An Efficient Algorithm for QTL Analysis by Multivariate LMM

Hyeonju Kim, Saunak Sen
Division of Biostatistics
University of Tennessee Health Science Center
E-mail: hkim@uthsc.edu

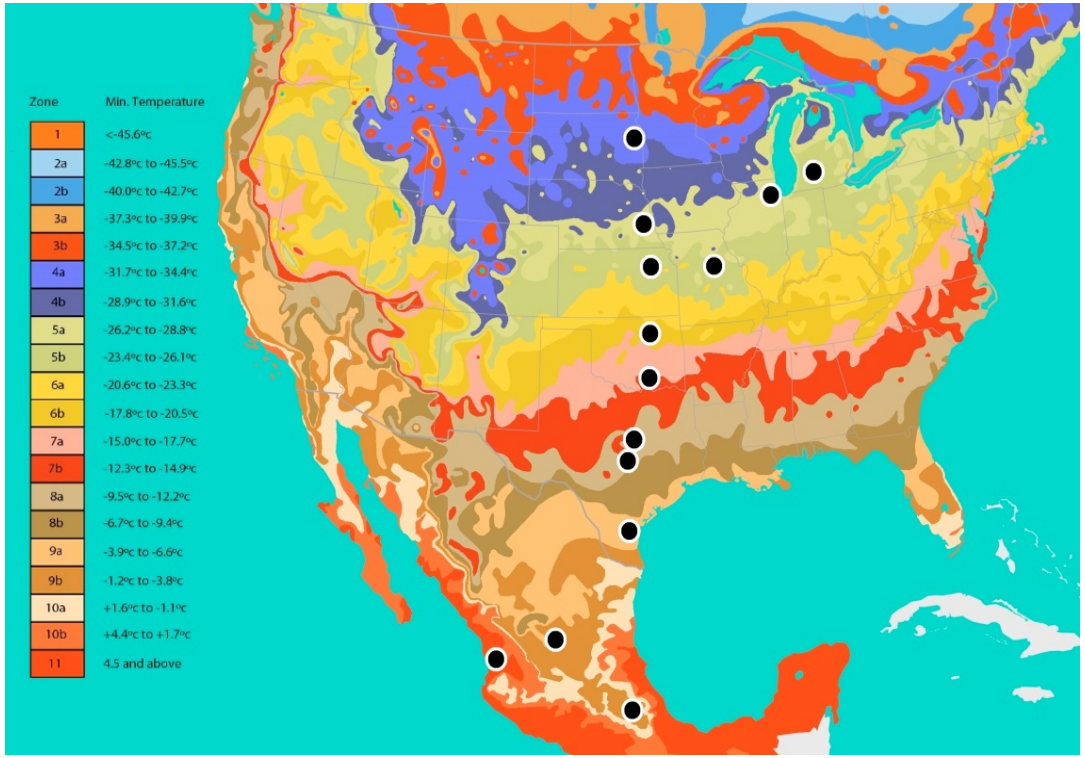
Motivation : High-dimensional data in genetics

- Prototype example datasets :

1. Switchgrass data

- Segregating plant populations grown in multiple sites in multiple years (10 latitudes by 3 yrs for 4 traits)
- Traits : Biomass, Flowering time, Height, # of Tillers
- Information annotating the environments (ex. Min/max/avg air temperatures, precipitation, Soil ingredients, latitudes/longitude)

- Goal: Identify the genomic regions (QTL) relying on latitude that accounts for weather patterns varying by year using GWAS



(figure & data by Dr. Tom Juenger in Integrative Biology at UT-Austin)

2. F2 intercross between Gough Island mice and WSB/EiJ

- Data from M. M. Gray, M. D. Parmenter, C. A. Hogan, I. Ford, R. J. Cuthbert, P. G. Ryan, K. Broman, and B. A. Payseur. Genetics of rapid and extreme size evolution in island mice. *Genetics*, 201(1):213-228, 2015.
- F2 intercross genotype probabilities (12,777 markers) excluding Chr X
- Traits: body weights weekly measured for 16 weeks (1212 individuals x 16 traits)
- Z : a matrix of B-spline(df=4), $K_C = I$

3. DO/HS data

- Collected from Jackson laboratory by Drs. Vivek and Chesler
- Subset of phenotypes : distance travel measured per 1 minute for 12 minutes (1452 individuals)
- Allele probabilities (8 alleles x 106,047 markers excluding Chr X)
- Challenges : large-sized allele probabilities (10GB) and genotype probabilities (115GB) >> out of memory issue when reading and difficult in doing permutation test
- May consider a way to estimate thresholds for LOD using diffusion processes

4. Arabidopsis thaliana data

- Data from J Agren et al. Genetic mapping of adaptation reveals fitness tradeoffs in Arabidopsis thaliana. *PNAS*, 110(52): 21077-21082, 2013.
- 400 individuals (recombinant inbred lines)
- Phenotype (the mean # of fruits per seedling planted in Italy & Sweden, July/2009-June/2012) : considered as a trait among 6 sites
- 2 genotypes : a(=0) from Italian parent, b(=1) from Swedish parent
- Total 5 Chromosomes : 699 markers after imputation
- Weather profile in both sites for K_C : soil, air temperatures, or draught index
- Our method: consistent results with detecting more QTL

Fig. 1 Map indicating the locations of the two study sites.

Figure from J. Agren et al. Reciprocal transplants demonstrate strong adaptive differentiation of the model organism Arabidopsis thaliana in its native range. *New Physiologist*,**194**:1112-1122, 2012.

Data Structure

$B_{p_1 \times q}$

$Z'_{q \times m}(phenotype\ covariates)$

$X_{n \times p_1}$ $(candidate\ markers)$

$Y_{n \times m}$ $(phenotypes/\ traits)$

$G_{n \times p}(background\ genotypes)$, $p \gg n$

$C'_{s \times m}(climatic\ information)$

Structure of B

- ex. In Arabidopsis data (n=400, m=6),

- $X = \begin{bmatrix} 1 & 0(=a) \\ \vdots & \vdots \\ 1 & 1(=b) \end{bmatrix}_{n \times p_1}$, $\mathrm{Z} = \begin{bmatrix} 1 & 1(Sweden) \\ \vdots & \vdots \\ 1 & -1(Italy) \end{bmatrix}_{m \times q}$

- $XBZ' \Longrightarrow$

$$B = \begin{matrix} \\ no\ qtl \\ qtl \end{matrix} \begin{matrix} overall\ site \\ mean\ site \\ qtl \end{matrix} \begin{matrix} site\ difference \\ mean\ difference\ in\ site \\ \boldsymbol{qtl \times site} \end{matrix}_{p_1 \times q}$$

Model : Multivariate Linear Mixed Model

$$\boldsymbol{Y = XBZ' + R + E},$$

where $\boldsymbol{E(Y^v) = (Z \otimes X)B^v}$, $\boldsymbol{var(R^v) = \tau^2 K_C \otimes K_G}$, $\boldsymbol{var(E^v) = \Sigma \otimes I_n}$.

