PNAS
www.pnas.org

Supplementary Information for

Genomic Regions Influencing Aggressive Behavior in Honey Bees are Defined by Colony Allele Frequencies

Arian Avalos, Miaoquan Fang, Hailin Pan, Aixa Ramirez Lluch, Alexander E. Lipka, Sihai Dave Zhao, Tugrul Giray, Gene E. Robinson, Guojie Zhang & Matthew E. Hudson

Corresponding Authors:
Gene E. Robinson; **Email:** generobi@illinois.edu.
Guojie Zhang; **Email:** Guojie.Zhang@bio.ku.dk.
Matthew E. Hudson; **Email:** mhudson@illinois.edu.

This PDF includes:

Captions for Dataset S1-2
Figs. S1-7

www.pnas.org/cgi/doi/10.1073/pnas.1922927117

Dataset S1 (annotation_data_table_s1.xlsx).

Data table in Excel format with three sheets corresponding to a legend and reference information, colony phenotype and collection information, a gene annotation list for the genes overlapping haplotypes containing a significant SNP, and a list of significant SNPs . The colony information sheet contains specific information including collection and sampling data as well as phenotype details for both measures of colony aggression. The gene annotation list contains information on the 254 genes within genomic regions of significant correlation. Information includes linkage group, number of haplotype blocks with significant SNPs in overlap with the gene, *A. mellifera* NCBI gene IDs and names, and gene ID, symbol and name for the nearest *D. melanogaster* homolog. Also highlighted are the subset of 56 genes in overlap with haplotype blocks also containing markers with evidence of selection. The SNP list contains information on the specific markers identified by our colony-level analysis. Included are positional information (linkage group, base pair position), nucleotide information, and resulting p-values for the individual- and colony-level analysis.

Dataset S2 (simulation_code_s5.r).

Annotated algorithm which conducts a P-value assessment using simulated Principal Components. Annotations summarize conceptual framework and provide the model under consideration for SNP x colony phenotype which mirrors the model considered in the manuscript.

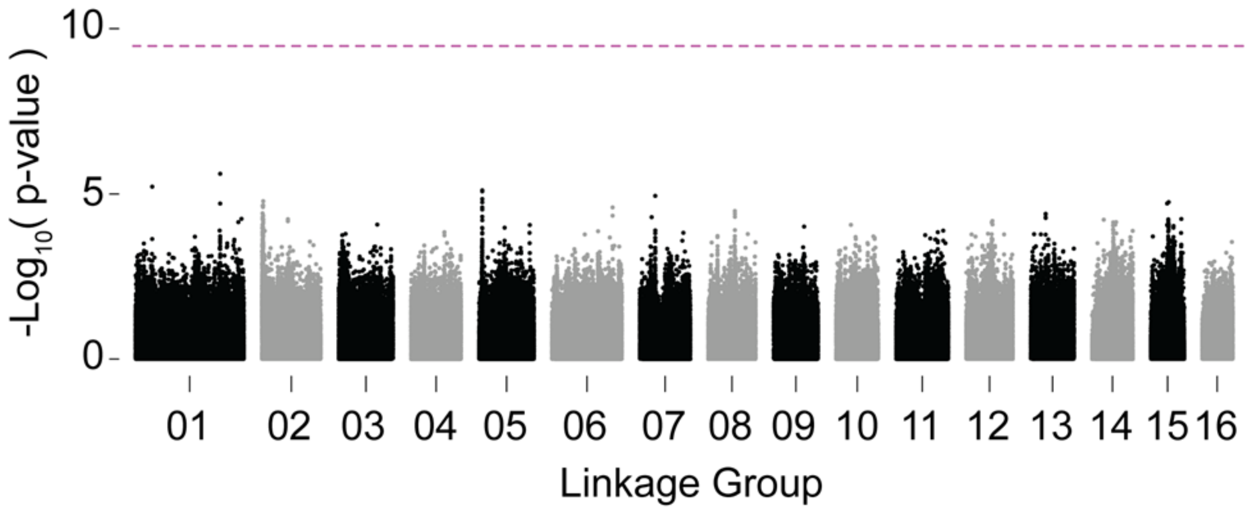

Fig. S1. Genome-wide associations of aggression at the individual level. Manhattan plot of P value distributions across the genome for the correlation of individual-level genotype to individual behavioral phenotype (Soldier vs. Forager). The dashed magenta line represents the Bonferroni adjusted threshold (α = 3.35E^{-10}) consistent between individual and colony level analyses.

