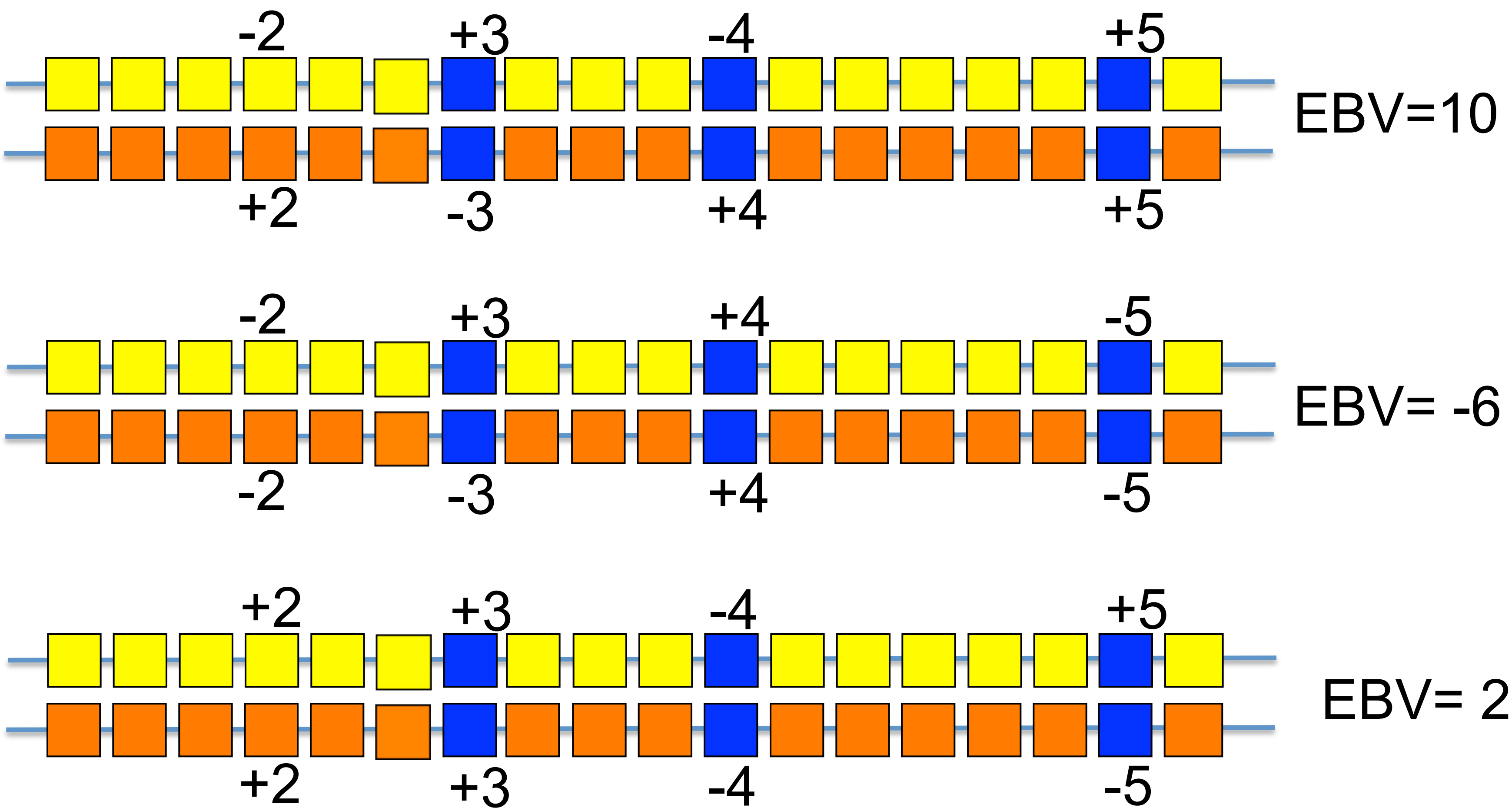
Errors in duplication
- Most are repaired
- Some will be transmitted
- Some of those may influence performance
    - Some will be beneficial, others harmful

Inspection of whole genome sequence
- Demonstrate historical errors
- And occasional new (de novo) mutations

Breeding Merit is sum of average gene effects

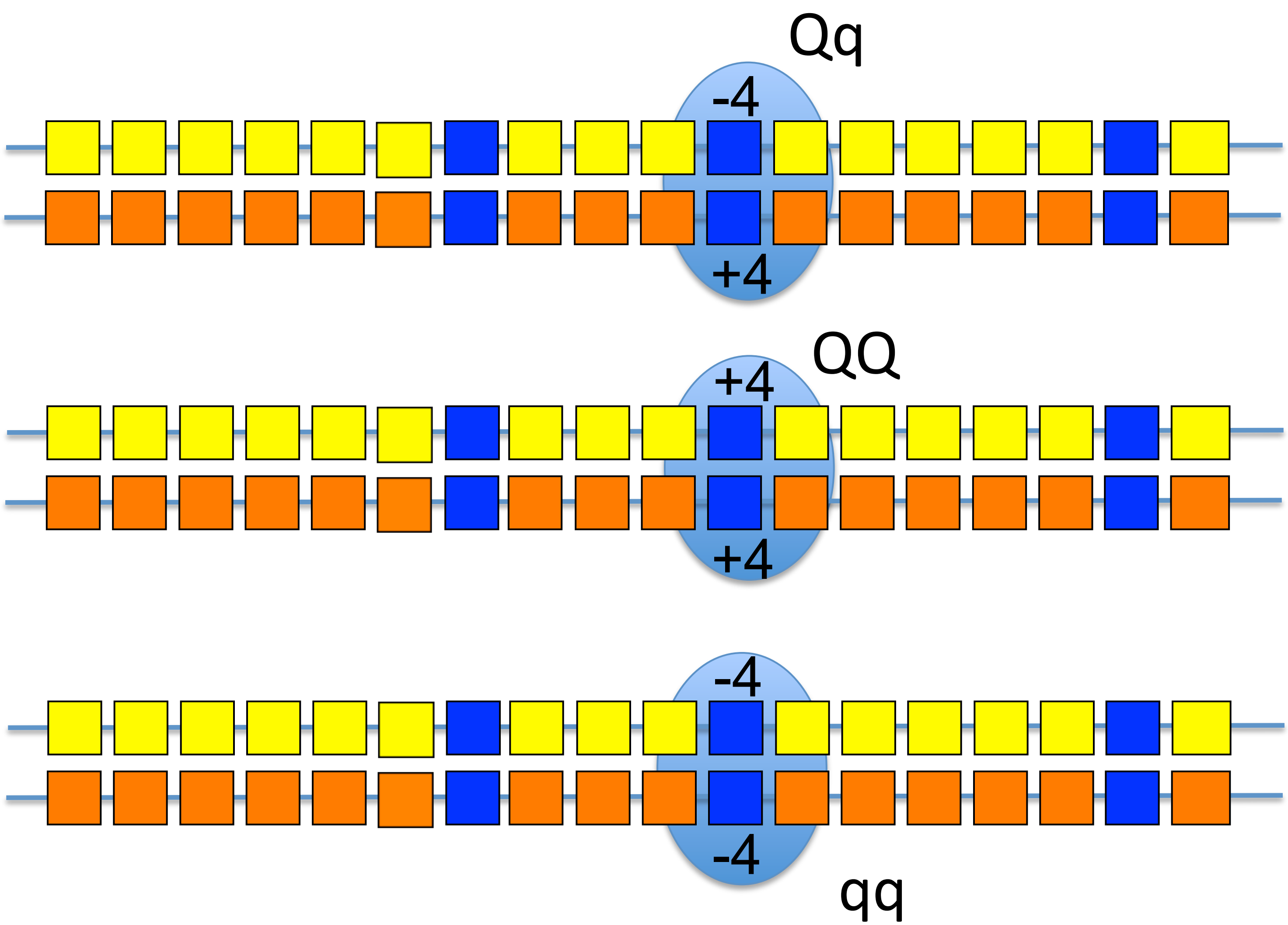


Consider 3 Bulls



Below-average bulls will have some above-average alleles and vice versa!

At any 1 locus there are 3 genotypes



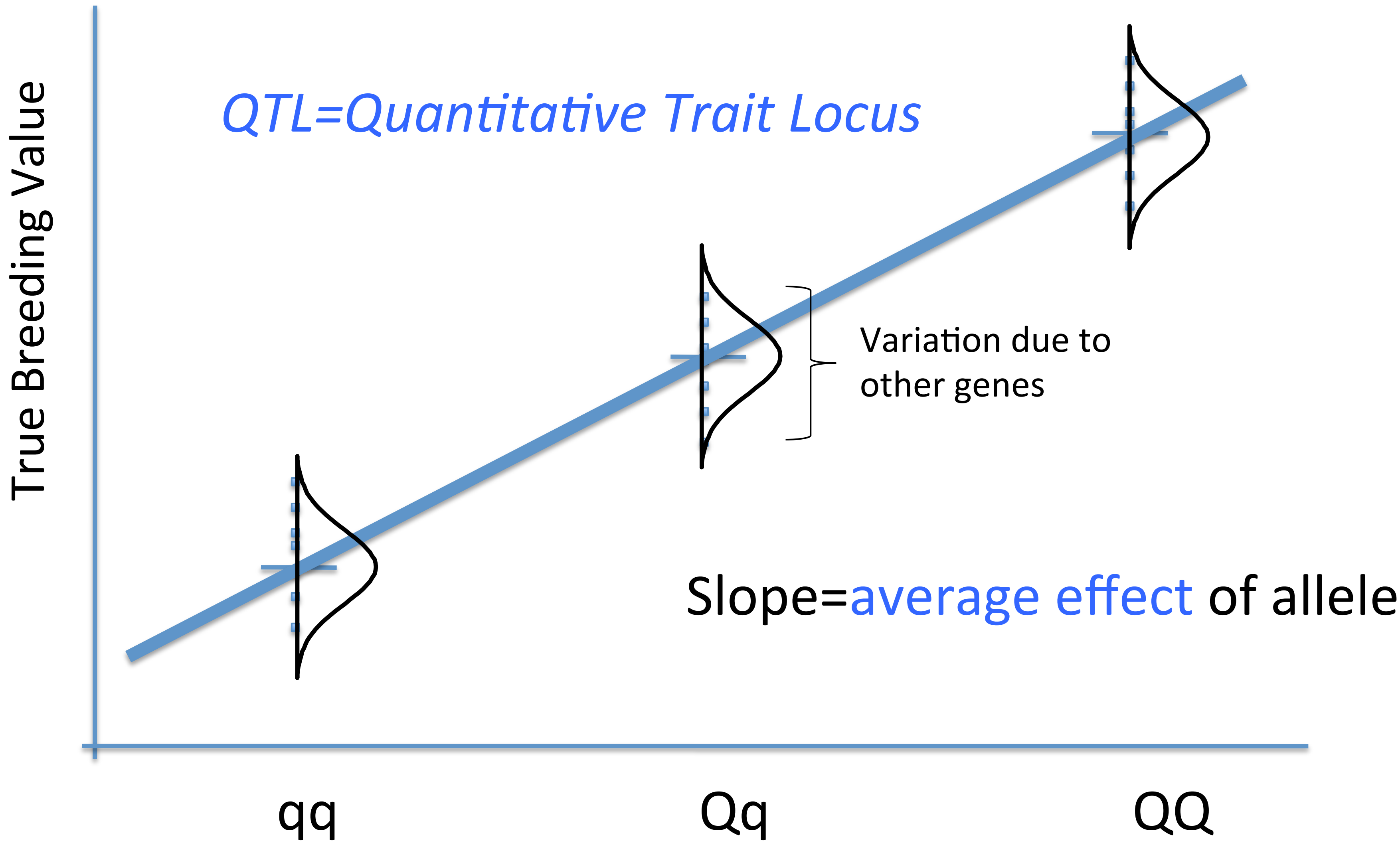
Contribution of this QTL

0

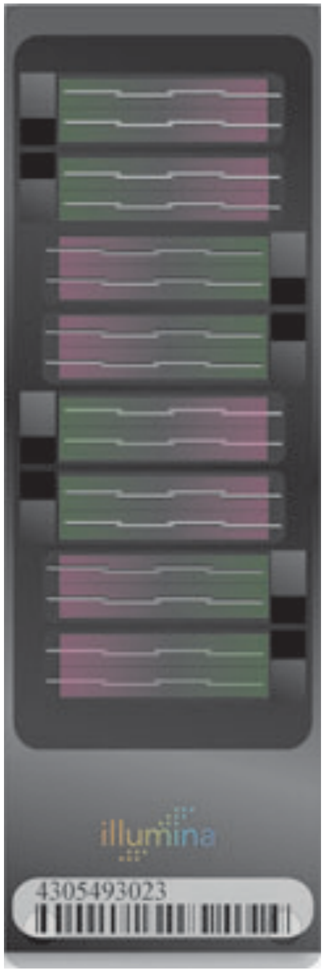
+8

-8

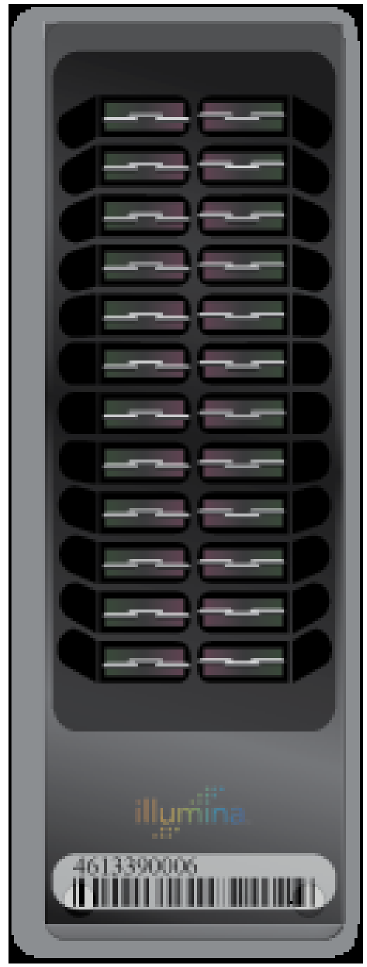
Regress BV on QTL genotype



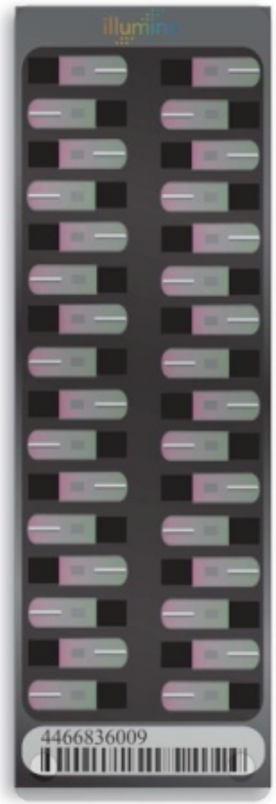
Illumina Bovine 770k, 50k (v2), 3k



700k (HD)

