



Übung 8 vom 27.11.2015

Linkage Disequilibrium

GWAS Single-SNP-Regression

Birgit Gredler-Grandl

Aufgabe 1: – Linkage Disequilibrium (LD)

- 224 Individuen genotypisiert für 8 SNP-Marker
- Berechnen des paarweisen LD zwischen Markern
- *install.packages("LDheatmap")*
- *install.packages("genetics")*
- *library(LDheatmap)*
- *library(genetics)*
- Daten einlesen:
- *genotypes<-read.table("SNP_LD.txt",sep=" ",header=T)*

Aufgabe 1: – Linkage Disequilibrium (LD)

- Daten einlesen:
- `genotypes<-read.table("SNP_LD.txt",sep=" ",header=T)`
- `> dim(genotypes)`
- `[1] 224 17`
- `> head(genotypes)`
- ID Snp1_1 Snp1_2 Snp2_1 Snp2_2 Snp3_1 Snp3_2
- 1 1 A A A A C C
- 2 2 A A A A C T
- 3 3 A A A A C C
- Genotypenobjekt erstellen für jeden SNP:
- `snp1<-genotype(genotypes$Snp1_1,genotypes$Snp1_2)`
- `snp2<-genotype(genotypes$Snp2_1,genotypes$Snp2_2)`
- `snp3<-genotype(genotypes$Snp3_1,genotypes$Snp3_2)`
-

Aufgabe 1: – Linkage Disequilibrium (LD)

- Genotypenobjekt erstellen für jeden SNP:
- `snp1<-genotype(genotypes$Snp1_1,genotypes$Snp1_2)`
- `snp2<-genotype(genotypes$Snp2_1,genotypes$Snp2_2)`
- `snp3<-genotype(genotypes$Snp3_1,genotypes$Snp3_2)`
-
- `>.snp1`
- `[1] "A/A" "A/A" "A/A" "A/A" "A/G" "A/A" "A/A" "A/A"`

- LD berechnen für SNP-Paare:

- `> LD(snp1,snp2)`

- *Pairwise LD*

	D	D'	Corr
<i>Estimates:</i>	0.02349079	0.9967644	0.7729406
	X^2	P-value	N
<i>LD Test:</i>	267.6518	0	224

D' ist ein weiteres Mass für LD (Lewontin, 1964)

Aufgabe 1: – Linkage Disequilibrium (LD)

- LD berechnen für SNP-Paare:
- > `LD(snp1,snp2)$"D"`
- [1] 0.02349079
- > `LD(snp1,snp2)$"R^2"`
- [1] 0.5974371

Aufgabe 1: – Linkage Disequilibrium (LD)

- Genotypenobjekte aller SNP in ein data.frame zusammenführen:
- `genoall<-(data.frame(snp1,snp2,snp3,snp4,snp5,snp6,snp7,snp8))`
- `> head(genoall)`
- `snp1.snp2.snp3.snp4.snp5.snp6.snp7.snp8`

1	A/A	A/A	C/C	T/T	A/A	C/C	C/C	T/T
2	A/A	A/A	C/T	T/T	A/T	C/C	C/C	T/T
3	A/A	A/A	C/C	T/T	A/A	C/C	C/C	T/T
4	A/A	A/A	C/T	T/T	A/T	C/C	C/C	T/T
- Paarweises LD auf gesamtes data.frame genoall berechnen:
- `> LD(genoall)`

Aufgabe 1: – Linkage Disequilibrium (LD)

- Paarweises LD auf gesamtes data.frame genoall berechnen:
- > *LD(genoall)*

Pairwise LD

	snp2	snp3	snp4	snp5	snp6	snp7	snp8
snp1 D	0.0234908	0.0258416	0.0233145	-0.0089199	0.0234908	0.0234908	0.0234232
snp1 D'	0.9967644	0.9970246	0.8370854	0.9945871	0.9967644	0.9967644	0.9110699
snp1 Corr.	0.7729406	0.2750162	0.7072855	-0.1090825	0.7729406	0.7729406	0.7387492
snp1 X^2	267.6518379	33.8839909	224.1132587	5.3307516	267.6518379	267.6518379	244.4961915
snp1 P-value	< 2.2204e-16	5.849808e-09	< 2.2204e-16	0.02095235	< 2.2204e-16	< 2.2204e-16	< 2.2204e-16
snp1 n	224	224	224	224	224	224	224
<hr/>							
snp2 D	0.0157630	0.0237648	0.0002749	0.0238744	0.0238744	0.0238196	
snp2 D'	0.9951858	0.9968016	0.0144133	0.9968162	0.9968162	0.9968089	
snp2 Corr.	0.2128679	0.9148235	0.0042659	0.9968162	0.9968162	0.9532793	
snp2 X^2	20.3001048	374.9320721	0.0081526	445.1518622	445.1518622	407.1161426	
snp2 P-value	6.619696e-06	< 2.2204e-16	0.9280555	< 2.2204e-16	< 2.2204e-16	< 2.2204e-16	
snp2 n	224	224	224	224	224	224	224

Aufgabe 1: – Linkage Disequilibrium (LD)

- Paarweises LD auf gesamtes data.frame genoall berechnen:
- > `LD(genoall)$"D"`

	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
snp1	NA	0.02349079	0.02584165	0.02331447	-0.0089198857	0.0234907899	0.0234907899	0.023423153
snp2	NA		NA	0.01576299	0.02376483	0.0002749026	0.0238744395	0.0238744395
snp3	NA		NA		0.01866969	0.1439327046	0.0157629918	0.0157629918
snp4	NA		NA		NA	0.0039150618	0.0237648253	0.0237648253
snp5	NA		NA		NA		0.0002749026	0.0002749026
snp6	NA		NA		NA		NA	0.0238744395
snp7	NA		NA		NA		NA	0.0238744395
snp8	NA		NA		NA		NA	NA

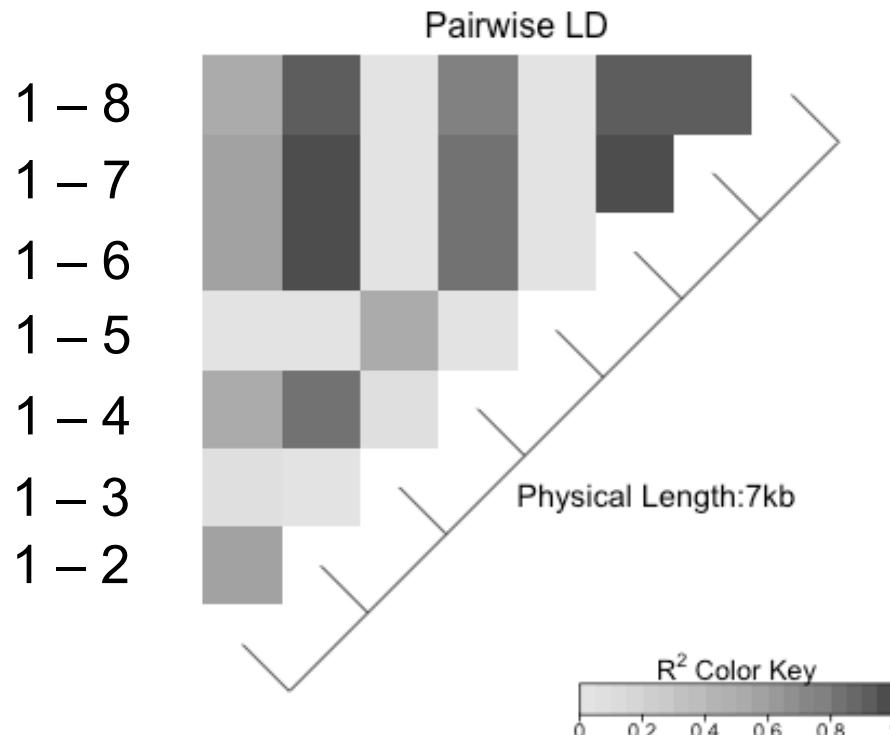
Aufgabe 1: – Linkage Disequilibrium (LD)

