Supplementary Information Text

SI. 1. Restriction Enzyme Choice

Although there were no published genomes for any *Eciton* species during the time of experimental design, eight freely-available ant genomes were downloaded from the Hymenopteran Database to choose the most ideal restriction enzyme for genotyping-by-sequencing (Elsik et al., 2015; Elshire et al., 2011). Previous studies utilizing the pyRAD pipeline mandated a minimum of 6x depth for any individual locus, and recommended coverage of 10x to be conservative (Eaton, 2014; Eaton, 2013; Eaton, pers. comm.). Due to the nature of sample and locus bias during GBS—thus creating a non-uniform distribution of read depth for both samples and loci—we aimed for more than triple the minimum coverage (~20x) across all samples and loci for restriction enzyme choice. Previous work of colleagues using GBS suggested that size-selection much improved sequence quality and locus capture, with an ideal size range of 300-800bp for 100bp sequencing (Winger et al., 2015; Eaton, 2013). Runs of 100bp sequencing were planned to be multiplexed with 50 individuals on an Illumina HiSeq2000, with a conservatively estimated 220 million reads.

Given these assumed parameters, *in silico* digest of the eight ant genomes with 280 potential restriction enzymes was performed, using custom python and R scripts. The vast majority of potential restriction enzymes were eliminated from consideration due to inappropriate cutting frequencies for the estimated 263.9Mb *E. burchellii* genome (Tsutsui et al., 2008). Among the remaining restriction enzymes *ApeK1* was selected as it maximized the theoretical number of loci captured while still providing ample coverage (**Table S1**). Specifically, with an estimated 220 million reads across 48 individuals, one run of Illumina sequencing could provide enough coverage for 229,100 loci with a target 20x depth. Across the eight ant genomes, *ApeK1 in silico* digests suggested a mean number of 221,540 loci (σ = 68,620), approximately equal to the targeted number of loci for our given coverage and sample multiplexing scheme. Coincidentally, the *E. burchellii* genome size estimate from flow cytometry (263.9 Mb [SE = 2.1pg, n = 4]), matched the mean assembled genome size for the 8 published ant genomes (**Table S1**; Tsutsui et al., 2008).

SI. 2. *In Silico* Digestion

While the choice of *ApeK1* for digestion within our GBS protocol was made in the absence of *Eciton* genomic sequences, recent sequencing of the *E. burchellii* genome (McKenzie et al., unpubl.) allows direct *in silico* digestion, facilitating comparisons of the theoretical maximum number of loci to actual loci harvested with GBS. Digestion of the 189.7Mb draft genome resulted in 552,700 fragments, with 124,910 fragments between the sizes of 300 – 800bp, translating to a theoretical maximum of 249,800 loci for GBS locus capture, since each fragment has two ends that can each be assembled into a different locus. Although the assembled genome size is smaller on average in comparison to the eight published ant genomes (**Table S1**), the cut frequency (0.0029) and proportion of fragments in the desired size range (0.226) were higher, resulting in a number of loci slightly greater (~250k) than the mean for the other genomes (~221k).

SI. 3. GBS Library Preparation and Locus Assembly

The low cost and scale of data offered by high-throughput sequencing approaches present an ideal opportunity for phylogenomics in non-model organisms (McCormack et al., 2013, Rowe et al., 2011, Wagner et al., 2012). Specifically, reduced-representation sequencing methods such as genotyping-by-sequencing (GBS) allow for highly cost-effective locus capture (Elshire et al., 2011; Hohenlohe et al., 2011), and recent computational advances have enabled accurate locus assembly without a reference genome (Eaton, 2014). In addition, *in silico* analyses have verified the utility of GBS for resolving clades younger than 60 Ma, demonstrating robust and accurate phylogenetic inference (Rubin et al., 2012). Empirically, several recent studies employing GBS for resolving difficult phylogenies have succeeded, further suggesting the efficacy of the method (Eaton & Ree, 2013).

DNA from all specimens was extracted using the Qiagen DNeasy Blood & Tissue Kit with a modified, optimized protocol for ants and then dried down (Moreau, 2014). Extractions were then quantified in triplicate using the Invitrogen Qubit 2.0 Fluorometer, with a minimum total DNA amount of 200ng necessary for inclusion in library preparation. All samples were standardized to 200ng. Size-selection was achieved using a high-melt agarose gel and the Qiagen gel extraction kit, followed by analysis of the size-selected library with the Agilent 2100 Bioanalyzer at the Field Museum of Natural History. All sequencing was performed on the Illumina HiSeq 2000 at the Center for Genome Research and Biocomputing at Oregon State University.