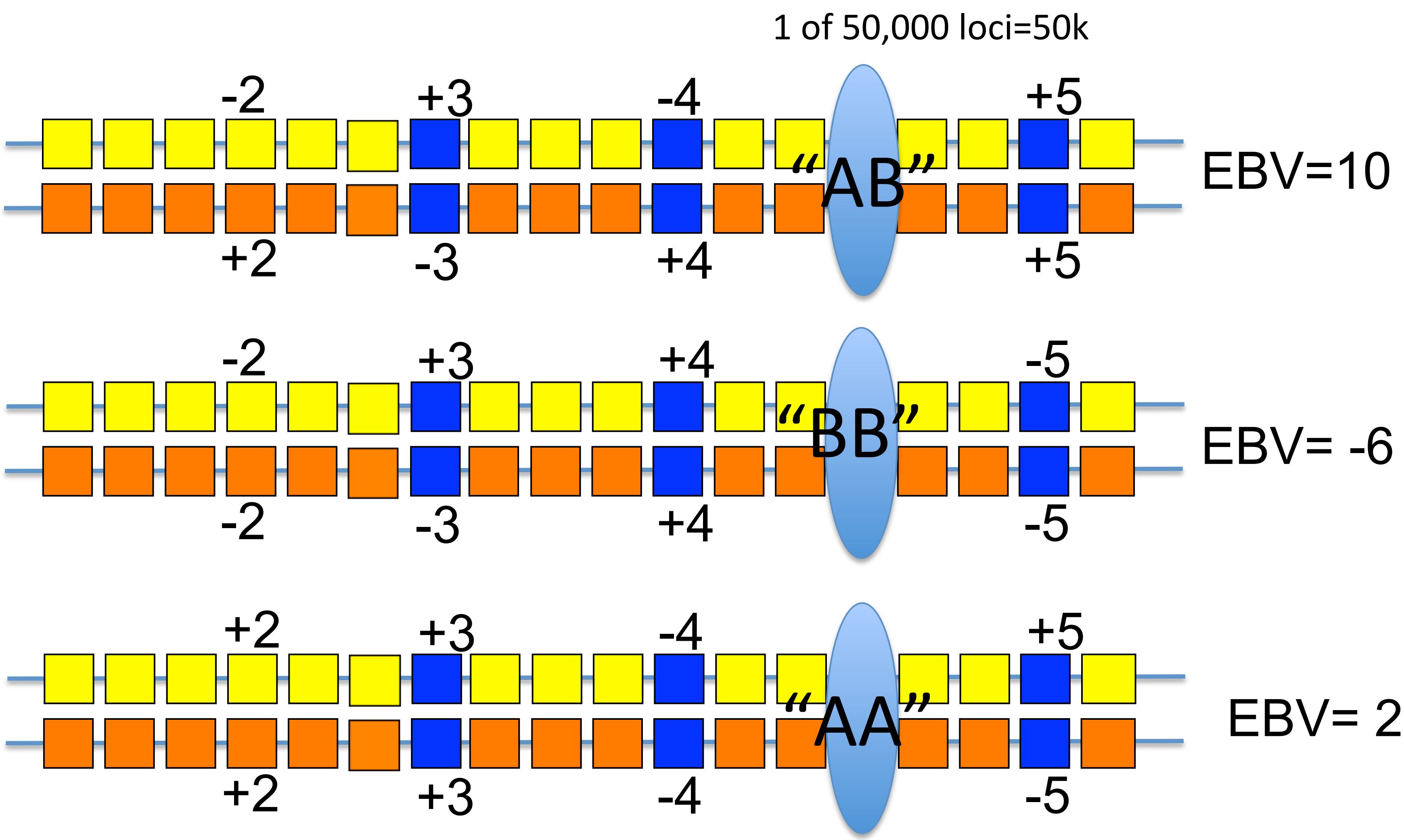


50k (Several versions)

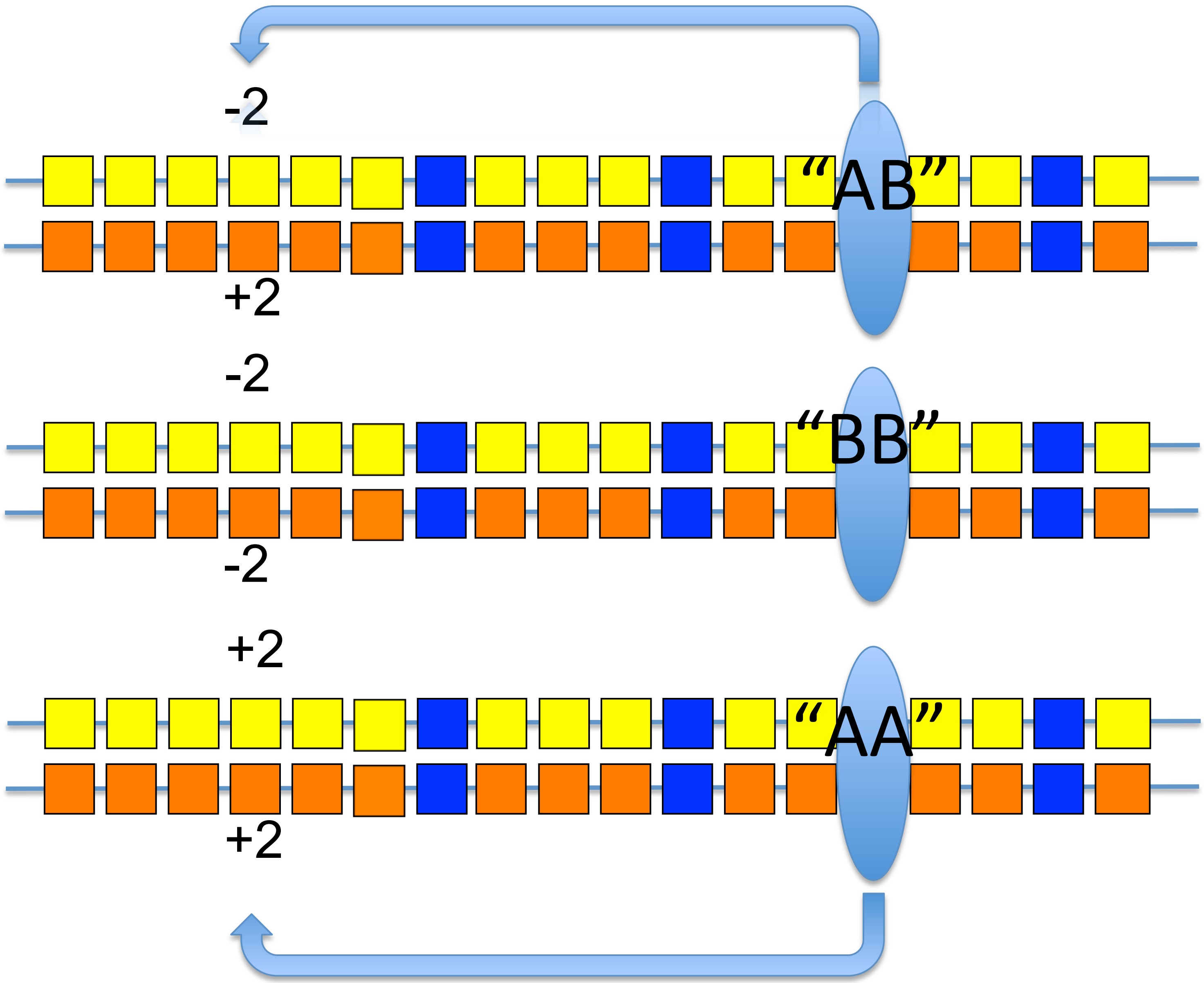


3k (LD)

SNP Genotyping the Bulls

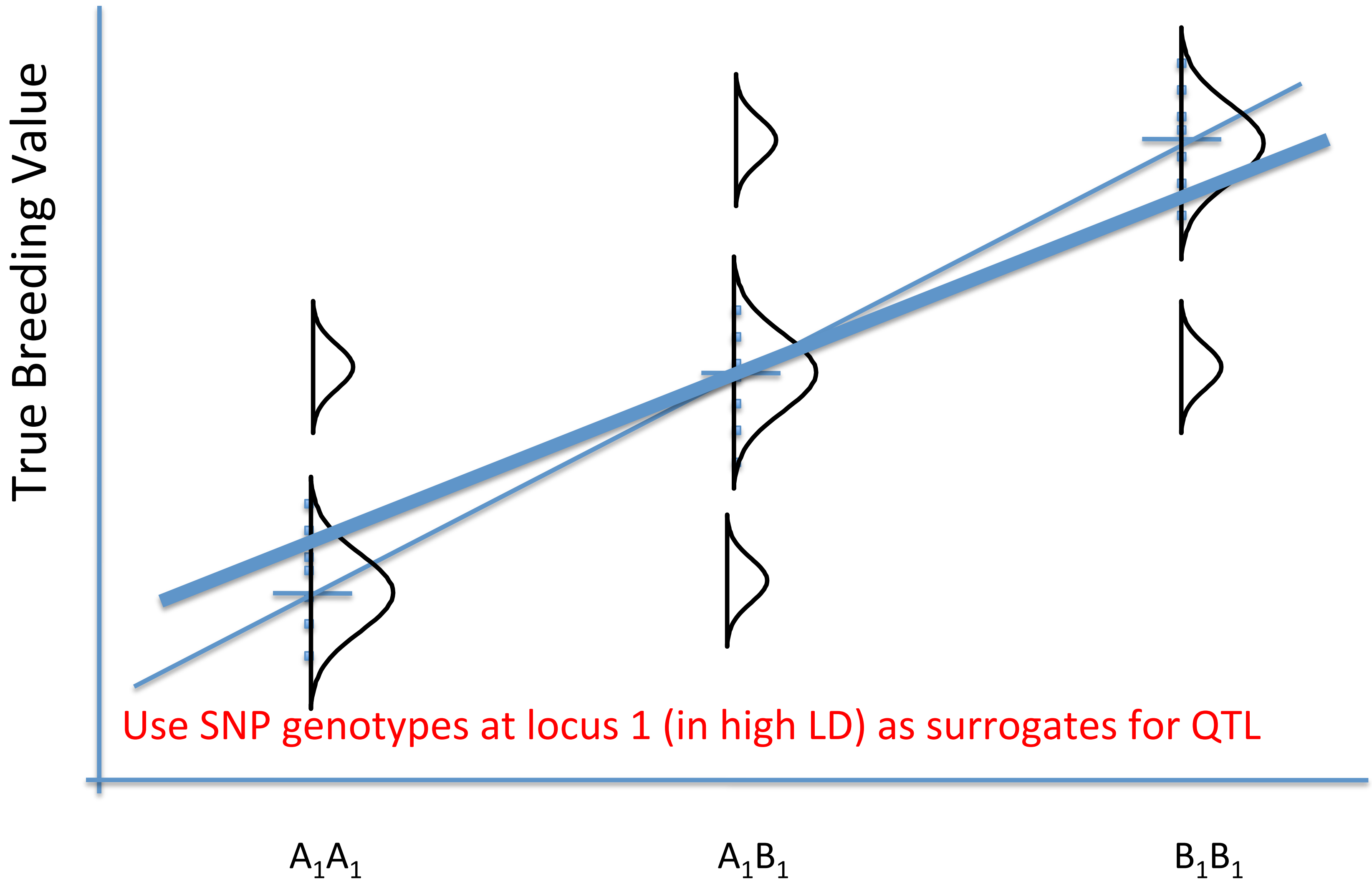


Linkage Disequilibrium (LD)

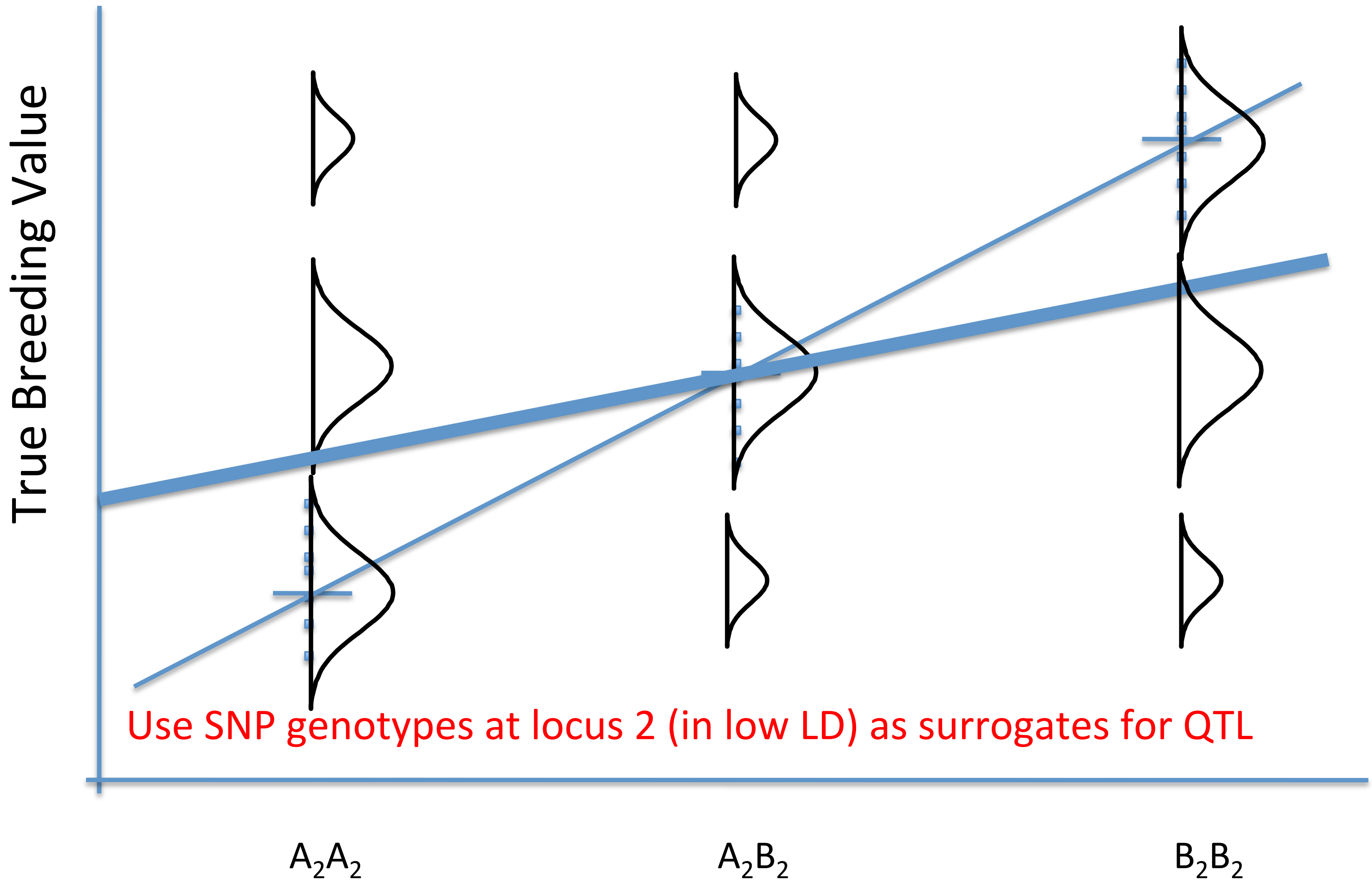


LD occurs when genotypes at one locus are predictive of genotypes at another

Practice – EBV on SNP



Practice – EBV on SNP



In practice fitting all SNP simultaneously
Meuwissen, Hayes and Goddard (2001)

www.23andme.com

23andMe

Health Risks

Alzheimer's Disease

Decreased Risk

NAME	CONFIDENCE	YOUR RISK	AVG. RISK	COMPARED TO AVERAGE
Alzheimer's Disease	★★★★	**4.9%**	7.2%	0.69x

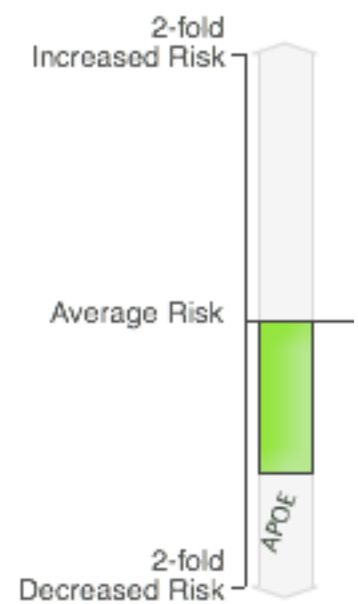
Your Data | How It Works | Technical Report | Community (162)

Technical Report

Gene or region: APOE

	SNPs used	Genotype	Allele	Adjusted Odds Ratio
Dorian Garrick	rs7412 rs429358	*CC* *TT*	ε3/ε3	European: 0.67

