Deregression and weighting information from various sources

Training on EBVs

Ideal Model (Equation) & data

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{Ma} + \varepsilon$$

$\mathbf{g}$ is (true) genetic merit (BV)

$\mathbf{M}$ is columns of covariates (genotypes)

$\mathbf{a}$ are substitution effects

ε is lack-of-fit (hopefully small)

Ideal Model & data

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{Ma} + \varepsilon$$

$\mathbf{g}$ is genetic merit (BV)

$\mathrm{var}(\mathbf{g}) = \mathbf{A}?$ *or* $\mathbf{G}?$

$\mathrm{var}(\mathbf{Ma}) = \mathbf{G}$ *genomic relationships*

$\mathrm{var}(\varepsilon) = \mathbf{I}\sigma_\varepsilon^2?$ *or* $c\mathbf{A}?$ for c= $\sigma_\varepsilon^2 / \sigma_g^2$

the fraction of var($\mathbf{g}$) unaccounted by markers

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{Ma} + \varepsilon$$

$\mathbf{g}$ is genetic merit (BV)

$\mathrm{var}(\mathbf{g}) = \mathbf{T}\sigma_g^2$ *where* $\mathbf{T}$ *from LD / LA*

$\mathrm{var}(\mathbf{Ma}) = \mathbf{G}\sigma_M^2$ *genomic relationships*

$\mathrm{var}(\varepsilon) = \mathbf{E}\sigma_\varepsilon^2$

(= 0 if markers competely explained merit)

approximate $\mathbf{E}$ as $c\mathbf{A}\sigma_g^2$, for c= $\sigma_\varepsilon^2 / \sigma_g^2$

$\mathbf{Ma}$ is random even if $\mathbf{a}$ is fixed

Towards a Practical Model

$$\mathbf{g} + \mathbf{e} = \mathbf{1}\mu + \mathbf{Ma} + (\varepsilon + \mathbf{e})$$

$\mathbf{g}$ is (true) genetic merit (BV)

$\mathbf{e}$ is usual $\mathbf{e}$

$(\mathbf{g} + \mathbf{e})$ is phenotype (no fixed effects)

$$\mathrm{var}(\varepsilon + \mathbf{e}) = \mathrm{c}\mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2 \quad \text{since } \mathrm{cov}(\varepsilon, \mathbf{e}') = \mathbf{0}$$

Practical Model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Ma} + (\varepsilon + \mathbf{e})$$

$\mathbf{Xb}$ are usual fixed effects

$$\mathrm{var}(\varepsilon + \mathbf{e}) = \mathrm{c}\mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$$

Not reasonable to assume $\mathrm{var}(\varepsilon + \mathbf{e}) = \mathbf{I}\sigma_e^2$

unless markers are fitting very well or

a polygenic effect is fitted

But we do all the time ! And the results are fairly similar

Repeated records on the Individual

$$\bar{\mathbf{y}}_k = \mathbf{Xb} + \mathbf{Ma} + \left(\varepsilon + \bar{\mathbf{e}}_k\right)$$

(i.e. $\mathbf{y}$ is means of varying k numbers of observations)

$$\mathrm{var}(\bar{\mathbf{e}}_k) = \left[\frac{1+(n-1)t}{n} - h^2\right]\sigma_P^2$$

$$\mathrm{var}\left(\varepsilon + \bar{\mathbf{e}}_k\right) = \mathbf{R} = \mathrm{var}(\varepsilon) + \mathrm{var}(\bar{\mathbf{e}}_k)$$

ignoring off-diagonals in $\mathbf{E}$,

$$\mathbf{R}^{-1} = \left[c\sigma_g^2 + \mathrm{var}(\bar{\mathbf{e}}_k)\right]^{-1}$$

$$\frac{w_n}{\sigma_e^2} = \frac{1-h^2}{ch^2 + \dfrac{1+(n-1)t}{n} - h^2} \quad (= 1 \ \ if \ \ c = 0, n = 1)$$

Family Data

When means are from relatives, rather than the same individuals, genetic relationships can contribute to the intraclass correlation

Half-sib offspring averages as data

$$\bar{\mathbf{y}}_p = \mathbf{Xb} + \mathbf{Ma} + \left(\varepsilon + \bar{\mathbf{e}}_p\right)$$

(i.e. $\mathbf{y}$ is means of observations on varying p offspring)

$$\mathrm{var}(\bar{\mathbf{e}}_p) = \left[\frac{0.75\sigma_g^2 + \sigma_e^2}{p}\right]$$

$$\frac{w_p}{\sigma_e^2} = \frac{1 - h^2}{ch^2 + \dfrac{4 - h^2}{p}}$$

EBVs as data

$$\mathbf{g} = \mathbf{Ma} + \varepsilon$$

$$\mathbf{g} + (\hat{\mathbf{g}} - \mathbf{g}) = \hat{\mathbf{g}} = \mathbf{Ma} + \varepsilon + (\hat{\mathbf{g}} - \mathbf{g})$$

$$\textit{with} \;\; \mathrm{var}(\hat{g} - g) = PEV > 0$$

Similar to previous

$$\mathbf{g} + \mathbf{e} = \mathbf{Ma} + \varepsilon + \mathbf{e}$$

where var(g+e) > var(g)

EBVs as data

$$\mathbf{g} = \mathbf{Ma} + \varepsilon$$

$$\mathbf{g} + (\hat{\mathbf{g}} - \mathbf{g}) = \hat{\mathbf{g}} = \mathbf{Ma} + \varepsilon + (\hat{\mathbf{g}} - \mathbf{g})$$

$$\textit{with} \;\; \mathrm{var}(\hat{g} - g) = PEV > 0$$

Generally $\mathrm{var}(\hat{g} - g) = \mathrm{var}(g) + \mathrm{var}(\hat{g}) - 2\,\mathrm{cov}(\hat{g}, g)$

But BLUP has special shrinkage properties

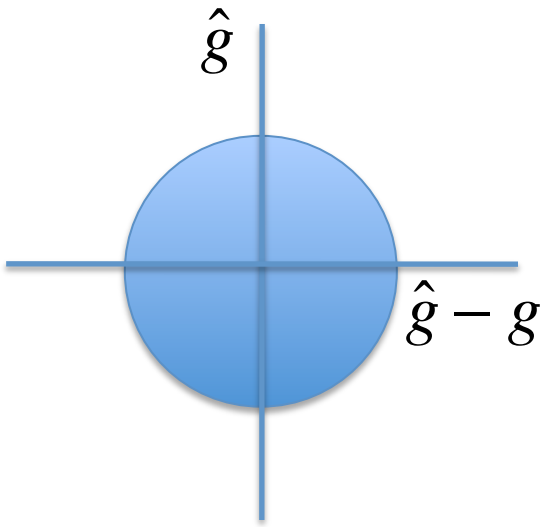
$\mathrm{cov}(\hat{g}, g) = \mathrm{var}(\hat{g})$ so that $\mathrm{var}(\hat{g} - g) = \mathrm{var}(g) - \mathrm{var}(\hat{g})$

$$r^2 = \frac{\mathrm{var}(\hat{g})}{\mathrm{var}(g)} \le 1 \quad \text{so} \quad 0 \le \mathrm{var}(\hat{g}) \le \mathrm{var}(g)$$

Other Relevant Properties of BLUP

$$\begin{aligned}\mathrm{cov}(\hat{g}, \hat{g} - g) &= \mathrm{var}(\hat{g}) - \mathrm{cov}(\hat{g}, g) \\ &= \mathrm{var}(\hat{g}) - \mathrm{var}(\hat{g}) = 0\end{aligned}$$

So prediction errors are uncorrelated with estimated merit



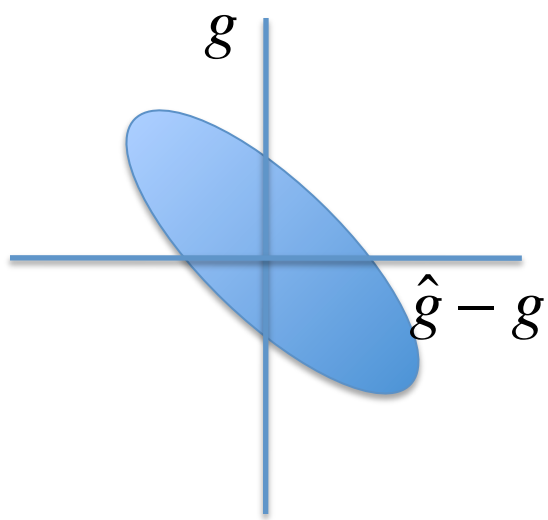
Other Relevant Properties of BLUP

But $\mathrm{cov}(g, \hat{g} - g) = \mathrm{cov}(\hat{g}, g) - \mathrm{var}(g)$

$$= \mathrm{var}(\hat{g}) - \mathrm{var}(g) < 0$$

Really good animals are underestimated

Really bad animals are overestimated



Genomic Prediction

usual linear regression of y on (fixed) x

$$\beta_{y.\mathrm{x}} = \frac{\mathrm{cov}(y,\ \mathrm{x})}{\mathrm{var}(x)},$$

(random) regression of EBV on markers

$\hat{\mathbf{g}}$ on $\mathbf{Ma}$ involves $\mathrm{cov}(\hat{\mathbf{g}},\ \mathbf{Ma}) \approx \mathrm{cov}(\hat{\mathbf{g}},\mathbf{g})$ for small c

But $\mathrm{cov}(\hat{\mathbf{g}},\mathbf{g})=\mathrm{var}(\hat{\mathbf{g}})$ will differ for every animal

according to its accuracy r^2

Need to "inflate" observations

$$\mathbf{g} = \mathbf{Ma} + \varepsilon$$

$$\mathbf{g} + (\mathbf{k}\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{k}\hat{\mathbf{g}} = \mathbf{Ma} + \varepsilon + (\mathbf{k}\hat{\mathbf{g}} - \mathbf{g})$$

Want to choose k so that

$$\mathrm{cov}(\mathbf{g}, \mathbf{k}\hat{\mathbf{g}} - \mathbf{g}) = 0$$

$$\mathrm{cov}(\mathbf{k}\hat{\mathbf{g}}, \mathbf{g}) \text{ to be constant}$$

Finding *k*

Want $\mathrm{cov}(\mathbf{g}, \mathbf{k}\hat{\mathbf{g}} - \mathbf{g}) = 0$

$$\mathrm{cov}(\mathbf{g}, \mathbf{k}\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{k}\mathrm{cov}(\mathbf{g}, \hat{\mathbf{g}}) - \mathrm{var}(\mathbf{g}) = \mathbf{k}\mathrm{var}(\hat{\mathbf{g}}) - \mathrm{var}(\mathbf{g})$$

so we want $\mathbf{k} = \dfrac{\mathrm{var}(\mathbf{g})}{\mathrm{var}(\hat{\mathbf{g}})} = \dfrac{1}{r^2}$

Want $\mathrm{cov}(\mathbf{k}\hat{\mathbf{g}}, \mathbf{g})$ to be constant (test above $\mathbf{k}$)

$$\mathrm{cov}(\mathbf{k}\hat{\mathbf{g}}, \mathbf{g}) = \mathbf{k}\,\mathrm{var}(\hat{\mathbf{g}}) = \frac{\mathrm{var}(\mathbf{g})}{\mathrm{var}(\hat{\mathbf{g}})}\mathrm{var}(\hat{\mathbf{g}}) = \mathrm{var}(\mathbf{g})$$

Implications

Deregress by dividing EBV by their reliability

$$\frac{\hat{g}}{r^2} = d, \;\; a\ deregressed\ EBV\ is$$

$$really\ an\ "observation"\ \ with\ h^2 = r^2$$

$$Observations\ have\ h^2 = \mathrm{cov}(g, y) / \mathrm{var}(p)$$

$$the\ regression\ of\ genotype\ on\ phenotype$$

$$\mathrm{cov}(g, \frac{\hat{g}}{r^2}) = \frac{1}{r^2}\mathrm{var}(\hat{g}) \quad and \;\; \mathrm{var}(\frac{\hat{g}}{r^2}) = \frac{1}{r^4}\mathrm{var}(\hat{g})$$

$$so\ "h^2" = r^4 \big/ r^2 = r^2$$

More Implications

But deregressed observations have heterogeneous variance

$$\mathrm{var}\left[\varepsilon+\left(\mathrm{k}\hat{\mathrm{g}}\text{-}\mathrm{g}\right)\right] \textit{with } k = r^{-2} \textit{ so } kr^2 = 1$$

$$\begin{aligned}\mathrm{var}\left(\varepsilon+\mathrm{k}\hat{\mathrm{g}}\text{-}\mathrm{g}\right) &= \mathrm{var}\left(\varepsilon\right) + \mathrm{var}(\mathrm{k}\hat{\mathrm{g}}\text{-}\mathrm{g}) \\ &= \mathrm{var}\left(\varepsilon\right) + k^2\mathrm{var}\left(\hat{\mathrm{g}}\right) + \mathrm{var}\left(\mathrm{g}\right) - 2k\,\mathrm{var}(\hat{\mathrm{g}}) \\ &= \mathrm{var}\left(\varepsilon\right) + k^2 r^2\mathrm{var}\left(\mathrm{g}\right) + \mathrm{var}\left(\mathrm{g}\right) - 2kr^2\,\mathrm{var}(\mathrm{g}) \\ &= \mathrm{var}\left(\varepsilon\right) + \left(k-1\right)\mathrm{var}\left(\mathrm{g}\right) \textit{ and } k-1 = \frac{1-r^2}{r^2}\end{aligned}$$

Therefore the weights representing diagonals of $\mathbf{R}^{-1}$ are

$$\frac{w}{\sigma_e^2} = \frac{1-h^2}{\left[c+\left(1-r^2\right)/r^2\right]h^2}$$

Removing Parent Average

During the deregression process, parent average effects should be removed

Why?

Animals with own and/or progeny information are shrunk towards the parent average

Imagine if many bulls had no own/progeny info

They should not contribute anything to training

Imagine if some parents were segregating a major effect

We dont want this effect shrunk in all the offspring

Deregression is no problem if deregressed information is derived directly from animal models during evaluation

Removing Parent Average

Deregression and removal of parent average effects can be approximately achieved using only the EBV and r^2 values from trios of the training animal, its sire and dam, by setting up mixed model equations for the parent average and offspring, reconstructing the left-hand side to obtain the published reliabilities, before reconstructing the implied right-hand side to determine the deregressed observation and its appropriate r^2 ignoring the parental contribution

Genomic Selection

Value in QTL detection and links to bioinformatics

Dorian Garrick

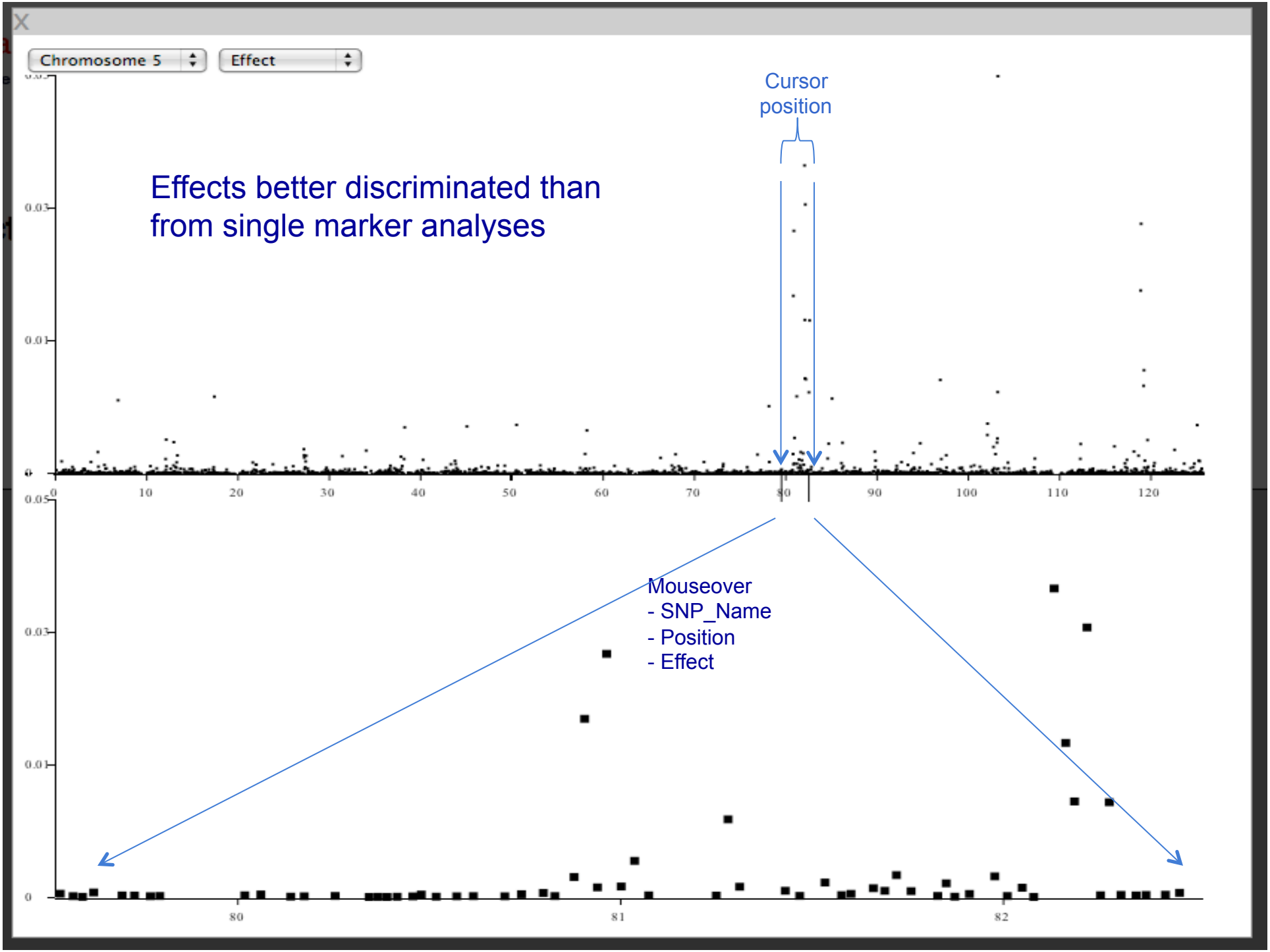
dorian@iastate.edu

Animal Science

1858 2008 150

Iowa State University

Animal Breeding & Genetics


Chromosome 5
Effect
Cursor position
Effects better discriminated than from single marker analyses
0.03
0.01
0
0
10
20
30
40
50
60
70
80
90
100
110
120
0.05
0.03
0.01
0
80
81
82
Mouseover
- SNP_Name
- Position
- Effect

Cattle genome browser (NCBI Data): Chr.15:4511269..14511268

http://www.animalgenome.org/cgi-bin/gbrowse/cattle/#search

Genetics Sel...n Evolution | ESPN Sports | Wolfram|Alpha | Country-Wide | WCGALP2010 | SNPLOTZ | GeneSeek - ...ta download | Flint Rodent Data | ISU ABG Seminars | Animal Bree...n p/garrick

Cattle Genome Track — QTL, Coding regions, transcripts, SNPs, etc.

Showing 10 Mbp from Chr.15, positions 4,511,269 to 14,511,268

Instructions

Searching: Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed.

Navigation: Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.

