Supplemental Materials:

Title: How do genomes create novel phenotypes? Insights from the loss of the worker caste in ant social parasites.

Authors:

Chris R. Smith[1*], Sara Helms Cahan[2], Carsten Kemena[3], Seán G. Brady[4], Wei Yang[5], E. Bornberg-Bauer[3], Ti Eriksson[6], Juergen Gadau[6], Martin Helmkampf[6], Dietrich Gotzek[7,8], Misato Okamoto Miyakawa[9], Andrew Suarez[8,10], and Alexander Mikheyev[9*]

[1] Dept. of Biology, Earlham College, USA

[2] Dept. of Biology, University of Vermont, USA

[3]Institute for Evolution and Biodiversity, WestfälischeWilhems-UniversitätMünster, Germany

[4]Dept. of Entomology, National Museum of Natural History, Smithsonian Institution, USA

[5] Dept. of Computer Science, University of Illinois, USA

[6] School of Life Sciences, Arizona State University, USA

[7] Dept. of Entomology, University of Georgia, USA

[8] Dept. of Animal Biology, University of Illinois, USA

[9] Ecology and Evolution Unit, Okinawa Institute of Science and Technology, Japan

[10] Dept. of Entomology, University of Illinois, USA

* Authors for correspondence: crsmith.ant@gmail.com, alexander.mikheyev@oist.jp

***Caste Exclusive Genes*:**

There were two exclusively queen expressed genes in *V. emeryi*. They were: fgenesh_masked-contig00927-processed-gene-0.8, and snap_masked-contig03825-processed-gene-0.9. The former gene has weak homology to a bacterial transcriptional regulator while the latter has no detectable homology. It is likely that both of these genes represent non-coding RNAs.

There were three exclusively queen expressed genes in *Camponotus floridanus*, Cflo_00584 (a likely transposase), Cflo_02637 (uncharacterized protein with homology across hymenopteran genomes), and Cflo_15228 (a predicted mevalonate kinase).

Signatures of Positive Selection:

We used a branch-site model to detect genes under positive selection in the social parasites compared to their hosts. There were only three genes with a signature of positive selection in the harvester ant social parasites (*P. colei,* and *P. anergismus*) : *ERC protein2*, *COX18*, and *tRNA (uracil-5)-methyltransferase-like protein A*, or PB10737, PB14156, and PB15272, respectively. Positive selection in social parasites for these genes suggests that neurological and metabolic changes may be important in the transition to social parasitism. ERC protein2 is a RIM-binding protein involved in development and regulation of neurotransmitter release in synaptic active zones; mutants in the *Drosophila* ortholog, *brp*, show defects in synaptic plasticity and olfactory memory (Knapek et al. 2011). COX18 plays a role in assembly and activity of cytochrome c oxidase in the mitochondrial membrane respiratory chain. In contrast, no genes exhibited significant positive selection in *V. nipponica* relative to its host, *V. emeryi*.

Tables

Table S1. Basic information on the genomes used in this study. Assembly size of the mapped genomes is limited to the template genome size and hence is not meaningful.

Species	Fold coverage	Assembly size (mb)	# contigs/scaffold/# contigs/scaffolds covered 90%**	No. of genes*	GC content	annotation	Source
Pogonomyrmex barbatus	12 (Roche 454)	235	4646	17,177	33.9	denovo	Smith et. al 2011
P. rugosus	71	n.a.	3861	17,093	33.9	mapped to P. barbatus	this study
P. colei	55	n.a.	3467	16,910	33.9	mapped to P. barbatus	this study
P. anergismus	40	n.a.	3743	16,968	33.9	mapped to P. barbatus	this study
Vollenhovia emeryi	11 (Roche 454)	269	46989	26,902	42.2	denovo	this study
V. nipponica	94	n.a.	40233	25,102	42.2	mapped to V. emeryi	this study

* for mapped genomes number of genes is defined as genes that have at least a 90% coverage of the cds.

** number of scaffolds for de novo genomes and number of scaffolds covered at least by 90% for the mapped genomes

Table S2. Library information for *P. barbatus* (J-lineage) samples. Sampled marked as adult queens were virgin queens (gynes).

Caste	Stage	LarvaSize	LarvaMass (mg)	Label	LibrarySize(reads)
Worker	Larva	small	1.8	41-10	589656
Worker	Larva	large	23.3	41-21	8884403
Queen	Larva	large	42.9	41-27	8958355
Worker	Larva	small	1.5	41-29	1377220
Queen	Larva	small	1.2	41-31	2850808
Queen	Larva	small	1.4	41-8	2254900
Worker	Larva	large	24.6	44-17	23717
Queen	Larva	large	46.3	44-30	738837
Queen	Adult			GA-5	20310060
Queen	Adult			GA-6	17558249
Queen	Pupa			GP-10	15255864
Queen	Pupa			GP-8	17078579
Worker	Adult			WA-11	24633962
Worker	Adult			WA-12	15542426
Worker	Pupa			WP-14	20041117
Worker	Pupa			WP-15	14146680

Table S3. RNA sequencing library information for *V. emeryi* samples. "Q" samples are derived from queens and "W" samples are derived from workers.

Sample	Library Size (reads)
Q1	35717244
Q2	43469416
Q3	38788964
Q4	41470112
Q5	41224582
W1	43877546
W2	38389236
W3	57789164
W4	43926009
W5	31908597

Table S4. Differential gene expression results for *P. barbatus*. The number of genes with higher expression in the first category (e.g., "Worker") are listed under "1" in each table, and those with higher expression in the second category (e.g., "Queen") are listed under "-1"; those genes not differentially expressed are listed under "0". The conservative analysis included only genes with at least 100 copies per million in at least two individuals while the liberal analysis included genes with at least 2 copies per million in at least four individuals. The total is the total number of genes used in the analysis, while "Total DE" is the total number of genes that were differentially expressed.