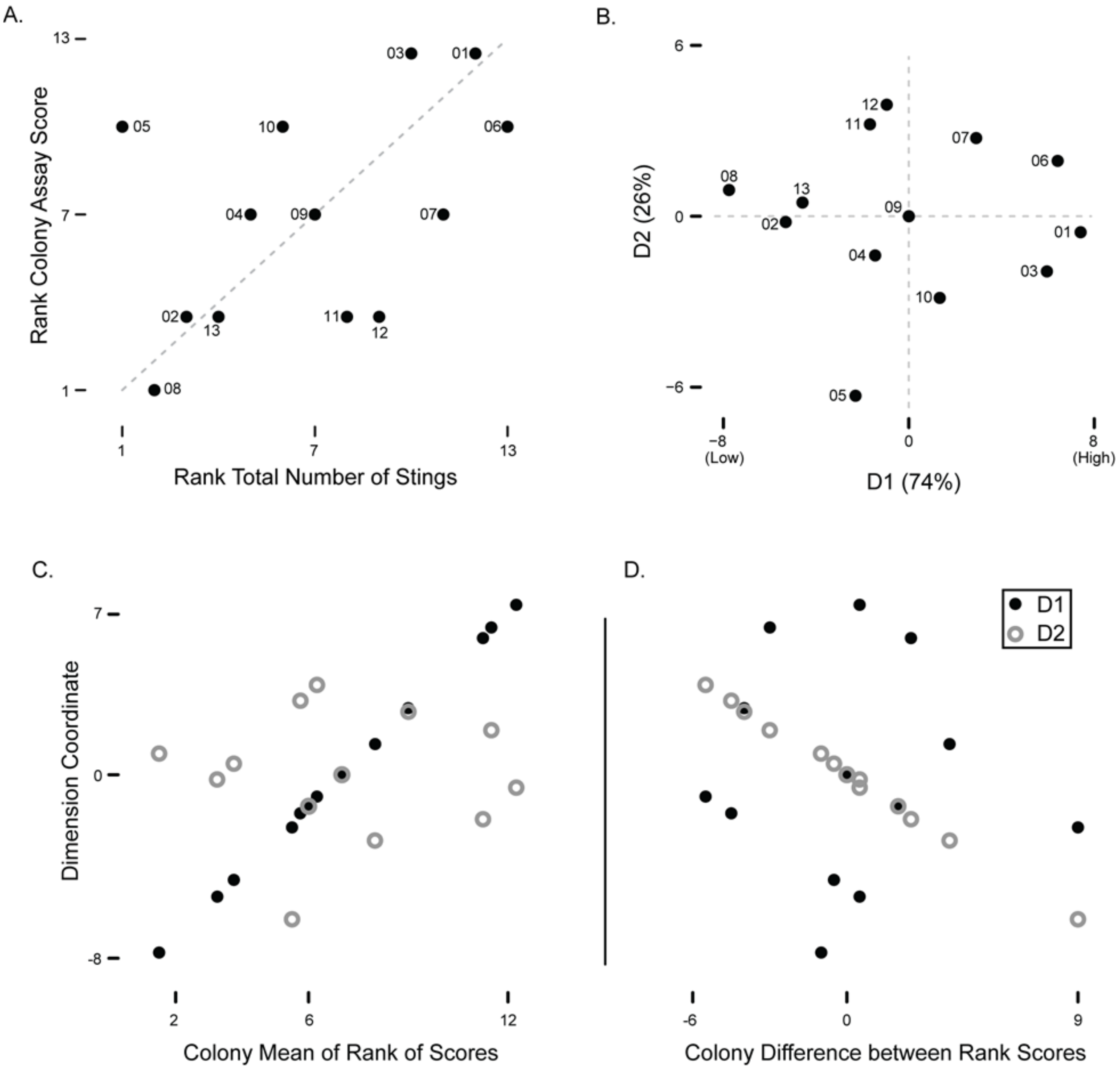

Fig. S2. Assessment of colony phenotype. Panel A: correlation between the rank values of the response for each colony in the two behavioral assays. Panel B: results from Multidimensional Scaling (MDS). In both panels each point corresponds to a colony and is labeled with its colony number. Panels C and D summarize the four correlations between the dimension coordinate value for each of the colonies and the corresponding summary of phenotype. Panel C summarizes the relationship between the dimension value and the per-colony mean of rank scores between the two assays. Panel D summarizes the correlation between the same dimensional position and the per-colony difference between the ranks of the behavioral assays. As in panels A and B, each point is a colony, black points are used when correlating against the first MDS dimension (D1) and open grey circles are used when correlating against the second MDS dimension (D2).