[Bookmark this] [Upload your own data] [Hide banner] [Share these tracks] [Link to Image] [High-res Image] [Help] [Reset]

Search

Chr.1, Chr.2, Chr.3, Chr.4, Chr.5, Chr.6, Chr.7, Chr.8, Chr.9, Chr.10, Chr.11, Chr.12, Chr.13, Chr.14, Chr.15, Chr.16, Chr.17, Chr.18, Chr.19, Chr.20, Chr.21, Chr.22, Chr.23, Chr.24, Chr.25, Chr.26, Chr.27, Chr.28, Chr.29, Chr.X

Landmark or Region:

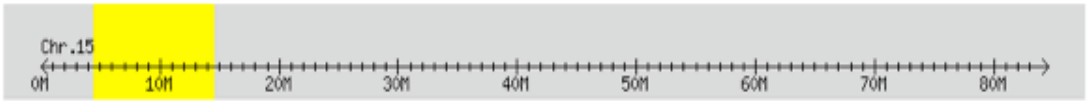
Chr.15:4511269..14511268 Search

Data Source

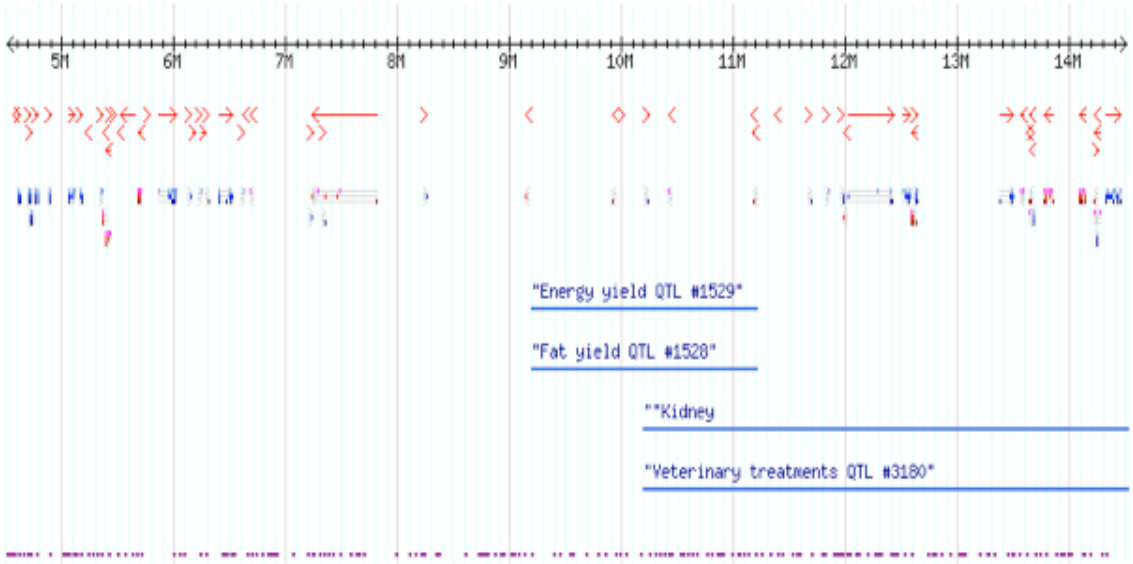
Cattle genome browser (NCBI Data)

Scroll/Zoom: Show 10 Mbp **Flip**

Overview



Details



Clear highlighting

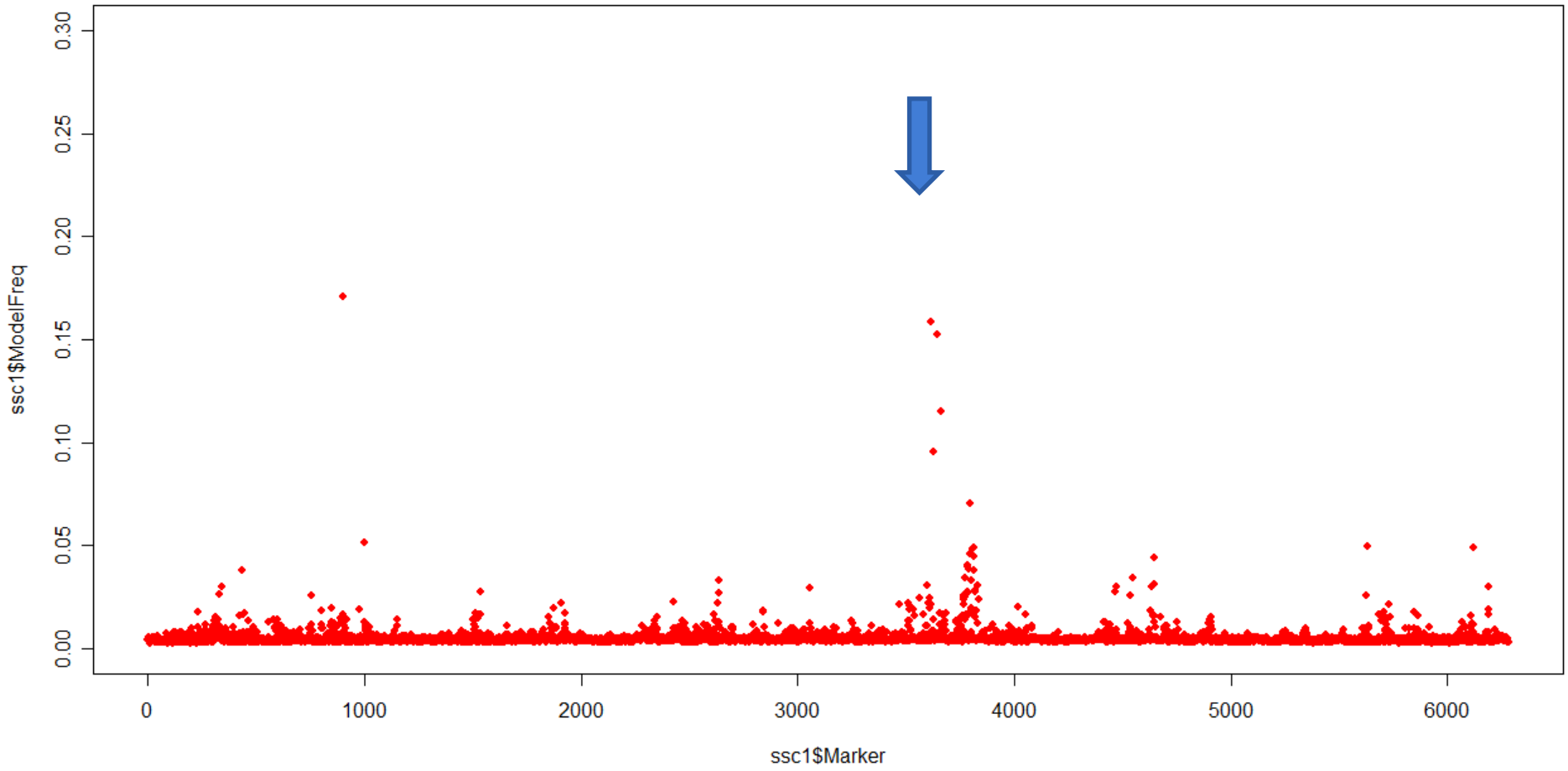
Update Image

Tracks

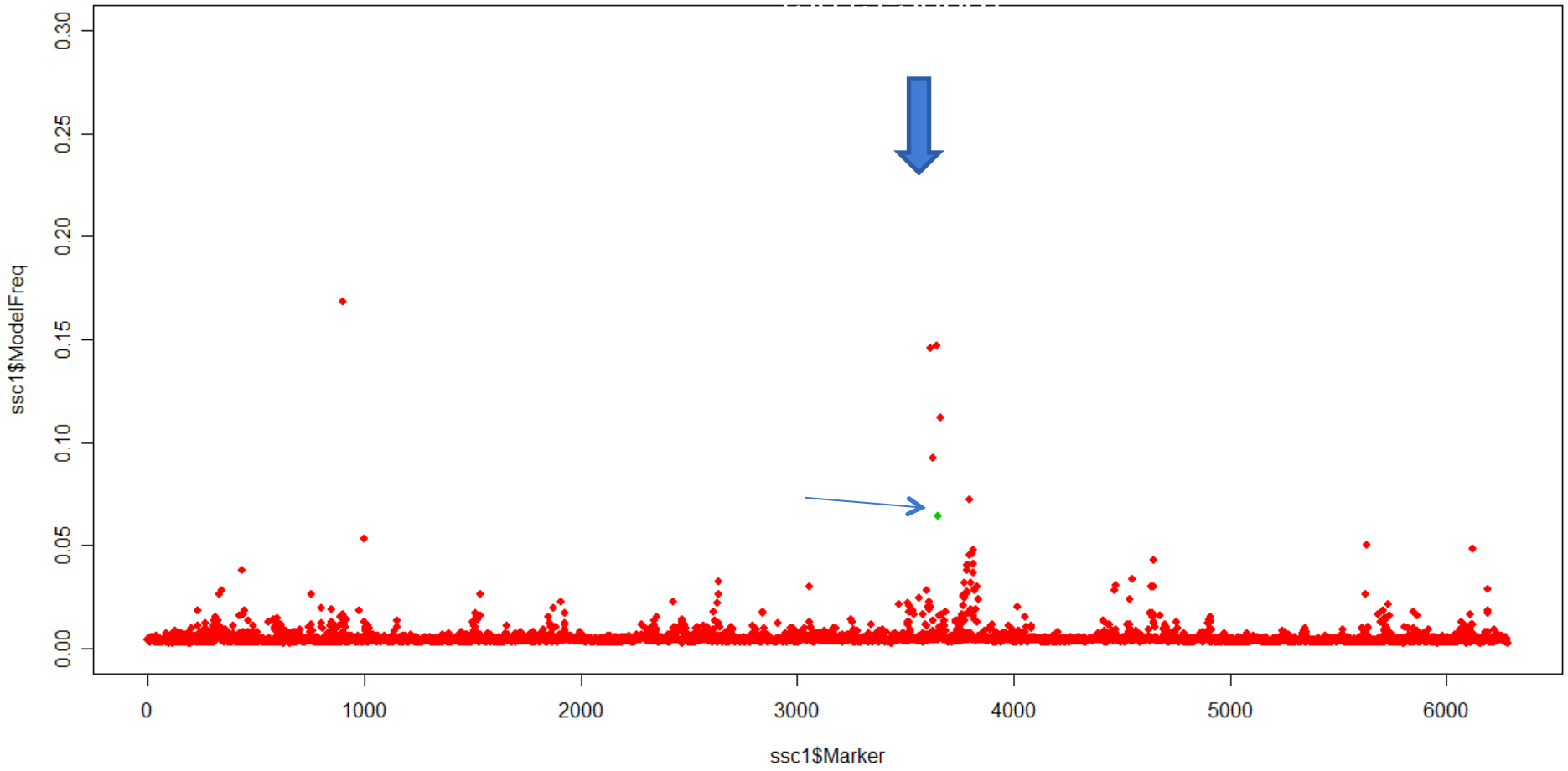
Display Settings

Add your own tracks

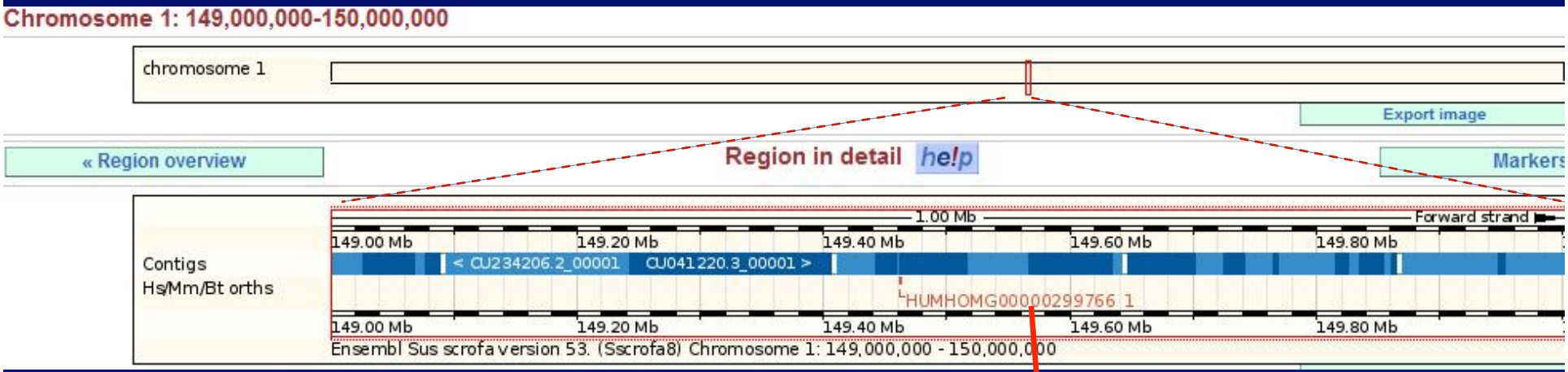
Strong signal for subcutaneous fatdepth near *MC4R* gene on SSC1

