1 Online Appendix A

1.1 Description of Simulation and ABC

At the start of each simulation, flies were distributed across all patches at random, and allowed to assort themselves for 100 fly moves (patch joining-or-leaving events). We then allowed the simulation to run for 200$\times$(n-samples) fly moves and recorded the state at each event. The time-points at

which we sampled within each simulation were chosen to match the empirical data collection (from 10-20 samples). Samples were collected at equally spaced intervals relative to the internal simulation clock against which movement rates were calculated. For each sample set, we calculated summary statistics S' exactly as we did the empirical summary statistics S in S&F. At each step of the ABC process $X(\cdot)$, we propose new parameter values Θ' and then measure similarity between $\mathcal{D}'$, data simulated with the current set of parameter values, and the observed data, $\mathcal{D}$. We do this via the summary statistics S' (S), given earlier. This forms part of a determination of whether to accept the new Θ' or revert to the last accepted Θ. Once the algorithm has reached *stationarity* (*i.e.*, the sequence of Θ-values is no longer influenced by the point from which it started), the samples are from the posterior distribution $f(\Theta \mid S)$, an approximation to the distribution of interest $f(\Theta \mid \mathcal{D})$ (Marjoram et al., 2003). All S' were normalised by subtracting the mean, and dividing by the standard deviation, of their respective empirical S to place all statistics on the same scale.

1.2 Summary statistics and model fit

Summary statistics for each genotype$\times$density treatment were calculated across a series of time points within each simulation. All summary statistics were standardized by subtracting the empirical mean value for that statistic, and dividing by the empirical standard deviation. This was done to help ensure that no single statistic dominated the calculation of the distance metric. The distance measure, d, between observed and simulated data was calculated as the sum of squared deviations of each summary statistic S_i' from the empirical statistic S_i (Euclidean distance), where 0 indicates a perfect fit. We derived 11 empirical fit statistics S_j from the data of S&F. We used the five descriptive statistics reported in S&F, plus an additional 4 individual-group size descriptors. To account for variation between samples, we fit the the standard deviation among samples for 2 of our metrics, mAv and fAv.

The following j group metrics M_j were calculated for each time point:

mAv, fAv— Male and female average group size was calculated as the average number of individuals of sex s on each p of P patches .

$$\frac{1}{P}\sum_{p=1}^{P} n_{sp} \tag{2}$$

mVar, fVar— Male and female variance was calculated as the variance between patches p of flies of

sex s.

$$\frac{1}{P}\sum_{p=1}^{P}(n_{sp}-\bar{n}_s)^2 \tag{3}$$

mfCov— The covariance between males m and females f, was calculated as the covariance between males and females across patches p.

$$\frac{1}{P}\sum_{p=1}^{P}(n_{mp}-\bar{n}_m)(n_{fp}-\bar{n}_f) \tag{4}$$

mPerM, fPerM, mPerF, fPerF— The so-called individual-group size metrics (Jovani and Mavor, 2011) are the mean number of (other) individuals of sex q on a patch with each individual of sex s. Let Q_p be the number of non-self individuals of sex q on patch p. That is, if q=s, $Q_p = n_{sp} - 1$; while if $q \neq s$, $Q_p = n_{qp}$. We then define these parameters as:

$$(\sum_{p=1}^{P} n_{sp})^{-1}\sum_{p=1}^{P} n_{sp}Q_p \tag{5}$$

The average of the M_j group level metrics was calculated across each timepoint t, for the N samples of treatment k, as:

$$\frac{1}{N_k}\sum_{t=1}^{N_k} M_{tjk} \tag{6}$$

The standard deviation of the M_j group level metrics was calculated across each timepoint t, for the N samples of treatment k

$$\frac{1}{N_k-1}\sum_{t=1}^{N_k}(M_{tjk}-\bar{M}_{jk})^2 \tag{7}$$

The set of summary statistics S_{ik} comprised the mean values of M_{jk}, as well as the standard deviations of mAv and fAv,

The statistics mVar, fVar, mPerM, fPerM, mPerF, and fPerF were log transformed. For each of these statistics, a nominal amount of 0.1 was added to the untransformed statistic to avoid taking the log

of 0. The statistic mfCovar was log transformed as well, after adding a nominal value of 2.2. For values of mfCovar lower than -2.2, representing a floor value which was only surpassed a single time in our empirical data, a transformed value of -8.8 was assigned. In regression analyses of the empirical data, using a floor value of -8.8 resulted in the most normally distributed residuals. No transformation was necessary for mAv or fAv, or for the standard deviations of these two statistics.

1.3 Acceptance threshold

It is important in ABC to have a low tolerance, ϵ. As ϵ goes up, the rejection rate decreases, and estimates of Θ will increasingly resemble the prior. We employed a two-part ABC process to balance the need to efficiently explore a large state space (7 dimensional) with the need to rigorously characterize the posterior. The first step in the process was a preliminary threshold estimation to establish an appropriate ϵ. Then we estimated Θ using the ϵ determined in the previous step.

