A combined physical-genetic map for dairy cattle

A. Hampel, F. Teuscher, D. Wittenburg

69th Annual Meeting of the EAAP in Dubrovnik, Croatia

August 29, 2018

Background: half-sib family

Family structure influences estimates of population-genetic parameters

⇨ **recombination rate (θ) and linkage disequilibrium (LD)**

Recombinant and non-recombinant offspring:

Probability: $\frac{1}{2}(1-\theta)$ $\frac{1}{2}\theta$ $\frac{1}{2}(1-\theta)$ $\frac{1}{2}\theta$

Paternal LD: $D^{sire} = \frac{1-2\theta}{4}$

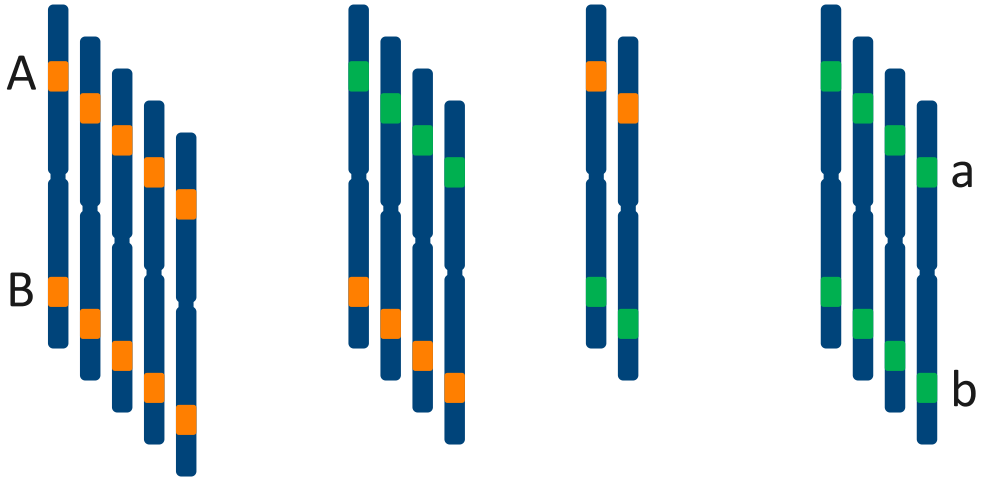
Background: half-sib family (cont.)

Family structure influences estimates of population-genetic parameters

⇨ **recombination rate (θ) and linkage disequilibrium (LD)**

Frequencies of maternal gametes: $p_{AB}, p_{aB}, p_{Ab}, p_{ab}$

$$D^{dam} = p_{AB}p_{ab} - p_{Ab}p_{aB}$$



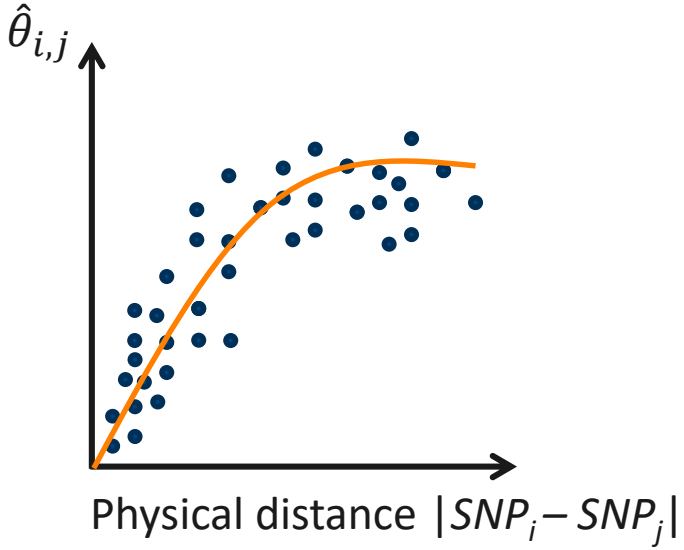
Objectives

(1) Study relationship between recombination rate and physical distance (consensus curve)
   - Estimates of recombination rate between all pairs of SNPs