- Paarweises LD auf gesamtes data.frame genoall berechnen:
- > `LD(genoall)$"R^2"`

	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
snp1	NA	0.5974371	0.07563391	0.50025281	0.0118989991	0.5974371382	0.5974371382	0.5457504274
snp2	NA		NA	0.04531273	0.83690195	0.0000181977	0.9936425495	0.9936425495
snp3	NA		NA		0.05403295	0.5218633734	0.0453127339	0.0453127339
snp4	NA		NA		NA	0.0031374545	0.8369019468	0.8369019468
snp5	NA		NA		NA		0.0000181977	0.0000181977
snp6	NA		NA		NA		NA	0.9936425495
snp7	NA		NA		NA		NA	0.9087413896
snp8	NA		NA		NA		NA	NA

Aufgabe 1: – Linkage Disequilibrium (LD)

- LD plot:
- [LDheatmap\(genoall\)](#)



	snp1	snp2	snp3	snp4	snp5	snp6	snp7	snp8
snp1	NA	0.5974371	0.07563391	0.50025281	0.0118989991	0.5974371382	0.5974371382	0.5457504274
snp2	NA		NA	0.04531273	0.83690195	0.0000181977	0.9936425495	0.9087413896
snp3	NA		NA		0.05403295	0.5218633734	0.0453127339	0.0453127339
snp4	NA		NA		NA	0.0031374545	0.8369019468	0.8369019468
snp5	NA		NA		NA		0.0000181977	0.0000181977
snp6	NA		NA		NA		NA	0.9936425495
snp7	NA		NA		NA		NA	0.9087413896
snp8	NA		NA		NA		NA	NA

Aufgabe 2: – Single-SNP-Regression

- 325 Stiere wurden für 10 SNP-Marker genotypisiert
- Für alle Stiere liegen Phänotypen für Eiweiss-% vor
- Gibt es eine Assoziation zwischen Phänotyp und SNP?
- Testen des Einflusses jedes SNP mittels Single-SNP-Regression

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

Aufgabe 2: Single-SNP-Regression

- Daten in R einlesen
- `setwd("/Pfad-zu-den-Daten/")`
- `phenotypes <- read.table("yvec.inp", header=F)`
- `genotypes <- read.table("xmatrix.inp", header=F)`
- `map <- read.table("map_markers.txt", header=F)`

```
> head(phenotypes)  > head(genotypes)          > head(map)
  V1                  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10      V1   V2
1 0.186              1 0 0 0 0 0 1 2 0 2             1 m1 47434
2 1.226              2 1 0 0 1 1 1 2 1 0 1             2 m2 123421
3 0.864              3 1 0 0 1 0 0 1 1 1 1             3 m3 237726
4 1.228              4 1 1 1 1 0 1 2 1 1 1             4 m4 367975
5 0.446              5 0 1 1 1 1 1 2 1 0 1             5 m5 496978
6 1.125              6 1 0 0 1 1 1 2 1 0 1             6 m6 558613

> dim(phenotypes)    > dim(genotypes)          > dim(map)
[1] 325  1            [1] 325  10           [1] 10  2
```

Aufgabe 2: Single-SNP-Regression

- **Single-SNP-Regression**
- *lm(phenotypes[, 1]~genotypes[, 1])*

Call:

`lm(formula = phenotypes[, 1] ~ genotypes[, 1])`

Coefficients:

(Intercept) genotypes[, 1]

0.1013 0.3826

- Summary: *summary(lm(phenotypes[, 1]~genotypes[, 1]))*

Aufgabe 2: Single-SNP-Regression

- Summary: `summary(lm(phenotypes[, 1]~genotypes[, 1]))`

Call:

```
lm(formula = phenotypes[, 1] ~ genotypes[, 1])
```

Residuals:

Min	1Q	Median	3Q	Max
-1.04889	-0.31149	-0.00549	0.30911	1.41251

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.10128	0.03897	2.599	0.00979 **
genotypes[, 1]	0.38261	0.03338	11.462	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4509 on 323 degrees of freedom

Multiple R-squared: 0.2891, Adjusted R-squared: 0.2869

F-statistic: 131.4 on 1 and 323 DF, p-value: < 2.2e-16

Aufgabe 2: Single-SNP-Regression

- **Nur den Effekt des SNP ausgeben lassen:**
- `lm(phenotypes[, 1]~genotypes[, 1])$coeff[2]`
genotypes[, 1]
0.382608
- **Ausgabe des genauen P-Wertes:**
- `anova(lm(phenotypes[, 1]~genotypes[, 1]))$P[1]`
[1] 9.492982e-26

Aufgabe 2: Single-SNP-Regression

- **Testen aller SNP im Loop und speichern der Ergebnisse in einem data.frame**
- Dimension des data.frame definieren:
- `results <- data.frame(rep(NA,10),rep(NA,10))`
- `names(results) <- c(„Effekt“, "p-Wert")`
- `for (i in 1:10){`
 `results[i,1] <- lm(phenotypes[,1]~genotypes[,i])$coeff[2]`
 `results[i,2] <- anova(lm(phenotypes[,1]~genotypes[,i]))$P[1]`
 `}`

Aufgabe 2: Single-SNP-Regression

> *results*

	Effekt	p-Wert	
1	0.3826080	9.492982e-26	
2	0.4535504	6.820006e-13	
3	0.4535504	6.820006e-13	Welcher SNP weist den grössten Effekt auf?
4	0.4972216	7.047841e-43	
5	0.4465850	8.634177e-30	
6	0.4926920	1.615060e-37	Welcher SNP hat einen signifikanten Einfluss?
7	0.2221599	1.632815e-04	
8	-0.4972216	7.047841e-43	
9	0.3527940	3.354479e-13	
10	-0.4972216	7.047841e-43	

Aufgabe 2: Single-SNP-Regression

> *results*

	Effekt	p-Wert
1	0.3826080	9.492982e-26
2	0.4535504	6.820006e-13
3	0.4535504	6.820006e-13
4	0.4972216	7.047841e-43
5	0.4465850	8.634177e-30
6	0.4926920	1.615060e-37
7	0.2221599	1.632815e-04
8	-0.4972216	7.047841e-43
9	0.3527940	3.354479e-13
10	-0.4972216	7.047841e-43

Aufgabe 2: Single-SNP-Regression

> *results*

	Effekt	p-Wert
1	0.3826080	9.492982e-26
2	0.4535504	6.820006e-13
3	0.4535504	6.820006e-13
4	0.4972216	7.047841e-43
5	0.4465850	8.634177e-30
6	0.4926920	1.615060e-37
7	0.2221599	1.632815e-04
8	-0.4972216	7.047841e-43
9	0.3527940	3.354479e-13
10	-0.4972216	7.047841e-43

Aufgabe 2: Single-SNP-Regression

> *results*

	Effekt	p-Wert
1	0.3826080	9.492982e-26
2	0.4535504	6.820006e-13
3	0.4535504	6.820006e-13
4	0.4972216	7.047841e-43
5	0.4465850	8.634177e-30
6	0.4926920	1.615060e-37
7	0.2221599	1.632815e-04
8	-0.4972216	7.047841e-43
9	0.3527940	3.354479e-13
10	-0.4972216	7.047841e-43

Warum haben manche SNP
den gleichen Effekt?