In the first step we conducted ABC with a fluctuating ϵ which tended towards lower values as the MCMC explored regions of better model fit. This process was analogous to simulated annealing, or the ABC scheme of (Bortot et al., 2007), and allowed the algorithm to "cool" to efficient levels of model fit without becoming stuck in local minima. We ran 100 chains of the annealing process in parallel for each treatment for an arbitrary time period (1 hour). From these chains, we chose a target value of ϵ equal to the lowest 5th percentile for each treatment independently. Next we ran our final ABC analysis, in which, after a burn-in period, ϵ was fixed at the value determined in the previous step. Only after fixing the threshold did we begin to record Θ. The fixed threshold values varied from 0.39 to 1.15 among treatments. The acceptance rates varied from 0.06 to 0.15 across treatments in this fixed ϵ, posterior estimation phase of the ABC analysis.

1.4 Stationarity diagnostics and autocorrelation

We used the autocorrelation diagnostic acf{stats} in R to establish a subsample rate for each Markov Chain. We ran 30 parallel MCMC chains, with separate random number seeds for each treatment, in order to maximize computational speed on the University of Southern California high performance computer cluster. Each chain was run on average for 7.8 million iterations. MCMC convergence

among chains was assessed using the Gelman-Rubin test via the gelman.diag{coda} function in R. Briefly, this test calculates the ratio of between-run variation to within-run variation—a Potential Scale Reduction Factor (PSRF). A value close to 1 means that all the variation is due to within-run variation. We tested all replicates in the full dataset, and the two models mfHigh and mfLow. For the full dataset, values of the Multivariate PSRF (MPSRF), which considers each of the 7 parameters simultaneously, ranged from 1.02 to 1.81, with an average value of 1.15 (sd=0.16). This reflects the difficulty of mixing effectively between the mfHigh and mfLow models. When calculated separately for the two models, the MPSRF was much lower. For mfHigh the values of MPSRF varied between 1.02 and 1.18, with a mean of 1.08 (sd=0.04). For mfLow, MPSRF varied between 1.02, and 1.28 with a mean of 1.07 (sd=0.06). We did not calculate one MPSRF score in mfLow because the sample size was too small (all chains had fewer than 500 samples). An additional 2 of the mfLow MPSRF scores were above 1.2, also a result of very small sample size.

1.5 Multimodality tests

We tested for multiple peaks (peaksIDPmisc in R), with PHmin set at $1/10$ the maximum peak height (qualitatively similar results were found for all values of PHmin tested). We found significant multimodality, in particular for parameters a_{fm}, a_{mf} and l_f (Fig. S2). Across all treatments, the number of times a parameter was bimodal (there were no trimodal peaks) was 0 for j_m, 15 for l_f, 1 for j_f, 0 for a_{mm}, 18 for a_{mf} , 23 for a_{fm}, and 0 for a_{ff}, for a total of 57 bimodal marginal posteriors. Because the bimodality in a_{mf} and a_{fm} was reflected around the 1:1 axis (Figure S1), we split the posterior of each treatment into an mfHigh and mfLow posterior, as described in the main text. In effect, we are evaluating our data under 2 different priors, where the joint prior of a_{mf}/a_{fm} is constrained to be above or below 1 for mfHigh and mfLow respectively. All other priors remain flat and uniform within their ranges. Upon testing the resulting posteriors of mfHigh and mfLow, there is no further evidence of bi-modality in any parameter in any treatment (assessed using peaksIDPisc in R). This is evidence that we have successfully resolved the identifiability issue.

1.6 Variation among posteriors across treatments

To evaluate differences between posteriors, we used the Wilcoxon Signed Rank test of R (wilcox.test{stats}). We did not test all pairs of Θ, but tested for differences in posteriors for pairs of treatments that either: were between genotypes that differed by one rank of aggression and were within the same density; or were within the same genotype but which differed by one unit of density. In a subsequent step, we used the marginal mean of each parameter as a point estimate of the location of the full posterior, allowing us to examine effects of genotype, aggression and density. Because information might be lost in summarizing a full posterior this way, we also performed the same tests using nonparametric (permutation) methods on the full posterior. We found the same relationships, with similar significance levels, when we used the full posterior for Θ and when we used the point estimate $\hat{\Theta}$. We therefore opted to present the simpler analysis in the main text.

1.7 Model validation

Any Bayesian model-based analysis paradigm will return a posterior distribution for its parameters, even if the chosen model is a poor fit. Therefore, subsequent to the analysis, it is important to test whether the model is capable of generating observed data values with reasonable probability when using parameters sampled from their estimated posterior distributions. We assess this in two ways. First, we take a number of parameter combinations, generate summary statistics from them, and run the full ABC analysis to test whether we can reliably recover the original parameters. Second, by using the *posterior predictive distribution* (cf Beaumont, 2010).

Full cycle ABC: In order to determine whether our analysis can recover known parameter estimates with any fidelity, we ran a full ABC cycle, starting by generating our summary statistics via simulation from 10 combinations of parameters. These 10 initial sets of Θ were drawn from the posterior of one of our replicates (Ag$\times$d10). We sampled from this region of parameter space, rather than randomly drawn values of Θ, because we are not interested in characterising the behavior and reliability of the entire parameter space. Much of this parameter space will correspond to uninformative behavior, where (for instance) only one sex moves, no flies spend time on patches, or all flies congregate on a single patch. In many of these regions, we may have effectively no power to recover reliable information about most of the initial parameters, or be subject to biases that have no relevance to the

problem we are considering.

