Spring 2017 – Epigenetics and Systems Biology
Discussion Session (Evolutionary Biology)
Michael K. Skinner – Biol 476/576
Week 15 (April 20)

Epigenetics and Evolutionary Biology

Primary Papers

1. Cunningham, et al. (2015) GBE 7:3383-96
2. Reid, et al. (2016) Science 354:1305-1308
3. Skinner, et al. (2014) Genome Biology and Evolution 6:1972-1989

Discussion

Student 10 – Ref #1 above

- What organisms and behavior involved?
- What epigenetic differences identified?
- What is the integration of epigenetics and evolution suggested?

Student 13 – Ref #2 above

- What is the experimental design?
- What molecular changes observed and adaptation response?
- How could epigenetics be involved in the adaptive response?

Student 23 – Ref #3 above

- What was the model system and experimental design?
- What epigenetic observations were provided and how might environmental epigenetics impact evolution?
- Is this a Lamarckian contribution to evolution?

The Genome and Methylome of a Beetle with Complex Social Behavior, *Nicrophorus vespilloides* (Coleoptera: Silphidae)

Christopher B. Cunningham[1,*], Lexiang Ji[2], R. Axel W. Wiberg[3], Jennifer Shelton[4], Elizabeth C. McKinney[1], Darren J. Parker[3], Richard B. Meagher[1], Kyle M. Benowitz[1], Eileen M. Roy-Zokan[1], Michael G. Ritchie[3], Susan J. Brown[4], Robert J. Schmitz[1], and Allen J. Moore[1,*]

[1]Department of Genetics, University of Georgia

[2]Institute of Bioinformatics, University of Georgia

[3]Centre for Biological Diversity, School of Biology, University of St. Andrews, Fife, United Kingdom

[4]Division of Biology & Bioinformatics Center & Arthropod Genomics Center, Kansas State University

*Corresponding author: E-mail: cbc83@uga.edu; ajmoore@uga.edu.

Accepted: October 5, 2015

Data deposition: This project has been deposited at NCBI under the BioProject accession PRJNA284849.

Abstract

Testing for conserved and novel mechanisms underlying phenotypic evolution requires a diversity of genomes available for comparison spanning multiple independent lineages. For example, complex social behavior in insects has been investigated primarily with eusocial lineages, nearly all of which are Hymenoptera. If conserved genomic influences on sociality do exist, we need data from a wider range of taxa that also vary in their levels of sociality. Here, we present the assembled and annotated genome of the subsocial beetle *Nicrophorus vespilloides*, a species long used to investigate evolutionary questions of complex social behavior. We used this genome to address two questions. First, do aspects of life history, such as using a carcass to breed, predict overlap in gene models more strongly than phylogeny? We found that the overlap in gene models was similar between *N. vespilloides* and all other insect groups regardless of life history. Second, like other insects with highly developed social behavior but unlike other beetles, does *N. vespilloides* have DNA methylation? We found strong evidence for an active DNA methylation system. The distribution of methylation was similar to other insects with exons having the most methylated CpGs. Methylation status appears highly conserved; 85% of the methylated genes in *N. vespilloides* are also methylated in the hymentopteran *Nasonia vitripennis*. The addition of this genome adds a coleopteran resource to answer questions about the evolution and mechanistic basis of sociality and to address questions about the potential role of methylation in social behavior.

Key words: burying beetle, epigenetics, parental care, sociality.

Introduction

Understanding phenotypic evolution necessitates investigating both the ultimate and proximate influences on traits; however, these investigations require the appropriate tools. Social behavior is a particularly thorny phenotype to study because of its complexity, variation, and its multilevel integration across an organism (Boake et al. 2002). In addition, social behavior also often displays unusual evolutionary dynamics arising from the genetic influences on interactions required for sociality (McGlothlin et al. 2010). Although single genes can influence behavior (Fischman et al. 2011), social behavior is often multifaceted and can reflect a complex genetic architecture (Walling et al. 2008; Mikheyev and Linksvayer 2015) including influences from epigenetic mechanisms (Cardoso et al. 2015). Genomes in particular are useful resources for evolutionary questions of social behavior because they grant access to both broad scale and fine scale details and mechanisms (Richards 2015). For social behavior, although there are multiple Hymenopteran genomes available to investigate fine scale detail, we lack sufficiently distantly related species to address broader patterns. It is therefore important to develop genomic resources for organisms that are particularly useful phenotypic models of social behavior but where genomic information is lacking.