(2) Identify regions on the genome affecting recombination rate (hot/cold spots)

⇨ Improvement of breeding designs

⇨ Management of genetic diversity



Part 1: Estimation of recombination rate and LD

Estimation of recombination rate and LD

- **Observed**: genotype frequencies of progeny at SNP pair ($n_{AA,BB}$, $n_{AA,Bb}$, …)
- Log-likelihood function (dependend on sire phase)

$$\log LF(\pi_{AA,BB},\pi_{AA,Bb},\ldots,\pi_{aa,bb}\,|\,n_{AA,BB},n_{AA,Bb},\ldots,n_{aa,bb}) = \sum_{\substack{i\in\{AA,Aa,aa\}\\ k\in\{BB,Bb,bb\}}} n_{i,k}\log\pi_{i,k} + \text{constant}$$

with, e.g., $\pi_{AA,Bb} = \frac{1}{2}(1-\theta)p_{Ab} + \frac{1}{2}\theta p_{AB}$

- Expectation maximisation algorithm (EM; Gomez-Raya 2012 Genetics)
- **Unknown**: $p_{AB}, p_{Ab}, p_{aB}, \theta$ or equivalently $p_1, p_2, D^{dam}, D^{sire}$

(relationship $p_{AB} = p_1 p_2 + D^{dam}$)