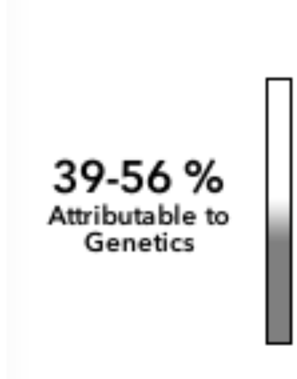
Marker Effects

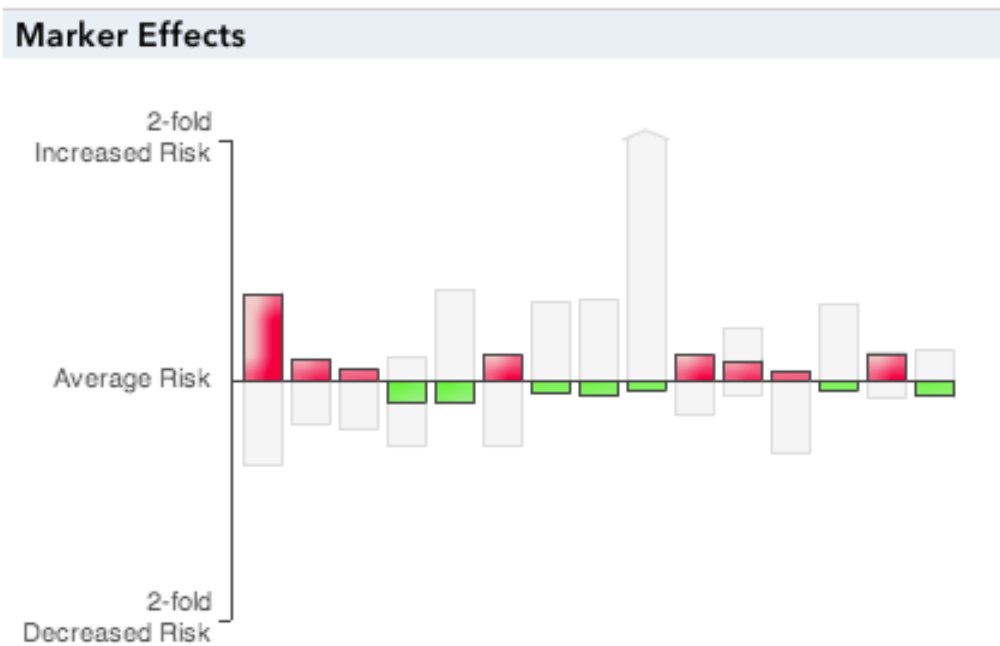


Only significant, validated GWAS findings used in prediction

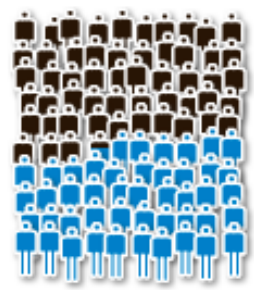
www.23andme.com

- Coronary Heart Disease





Dorian Garrick
55.0 out of 100
men of European ethnicity who share Dorian Garrick's genotype will develop Coronary Heart Disease between the ages of 45 and 79.



Average
46.8 out of 100
men of European ethnicity will develop Coronary Heart Disease between the ages of 45 and 79.

Each bar represents a different risk QTL allele
(mouseover shows the allele and links to the research publications)
QTL=Quantitative Trait Locus

Only significant, validated GWAS findings used in prediction

Plant & Animal Perspective

- Typically more SNP loci than subjects
- Landmark concepts were suggested by Meuwissen, Hayes & Goddard (2001)
  - Could simply fit all the SNP together (regardless of "significance") by treating as random effects
    - They referred to these methods as "BLUP" or "BayesA"
  - Or use a variable selection model to fit as random effects some subset of the most informative SNP
    - They proposed a method called "BayesB"

Genomic Prediction

$$y = Xb + Ms + e$$

Like Ridge Regression

$$\begin{bmatrix} X'X & X'M \\ M'X & M'M + \lambda I \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X'y \\ M'y \end{bmatrix}$$

$$\hat{u} = M\hat{s}$$

Regardless of “significance” of s-hat

These equations have order = number of SNP+means and are dense

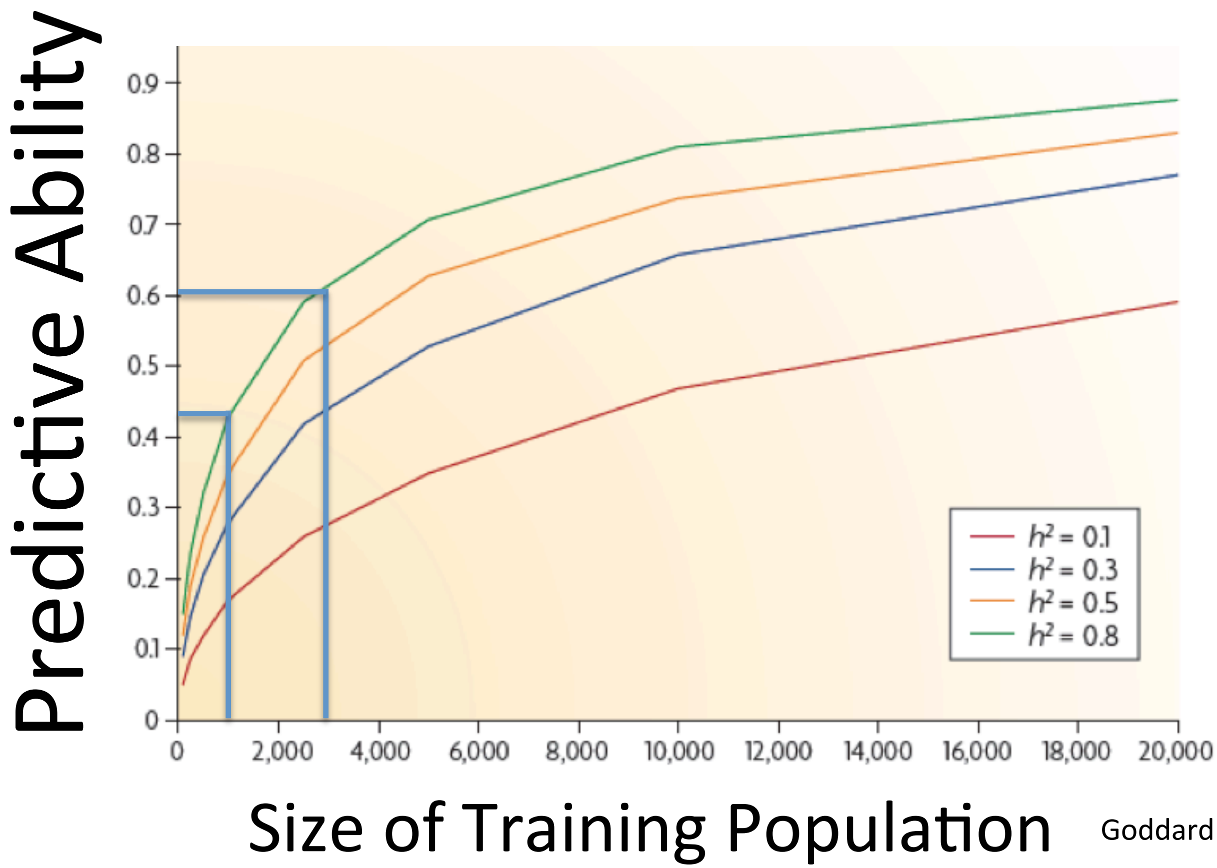
λ is a known constant = “BLUP”

λ unknown & varies for each marker = Bayes A

and marker effects from mixture distribution = Bayes B

Meuwissen, Hayes & Goddard (2001)

Theoretical Basis for Accuracy



Heritability=0.8

N_e=100
like Holsteins & Jerseys

1,000 training animals
r=0.43 20% genetic variance

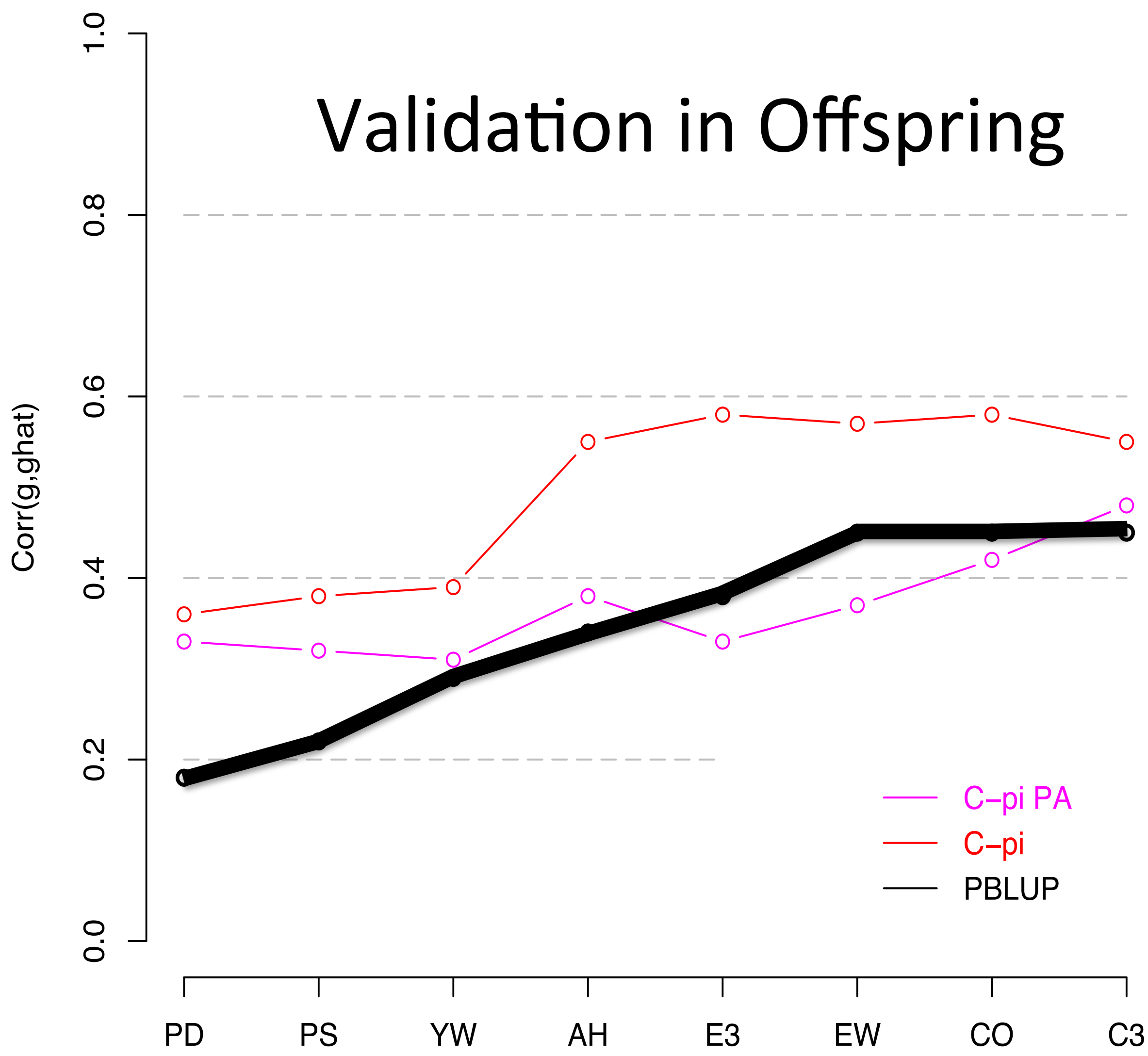
3,000 training animals
r=0.6 36% genetic variance

Goddard & Hayes (Nature Reviews Genetics, 2009)

Reliable prediction requires large training populations of genotyped and phenotyped individuals