Since all sequencing runs were multiplexed with 50 individuals, the first step in assembly was to demultiplex the raw Illumina reads using cut-site matching and known barcodes ligated during the GBS protocol [*pyRAD step 1*]. All reads that failed to match an ApeK1 cut-site or known barcode were discarded. Next, the demultiplexed reads were filtered based on Phred quality scores and presence of adapter sequences [*pyRAD step 2*]. All reads containing misplaced adapter sequences or more than 4 sites with Phred scores 20 or lower were discarded from further analysis. After quality filtering, all reads for each sample were de-replicated and clustered with a sequence similarity threshold of 0.9 [*pyRAD step 3*]. Clustering is accomplished with a global alignment clustering algorithm that first assembles draft loci (clusters) that are then filtered in downstream processing (Edgar, 2004; Edgar, 2010; Rognes, 2016). All clusters with read depths under 10 in a given sample were discarded from downstream processing. Joint inference of error-rate (E) and heterozygosity (π) was then performed with the remaining clusters across samples [*pyRAD step 4*]. Utilizing these estimates of error-rate and heterozygosity, we then applied rigorous paralog filters and called consensus sequences for each cluster [*pyRAD step 5*]. Consensus sequences were then clustered across samples using the same pyRAD clustering algorithm, which formed the set of final loci used for downstream analyses [*pyRAD step 6*]. The final set of loci was subset by (i) the minimum number of samples for which a locus genotype was available, (ii) maximum number of individuals with shared mutation in the locus, and (iii) the exclusion of desired taxa, depending on the analysis [*pyRAD step 7*].

Illumina HiSeq 2000 sequencing yielded 674.1 million raw reads (66.0 Gb), which were then demultiplexed by the 150 samples, resulting in 634.4 million reads (62.1 Gb: *94.1% of raw reads*) that matched both the appropriate *ApeK1* cut-site and one of the possible barcodes. Three samples were dropped due to insufficient read coverage, resulting in 147 samples with sufficient coverage. Filtering of the demultiplexed reads for Phred quality, read length, and adapter sequence resulted in 441.3 million utilizable reads for clustering (43.2 Gb: *69.6% of demultiplexed reads*) under the strict filters, with a mean 3.0 million reads per individual [μ = 3.00M; σ = 1.79M; n=147]. Within-sample cluster formation with the *pyRAD* algorithm yielded a mean 277k clusters per sample [μ = 276,837; σ = 78,616; n=147], which was reduced to a mean 84k clusters per sample [μ = 84,083; σ = 42,602; n=147] with a minimum of 10x coverage, utilizing 322.3 million of the initial reads for clustering [73.0%]. Because the algorithm operates without a reference (*de novo*), decent coverage (>5x) is recommended for confidence in locus assembly (Eaton, 2014). Joint inference using 10x clusters for each sample provided estimates of heterozygosity [$\boldsymbol{\pi}$: μ = 0.0049; median = 0.0023; σ = 0.0062; n=144] and error rate [**E**: μ = 0.0018; median = 0.0013; σ = 0.0013; n=144]. Estimates were then utilized for paralog filtering and calling consensus sequences, which were clustered across samples to produce the final set of loci (**Fig. S1**). Across all samples, the mean number of final loci was highly variable [μ = 58,095; median = 54,550; σ = 30,950; n=147]. This finding is not surprising as sample bias in library preparation is common, resulting in insufficient coverage in some samples for locus assembly. Locus assembly statistics robustly support this hypothesis, with a highly significant correlation between log number of reads and log final loci across samples [r_{adj}^2 = 0.689; p < 2.2e-16] (**Fig. S1**). To identify the sources of variation in the number of shared loci between samples, we developed an effective linear model using data from locus assembly and phylogenetic distance [N = 10,731; r = 0.91, p = 0] (**Fig. S1**).

SI. 4. Phylogenomic Inference

Method Details

Maximum likelihood inference was accomplished using Randomized Axelerated Maximum Likelihood with High Performance Computing (RAxML-HPC) pipeline, a highly optimized program that can exploit both fine-grained and coarse-grained parallelism (Stamatakis, 2006). Unlike most phylogenetic pipelines, RAxML-HPC can easily handle extremely large datasets, such as those created by GBS. Given the known rate heterogeneity across reduced representation loci, our concatenated dataset was run with a general time-reversible model of nucleotide substitution and a gamma model for rate heterogeneity (GTR-GAMMA) (Yang, 1993; Yang & Rannala, 2012). In order to assess the statistical robustness of the ML tree 100 rapid bootstrap trees were also inferred using different subsets of the larger dataset.