Surface of log likelihood

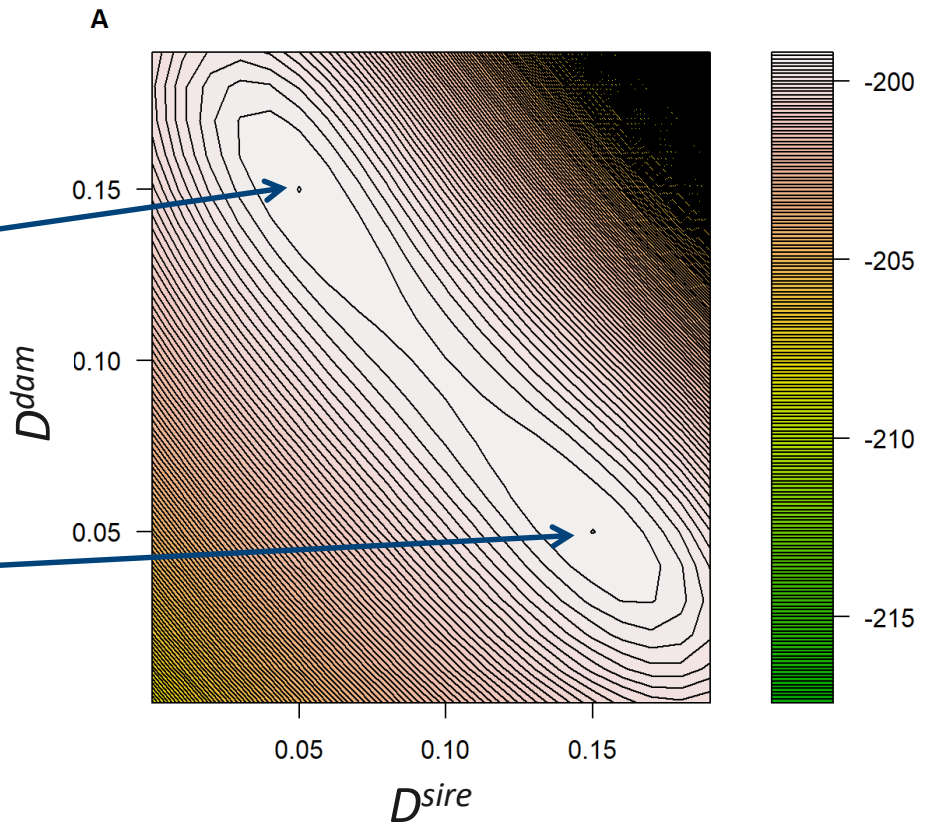
$p_1 = p_2 = 0.5$

$D^{dam} = 0.15$

$D^{sire} = 0.05$

$D^{dam} = 0.05$

$D^{sire} = 0.15$



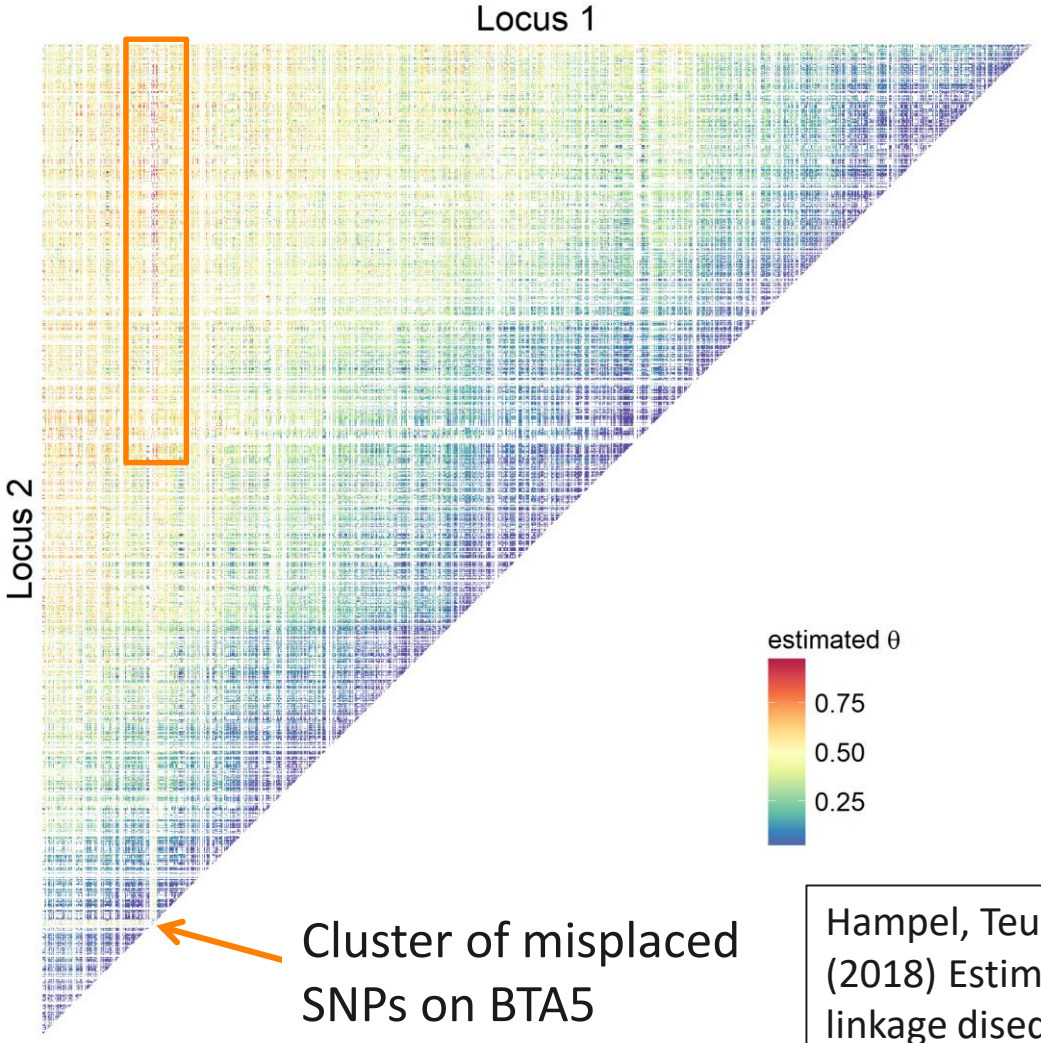
EM approach based on the log-**likelihood function** converges to a local mode. Which one depends on the start values.