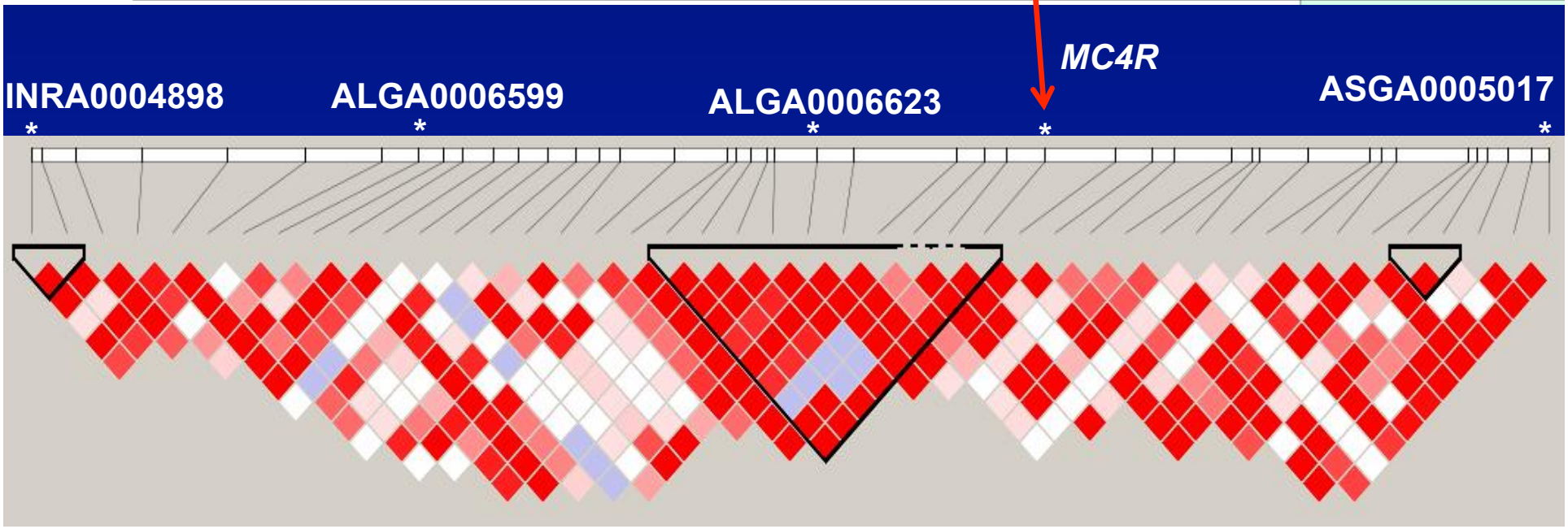


Effect not due to patented polymorphism in *MC4R* gene

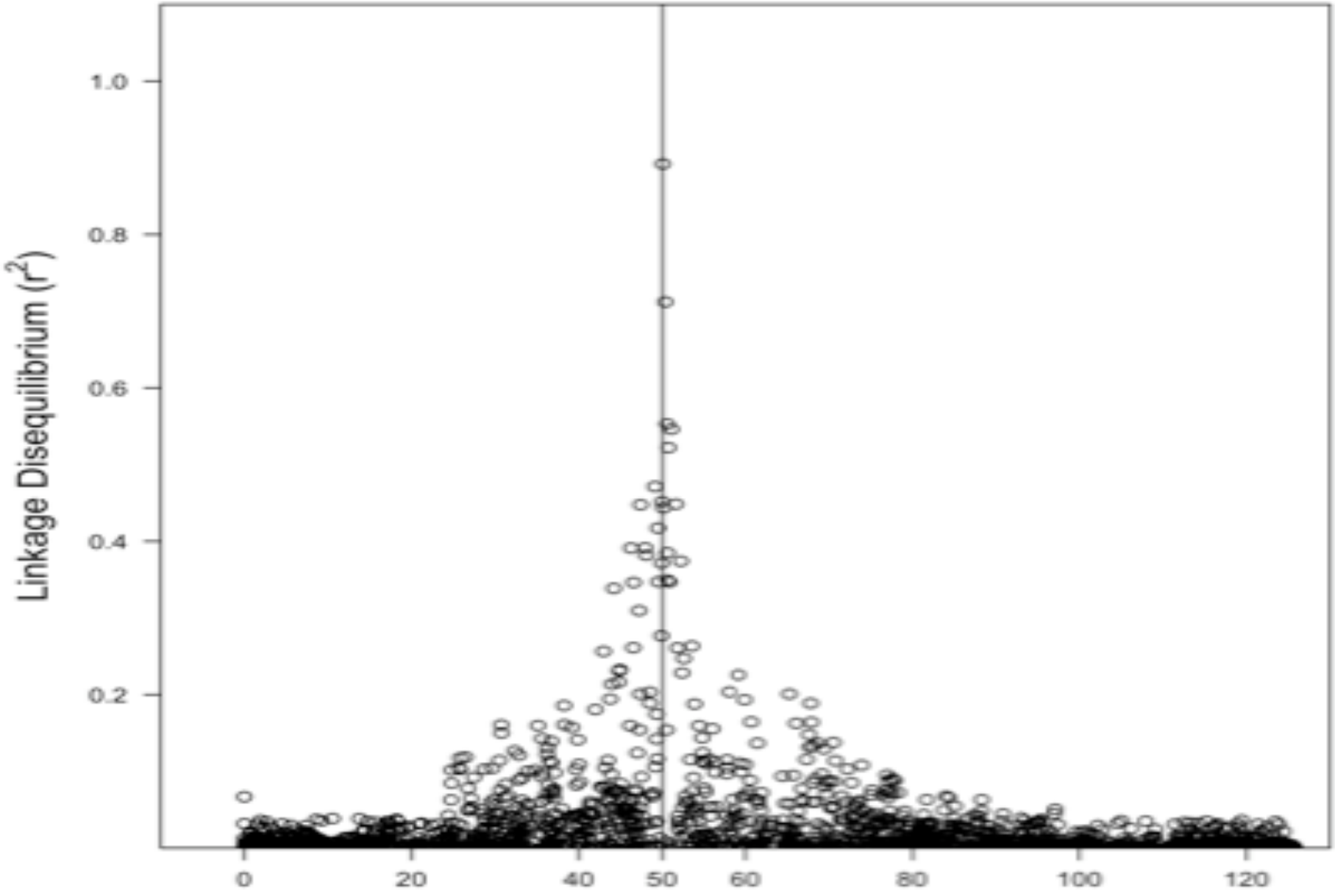


Strong LD detected in the region containing the *MC4R* gene on SSC1

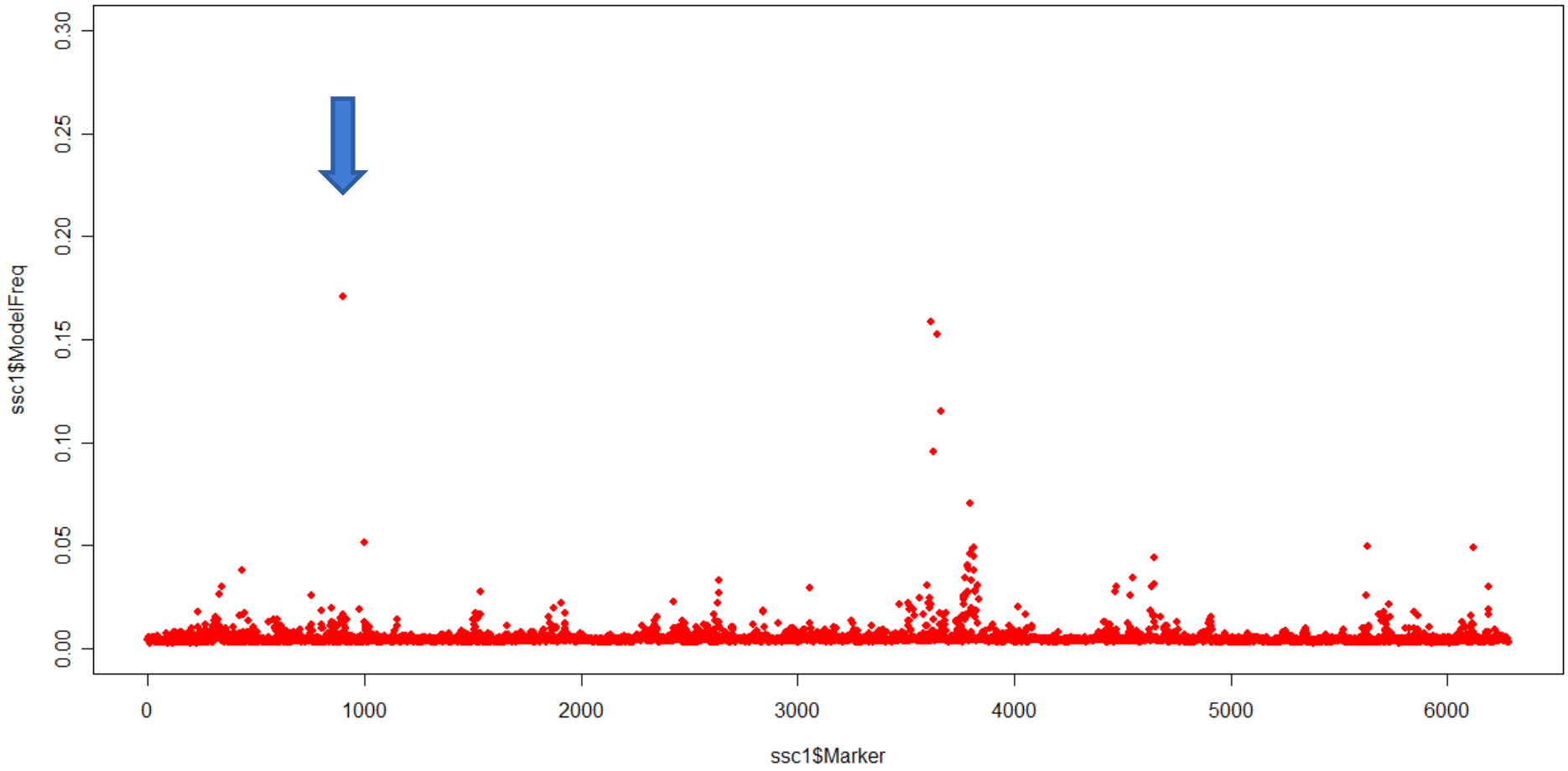


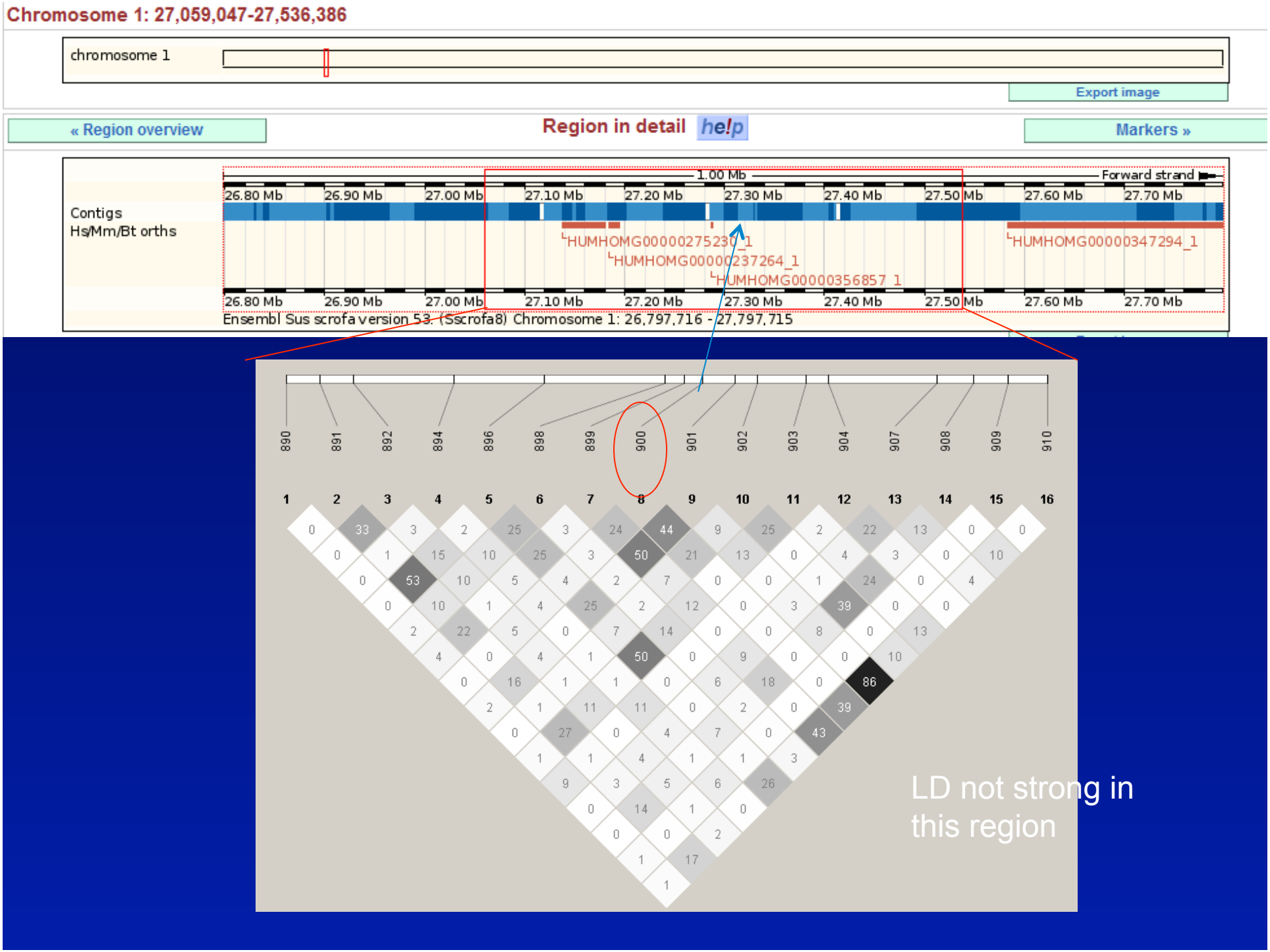


One Informative Locus

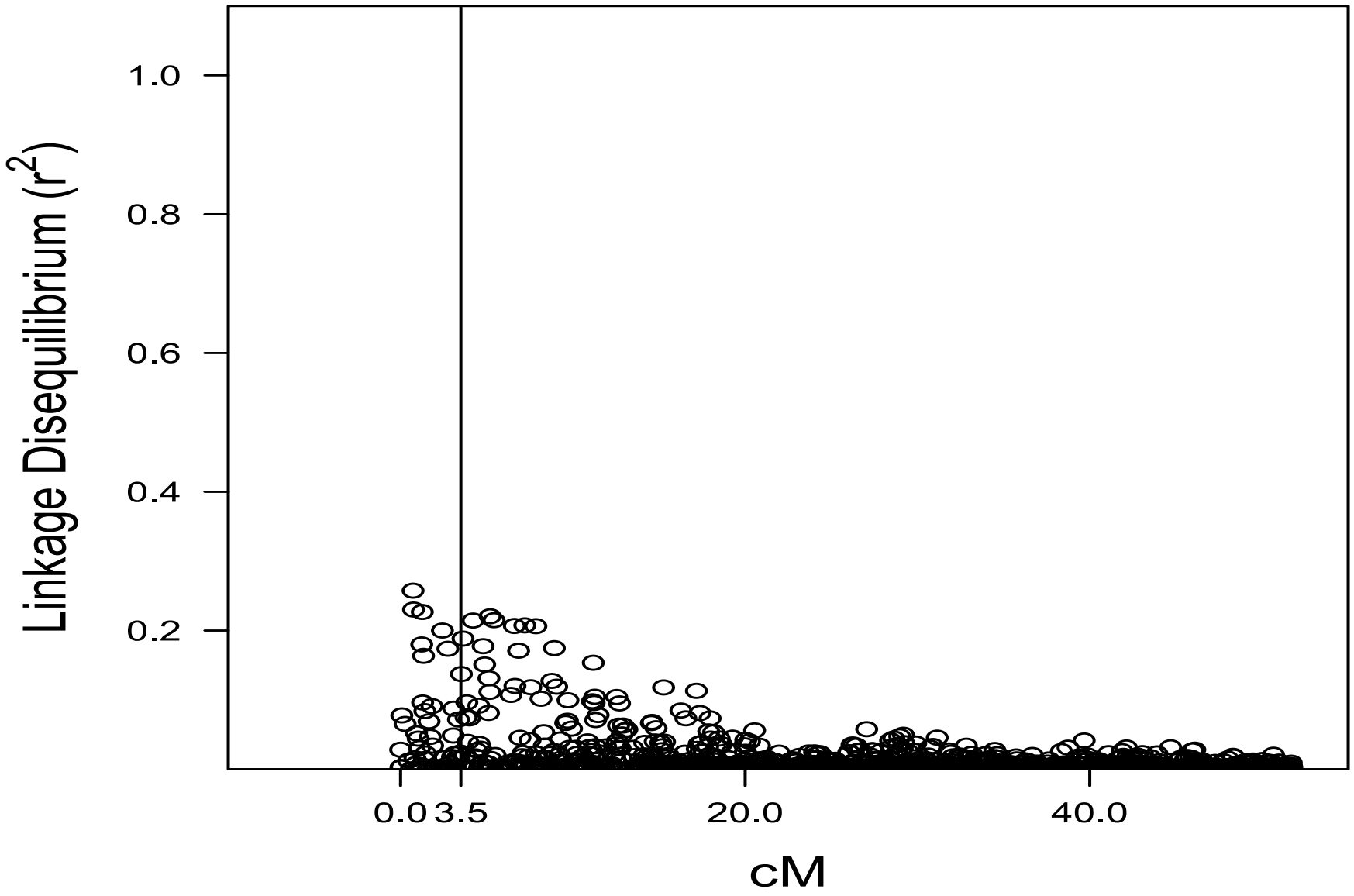


Strong signal for subcutaneous fatdepth near *MC4R* gene on SSC1




Chromosome 1: 27,059,047-27,536,386
chromosome 1
Export image
« Region overview
Region in detail
help
Markers »
1.00 Mb
Forward strand
26.80 Mb
26.90 Mb
27.00 Mb
27.10 Mb
27.20 Mb
27.30 Mb
27.40 Mb
27.50 Mb
27.60 Mb
27.70 Mb
Contigs
Hs/Mm/Bt orths
HUMHOMG00000237264_1
HUMHOMG00000356857_1
HUMHOMG00000347294_1
Ensembl Sus scrofa version 53. (Sscrofa8) Chromosome 1: 26,797,716 - 27,797,715
890
891
892
894
896
898
899
900
901
902
903
904
907
908
909
910
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
LD not strong in this region

Another informative locus



Conclusion

- Genomic Selection information should be part of a large scale bioinformatics system to properly exploit the gene discovery knowledge generated

Using real-life (Illumina) genotypes

```
[Header]
BSGT Version    3.3.4
Processing Date 3/20/2009 11:20 PM
Content         Kit-OvineSNP50_11330224_D.bpm
Num SNPs        54977
Total SNPs      54977
Num Samples     60
Total Samples   60
[Data]
SNP Name                SampleID Allele1-Forward Allele2-Forward GCScore X Y
250506CS3900065000002_1238.1    1       C       C       0.9239  0.039   1.031
250506CS3900140500001_312.1     1       C       C       0.9613  0.003   0.631
250506CS3900176800001_906.1     1       T       T       0.9573  0.869   0.023
250506CS3900211600001_1041.1    1       G       G       0.9504  0.006   0.772
250506CS3900218700001_1294.1    1       T       C       0.9061  0.380   0.545
250506CS3900283200001_442.1     1       A       C       0.9622  0.334   0.353
250506CS3900371000001_1255.1    1       T       C       0.9705  0.226   0.302
250506CS3900386000001_696.1     1       T       C       0.9024  0.776   0.781
250506CS3900414400001_1178.1    1       C       C       0.9593  0.011   0.863
250506CS3900435700001_1658.1    1       A       A       0.4358  0.797   0.002
250506CS3900464100001_519.1     1       T       T       0.9461  0.939   0.008
250506CS3900487100001_1521.1    1       A       G       0.8954  0.803   0.784
250506CS3900539000001_471.1     1       T       T       0.9596  0.553   0.016
250506CS3901012300001_913.1     1       T       C       0.8424  1.091   0.993
250506CS3901300500001_1084.1    1       C       C       0.8444  0.047   1.303
```

```
[Header]
BSGT Version     3.3.4
Processing Date  1/12/2009 1:06 AM
Content          BovineSNP50_B.bpm
Num SNPs         54001
Total SNPs       54001
Num Samples      1319
Total Samples    1319
[Data]
SNP Name         Sample ID        Allele1 - AB     Allele2 - AB     X        Y        GC Score
BFGL-NGS-109695  157      B       B        0.020   1.450    0.7266
BFGL-NGS-109696  157      B       B        0.009   1.206    0.8765
BFGL-NGS-109701  157      B       B        0.047   1.185    0.8113
BFGL-NGS-109702  157      A       B        0.521   0.886    0.3152
BFGL-NGS-109705  157      B       B        0.035   1.115    0.7284
BFGL-NGS-109707  157      A       B        1.052   0.906    0.7790
BFGL-NGS-109711  157      B       B        0.019   1.137    0.8765
BFGL-NGS-109712  157      A       B        0.308   0.656    0.7617
BFGL-NGS-109714  157      A       A        1.254   0.060    0.9328
BFGL-NGS-109716  157      A       B        0.540   0.804    0.8402
BFGL-NGS-109720  157      A       A        0.875   0.028    0.8809
BFGL-NGS-109722  157      B       B        0.016   0.937    0.8081
```

2½ - 3 Gb files per 1,000 animals

Index	Name	Chromosome	Position	GenTrain	SNP	ILMN Strand	
1	250506CS3900065000002_1238.1	15	5327353	0.8867	[A/G]	TOP	BOT
2	250506CS3900140500001_312.1	23	27428869	0.9323	[A/G]	TOP	BOT
3	250506CS3900176800001_906.1	7	89002990	0.9266	[T/C]	BOT	BOT
4	250506CS3900211600001_1041.1	16	44955568	0.9173	[A/C]	TOP	BOT
5	250506CS3900218700001_1294.1	2	157820235	0.8692	[A/G]	TOP	BOT
6	250506CS3900283200001_442.1	1	203289635	0.9335	[A/C]	TOP	BOT
7	250506CS3900371000001_1255.1	11	37632867	0.9464	[T/C]	BOT	BOT
8	250506CS3900386000001_696.1	16	68297712	0.8658	[A/G]	TOP	TOP
9	250506CS3900414400001_1178.1	1	111100644	0.9294	[T/C]	BOT	TOP

Roughly 1m bp per cM

Recode SNP names to your own index identifier

```
20071018I10973 AA AB BB BB AB AB BB AB AA
20071018I10975 AA BB BB BB BB AA BB AB AB
20071018I10977 AA BB BB BB BB AA BB AA AA
20071018I10979 AA AB BB BB BB AA BB AB AA
20071018I10981 AA BB BB BB BB AA BB AA BB
20071018I10983 AA BB BB BB BB AA BB AB AB
20071018I10985 AA BB BB BB AB AA BB BB AA
20071018I10987 AA BB BB BB BB AA BB AB AA
20071018I10989 AA BB BB BB BB AA BB AB AB
20071018I10991 AA BB BB BB AB AA BB AB AB
20071018I10993 AA BB BB BB BB AA BB BB AA
20071018I10995 AA BB BB BB AB AA BB AB AA
```

Quality control checks
- minor allele frequency
- Hardy-Weinberg equilibrium
- parentage agreement with pedigree

Quality control files
- by locus
- by sample

```
20071018I10973 AA AB BB BB AB AB BB AB AA
20071018I10975 AA BB BB BB BB AA BB AB AB
20071018I10977 AA BB BB BB BB AA BB AA AA
20071018I10979 AA AB BB BB BB AA BB AB AA
20071018I10981 AA BB BB BB BB AA BB AA BB
20071018I10983 AA BB BB BB BB AA BB AB AB
20071018I10985 AA BB BB BB AB AA BB BB AA
20071018I10987 AA BB BB BB BB AA BB AB AA
20071018I10989 AA BB BB BB BB AA BB AB AB
20071018I10991 AA BB BB BB AB AA BB AB AB
20071018I10993 AA BB BB BB BB AA BB BB AA
20071018I10995 AA BB BB BB AB AA BB AB AA
```

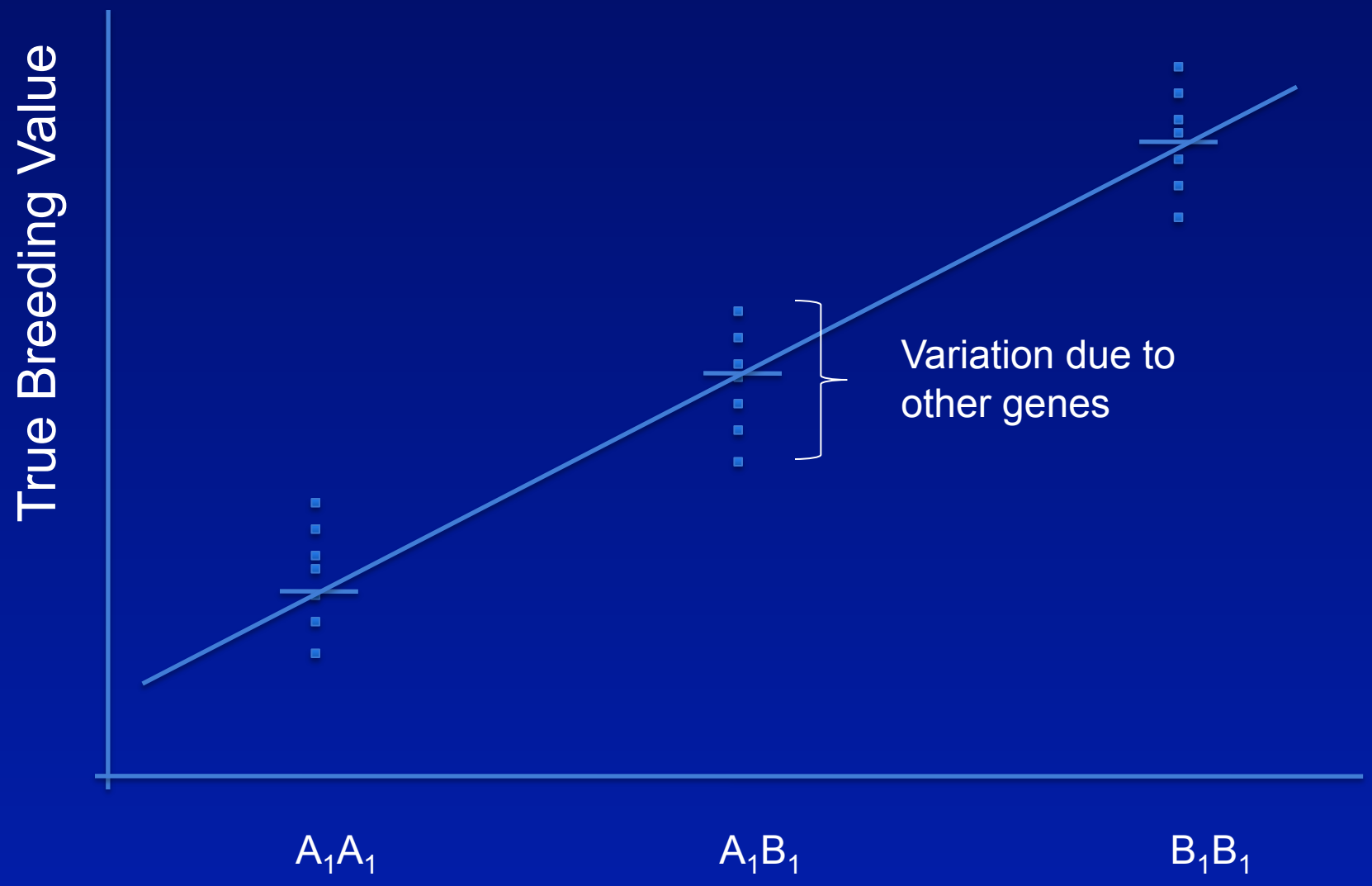
Convert every pair of alleles to a covariate

Consistent allele calling e.g. AA= -10, AB=0, B+10
1Gb storage for 10,000 animals

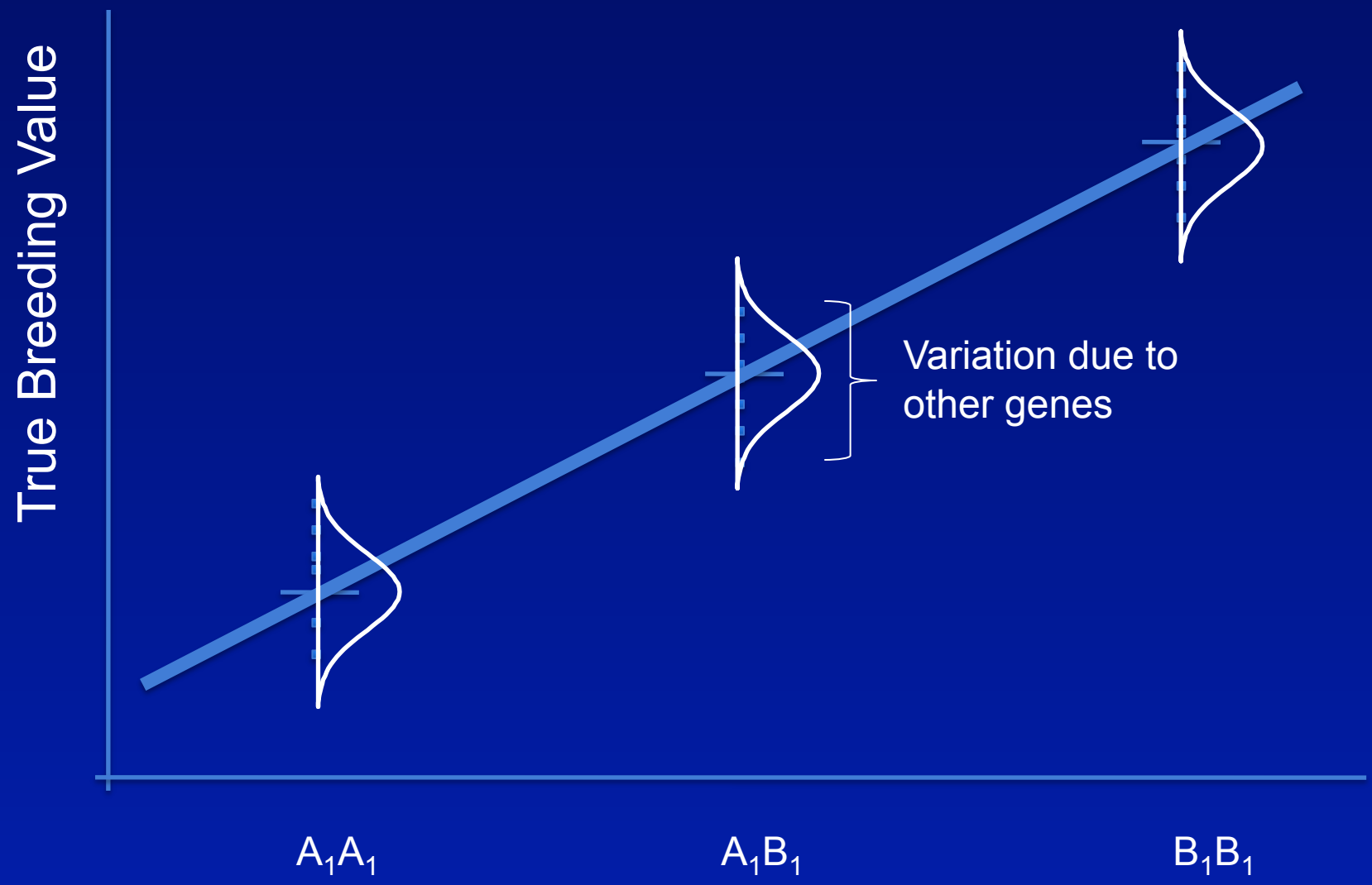
```
WG0056939-DNAA02_A990182 -10 10 10 0 0 -10 10 0 0
WG0056939-DNAA03_A990761 -10 10 10 10 10 -10 10 -10 0
WG0056939-DNAA04_A990802 -10 0 10 10 0 -10 10 -10 -1
WG0056939-DNAA05_A990027 -10 10 10 0 10 -10 10 0 -1
WG0056939-DNAA06_A990038 -10 0 10 0 10 -10 10 -10 0
WG0056939-DNAA07_A990770 -10 10 10 10 0 -10 10 0 10
WG0056939-DNAA08_A990502 -10 0 10 10 10 10 10 0 0
WG0056939-DNAA09_A990515 -10 10 10 10 10 0 10 -10 -10
WG0056939-DNAA10_A990564 -10 10 10 10 0 -10 10 -10 -10
WG0056939-DNAA11_A000684 -10 10 10 10 -10 -10 10 0 10
WG0056939-DNAB01_A001214 -10 10 10 10 -10 -10 10 0 10
```