The genomes of several social insects are now available, including eusocial species such as honey bees (Elsik et al.

2014), stingless bees (Kapheim et al. 2015), several ant species (Gadau et al. 2012), primitively eusocial species including bumble bees (Sadd et al. 2015), a sweat bee (Kocher et al. 2013) and a euscocial termite (Terrapon et al. 2014). There is an assembled and annotated genome for the African social velvet spider (Sanggaard et al. 2014). Although enormous progress has been made in identifying genes associated with the behavioral division of labor and developmental shifts in social and other behavioral tasks in eusocial insects (Zayed and Robinson 2012; Rehan and Toth 2015), and the influence of epigenetic inheritance on developmental plasticity and behavior (Glastad et al. 2015; Yan et al. 2015), the generality of any mechanism underlying social interactions requires information from insects reflecting other levels of sociality and from other orders. Sociality occurs in nearly every insect order (Wilson 1971; Costa 2006), with eusociality representing an extreme on a social continuum. Outside Hymenoptera there are many subsocial species that have highly developed social behaviors, including parental care, but no division of labor (Costa 2006). To begin to address this gap, we assembled and annotated the genome of *Nicrophorus vespilloides*, a subsocial beetle that serves as a behavioral model species for many types of complex social interactions, including elaborate and advanced parental care with direct regurgitation of food to begging offspring (Walling et al. 2008<AQ7>), parent–offspring conflict (Kilner and Hinde 2012), sibling competition (Smiseth et al. 2007), and adult competition for resources (Hopwood et al. 2013). By sequencing, assembling, and annotating the genome of *N. vespilloides*, we were able to address two questions: First, is the gene complement of *N. vespilloides* more reflective of phylogeny or life history? Second, given methylation has been implicated in the success of eusocial species and facilitates plasticity, could this mechanism play a role in *N. vespilloides* social plasticity as well? *Tribolium castaneum*, the model beetle species, seems to lack DNA methylation (Zemach et al. 2010). This has led to the assumption that methylation may be unimportant in beetles generally. However, methylation has been suggested to regulate behavioral states in social insects (Glastad et al. 2015; Yan et al. 2015). *Nicrophorus vespilloides* is an unusual beetle in that it is highly social, with extensive interactions between parents and offspring, but males in the presence of females do not care for offspring or show the same levels of gene expression as caring parents (Parker et al. 2015). There is a rapid transition between behavioral states if the female parent is removed (Smiseth et al. 2005), with extensive changes in gene expression in the male (Parker et al. 2015). Given this, we sought to test for the presence of DNA methylation in *N. vespilloides*, which could provide a mechanism for this rapid behavioral transition.

Burying beetles (*Nicrophorus* spp.) are a group of about 85 species that are subsocial, showing a usual level of direct and indirect parental care of offspring (Eggert and Müller 1997; Scott 1998; Sikes and Venables 2013). Burying beetles use vertebrate carcasses as food for their offspring, and go well beyond simple forms of parental care with direct regurgitation of food by parents to begging offspring (Eggert and Müller 1997; Scott 1998; Walling et al. 2008; Trumbo 2012). There is also indirect parental care including depositing antimicrobial excretions to retard decomposition and microbial growth on the carcass used as food. Parents continuously maintain the carcass against microbial growth and interspecific competitors (e.g., fly larvae). The most extensively studied burying beetle, *N. vespilloides*, has proven an excellent model for investigating the ecology and evolution of social interactions between family members (Eggert and Müller 1997; Scott 1998; Trumbo 2012; fig. 1). Although parental care is essential, especially in the first 24 h of larval life (Eggert et al. 1998; Smiseth et al. 2003), care in this species can be uniparental, either male or female, or biparental. All forms of care are equivalently beneficial for offspring (Parker et al. 2015).