Predictive Ability = Accuracy (r) = correlation true & predicted merit

Accuracy of Genomic Prediction

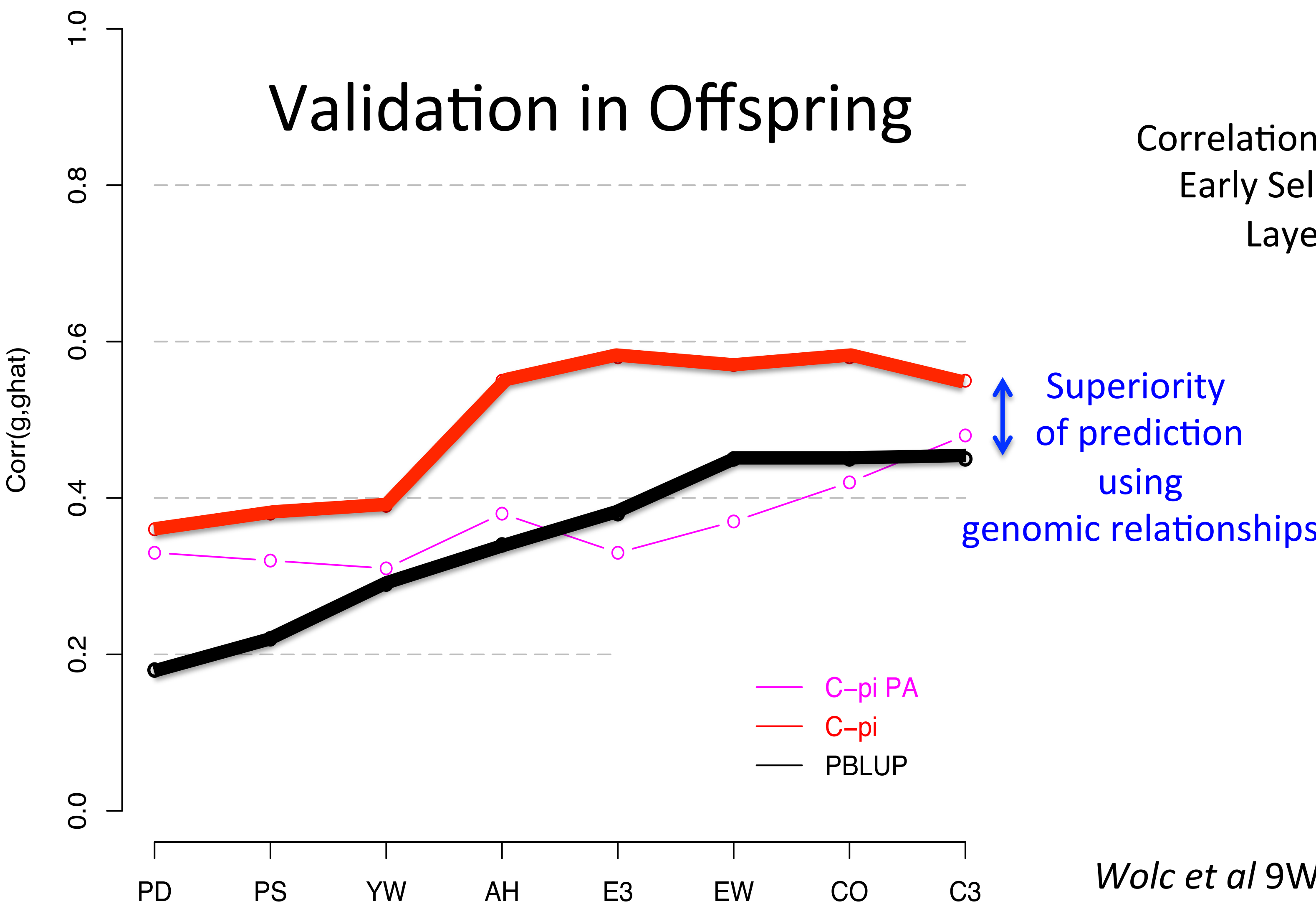


Correlation(g,g-hat)
Early Selection
Layers

Conventional pedigree relationships

Wolc et al 2010 9WCGALP

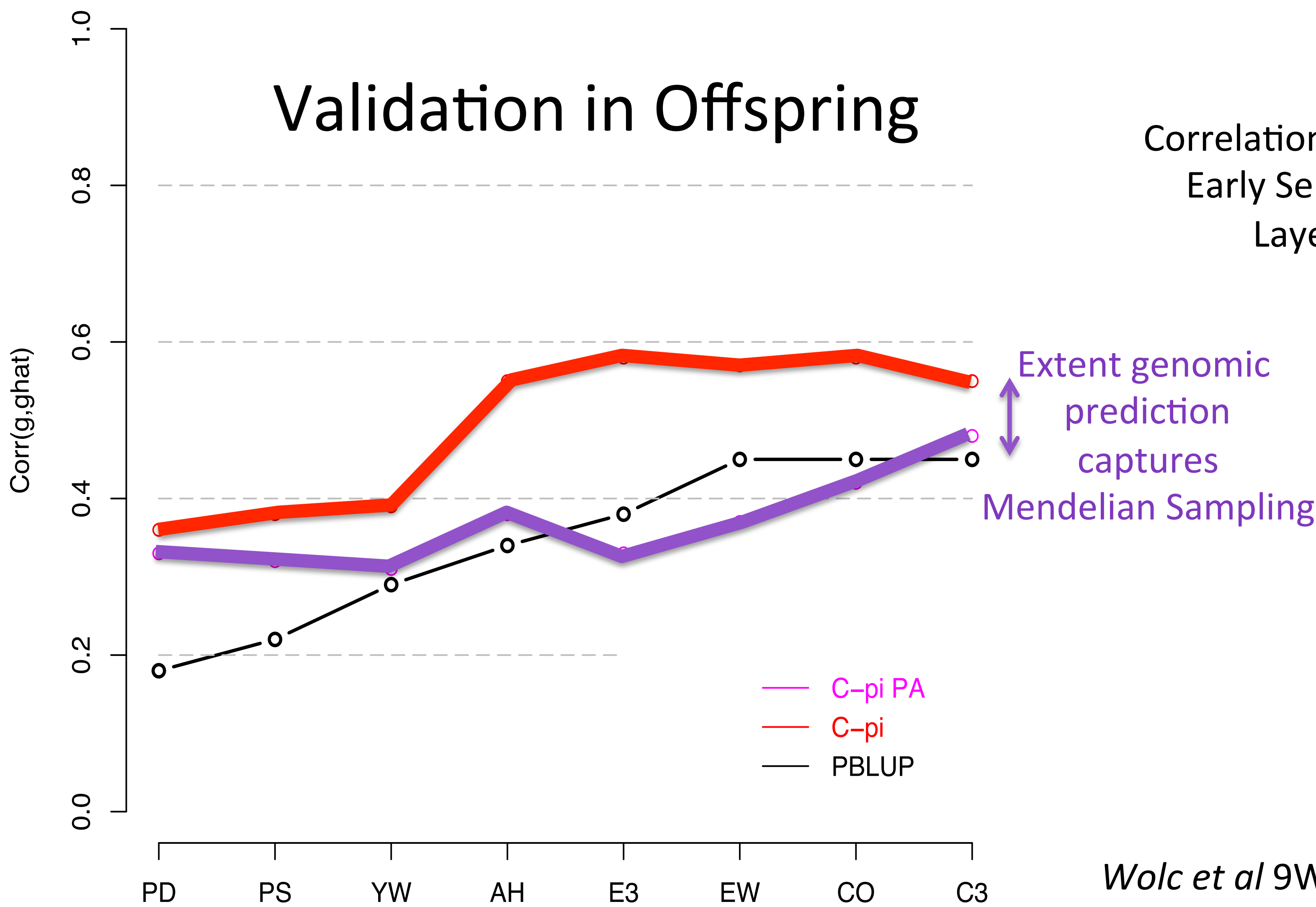
Accuracy of Genomic Prediction



Correlation(g,g-hat)
Early Selection
Layers

Wolc et al 9WCGALP

Accuracy of Genomic Prediction



Correlation(g,g-hat)
Early Selection
Layers

Wolc et al 9WCGALP

Layer Hens – Dekkers scheme

Strategy	Traditional	
	Male	Female
#candidates with phenotype	1000	3000
# selected	60	360
Generation interval (months)	13	
Information	Own Phenotype	

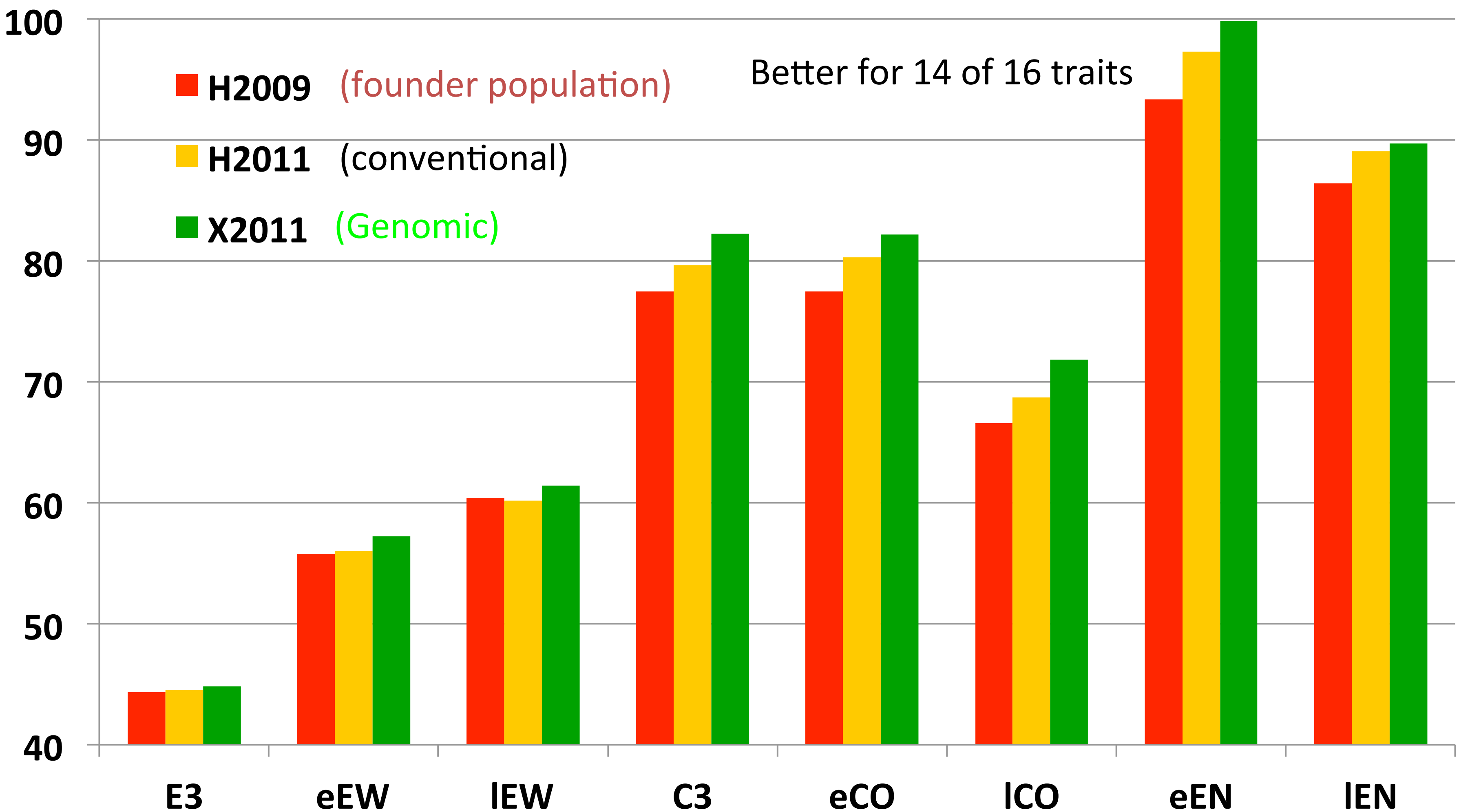
Layer Hens – Dekkers scheme

Strategy	Traditional		GS	
	Male	Female	Male	Female
#candidates with phenotype	1000	3000	300	300
# selected	60	360	50	50
Generation interval (months)	13		6-7	
Information	Own Phenotype		Genotype+Phenotype	

Crossclassified mating

Halve the generation interval and reduce costs by (less phenotyping) to get same gain & same inbreeding

Selection Response - Difference between the lines



After 3 generations of conventional or 6 gens of genomic selection

Genomic selection was as good, if not better in terms of realized response

Predictions in Beef Cattle Breeds

Trait	RedAngus (6,412)	Angus (3,500)	Hereford (2,980)	Simmental (2,800)	Limousin (2,400)	Gelbvieh (1,321)+
BirthWt	0.75	0.64	0.68	0.65	0.58	0.62
WeanWt	0.67	0.67	0.52	0.52	0.58	0.52
YlgWt	0.69	0.75	0.60	0.45	0.76	0.53
Milk	0.51	0.51	0.37	0.34	0.46	0.39
Fat	0.90	0.70	0.48	0.29		0.75
REA	0.75	0.75	0.49	0.59	0.63	0.61
Marbling	0.85	0.80	0.43	0.63	0.65	0.87
CED	0.60	0.69	0.68	0.45	0.52	0.47
CEM	0.32	0.73	0.51	0.32	0.51	0.62
SC		0.71	0.43		0.45	
Average	**0.67**	**0.69**	**0.52**	**0.47**	**0.57**	**0.56**

Genetic correlations from k-fold validation Saatchi et al (GSE, 2011; 2012; J Anim Sc, 2013)

PA+DYD better than DYD

Train	PA+DYD	DYD
Validate	**DYD**	**DYD**
Nellore (BWT) (1206)	0.71	0.58
Nellore (BWT) (791)	0.51	0.45
Brangus (BWT)	0.65	0.61
Brngus (WWT)	0.52	0.45
	0.60	0.52
	36%	27%

GGP-HD better than 50k

Train	PA+DYD	PA+DYD	DYD		Current
Validate	DYD	DYD	DYD	NextGen	GeneSeek
Training Size	10,000	10,000	3,000		
Panel	New50K	NewGGP_HD	Old50k	Variance	Variance
bw	0.83	0.86	0.68	74%	46%
ced	DNC	0.84	0.68	71%	46%
cem	0.46	0.55	0.51	30%	26%
fat	0.32	0.38	0.48	14%	23%
mcw	0.77	0.80	0.64	64%	41%
milk	0.47	0.50	0.37	25%	14%
mrb	0.64	0.71	0.43	50%	18%
rea	0.58	0.58	0.49	34%	24%
sc	0.58	0.60	0.43	36%	18%
ww	0.64	0.67	0.52	45%	27%
yw	0.71	0.75	0.60	56%	36%
	0.60	**0.66**	**0.53**	**0.45**	**0.29**

DNC=did not converge

Blending

- Use DGV along with EBV in selection index
- Use DGV as a correlated trait
- Use DGV as “external EBV”
  - Same concept as using interbull EBV in local
- Combine genotyped and nongenotyped
  - Known as “Single Step”

Blending is a Selection Index Problem

$\text{Blended\_EPD} = \text{mean} + b_1\text{EBV} + b_2\text{DGV}$

- Need to determine the weights (b_1 and b_2) to combine the information sources
  - Based on variance-covariance assumptions

- And determine the accuracy of the blended EPD which must be greater than either of the component EPDs

Selection Index Assumptions

Pb = g

$$var\begin{bmatrix}\widehat{u}\\ \widehat{m}\\ u\end{bmatrix}=\begin{bmatrix} r_p^2 & r_p^2 r_m^2 & r_p^2\\ r_p^2 r_m^2 & r_m^2 & r_m^2\\ r_p^2 & r_m^2 & 1\end{bmatrix}\sigma_g^2$$

$$var\begin{bmatrix}u-\widehat{u}\\ m-\widehat{m}\end{bmatrix}=\begin{bmatrix}1-r_p^2 & \left(1-r_p^2\right)\left(1-r_m^2\right)\\ \left(1-r_p^2\right)\left(1-r_m^2\right) & 1-r_m^2\end{bmatrix}$$

Kachman (unpublished) – Assumes well behaved predictions

Blending

$$\widehat{u_n} = \frac{(1-r^2)(\widehat{u_p} - \mu_{u_p}) + (1-a^2)(\widehat{m} - \mu_m)}{1 - r^2 a^2}$$

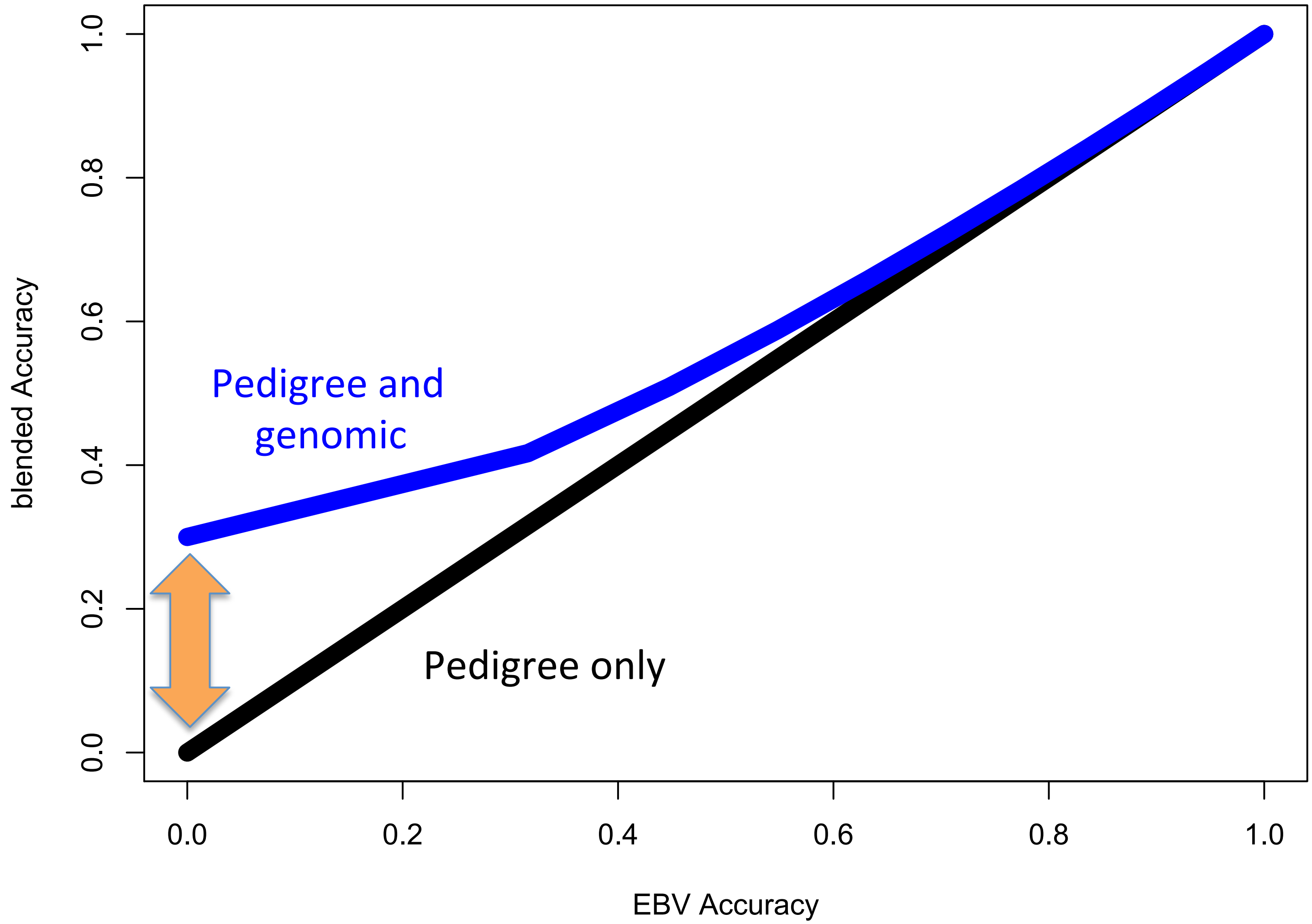
$$Rel_n = 1 - \frac{(1-r^2)(1-a^2)}{1-r^2 a^2}$$

