

**A re-look at a QTL study for stinging behavior and body size in honey bees using R.
An. Sci. 675 (Statistical Genomics) final project
Michael Wilson, Nov 19th, 2010
Instructor: Arnold Saxton, Data provided by: Greg Hunt**

Abstract:

Data from a study (Hunt, G. J., Guzman-Nova, E., Fondrk, M. K., and Page, R.E. (1998). Quantitative trait loci for honey bee singing behavior and body size. *Genetics* 148: 1203-1213. <http://www.genetics.org/cgi/content/abstract/148/3/1203>) was obtained from Greg Hunt, Purdue University. In R/qtl, a marker map was reconstructed and the data was analyzed to approximate the methods in the original publication which utilized MapQTL, interval mapping, and multiple interval mapping (MQM). Opportunities for alternate analysis were explored. Specifically, MQM and single QTL interval mapping was compared for all phenotypes and a 2-part model was used to handle zero inflated values in one phenotype (number of stings) in a single QTL model.

The Data and QTL *Sting-1*:

The data obtained was not exactly the same as reported in Hunt et al 1998. In the published paper, additional markers were added to the map and an additional phenotype was included. Also, the marker map file was separate from the genotype file and several of the map names did not match and were in different orders. Two populations of bees, a European and an Africanized line were backcrossed. Data from 179 bees from the backcross experiment were represented with 1073 RAPD markers. Each bee is representative of its parent colony, 179 colonies. Four phenotypes measured were indicators of Africanization, which include flightiness, a tendency to sting score, the number of stings in a leather patch, and forewing length. Twenty eight linkage groups were represented in the map.

QTL *Sting-1*

Since Hunt et al 1998 was published, a QTL discovered in that experiment, *Sting-1* (on linkage group 4, position ~300cM), has become well understood to be a QTL influencing stinging behavior (Hunt et al 2007). We should therefore be able to assume that a signal from that region in this analysis is a real biological signal and this may be useful in comparing analytical methods. Other suggestive QTL in the original study which were later confirmed are *Sting-2* and *Sting-3*.

Make a Marker Map in R/qtl:

The marker map, which was originally in a separate file, was copied over to the gene file in a spreadsheet. Since many of the marker names did not match, or were missing, or out of order, making a new map from scratch was explored. As explained by Karl Browman (2010) to make a new map, use the following steps and functions.

1. Estimate inter-marker recombination fractions
 - `est.rf()`
2. Divide markers into linkage groups
 - `formLinkageGroups()`
3. Get an initial ordering of the markers within the linkage groups
 - `orderMarkers()`

I was able to form the linkage group and make an initial order of the markers using the old, provided marker map, but many positions were guessed, so Step 2 above was attempted. However, I exceeded my memory limit quickly. To increase memory limit in R, see function 'memory.limit'. Making a map completely from scratch was not necessary, since the provided map gave a good start to form linkage groups and marker orders, so I picked up the process below. Also, other data checking (not included in this report) was done following Broman and Sen (2009) and , see supplementary R file additional steps taken.

4. Estimate inter-marker recombination fractions (If not already done)
 - est.rf()
5. Use a plot tool to see if some markers might need to be moved to another linkage group
 - plot.rf()
6. Move any markers necessary
 - movemarker()
7. Re-plot until linkage groups indicator lines line up with more obvious linkage groups, see figure 1.