	Liberal Analysis					**Conservative Analysis**				
	1	0	-1	Total	Total DE	1	0	-1	Total	Total DE
Adult - Large Larva	1683	8767	2010	12460	3693	763	2239	787	3789	1550
Adult - Pupa	1024	10703	733	12460	1757	431	3036	322	3789	753
Adult - Small Larva	1749	7160	3551	12460	5300	844	2039	906	3789	1750
Large Larva - Pupa	1411	10285	764	12460	2175	568	2786	435	3789	1003
Large Larva - Small Larva	216	11054	1190	12460	1406	100	3435	254	3789	354
Pupa - Small Larva	1176	7710	3574	12460	4750	633	2307	849	3789	1482
Worker-Queen (Adult)	65	12351	44	12460	109	40	3724	25	3789	65
Worker-Queen (Pupa)	121	12286	53	12460	174	55	3697	37	3789	92
Worker-Queen (Large Larva)	87	12260	113	12460	200	100	3604	85	3789	185
Worker-Queen (Small Larva)	466	11892	102	12460	568	46	3724	19	3789	65

Table S5.
Differential gene expression results for *V. emeryi*. The number of genes with higher expression in the workers are listed under "1" in each table, and those with higher expression in queens are listed under "-1"; "; those genes not differentially expressed are listed under "0". The conservative analysis included only genes with at least 100 copies per million in at least two individuals while the liberal analysis included genes with at least 2 copies per million in at least four individuals. The total is the total number of genes used in the analysis, while "Total DE" is the total number of genes that were differentially expressed.

	Liberal Analysis					**Conservative Analysis**				
	1	0	-1	Total	Total DE	1	0	-1	Total	Total DE
Worker -Queen (Adult)	2065	4278	3844	10187	5909	543	830	1139	2512	1682

Table S6. Orphan genes were not enriched in genes with worker-biased gene expression ($\chi^2 = 2.2$, df = 1, P = 0.13). The worker up-regulated gene list was all genes up-regulated in workers at any developmental stage (see Table S4).

Comparison	**Orphan**	**Non-Orphan**	**Total**
Worker-Biased	45	788	833
Total	833	11627	12460

Table S7. Orphan genes were not enriched in genes with caste-biased gene expression ($\chi^2 < 0.01$, df = 1, P = 0.99). The caste up-regulated gene list was all genes up-regulated in either workers or queens, at any developmental stage (see Table S4).

Comparison	**Orphan**	**Non-Orphan**	**Total**
Caste-Biased	65	907	972
Total	833	11627	12460

Figures

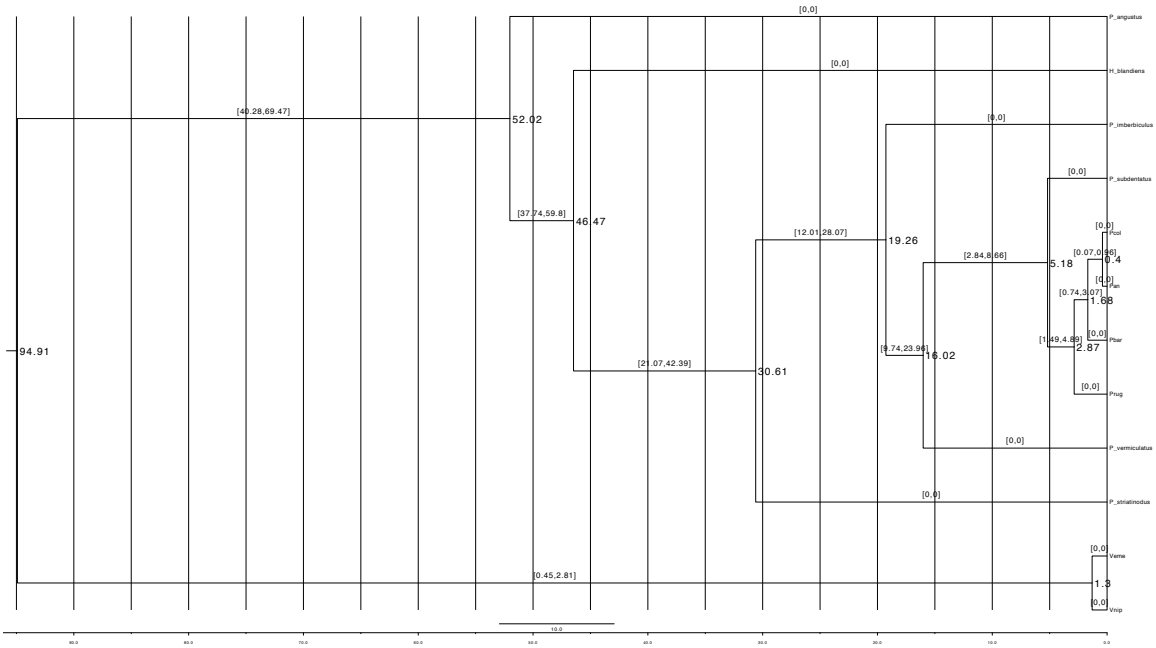

Figure S1. Dated phylogenetic reconstruction of the host-parasite pairs for *Pogonomyrmex* and *Vollenhovia*. All nodes have 1.0 posterior support. The dating (in MYA) are at the nodes while 95% confidence intervals are on the branches. The species abbreviations (focal species from this study) on the branch tips are: Pcol = *Pogonomyrmex colei*, Pane = *P. anergismus*, Pbar = *P. barbatus*, Prug = *P. rugosus*, Veme = *Vollenhovia emeryi*, Vnip = *V. nipponica*. All non-*Vollenhovia* species are *Pogonomyrmex* except *Hylomyrma blandens*, which is clearly nested within the *Pogonomyrmex* clade.