We simulated data one time for each of these sets of Θ, sampled them 18 times each, and calculated S. We then used each replicate value of S in a full ABC analysis to recover Θ'. The original values of Θ were appropriately distributed within the range of their posterior estimates. Out of the 70 parameter estimates (for the 10 replicates of the 7 non-zero parameters in Θ), only one fell outside the 95% confidence interval. The lowest quartile of Θ' contained the true Θ 21% of the time; the next quartile contained the true Θ 36% of the time; the next quartile 23% of the time, and the highest quartile 20% of the time.

Posterior predictive distribution: Having established that our simulation robustly returns reasonable estimates of a given Θ, we then use the posterior predictive distribution to establish whether we have the resolution to measure differences in S between replicates. For a given treatment, we repeatedly sample parameter values from their (joint) posterior distribution, given the observed data. For each set of sampled parameter values we then simulate an observed dataset and calculate the values of the summary statistics. The collection of statistic values we observe using this procedure forms the posterior predictive distribution. Poor model fit is indicated if the statistic values observed in the original dataset consistently fall in the tails of the posterior predictive distribution. We performed this validation on the mfHigh and mfLow models separately, and on the combined model.

The empirical values of S for most treatments lie comfortably within the interquartile range of S' (Table S1). For example, in the dataset derived from the full posterior, mAve, fAve, mfCOv, fPermM and mPerF, 100% of S for all treatments lie within the interquartile range. This indicates good model fit. However, we do see some evidence of a tendency to generate values of mAveSD and fAveSD that tend to the high side. This indicates that, despite the good overall fit, our models appears to favor less-dynamic fly movement rates. This suggests possible future refinements to our model, that better capture this feature.

1.8 Test of difference between male and female social preferences

It is a known drawback of ABC that typical model selection approaches may fail, because the (intractable) term $P(D|\Theta)$ will vary between models with different parameters (Robert et al., 2011). A typical hierarchical model building approach might start by modeling male and female social pref-

Table S1: Summary of model validation results for the full prior, and the two restricted models mfHigh and mfLow. For each of the 11 summary statistics, the proportion of treatments for which the empirical statistics lie within the interquartile range of their posterior predictive distribution is shown

Stat	Full Prior	mfHigh	mfLow
mAve	1.00	1.00	1.00
fAve	1.00	1.00	0.97
mAveSD	0.37	0.33	0.33
fAveSD	0.50	0.57	0.23
mVar	0.90	0.93	0.73
fVar	0.87	0.87	0.93
mfCov	1.00	0.93	0.97
mPerM	0.97	0.93	0.83
fPerM	1.00	1.00	0.77
mPerF	1.00	1.00	0.87
fPerF	0.97	0.90	1.00

erences as identical, and then test for the significance of subsequent sex-specific terms. Because we were interested in measuring the differences in male and female social behavior from the outset, and given that we could not perform model selection, we constructed our minimal model including the sex$\times$sex interaction terms.

If these terms were, in fact, not sex specific, we would not expect to see consistent differences between our estimated social preference terms a_{mf} and a_{ff}; or a_{mm} and a_{fm}. We performed model selection, testing for differences in these terms among replicates, using the mean values of the social preference parameters in linear models, with sex (1 or 0), density, and the interaction between them as possible predictors; ranking models by BIC.

Model selection suggested that males and females differed in both their preferences for males and females, and that there was an interaction with density on these preferences. In the contrast between male and female social preferences for females, the full model was preferred, with sex (est=7.76, t=5.62, P$<$0.001), density (est=0.59, t=4.91, P$<$0.001) and the interaction (est=-0.50, t=-2.96, P=0.005) all significant. Similarly, for the social preference for males the full model was preferred. Sex (est=-6.12, t=-2.98, P=0.004), density (est=-0.60, t=-3.33, P=0.002) and the interaction (est=0.59, t=2.33, P=0.023) were all significant.

Given the degree of significance on the consistency of the social preference terms, and the complexity of the interactions, we are comfortable that there is support for our choice to model social preferences separately for the sexes.

1.9 Hypothesis testing in an ABS framework

As described in the main text, we tested hypotheses for mechanisms of group structure regulation in our ABS framework. To test for effects of aggression and density, we first created a "baseline" model, where all parameters were set to their overall means across genotype and density treatment. To model the effects of density or aggression on parameters of interest, we took the linear expectation of these parameters at the extremes of density or aggression. In all simulations we used 20 flies: 10 males and 10 females (equivalent to our highest density treatment). We sampled from the simulation in the same way as during the ABC process: from each simulation 20 times, at even intervals relative to "fly time". We calculated our summary statistics from these samples for each replicate independently, again as during the ABC process. We ran 30 replicates for each simulated "treatment", varying the random seed for each replicate. Thus, the sample size for estimating effects was identical to that of our empirical data, and our ABC estimates. When we wanted to minimize the effect of a_{mf} on female behavior, we set a_{mf} to 1, rather than 0. Setting a_{mf} to 0 invariably resulted in nearly all females joining a single patch and not moving at all, such that it was difficult to detect the effects of variation in other parameters.