- $Y_{n\times m}$: phenotypic trait values over m sites for n individuals
- $X_{n\times p_1}$: p_1 genotypes or genotype probabilities (and/or covariates) with intercepts for a candidate marker (total candidate markers = p)
- $Z_{m\times q}$: q phenotypic low-dimensional covariates (e.g. contrasts, basis functions, etc.)
- R: GxE random effects including high-dimensional column covariates
- E : model or environment errors independent of R
- K_G, K_C (symmetric positive definite): genetic relatedness (genetic kinship), climate relatedness (climate kinship)

▶Multivariate linear mixed model (Kernel regression) : GEMMA (X. Zhou & M. Stephens. Genome-wide efficient mixed model association, *Nature Methods*, 2014)

$$\boldsymbol{Y = XB + R + E},$$

$$\boldsymbol{Y^v \sim MVN\left(\ (I_m \otimes X)B^v,\quad \Sigma_1 \otimes K_G\ +\ \Sigma_2 \otimes I_n\ \right).}$$

Challenges are :

- high-dimensional fixed effects : $\boldsymbol{B_{p\times m}}$(not accounting for interaction between column and row covariates) ⇒ $\boldsymbol{B_{p\times q}}$ using low covariates (basis functions, contrasts, etc.), $\boldsymbol{Z_{m\times q}(q << m)}$
- High-dimensional random effects : $\boldsymbol{\Sigma_1 \Rightarrow \tau^2 K_C}$ using climatic information (generated by infinitesimal random GxE effects)
- Parameter estimation : $pm + m(m+1) \Longrightarrow pq + 1 + \frac{m(m+1)}{2}$

Goal

- How to efficiently estimate $\mathbf{B}$ (a coefficient matrix: QTL main and interaction effects), $\boldsymbol{\tau}^{\mathbf{2}}$(an overall genetic variance among sites), and $\boldsymbol{\Sigma}$ (residual covariance matrix)

- The objective is to maximize a log-likelihood function

$\Longrightarrow$ find maximum likelihood estimates (B, τ^2, Σ)

- Methods used : Expectation Conditional Maximization (ECM) + Speed restarting Nesterov's Accelerated Gradient scheme

Comparison in performance time and accuracy

- Competing method : GEMMA (general multivariate LMM run in C++)
- Our method (TLMM) : run in Julia v1.0.3 & also a general multivariate LMM for comparison
- Mouse_HS1940 data : n=1940 individuals, m=3 traits, p= 967 (Chr 1) for TLMM, 950 for GEMMA in 11182 candidate markers (computing a genetic kinship ($\boldsymbol{K_G}$)), set $\boldsymbol{K_C} = \boldsymbol{I_m}$ for TLMM (no gxe in data)
- Performance time : alternating full update (all parameters) and partial update (B only)

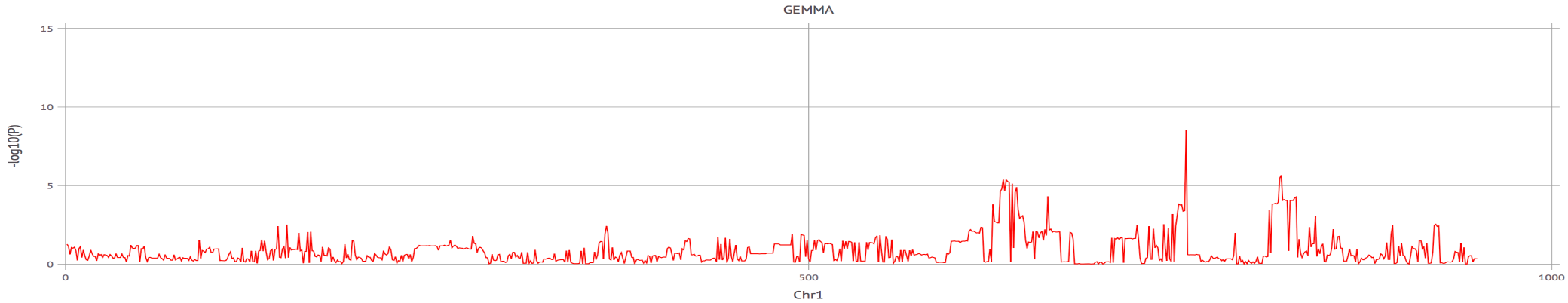
Model	implementation	Time
MVLMM	GEMMA (C++)	10.9 sec
MVLMM	Julia	7.5 sec
TLMM (ours)	Julia	6.5 sec

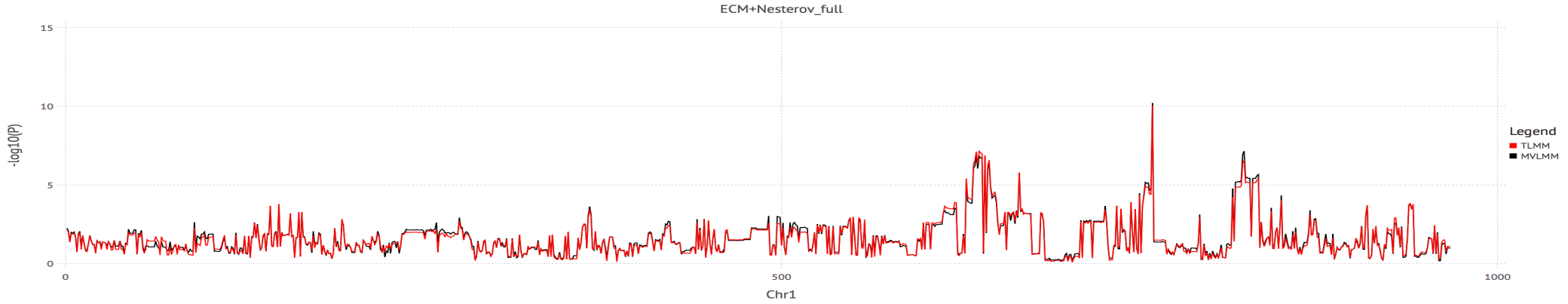
- Version info: OS: Linux Debian 4.19.37-5,

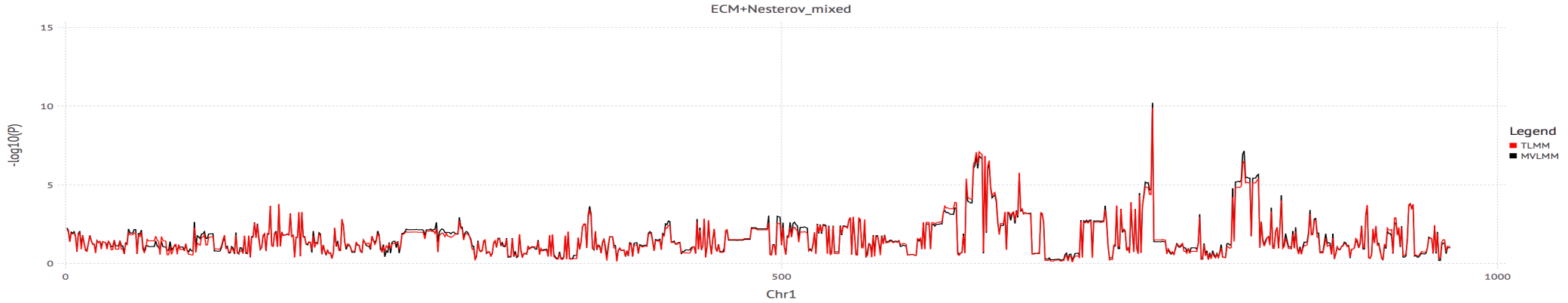
  CPU: Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz (8 threads)

- Accuracy

GEMMA	Nesterov+ ECM (alternating two types of updates)	
MVLMM	MVLMM	TLMM(ours)
• MLE for $\boldsymbol{\Sigma_1}$ in the null model	• MLE for $\boldsymbol{\Sigma_1}$ in the null model	• MLE for $\boldsymbol{\tau^2}$ in the null model
0.4076 -0.1665 0.4519 -0.1538 0.0028 1.2235	0.4526 -0.1848 0.5019 -0.1707 0.0030 1.3594	0.7160 ($\boldsymbol{\Sigma_1} \approx \tau^2 I$)
• MLE for $\boldsymbol{\Sigma_2}$ in the null model	• MLE for $\boldsymbol{\Sigma_2}$ in the null model	• MLE for $\boldsymbol{\Sigma_2}$ in the null model
0.7487 -0.1690 0.9055 -0.0400 -0.0003 0.6376	0.7034 -0.1505 0.8553 -0.0229 -0.0005 0.5016	0.6589 -0.1814 0.8186 -0.0546 0.0016 0.6183
• MLE log-likelihood	• MLE log-likelihood	• MLE log-likelihood
-7858.4450	-7858.4450	-7883.2440