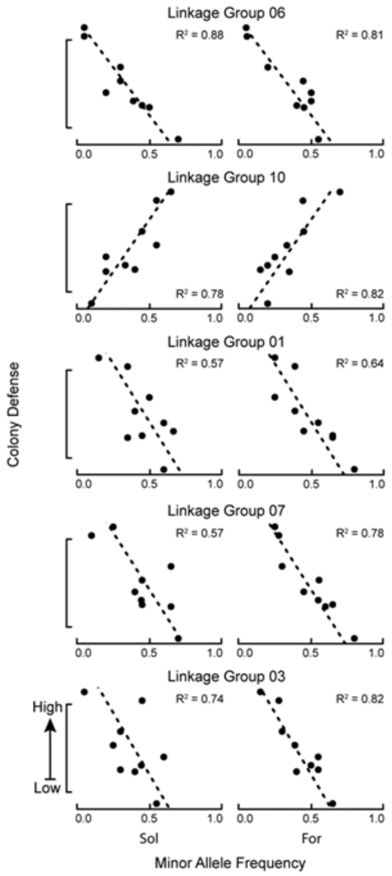

Fig. S3. Correlation between minor allele frequencies (MAF) for each behavioral group and the colony phenotype. Goodness of fit (R^2) was estimated to the model function derived from the colony-level fit (dashed line). A paired set of plots is provided for each of the top 5 focal SNPs of the top 5 peaks of association candidate SNPs (Fig. 1C) one for Soldiers (Sol) and the other for Foragers (For). Y axis corresponds to colony defense (D1; SI Appendix, Fig. S2), X axis to the MAF, and each point represents a colony.

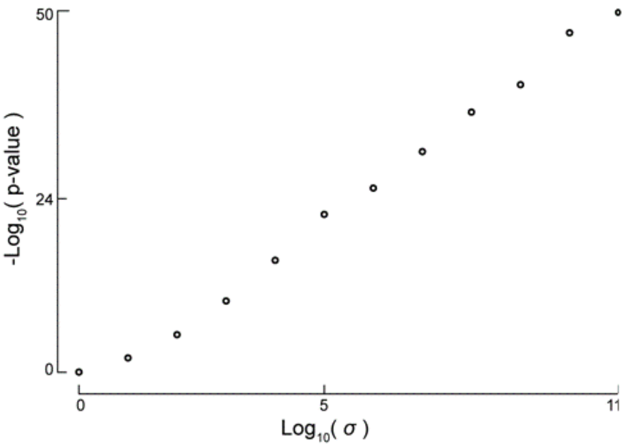

Fig. S4. Relationship between a range of simulated residual variances and resulting P values. In our analysis, even with N = 9, as residual variance decreases, p-values reach levels of 10 - 50 (1.00E^{-10} – 1.00E^{-50}), as we detected (Fig. 1).