1.10 Aggression effects, and the role of a_{mf} on changes in female behavior

We hypothesized that several of the changes in parameter estimates with aggression were related to disruption of group structure by mobile, displaced males. In particular, higher male-male aggression was correlated with higher female group joining rates (j_f) but with no effect on fAv (the average number of females on patches). We inferred that these changes might be reflected in higher male and female movement rates, but that the effect of aggression on females would be minimized if a_{mf} was negligible.

We created two combinations of parameters that differed in only the 3 parameters directly correlated with aggression: j_f, j_m, and a_{mm}. In both cases, l_m=0, l_f=-2.96, a_{mf}=6.4, a_{fm}=-2.5, and a_{ff}=2.6. These are the mean values for these parameters across all trials. In the low aggression mimic, j_m=-0.5, j_f=-0.5, and a_{mm}=-3. In the high aggression case, j_m=0.4, j_f=0.1, and a_{mm}=-5, which are approximately the linear expectations in high and low aggression. To explore the effect of a_{mf} on overall movement rate, and particularly female movement rate, we set a_{mf} to 1 (nearly neutral) in the high-aggression

combination of parameters.

All treatments were highly significantly different from each other ($P<0.001$) in both time and the proportion of male moves (arcsin transformed) under standard linear regression. At high simulated aggression, flies moved more often overall, and males increased their movement rate more than females. When the effect of male preference for females was removed, flies moved less often, and males accounted for almost all the movement (Figure S2). This suggests that the greatest contributor to female movement among patches could be male presence on those patches. However, it is important to note that what we are testing here is whether it is possible that changes in male-male interactions can lead to the apparent effects we detect in other male-and-female parameters, and in overall group dynamics. Our proposed mechanism—that aggression changes increases overall male movement rates, and thus indirectly drives an increase in female overall movement rates—is only one possible interpretation of our results. It is possible that there are other, more subtle effects that our minimal simulation and univariate regression analysis is not detecting. Full testing of our hypothesis will require further experimentation and analysis. The same qualified interpretation should be applied to our analysis in the next section.

1.11 Density Effects, and the interaction between a_{mf} and changes in female preferences

We found that, at higher densities, female social parameter estimates were more extreme. We did not find changes in the summary statistics that reflected these changes. We hypothesized, first, that the changes in a_{fm} and a_{ff} canceled each other out (that if we changed one or the other, group outcomes would change dramatically). We hypothesized further, that they canceled-out via the effects of male behavior. We test this as follows.

Under high density, female attraction to females a_{ff} increases, and female attraction to males a_{fm} decreases. We set the parameters in our baseline model: l_m=0, l_f=-2.96, j_f=-0.5, a_{mm}=-3.95, a_{mf}=6.1, a_{fm}=-1.2, a_{ff}=1.2. We varied two parameters to their high density values—a_{fm}=-3.9 and a_{ff}=3.9—separately and together. Then, to examine whether the observed buffering of female group statistics was due to the effect of males, we tested the high versus low density values of a_{ff} and a_{fm} with a_{mf} set to 1 (nearly neutral). Note that we do not change the number of flies in these models.

We tested the 3 female-specific group statistics: fAve, fVar, and fPerF for the effects of these manipulations. We compared the "low density" statistics to the three cases where we varied a_{fm}, to its low "high density" value, a_{ff} to its high "high density" value, and both parameters together. For one statistic, fAve, the "low density" values and the "high density" values were significantly different (est=-0.13, t=-5.96, P¡0.001). For the other two statistics, the difference was not significant. The statistics were all highly significantly different from the "low density" baseline when we varied a_{fm} and a_{ff} independently (Figure S3). When we removed the effect of a_{mf}, the two low a_{mf} treatments were significantly different in all statistics. This suggests that male attraction to females is a necessary condition for the buffering effect on female group structure with changes in preferences due to density.

We only report the female-specific parameters, because the effects of changing female preferences and a_{mf} on male-female covariance statistics were difficult to interpret, and will need to be explored further in conjunction with additional experimental manipulations in future.