Aufgabe 2: Single-SNP-Regression

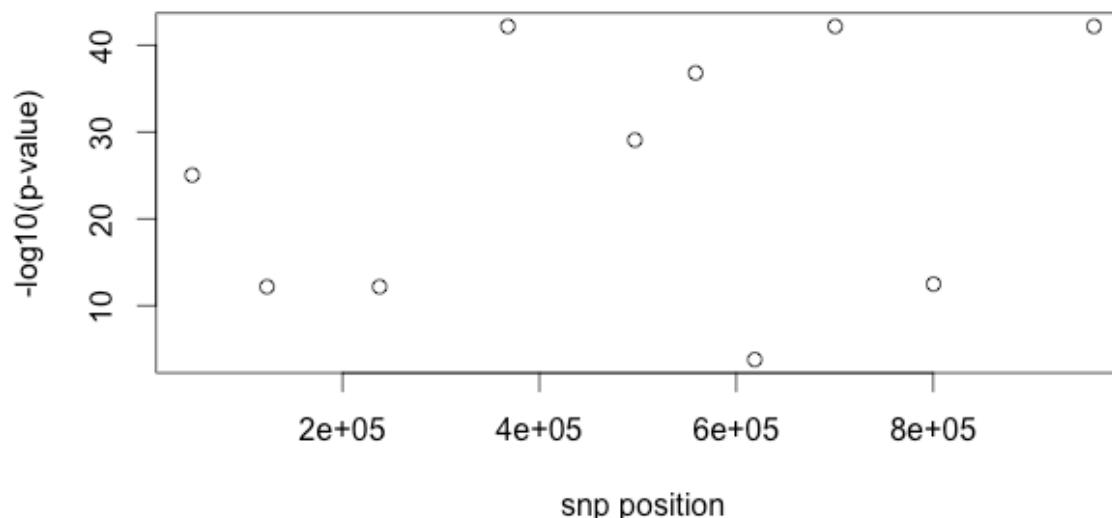
- Warum haben manche SNP den gleichen Effekt?
- Zusatzaufgabe: Berechnen Sie das LD (r^2) zwischen allen 10 SNP
- Genotypen sind jetzt 0,1,2 kodiert
- Für die LD-Berechnung mit der Funktion LD müssen diese in Character-Allelkodierungen (z.B. CC, CT, TT) vorliegen
- Mit der Funktion „as.genotype.allele.count“ können die Allele umkodiert werden:
 - `genoneu<-data.frame(rep(NA,325),rep(NA,325),rep(NA,325),rep(NA,325),rep(NA,325),rep(NA,325),rep(NA,325),rep(NA,325),rep(NA,325))`
 - `names(test)<-c("snp1","snp2","snp3","snp4","snp5","snp6","snp7","snp8","snp9","snp10")`
 - `for (i in 1:10){`
 - `genoneu[,i]<-as.genotype.allele.count(genotypes[,i], alleles=c("C","T"))`
 - `}`

Aufgabe 2: Single-SNP-Regression

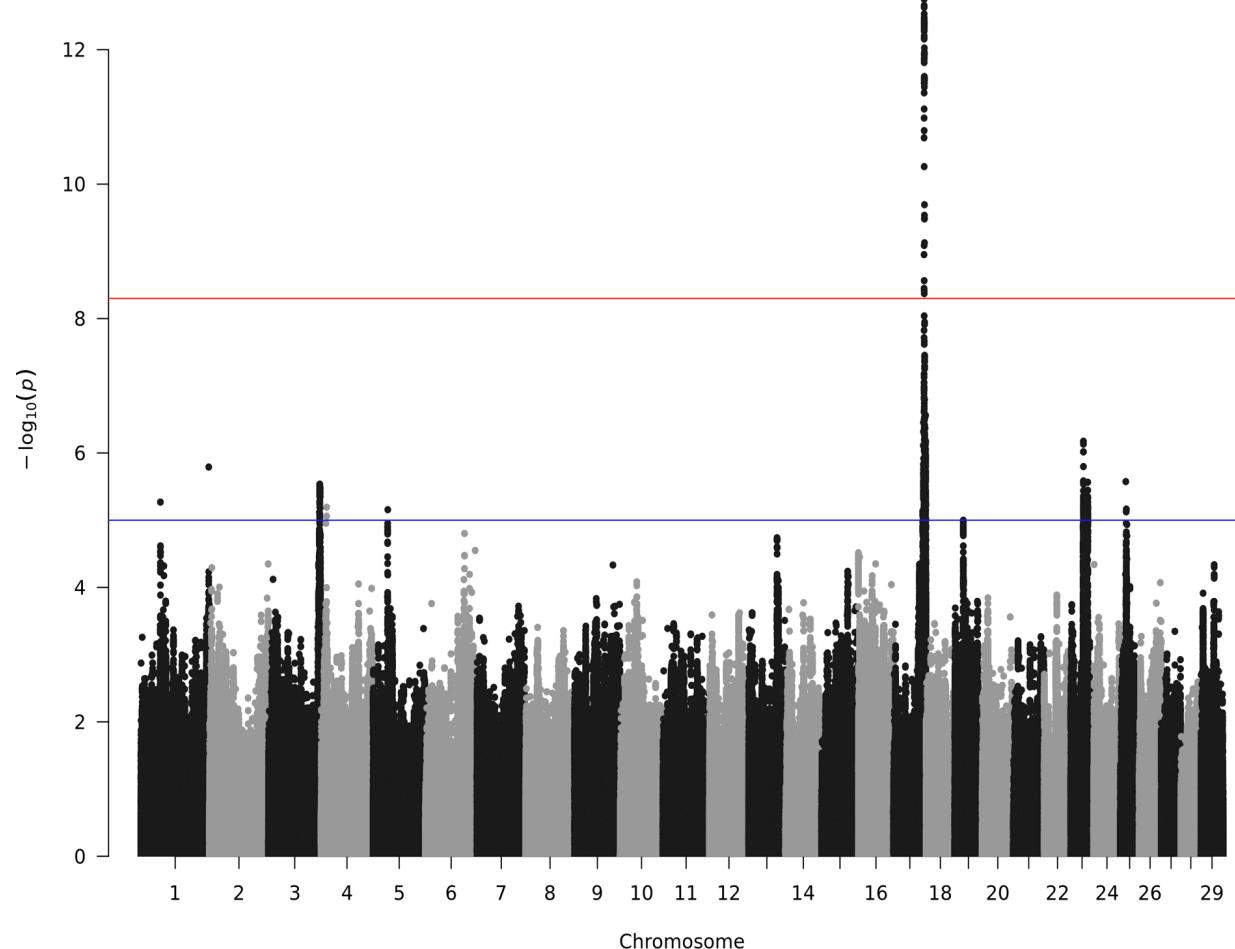
- LD (r^2) für gesamtes data.frame berechnen und mittels Ldheatmap visualisieren

Aufgabe 2: Single-SNP-Regression

- **Visualisierung der Ergebnisse (Wo sind die signifikanten SNP?)**
- P-Werte sehr klein → logarithmieren
- Plotten der Position gegen die P-Werte
- `plot(x, y, ...)`
- `plot(map[,2], -log10(results[,2]), xlab="snp position", ylab="-log10(p-value)")`