Linkage Disequilibrium

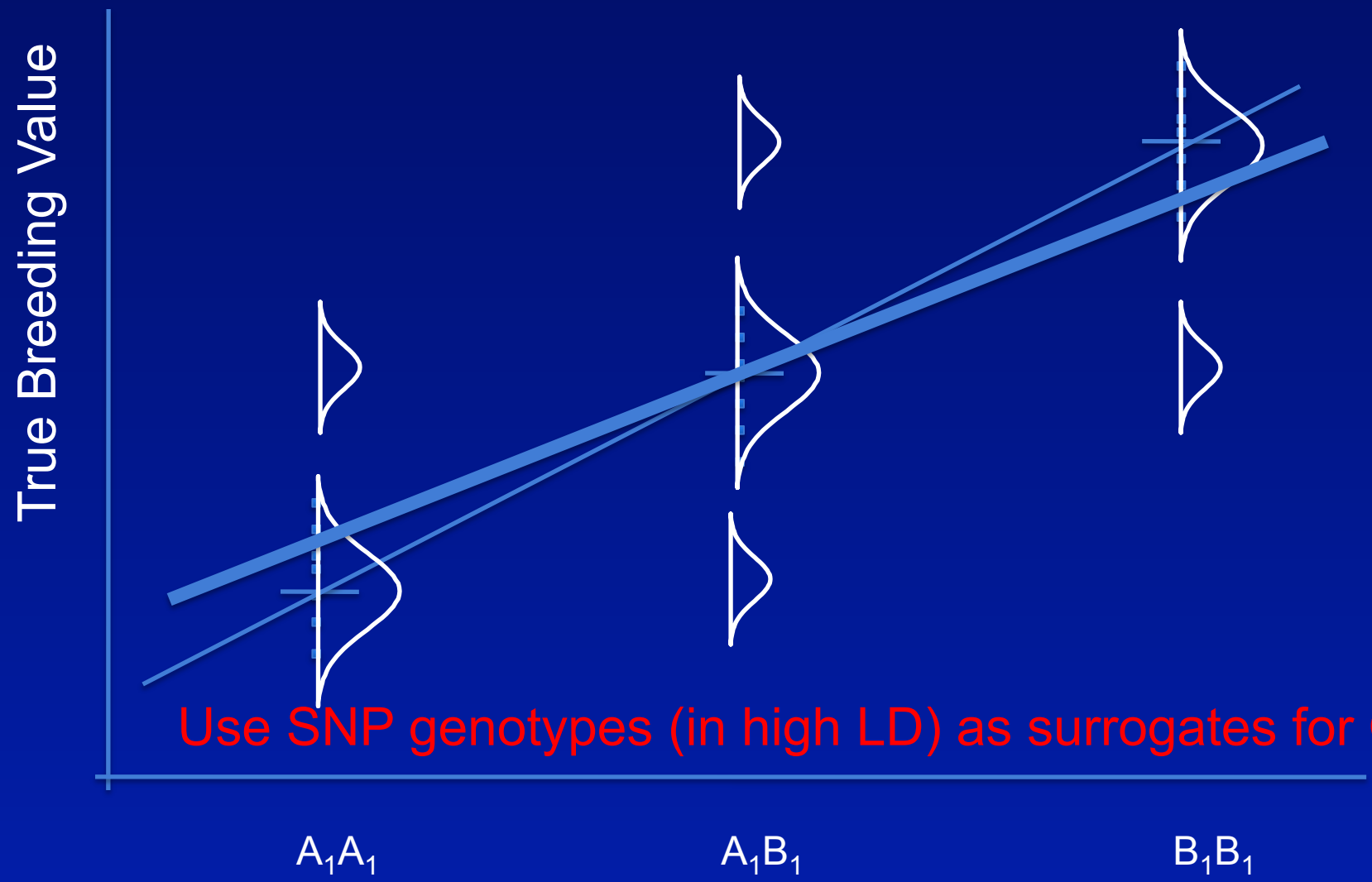
Overall intent – BV on QTL



Overall intent – BV on QTL

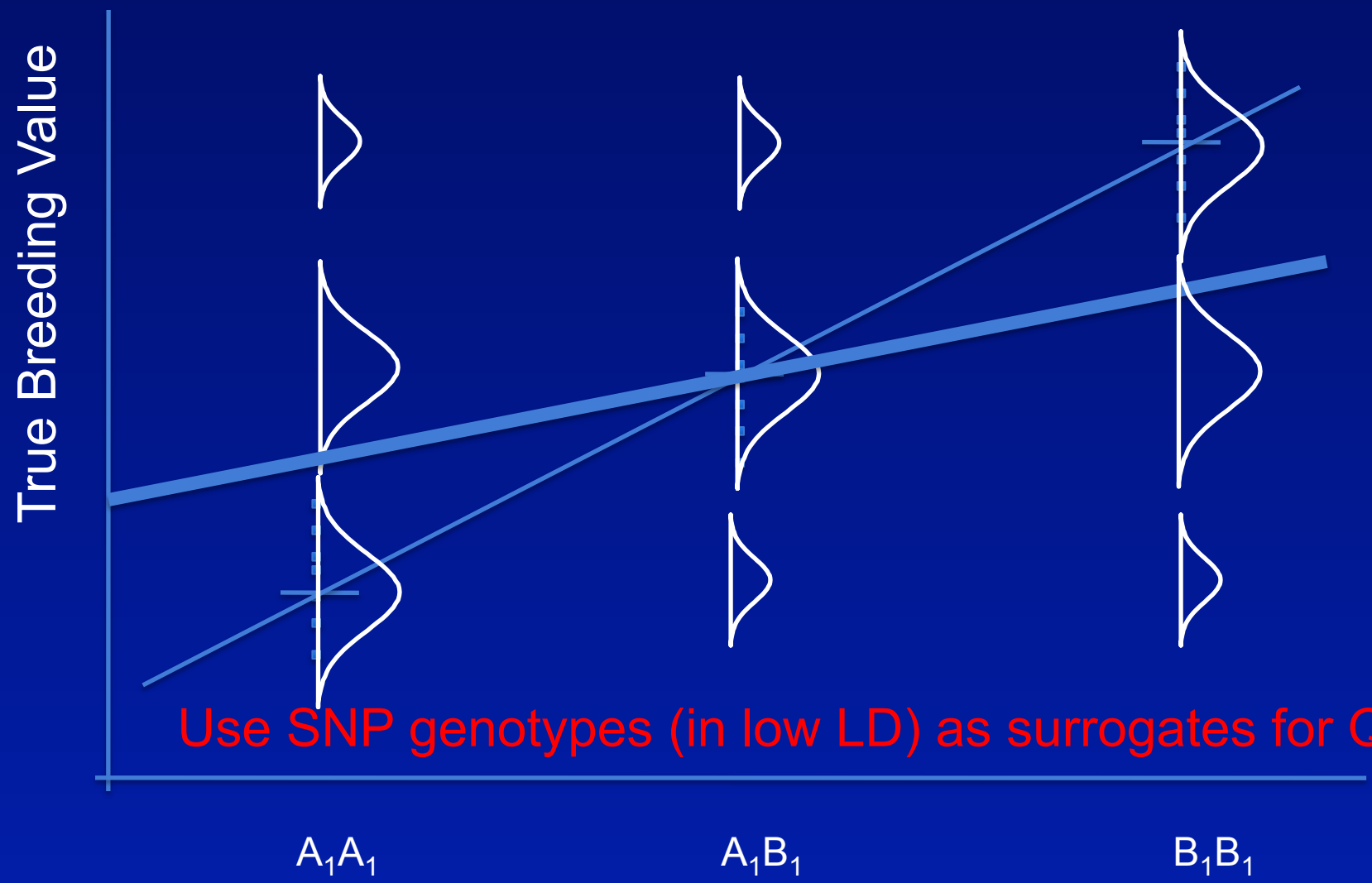


Practice – BV on SNP



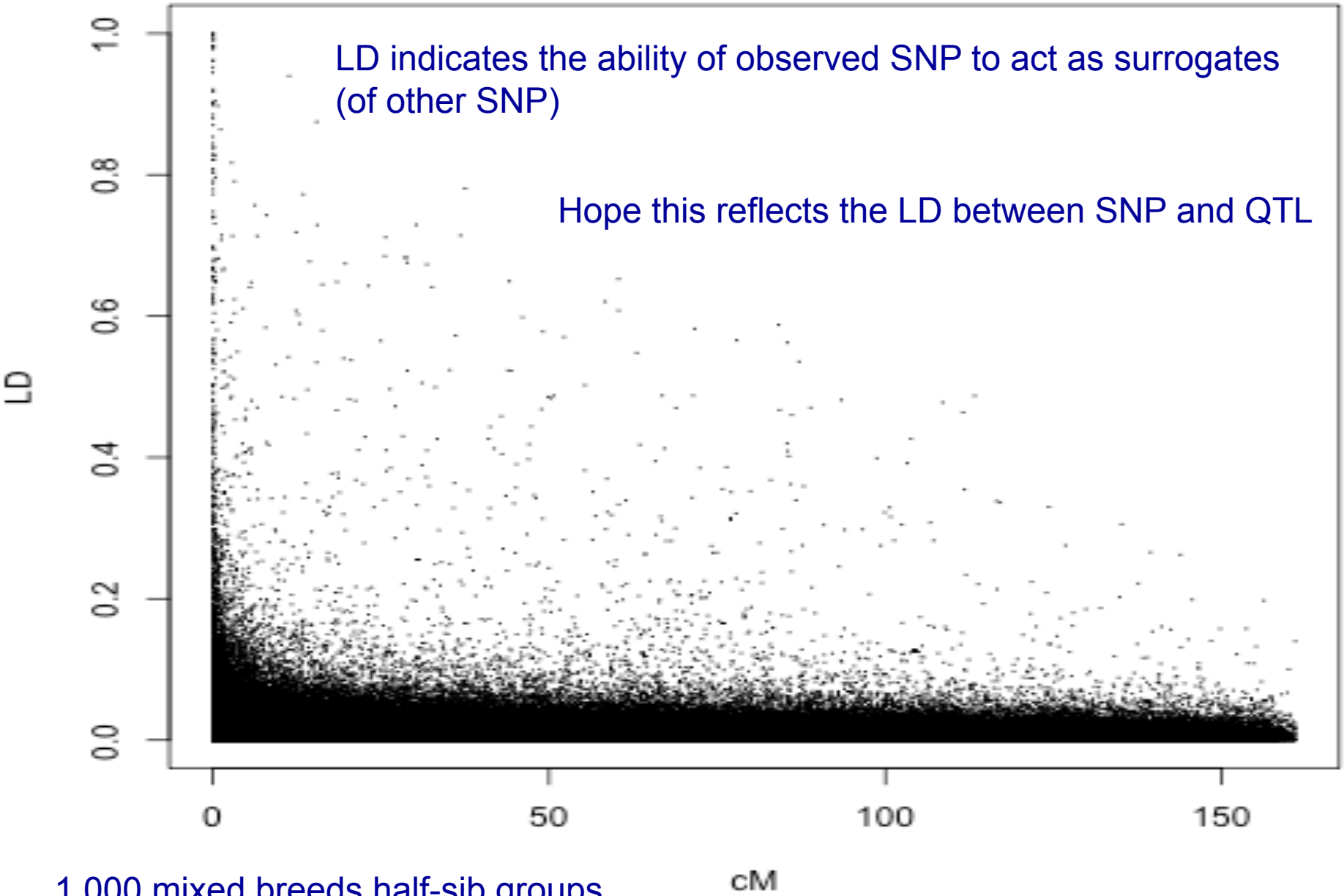
Use SNP genotypes (in high LD) as surrogates for QTL

Practice – BV on SNP



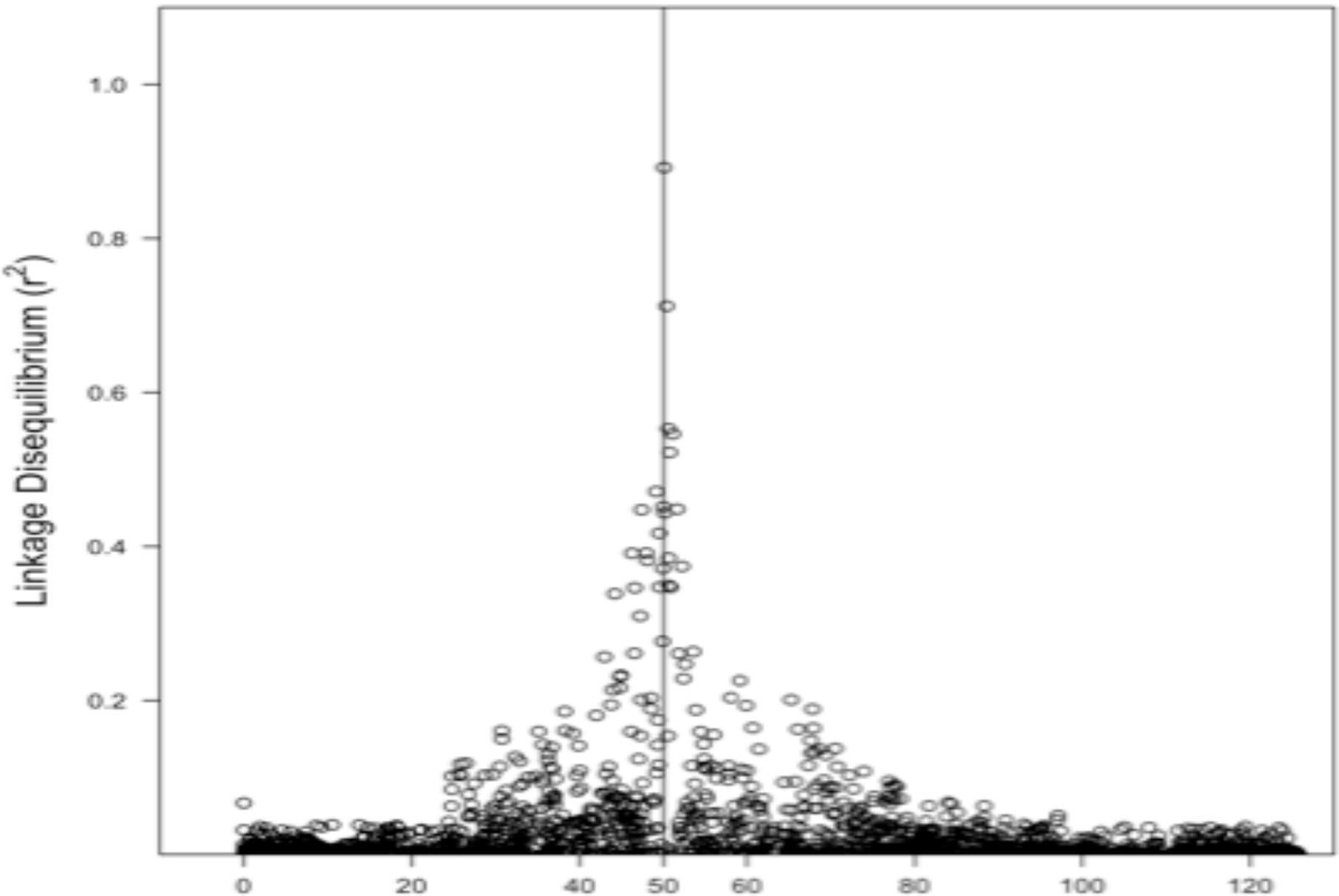
Use SNP genotypes (in low LD) as surrogates for QTL

Linkage Disequilibrium (LD) on bovine chromosome 1

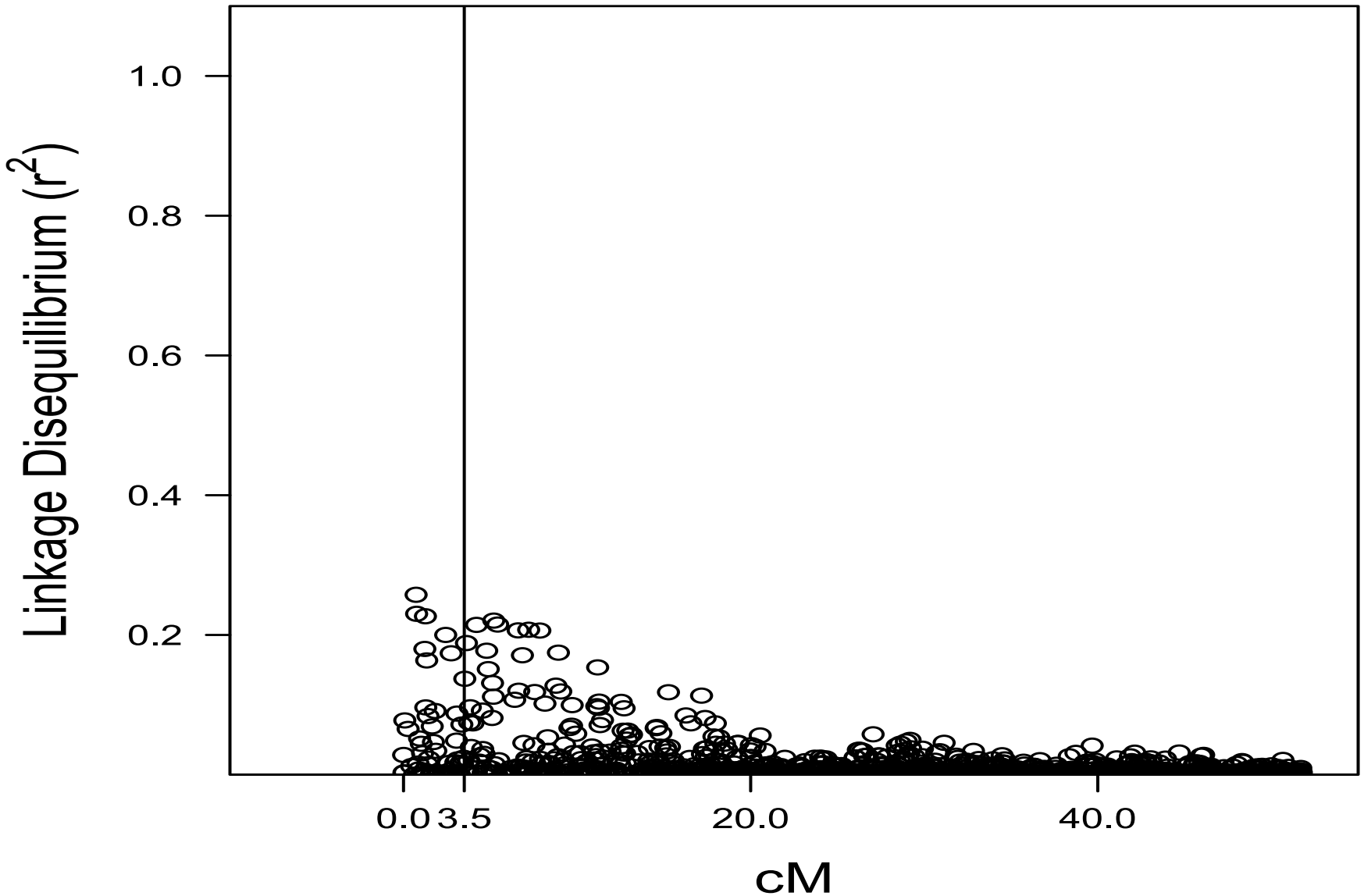


1,000 mixed breeds half-sib groups

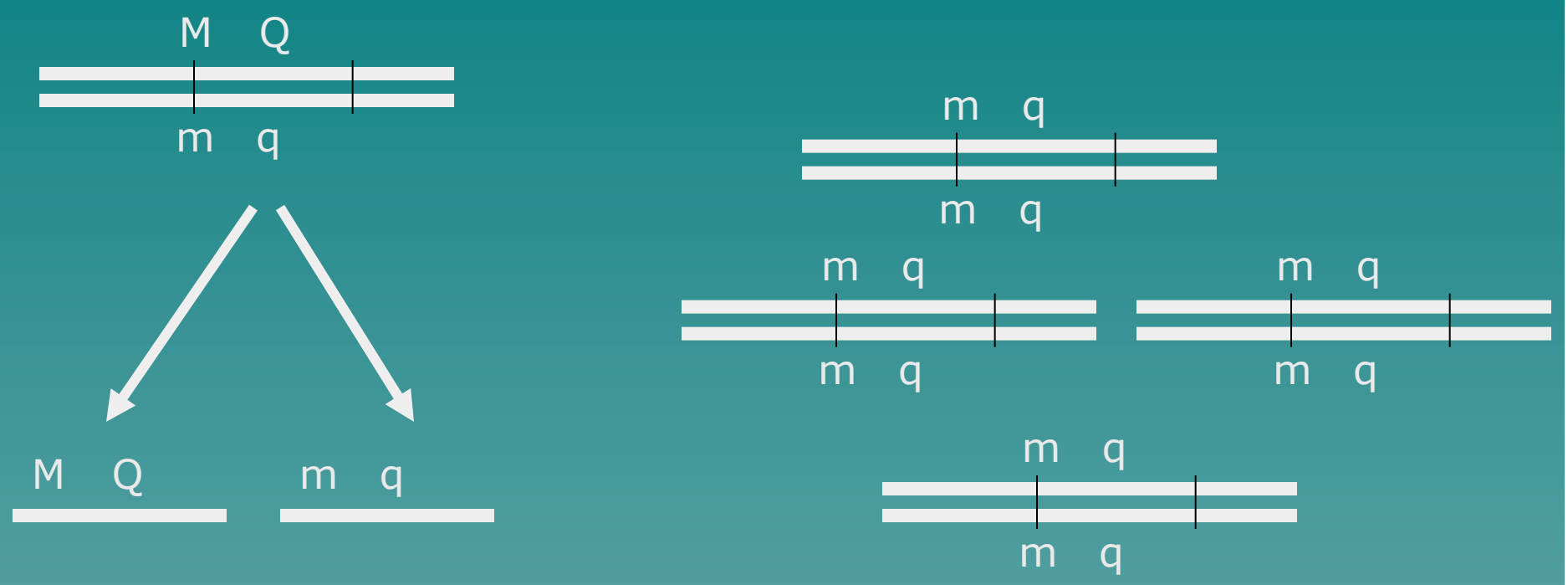
One Informative Locus



Another informative locus



Linkage Disequilibrium (LD)



Hardy-Weinberg Equilibrium

After a few generations, suppose freq(M)=0.2

Marker genotypes

MM	Mm	mm
0.04	0.32	0.64

QTL genotypes

- QQ
- Qq
- qq

Hardy-Weinberg Equilibrium

After a few generations, suppose freq(M)=0.2

Marker genotypes

MM	Mm	mm
0.04	0.32	0.64

QTL genotypes

- QQ
- Qq
- qq

And suppose M was "close enough" to Q that a crossover between them never occurred then freq(Q)=0.2

Hardy-Weinberg Equilibrium

After a few generations, suppose freq(M)=0.2=freq(Q)

Marker genotypes

MM	Mm	mm
0.04	0.32	0.64

QTL genotypes

QQ	0.04
Qq	0.32
qq	0.64

Hardy-Weinberg Equilibrium (& LD)

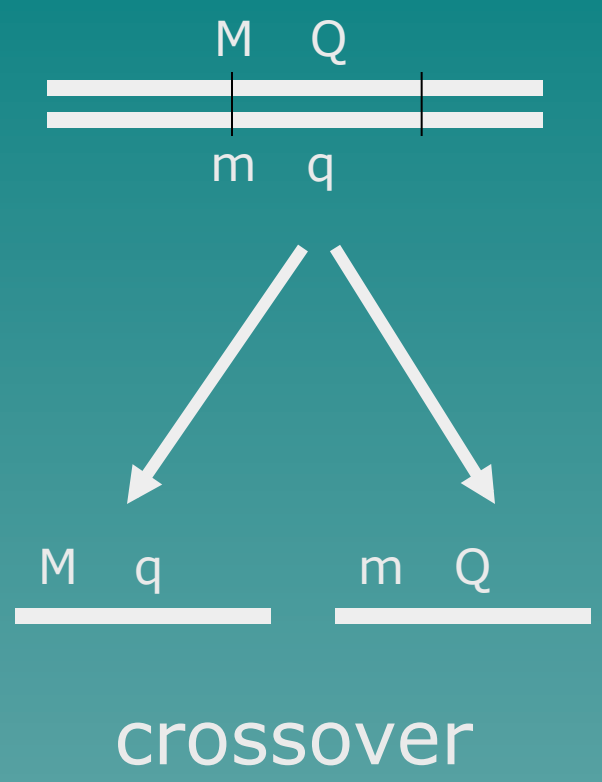
After a few generations, suppose freq(M)=0.2=freq(Q)