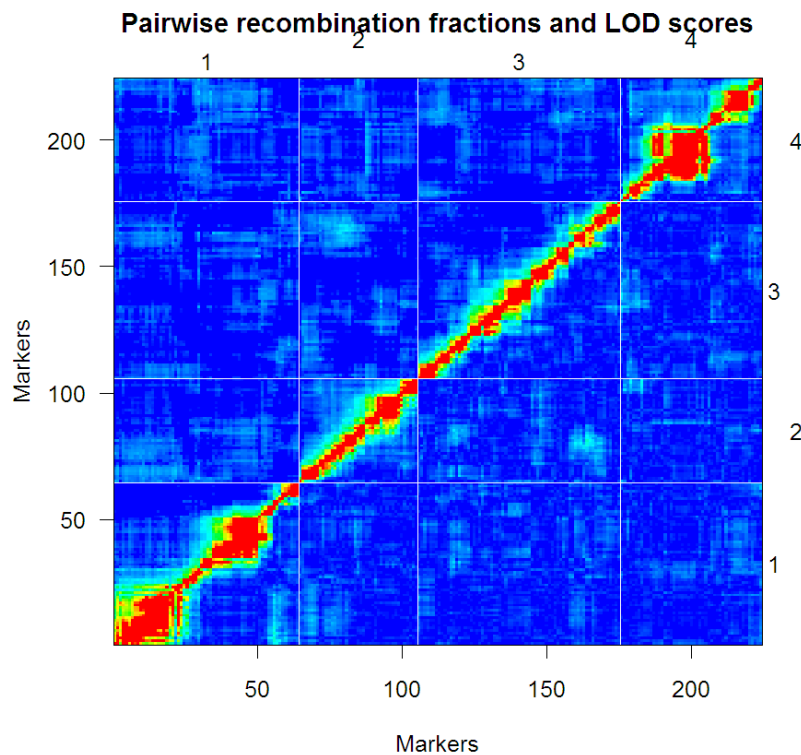


Figure 1- Plot of recombination fractions for first 4 linkage groups after moving some bordering markers into the correct linkage group. Once the markers are in the correct linkage group, we can order the markers within the linkage group. Note the absence of recombination between linkage groups. This was consistent across the entire data set.

8. Explore alternate orders for the markers, within the linkage group.
 - `Ripple()`
 - Since there are $n!/2$ possible orders of markers per linkage group, when $n = \#$ of markers, ripple only considers the order within a set window and then moves to the next window to find an order. Otherwise, computing times would be excessive. Two methods can be used to determine the best order within the window. First use counting crossovers.
 - Counting crossovers: `method=c("countxo")` This is a quick calculation, but not the best option. Use it first then move on to the next option. Although quick, the window size must still be small due to memory issues.
9. Switch to a better marker order as applicable. Fewer crossovers are better.
 - `switch.order()`
10. Repeat exploration of alternate orders for the markers, but this time, use maximum likelihood
 - `ripple()`
 - Maximum likelihood using option: `method=c("likelihood")` This gives a LOD score. A general cut-off of 2 can be used to choose the best order. The length of the 'chromosome' is also given so if in doubt, check to make sure you are not choosing an order that creates a longer chromosome. Computing times can be long, ~3 hrs for the biggest linkage group in this study with `window=4`.
11. Repeat `ripple()` and `Switch.order()` for each linkage group.
12. Use the plot tool again to make sure markers distant from each other (in a linkage group) are not showing recombination.
 - `plot.rf()`
13. Once a satisfactory linkage grouping and order is achieved, its time to estimate the distances between the markers.
 - `est.map()`
14. Compare these newly estimated distances to the previous distances.
 - `plot(newmap, crossobject)`
15. Use the new map, if applicable, which it certainly was in this case.
 - `replace.map()`

Identifying QTL:

The 1998 publication used single QTL interval mapping for flying and stinging phenotypes and multiple QTL interval mapping (MQM) for the number of stings in a leather patch and forewing length. The lack of a normal distribution in all but wing length was a concern identified in the paper. Particularly when there is a spike in the phenotype distribution (as in the case of number of stings), spurious LOD peaks can occur in areas lacking genotype information (Broman 2003). To address this, they used non-parametric methods (Kruskal-Wallis rank test) for the two single QTL interval mappings. In R, I first conducted Kruskal-Wallis rank test on the three non-normal phenotypes, then used a parametric single QTL model for wing length. I then ran a 2-part normal model on number of stings. In the 1998 paper, for the number of stings was analyzed with MQM and permutation testing was conducted to get a LOD threshold which is applicable to non-normal data (Churchill and Doerge 1994). I conducted MQM on all four phenotypes for comparison purposes.