Bayesian inference was accomplished with the *ExaBayes* pipeline, which implements a Markov Chain Monte Carlo (MCMC) approach that simultaneously estimates the posterior probability of a phylogeny and various evolutionary model parameters (Aberer et al., 2014). Similar approaches are implemented in *BEAST* (Drummond et al., 2012) and *MrBayes* (Ronquist et al., 2012), however, the

parallelizable and computationally-efficient approach offered by *ExaBayes* makes it an ideal pipeline for large phylogenomic datasets such as the one attained in this study. Two independent chains were run with a GTR-GAMMA model of evolution, each for a minimum of one million generations, checking for convergence every 5,000 runs, sampling every 500 generations, and tuning parameters every 100 generations. Parsimony trees were used in the initialization of the independent runs. Convergence criteria were normality and stability of parameter estimation as well as consistency at known interspecific nodes. Burn-in proportion was 0.25, and each run had an even Dirichlet prior for all nucleotides.

Result Details

BI and ML inference converged on the same tree (**Fig. S2**), which agreed with many previous studies at several levels of evolutionary scale (**Fig. S3**). The ML tree search on the alignment of 147 individuals with 6,700,494 distinct nucleotide sites with RAxML-HPC took 485.8 hours, including 26.8 hours for model parameter optimization (**Fig. S2**). The two independent Bayesian MCMC chains were each run for 1,000,000 generations using Exabayes, although topological and parameter convergence was reached after 60,000 generations (**Fig. S4**). Lastly, one technical replicate was included in the study as an independent check of method efficacy. In all phylogenies the technical replicate was inferred to be sister to its replicate sample, and was discarded for future analyses.

Based on the initial traversal time (144.7s), Subtree Equality Vector (SEV) likelihood implementation was used for the ML tree search. The average time for each rapid bootstrap was 3.43 hours, totaling 343.5 hours for all 100 bootstraps (**Fig. S2**). The proportion of gaps and completely undetermined characters in the aligned matrix was 87.23%.

Estimated parameter means and standard deviations for each independent Exabayes MCMC run were: Ln Pr *(μ_1 = 668.0, σ_1 = 0.006; μ_2 = 668.0, σ_2 = 0.004*), Ln L *(μ_1 = -7.1e7, σ_1 = 131; μ_2 = -7.1e7, σ_2 = 106*), TL *(μ_1 = 0.331, σ_1 = 5.6e-4; μ_2 = 0.331, σ_2 = 3.6e-4*), α *(μ_1 = 0.100, σ_1 = 9.9e-4; μ_2 = 0.100, σ_2 = 9.8e-4*), r_{AC} *(μ_1 = 0.054, σ_1 = 1.5e-4; μ_2 = 0.054, σ_2 = 1.6e-4*), r_{AG} *(μ_1 = 0.359, σ_1 = 3.3e-4; μ_2 = 0.359, σ_2 = 3.3e-4*), r_{AT} *(μ_1 = 0.059, σ_1 = 1.7e-4; μ_2 = 0.058, σ_2 = 1.5e-4*), r_{CG} *(μ_1 = 0.110, σ_1 = 2.2e-4; μ_2 = 0.110, σ_2 = 2.2e-4*), r_{CT} *(μ_1 = 0.360, σ_1 = 3.3e-4; μ_2 = 0.360, σ_2 = 3.4e-4*), r_{GT} *(μ_1 = 0.059, σ_1 = 2.3e-4; μ_2 = 0.059, σ_2 = 2.1e-4*), π_A *(μ_1 = 0.239, σ_1 = 7.9e-5; μ_2 = 0.239, σ_2 = 6.4e-5*), π_C *(μ_1 = 0.261, σ_1 = 7.9e-5; μ_2 = 0.261, σ_2 = 6.3e-5*), π_G *(μ_1 = 0.267, σ_1 = 7.5e-5; μ_2 = 0.267, σ_2 = 6.1e-5*), π_T *(μ_1 = 0.233, σ_1 = 6.0e-5; μ_2 = 0.233, σ_2 = 5.9e-5*). Average deviation of split frequencies was 6.23%, which is extremely low considering the sizeable number of population-level samples included in the study.