GEMMA
-log10(P)
15
10
5
0
0
500
1000
Chr1


ECM+Nesterov_full
-log10(P)
15
10
5
0
0
500
1000
Chr1
Legend
TLMM
MVLMM


ECM+Nesterov_mixed
-log10(P)
15
10
5
0
0
500
1000
Chr1
Legend
TLMM
MVLMM

Analysis of F2 intercrosses between Gough Island mice and WSB/EiJ

- Data from M. M. Gray, M. D. Parmenter, C. A. Hogan, I. Ford, R. J. Cuthbert, P. G. Ryan, K. Broman, and B. A. Payseur. Genetics of rapid and extreme size evolution in island mice. *Genetics*, 201(1):213-228, 2015.
- F2 intercross genotype probabilities (12,777 markers) excluding Chr X
- Traits: body weights weekly measured for 16 weeks (1212 individuals x 16 traits)
- Z : a matrix of B-spline(df=4), $K_C = I$

Body weight graphs

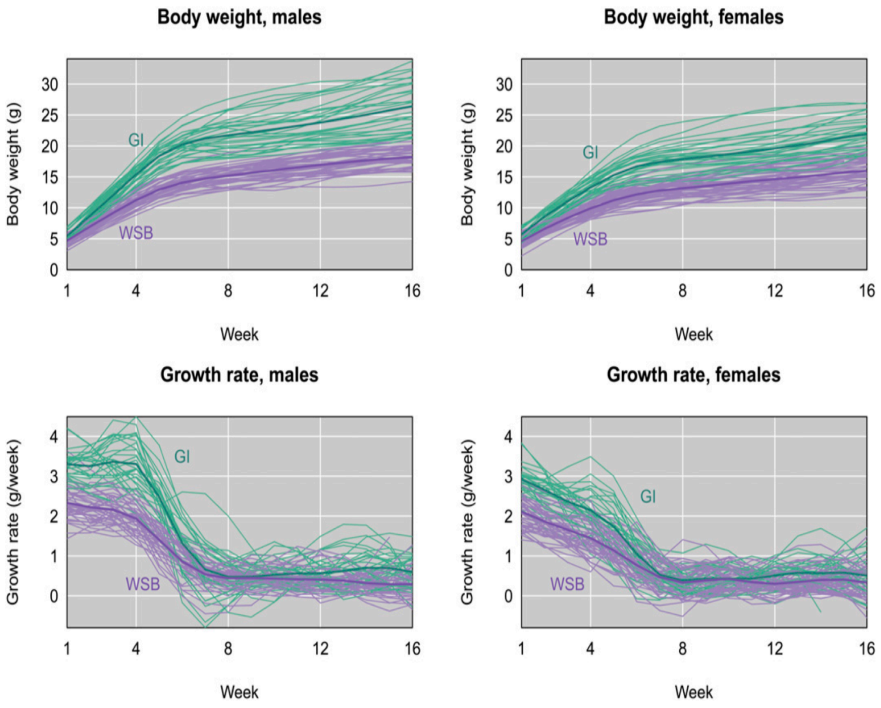


Figure 3 Body weight (top panels) and growth rate (bottom panels) for males (left) and females (right), as a function of age in weeks, for a sample of Gough Island mice (GI, green) and WSB mice (purple) raised in the lab. Individual body weight curves were lightly smoothed using cubic splines; the growth rate curves were estimated as the first derivative of the fitted splines. Thicker curves follow the group averages.

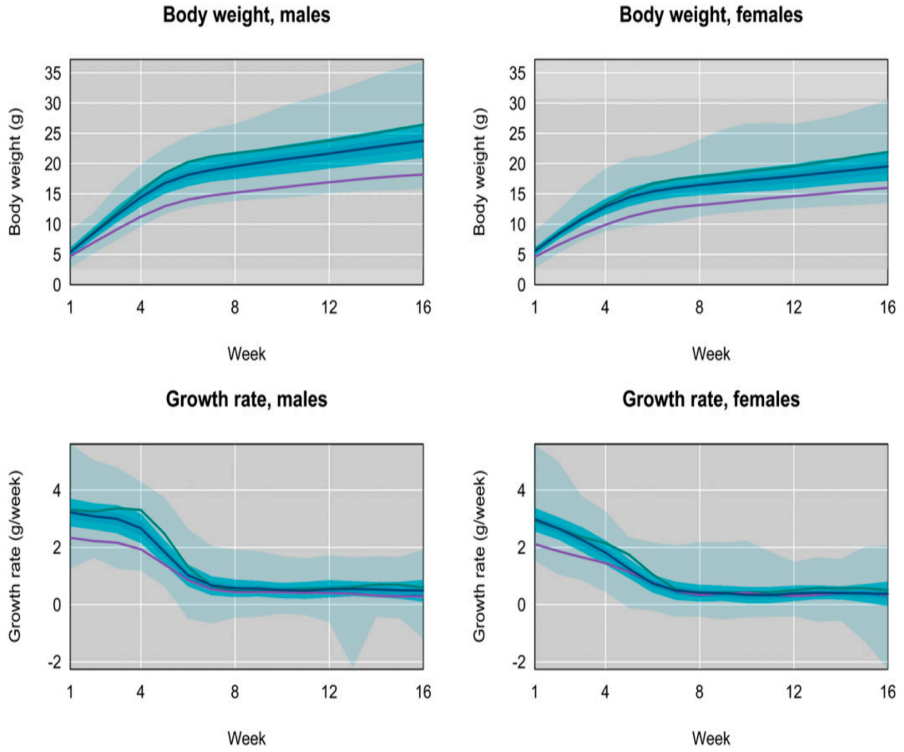
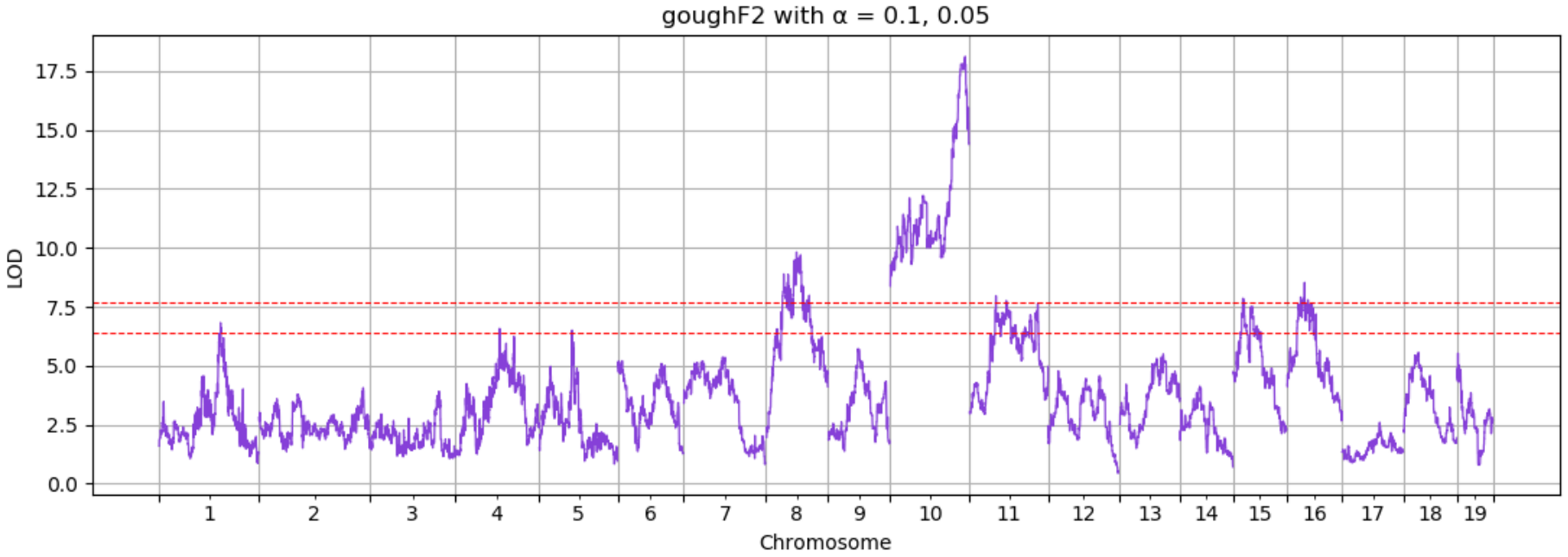