Single QTL Interval Mapping

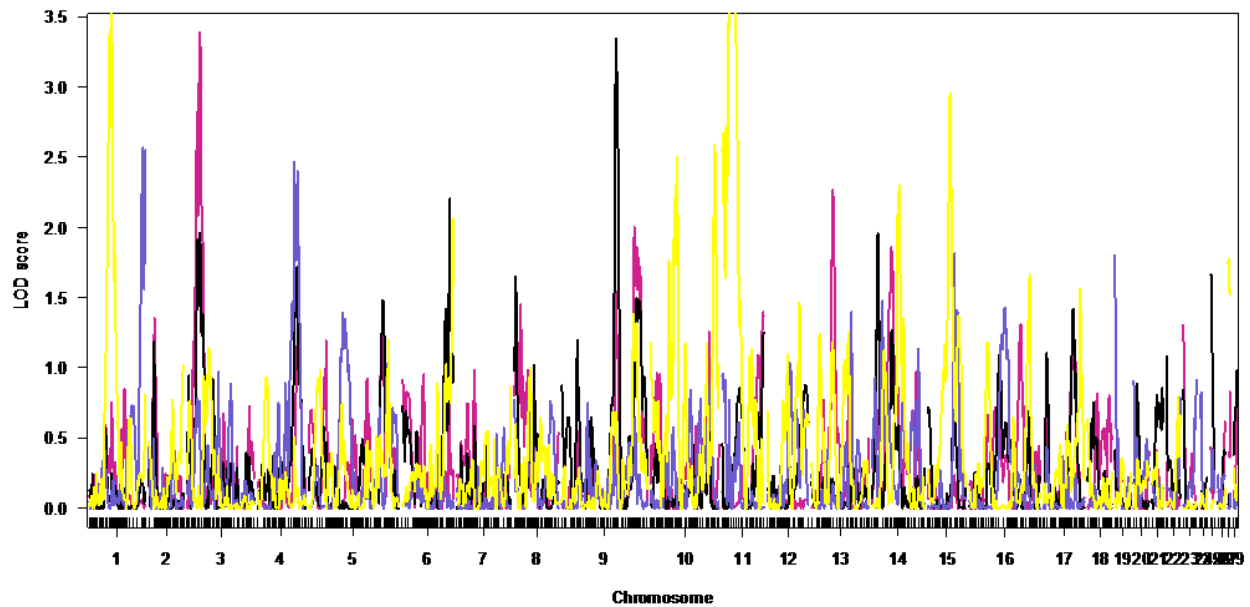


Fig2. - Plot of single interval mapping using function: scanone

LOD thresholds with 1000 permutations single QTL scanning				
	Flying	Stinging	# stings	Wing length
Plot color	Violetred	Black	Slateblue	Yellow
5%	3.37	3.23	3.25	3.47
10%	3.04	2.96	3.03	3.09

Markers above LOD 2.0

(markers within 90% confidence highlighted)

```
--Flying --
chr   pos   lod
Z8_1_11    3     90.4  3.39 (QTL Sting-2)
a64_169_S_ 13    153.1  2.27

--Stinging--
chr   pos   lod
_320_93f  6     347    2.21
c9.loc388  9     388    3.35

--Number Stings--
chr   pos   lod
c1.loc385  1     385    2.57
a43_145_S_ 4     266    2.46 (QTL Sting-1)

--Wing Length--
chr   pos   lod
AJ509344_2 1     159    3.57
a23_119    6     365    2.07
a17_142_S_ 10    462    2.59
c11.loc81  11    81     5.38
c14.loc196 14    196    2.30
a73_217_S_ 15    161    2.95
```

Log Normalization:

The 1998 paper did not mention if the numbers were log transformed or not for the data used in MQM. For this project, log normalization was attempted for the three non-normal phenotypes, but it did not help the data when testing with Shapiro's normality test using function:

- `mqmtestnormal`.