$where\ \widehat{u_p}\ is\ the\ previous\ national\ EBV\ with\ Rel_p = a^2$

$and\ \widehat{m}\ is\ the\ MBV\ (DGV)\ with\ genetic\ correlation\ r^2$

Impact on Accuracy--%GV=10%

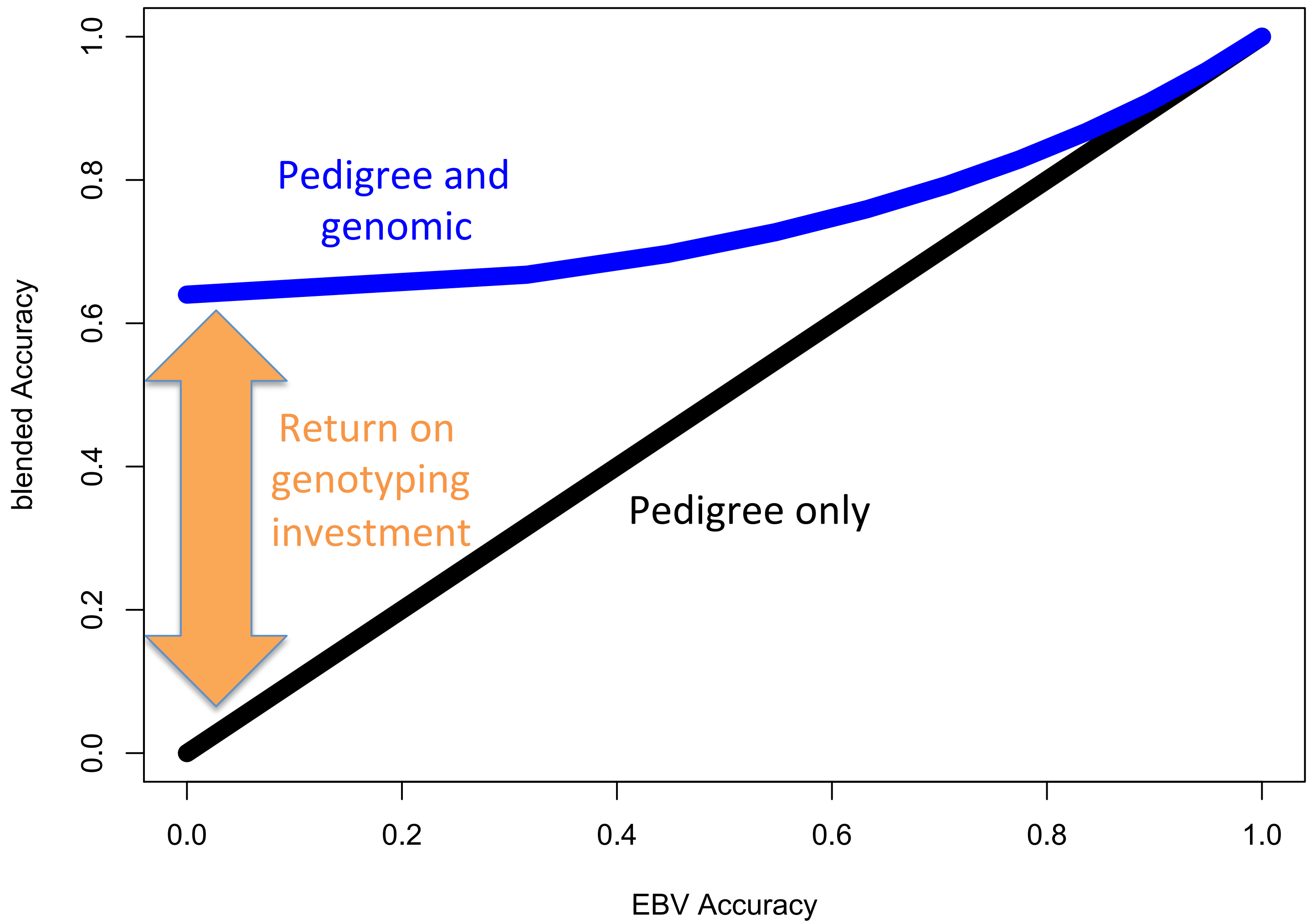
Genetic correlation=0.3



Blending will not improve the accuracy of a bull that already has a reliable EBV

Impact on Accuracy--%GV=40%

Genetic correlation=0.64



Blended EBVs are equally likely to be better or worse than the preblended EBVs

Properties of BLUP (1 of 2)

- Provided the model is correct:

$$\operatorname{cov}(u,\hat{u}) = \operatorname{var}(\hat{u})$$

Quantify from inverse MME
Or approximate from MME

- Then

$$\beta_{u/\hat{u}} = \frac{\operatorname{cov}(u,\hat{u})}{\operatorname{var}(\hat{u})} = 1 \quad (exactly)$$

$$Although\ E[u] = 0,\ E[u / \hat{u}] = \hat{u}$$

Properties of BLUP (2 of 2)

- Provided the model is correct:

$$\mathrm{cov}(u,\hat{u}) = \mathrm{var}(\hat{u})$$

- Then

$$r_{u,\hat{u}} = \frac{\mathrm{cov}(u,\hat{u})}{\sqrt{\mathrm{var}(\hat{u})\,\mathrm{var}(u)}} = \sqrt{\frac{\mathrm{var}(\hat{u})}{\mathrm{var}(u)}}$$

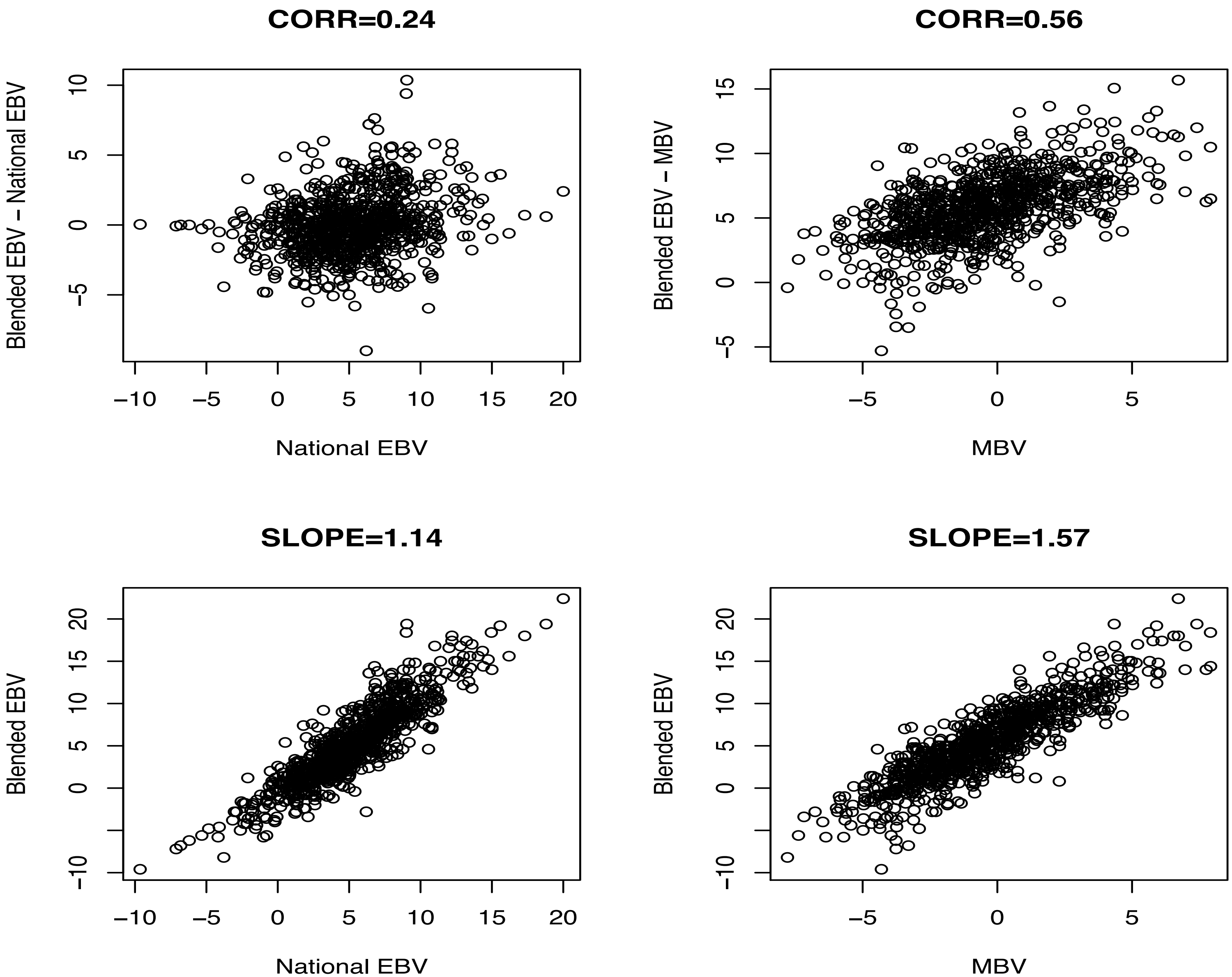
- And

$$\mathrm{var}(\hat{u}) = r^2_{u,\hat{u}}\ \mathrm{var}(u)$$

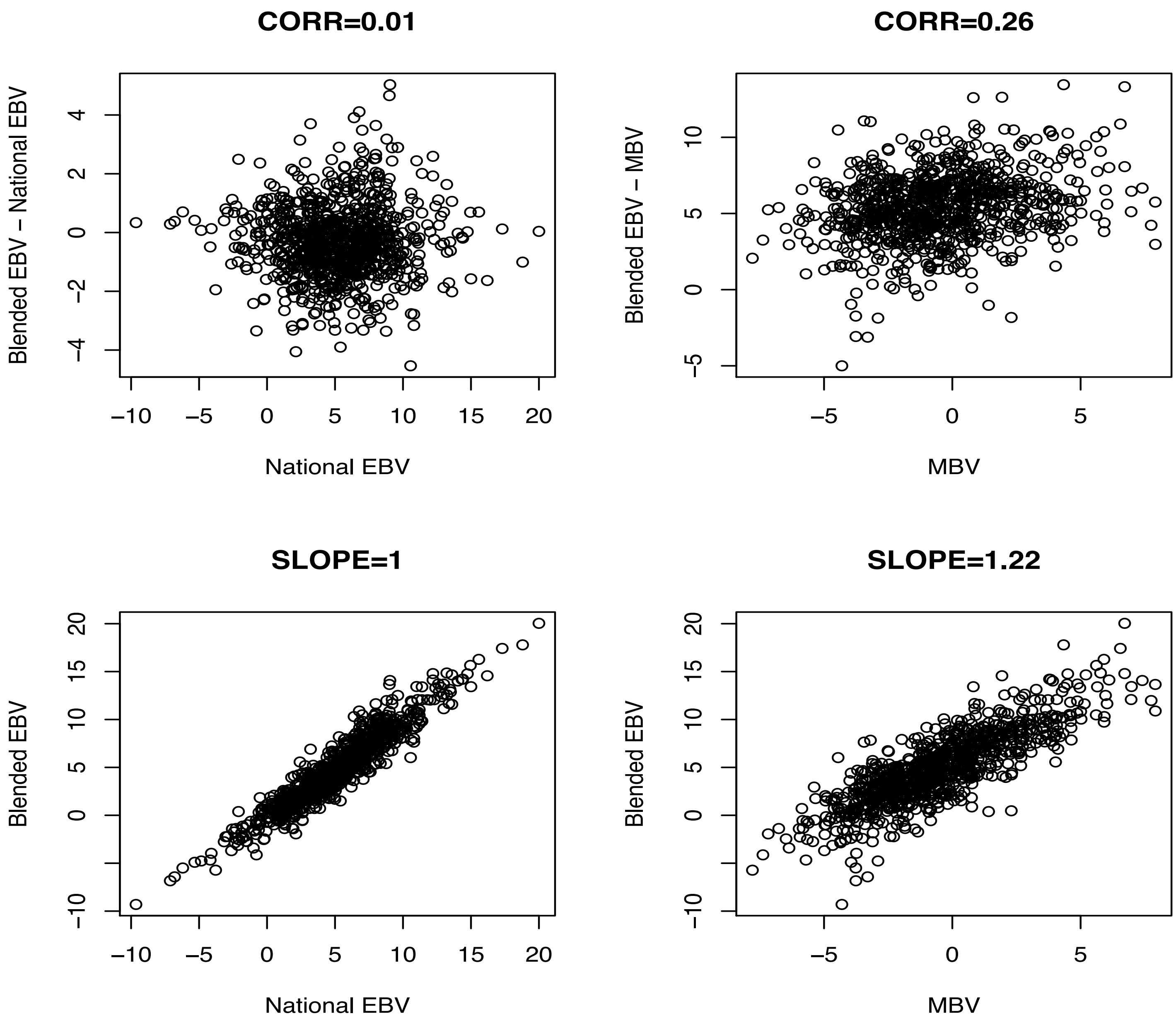
Diagnostics of Good Behavior

- Regression of more accurate (blended) on less accurate (EBV or MBV) should be 1
- Correlation of less accurate EBV with change in EBV (from less accurate to more accurate) should be zero

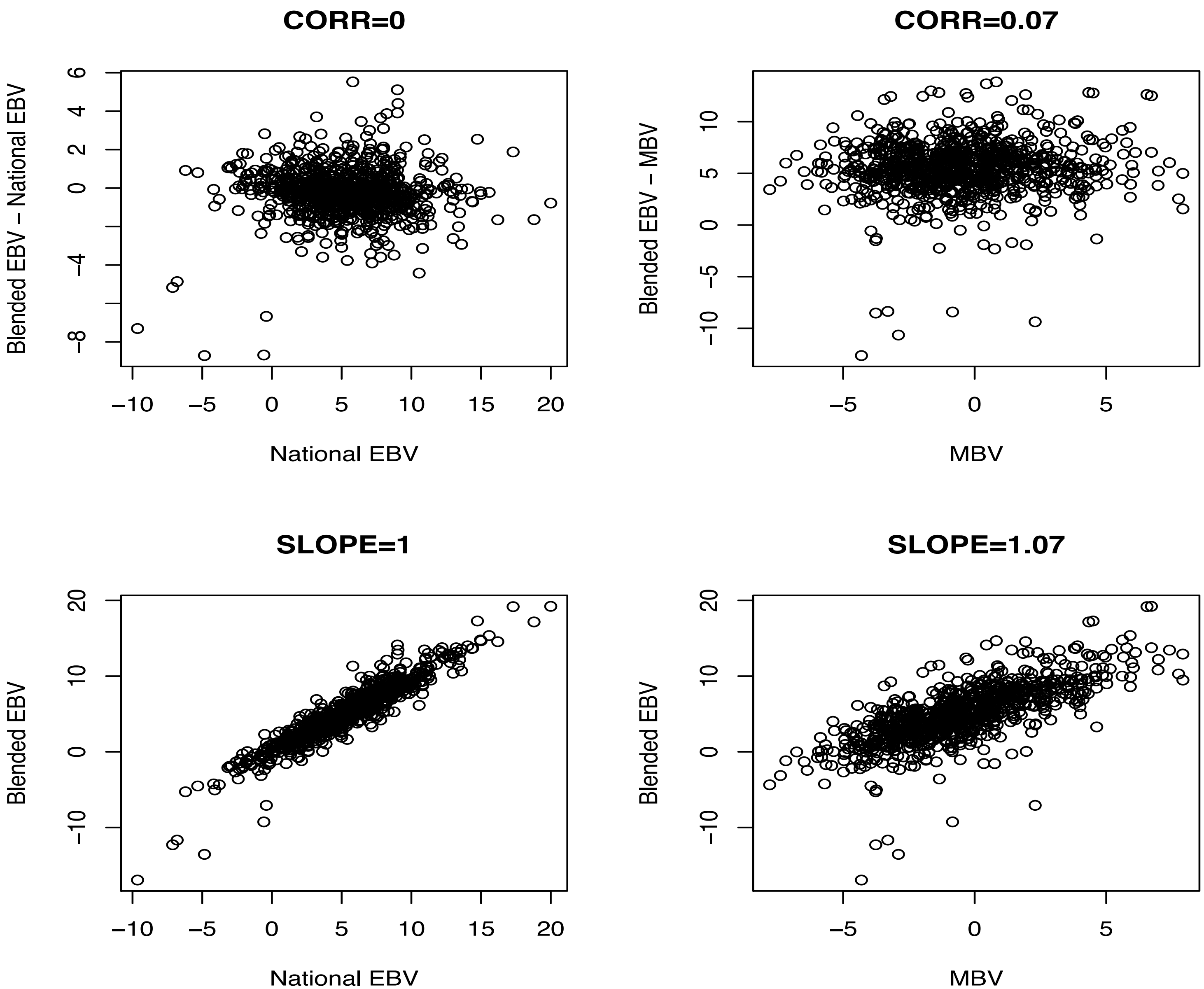
Validation of Breedplan Blending



Validation of Birth Weight



Inflation of EBV/MBV covariance

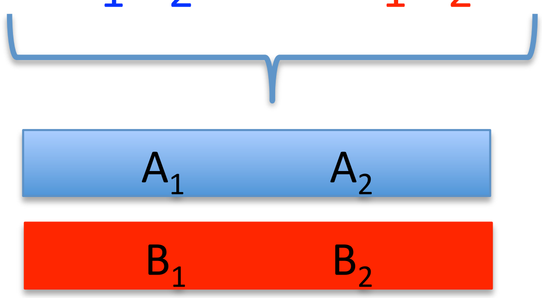


Genotypes vs Haplotypes

- Suppose an animal is
  - heterozygous at locus 1 (genotype A_1B_1) and
  - heterozygous at locus 2 (genotype A_2B_2)