Figure 4 Body weight (top panels) and growth rate (bottom panels) for males (left) and females (right), as a function of age in weeks, for the F_2 mice. In each panel there are three shaded regions; the darkest region covers the middle third of the individuals; the next-darkest two-thirds, and the lightest region all mice. The blue curve is for the average of the F_2 mice. The green and purple curves are for the averages of Gough Island and WSB mice, respectively, as in Figure 3.

MULTIVARIATE 1D genome scan with $\alpha = 0.1, 0.05$ vs. r/qtl

- Z=B-spline(df=4)



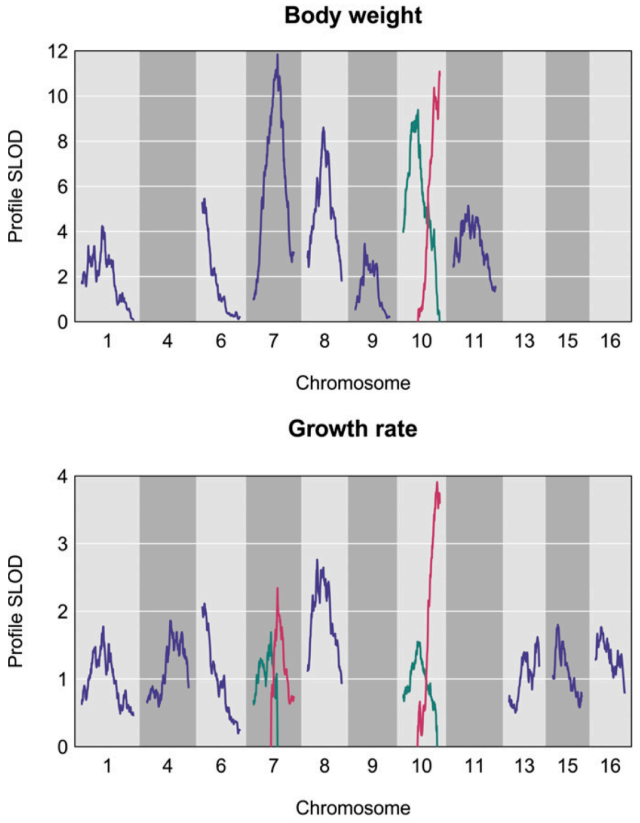
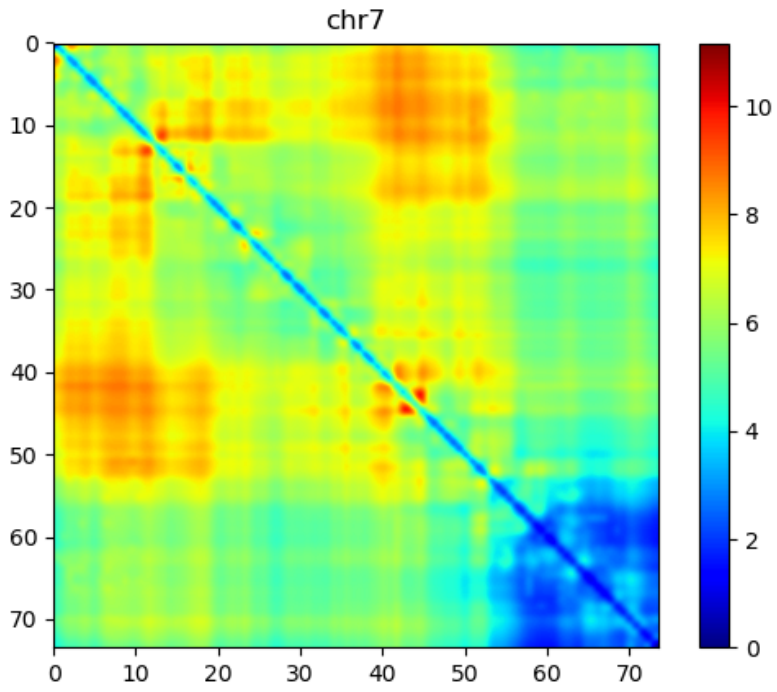
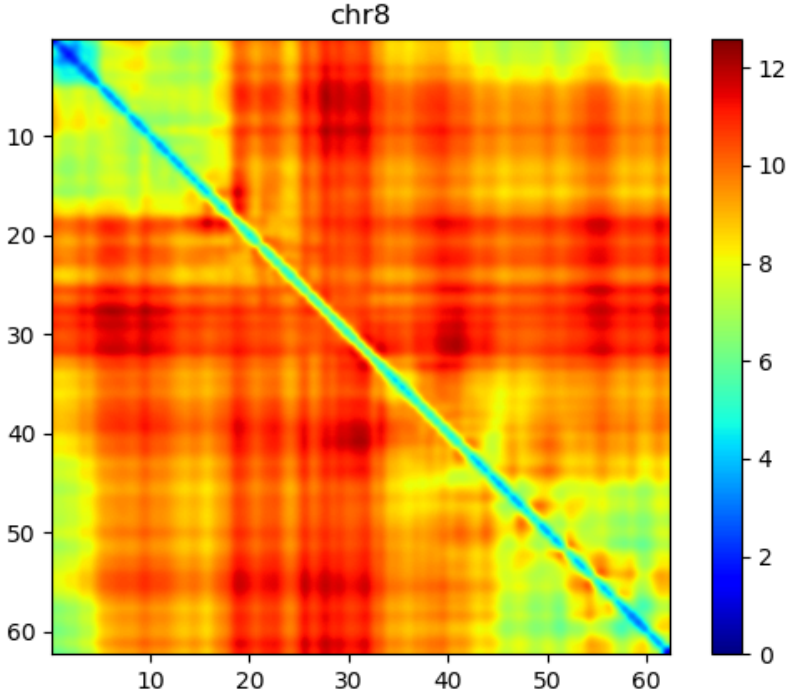
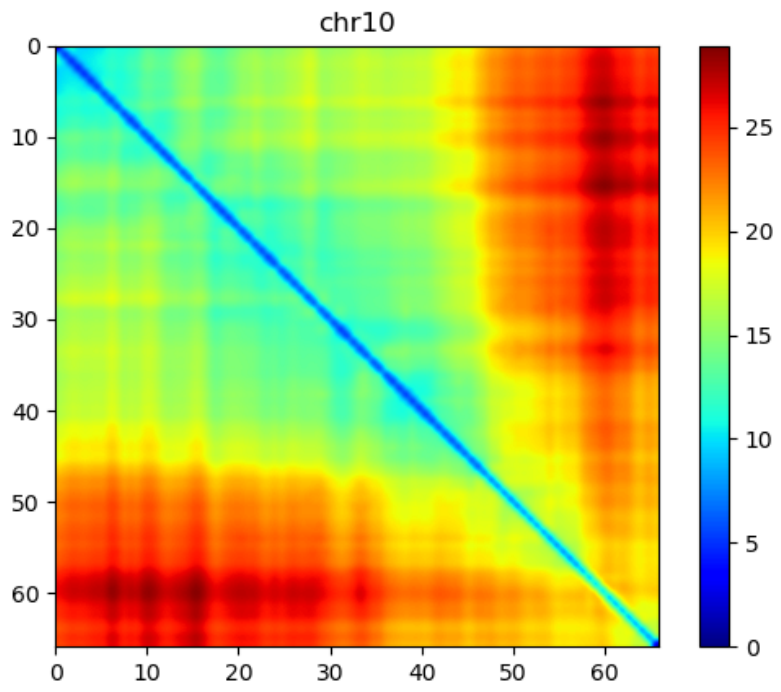


Figure 6 Profile SLOD curves for the selected multiple-QTL models for body weight and growth rate. The location of each QTL was varied, one at a time, with all other QTL fixed at their estimated locations, and the multiple-QTL model was compared to the model with the given QTL omitted.