For the number of stings in a leather patch, we can see why in the following graphs. The zero values are heavily inflated. Zero inflation was not as bad in the other 2 phenotypes. Since taking the log did not help any of the 3 problematic distributions, I used the real numbers for all analysis, except for the 2part analysis.

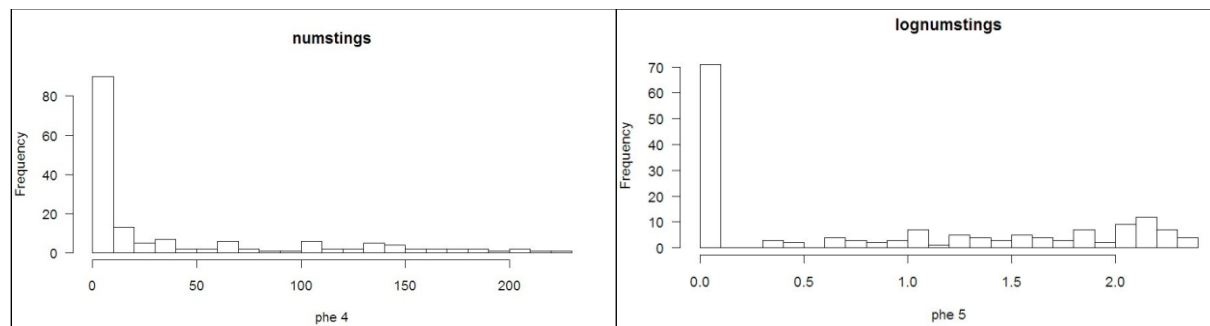


Fig. 3 – shows number of stings before and after log normalization, with zero inflated values.

2-Part Interval Mapping:

R/QTL has an interval mapping function designed specifically for phenotypes with an inflated or spiked value, in this case at zero. One strategy to analyze zero inflated data like this is to analyze it in two separate models. First, a model would be tested on a binary trait of whether or not the bees would sting (compare observations with zero stings to observations with stings >0). Then, for those individuals that do sting, a model could be used to analyze a quantitative trait of the number of stings. The 2-part method in R/QTL combines these two models into a single model (Broman 2003). First the null hypothesis is considered as $\pi_j \equiv \pi$, while μ_j varies to consider QTL that influence stings 0 or stings >0 (output is lod.p). Then, the null hypothesis is considered where $\mu_j = \mu$ while π_j varies to generate LODs for the quantitative trait when stings >0 (output is lod.mu), (see Broman and Sen (2009) p. 142 and R help files) when π_0 is the proportion of individuals with a positive phenotype and μ is the sample mean among individuals with the positive phenotype (Broman 2003). The output lod.p.mu is simply the sum of the two LOD scores. To call the 2 part model, specify it in the 'model' option of scanone.

- `Scan2parts <- scanone(beeqtl, pheno.col=5, model="2part", upper=FALSE)`

Broman (2003) identifies scenarios where the 2-part model is advantageous over non-parametric analysis. However, it is often not due to a loss in power. In the plot in Fig. 4., we see some additional spikes compared to the non-parametric analysis. There is a clear distinction between the two parts of the model at the marker for Sting-1 and the total LOD is slightly higher than in the non-parametric analysis. However, again none of the spikes LOD scores are within the 90% confidence interval.