⇨ exploit **relationship between modes**

Empirical data set

- 5 half-sib families of Holstein cows, 265 offspring in total (minimum family size 30)
- 39,780 SNPs on the autosomes (e.g. BTA1 with 2,560 SNPs)
- Physical order according to genome assembly **Btau 4.2**
- Parameter estimation for SNP pair if (at least one) sire is double heterozygous
  - Run EM algorithm twice
  - Select most likely set of estimates,
    or employ information from neighbouring SNPs

Results: recombination rate



Cluster of misplaced SNPs on BTA5

BTA1-29: 12.8 million estimates

Pattern search for unusually large estimates of recombination rate (θ) to close SNPs or low estimates to distant SNPs revealed **candidates of misplacement**.

⇨ partly proved with newer assembly UMD 3.1.1

Hampel, Teuscher, Gomez-Raya, Doschoris, Wittenburg (2018) Estimation of recombination rate and maternal linkage disequilibrium in half-sibs. *Front Genet.*

Part 2: Consensus curve

Smoothing of estimates

Categorical variable: chromosome window (500 kb)

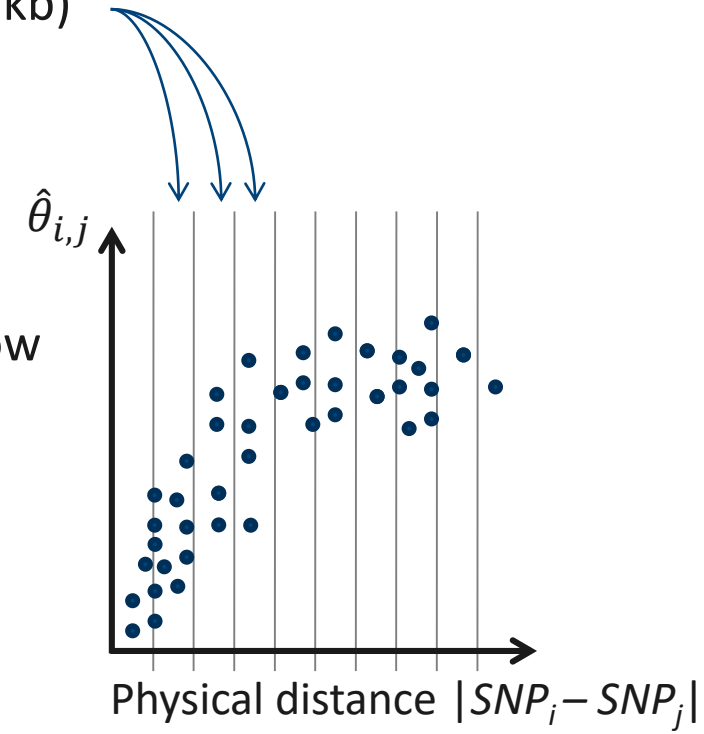
Model: $\theta = W\alpha + (X_1 + X_2)\beta + e$

W: assign distance between SNPs to window

X_1, X_2: assign position of 1st and 2nd SNP to window

⇨ Account for abundance of estimates by **weight**=1/(relative distance)

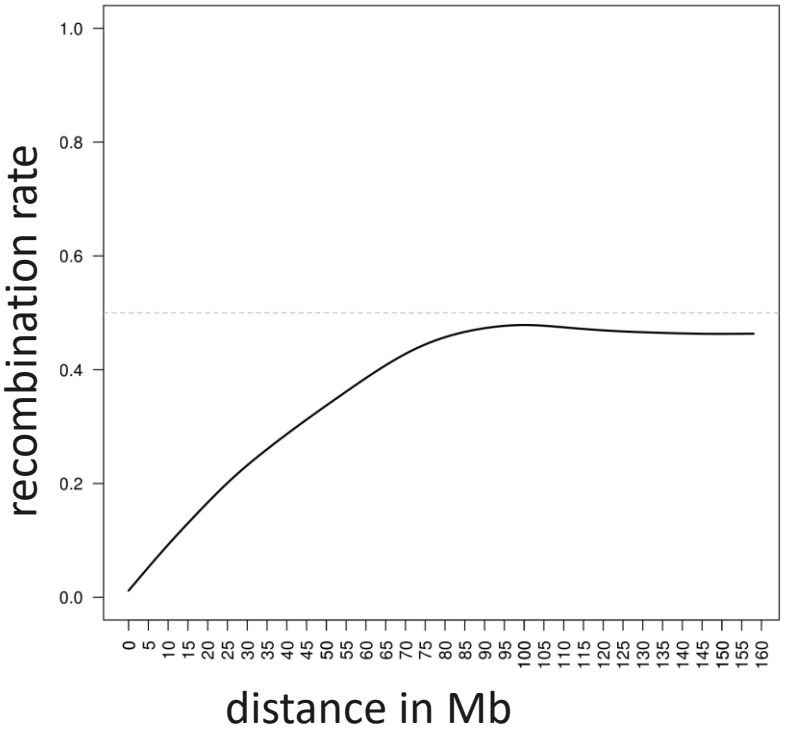
⇨ Account for proximity of SNPs by **scaling** of entries in X_1 and X_2 depending on distance