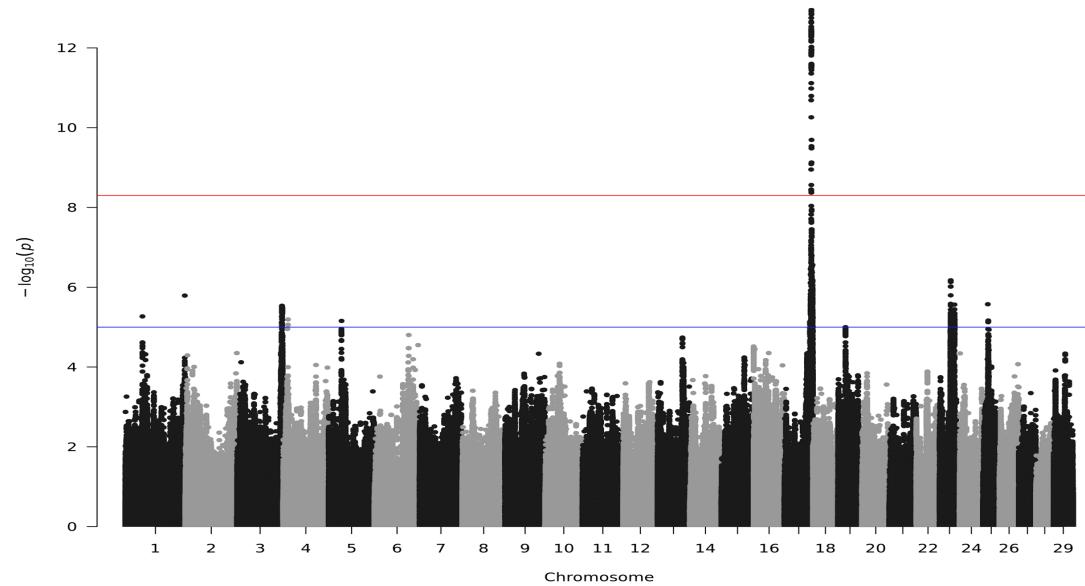


Manhattan Plots



Non-Return-Rate
Rind bei Braunvieh
9'972'582 SNP
werden getestet

Manhattan Plots





Genomische Selektion

Inhalt grossteils aus:

Kursunterlagen Genomic Selection, Ben Hayes, 2015

Mrode, 2015: Linear Models for the Prediction of Animal Breeding Values

Heutige Vorlesung

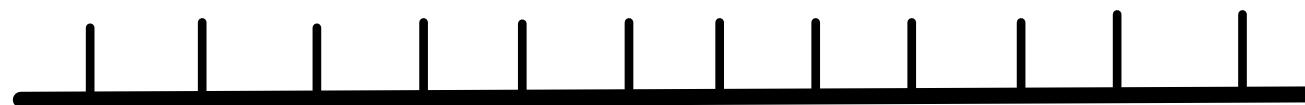
- Was ist genomische Selektion?
- Methoden zur Schätzung genomischer Zuchtwerte
 - SNP-BLUP
 - G-BLUP
 - Übung: genomische Zuchtwerte mit SNP-BLUP schätzen

Genomische Selektion

- Genetische Marker könnten in der Tierzucht/Pflanzenzucht folgendermassen genutzt werden:
 - Durchführen einer GWAS
 - Höchst signifikante SNP werden in Markergestützten Selektion verwendet
- Merkmale werden aber oft von vielen QTL (mit kleinen Effekten) kontrolliert → MAS nur limitiert erfolgreich, weil nur ein Teil der genetischen Varianz abgedeckt werden würde
- Idee: Alle Chromosomensegmente mit vielen Markern (tausenden) über das gesamte Genom verteilt verfolgen und **gleichzeitig** berücksichtigen
 - → somit sollte es theoretisch möglich sein, alle QTL zu verfolgen
 - **Voraussetzung:** Marker sind im **Kopplungsungleichgewicht (LD)** mit QTL
 - Markerdichte muss hoch genug sein, damit LD gewährleistet wird

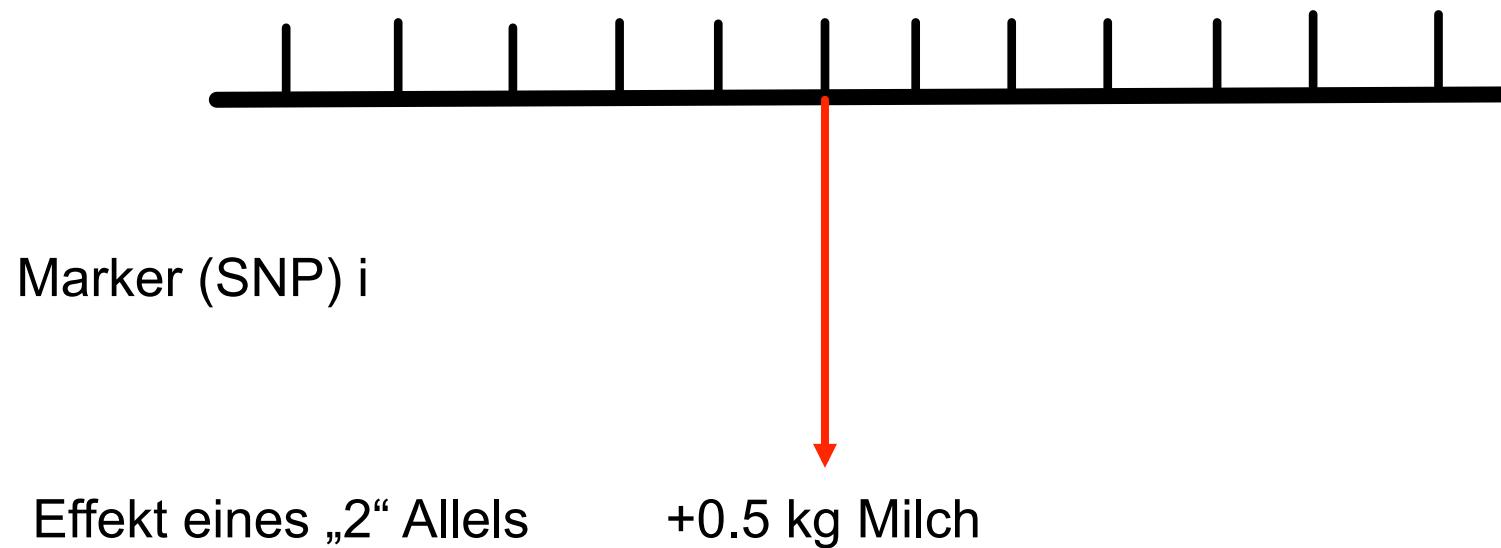
Genomische Selektion

Marker über das gesamte Chromosome verteilt



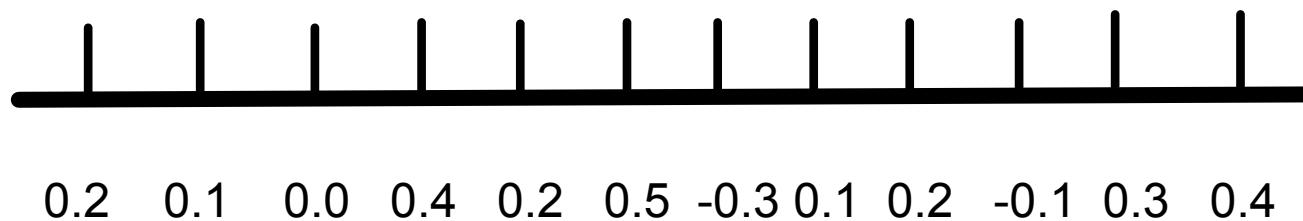
Genomische Selektion

Marker über das gesamte Chromosome verteilt



Genomische Selektion

Marker über das gesamte Chromosome verteilt



Schätzen aller SNP-Effekte

Direkt genomicscher Zuchtwert:

$$DGZW = \sum_i^n X_i \hat{g}_i$$

n = Anzahl SNP
X = Designmatrix
 \hat{g}_i = SNP-Effekt

Genomische Selektion: eine neue Ära ...

- Theorie zur Genomischen Selektion wurde publiziert bevor technische Möglichkeiten gegeben waren:
- *Meuwissen, T., Hayes, B., Goddard, M. 2001: Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819-1829*
- Zeigten züchterische Möglichkeiten von hochdichten Markerkarten für die Zuchtwertschätzung auf (Simulationsstudien).
- Sicherheit des genomischen Zuchtwertes liegt bei 80%

Genomische Selektion: eine neue Ära ...