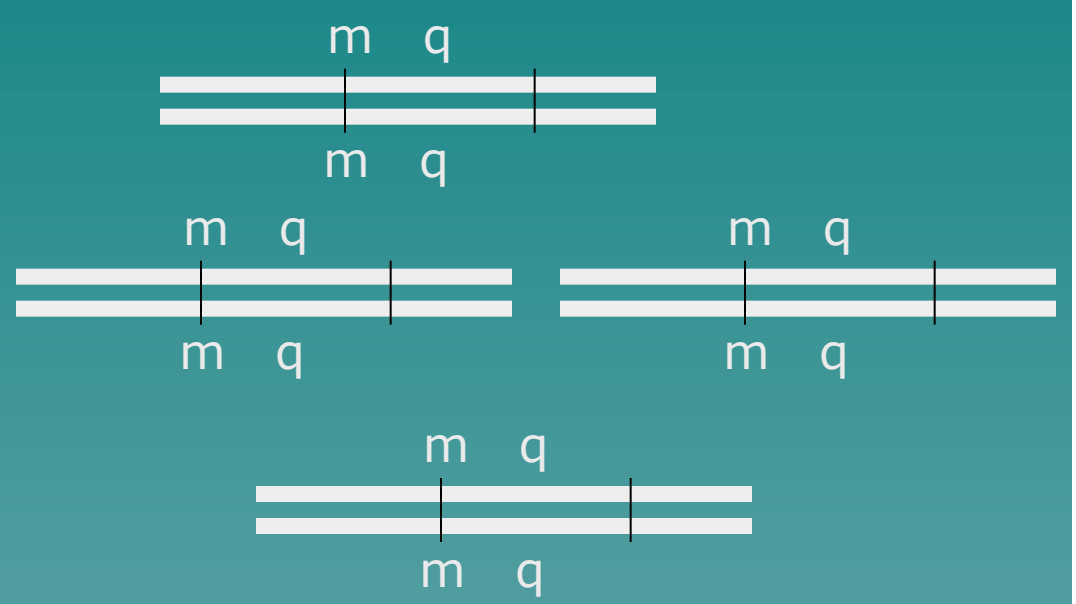
			Marker genotypes		
			MM	Mm	mm
			0.04	0.32	0.64
QTL genotypes	QQ	0.04	0.04		
	Qq	0.32		0.32	
	qq	0.64			0.64

Linkage Disequilibrium

Then LD is perfect & M is a direct indicator of the presence of Q

Linkage Equilibrium (LE)





Hardy-Weinberg Equilibrium

After more generations with no change in gene frequencies

Marker genotypes

MM	Mm	mm
0.04	0.32	0.64

QTL genotypes

QQ	0.04
Qq	0.32
qq	0.64

Hardy-Weinberg Equilibrium (& LE)

After more generations with no change in gene frequencies

			Marker genotypes		
			MM	Mm	mm
			0.04	0.32	0.64
QTL genotypes	QQ	0.04	.0016	.0128	.0256
	Qq	0.32	.0128	.1024	.2048
	qq	0.64	.0256	.2048	.4096

Linkage Equilibrium

Then LE is perfect & M tells nothing about the presence of Q

Linkage Equilibrium (LE)

- But individual chromosome segments can only be one of four

M Q

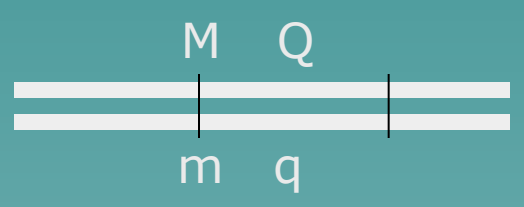
M q

m Q

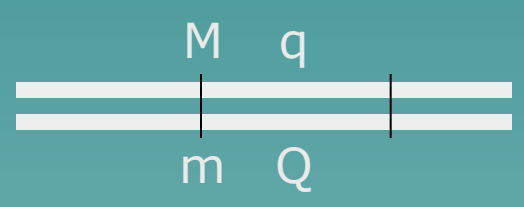
m q

Linkage Equilibrium (LE)

- So provided an animal is heterozygous for the marker and heterozygous for the QTL allele then we can use the marker provided we know the phase or marker-QTL haplotype



M indicates Q



m indicates Q

Forces modifying LE/LD

- Continuously operating factors
  - Drift/inbreeding
    - Especially small populations
  - Recurrent migration
    - Continuous mixing of populations with haplotypes at different frequencies
  - Selection
    - Natural or artificial selection
    - Can create LD between chromosomes (Bulmer effect)

Forces Modifying LE/LD (cont)

- Sporadic factors
  - Mutation – when occurring in a specific haplotype
  - Admixture/migration/crossing
  - Population bottlenecks/founder effects

Simulated LD

- Although much is known about the impact of these continuous and sporadic effects on LD, it is hard to simulate LD that behaves in an identical manner to that we observe in real life data
  - Genomic selection
  - Haplotype construction

Low Density Panels

Faculty of Agriculture and Nutritional Science

Christian-Albrechts-University of Kiel
Institute of Animal Breeding and Husbandry

Genomic Selection using Low-Density SNPs

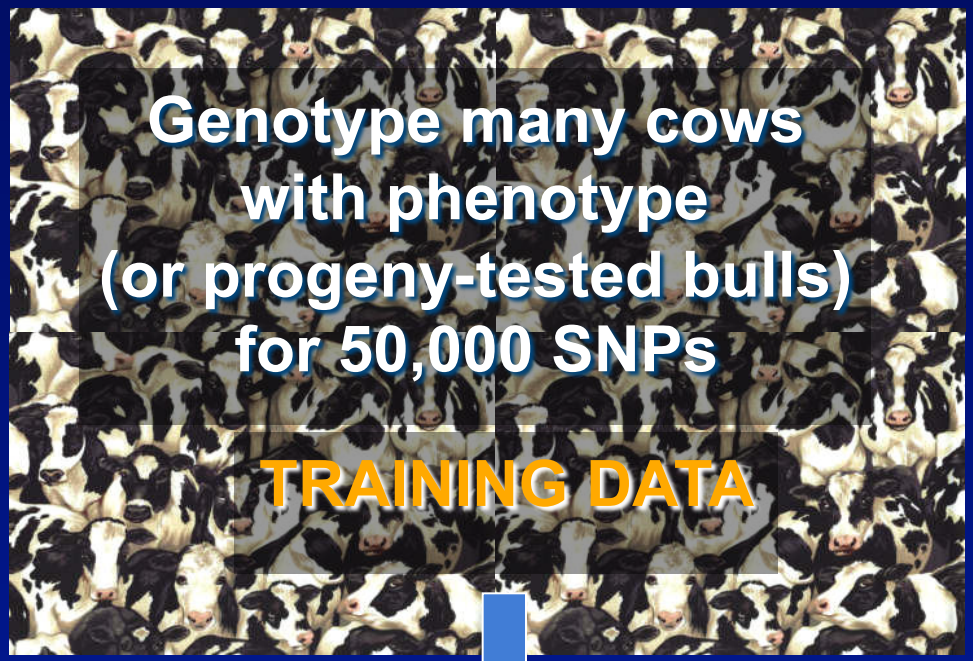
David Habier
Napapan Pyiasatian
Jack Dekkers
Rohan Fernando

***Habier et al. 2009 Genetics** 182: 343 - 353*

Genomic selection

Meuwissen et al. 2001

Genetic Evaluation using high-density SNPs



Training analysis

$$y_i = m + \sum_{\text{SNP } k} a_k g_{ik} + e_i$$

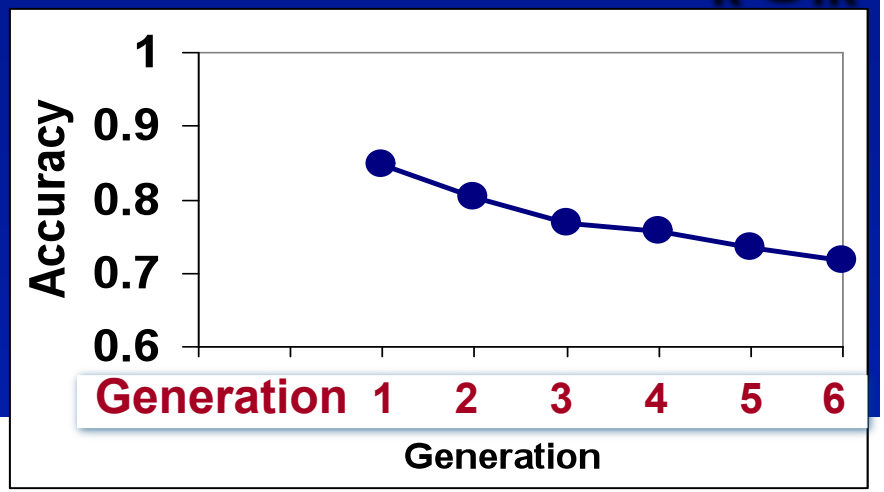
SNP effect

'0' alleles (0/1/2)

Estimates of SNP effects $\hat{a}_k$

Genotype for 50,000 SNPs

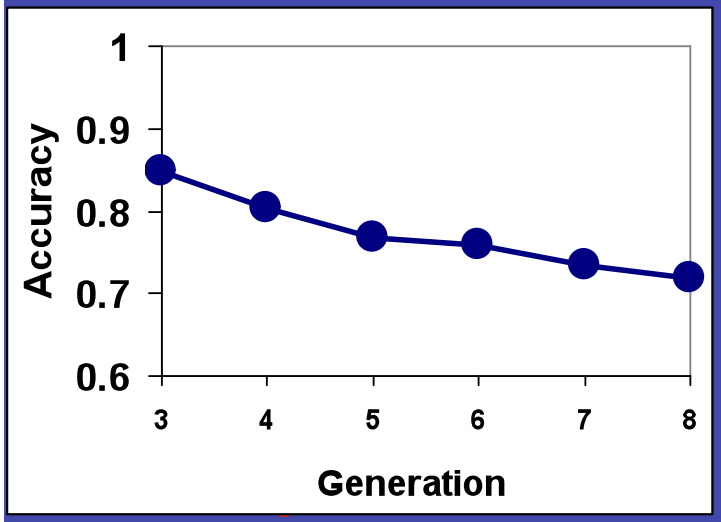
$$\text{Genomic EBV} = \sum \hat{a}_k g_{ik}$$



Genomic selection

Meuwissen et al. 2001

Genetic Evaluation using high-density SNPs



Phenotype

Genotype for >50,000 SNPs

Training data

Estimate marker effects

Genotype for >50,000 SNPs

Predict BV from marker genotypes at early age

Genotype for >50,000 SNPs

Predict BV from marker genotypes at early age

Introduction
Implementation of GS

Original principle of Genomic Selection (GS)

High-density (HD) SNP genotypes used for both

- Estimation of marker effects (training)
- Prediction of GS-EBV for selection candidates

Not feasible for many species

Need Low- (<380) vs. High-density panel for routine implementation

?? $50 vs. $250 per animal ??

'Standard' approach to developing Low-density panels:

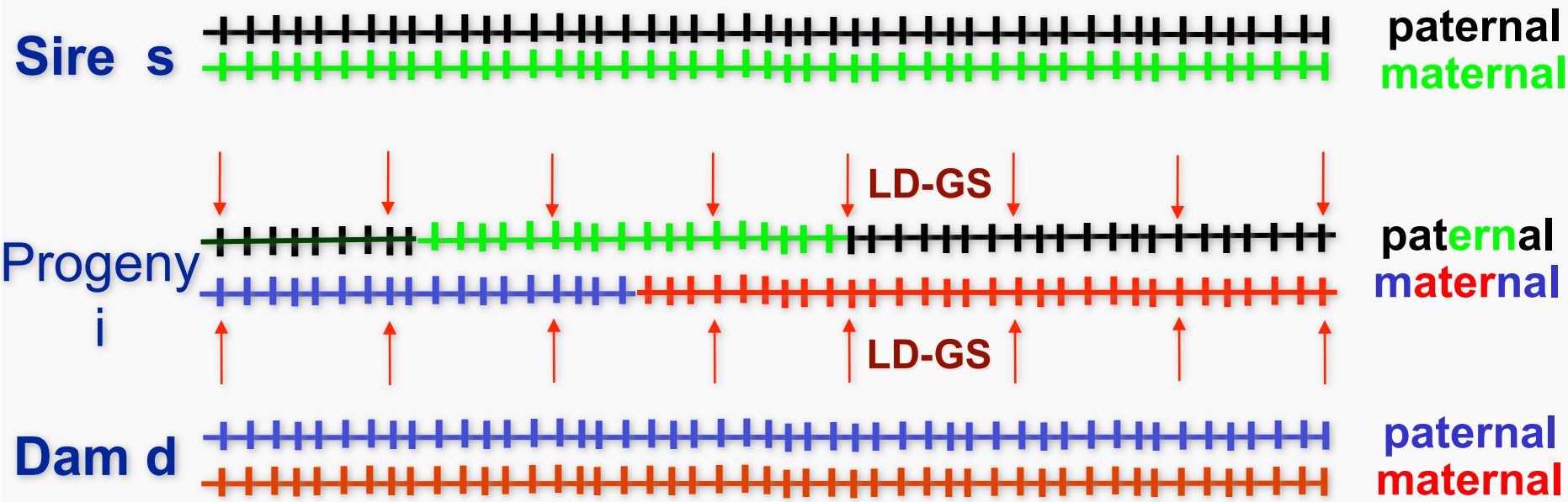
- **Select the 'best' SNPs from the HD-panel**
  - **Trait and population specific**

Proposed approach: use well-spaced Low-density SNP genotypes on selection candidates to 'fill in' missing HD SNP genotypes

Outline

- Introduction – What is ELD-GS?
- Methods
- Published results
- Unpublished results
  - Criteria for loss of accuracy
  - Factors affecting loss of accuracy of ELD-GS
    - Precision of PDMs
  - Simulations – Results
- Conclusions & outlook

Concept of Low-Density Genomic Selection

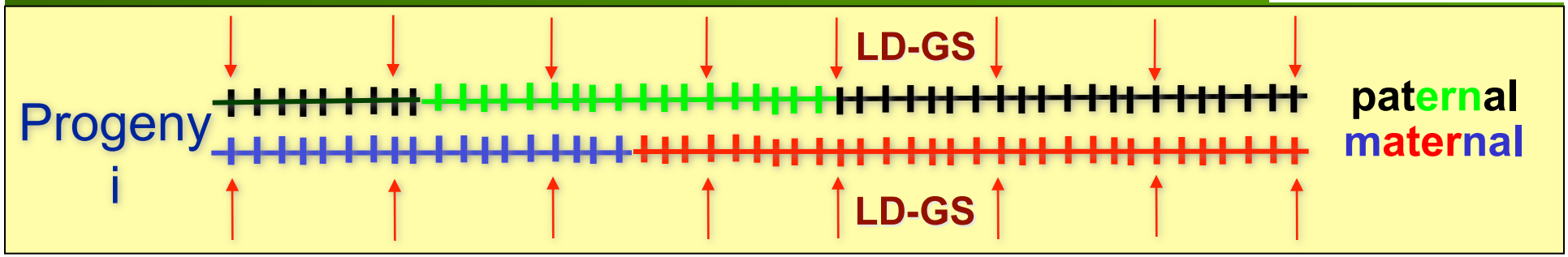


HD-GS → $EBV_i = \sum_{SNP\ k} \left(g^m_{ik} + g^p_{ik} \right)$ Sum estimates of effects of maternal and paternal SNP alleles

LD-GS → $EBV_i = \sum_{SNP\ k} \left(p^{md}_{ik,dk} g^m_{ik} + p^{pd}_{ik,dk} g^p_{ik} + p^{ms}_{ik,sk} g^m_{ik} + p^{ps}_{ik,sk} g^p \right)$

Probability that i received dam's maternal allele at SNP k

Methods



Steps of proposed low-density genomic selection method:

1. Estimate marker allele effects of HD-SNPs – Bayes-B
2. Infer HD-SNP haplotypes of training individuals
   - Requires parental HD-SNP genotypes
3. Trace HD-SNP alleles of selection candidates based on their LowD-SNP genotypes
   - Probability of descent of marker alleles
4. Predict GS-EBV of selection candidates
   - Weighted sum of effects of parental HD-SNP alleles

I. Estimation of HD-SNP effects

General statistical model:

$$\mathbf{y} = \mathbf{1}\mu + \sum_k \mathbf{x}_k \beta_k \delta_k + \mathbf{e}$$

$\mathbf{x}_k$ **= # "1" alleles carried at SNP** k

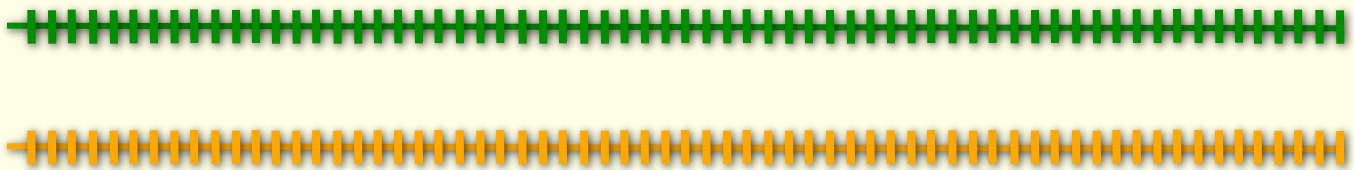
b_k **= substitution effect of SNP** k
d_k **= indicator variable for SNP** k **to be in (=1) or out (=0) of the model**

BayesB is used here, but other methods modeling disequilibrium and co-segregation, dominance or epistasis can be used also.

II. Infer HD-SNP haplotypes

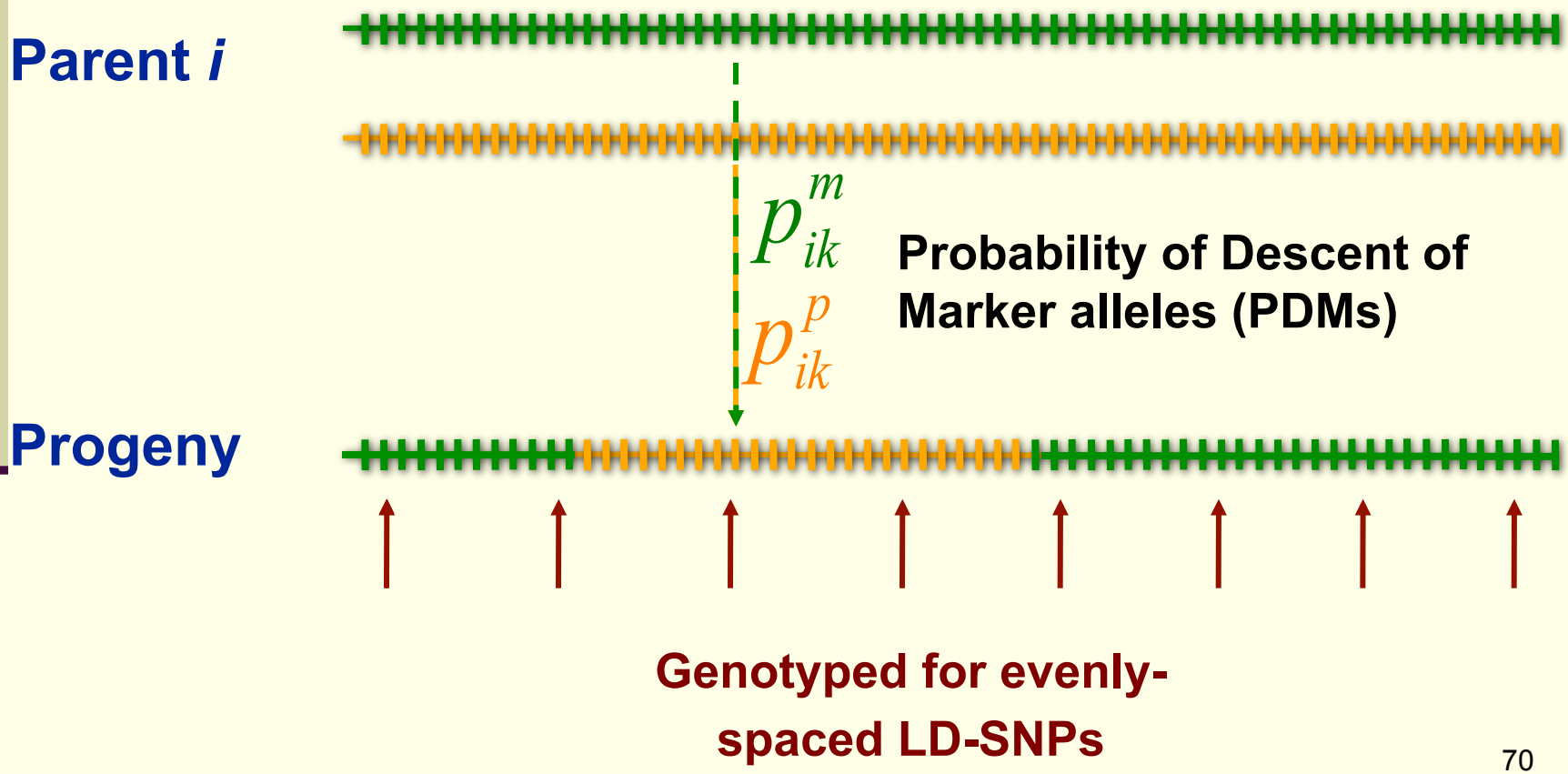
In the training generation, haplotypes must be inferred for males and females

Parent *i*



x_{ik}^{m}, x_{ik}^{p} = maternal and paternal allele states of individual *i* at SNP *k*

III. Track HD-SNP alleles



Estimation of PDMs

- MCMC sampling:
  - Joint probabilities of sampled allele origins for adjacent ELD-SNP pairs were estimated
  - Information from all ELD-SNPs is utilized
  - Haplotype phases of HD-genotyped ancestors assumed known

IV. Prediction of GEBVs

- ELD-SNP genotyped offspring:

$$GEBV_{ELD} = \sum_{k}^{loci} \left(\hat{x}_k^m + \hat{x}_k^p \right) \hat{b}_k$$

Generation after training: $\hat{x}_k^m = p_k^m * x_k^m \quad \hat{x}_k^p = p_k^p * x_k^p$

Later generations: $\hat{x}_k^m = p_k^m * \hat{x}_k^m \quad \hat{x}_k^p = p_k^p * \hat{x}_k^p$

- HD genotyped parents:

$$GEBV_{HD} = \sum_{k}^{loci} X \hat{b}_k$$

$$= \sum_{k}^{loci} \left(x_k^m + x_k^p \right) \hat{b}_k$$

Tested by Simulation

Population

Generation -1060	Random Mating (N_e=500)
Generation -60	Random Mating (N_e=100)
Generation -10	Population Growth (N=100 to N=1000)
Generation 1-3	50 males x 500 females (N=1000)
Pedigree recording and genotyping starts	
Generation 4	Training data (N=1000)
Generation 4-7	10 males x 100 females

GS-EBV using HighD SNPs

GS-EBV using LowD SNPs

Genome

10 chromosomes of 1 M

20,000 SNPs ; 500 QTL

1,000 SNPs selected after 1060 gener.