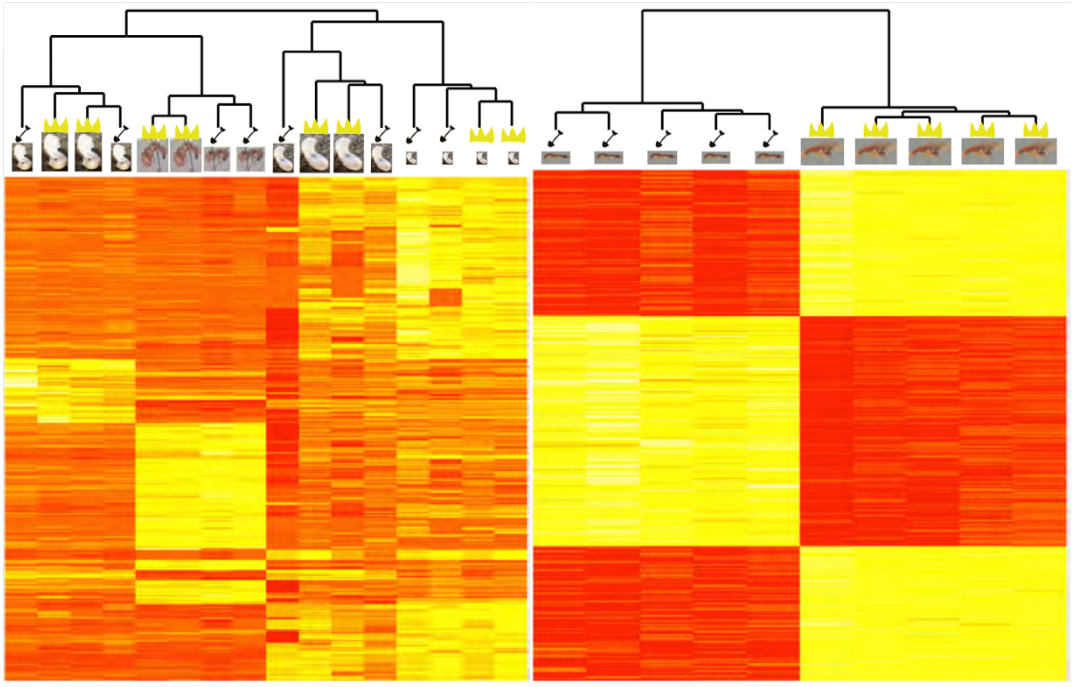

Figure S2. Heatmaps of gene expression (log normalized counts) for the top 500 genes in each *P. barbatus* (left) and *V. emeryi* (right). Stages are represented by pictures and caste is represented by symbol (crown for queen and shovel for worker).

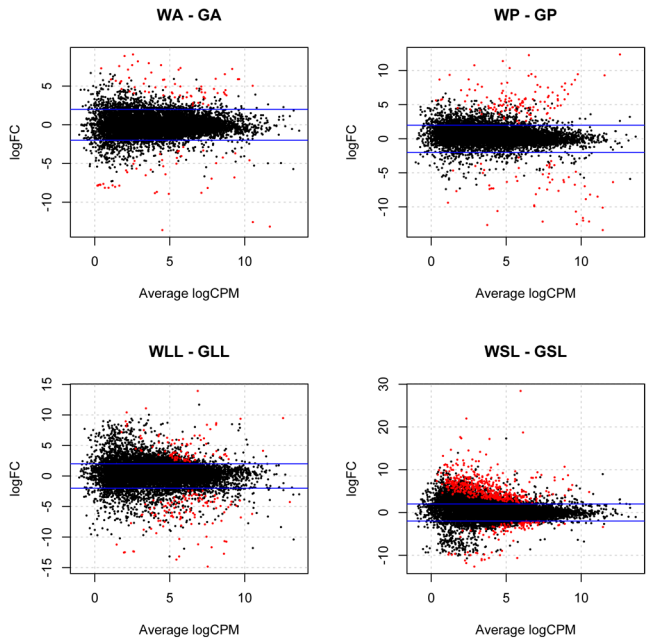

Figure S3. Smear plots showing the relationship between expression level (log copies per million, CPM) and the level of differential expression (log fold change, FC) for comparisons between workers (W) and queens (G – for gyne) of *P. barbatus* in four developmental stages: A = adults, P = pupal, LL = large larval, SL = small larval. Each point represents a gene, with red genes having differential expression at FDR < 0.05. The horizontal blue lines represent a four-fold difference in expression either up (worker bias) or down (queen bias).

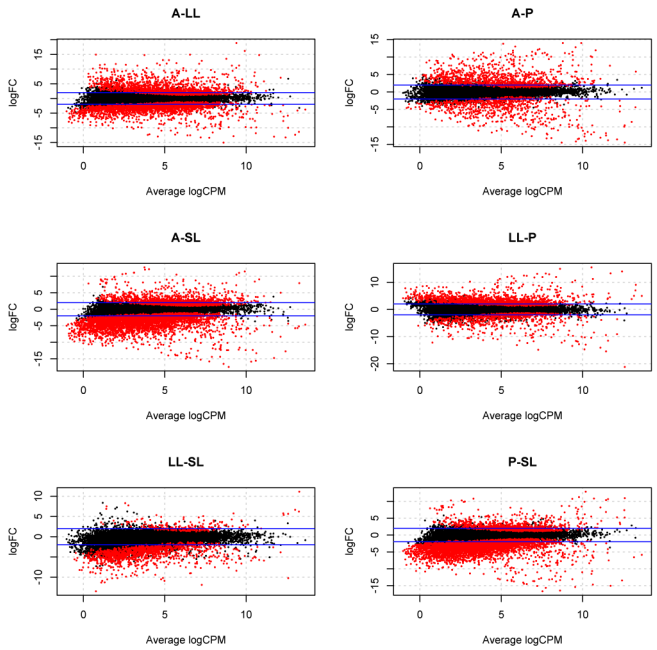

Figure S4. Smear plots showing the relationship between expression level (log copies per million, CPM) and the level of differential expression (log fold change, FC) for comparisons between developmental stages of *P. barbatus*: A = adults, P = pupal, LL = large larval, SL = small larval. Each point represents a gene, with red having differential expression at FDR < 0.05. The horizontal blue lines represent a four-fold difference in expression (the first stage listed is up in the plot, e.g., A-SL means that positive logFC values are genes up-regulated in adults).