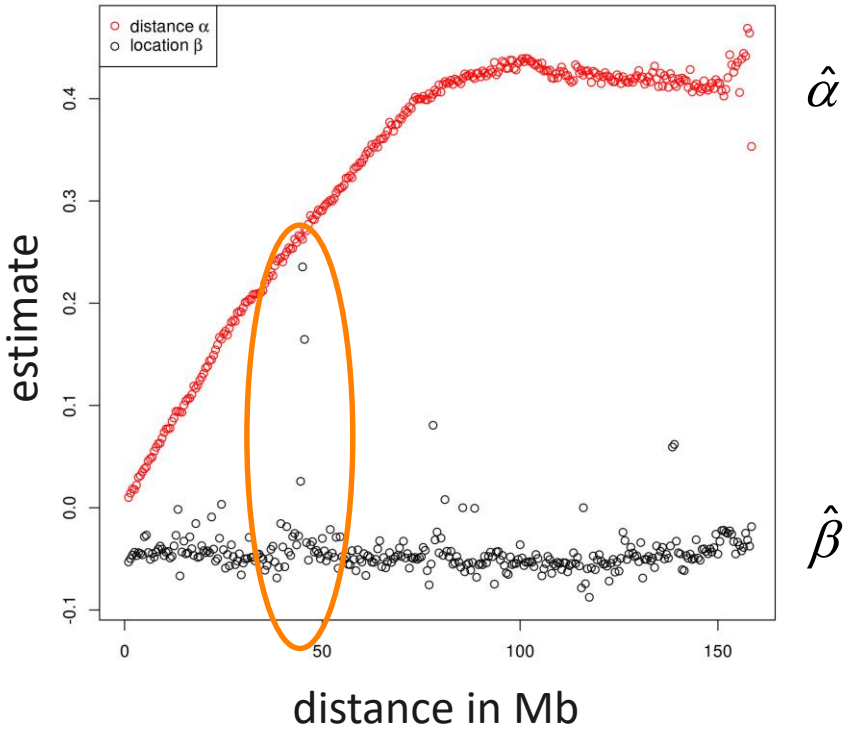
⇨ Allocation of large vectors/matrices and least-squares method with **biglm** (R)

Results: empirical data on BTA1 (UMD 3.1.1)

Smooth curve through fitted values



Coefficients of linear model



Discussion: consensus curve

Ongoing work on combined physical-genetic map

SNP-ID	Physical position (bp)	Genetic position (cM)
…	…	…

- Consider estimates of local effects and employ genetic map function
  - Proper scaling of entries in X_1 and X_2 required
  - Update analysis (genome assembly ARS-UCD 1.2)
- Evaluate precision/quality of consensus curve

  ⇨ Project starting in **April 2019 (postdoc)**

Summary

- Estimates of population parameters using genotypes of half-siblings
- Detection of misplaced SNPs in the genome assembly
- Smooth curve of $\hat{\theta}$ vs. physical distance accounting for local effects (verification with simulated data)
  - Deeper verification of weighting and scaling terms is required

Thank you for your attention!