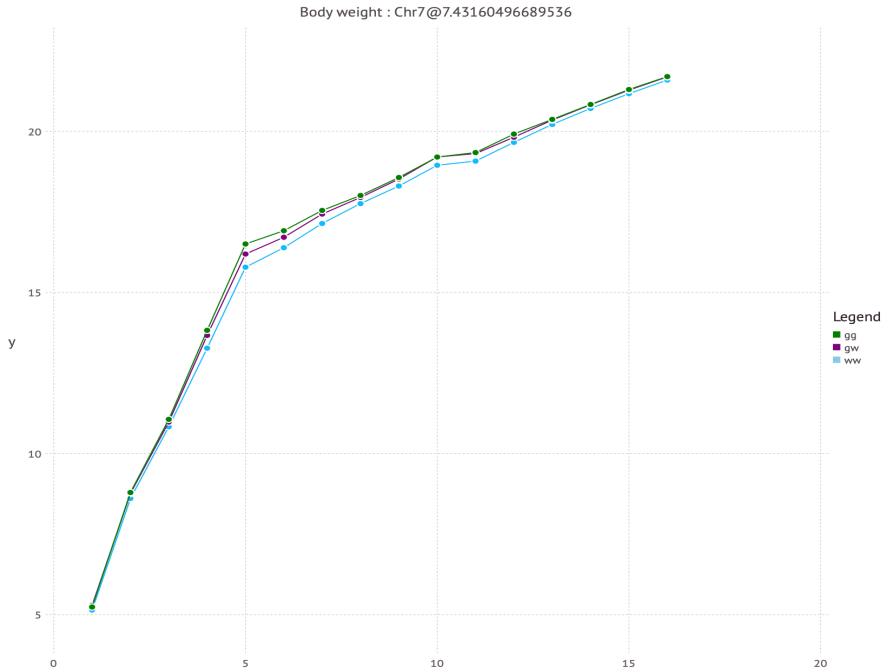
Multivariate (additive) 2D genome scan : Z= B-spline(df=4)