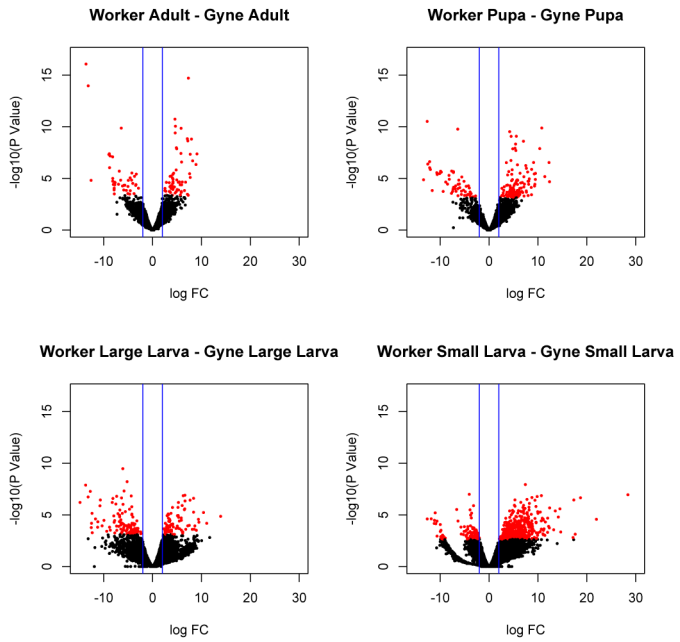

Figure S5. "Volcano" plots (expression fold change by statistical significance) for worker to queen (written as "gyne" on the figure as they are virgin queens) caste comparisons across four developmental stages. Each dot is a gene and those highlighted in red are differentially expressed at FDR < 0.05. The vertical blue lines represent a fold change in expression of +/- 4 fold; the direction of fold change is represented in the title of each graph (positive values represent genes upregulated in workers).

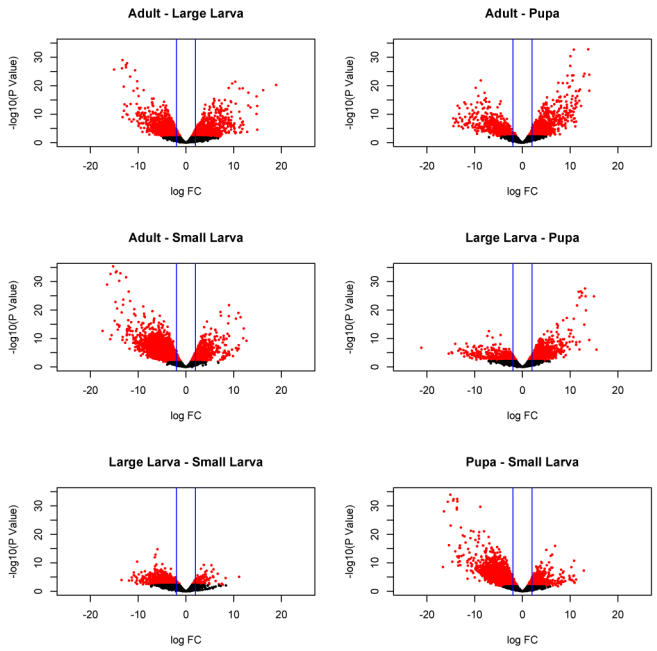

Figure S6. "Volcano" plots (expression fold change by statistical significance) for comparisons among four developmental stages. Each dot is a gene and those highlighted in red are differentially expressed at FDR < 0.05. The vertical blue lines represent a fold change in expression of +/- 4 fold; the direction of fold change is represented in the title of each graph (positive values represent genes upregulated in the first stage in the title, e.g., in Adults for "Adults – Large Larva").

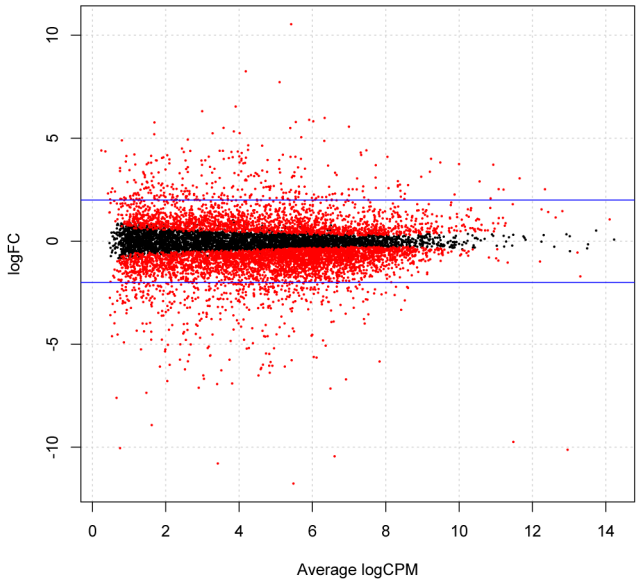

Figure S7. Smear plot showing the relationship between expression level (log copies per million, CPM) and the level of differential expression (log fold change, FC) between worker and queen adults of *Vollenhovia emeryi* (positive logFC values are genes upregulated in workers). Each point represents a gene, with red having differential expression at FDR < 0.05. The horizontal blue lines represent a four-fold difference in expression.