LEIBNIZ INSTITUTE
FOR FARM ANIMAL BIOLOGY

Leibniz-Institut für Nutztierbiologie FBN

Wilhelm-Stahl-Allee 2
18196 Dummerstorf
Germany

Contact

Dr. Dörte Wittenburg

Phone: +49 38208 68 930
Fax: +49 38208 68 902
E-Mail: wittenburg@fbn-dummerstorf.de
Internet: www.fbn-dummerstorf.de

Relationship between modes

System of equations for the reduced problem, i.e. p_1, p_2 known (Bonk et al. 2016)

- Covariance between codes for additive SNP effects at two loci

$$\text{cov}_{add} = D^{sire} + D^{dam}$$

AA	Aa	aa
1	0	-1

- Covariance between codes for dominance effects

$$\text{cov}_{dom} = 16D^{sire}D^{dam} + 4D^{sire}\left(1-2p_1\right)\left(1-2p_2\right)$$

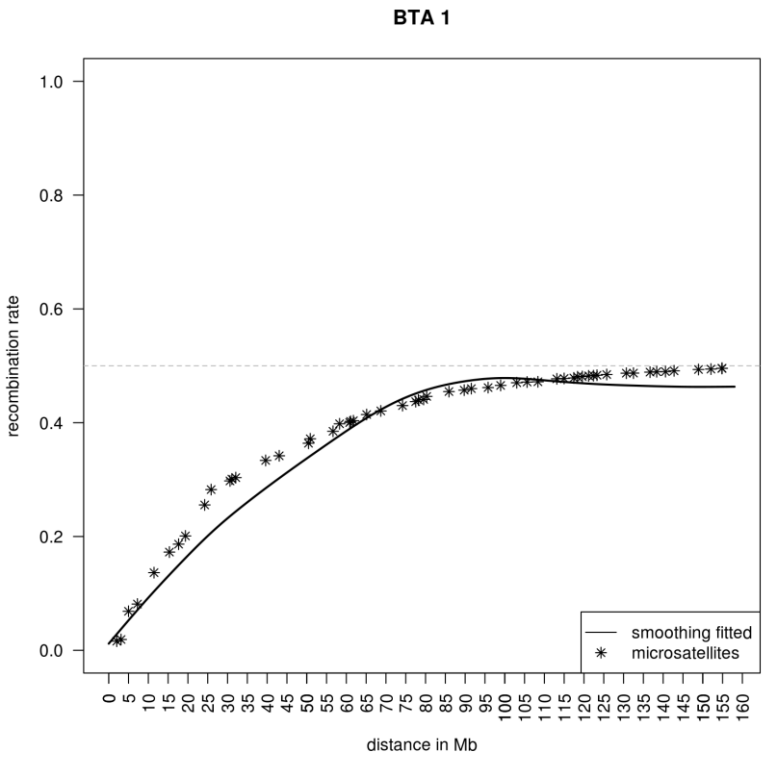
AA	Aa	aa
-1	1	-1

⇨ Complementary solutions

$$D_{II}^{sire} = D_{I}^{dam} + \frac{1}{4}(1-2p_1)(1-2p_2) \quad \text{and} \quad D_{II}^{dam} = D_{I}^{sire} - \frac{1}{4}(1-2p_1)(1-2p_2)$$

Validation with microsatellites

- 1,281 microsatellites from the NCBI database (UMD 3.1)
- Information about physical and genetic positions
- Quality and plausibility check: **761** micros
- Kosambi map function for calculation of θ
  - Based on 206 animals of different spe
  - θ between microsatellites is average c rate



Simulation study

- $n = 1{,}000$ meioses, $m = 1{,}000$ markers
- Chromosome length 200 Mb (roughly equivalent to $L = 2$ Morgan)
- Number of cross-overs (x) according to Poisson distribution (Haldane 1919)

k	Pr(x=k)
0	0.014
1	0.271
2	0.271
…	…
10	$3.8 \cdot 10^{-5}$

$$\Pr(x = k) = \frac{L^k e^{-L}}{k!}$$

- Uniform distribution of cross-over positions
- Hot spot between 50 and 70 Mbp with 50% probability

Results: simulated data