ORIGINAL ARTICLE

Strategy for applying genome-wide selection in dairy cattle

L.R. Schaeffer

Department of Animal and Poultry Science, Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada

(Journal of Animal Breeding and Genetics, 2006, 123, 218-223)

- Schaeffer errechnete für Rinderzuchtprogramme eine Kostenersparnis von 92% und eine Verdoppelung des Zuchtfortschritts
- bei Anwendung der Methode nach Meuwissen et al. (2001) unter der Annahme Sicherheit des genom. Zuchtwertes liegt bei 80%
- Genomischer Zuchtwerte können bereits bei Geburt von Stieren berechnet werden.
- Entwicklung kostengünstiger Hochdurchsatzverfahren zur Genotypisierung → erster SNP-chip 2005/6



Seitdem Fülle an SNP-Chips entwickelt ...

SNP-Chip Name	GGPLDv2 (9K)	GGPLDv3 (26K)	50Kv1 (54K)	50Kv2 (54K)	GGPHD (80K)	GGPHD (150K)	HD (850K)
GGPLDv2 (9K)	8.762	92.9%	92.6%	92.3%	92.1%	93.6%	95.8%
GGPLDv3 (20K)		19.725	42.1%	41.9%	46.3%	95.0%	98.8%
50Kv1 (54K)			54.001	96.6%	52.4%	77.2%	90.2%
50Kv2 (54K)				54.609	51.7%	75.8%	90.5%
GGPHD (80K)					76.999	95.9%	96.8%
GGPHD (150K)						139.481	96.3%
HD (850K)							777.962

Auslösen einer weltweiten Genotypisierungswelle ...

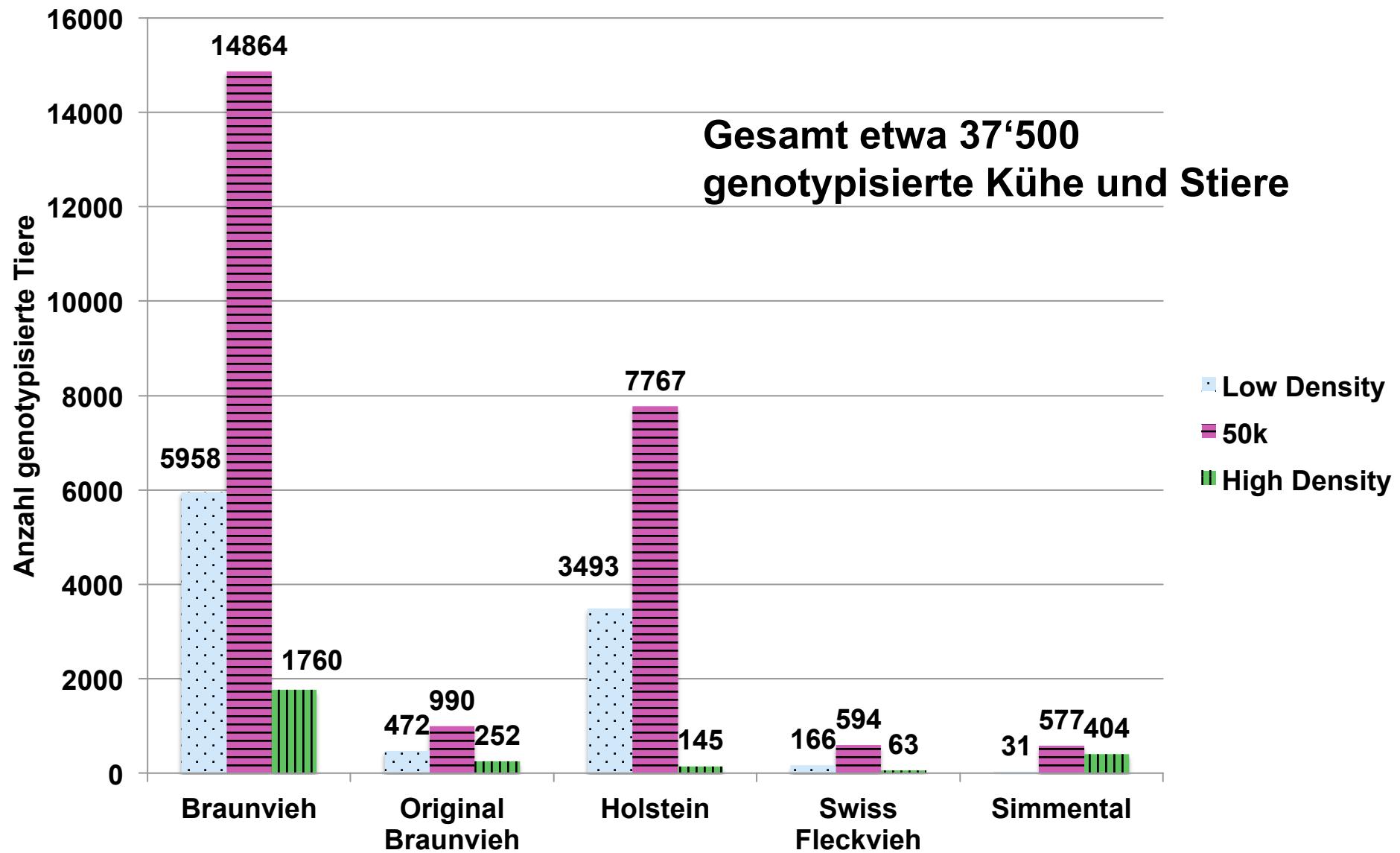
Anzahl Genotypen USA: https://www.cdcb.us/Genotype/cur_density.html

**Genotypes included in evaluations by breed, chip density,
presence of phenotypes (old vs. young),
and evaluation year-month (cumulative).**

HO	JE	BS	AY	>= 50K				Lower Density (< 50K)				Imputed*	Total		
Run	Old		Young		Old		Young		Old		Young				
Date	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female			
1512	28806	31695	57330	43749	111	157896	73375	586025	3391	1582	983960				
1511	28181	29937	57687	44488	97	134563	71713	589118	3364	1587	960735				
1510	28182	29852	57390	43841	97	133673	69749	567051	3367	1578	934780				

0906	7883	3049	11459	2974	0	0	0	0	0	0	0	25365
0904	7600	2711	9685	1937	0	0	0	0	0	0	0	21933
0901	7330	1319	6448	1560	0	0	0	0	0	0	0	16657