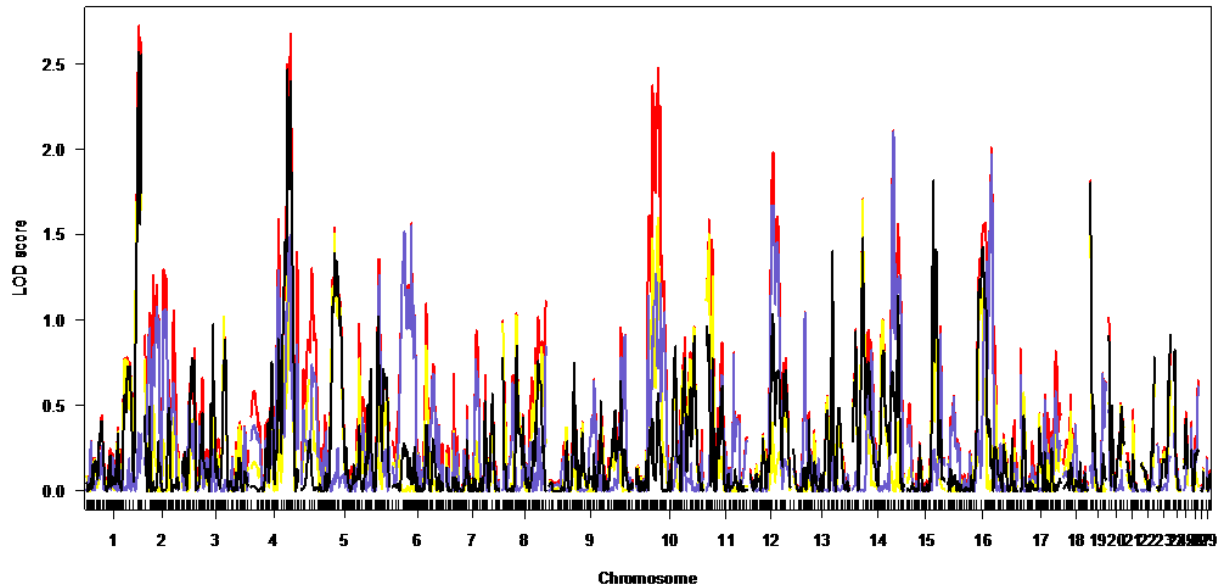


Fig4 – 2part scan plot in 3 colors. Black is non-parametric analysis plot.

Markers above LOD 2.0

(No markers in 90% confidence, see plot color key below. Black is the previous non-parametric analysis.)

--# Stings --	Chr	pos	lod.p.mu	lod.p	lod.mu
c1.loc385	1	385	2.72	2.38	0.287
N4_27_S_	4	292	2.68	1.31	1.363 (QTL <i>Sting-1</i>)
_361ND34	10	164	2.47	1.59	0.882
c14.loc303	14	303	2.11	<0.01	2.103
c16.loc320	16	320	2.01	0.06	1.970

MQM in R

To conduct MQM, first missing genotypes are augmented with multiple genotypes and their estimated probabilities.

- `augmentedcross <- mqmaugment(beeqtl, minprob=0.1)`

Then, important markers are chosen as cofactors. Two options are available in R/qtl, I chose unsupervised cofactor selection through backward elimination, which accounts for marker density.

- `autocofactors <- mqmautocofactors(augmentedcross)`

Genetic map

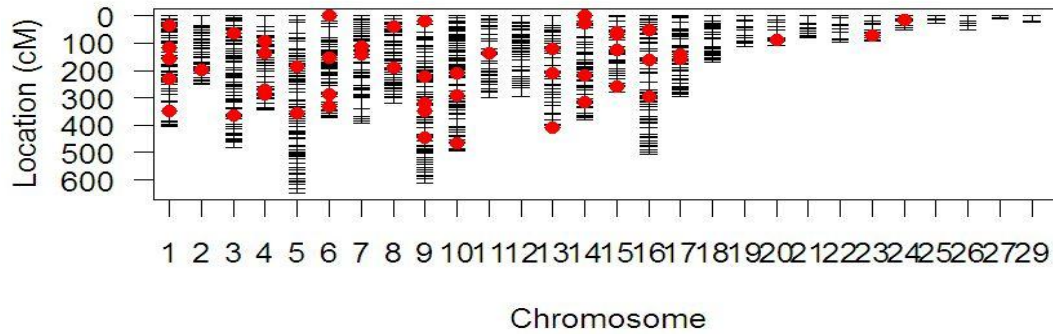


Fig.5 – Cofactors chosen are indicated with red dots

MQM scan can then be run on individual phenotypes. See plots in figure 6.