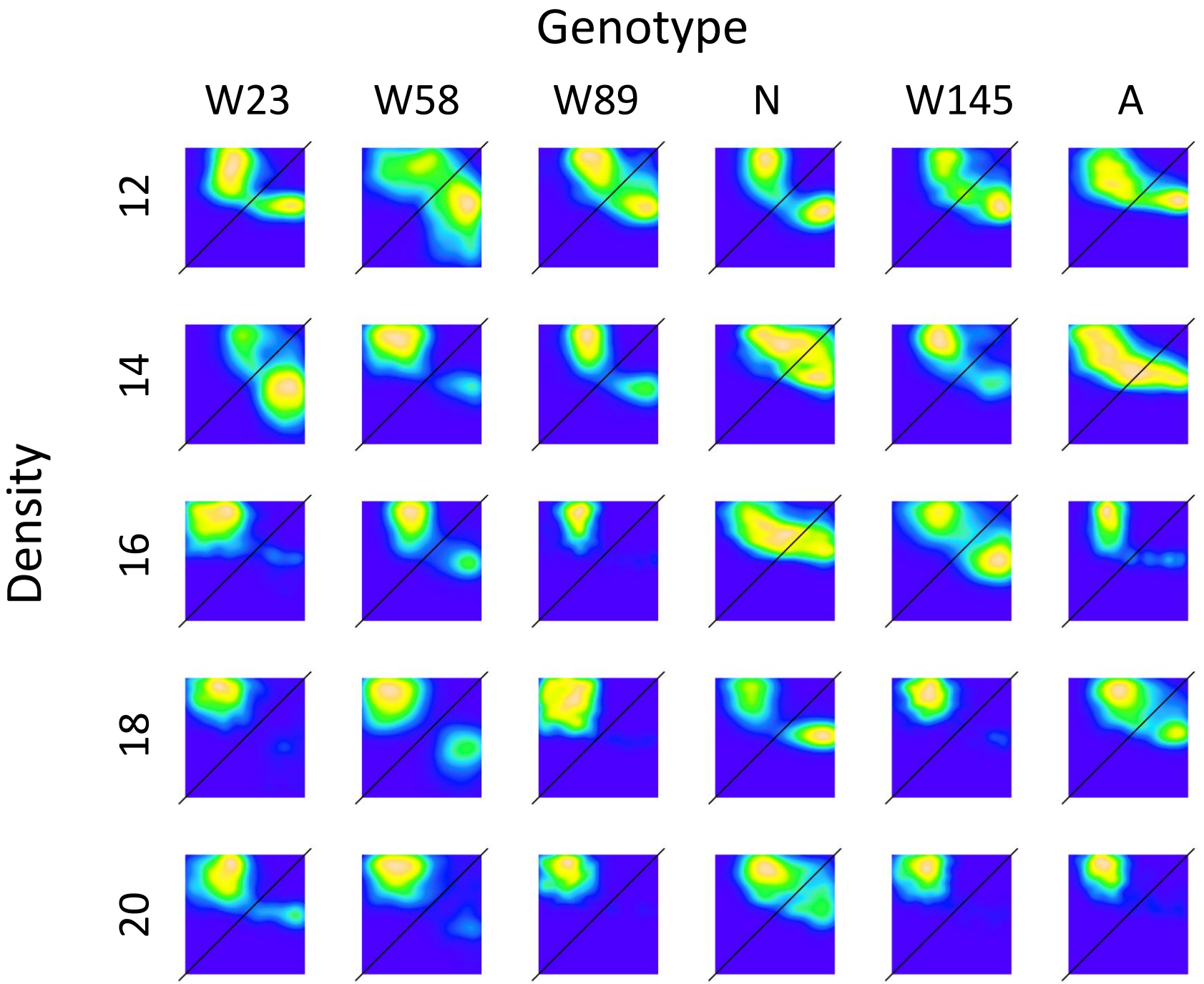
Online Figure S1: Bivariate density plots of the joint density estimates of y=a_{mf} and x=a_{fm} for each treatment. Low posterior densities are indicated with blue, and high with yellow. All axese are scaled -10:10, and the identity line is indicated in black.

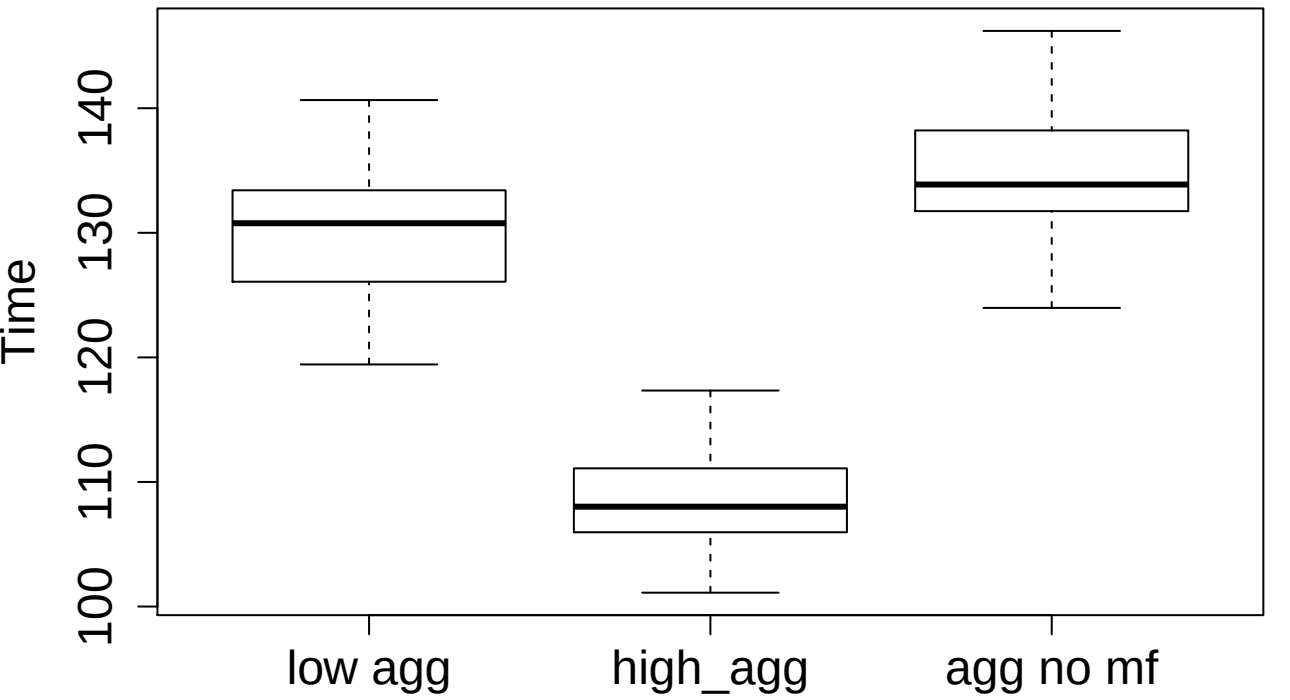
Online Figure S2: The effect of simulated aggression, and removing the effects of a_{mf}, on rates of overall movement, and relative male movement rate. Time is expressed in relative fly 'seconds', moves are the proportion of total moves made by males.