SI. 5. Population Genomic Inference

In order to avoid issues of linkage disequilibrium, we used the ".unlinked_snps" output from pyRAD for population genomic inference, which samples a single SNP from each locus. We then wrote custom scripts to employ a modified Wright's F_{ST} estimator on a mean of 29,370 loci (σ = 11,327) for each of the five parapatric lineage pairs on the unlinked SNPs (**Table S2**), weighted by population sample size,

so that the estimator could provide comparable estimates across our loci as defined in Chen et al. (2015). Subsequently, to test population genomic differentiation on a more conservative set of loci, we limited our analysis to variable loci with high nucleotide diversity ($\pi > 0.8$) and high sample coverage ($n > 9$). The significant results demonstrated with all variable loci were only stronger and more significant with the conservative set of loci (**Table S2**; **Fig. S5**).

SI. 6. Tree Dating

We used a penalized likelihood approach for divergence dating and for estimating absolute rates of molecular evolution, known as non-parametric rate smoothing (NPRS), paired with a two-fold cross-validation step for the optimal choice of the smoothing parameter (λ). The two-fold cross-validation step resulted in a clear global minimum ($\lambda = 0.35$), which was then applied to NPRS with the 95% credible interval (CI) derived from the marginal posterior probability for generic nodes from Brady et al. (2014) for prior calibrations and external validation of our estimates (**Fig. S3**). Specifically, using the most recent common ancestor (MRCA) of Neotropical army ants from Brady et al. (2014) as our fixed prior calibration, we estimated distributions for the MRCA of *Eciton* and the other Neotropical army ant genera for comparison with the remaining prior distributions from the same paper (**Fig. S3**). All of our estimated MRCA ages were within the original confidence intervals of Brady et al. (2014), suggesting general concordance across studies. Divergence date distributions were estimated for all five pairs of sister lineages across the IP (**Fig. 2; Fig. S3**).

SI. 7. Shared Loci Model and Locus Sparsity

Reduced representation sequencing methods such as GBS produce various degrees of locus bias, which we define here as the non-uniform distribution of reads across assembled loci. Along with other factors such as phylogenetic distance and sample bias, locus bias frequently results in sparse data matrices, where each sample shares some subset of the total number of assembled loci with other samples in the dataset (**Fig. S1**). Although it is possible to restrict the dataset to markers with full coverage, recent *in silico* work has demonstrated substantial pitfalls of this method (Huang & Knowles, 2014), and thus *total evidence* is preferable for phylogenomic studies. Due to this body of research, we implemented a "total-evidence" approach for phylogenetic inference, using a larger, sparser matrix instead of dropping loci for a more complete matrix.

To understand the relative influences of phylogenetic, sample, and locus bias, we used locus assembly data and several custom scripts to build a linear model of pairwise shared loci for all sequenced samples. First, we assembled a matrix (n x n; n = 147) containing the number of pairwise shared loci for all sequenced samples from the *pyRAD.loci* file [$\mu = 17{,}465$, *median* $= 14{,}219$, $\sigma = 12{,}475$, $N = 10{,}731$] (**Fig. S1**). Next, using the total number of reads passing quality filters for each sample, we modeled the log-transformed number of shared loci for any two samples with the log-root of the product of the filtered reads for each sample as the primary independent variable [N = 10,731; r = 0.71, p = 0]. After observing locus drop out due to phylogenetic distance, we included phylogenetic distance as a second

independent variable, which greatly improved our linear model [N = 10,731; r = 0.88, p = 0] (**Fig. S1**). Remarkably, this linear model, which was based solely on the total amount of data pre-clustering, was almost as predictive as a similar model we produced using the log-product of total final loci for each sample as the primary independent variable [N = 10,731; r = 0.91, p = 0].

SI. 8. Biogeographic Inference

All samples sequenced in this study were provided with GPS coordinates (**Table S3**). Distinct phylogenetic lineages were then mapped to an adaption of the biogeographic areas of Morrone (2006) with the geographic coordinates associated with the samples of that lineage. To deal with the sparse geographic sampling in certain areas and the high resolution of the biogeographic areas from Morrone (2006) across the Neotropics, any areas that lacked substantial occurrence data were merged into broader biogeographic areas to facilitate visualization. Thus, the biogeographic areas featured in this study are an adaptation of Morrone (2006), and not an exact replicate of the areas *sensu stricto*. Although in many cases distinct sister lineages are known to have parapatric overlap along the borders of some biogeographic areas (Watkins, 1976), for visualization these overlapping areas were treated as borders for biogeographic inference in the main text (**Fig. 1**). Mapping of our samples on Neotropical geography with their identity defined by the inferred clade from phylogenomic inference demonstrates the magnitude of broad sampling achieved by this study and the strength of our results (**Fig. S6**).