Anzahl genotypisierte Stiere und Kühe Schweiz

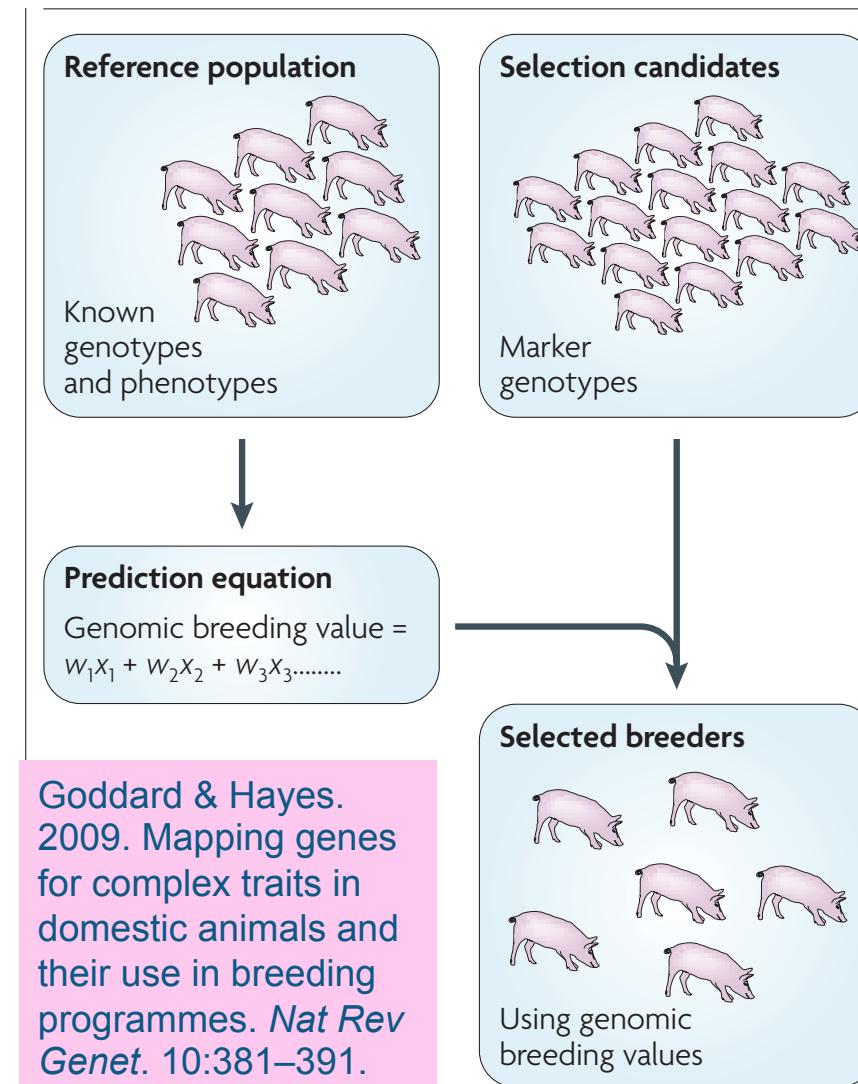


Was ist Genomische Selektion?

- **Selektionsentscheidungen** werden aufgrund von **genomischen Zuchtwerten** (bzw. einer Kombination von genomischen und traditionellen Zuchtwerten) getroffen
- Prinzip der genomischen Zuchtwertschätzung beinhaltet zwei Schritte:
- 1. Schätzen aller SNP-Effekte in einer Referenzpopulation (Genotypen und sichere geschätzte traditionelle Zuchtwerte)
- 2. SNP-Effekte werden genutzt um für junge genotypisierte Tiere (ohne traditionellen Zuchtwert) einen genomischen Zuchtwert zu schätzen

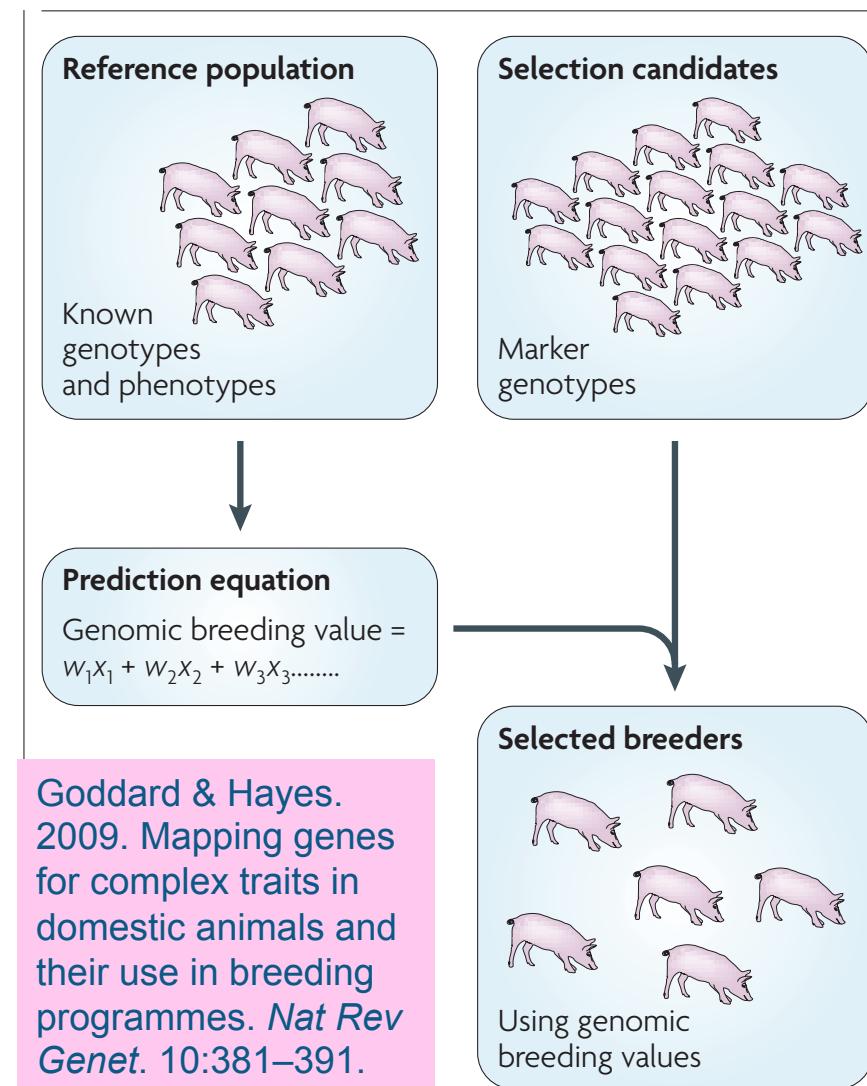
Was ist Genomische Selektion?

Referenzpopulation
(Referenzstichprobe,
Kalibrierungsstich-probe)
mit sicher geschätzte
Zuchtwerte → als
Phänotypische
Beobachtung werden
sicher geschätzte
Zuchtwerte verwendet



Was ist Genomische Selektion?

- Herausforderung in der Kalibrierung: Sehr viele SNP-Effekte (tausende) müssen anhand von vergleichsweise wenigen phänotyp. Beobachtungen geschätzt werden
- $p >>>> n$
- überparameterisiert
- → Schätzmethoden-entwicklung



Methoden genomische Zuchtwertschätzung

Lineares Modell – Schätzung SNP-Effekte mittels BLUP

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{i=1}^m \mathbf{M}_i \mathbf{g}_i + \mathbf{e}$$

\mathbf{y} = Vektor mit Beobachtungen

\mathbf{b} = Vektor der fixen Effekte

\mathbf{g} = zufällige genetischer Effekt von SNP i

\mathbf{X} = Designmatrix fixe Effekte

\mathbf{M} = Designmatrix zufällige SNP-Effekte (Anzahl Tiere X Anzahl SNP-Marker)

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{M} \\ \mathbf{M}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{M}'\mathbf{R}^{-1} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{M}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Methoden genomische Zuchtwertschätzung

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

y = Vektor mit Phänotyp (Merkmalswerte)

1_n = Vektor von „1“, teilt die Phänotypwerte y dem Mittelwert zu

μ = globaler Mittelwert

X = Designmatrix, teilt Phänotypwerte den Genotypen (SNP Marker) zu

g = zufälliger Effekt des SNP (Markers)

e = zufälliger Fehler

Annahme: SNP-Marker Effekte sind sehr klein und folgen einer

Normalverteilung: $V(g) \sim N(0, I\sigma_g^2)$

Methoden genomische Zuchtwertschätzung

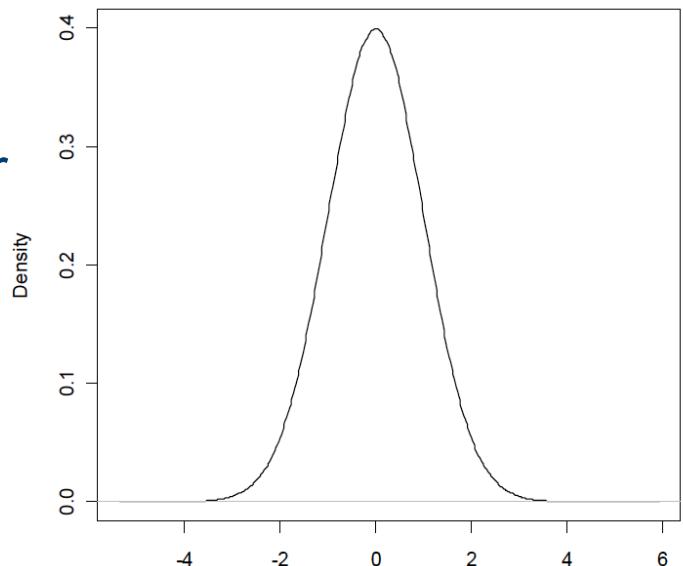
$$y = \mathbf{1}_n \boldsymbol{\mu} + \mathbf{X}\mathbf{g} + \mathbf{e}$$