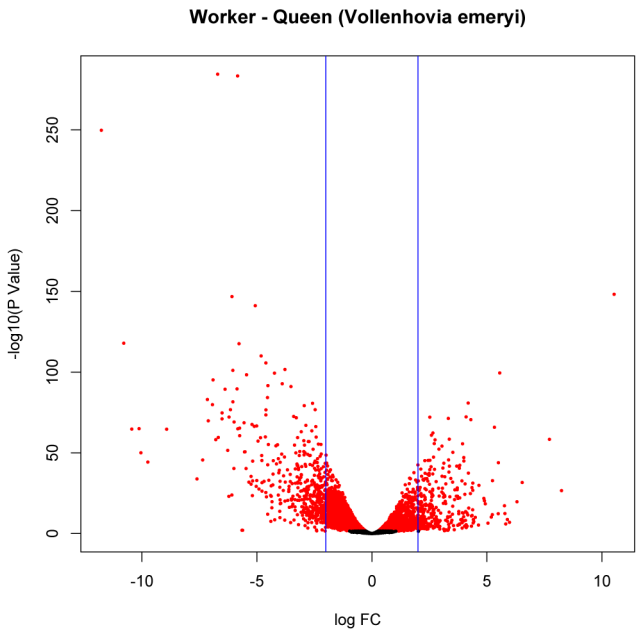

Figure S8. "Volcano" plot (expression fold change by statistical significance) comparing worker and queen differential gene expression of *V. emeryi* (positive

logFC values are genes upregulated in workers). Each dot is a gene and those highlighted in red are differentially expressed at FDR < 0.05.

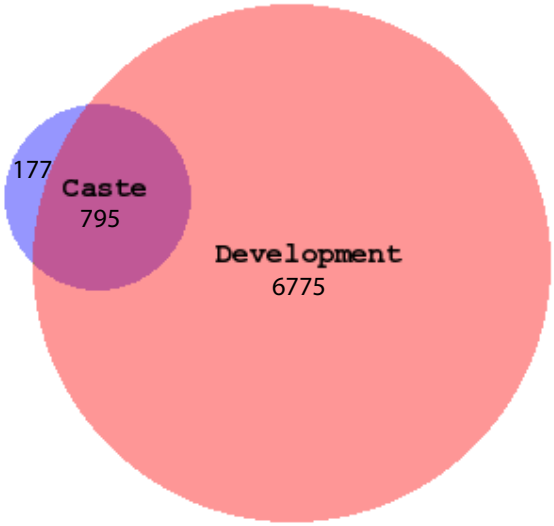

Figure S9. Venn diagram of overlapping gene sets using the more liberal analysis of *P. barbatus* RNAseq data. The sets are, genes differentially expressed between developmental stages (red) and those differentially expressed between castes (within developmental stages, blue). Analysis done using BioVenn (Hulsen et al. 2008).