Existing geological data on the development of the Central American landmass provide a probable scenario for dispersal from South America (Coates et al., 2004; Gutierrez-Garcia & Vazquez-Dominguez, 2013). Beginning around 12 Ma, the volcanic archipelago now known as the Chorotega block was forming due to the Central American Arc Collision (CAAC), growing in area and rising in elevation until completion of the CAAC by 7.1 Ma (Coates et al., 2004). At the time of completion, the shallow depths of the CAAC are known to have extended to the region between South and Central America, and it is likely that ephemeral connections would have provided the necessary dispersal route (Coates et al., 2004; Gutierrez-Garcia & Vazquez-Dominguez, 2013). Subsequently, an area known as the Boca break on the developing volcanic landmass was subsiding, leading to marine transgressions beginning ~ 5 Ma that would have separated any dispersing lineages from their source population (Gutierrez-Garcia & Vazquez-Dominguez, 2013). Since the marine transgressions lasted for over 3 Ma, we suggest that this barrier along with other barriers in this complex, fragmented landscape provided enough reproductive isolation to inhibit mating upon the eventual secondary contact after full closure of the Isthmus of Panamá.

Interestingly, there are other reasons to support an ephemeral land connection besides concordance with the established geological record. First, the timing of this event would have been at the peak of the CAAC, a time when the likelihood for massive tectonic change in geologically short time scales is at its highest (Coates et al., 2004). Evidence from other complex regions such as the Sunda Shelf has shown that similarly complex subduction zones can result in rapid vertical movements that might provide temporary terranes for dispersal and be

hard to detect with standard tectonic models (Hall, 2012). Second, the time and region in question is part of a large unconformity where sedimentary data and geological history is lacking, suggesting that lack of evidence for a land connection could likely be a false negative due to weak geological resolution (Coates et al., 2004). Lastly, and perhaps the reason with the largest implications, is that recent work featuring reanalysis of over 420 studies across several plant and animal phyla has shown that rate shifts in migration peaked much earlier than the traditional timing of 3 Ma. In fact, the most dramatic shifts in migration rates match the timing of these events associated with the CAAC at 8.0 Ma (0.036 to 0.142 migrations/Ma) and at 5.1 Ma (0.142 to 0.371 migrations/Ma), and moreover, it was during this time that these rates became asymmetric in favor of migration from South America into Central America (Bacon et al., 2015). We would expect these results under the suggested scenario, as a temporary CAAC land bridge between South America and the Chorotega block followed by marine transgressions would allow terrestrial South American lineages to colonize North America far before the reverse migration occurred during the traditional GABI.

Custom Scripts

All custom scripting for this project can be found on the first author's GitHub repository (https://github.com/mewinsto).

SI References

Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era. *Mol Biol Evol* 1–8. doi:10.1093/molbev/msu236

Bacon CD, et al. (2015) Biological evidence supports an early and complex emergence of the Isthmus of Panama. *P Natl Acad Sci USA* 112(19). doi:10.1073/pnas.1423853112

Bonasio R, et al. (2010) Genomic comparison of the ants Camponotus floridanus and Harpegnathos saltator. *Science 329*(5995): 1068–1071.

Brady SG, Fisher BL, Schultz TR, Ward PS (2014) The rise of army ants and their relatives : diversification of specialized predatory doryline ants. *BMC Evol Bio* 14(1): 1–14. doi:10.1186/1471-2148-14-93.

Chen G, et al. (2015) An Improved F(st) Estimator. *PloS one* 10(8): e0135368. doi:10.1371/journal.pone.0135368

Coates AG, Collins LS, Aubry MP, Berggren WA (2004) The Geology of the Darien, Panama, and the late Miocene-Pliocene collision of the Panama arc with northwestern South America. *Geol Soc Am Bull* 116(11): 1327. doi:10.1130/B25275.1

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol 29*(8): 1969–73. doi:10.1093/molbev/mss075

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30(13): 1844–9. doi:10.1093/bioinformatics/btu121

Eaton DAR, Ree RH (2013) Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Syst Biology* 62(5): 689–706. doi:10.5061/dryad.bn281

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792-1797.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460-2461.

Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE (2015) Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Research*

Elshire RJ, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* 6(5): 1937-9. doi:10.1371/journal.pone.0019379

Gutiérrez-García TA, Vázquez-Domínguez E (2013) Consensus between genes and stones in the biogeographic and evolutionary history of Central America. *Quaternary Res* 79(3): 311–324. doi:10.1016/j.yqres.2012.12.007

Hall R (2012) Sundaland and Wallacea: geology, plate tectonics, and palaeogeography. *Biotic Evolution and Environmental Change in Southeast Asia*.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol Ecology Res* 11(1): 117–22. doi:10.1111/j.1755-0998.2010.02967.x

Huang H, Knowles LL (2014) Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Syst Biology*, *0*(0), 1–9. doi:10.1093/sysbio/syu046

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol* 66(2): 526–38. doi:10.1016/j.ympev.2011.12.007

Moreau CS (2014) A practical guide to DNA extraction, PCR, and gene-based DNA sequencing in insects. *Halteres* 5: 32–42.

Morrone JJ (2006) Biogeographic areas and transition zones of Latin America and the Caribbean islands based on panbiogeographic and cladistic analyses of the entomofauna. *Annu Rev Entomol* 51(125): 467–94. doi:10.1146/annurev.ento.50.071803.130447

Nygaard S, et al. (2011) The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. *Genome Res* doi:10.1101/gr.121392.111

Rognes T (2016) VSEARCH: Github repository. https://github.com/torognes/vsearch.

Ronquist F, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biology 61*(3): 539–42. doi:10.1093/sysbio/sys029

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Mol Ecol* 20(17): 3499–502. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21991593.

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PloS one* 7(4): e33394. doi:10.1371/journal.pone.0033394

Schrader L, et al. (2014) Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nat Comm 5*: 5495. doi:10.1038/ncomms6495

Smith CD, et al. (2011) Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). *P Natl Acad Sci USA* 108(14): 5673–8. doi:10.1073/pnas.1008617108

Smith CR, et al (2011) Draft genome of the red harvester ant Pogonomyrmex barbatus. *P Natl Acad Sci USA* 108(14): 5667–72. doi:10.1073/pnas.1007901108

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21): 2688–90. doi:10.1093/bioinformatics/btl446

Suen G, et al. (2011) The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. *PLoS Genetics*, *7*(2), e1002007. doi:10.1371/journal.pgen.1002007

Tsutsui ND, Suarez AV, Spagna JC, Johnston JS (2008) The evolution of genome size in ants. *BMC Evol Bio 8*, 64. doi:10.1186/1471-2148-8-64

Wagner CE, et al. (2012) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol*. doi:10.1111/mec.12023

Watkins JF (1976) *The identification and distribution of New World Army Ants*. (pp. 1–109).

Winger BM, et al. (2015) Inferring speciation history in the Andes with reduced-representation sequence data: an example in the bay-backed antpittas (Aves; Grallariidae; Grallaria hypoleuca s. l.). *Mol Ecol 24*(24), 6256–6277. doi:10.1111/mec.13477

Wurm Y, et al (2011) The genome of the fire ant Solenopsis invicta. *P Natl Acad Sci USA* 108(14): 5679–84. doi:10.1073/pnas.1009690108

Yang Z (1993) Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites. *Mol Biol Evol 1*(6): 1396–1401.

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. *Nat Rev Gen* 13(5):303–14. doi:10.1038/nrg3186

SI Figure Legends

Fig. S1. *De novo* locus assembly and quantitative modeling. (a) *Density distribution of log number of final loci for all samples.* Dashed lines indicate minimum (15,252) and maximum (164,492) numbers of final loci. Black solid line indicates mean number of loci (58,095), while red solid line indicates theoretical maximum from *in silico* digest (250,000 loci). (b) *Density distribution of pairwise shared loci for 147 individuals.* Mode is ~8,000 and indicated by the solid black line, mean is 17,465 and indicated by the dashed line, median is 14,219 and indicated by the dotted line. Mean number of loci for each individual is 58,095. (c) *Relationship of phylogenetic distance between pairwise samples and shared loci.* Note that both the mean and maximum number of shared loci between samples is reduced with increased phylogenetic distance [N = 10,731]. (d) *Log number of final loci in comparison*