Annahme:

SNP-Marker Effekte sind sehr klein und folgen einer Normalverteilung:

$$\mathbf{V}(\mathbf{g}) \sim N(0, I\sigma_g^2)$$

σ_g^2 Varianz der Markereffekte für alle Marker



Methoden genomische Zuchtwertschätzung

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

- BLUP Mischmodellgleichung
- \mathbf{I} = Einheitsmatrix („1er“ auf Diagonale, „0er“ Offdiagonalen)
- λ = Varianzenquotient

$$\lambda = \frac{\sigma_e^2}{\sigma_g^2} \quad \begin{array}{l} \sigma_e^2 \text{ Varianz der Residuen (zufälliger Rest)} \\ \sigma_g^2 \text{ Varianz der Markereffekte für alle Marker} \end{array}$$

SNP-BLUP – ein Beispiel ...

- 5 Stiere sind genotypisiert für 10 SNP-Marker, sichere Zuchtwerte
- SNP-Marker Allele sind A und T → Genotypen AA, AT, TT werden mit 0, 1, 2 kodiert (zählen wieviele Kopien des T-Allels)



Tier	Phänotyp	SNP									
		1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

SNP-BLUP – ein Beispiel ...

Tier	Phänotyp	SNP									
		1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

$$y = \mathbf{1}_n \hat{\mu} + \mathbf{X}g + e \quad \rightarrow \quad \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' y \\ \mathbf{X}' y \end{bmatrix}$$

Annahme $\lambda = 1$

SNP-BLUP – ein Beispiel ...

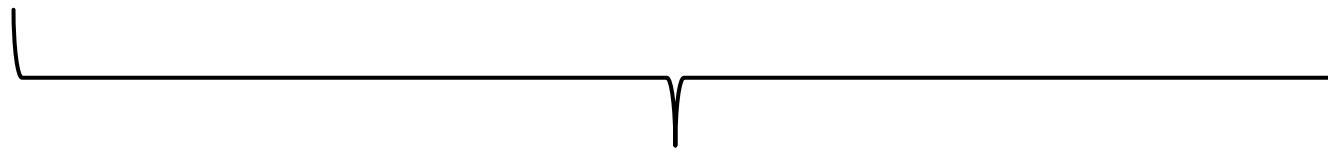
$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

$$\mathbf{1}'_n = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 5$$

SNP-BLUP – ein Beispiel

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

Tier	Phänotyp	1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1



X-Matrix

SNP-BLUP – ein Beispiel

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + I\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

Tier	Phänotyp	1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

SNP

X-Matrix

SNP-BLUP – ein Beispiel

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} - \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

Tier	Phänotyp	SNP									
		1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

X-Matrix

\mathbf{I} = Einheitsmatrix (Anzahl Marker x Anzahl Marker)

SNP-BLUP – ein Beispiel

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + I\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} 0.47 \\ 0.29 \\ -0.05 \\ -0.05 \\ 0.08 \\ -0.02 \\ 0.13 \\ 0.13 \\ -0.08 \\ 0.11 \\ -0.08 \end{bmatrix}$$

Mittelwert = 0.47

Effekt des „2“ – Allels = 0.29 für SNP 1

Wird auch **Kalibrierungsgleichung**
(Prediction equation) bezeichnet

SNP-BLUP – ein Beispiel ...

Mit Hilfe der Kalibrierungsgleichung können nun genomische Zuchtwerte für junge genotypisierte Selektionskandidaten geschätzt werden:



www.natursicht.ch

Nachkommen	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	2	1	0	1
2	1	0	0	1	1	1	2	1	0	1
3	1	0	0	1	1	1	2	1	0	1
4	1	0	0	1	1	1	2	1	0	1
5	0	0	0	0	0	0	1	2	0	2

SNP-BLUP – ein Beispiel ...

Mit Hilfe der Kalibrierungsgleichung können nun genomische Zuchtwerte für junge genotypisierte Selektionskandidaten geschätzt werden:



www.natursicht.ch

Nachkommen	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	2	1	0	1
2	1	0	0	1	1	1	2	1	0	1
3	1	0	0	1	1	1	2	1	0	1
4	1	0	0	1	1	1	2	1	0	1
5	0	0	0	0	0	0	1	2	0	2

X-Matrix

SNP-BLUP – ein Beispiel ...

Direkt genomischer Zuchtwert

$$DGZW = X\hat{g}$$



www.natursicht.ch

SNP										
1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	2	1	0	1	
1	0	0	1	1	1	2	1	0	1	
1	0	0	1	1	1	2	1	0	1	
1	0	0	1	1	1	2	1	0	1	
0	0	0	0	0	0	1	2	0	2	

*

$$= \begin{vmatrix} 0.29 \\ -0.05 \\ -0.05 \\ 0.08 \\ -0.02 \\ 0.13 \\ 0.13 \\ -0.08 \\ 0.11 \\ -0.08 \end{vmatrix} \quad \text{DGZW} \begin{vmatrix} 0.47 \\ 0.58 \\ 0.58 \\ 0.58 \\ -0.20 \end{vmatrix}$$

SNP-BLUP – ein Beispiel ...

- Hier im Beispiel Annahme, dass σ_g^2 bekannt ist ($\lambda = 1$)
- In der Praxis ist dies sehr wahrscheinlich nicht der Fall
- Schätzung der additiv genetischen Varianz(z.B. REML)
- Diese wird durch die Anzahl an Marker geteilt $\sigma_g^2 = \sigma_a^2 / m$
- Hier werden allerdings unterschiedliche Allelfrequenzen der Marker nicht berücksichtigt.
- Deshalb:
$$\sigma_g^2 = \sigma_a^2 / 2 \sum_{j=1}^n p_j(1 - p_j)$$
- Kreuzvalidierungen durchführen

Eine äquivalente Methode ...

- ... Ist die Schätzung von genomischen Zuchtwerten mit einer **genomischen Verwandtschaftsmatrix**
- Dabei wird in den BLUP Mischmodellgleichungen die additiv genetische Verwandtschaftsmatrix A^{-1} durch die genomische Verwandtschaftsmatrix G^{-1} ausgetauscht → **G-BLUP**
- Die G-Matrix beschreibt die Ähnlichkeit von Tieren auf Markerebene
- Es werden die „wahren“ Verwandtschaftsbeziehungen zwischen Tieren abgebildet

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

The term $\lambda\mathbf{G}^{-1}$ in the second matrix is circled in red.