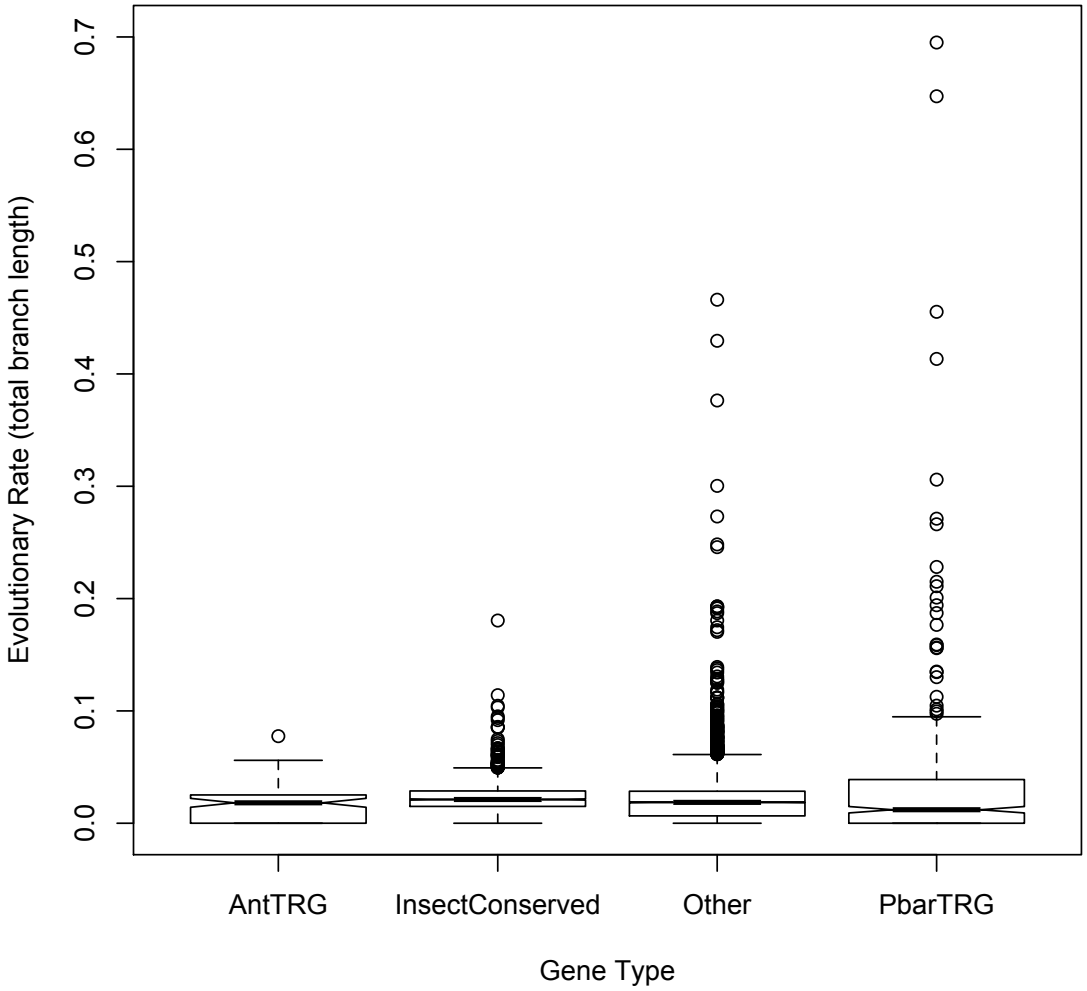

Figure S10. Boxplot of the evolutionary rate (total branch length) of orphan (taxonomically restricted genes, TRG) for both ants and *P. barbatus* compared to non-orphan genes (highly conserved genes across insects - InsectConserved, and other genes of intermediate conservation - Other). Orphan genes are evolving at a faster rate than the other categories, $F_{3,9594} = 16.83$, $P < 0.0001$), though this seems largely driven by extreme values. In post-hoc comparisons using Tukey HSD, P. barbatus TRGs were significantly greater ($P < 0.005$ in all comparisons) than all other groups, but other groups were not significantly different.

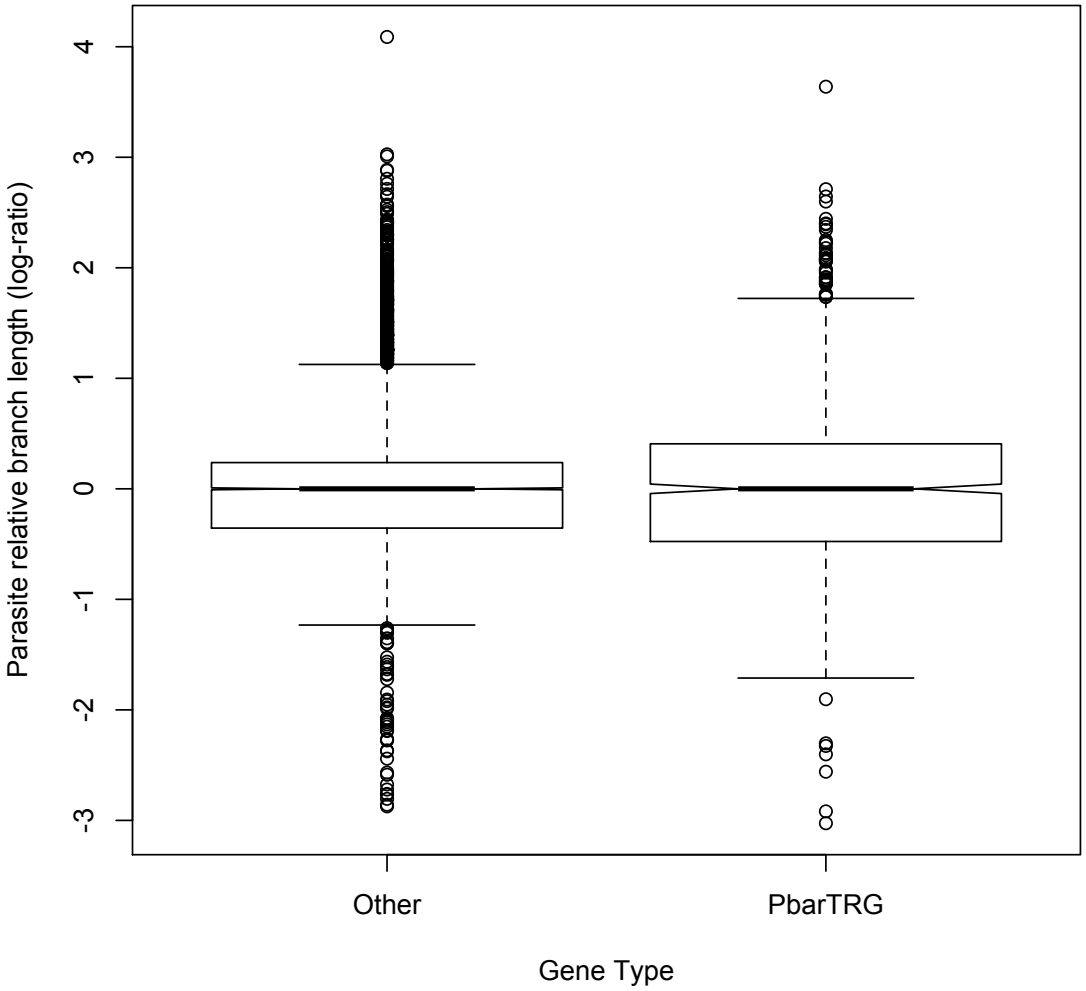

Figure S11. Boxplot of host-parasite relative branch lengths between *P. barbatus* taxonomically restricted "orphan" genes (PbarTRG) and other genes (Other). Taxonomically restricted genes are not evolving at a significantly faster rate than other genes in the social parasites compared to their hosts (t = -1.2, df = 1170, P = 0.3).