to initial log number of reads. Red line indicates theoretical maximum from *in silico* digest (250,000 loci). Black dotted line demonstrates expected values by linear model ($r_{adj}^2 = 0.689$; $p < 2.2e\text{-}16$; $a = -1.20$, $b = 0.795$). (e) *Plot of log number of loci shared by at least a given number of taxa, demonstrating distribution of 419,804 loci among 146 samples.* Plot shows both sparseness of concatenated data matrix by the proportion of loci shared among fewer number of taxa, as well as a strong log-linear relationship. Dotted lines show orders of magnitude (10, 10^2, 10^3, 10^4, 10^5). Note that among 89 taxa or more there are over 10,000 shared loci, whereas there are fewer than 10 loci shared among 137 or more samples. (f) *Linear model of log expected pairwise shared loci given phylogenetic distance and number of reads for each sample.* The expected number of pairwise shared loci were log-transformed, and number of reads was modeled as the sample-wise product of the root-transformed reads. The number of reads used for the model was after quality filtering [N = 10,731; r = 0.88; p = 0].

Fig. S2. Full phylogeny inferred from both ML and Bayesian methods with bootstrap support and BPP. The phylogeny features 146 Neotropical army ant samples (1 technical replicate removed), including 11 samples from outgroup genera (*Neivamyrmex, Cheliomyrmex, Labidus, Nomamyrmex*) and 135 *Eciton* samples. Sample tips feature the FMNH ID, sample country in parentheses, and taxon groups in brackets. Stars indicate perfect support at that node for both ML and Bayesian methods (100/100 bootstrap support and 1.0 BPP). Number of bootstraps supported at each node is out of 100 and is the first number indicated, the BPP of that node is indicated by the second node. Support values for any nodes with lower than 95/100 bootstrap support and 0.95 BPP has been omitted. Scale bar indicates phylogenetic distance in units of number of substitutions per site (0.005). Additional sample information can be found in Table S1.

Fig. S3. Tree-dating procedure using nonparametric rate smoothing (NPRS) and absolute calibration and verification using credible intervals from Brady et al. (2014). (a) *Two-fold cross-validation NPRS procedure for optimal lambda search, plotting error measurement (D_{CV}) against a range of lambda values.* The two plots demonstrate the broad-scale search for the rate-smoothing parameter (lambda) across multiple orders of magnitude ($10^{-6} < \lambda < 10^6$), and then the finer-scale search ($0.1 < \lambda < 2.0$) for a more exact value of lambda. The clear global minimum is demarcated by the dotted line in the lower plot ($\lambda = 0.35$). The equations for nonparametric rate smoothing (NPRS) and the cross-validation procedure are given on the left. (b) *Calibration and verification of divergence time distributions for generic nodes of Neotropical army ants.* Calibration for absolute dates was performed by using the posterior density for the MRCA of *Eciton* and *Neivamyrmex* from Brady et al. (2014), and validated by assessing concordance with the credible intervals from the other generic nodes.

Fig. S4. Plots of parameter estimates from converged, independent Bayesian inference MCMC chains. (a) *Trace of 14 measured parameters for two independent runs (*Run 1 = Circles, Run 2 = Crosses). (b) *Density distributions of 14 measured parameters following the 60,000 generation burn-in* (Run 1 = Red, Run 2 = Black). For mean values inferred from runs, refer to SI. 3.

Fig. S5. Genetic relationships between the Central American and South American army ant lineage pairs. (a) *Distribution of F_{ST} values for conservative sets of loci for each of the lineage pairs.* Starting at the top left, F_{ST} distributions for *E. burchellii* (N = 9,391 loci), *E. vagans* (N = 6,490), *E. lucanoides* (N = 6,397), *E. hamatum* (N = 6,804), and *E. mexicanum* (N = 13,112). Note in all cases the small proportion of loci with $F_{ST} = 0$. Due to broad geographic coverage in both lineage pairs and known population structure, many loci with F_{ST} values between 0 and 1 are to be expected without gene flow between lineages. (b) *Pairwise plots of phylogenetic and geographic distance for specimens from each lineage pair.* Each pairwise comparison between two specimens from the same regional lineage (Central or South American) are colored red, while pairwise comparisons between two specimens from different regional lineages are colored blue. Starting at the top left, number of pairwise comparisons for *E. burchellii*, *E. vagans*, *E. lucanoides*, *E. hamatum*, and *E. mexicanum*. Note in some cases (*E. burchellii*), there is known phylogenetic structure as there are two South American clades (Fig. 1), and that comparison between these two South American clades is colored in purple.