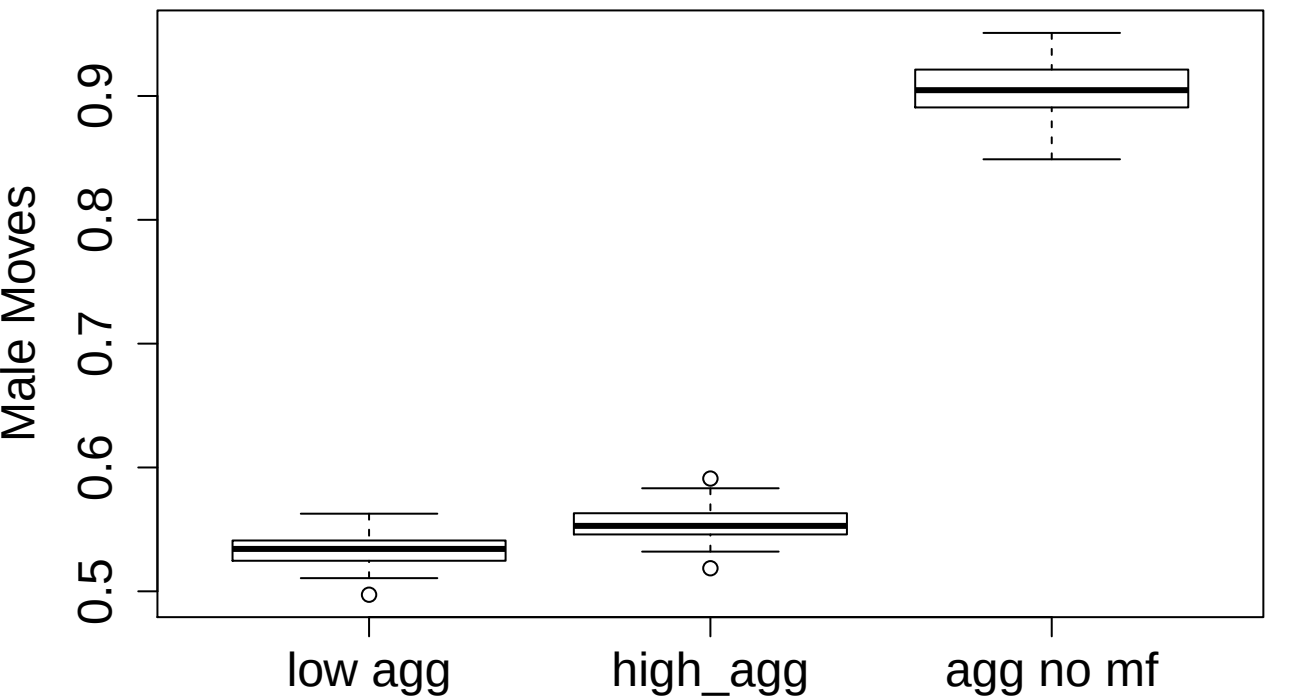
Online Figure S3: The effect of simulated density changes in female group preferences from "low density" (Low Dens), when parameters vary separately (Low fm and High ff) and together (High Dens). We also tested whether male preferences for females was a possible mechanism for the lack of change in statistics between the Low Dens and High Dens treatments. We tested the effects on 3 female specific statistics, the average number of females among patches (Female Ave), the variance in number of females among patches (Female Var) and the number of females per female in a group (F per F). Of all comparisons made, only the Low Dens - High Dens comparisons for Female Var and F per F were non-significant.