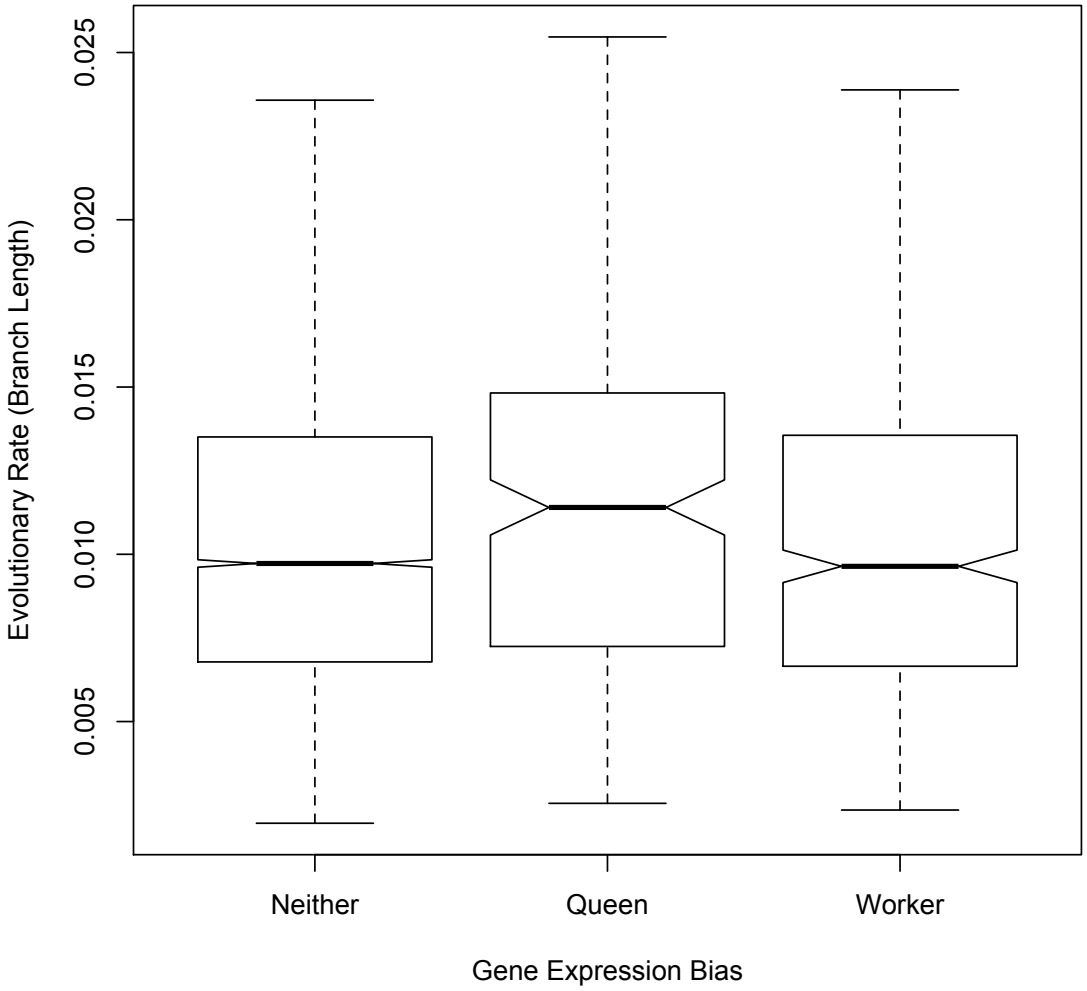

Figure S12. Boxplot of the evolutionary rate (total branch length) compared across genes with significantly greater expression in workers queens, or that were not differentially expressed (Neither). Genes in each category did not differ in evolutionary rate. Outliers are not shown in order to better visualize group differences (ANOVA: $F_{2,9678} = 2.1$, $P = 0.12$).