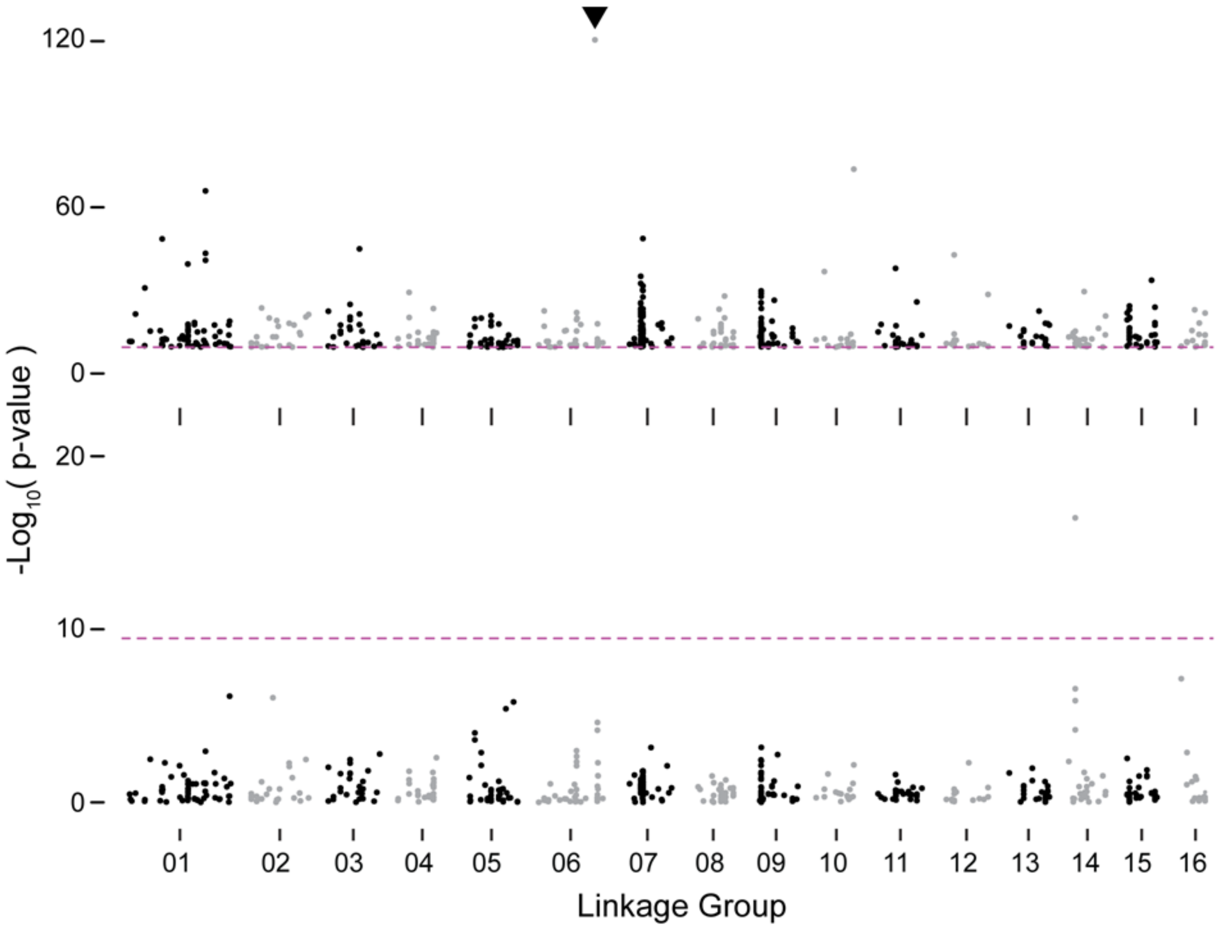

Fig. S5. Analysis of covariation across top candidate SNPs. Top panel: P values of SNPs from Fig 1C that pass the significance threshold, plotted against genomic location. Bottom panel: P values of significant SNPs from Fig 1C with the most significant SNP (highlighted by the black triangle) included in the model as a covariate.