Fig. S6. Maps of clade assignments and geographic ranges for each Neotropical army ant lineage pair. The Central American clade is always colored in blue, with the South American clade colored in red (in the case of *E. burchellii* where there are two South American clades, one is colored in green). Points represent geographic locations for samples sequenced and assigned to clades from this study, and that colored areas are approximate, based on our data and known geographic ranges. Note that many of the points represent multiple samples and some sites are obscured by the large scale of the map. For three lineage pairs with known secondary contact zones in Costa Rica and Nicaragua (*E. burchellii*, *E. mexicanum*, and *E. vagans*), the contact zones are indicated by thick black lines. The known subspecies break for *E. lucanoides* is indicated by a black dashed line. For sample sizes refer to Fig. 1.

Table S1. *In silico* digest results for ApeK1 on eight published ant genomes. Genome size gives the actual size of the full genome assemblies (Suen et al., 2011; Nygaard et al., 2011; Bonasio et al., 2010; Schrader et al., 2014; Smith et al., 2011a; Smith et al., 2011b; Wurm et al., 2011). Locus proportion gives the proportion of digest fragments within the size selection range (300 – 800bp), whereas total loci was calculated by twice the number of digest fragments within the same size selection range.

Taxa	Cut Frequency	Genome Size (Mb)	Loci Proportion	Total Loci
A. cephalotes	0.0011	315.9	0.196	137,404
A. echiniator	0.0015	295.4	0.198	172,650
C. floridanus	0.0017	232.9	0.208	171,320
C. obscurior	0.0030	175.3	0.230	240,460
H. saltator	0.0038	291.3	0.169	249,828
L. humile	0.0021	217.5	0.232	375,992
P. barbatus	0.0020	233.0	0.211	216,284
S. invicta	0.0017	349.8	0.199	198,372
MEAN	**0.0021**	**263.9**	**0.205**	**221,540**
SD (σ)	**0.0009**	**58.3**	**0.020**	**62,620**

Table S2. Partitioning of genetic diversity between parapatric pairs of sister lineages. L_{total} is the total number of loci variable across the respective pair of sister lineages; L_{fixed} is the number of variable loci that are fixed between sister species; K_{fixed} is the proportion of total loci fixed between sister species [$K_{fixed} = L_{fixed}/L_{total}$]; $L_{total}(\pi, n)$ is the number of variable loci after filtering out loci with low coverage and/or nucleotide diversity; $L_{fixed}(\pi, n)$ is the number of conservative loci fixed between sister species; $K_{fixed}(\pi, n)$ is the proportion of conservative loci fixed between sister species; $L_{null}(\pi, n)$ is the number of conservative loci fixed between sister species with pseudo-null assignments; $K_{null}(\pi, n)$ is the proportion of conservative loci fixed between sister species with pseudo-null assignments. Total sample size (N) is the sum of samples in lineage 1 (N_1) and lineage 2 (N_2). (*) indicates that due to lower sample size, minimum coverage in *E. lucanoides* was decreased to six instead of ten. Bold font indicates the statistics from the more conservative set of loci used for inference.

	μ	**σ**	***burchellii***	***mexicanum***	***vagans***	***hamatum***	***lucanoides***
L_{total}	29,370	11,327	31,092	45,336	30,673	25,927	13,821
L_{fixed}	11,505	5,164	12,649	19,332	8,088	5,813	11,644
K_{fixed}	0.43	0.25	0.41	0.43	0.26	0.22	0.84

L_{total} (π, n)	**8,498**	**2,877**	**9,532**	**13,112**	**6,556**	**6,888**	**6,403**
L_{fixed} (π, n)	**4,620**	**2,015**	**5,344**	**6,685**	**2,966**	**2,081**	**5,842**
K_{fixed} (π, n)	**0.55**	**0.23**	**0.56**	**0.52**	**0.45**	**0.30**	**0.91**
L_{null} (π, n)	2.6	3.8	3	0	0	1	9
K_{null} (π, n)	0	0	0	0	0	0	0
N	23	8.5	30	22	25	29	9

Table S3. Sample information. This table contains information for: SampleID, FMNH ID, Genus, Species, Latitude, Longitude, Country, and pyRAD stats. pyRAD stats are as follows: *total* denotes total number of within-sample clusters constructed from quality reads; *dpt.me* is the mean depth across the within-sample clusters; *dpt.sd* is the standard deviation of the depth for the within-sample clusters; *d.9.total* is the total number of within-sample clusters with a minimum of 10x coverage; *d.9.me* is the mean depth for the 10x within-sample clusters; *d.9.sd* is the standard deviation for the depth of the 10x within-sample clusters; *loci* is the total number of final loci.