- `mqmscan()`

MQM assumes a normal distribution, which the first three phenotypes are not. However, a LOD threshold generated by permutation data should be applicable to non-normal data (Churchill and Doerge 1994). Compared to the single QTL model and the 2-part model, we can see a drastic increase in the LOD score for QTL *Sting-1* in the #stings. This is the only model tested where *Sting-1* on the #stings was statistically significant (LOD 4.3).

With the flying phenotype, Hunt et al (1998) found that QTL *Sting-1* was significant with a single QTL model, but at a marker that is not included in the data set analyzed here (stsN4-.245). With this data set, I did not find *Sting-1* as a significant QTL with the flying phenotype with either MQM or scanone. Information around the marker not in this dataset could

explain the different results. With this data set, MQM did more clearly recognize *Sting-1* with the flying phenotype, but with 1000 permutations, now none of the markers indicate a significant QTL. With the single QTL model, a significant QTL was found on linkage group 3.

Markers above LOD 2.0

(1000 permutation confidence intervals)

--Flying--	chr	pos	LOD
(LOD threshold 5%=3.95 10%=3.53)			
c3.loc90	3	90	2.65(QTL <i>Sting-2</i>)
c4.loc300	4	300	2.25 (QTL <i>Sting-1</i>)
c13.loc155	13	155	2.09

--Stinging--

(permutations not run)

c6.loc345	6	345	2.58
c9.loc390	9	390	2.20

--#Stings

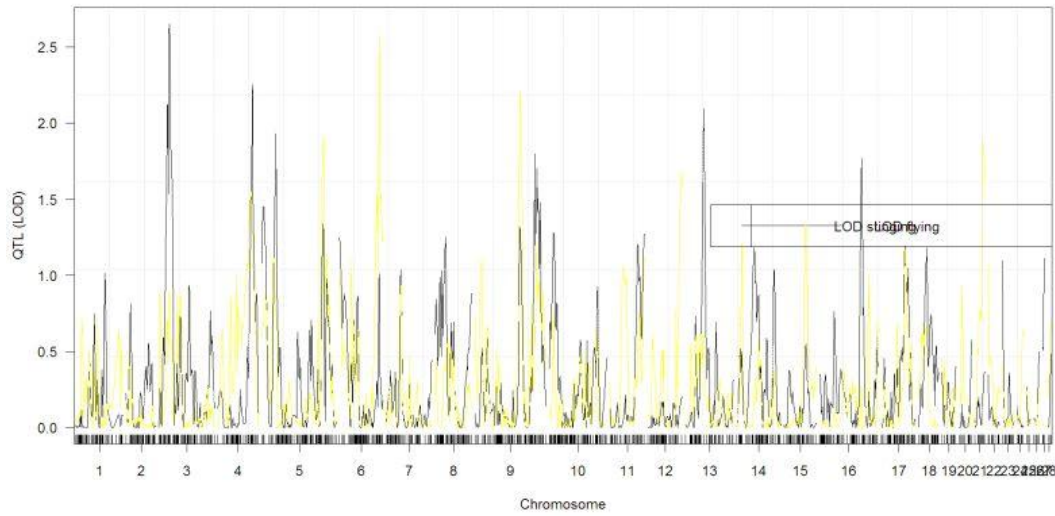
(LOD threshold 5%=3.35 10%=3.06)

c4.loc275	4	275	4.3 (QTL <i>Sting-1</i>)
-----------	---	-----	---------------------------