Verwandtschaftsmatrix A

1.00	0.25	0.25	0.25	0.00	0.00	0.00	0.00	0.00
0.25	1.00	0.25	0.25	0.00	0.00	0.00	0.00	0.00
0.25	0.25	1.00	0.25	0.00	0.00	0.00	0.00	0.00
0.25	0.25	0.25	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.25	0.25	0.25	0.25
0.00	0.00	0.00	0.00	0.25	1.00	0.25	0.25	0.25
0.00	0.00	0.00	0.00	0.25	0.25	1.00	0.25	0.25
0.00	0.00	0.00	0.00	0.25	0.25	0.25	1.00	0.25
0.00	0.00	0.00	0.00	0.25	0.25	0.25	0.25	1.00

Realisierte genomische Verwandtschaftsmatrix

1.00	0.30	0.21	0.28	0.00	0.00	0.00	0.00	0.00
0.30	1.00	0.20	0.27	0.00	0.00	0.00	0.00	0.00
0.21	0.20	1.00	0.22	0.00	0.00	0.00	0.00	0.00
0.28	0.27	0.22	1.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.31	0.24	0.26	0.29
0.00	0.00	0.00	0.00	0.31	1.00	0.22	0.30	0.23
0.00	0.00	0.00	0.00	0.24	0.22	1.00	0.26	0.22
0.00	0.00	0.00	0.00	0.26	0.30	0.26	1.00	0.29
0.00	0.00	0.00	0.00	0.29	0.23	0.22	0.29	1.00

Realisierte genomische Verwandtschaftsmatrix

1.00	0.30	0.21	0.28	0.01	0.03	0.07	0.02	0.01
0.30	1.00	0.20	0.27	0.05	0.02	0.01	0.03	0.06
0.21	0.20	1.00	0.22	0.02	0.01	0.08	0.04	0.02
0.28	0.27	0.22	1.00	0.07	0.03	0.01	0.05	0.04
0.01	0.05	0.02	0.07	1.00	0.31	0.24	0.26	0.29
0.03	0.02	0.01	0.03	0.31	1.00	0.22	0.30	0.23
0.07	0.01	0.08	0.01	0.24	0.22	1.00	0.26	0.22
0.02	0.03	0.04	0.05	0.26	0.30	0.26	1.00	0.29
0.01	0.06	0.02	0.04	0.29	0.23	0.22	0.29	1.00

G-Matrix

- X-Matrix mit Genotypen (0, 1, 2) wird mit Allelfrequenz skaliert: $W_{ij} = X_{ij} - 2p_j$

$$G = WW' / 2 \sum_{j=1}^n p_j(1 - p_j)$$

- Zuchtwerte aus unserem SNP-BLUP Beispiel:

$$\begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{Z} \\ \mathbf{Z}' \mathbf{1}_n & \mathbf{Z}' \mathbf{Z} + \mathbf{G}^{-1} \frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{Z}' \mathbf{y} \end{bmatrix}$$

G-BLUP

- G-BLUP wird in der Routine für die genomische Zuchtwertschätzung sehr häufig angewendet
- ... Häufiger als die direkte Schätzung von SNP-Effekten
- „Nur“ A durch G ersetzen
- Nützlich für Populationen ohne gute Pedigreeinformation
→ Marker bilden Verwandtschaft zwischen Tieren ab

Weitere Methoden

- Bayessche Methoden mit unterschiedlichen Prior-Annahmen für die SNP-Effekte
- Eine kurze Übersicht dazu gibt es das nächste Mal ...

Wie gut funktioniert die genomische Zuchtwertschätzung?

- Durchführen von Kreuzvalidierungen



Wie gut funktioniert die genomische Zuchtwertschätzung?

- Durchführen von Kreuzvalidierungen

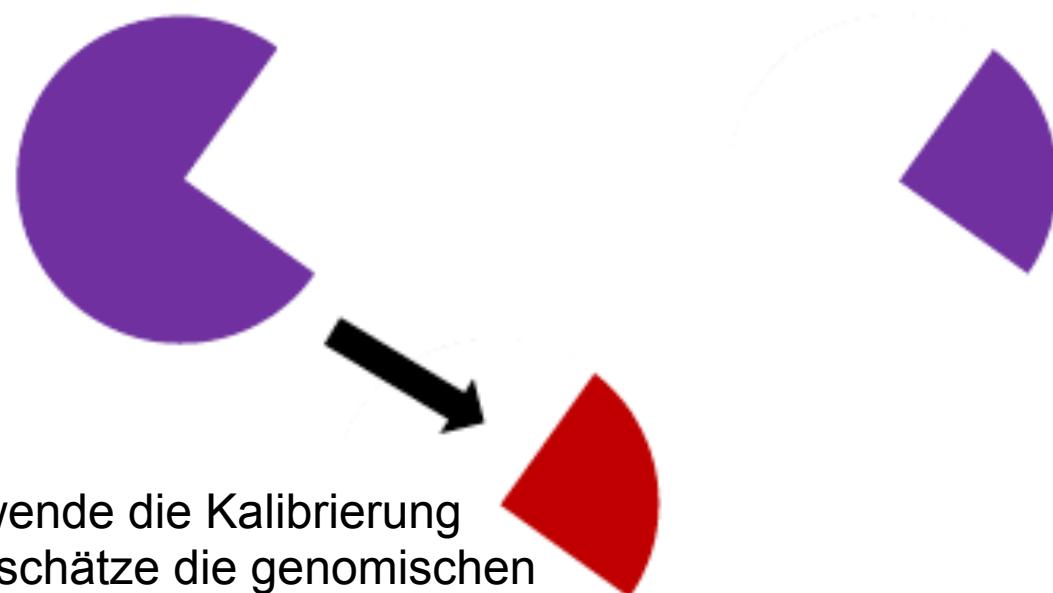


Aufteilung erfolgt zufällige und wiederholt.

Christian Edel, 2010

Wie gut funktioniert die genomische Zuchtwertschätzung?

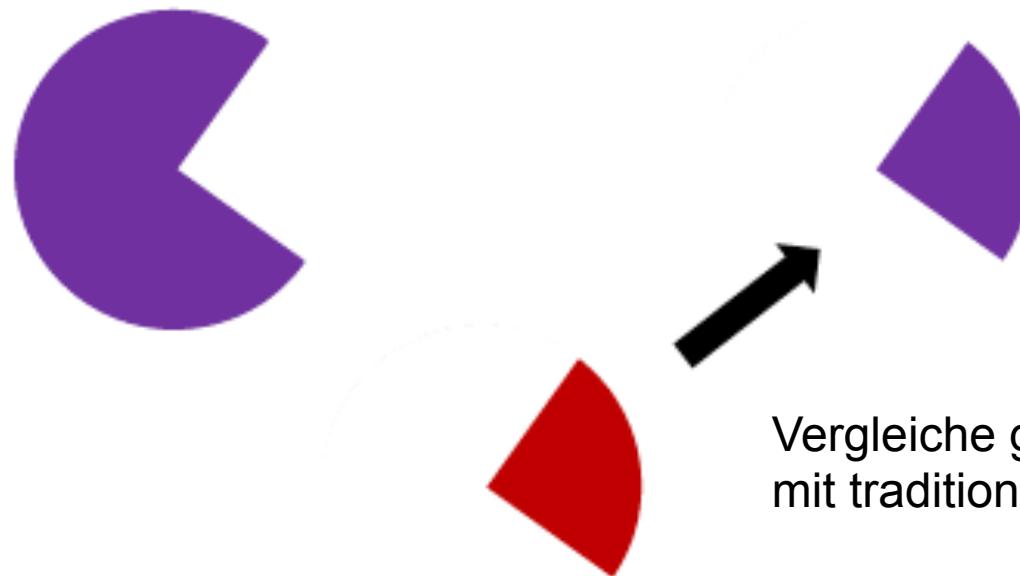
- Durchführen von Kreuzvalidierungen



Christian Edel, 2010

Wie gut funktioniert die genomische Zuchtwertschätzung?

- Durchführen von Kreuzvalidierungen



Christian Edel, 2010

Nächste Vorlesung 11. Dezember 2015

- Wiederholung gewünschter Themen
 - Direktes Aufstellen der A^{-1}
 - ...
 - Weitere gewünschte Themen bitte bis Sonntag, 6.12.2015 melden: [birgit.gredler\(at\)qualitasag.ch](mailto:birgit.gredler(at)qualitasag.ch)