HD SNP spacing ~ 1 cM

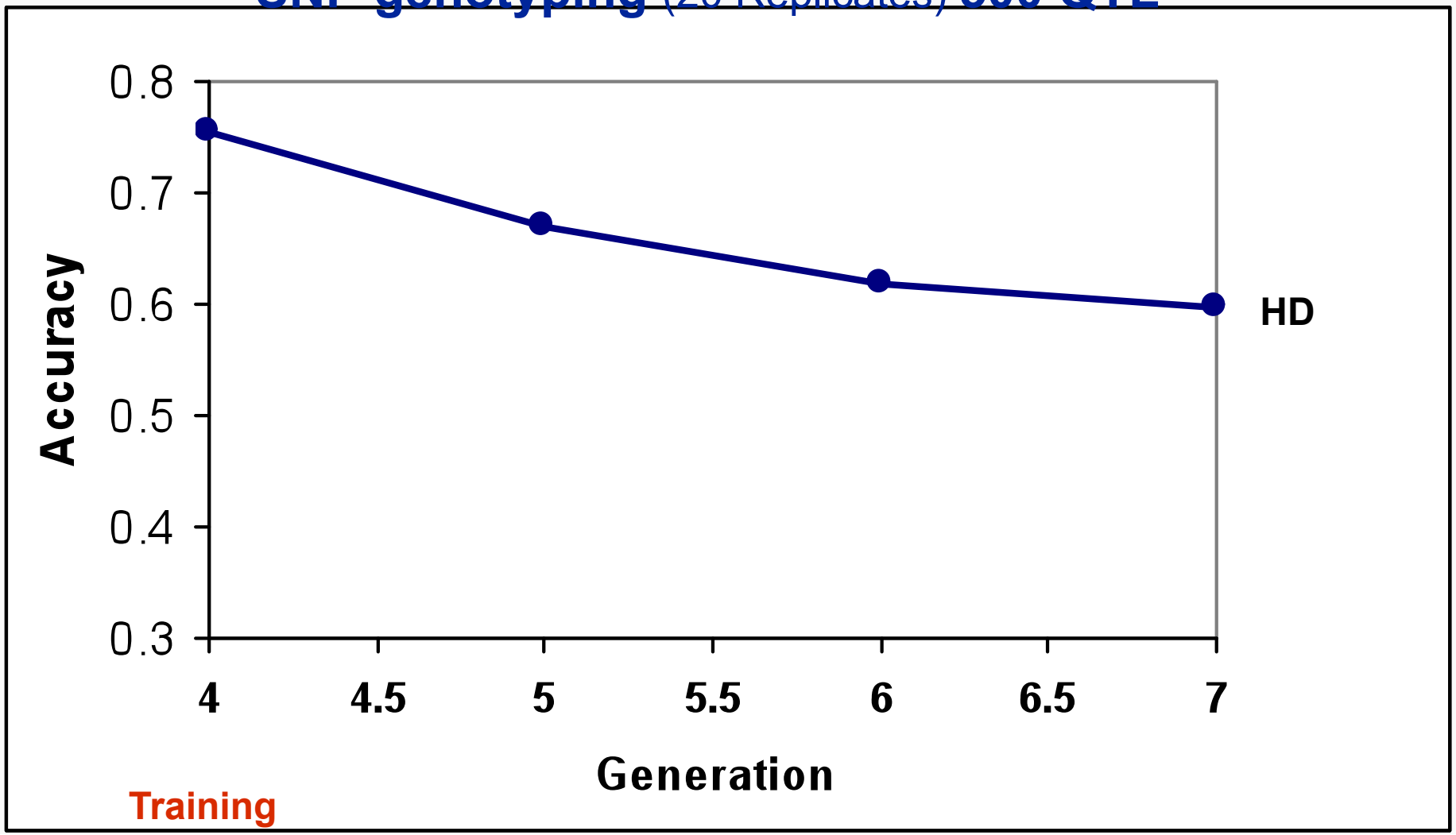
LD SNPs at 10 or 20 cM

Trait h^2 = 0.5

Bayes-B (Meuwissen et al. '01)

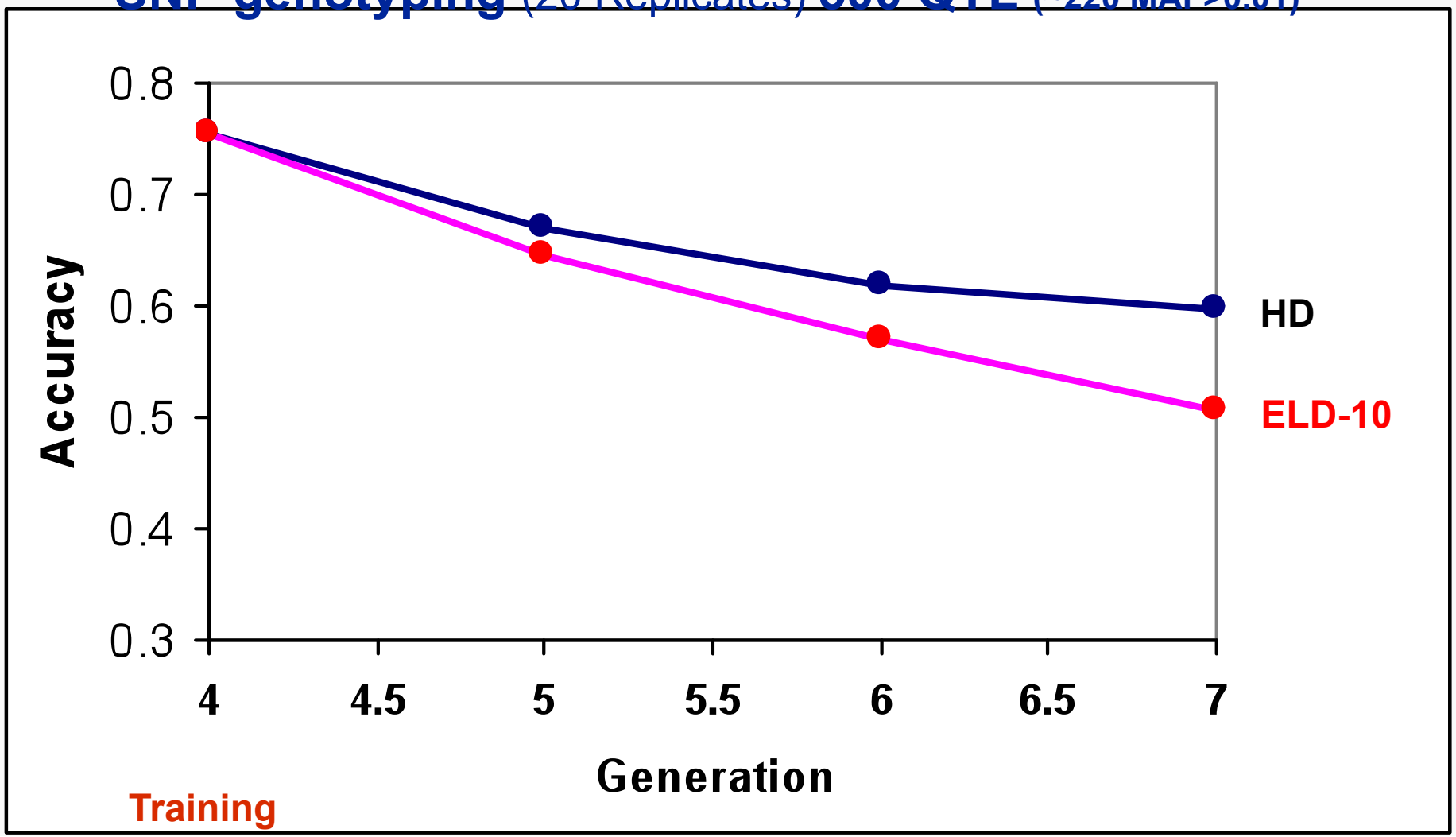
Results

Accuracy of GS-EBV based on High- and Low-Density SNP genotyping (20 Replicates) 500 QTL



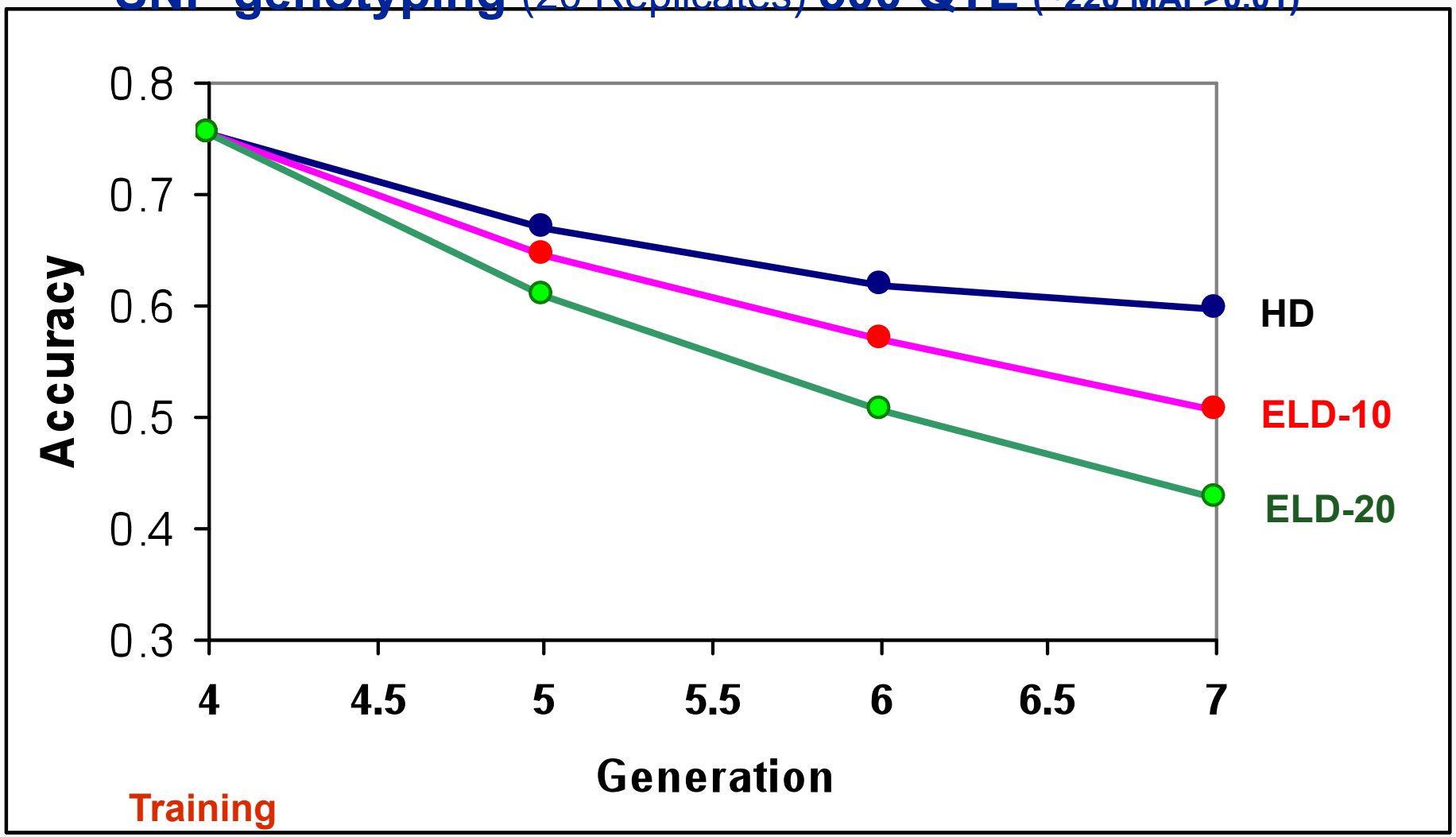
Results

Accuracy of GS-EBV based on High- and Low-Density SNP genotyping (20 Replicates) 500 QTL (~220 MAF>0.01)



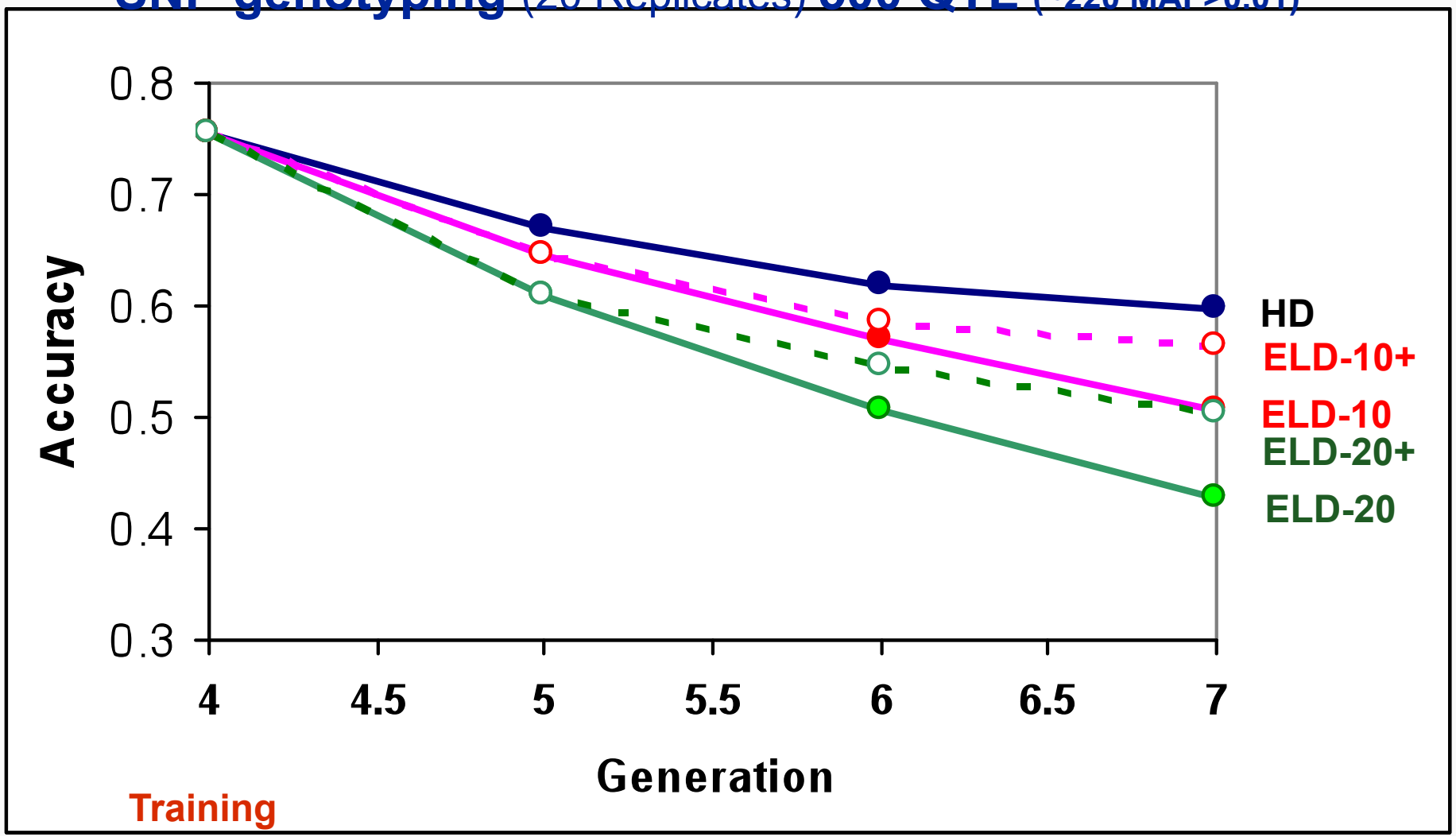
Results

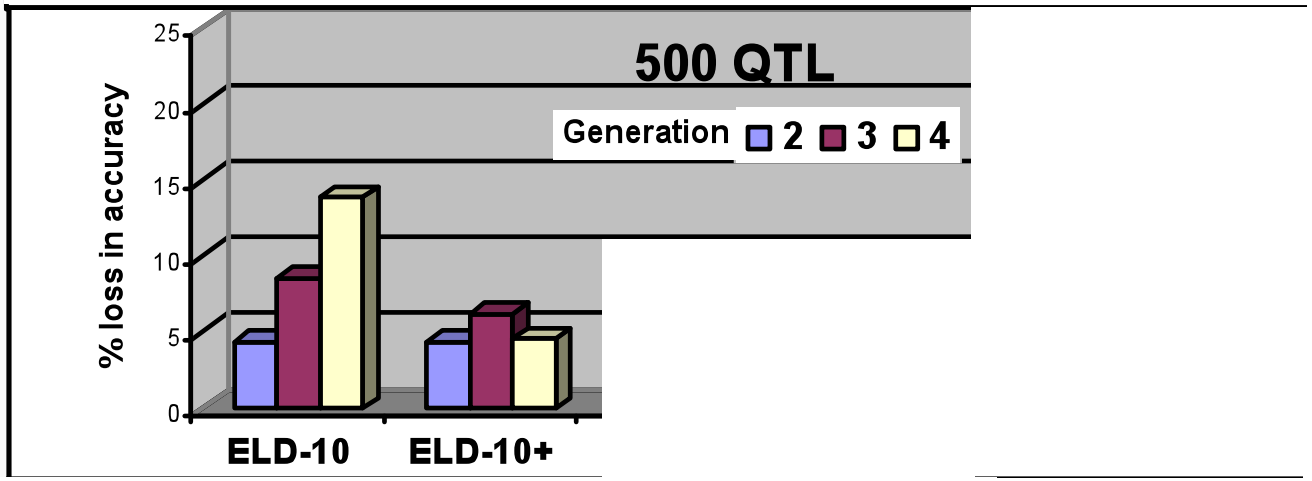
Accuracy of GS-EBV based on High- and Low-Density SNP genotyping (20 Replicates) 500 QTL (~220 MAF>0.01)



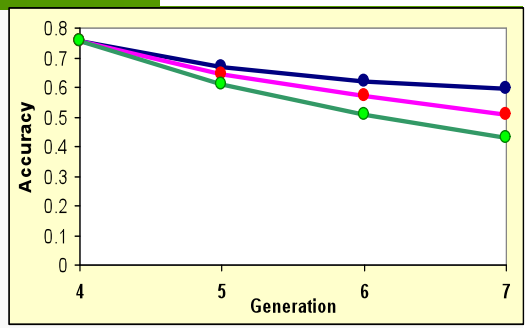
Results

Accuracy of GS-EBV based on High- and Low-Density SNP genotyping (20 Replicates) 500 QTL (~220 MAF>0.01)




500 QTL
Generation 2 3 4
% loss in accuracy
25
20
15
10
5
0
ELD-10
ELD-10+

Discussion & Conclusions



Genomic Selection can be implemented with low-density SNP genotyping of selection candidates

- Loss in accuracy limited: < 3.5% after 1 generation

  < 8 % after 2 generations

  with 300 equally spaced SNPs (10 cM)
- Loss in accuracy ~ independent of # QTL and # traits
- Lower rate of fixation of panel SNPs with selection → slower accuracy decline
- Cost effectiveness needs to be analyzed
  - Depends on costs of Low- vs. High-density genotyping

    $40 ←??→ $180
- Optimal implementation needs to be further analyzed
  - Which individuals to genotype – HD / LD

Outline

- Introduction – What is ELD-GS?
- Methods
- Published results
- Unpublished results
  - Criteria for loss of accuracy
  - Factors affecting loss of accuracy of ELD-GS
    - Precision of PDMs
  - Simulations – Results
- Conclusions & outlook

Objectives of recent work

- Analyze factors affecting loss of accuracy with ELD-GS
  - Type and extent of LD
  - Precision of PDMs
- Analyze loss of accuracy under more realistic assumptions
  - LD based on a real pedigree

Funding from Aviagen

Criteria for loss of accuracy

- Accuracy of $GEBV_{HD}$ and $GEBV_{ELD}$

$$GEBV_{HD} = \sum_{k}^{loci} X\hat{b}_k$$

$$= \sum_{k}^{loci} \left(x_k^m + x_k^p \right) \hat{b}_k$$

$$GEBV_{ELD} = \sum_{k}^{loci} \left(\hat{x}_k^m + \hat{x}_k^p \right) \hat{b}_k$$

- Uncertainty in tracking HD-SNP alleles
  - Assumption: Only precision of PDMs affects loss of accuracy → $\hat{b} = 1$
  - Correlation between $GEBV_{HD}$ and $GEBV_{ELD}$ (lower bound)

Factors affecting accuracy from ELD-GS

- Precision of PDMs
  - HD-genotyping of parents (see previous)
  - ELD-SNP spacing
  - Family structure

Simulations – Genome structure

- 8 chromosomes of 75 cM
  - 8000 HD-SNPs (Spacing 0.075 cM)
    - MAF > 0.05
  - 800 QTL
- Mutation rate 0.005 (important when historic LD simulated)
  - → # segregating QTL similar to no-LD case
- ELD-spacing: 5, 8, 10, 12, 20 cM
    - MAF > 0.40

Simulations – Population

With historic LD

Generation -1060	Random mating (N=500)
Generation - 60	Random mating (N=100)
Generation -10	Population growth until N=1000
Generation 0	50 sires + 500 dams