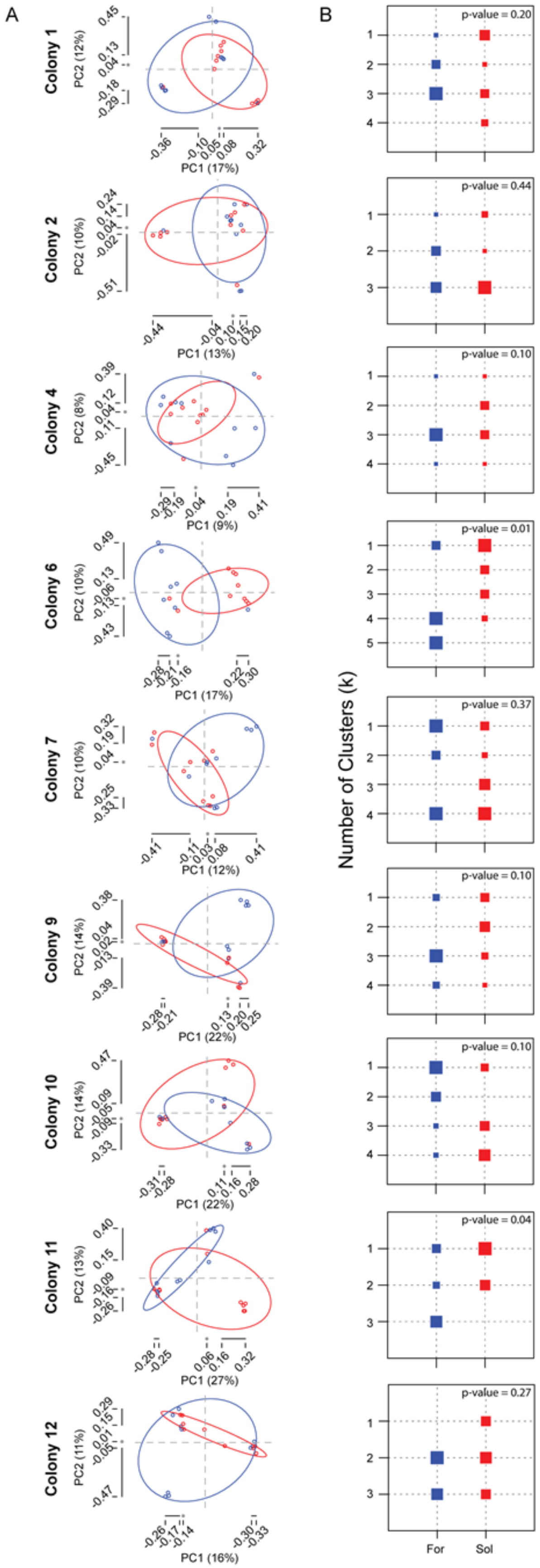
A
B
Colony 1
PC2 (12%)
PC1 (17%)
p-value = 0.20
Colony 2
PC2 (10%)
PC1 (13%)
p-value = 0.44
Colony 4
PC2 (8%)
PC1 (9%)
p-value = 0.10
Colony 6
PC2 (10%)
PC1 (17%)
p-value = 0.01
Colony 7
PC2 (10%)
PC1 (12%)
p-value = 0.37
Colony 9
PC2 (14%)
PC1 (22%)
p-value = 0.10
Colony 10
PC2 (14%)
PC1 (22%)
p-value = 0.10
Colony 11
PC2 (13%)
PC1 (27%)
p-value = 0.04
Colony 12
PC2 (11%)
PC1 (16%)
p-value = 0.27
Number of Clusters (k)
For
Sol

Fig. S6. Analysis of concordance of genetic diversity with aggressive phenotype within each colony. (A) Principal component analysis of genetic variation is summarized by the first two principal components derived from the genotype matrix for each colony. Each point corresponds to a sample, and each sample is colored by behavioral group: blue = Forager (For), red = Soldier (Sol). An ellipse encapsulating 65% of the samples within a behavioral group (~7 in each group) is provided to further highlight distribution of the behavioral groups across the PC space. (B) Optimal number of clusters was determined via iterative k-means clustering and the elbow method using the within-group total sum of squares. The distribution of behavioral groups between genetic clusters was assessed for each colony using a Fisher's exact test (P value at top right of table).