Genotypes vs Haplotypes

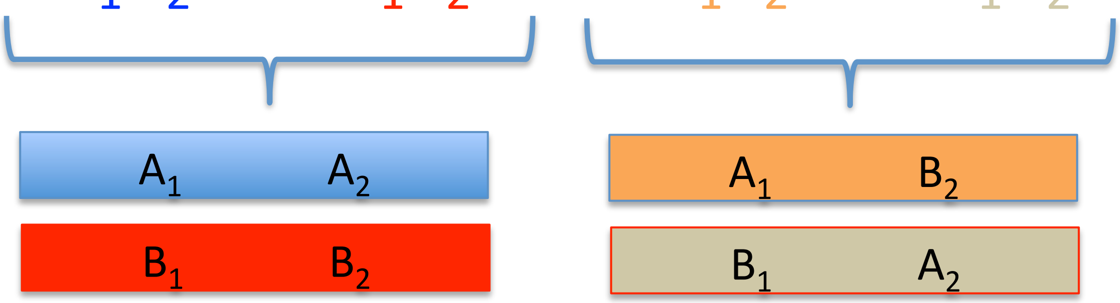
- Suppose an animal is
  - heterozygous at locus 1 (genotype A_1B_1) and
  - heterozygous at locus 2 (genotype A_2B_2)
- Its diplotype (pair of haplotypes) might be
  - Either A_1A_2 and B_1B_2



Alleles are in coupling

Genotypes vs Haplotypes

- Suppose an animal is
  - heterozygous at locus 1 (genotype A_1B_1) and
  - heterozygous at locus 2 (genotype A_2B_2)
- Its diplotype (pair of haplotypes) might be
  - Either A_1A_2 and B_1B_2 or A_1B_2 and B_1A_2



Alleles are in coupling

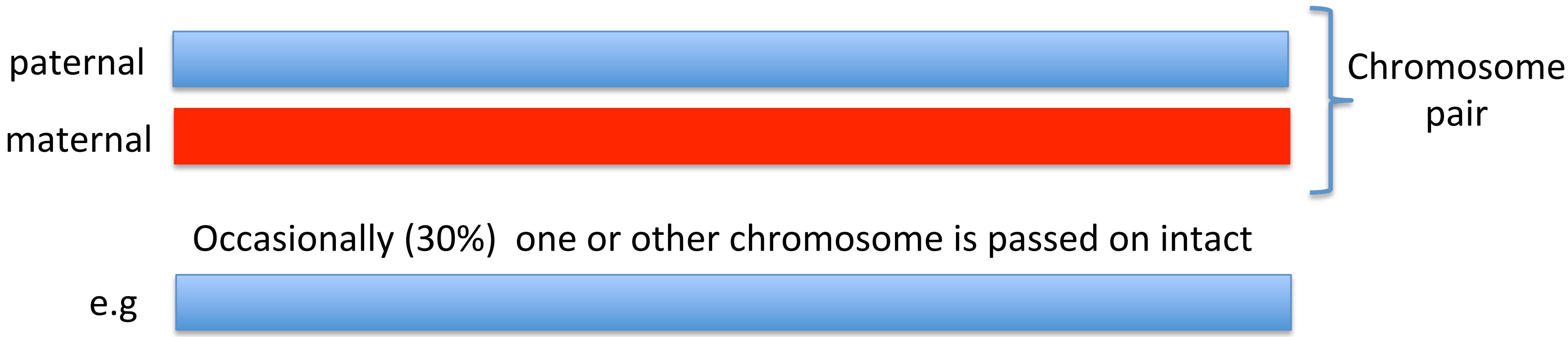
Alleles are in repulsion

Many Potential Haplotypes

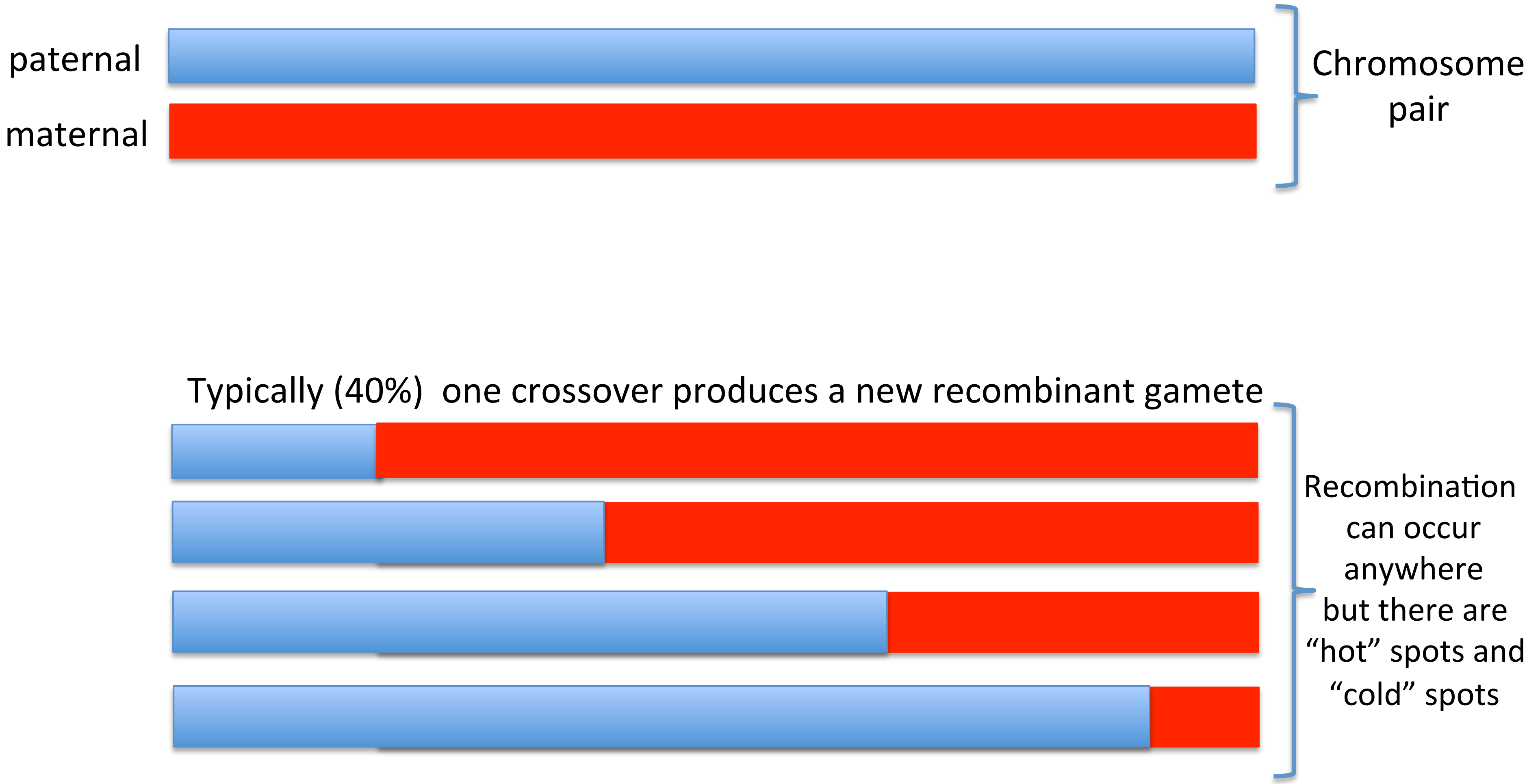
- At 2 loci there are 4 possible haplotypes
  - "A_1A_2", "A_1B_2", "B_1A_2", and "B_1B_2"
- At 3 loci there are 8 possible haplotypes
  - "AAA", "AAB", "ABA", "ABB", "BAA", "BAB", "BBA", "BBB"
- At *k* loci there are 2^*k* possible haplotypes
- At 20 loci (e.g. 1% or 1 Mb chromosome on 50k) there are >1 million possible haplotypes
  - In a population of <1 million they can't all be present!

SNP Alleles are inherited in blocks

SNP Alleles are inherited in blocks

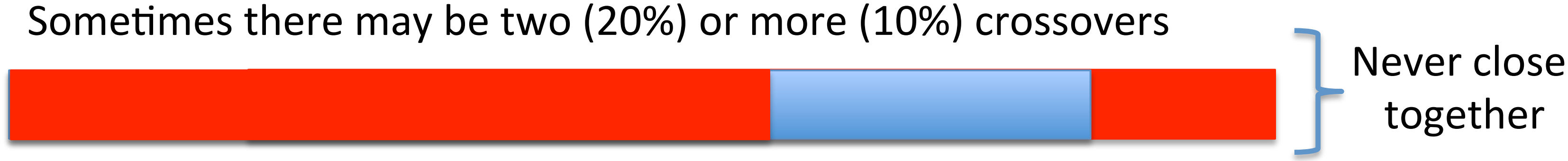


SNP Alleles are inherited in blocks

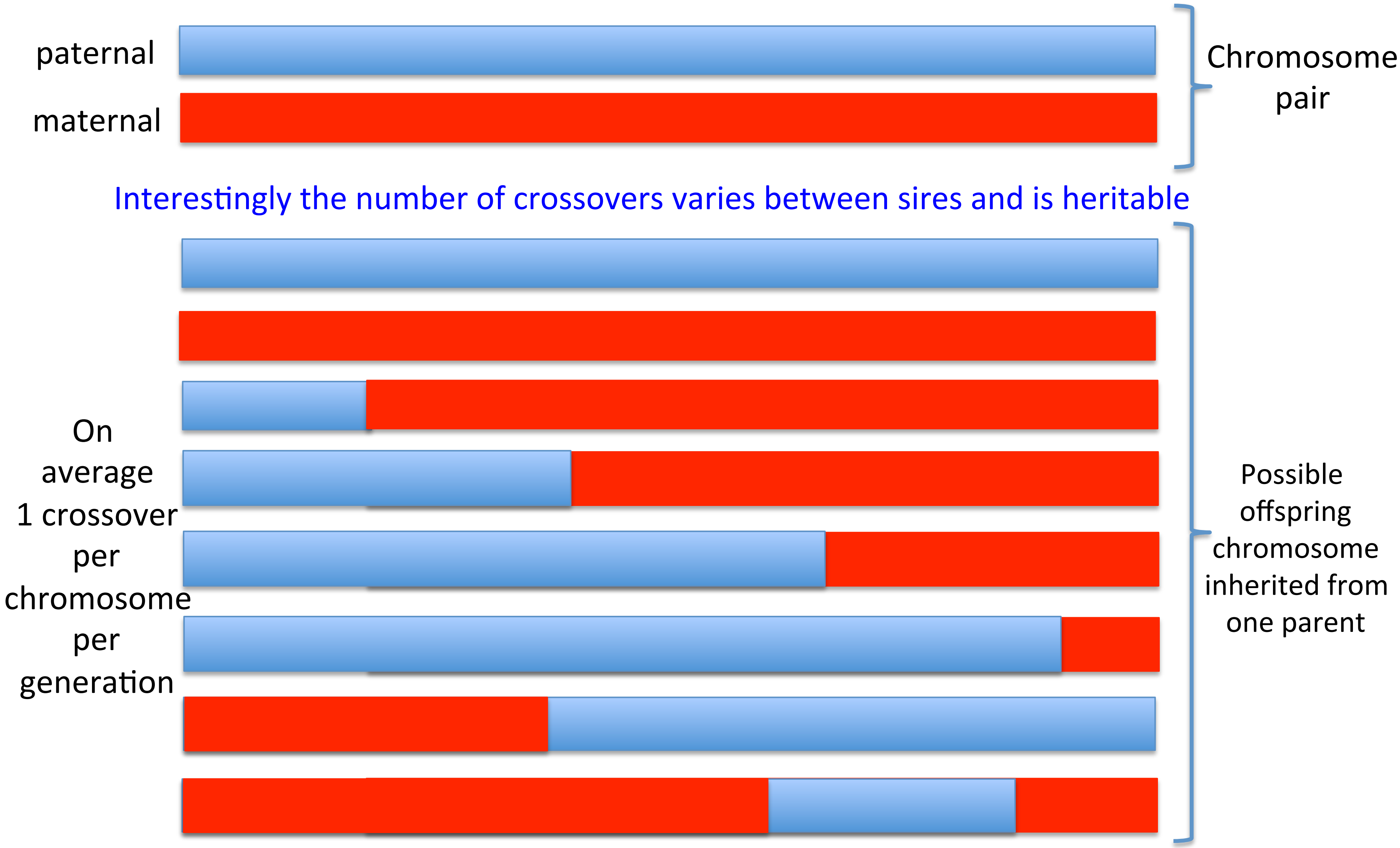


SNP Alleles are inherited in blocks

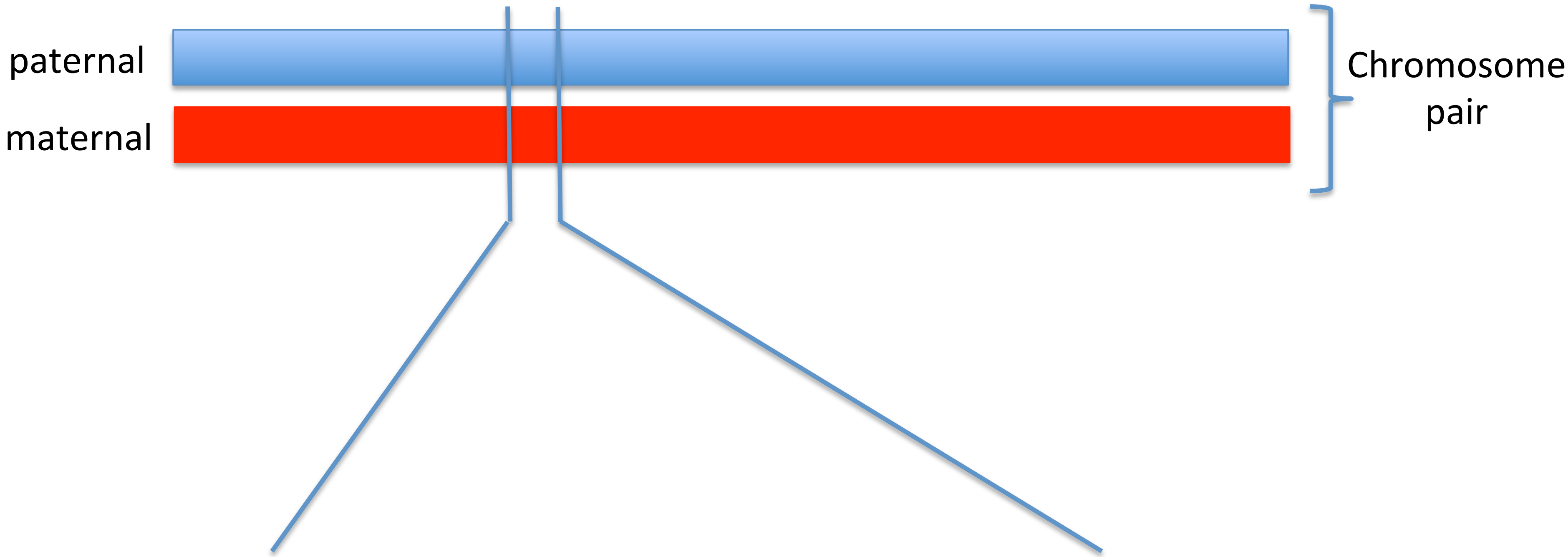
Sometimes there may be two (20%) or more (10%) crossovers