Here, we report a genome assembly and annotation of *N. vespilloides* and use this to investigate hypotheses regarding evolution associated with social behavior and the unusual life history of this beetle. Our assembly integrates high-throughput short reads, long reads, and a genome map providing sequence for greater than 90% of the predicted genome size. We annotated 13,526 protein-coding genes and compared these genes to social insects, another beetle, and a fly that uses vertebrate carcasses as food but lacks sociality. The rationale was to test whether social evolution, shared aspects of life history such as using carcasses for developing larvae, or shared evolutionary history is associated with similar molecular evolution. The overlap of shared number of orthologs was similar between *N. vespilloides* and all other insect groups regardless of the use of carcass for reproduction or highly developed social interactions. We then tested whether *N. vespilloides* has DNA methylation by looking for sequences coding for the enzymes responsible, DNA methyltransferases, and by using whole-genome bisulfite sequencing with the hypothesis that like *T. castaneum* (Zemach et al. 2010), *N. vespilloides* would lack DNA methylation. We confirmed the lack of methylation in *T. castaneum* but we did find evidence of DNA methylation in *N. vespilloides*. We found that the genes methylated in *N. vespilloides* showed considerable overlap with those methylated in a Hymenopteran, the jewel wasp, *Nasonia vitripennis*. Thus, the *N. vespilloides* genome adds the first coleopteran resource to investigators interested in the genomic and molecular signatures of social interactions, parent–offspring conflict, social tolerance, mate choice, and mate cooperation with an experimentally tractable and evolutionarily divergent model to use in comparative studies.

Fig. 1.—An adult female *N. vespilloides* regurgitating food into the mouth of her begging larvae on a prepared mouse carcass. Photograph by A. J. Moore.

Materials and Methods

Animals Samples

All *N. vespilloides* used in this research were obtained from an outbred colony maintained at the University of Georgia under laboratory conditions for this species (see Cunningham et al. 2014 for a full description of conditions).

Genome Size Estimation, Sequencing, Assembly, and Quality Control

We used flow cytometry with propidium iodide staining to estimate the genome size of *N. vespilloides* using *T. castaneum* as a standard. Nuclei from insect heads and whole insects, respectively, were prepared as described in Yu et al. (2015), stained as in Hare and Johnston (2011), and analyzed with a CyAn Flow Cytometer (Beckman Coulter, Brea, CA) at the UGA's Center for Tropical and Emerging Global Diseases Flow Cytometry Core Facility. Data were processed with FlowJo software (Treestar, Inc., Ashland, OR).

Genomic DNA was extracted from a single larva derived from a single sibling–sibling mating using a sodium dodecyl sulfate-lysis buffer and a phenol–chloroform extraction. A 275-bp Illumina (San Diego, CA) TruSeq library was prepared and run on one lane of an Illumina HiSeq 2000 using a paired-end (2×100 bp) sequencing protocol at the HudsonAlpha Genome Sequencing Center (Huntsville, AL).

We used FastQC (v0.11.2; Babraham Institute; default settings) to create summary statistics and to identify possible adapter contamination of raw Illumina paired-end reads. No adapter contamination was reported, a result supported by analysis with CutAdapt (v1.2.1; Martin 2011), which only found evidence for adapters in less than 0.01% of the raw reads. Because sequencing library construction can generate inserts of genomic DNA that are less than twice the average read length, overlapping paired-end reads were first merged using FLASH (v1.2.4; Magoc and Salzberg 2011; default settings, insert size: 278 bp with SD of 53 bp [estimate from Platanus scaffolding step]). Quality control was performed with PrinSeq (Schmieder and Edwards 2011b). Reads were required to have a mean overall Phred quality score of $\geq$25, read ends were trimmed to >20 Phred quality score, a minimum length of 90 bp and a maximum length of 99 bp were allowed, and reads were allowed only one unidentified (N) nucleotide per read.

To obtain Pacific Bioscience (PacBio; Menlo Park, CA) continuous long reads (CLRs), we extracted genomic DNA using the same phenol–chloroform extraction as used to extract gDNA for the Illumina sequencing from a brother/sister pair of adult beetles that had been inbred for six generations. The University of Maryland Institute for Genomic Sciences prepared a 14.4-kb-long insert PacBio library. This library was sequenced with 22 PacBio's RS II P5-C4 Single Molecule, Real Time (SMRT) cells to generate CLRs to scaffold the assembly to increase long-range connectivity of the assembly. PacBio reads greater than 6,300 bp (36.4× coverage) were error corrected with the PBcR pipeline (Koren et al. 2012) using 49× coverage of the quality-controlled Illumina reads with default settings, which after error correction and assembly produced an estimated 20.9× coverage of CLRs.

To increase the long-range scaffolding (i.e., superscaffold) of our draft genome, we generated a BioNano Genomics (San Diego, CA) genome map. High molecular weight (HMW) genomic DNA was extracted from a single pupa as previously described (Shelton et al. 2015). HMW gDNA was nicked with a nicking restriction digest by *Bsp*QI and *Bbv*CI restriction enzymes that had been converted to nickases (New England Biolabs, Ipswich, MA). Restriction sites were labeled with fluorescent nucleotides and imaged on the Irys system (BioNano Genomics) according to the manufacturer's instructions.