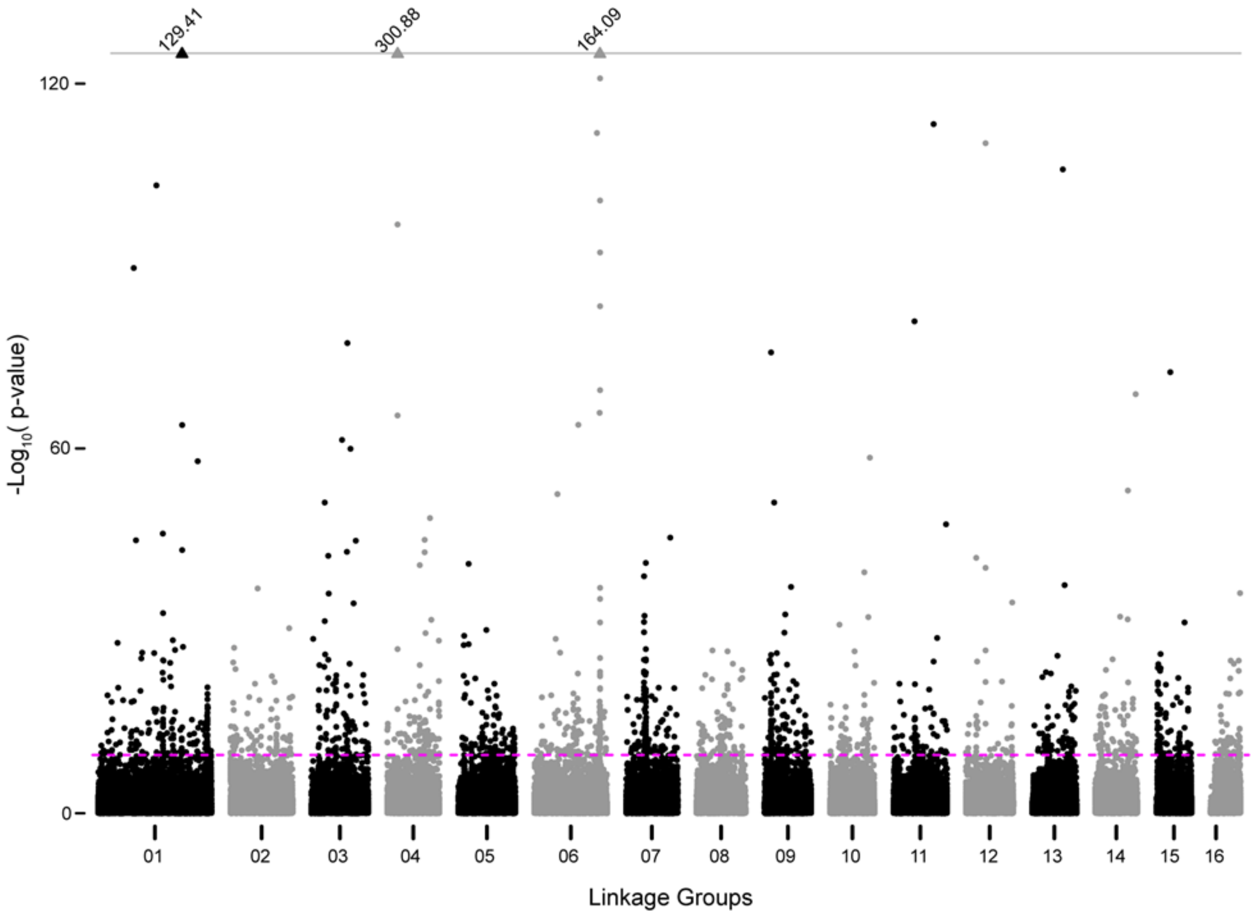

Figure S7. Genome-wide associations of aggression at the group level considering original geographic source of colony. Manhattan plot of minimum p-value in each haplotype block resulting from an association of colony-level minor allele frequency across the genome to colony aggression phenotype, considering geographic source of colony as a covariate. Dashed line shows significance threshold (Bonferroni corrected α = 7.67E^{-10}). Outlier haplotype blocks are highlighted as triangles aligned on an arbitrary limit bar, and the -log10(p-value) is provided separately above the triangle.