SNP Alleles are inherited in blocks

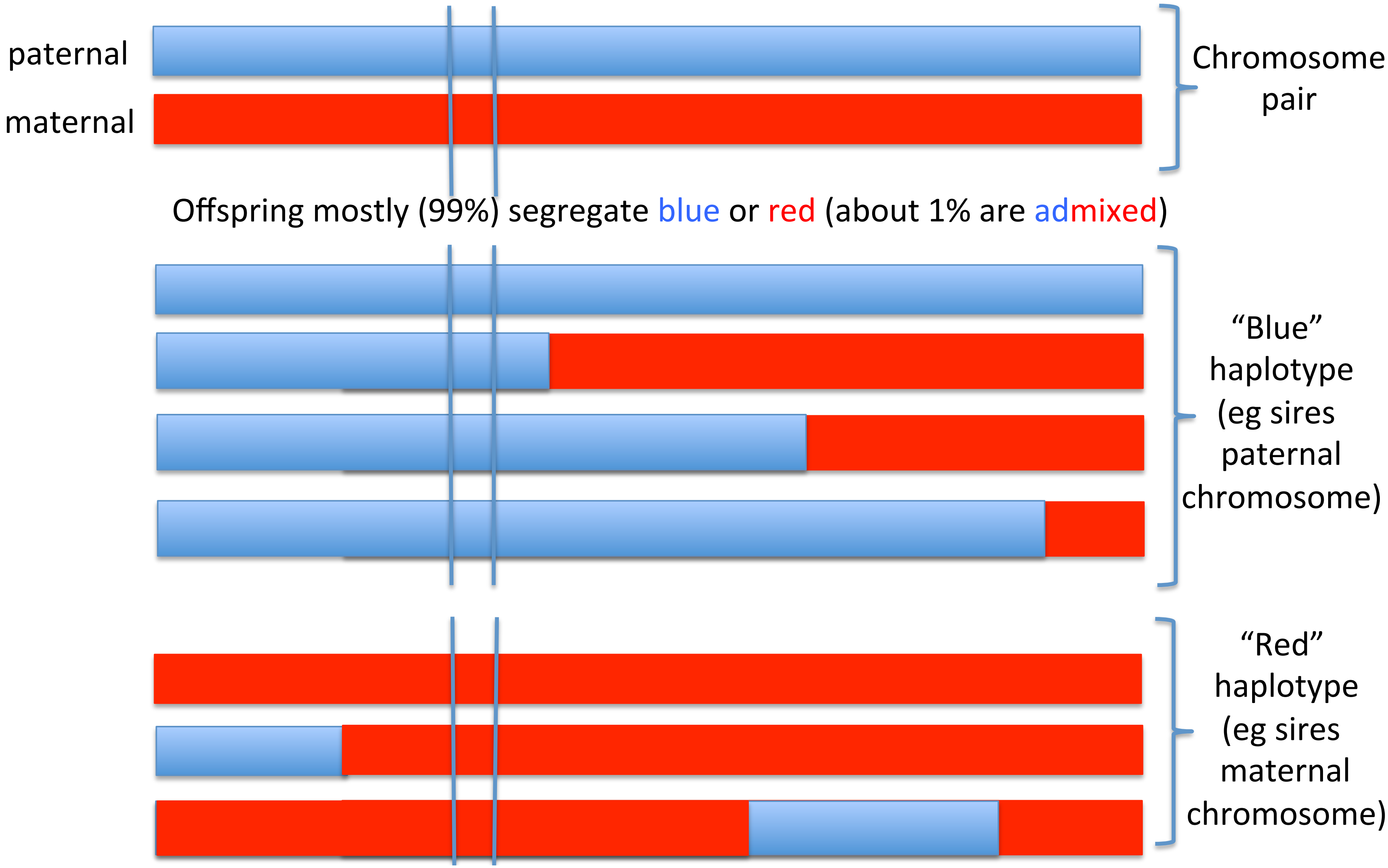


SNP Alleles are inherited in blocks

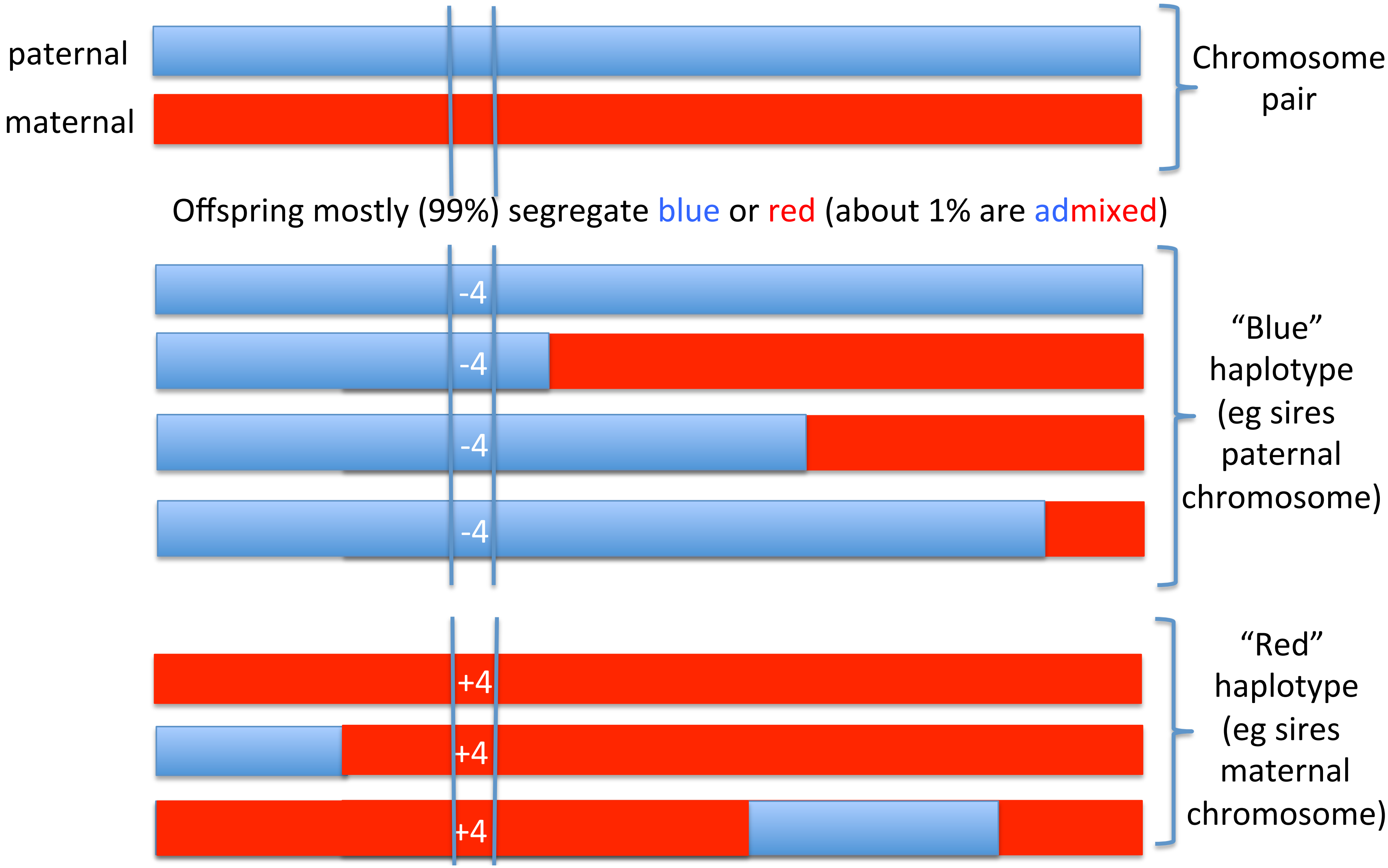


Consider a small window of say 1% chromosome (1 Mb)

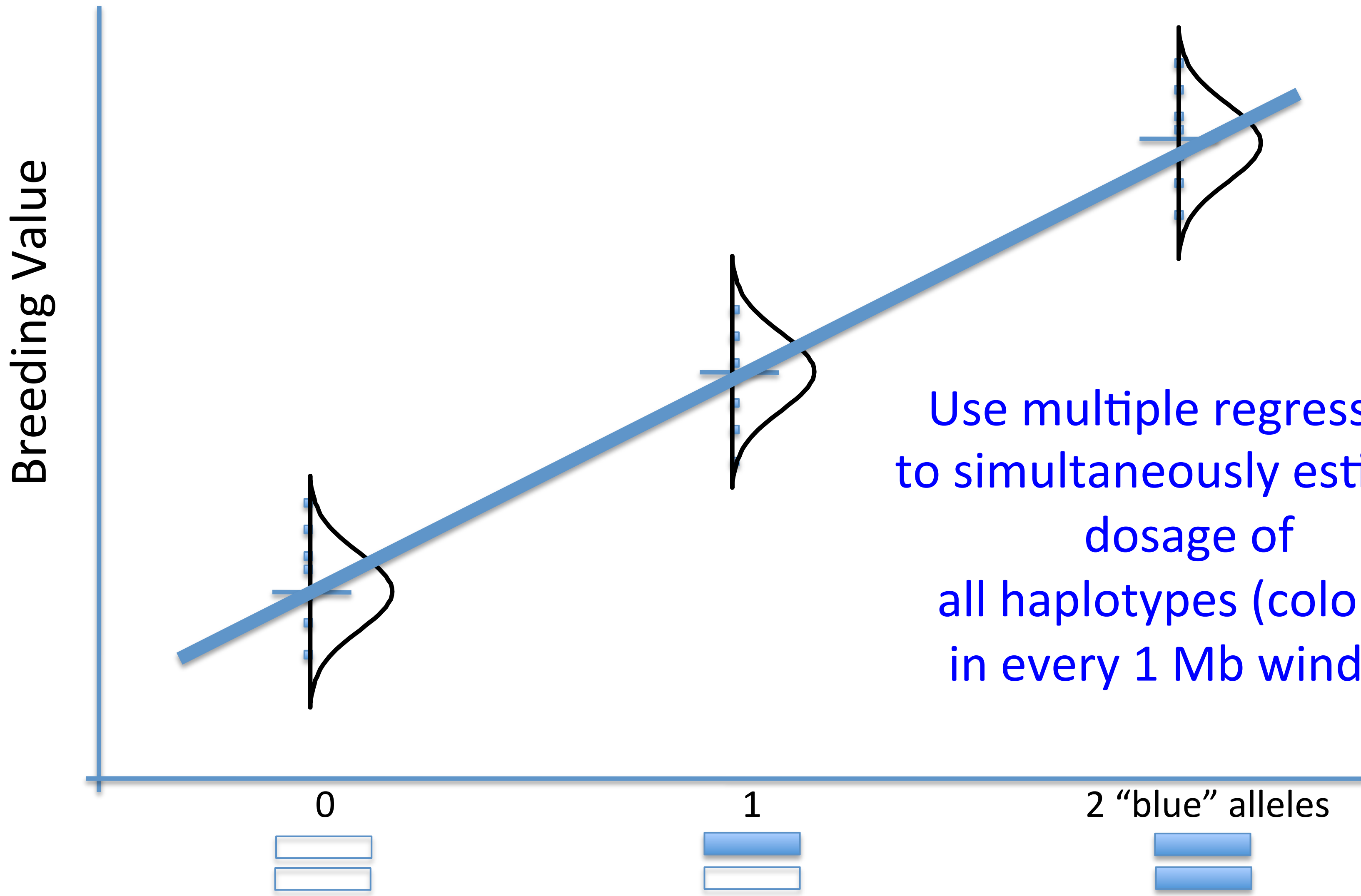
SNP Alleles are inherited in blocks



SNP Alleles are inherited in blocks



Regress BV on haplotype dosage



Use multiple regression
to simultaneously estimate
dosage of
all haplotypes (colours)
in every 1 Mb window

Few Haplotypes are Present

- In *Bos taurus* breeds we seldom see more than 30 common haplotypes in any 1Mb chromosome region (i.e. 1% chromosome)
  - Common haplotypes are those seen more often than once every 50 individuals (≥ 1% frequency)
  - On average there are 20 such common haplotypes
    - We could assign these 20 “colours” like “blue”, “red” etc to represent their ancestral origins in the breed
  - We only need enough SNP to identify haplotypes

Prediction of Shorthorn only from other Breeds

	Angus	Brangus	Gelbvieh	Hereford	Limousin	Red Angus	Simmental
Birth Weight	0.08	-0.05	0.09	0.23	0.18	0.40	0.37
Calving ease direct	0.05	-0.01	-0.16	0.17	0.15	0.23	0.30
Calving ease maternal	0.09	0.00		0.08	0.15	0.06	0.07
Carcass Weight	0.20	0.05	0.07		-0.10	0.23	0.20
Fat tickness	0.17	0.02		0.11		0.08	0.01
Milk	0.09	-0.04	0.16	-0.06	0.02	0.03	-0.06
Marbling	0.03	-0.04	0.11	-0.07	-0.08	0.09	0.17
Rib eye area	0.03	0.01	0.12	-0.07	-0.01	0.05	0.08
Weaning weight	0.12	-0.10	0.07	0.15	-0.02	0.15	0.09
Yearling weight	0.09	0.00	-0.08	0.14	0.02	0.13	0.13

Across breed prediction does not work if the breed is not in training

See also Kachman et al., 2013 GSE

Training on AANUSA

Trait	Predict AANUSA	Predict RANUSA
BirthWt	0.64	0.27
WeanWt	0.67	0.28
YearlingWt	0.75	0.23
Fat	0.70	0.21
RibEye Area	0.75	0.29
Marbling	0.80	0.21
CalvEase (D)	0.69	0.14
CalvEase (M)	0.73	0.18
Average	**0.71**	**0.23**

Cannot predict US Red Angus (RANUSA) very well from US Black Angus (AANUSA)
There is some predictive power because RANUSA exhibit some AANUSA haplotypes

Predicting American Simmental

Trait	Simmental from Single Breed	Simmental from Pooled Breeds
Birth weight	0.67	0.73
Calving ease direct	0.46	0.49
Calving ease maternal	0.31	0.29
Carcass weight	0.61	0.75
Docility	0.10	0.18
Fat thickness	0.19	0.26
Marbling	0.60	0.69
Rib eye muscle area	0.55	0.72
Shear force	0.52	0.60
Stayability	0.51	0.51
Weaning weight direct	0.56	0.63
Weaning wt maternal	0.32	0.28
Yield grade	0.73	0.91
Yearling weight	0.45	0.67

Average 22%

30% GV

Pooling uses
ASA multibreed DEBV
and not external data

Pooling breeds
does not typically
hurt predictions

and can provide
modest increases

Saatchi & Garrick, WSASAS 2013

Pooling Breeds (to Predict Brangus)

Trait	Train BRGUSA	BRGUSA+AANUSA+RANUSA
Birth Weight	0.82	0.83
Weaning Weight	0.66	0.65
Milk	0.51	0.44
Yearling Weight	0.70	0.69
Carcass Weight	0.64	0.63
Marbling IMF (U/S)	0.53	0.79
Fat (U/S)	0.53	0.52
Rib Eye Area (U/S	0.79	0.79
Scrotal Circumference	0.39	0.43
Average	**0.62**	**0.64**