4 pedigree generations start

Effect of HD genotyping of parents

- 4 scenarios

Dam	Sire
✗	✗
✗	✓
✓	✗
✓	✓

Assumption for **HD-genotyped individuals:**

- HD-SNP haplotypes are known

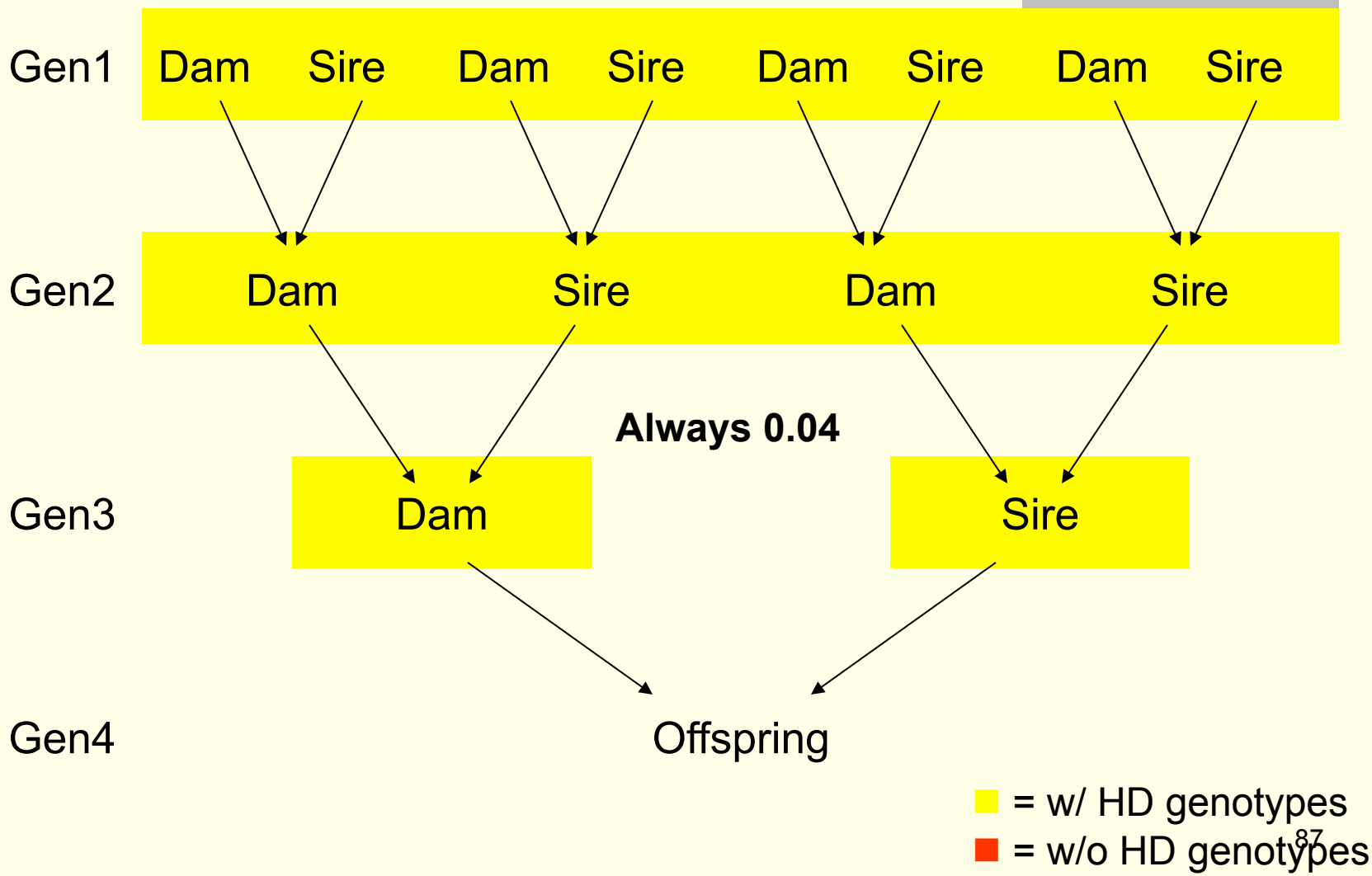
  ➔ $\hat{x}^m, \hat{x}^p$ becomes x^m, x^p

  (uncertainty from previous generations removed)

  ➔ Phases of ELD-SNPs assumed known also

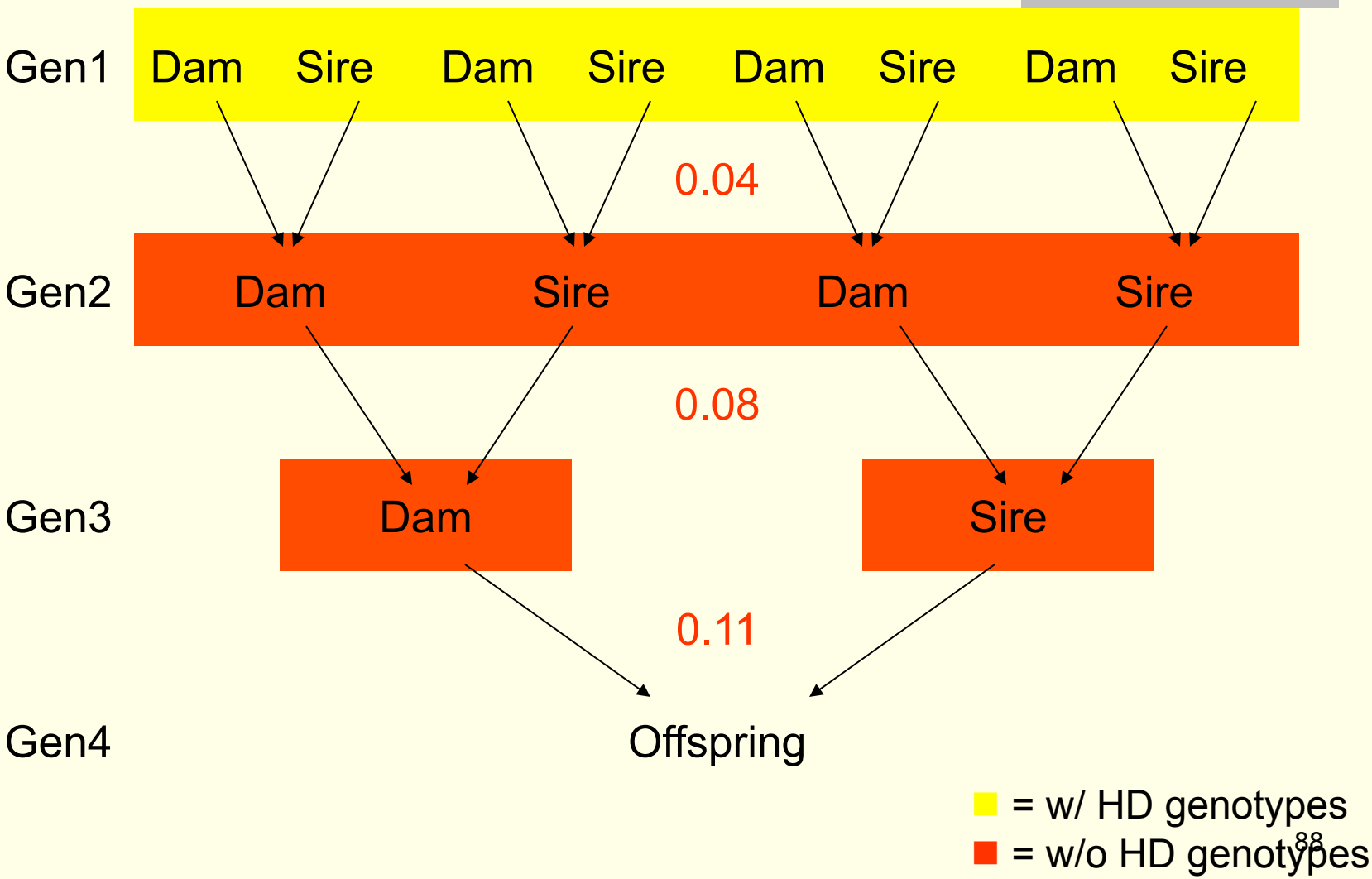
HD genotypes in dams and sires

3 SNPs, 10cM spacing, 1dam/sire, 100sires, on $\mathrm{MSE}(x_i^m, \hat{x}_i^m)$ & $\mathrm{MSE}(x_i^p, \hat{x}_i^p)$



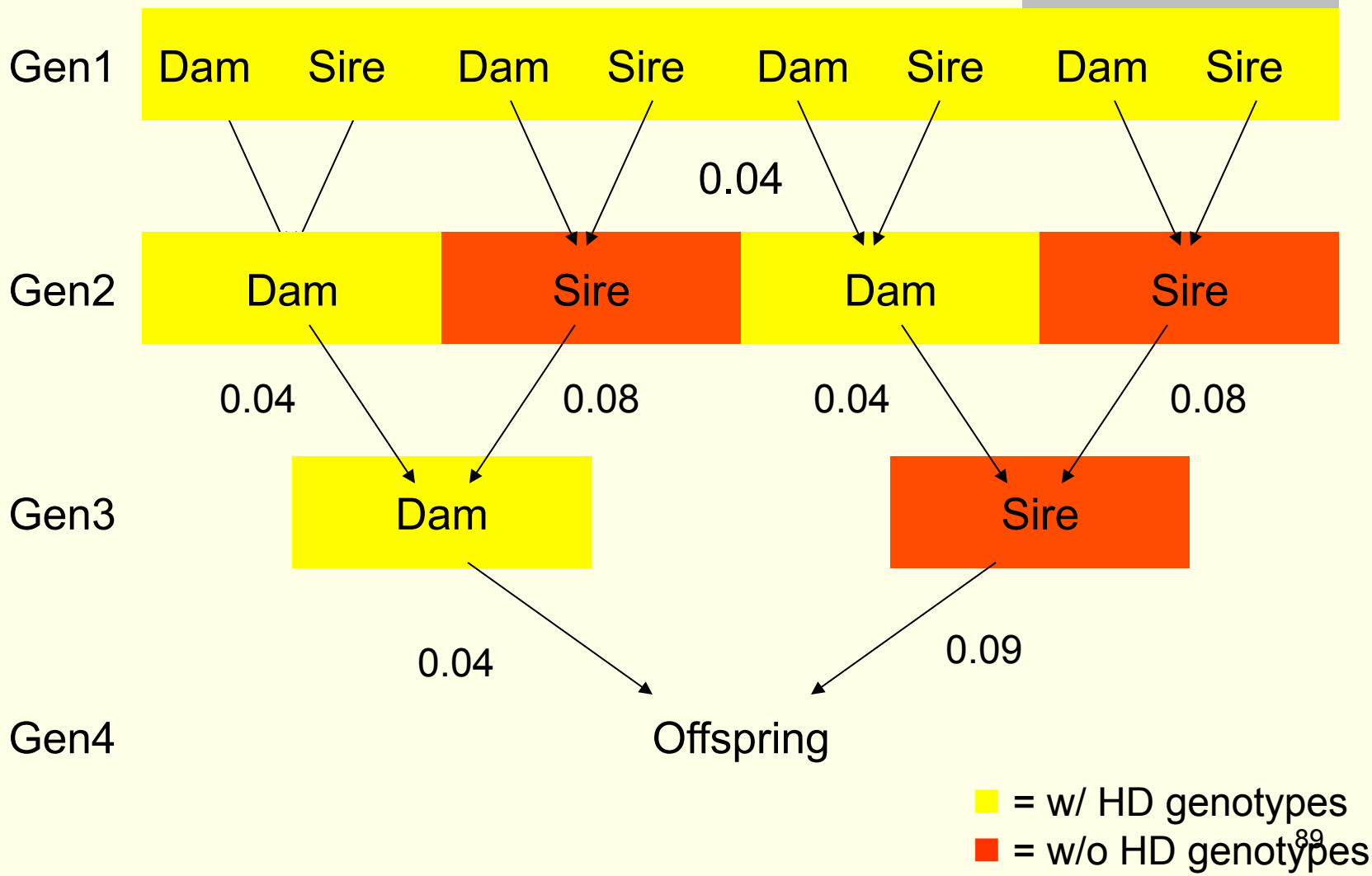
No HD genotypes on parents

3 SNPs, 10cM spacing, 1dam/sire, 100sires, on $\mathrm{MSE}(x_i^m, \hat{x}_i^m)$ & $\mathrm{MSE}(x_i^p, \hat{x}_i^p)$



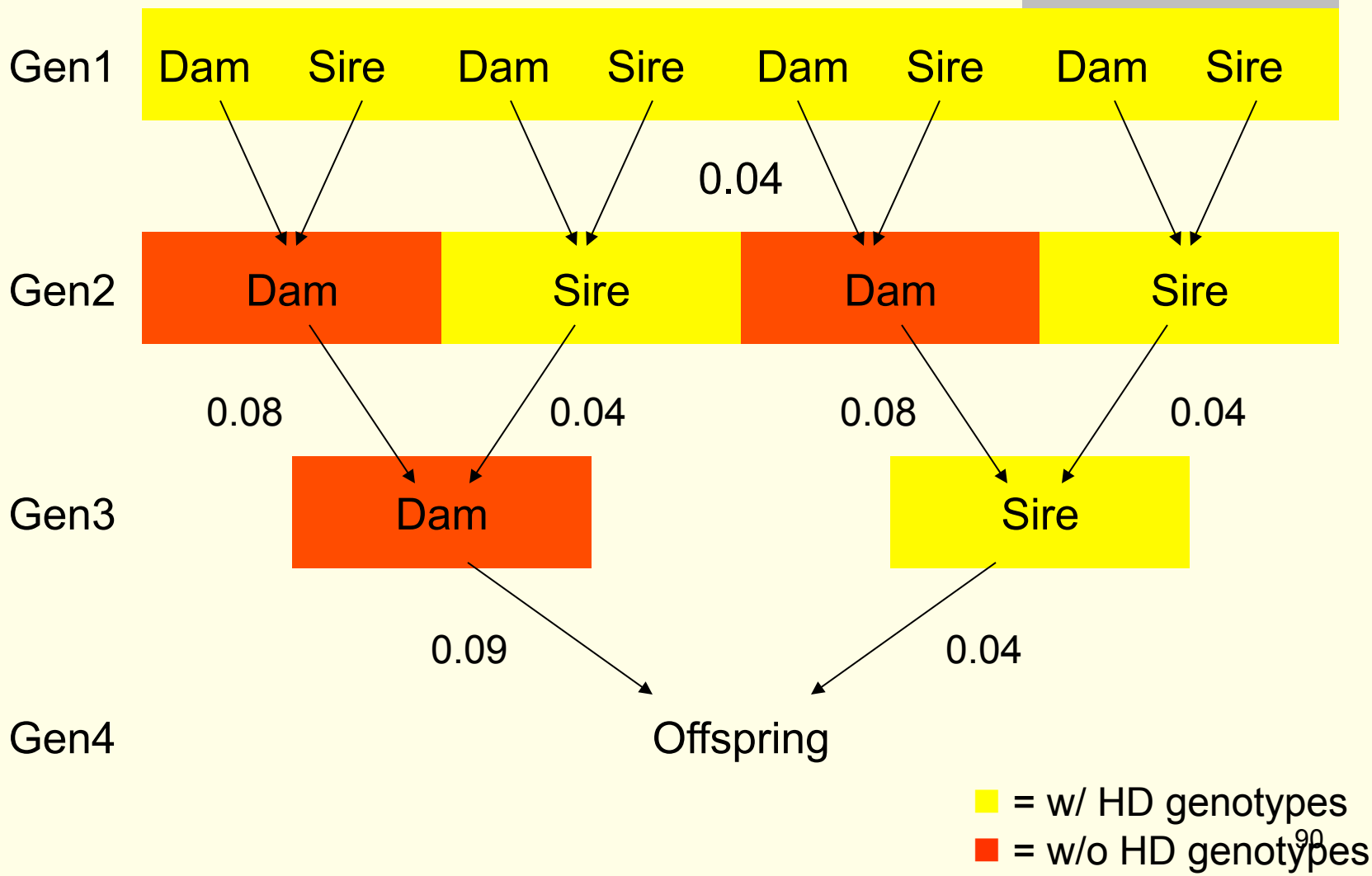
HD genotypes in dams

3 SNPs, 10cM spacing, 1dam/sire, 100sires, on $\mathrm{MSE}(x_i^m, \hat{x}_i^m)$ & $\mathrm{MSE}(x_i^p, \hat{x}_i^p)$



HD genotypes in sires

3 SNPs, 10cM spacing, 1dam/sire, 100sires, on $\mathrm{MSE}(x_i^m, \hat{x}_i^m)$ & $\mathrm{MSE}(x_i^p, \hat{x}_i^p)$



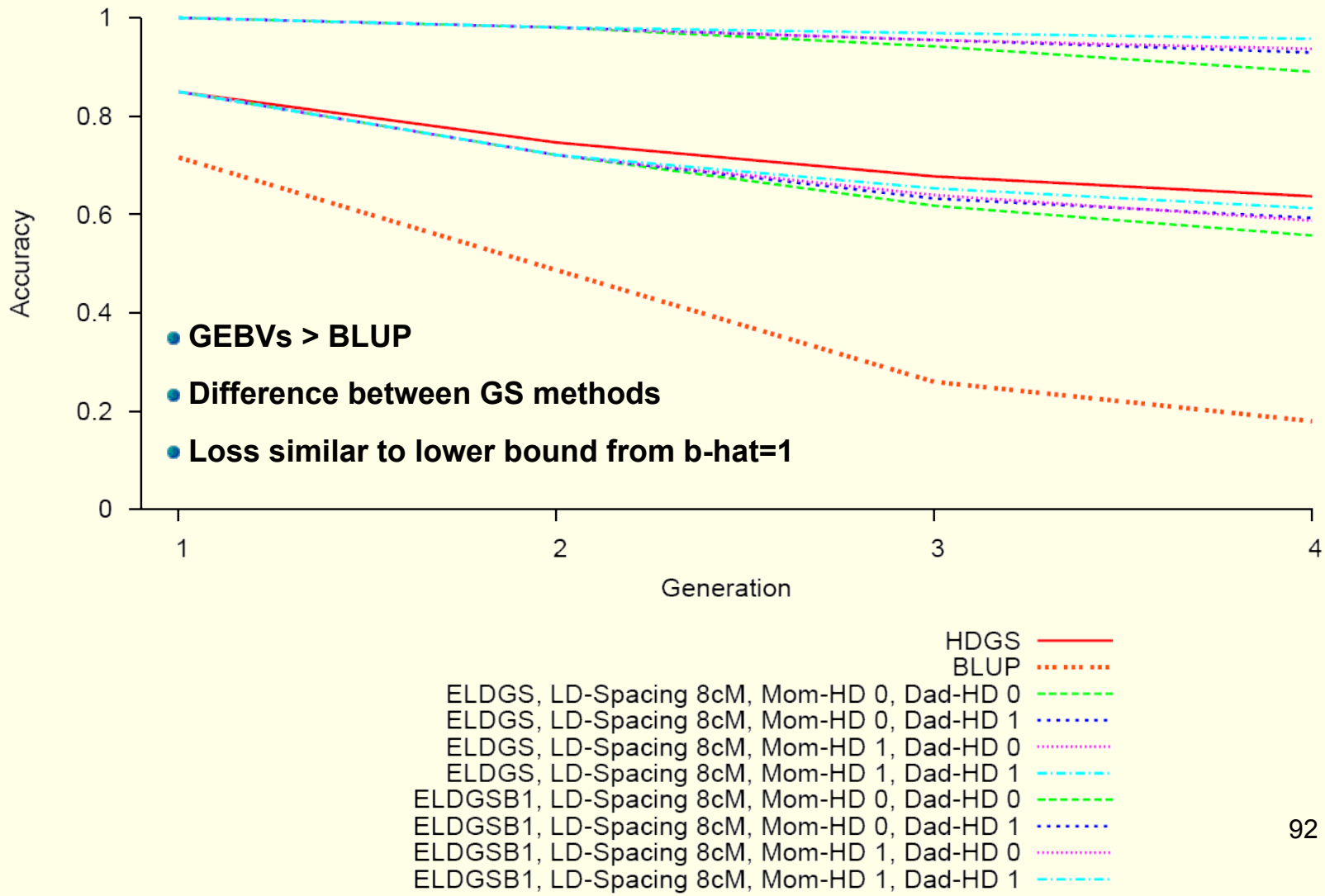
Accuracy and % - loss of accuracy

Historic LD – 8 chromosomes & 8 cM spacing (20 reps)

Method	HD-SNPs		Generation		
	Dam	Sire	2	3	4
HD-GS	-	-	74.6	67.7	63.7
			% loss from HD-GS		
BLUP	-	-	34.8	61.7	71.9
ELD-GS	✗	✗	3.4	8.8	12.5
	✗	✓	3.4	6.6	6.9
	✓	✗	3.4	5.6	7.8
	✓	✓	3.4	3.6	3.8

Accuracy of GEBVs

Historic LD – 8 chromosomes & 8cM spacing (20 reps)



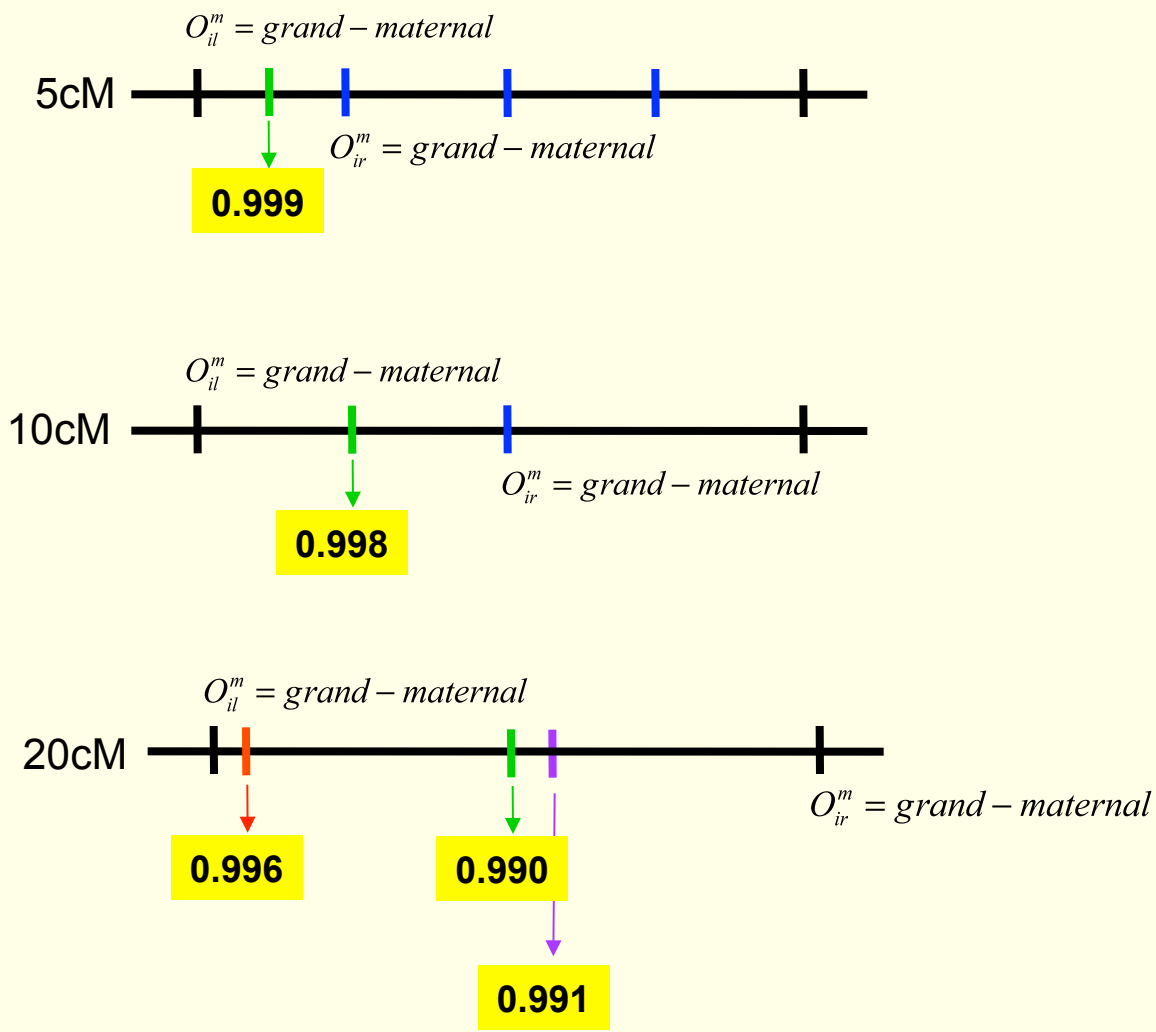
Impact of ELD-SNP spacing/density

- Effects of greater density:
  - Number of ELD-SNPs ↑
    - Adjacent SNPs help infer phases and origins
  - Recombination between adjacent ELD-SNPs ↓