All Illumina reads passing quality control were used as input for the Platanus assembler (v1.2.1; Kajitani et al. 2014). First, reads were assembled into contigs using the assemble protocol (nondefault settings: -s 3 -u 0.2 -d 0.3 -m 128). Next, contigs were scaffolded using the scaffold protocol (nondefault settings: -u 0.2). This step was iterated a total of five times using the same settings to extend the scaffold as much as possible with the Illumina reads. Gaps in the assembly were filled using the gap_close protocol with default settings. This step was iterated twice. Only contigs/scaffolds 1 kb or greater in length were used for further analysis and assembly.

PacBio reads were used to gap fill and scaffold the Platanus assembly with PBJelly2 (v14.9.9; English et al. 2012) using default settings and the error-corrected PacBio CLRs.

A genome map created from BioNano Genomic single molecule maps was used to superscaffold the Platanus/PBJelly assembly (Shelton et al. 2015). The BioNano Genomics genome map provides a means to "superscaffold" an assembly by using HMW DNA that has been fluorescently labeled at specific sequence recognition sites that is then compared with in silico maps of the assembly to link scaffolds over very large

genomic distances. It also provides an independent validation of a genome assembly. Briefly, the images were assembled into a consensus map based on the labeling pattern of each molecule imaged. These in silico maps, with a cumulative length of 133.7 Mb, were compared with the predicted labeling pattern of the Platanus/PBJelly that passed a quality filter (length > 150 kb and number of labels ≥ 8) to further scaffold and orient the Platanus/PBJelly assembly.

DeconSeq (v0.4.2; Schmieder and Edwards 2011a) was used to assess our draft assembly for possible contamination. Besides the 1,126 bacterial species included in the distribution, we also updated the human genome sequence (h37) and added the genomes of *Caenorhabditis elegans*, *Ralstonia pickettii*, and *Yarrowia lipolytica*. *Caenorhabditis elegans* was included because it is the closest genome available to the nematode symbiont of *N. vespilloides*, *Rhabditis stammeri* (Richter 1993). *Ralstonia pickettii* and *Y. lipolytica* were included because they were two species that showed up when the RNA-Seq experiment was assessed for possible contamination (Parker et al. 2015). *Tribolium castaneum* was used as a retention database. Only one contig was flagged and removed during our contamination search; belonging to *Morganella morganii*, a common bacterium found in vertebrate intestinal tracts.

Genome assembly quality and completeness were assessed with multiple benchmark data sets. First, the CEGMA analysis pipeline (v2.4.010312; Parra et al. 2009) was used to assess the completeness of 248 ultra conserved eukaryotic genes within our assembly. Next, we used the *T. castaneum* set of Benchmarking sets of Universal Single-Copy Orthologs (BUSCOs; 2,787 genes) to further assess the assembly completeness (Waterhouse et al. 2013). We also mapped the RNA-Seq reads back to the assembly to estimate coverage of the transcriptome of our assembly using the TopHat (v2.0.13) pipeline with Bowtie2 (v2.2.3) as the read aligner.

Genome Annotation

To begin genome annotation, we first generated a de novo library of repeats using Repeat-Modeler (v1.0.8; Smit and Hubley 2014) that integrates three separate repeat finder programs; RECON (v1.08; Bao and Eddy 2002), RepeatScout (v1.05; Price et al. 2005), and TRF (v4.07b; Benson 1999) with default parameters. Because some gene fragments, especially low-complexity motifs, might be captured in the repeat analysis, we used BLASTx to remove any matches to *T. castaneum* proteins in the UniProtKB database (Wang et al. 2008; Jiang 2014). The repeat analysis of the *T. castaneum* genome was carried out with RepeatMasker (v4.0.5; Smit et al. 2015) using default settings.