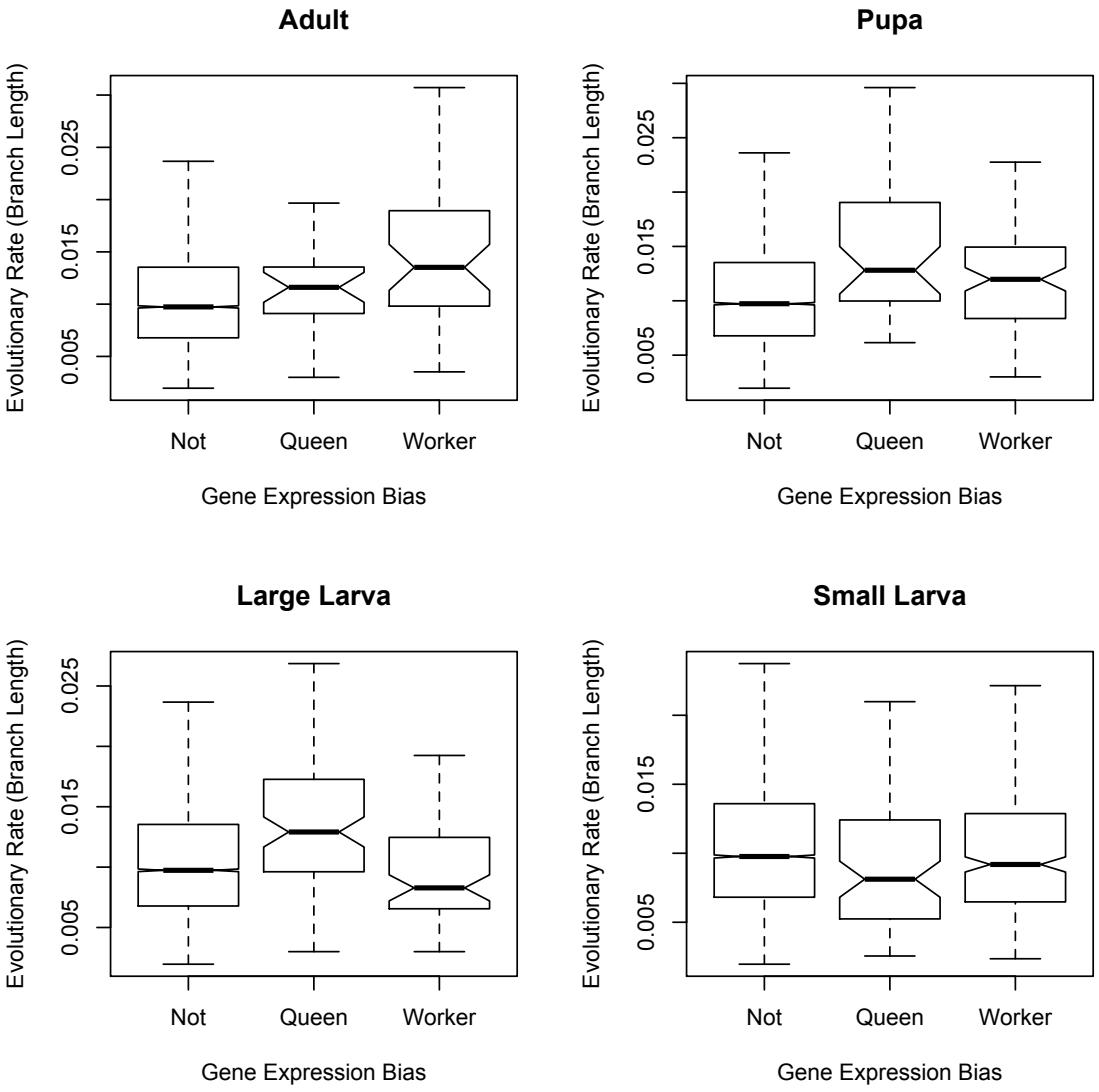

Figure S13. Boxplots of branch length (evolutionary rate) variation for differentially expressed genes with greater expression in queens and workers, or that are not differentially expressed (Not), across four developmental stages. Whether a category of gene was associated with an elevated evolutionary rate was dependent on developmental stage. For example, worker up-regulated genes evolved faster than non-differentially expressed genes in the adult stage ($P < 0.001$), but queen up-regulated genes evolved fastest in the pupal and large larval stages ($P < 0.05$ and $P < 0.001$, respectively); no groups were different among small larvae. These data suggest that differences in differential expression of genes across development may skew estimates of evolutionary rate. P-values in all comparisons are from ANOVA, $F_{2,9762}$, using Tukeys HSD for pair-wise comparisons.

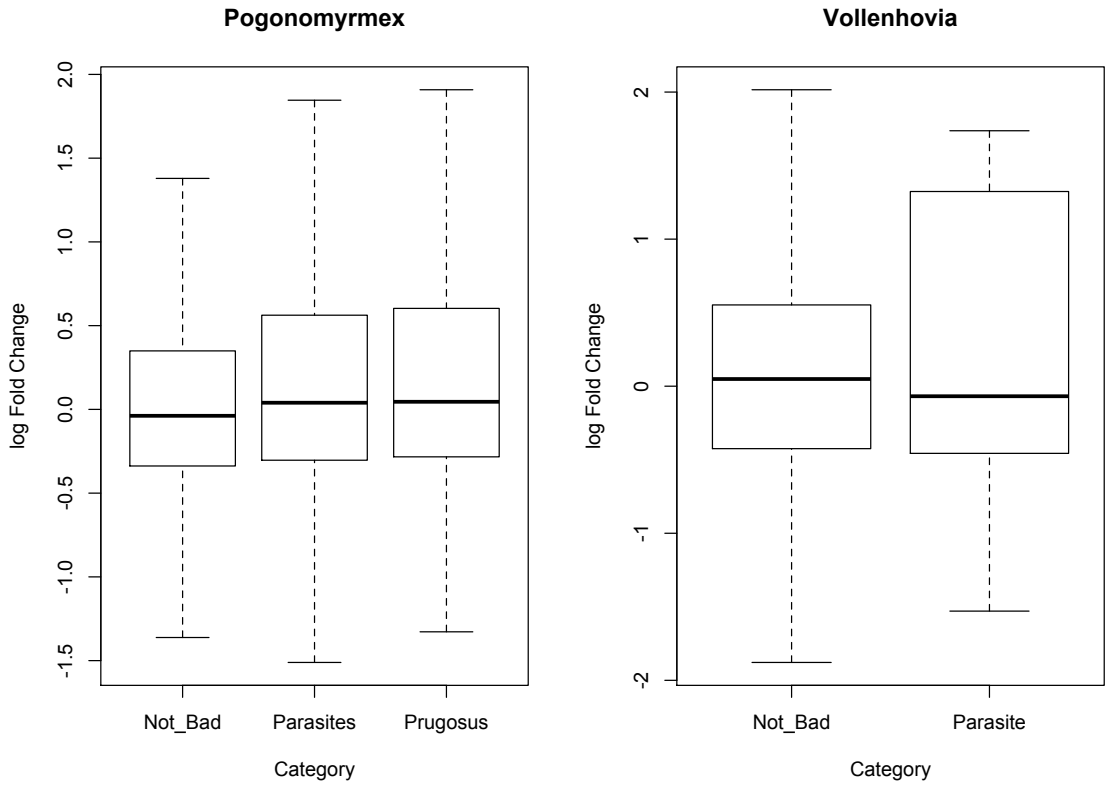

Figure S14. Boxplots of differential gene expression (log fold change, positive values represent worker bias) in host (*P. barbatus* and *V. emeryi*) genes with at least one loss of function mutation in the social parasites (left: *P. colei* or *P. anergismus*, right: *V. nipponica*) or the second host, *P. rugosus*, compared to genes without loss of function mutations (Not Bad). Taken together, genes with loss of function mutations were more worker-biased than normal genes in *Pogonomyrmex* (ANOVA: $F_{2,8561}$ = 4.4, P < 0.05). Genes with lost function in the *Pogonomyrmex* parasites were marginally more worker-biased in expression compared to genes without loss of function mutations (ANOVA: $F_{2,8561}$ = 2.9, P = 0.055, Tukey HSD: P = 0.057); however, there was no difference in expression bias between the parasites and *P. rugosus* (P = 0.45). There was also no detectable difference in logFC between normal genes and genes with lost function in *Vollenhovia* (t = 0.91, df = 21, P = 0.37). Outliers are not shown in order to better visualize group differences.

References

Knapek S, Sigrist S, Tanimoto H (2011) Bruchpilot, a synaptic active zone protein for anesthesia-resistant memory. *J Neurosci Off J Soc Neurosci* 31(9):3453–3458.

Hulsen T, J de Vlieg and W Alkema. 2008. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics 9:488.