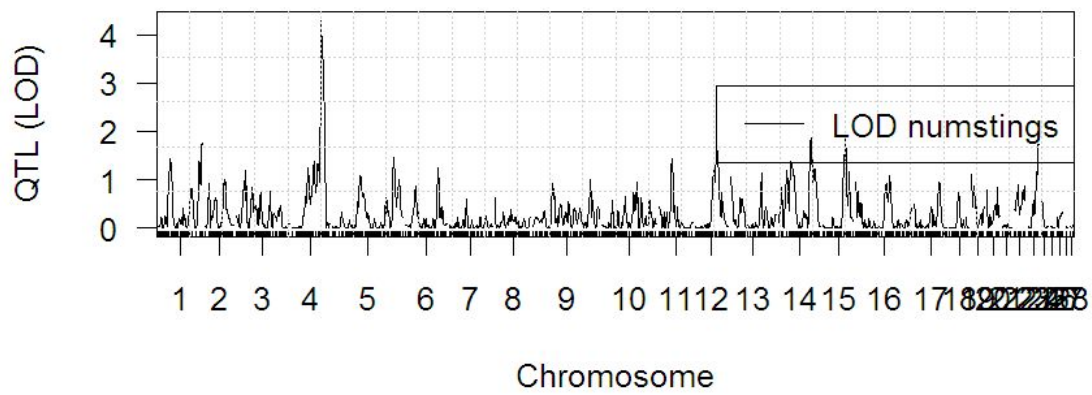
--Wing length--

(permutations not run)

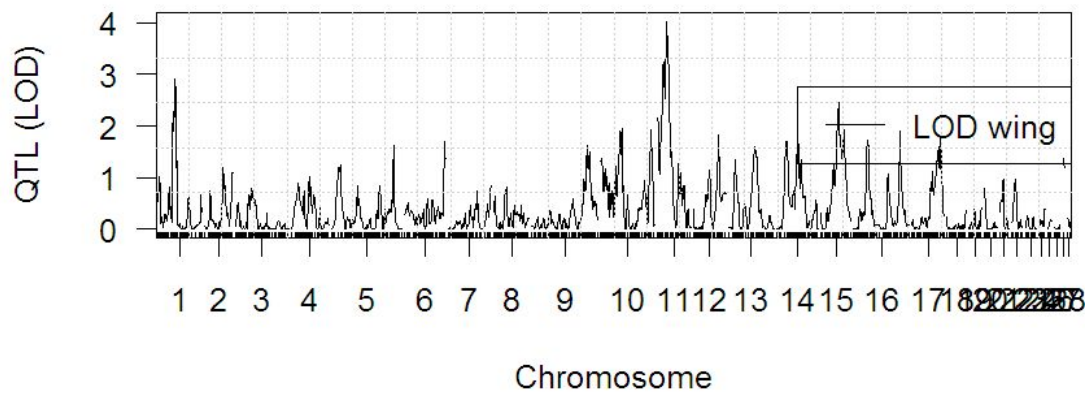
c1.loc160	1	160	2.90
c11.loc80	11	80	4.03
c15.loc170	15	170	2.45



MQM Plot results for Flying and Stinging rating



MQM Plot results for the Number of Stings in a Patch



MQM Plot results for Wing length

Fig. 6 – MQM plots from the 4 phenotypes

References:

Arends D., Prins, P., Broman, K. W., and Jansen, R. C. 2010. Tutorial – Multiple-QTL Mapping (MQM) Analysis. Online publication. <http://www.rqtl.org/tutorials/MQM-tour.pdf> accessed November 19th, 2010.

Broman, K. 2010. Google groups discussion board. R/qtl discussion. http://groups.google.com/group/rqtl-disc/browse_thread/thread/22a9c08b52465349?pli=1 Accessed: Nov 10, 2010.

Broman, K. W. 2003. Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genetics* 163: 1169-1175.

Broman and Sen. 2009. *A Guide to QTL Mapping with R/qtl*. Springer Dordrecht Heidelberg London.

Churchill, G. A. and Doerge, R. W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138: 963-971.

Hunt, G. et al. 2007. Behavioral genomics of honeybee foraging and nest defense. *Naturwissenschaften* 94: 247-267.

Hunt, G. J., Guzman-Nova, E., Fondrk, M. K., and Page, R.E. 1998. Quantitative trait loci for honey bee singing behavior and body size. *Genetics* 148: 1203-1213. <http://www.genetics.org/cgi/content/abstract/148/3/1203>