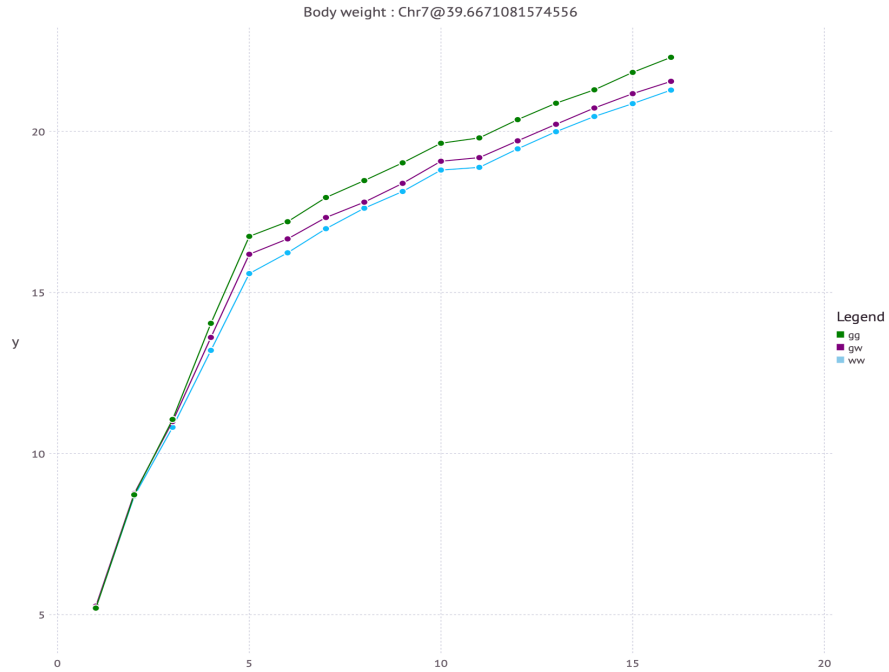


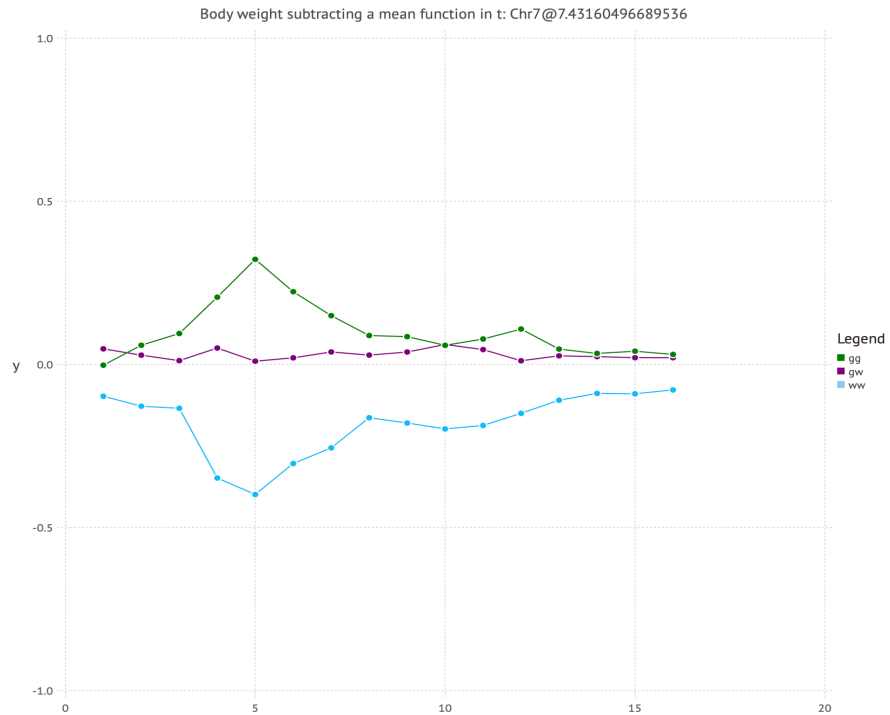


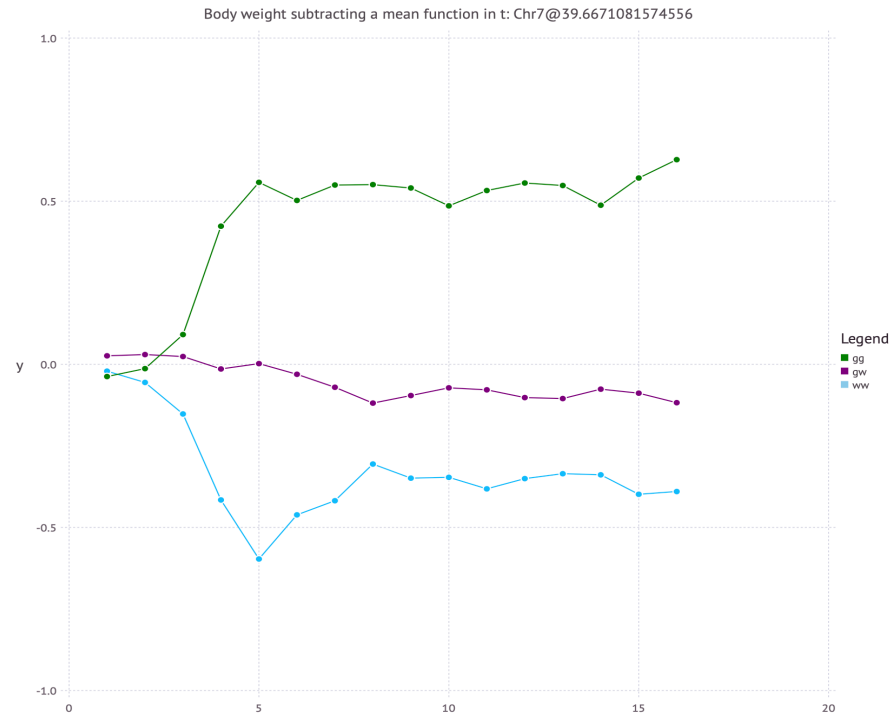


Max Effect Plots @ Chr 7

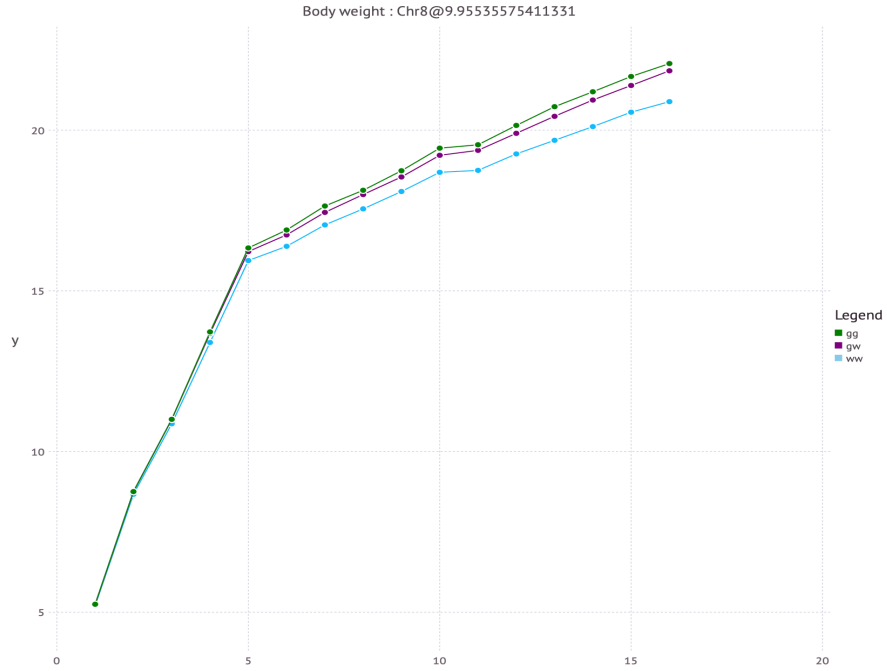


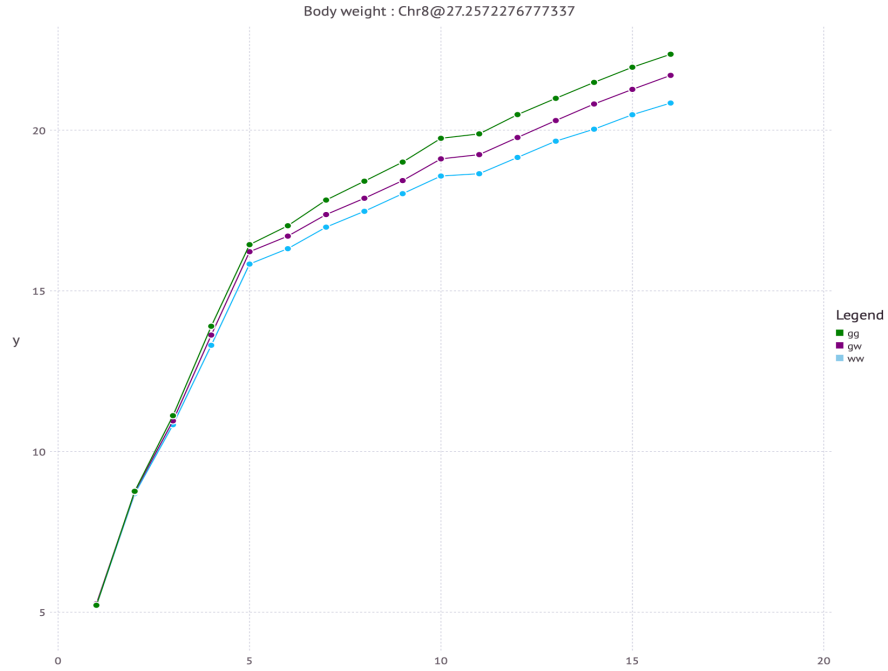


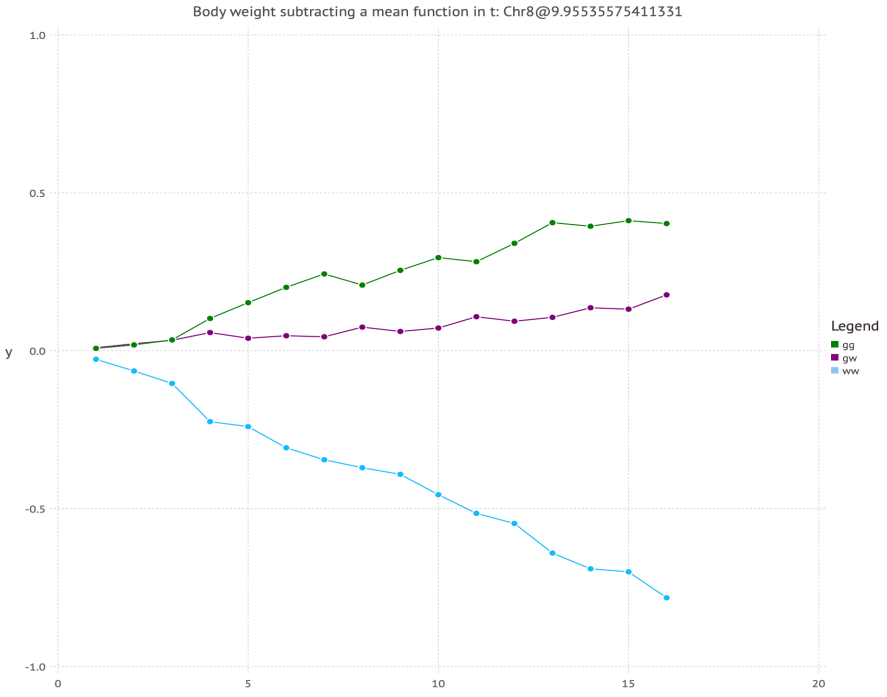


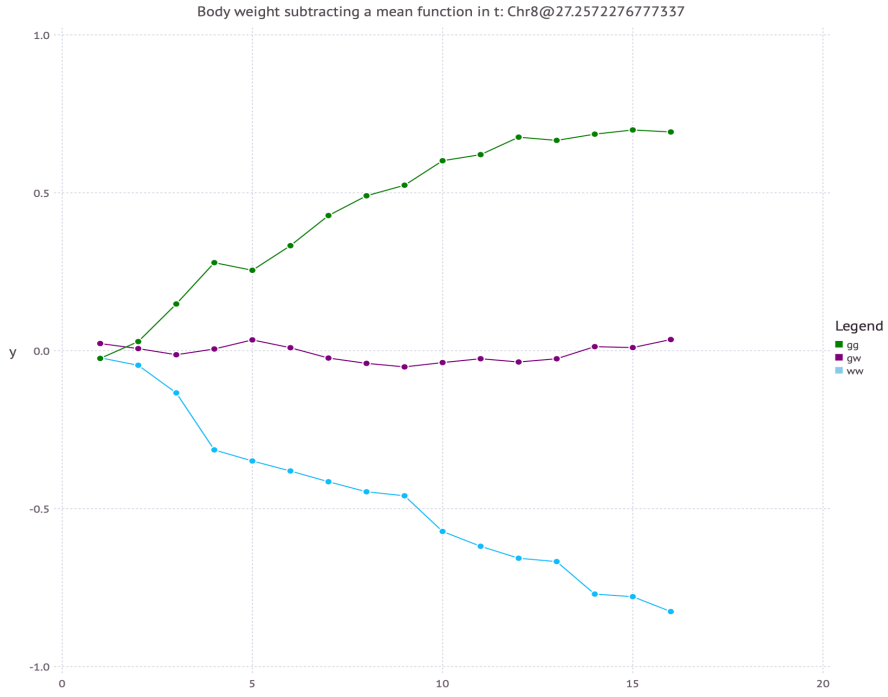


Max Effect Plots @ Chr 8

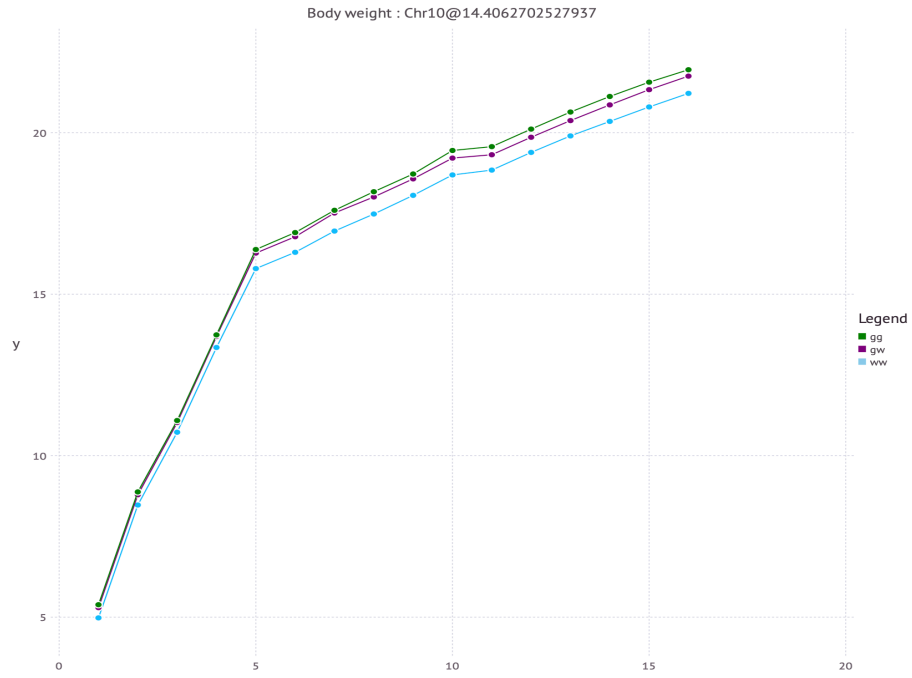


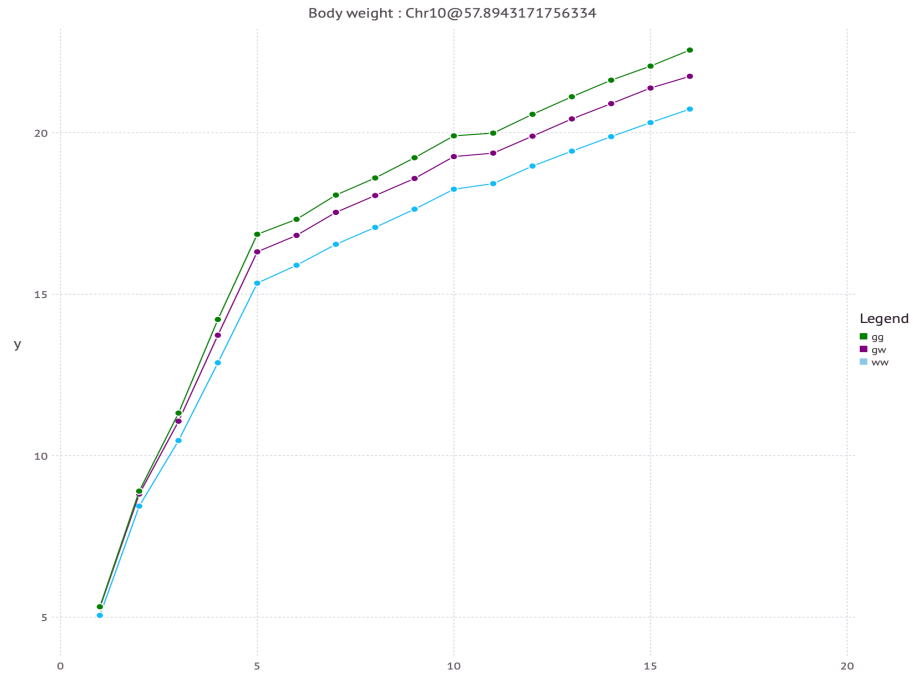


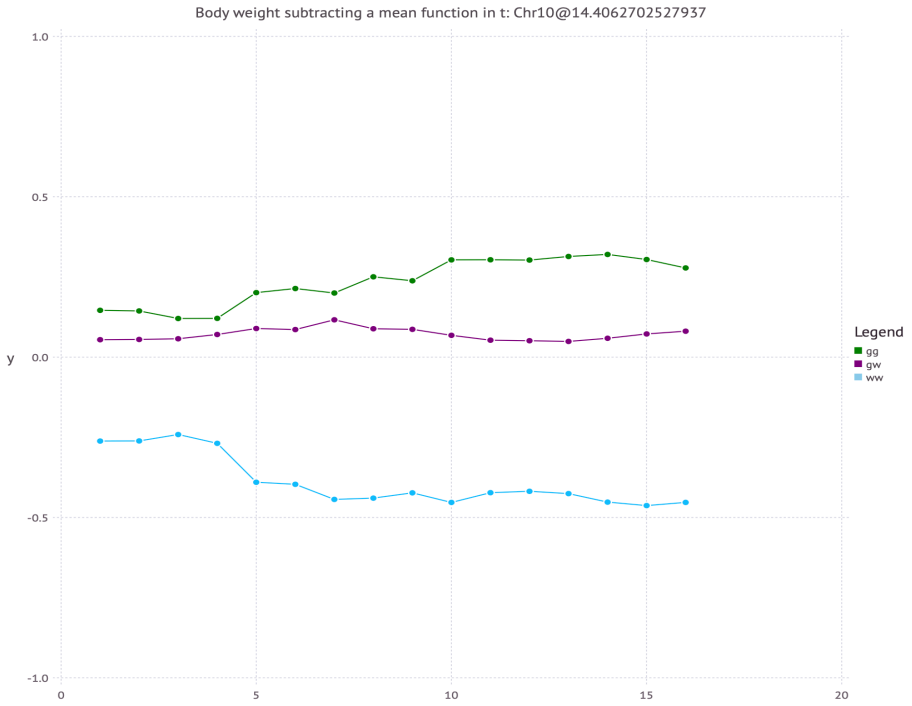


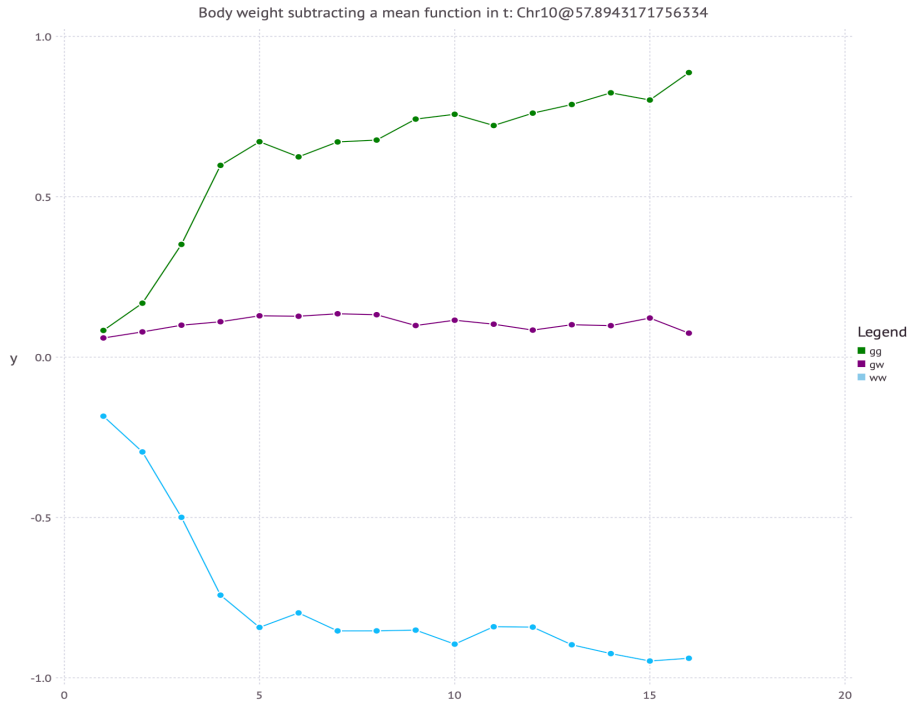


Max Effect Plots @ Chr 10









Summary

- Fast implementation for multivariate LMM handling low- and high- dimensional covariates
- Applications : multivariate LMM for multiple traits, time-valued phenotypic curves (functional data analysis approach)