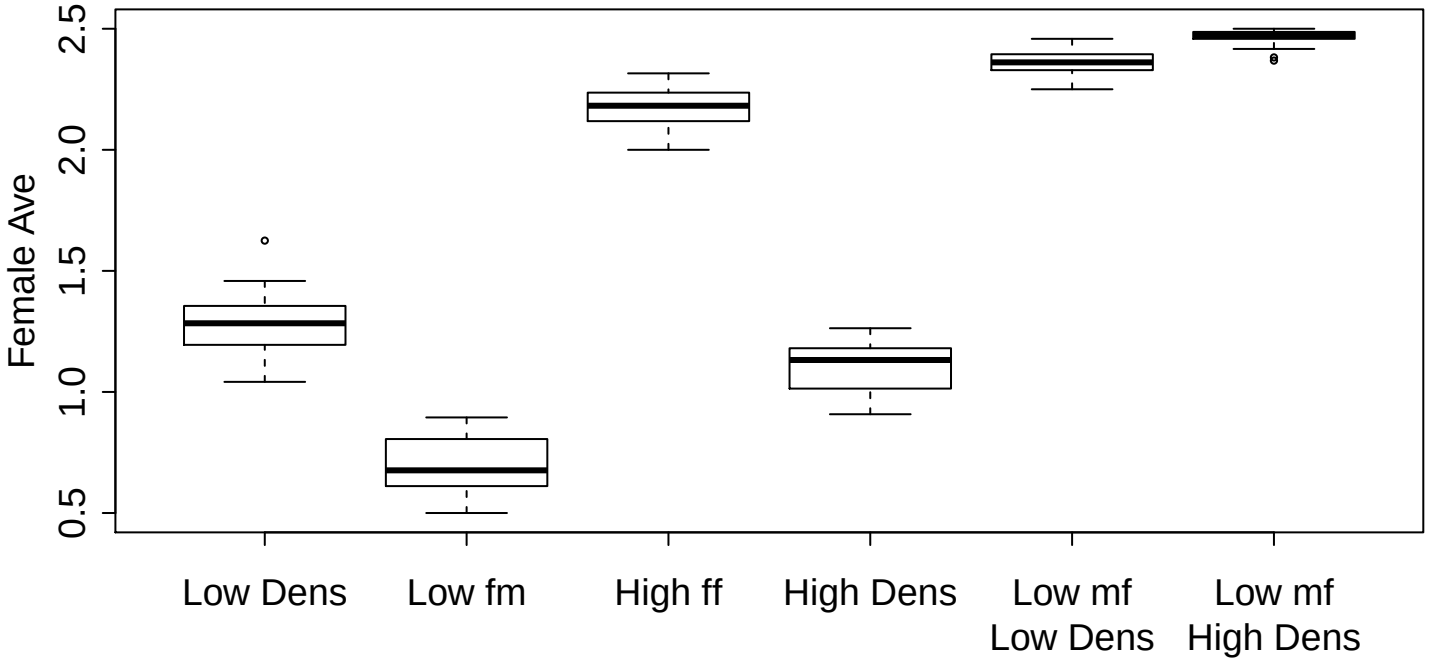
Statistic	Description
mAv/fAv	Average number of males, or females, in groups
mAvSD/fAvSD	Sampling variance in mAv, or fAv
mVar/fVar	Variance in the number of males, or females, among groups
mfCor	Correlation in the number of males and females among groups
mPerM/fPerM /mPerF/fPerF	The average number of (other) males or females each male, or female, experiences in their group.

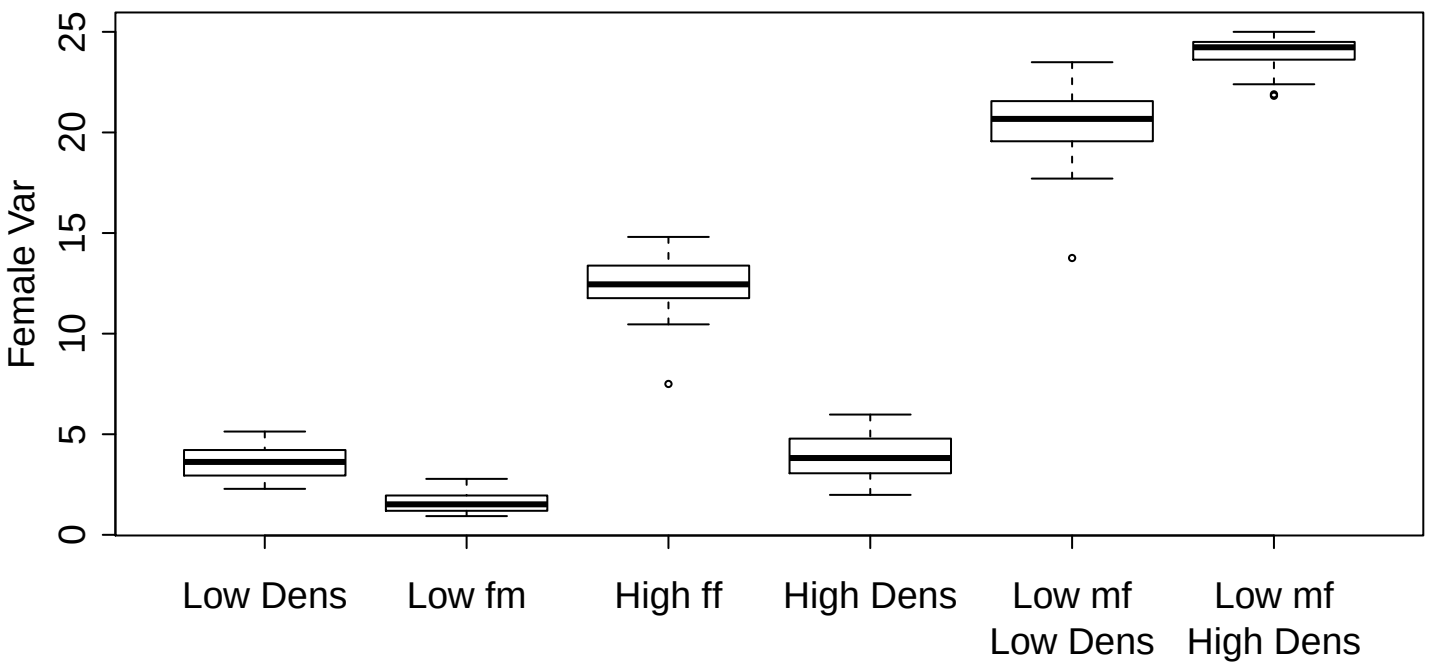
Θ	Parameter Description
l_m	Group leaving rate for males (=0)
j_m	Group joining rate for males
l_f	Group leaving rate for females
j_f	Group joining rate for females
a_{mm}	Effect of other males on male-group leaving rate
a_{mf}	Effect of females on male-group leaving rate
a_{fm}	Effect of males on female-group leaving rate
a_{ff}	Effect of other females on female-group leaving rate