Pooling breeds seldom improves accuracy in any one breed

Pooling Breeds

Trait	Limousin from Single Breed	Limousin from Pooled Breeds
Fat thickness	0.54	0.45
Marbling	0.75	0.58
Rib eye muscle area	0.68	0.57
Yield grade	0.67	0.35
Average	0.66	0.49

Pooling breeds does not typically hurt predictions

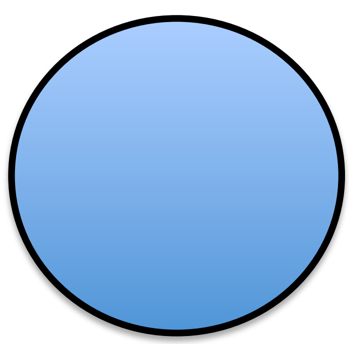
(exception is for LIM) For meat quality

Pooled breeds for LIM include AAN and RAN sires used in LIM database (LimFlex)

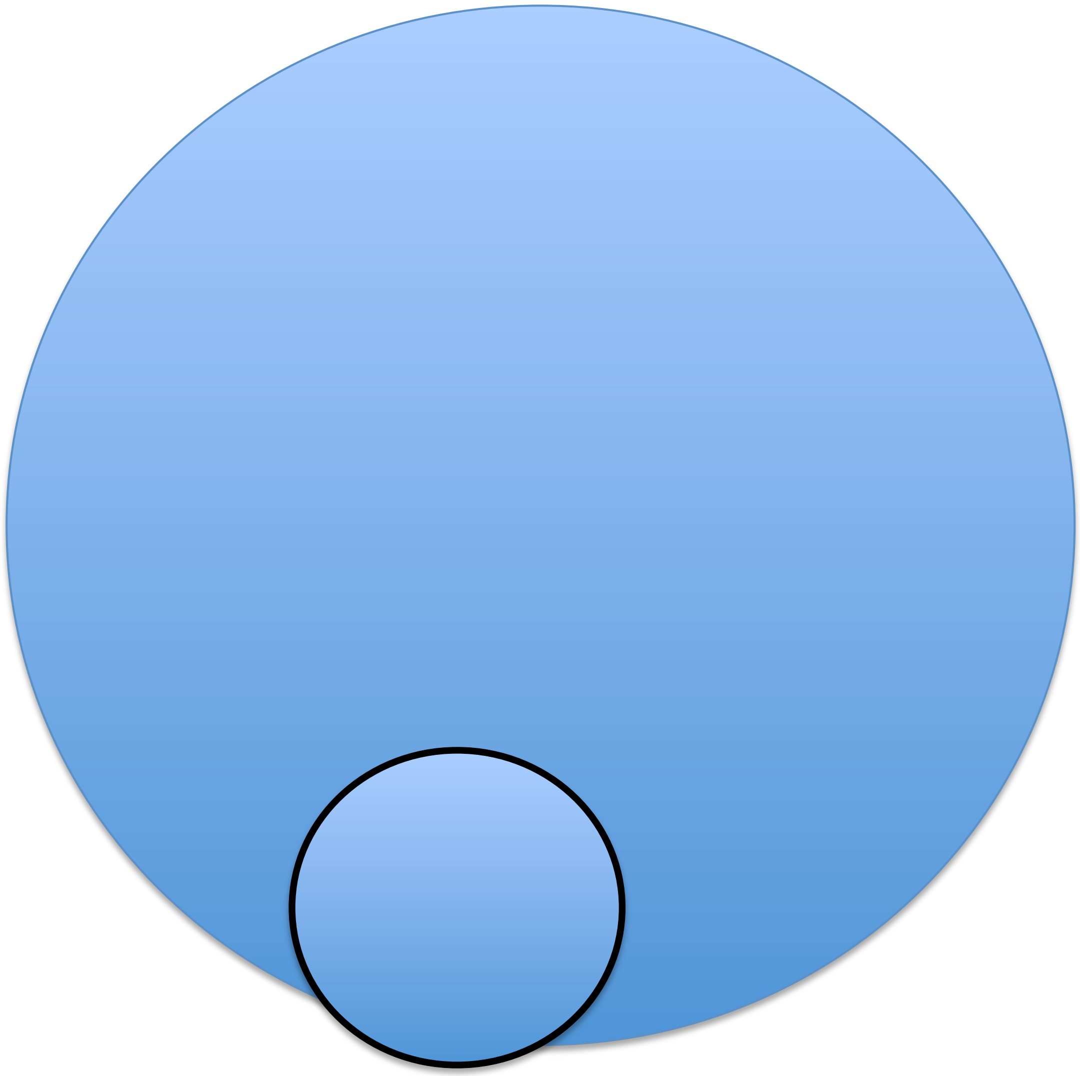
Have now genotyped the myostatin mutation to add the marker panel

Panel Comparison

Black = Illumina 50K



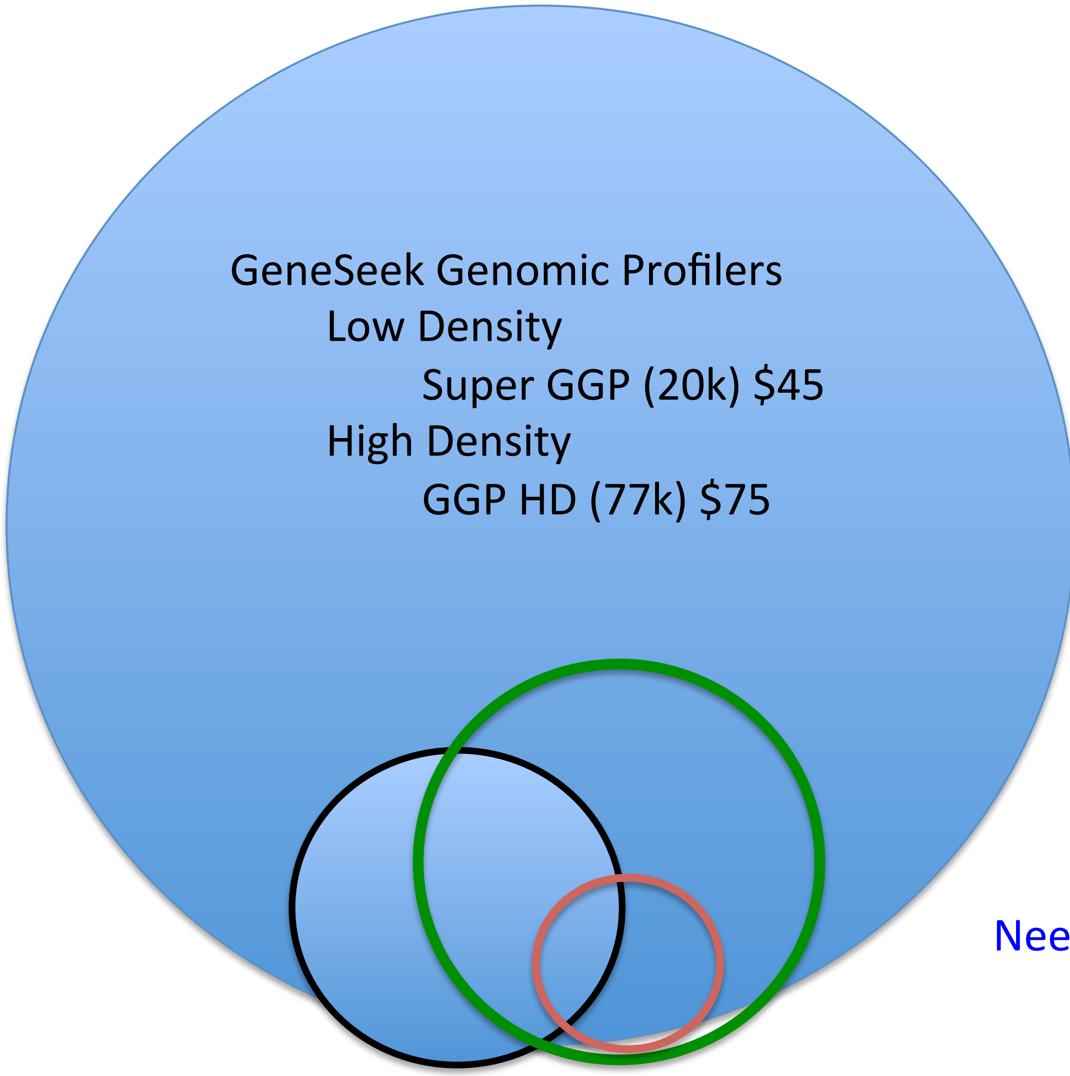
Panel Comparison



Black = Illumina 50K
Blue = Illumina HD (700K)

Panel Comparison

No longer using Illumina 50k



Orange = GGP-Super LD 19k
Green = GGP-HD (taurus) 70k
Black = Illumina 50K

GGP also include custom SNP

50k and GGP-HD share 28K
50k and GGP-Super LD share 8k

Need to genotype more individuals/yr
Need cheaper genotyping

There are multiple minor variants of all these panels!

Also a separate GGP-HD-I (Indicus)

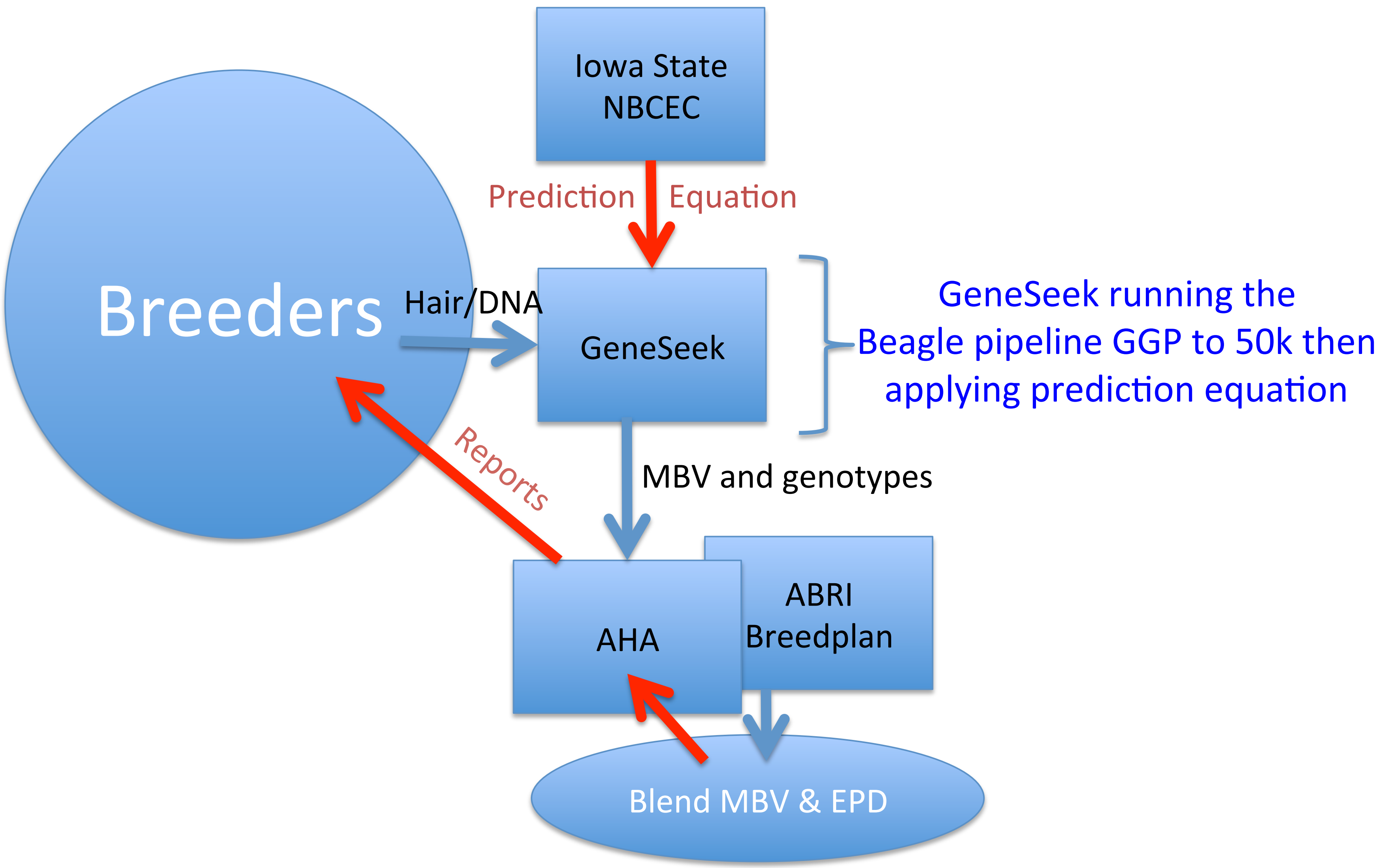
Lower Density Panels

AHA Predictive Accuracy 2,980 6-fold

Trait	Actual	Imputed
Birth Weight	0.67	0.65
Calving Ease Direct	0.68	0.67
Calving Ease Maternal	0.51	0.50
Fat Thickness	0.47	0.46
Marbling	0.42	0.42
Mature cow weight	0.64	0.62
Rib Eye Muscle Area	0.49	0.46
Scrotal Circumference	0.43	0.42
Weaning Weight Direct	0.53	0.50
Weaning Weight Maternal	0.37	0.35
Yearling Weight	0.61	0.59
Mean	0.53	0.51

Actual = 50k
Imputed = 10k
(from GGP-LD)

Genomic Prediction Pipeline



Current Genotype Counts

Breed	9k	GGP-LD	50k	GGP-HD	BOS-1	700k HD	TOTAL
AAN		911	13,409	787		947	16,054
BRG			1,128	173		243	1,544
BSH			325			136	461
CHA			1,617			525	2,142
GVH	186	209	1,643	371	414	430	3,253
HER			7,064	1,887	471	850	10,272
LIM		429	3,420	8	461	675	4,993
NEL						2,571	2,571
RAN			1,931	1,183	226		3,340
RDP			1,394				1,394
SIM	5,223	7,026	6,501	1,347	1,601	674	22,372
TOTALS	5,409	8,575	38,432	5,756	3,173	7,051	68,396

Major Regions for Birth Weight

Genetic Variance %

Chr_mb	Angus	Hereford	Shorthorn	Limousin	Simmental	Gelbvieh
7_93	7.10	5.85	0.01	0.02	0.18	0.02
6_38-39	0.47	8.48	11.63	5.90	16.3	4.75
20_4	3.70	7.99	1.19	0.07	1.53	0.03
14_24-26	0.42	0.01	0.01	0.71	3.05	8.14

Adding Haplotypes
3.20%
5.90%

Imputed 700k
Collective 3 QTL
30% GV

Some of these same regions have big effects on one or more of weaning weight, yearling weight, marbling, ribeye area, calving ease

Sequence

- Now sequencing individual sires
  - Identify loss-of-function alleles to compare to underrepresented haplotype alleles
  - Identify mutations that are perfectly concordant with haplotype allelic effect
    - More powerful across breed

Genomic Prediction

- Exploits advances in quantitative genetics, statistical genetics, computing, molecular biology, and bioinformatics
- Is the basis for some aspects of personalized medicine
- Will revolutionize plant and animal improvement programmes, but to different extents in different industries

Genomic Prediction

- Its application in humans, plants and animals is still an immature but maturing technology
  - Need trait and population specific validation
  - Cannot typically predict “unseen” populations
  - Regression of performance on prediction not 1
  - Reliability upwards biased in “distant” predictions
- Improving the accuracy of genomic prediction will require collaborative efforts

Acknowledgments

- Dr. Rohan Fernando
- Dr. Jack Dekkers
- Dr. Max Rothschild

- Dr. Ania Wolc
- Dr. Bruce Golden

- Dr. Mahdi Saatchi
- Dr. Kadir Kizilkaya
- Dr. David Habier
- Dr. Hailin Su
- Dr. Jungjae Lee
- Dr. Jingjing Yan
- Ziging Weng

- GeneSeek
- Beef Breed Associations
  - American Angus Assoc
  - American Hereford Assoc
  - American Simmental Assoc
  - American Gelbvieh Assoc
  - Red Angus Association
- Aviagen (Broilers)
- HyLine (Layers)
- Livestock Improvement Corp