- Our LMM set up (Z, $\tau^2 K_c$) : dimension reduction in parameter estimation, faster than GEMMA, and approximately the same result as GEMMA's
- Simulation results: relatively insensitive to different K_c's; currently, recommend $K_c = I$ for fast computation
- 1D &2D genome scans with LOCO option, permutation test, stepwise model selection(forward selection/backward elimination) by 1D & 2D scans, scan for environment factors (ongoing!)
- Future research: 3-d array data (multiple location/year or multiple location/trait combination, images, etc.) using tensors, threshold estimation for LOD using diffusion processes

References

[1] W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Advances in Neural Information Processing Systems,* pp. 2510–2518, 2014.

[2] X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* 11(4):407–409, 2014.

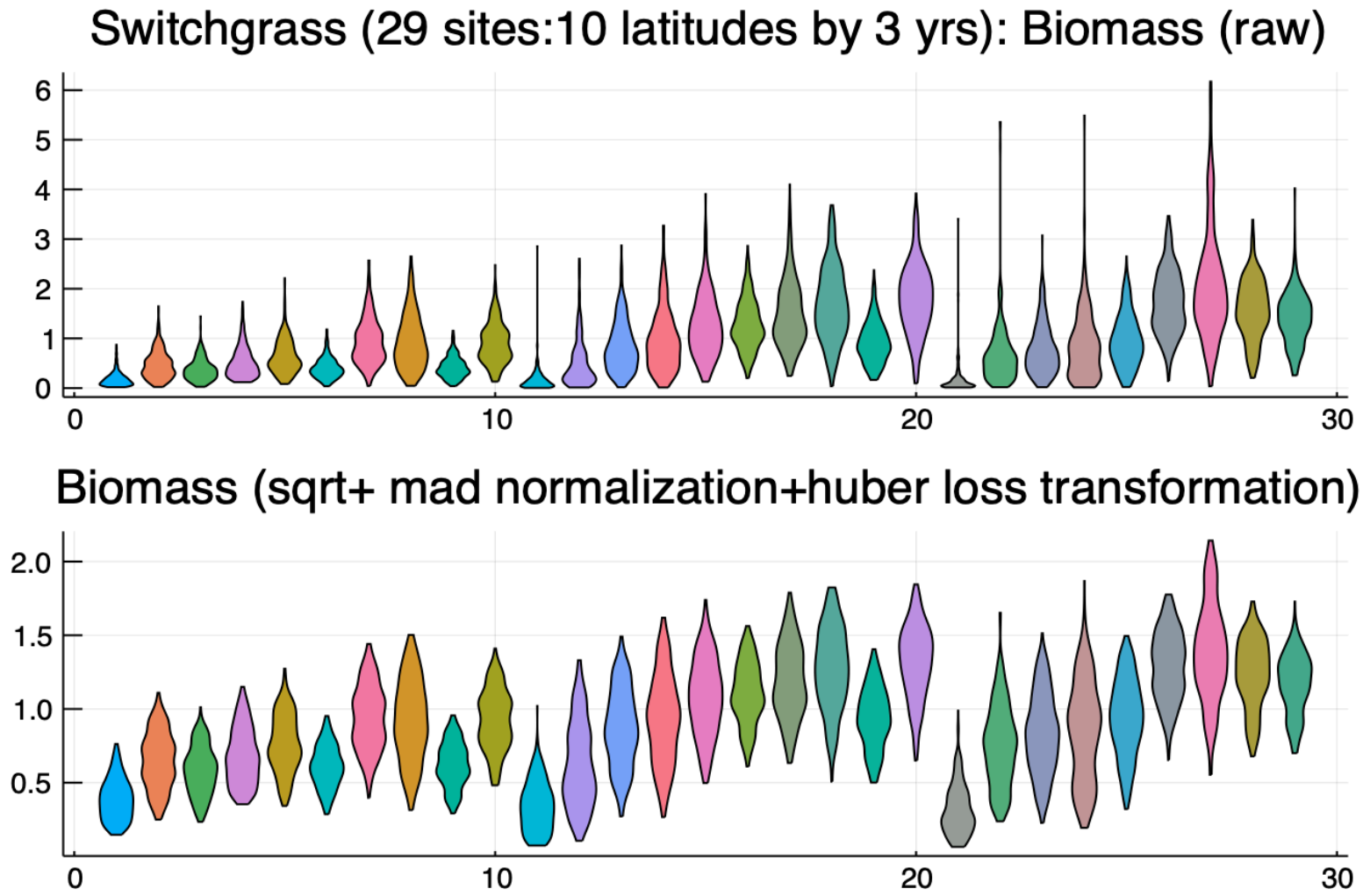
[3] C. R. Henderson. Applications of linear models in animal breeding. *University of Guelph Press*, Guelph 11, 652–653, 1984.

[4] J. Schä¨fer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology,* 4(1), 2005.

[5] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis,* 88(2):365–411, 2004.

[6] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika,* 80(2):267–278, 1993.

[7] A. Manichaikul, J. Y. Moon, S. Sen, B. S. Yandell, and K. W. Broman. A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics,* 181:1077-1086, 2009.


Switchgrass (29 sites:10 latitudes by 3 yrs): Biomass (raw)
6
5
4
3
2
1
0
0
10
20
30
Biomass (sqrt+ mad normalization+huber loss transformation)
2.0
1.5
1.0
0.5
0
10
20
30