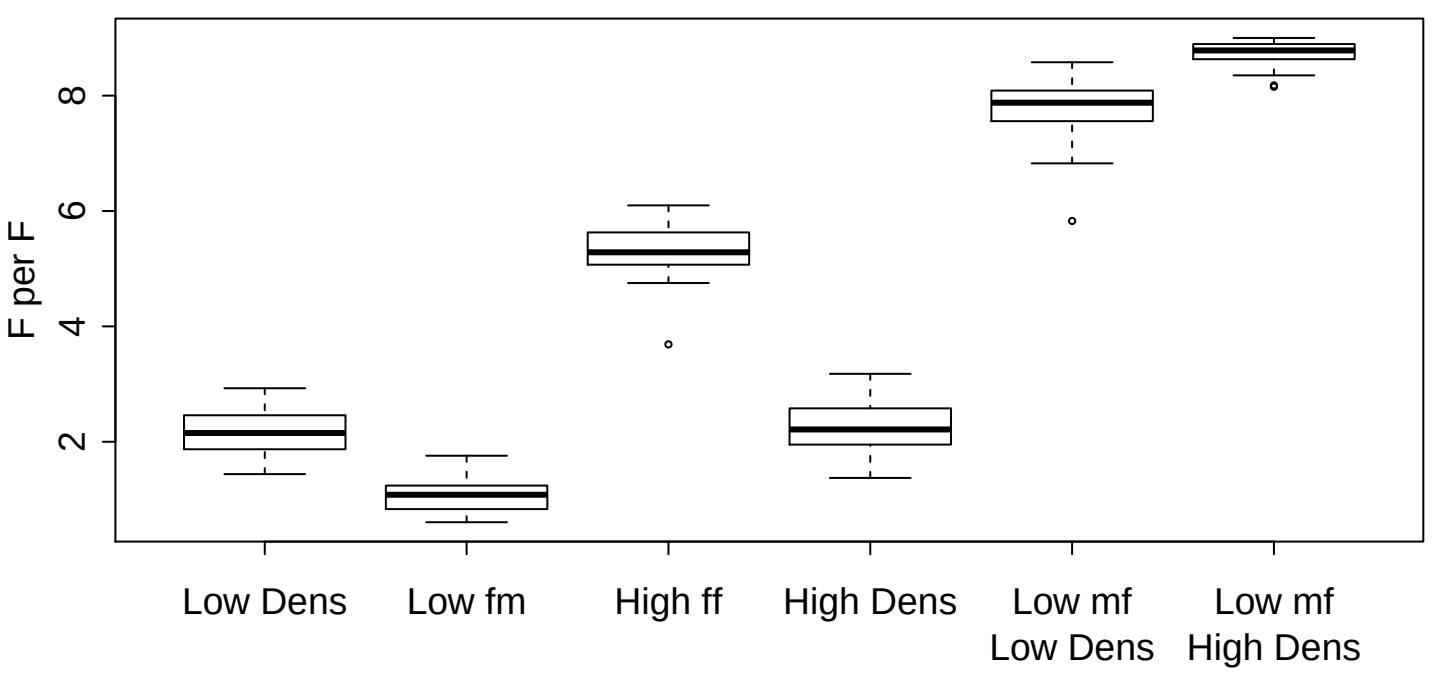

Genotype
W23
W58
W89
N
W145
A
Density
12
14
16
18
20


Time
100
110
120
130
140
low agg
high_agg
agg no mf


Male Moves
0.5
0.6
0.7
0.8
0.9
low agg
high_agg
agg no mf


Female Ave
0.5
1.0
1.5
2.0
2.5
Low Dens
Low fm
High ff
High Dens
Low mf
Low Dens
Low mf
High Dens


Female Var
0
5
10
15
20
25
Low Dens
Low fm
High ff
High Dens
Low mf
Low Dens
Low mf
High Dens


F per F
2
4
6
8
Low Dens
Low fm
High ff
High Dens
Low mf
Low Dens
Low mf
High Dens