No crossover:

Probability of receiving the HD grand-maternal allele

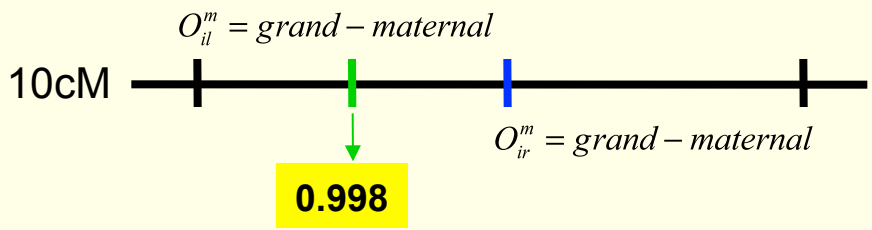
LD-SNPs



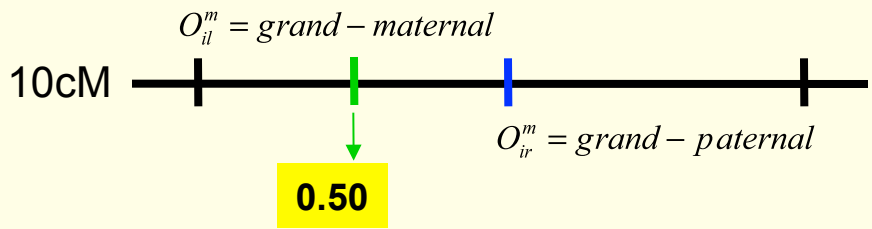
Crossovers

Probability of receiving the HD grand-maternal allele

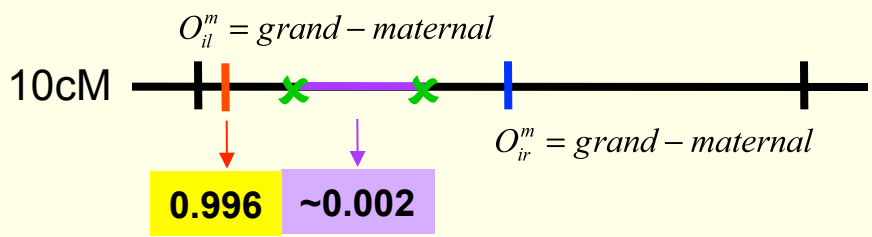
No crossover



1 crossover
= recombination
at LD-SNPs



2 crossovers
= no recombination
at LD-SNPs



ELD-SNP spacing:

% – Loss of accuracy

Both parents HD-genotyped

ELD-SNP Spacing (cM)	No. reps	Generation		
		2	3	4
5	48	1.5	2.1	2.8
8	48	2.4	3.7	3.0
10	48	4.9	4.3	4.9
12	48	4.1	4.1	6.6
20	48	8.5	8.1	9.1

Clear trend of loss of accuracy

Effect of Family structure

- Number of maternal and paternal sibs
- If a parent is HD-genotyped

  ➔ ELD-SNP phases of parent assumed known

  ➔ # parental sibs has no effect on precision of PDMs

Family structure

% – Loss of accuracy (8 cM)

No. dam/sire ↑ → No. paternal sibs ↑
same No. maternal sibs

Only females HD-genotyped

No. dams/sire	No. reps	Generation		
		2	3	4
1	48	2.4	4.4	5.0
2	48	2.8	5.6	6.6
3	43	2.5	5.6	5.5

Family structure

% – Loss of accuracy (8 cM)

No. full sibs ↑ → No. maternal and paternal sibs ↑

Parents not HD-genotyped

No. full sibs	No. reps	Generation		
		2	3	4
2	48	2.4	5.6	8.9
4	48	3.2	6.7	12.5
6	13	6.5	9.2	12.2

So far there is no trend. Again need more replicates!

Simulation with real pedigree

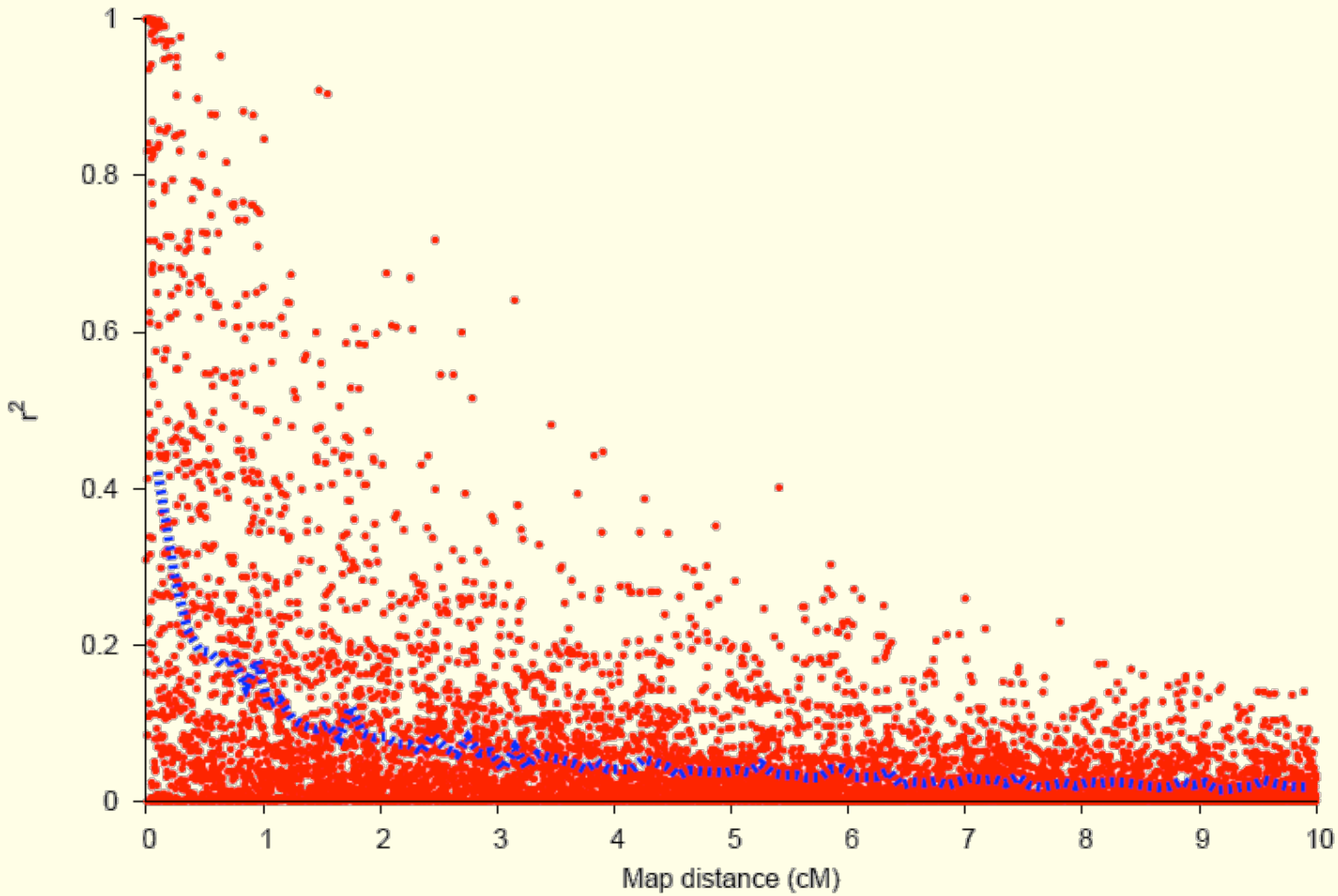
- 8 chromosomes
- 200 QTL/chromosome
- Heritability 0.5 for female phenotypes, 0.8 for male phenotypes
- No historic LD, only LD from the pedigree

Simulations – Population With Historic LD

Generation -1050	Random mating (N=500)
Generation - 50	Random mating (N=100)
Real pedigree (13 generations)	1500 males + 1500 females

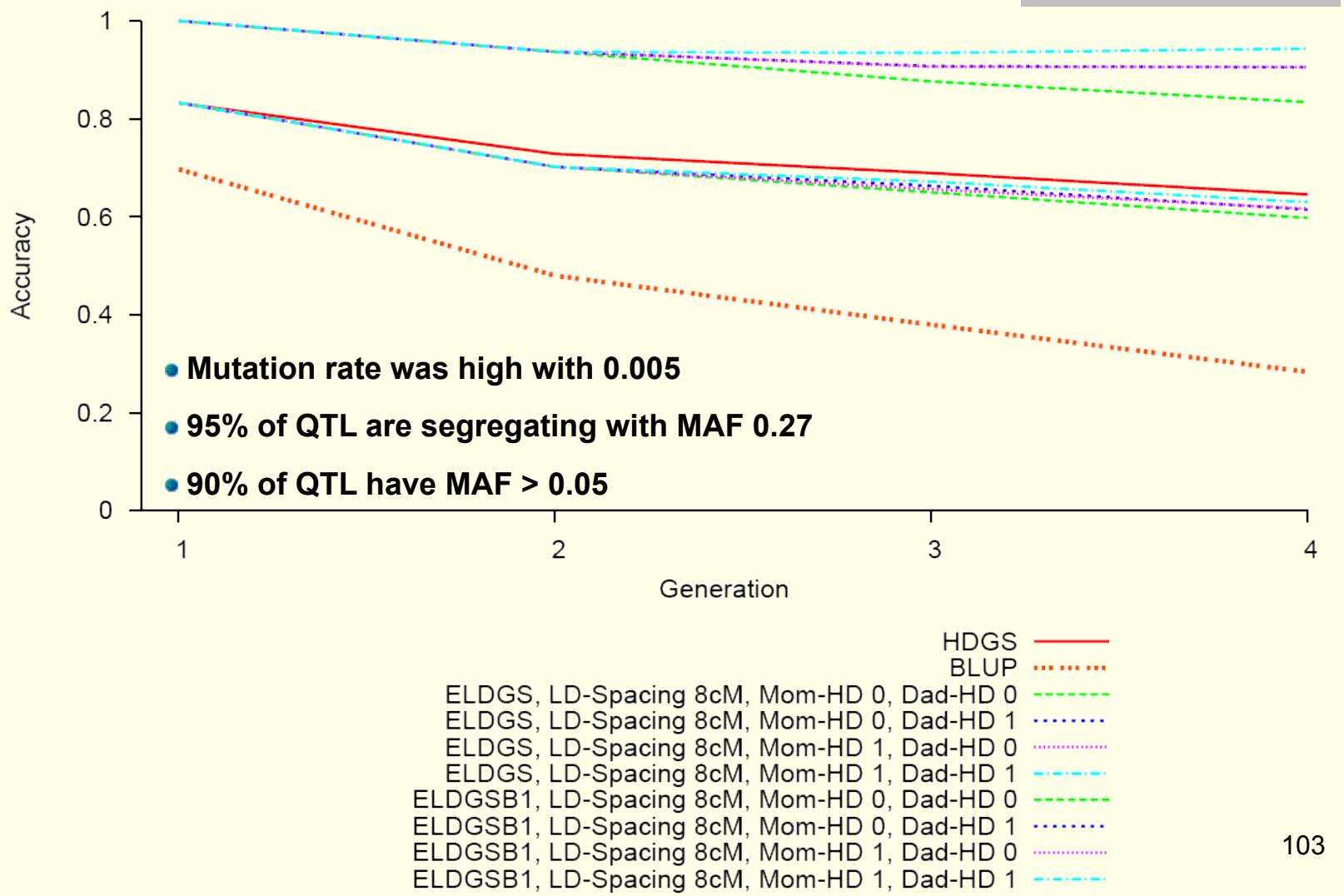
4 pedigree generations start

Linkage disequilibrium
Historic LD – Real pedigree



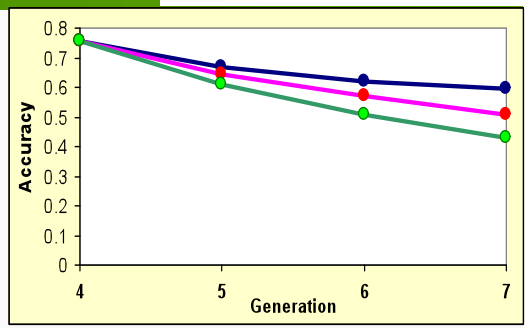
Accuracy of GEBVs

Historic LD – real pedigree (8 chromosomes & 8cM)



Discussion & Conclusions

Genomic Selection can be implemented with low-density SNP genotyping of selection candidates



- Loss in accuracy limited: < 3.5% after 1 generation

  < 8 % after 2 generations

  with 300 equally spaced SNPs (10 cM)
- Loss in accuracy ~ independent of # QTL and # traits
- Lower rate of fixation of panel SNPs with selection → slower accuracy decline
- Cost effectiveness needs to be analyzed
  - Depends on costs of Low- vs. High-density genotyping

    $40 ←??→ $180
- Optimal implementation needs to be further analyzed
  - Which individuals to genotype – HD / LD

Pooling Genomic and Pedigree Predictions

One-step assumptions

$$\mathrm{var}\begin{bmatrix} u_{pedigree} \\ u_{genotyped} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{?} \\ \mathbf{?} & \mathbf{G} \end{bmatrix} \sigma_g^2$$

What is covariance between genotyped and ungenotyped?
Is **A** an appropriate scaled var-covariance matrix given **G** on relatives?

First attempt

$$\mathrm{var}\begin{bmatrix} u_{pedigree} \\ u_{genotyped} \end{bmatrix} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{G} \end{bmatrix} \sigma_g^2$$

$$= \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G - A_{22}} \end{bmatrix} \sigma_g^2$$

Misztal et al, 2009 JDS 92:4648

Problematic

- It doesn't seem right that knowledge of genotyped animals cannot contribute to any modification of the relationships among non genotyped individuals
- For example, if parents are genotyped and shown to be more or less inbred and/or related than expected, progeny relationships should be suitably modified to reflect this information
  - This would happen, for example, if the tabular method to construct **A** was being used

Second Attempt

$$\mathrm{var}\begin{bmatrix} u_{pedigree} \\ u_{genotyped} \end{bmatrix} = \mathbf{H}$$

$$= \begin{bmatrix} \mathbf{A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21}} & \mathbf{A_{12}A_{22}^{-1}G} \\ \mathbf{GA_{22}^{-1}A_{21}} & \mathbf{G} \end{bmatrix} \sigma_g^2$$

Legarra et al, 2009 JDS 92:4656

Second Attempt

Which has a straightforward inverse

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} \sigma_g^2$$

But did not work very well in practice

Third Attempt

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \end{bmatrix} \sigma_g^2$$

Which worked better for an arbitrary (ad-hoc) λ from trial and error and is somewhat computationally attractive (for small order G)

Note that **G** can be regressed towards **A** to improve stability

Legarra et al. jds.2009-2730

Implications of Second Attempt

$$If \quad \mathrm{var}\left[u_{pedigree}\right] = \left[\mathbf{A_{11}} + \mathbf{A_{12}}\mathbf{A_{22}^{-1}}(\mathbf{G} - \mathbf{A_{22}})\mathbf{A_{22}^{-1}}\mathbf{A_{21}}\right]\sigma_g^2$$

Then we could improve the evaluation of pedigree animals by updating their var-covariance matrix according to genotyped offspring without any of their own performance information

In place of the inverse-NRM $\mathbf{A}_{11}^{-1}\sigma_g^{-2}$ we would use

$$\left[\mathbf{A_{11}} + \mathbf{A_{12}}\mathbf{A_{22}^{-1}}(\mathbf{G} - \mathbf{A_{22}})\mathbf{A_{22}^{-1}}\mathbf{A_{21}}\right]^{-1}\sigma_g^{-2}$$

How do these two alternatives compare (when $\mathbf{G} \neq \mathbf{A_{22}}$)?

Simple Example

- Suppose we have two non-inbred unrelated parents that produce two full-sib offspring
- The full A-matrix is $\begin{bmatrix} 1 & 0 & .5 & .5 \\ 0 & 1 & .5 & .5 \\ .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & 1 \end{bmatrix}$
- And the parental A-matrix that is relevant if the offspring have no records of their own is the leading 2x2 submatrix

  (an identity matrix of order 2)

Genomic matrix for offspring

- The genomic matrix might differ from the pedigree-based relationship matrix by demonstrating the
  - full-sibs have an additive relationship > 0.5
  - full-sibs have an additive relationship < 0.5
  - One or more of the fullsibs is inbred $a_{ii}<1$
- How do these modifications alter the additive variance-covariance matrix among the two parents?

Consider the exact solution

- Suppose the genotyping is for two loci, A & B, that completely determine the trait
  - Fullsib1 is A_1A_1 B_1B_2
  - Fullsib2 is A_2A_2 B_1B_2

$$\text{locus A is} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \text{ locus B is} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

$$\text{giving pooled } \mathbf{G} = \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$

  - How would this modify our assessment of the sire and dam?

Locus A in the parents

- At locus A, both parents must be heterozygous since they have offspring homozygous for the alternate forms

$$\text{locus A the parents are} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Locus B in the parents

Offspring Frequencies	Dam	B_1B_1	B_1B_2	B_2B_2
Sire	HW freq	0.25	0.5	0.25
B_1B_1	0.25	1/16	1/8	1/16
B_1B_2	0.5	1/8	1/4	1/8
B_2B_2	0.25	1/16	1/8	1/16

Locus B in the parents

Probability each parent combination produces B_1B_2

Offspring Frequencies	Dam	B_1B_1	B_1B_2	B_2B_2
Sire	HW freq	0.25	0.5	0.25
B_1B_1	0.25	1/16 (0)	1/8 (0.5)	1/16 (1)
B_1B_2	0.5	1/8 (0.5)	1/4 (0.5)	1/8 (0.5)
B_2B_2	0.25	1/16 (1)	1/8 (0.5)	1/16 (0)

Locus B in the parents

Probability each parent combination produces B_1B_2

Offspring Frequencies	Dam	B_1B_1	B_1B_2	B_2B_2
Sire	HW freq	0.25	0.5	0.25
B_1B_1	0.25	1/16 (0)	1/8 (0.5)	1/16 (1)
B_1B_2	0.5	1/8 (0.5)	1/4 (0.5)	1/8 (0.5)
B_2B_2	0.25	1/16 (1)	1/8 (0.5)	1/16 (0)

Probability each parent combination produces two full sibs that are B_1B_2

0	1/8 (0.5)^2	1/16 (1) ^2
1/8 (0.5) ^2	1/4 (0.5) ^2	1/8 (0.5) ^2
1/16 (1) ^2	1/8 (0.5) ^2	0

Locus B in the parents

	Dam	B_1B_1	B_1B_2	B_2B_2
Sire	HW freq	0.25	0.5	0.25
B_1B_1	0.25	0	1	2
B_1B_2	0.5	1	2	1
B_2B_2	0.25	2	1	0

32nds

Probability each parent combination produces two full sibs that are B_1B_2

0	1/8 (0.5)^2	1/16 (1) ^2
1/8 (0.5) ^2	1/4 (0.5) ^2	1/8 (0.5) ^2
1/16 (1) ^2	1/8 (0.5) ^2	0

Locus B in the parents

	Dam	B_1B_1	B_1B_2	B_2B_2
Sire	HW freq	0.25	0.5	0.25
B_1B_1	0.25	0	1	2
B_1B_2	0.5	1	2	1
B_2B_2	0.25	2	1	0

32nds

We need to calculate the parents genomic matrix for locus B by deriving the genomic matrix B for each of the above 9 parental combinations (or 7 cells with probabilities>0) and weight each genomic matrix by its probability (NB symmetry)

	1/32 $B_1B_1 \times B_1B_2$	2/32 $B_1B_1 \times B_2B_2$
1/32 $B_1B_2 \times B_1B_1$	2/32 $B_1B_2 \times B_1B_2$	1/32 $B_1B_2 \times B_2B_2$
2/32 $B_2B_2 \times B_1B_1$	1/32 $B_2B_2 \times B_1B_2$	

Possible B-locus Parental Genomic Matrices

1/32 $B_1B_1 \times B_1B_2$

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

2/32 $B_1B_1 \times B_2B_2$

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

1/32 $B_1B_2 \times B_1B_1$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

2/32 $B_1B_2 \times B_1B_2$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

1/32 $B_1B_2 \times B_2B_2$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

2/32 $B_2B_2 \times B_1B_1$

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

1/32 $B_2B_2 \times B_1B_2$

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

Possible B-locus Parental Genomic Matrices

$$\frac{\begin{bmatrix} 16 & 6 \\ 6 & 16 \end{bmatrix} \times \frac{1}{32}}{\frac{10}{32}}$$

$$= \begin{bmatrix} 1.6 & 0.6 \\ 0.6 & 1.6 \end{bmatrix}$$

1/32 $B_1B_1 \times B_1B_2$

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

2/32 $B_1B_1 \times B_2B_2$

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

1/32 $B_1B_2 \times B_1B_1$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

2/32 $B_1B_2 \times B_1B_2$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

1/32 $B_1B_2 \times B_2B_2$

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

2/32 $B_2B_2 \times B_1B_1$

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

1/32 $B_2B_2 \times B_1B_2$

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

Parental Genomic Matrix
Pooled across the A & B loci

Locus A Locus B Pooled

$$\frac{\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 1.6 & 0.6 \\ 0.6 & 1.6 \end{bmatrix}}{2} = \begin{bmatrix} 1.3 & 0.8 \\ 0.8 & 1.3 \end{bmatrix}$$

Summary

Pedigree A

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Legarra et al A

$$\begin{bmatrix} 1.1 & 0.1 \\ 0.1 & 1.1 \end{bmatrix}$$

Exact A

$$\begin{bmatrix} 1.3 & 0.8 \\ 0.8 & 1.3 \end{bmatrix}$$

Clearly, the Legarra et al approach is not giving the exact answer