We annotated putative protein-coding genes using the Maker2 annotation pipeline (v2.31.7; Holt and Yandell 2011) using an iterative process. After masking putative repeats within a genome, this pipeline generates gene models, including 5′- and 3′-untranslated regions (UTRs), by integrating ab initio gene predictions with aligned transcript and protein evidence. First, we mapped and assembled transcripts using the RNA-Seq data from an experiment of *N. vespilloides* in multiple behavioral states over a breeding cycle (mated, caring, and postcaring; see Parker et al. 2015 for full details) using the Bowtie (v2.2.3)/TopHat (v2.0.13)/Cufflinks (v2.2.1) pipeline (Langmead et al. 2009; Trapnell et al. 2010; Kim et al. 2013). To begin the annotation process, we annotated the genome exclusively with the *N. vespilloides* Cufflinks-assembled transcripts and the proteomes from five insects (*T. castaneum, Na. vitripennis, Apis mellifera, Musca domestica,* and *Drosophila melanogaster*; downloaded from UniProtKB, including all isoforms for comprehensive coverage) using default parameters, except for est2genome=1, protein2genome=1. After this first iteration of annotation (and every subsequent iteration), three scaffolds were inspected to visually check for annotation biases (Hoff and Stanke 2015) using the Apollo genome browser (Lewis et al. 2002). The next iteration used the same input data and parameters, except changes to split_hit=2000, correct_est_fusion=1, which corrected for the smaller intron size observed and the propensity of MAKER to fuse gene models that likely should be separate as inferred by visual inspection of BLAST evidence. For the next iteration, three ab initio gene predictors were included in the annotation process: Augustus (v2.5.5; Stanke et al. 2006), GeneMark-ES (v4.21; Lomsadze et al. 2005), and SNAP (v2010-7-28; Korf 2004; using est2genome=0, protein2genome=0). With AUGUSTUS, we used the "tribolium" gene set provided with its distribution to guide gene predictions. GeneMark was trained on the draft assembly of the *N. vespilloides* genome sequence using its automated training routine. SNAP was trained using the MAKER2 gene models produced during the first round of annotation. All gene predictors were run with default parameter values. The annotation was iterated twice with the gene predictors, updating the SNAP HMMs between the two iterations. Transfer RNAs were identified using tRNAscan-SE (v1.23; Lowe and Eddy 1997) within the Maker2 pipeline during the last iteration. Other noncoding RNA (ribosomal RNA, microRNA, small nuclear RNA, and small nucleolar RNA) were predicted and annotated with INFERNAL (v1.1.1; Nawrocki and Eddy 2013) using the complete Rfam database (v12.0; Nawrocki et al. 2014; supplementary table S4, Supplementary Material online).

Functional Annotation of Predicted Protein-Coding Genes

To gain insight into the putative function of each gene model, we annotated our gene models with three pipelines. First, we used BLASTp (v2.2.26; Altschul et al. 1997) to find the best hit against the entire UniProtKB database (vJan15; *E* value: 10e-5). Next, we used InterProScan (v5.8-49.0; Hunter et al. 2009) to find the known protein domains within every gene model from the TIGRFAM, ProDom, SMART, TMHMM, Phobius,

PANTHER, PrositeProfiles, SignalP-EUK, SuperFamily, PRINTS, Gene3d, PIRSF, Pfam, and Coils databases. We also used InterProScan5 to assign gene ontology (GO) terms to further characterize each protein. KEGG pathway analysis was also performed using the KEGG Automatic Annotation Server (KAAS; Moriya et al. 2007) using the single-directional best hit method to assign orthology with default parameters and the default Eukaryote gene sets plus all available arthropod gene sets.

Ortholog Comparison

To compare the orthology of our gene models to other insects, we analyzed our final MAKER2 proteome using OrthoMCL (v2.0.9; Li et al. 2003) against five other insect proteomes. We compared with *T. castaneum* and *D. melanogaster* as model inset genomes, *Na. vitripennis* and *A. mellifera* as other insects that share a social life history, and *M. domestica* as an insect that shares the use of carrion for reproduction and food for developing young. If a gene was represented by more than one isoform in its respective official gene set (OGS), the longest isoform was chosen for this analysis. We used BLASTp (*E* value: 1e-5) to characterize the homology among all proteins. The output from this analysis was used by OrthoMCL to cluster proteins into orthologous groupings. Results are presented as Venn diagrams generated using the University of Ghent Bioinformatics Evolutionary Genomics' Venn Diagram webtool (http://bioinformatics.psb.ugent.be/webtools/Venn/, last accessed April 16, 2015).

Gene Family Expansion/Contraction Analysis

To investigate possible expansion and contraction of shared gene families of the six insects that we used in the OrthoMCL analysis, we used CAFÉ (v3.1; default settings; Han et al. 2013) with phylogenetic relationships from Trautwein et al. (2012) and divergence times from TimeTree (Hedges et al. 2006). Only gene families with at least one representative from *N. vespilloides* were considered as gene family contractions.

Enrichment of GO terms among the expanded gene family members was performed using argiGO's web-based Singular Enrichment Analysis (Du et al. 2010) of customized annotations by comparing the GO terms associated with methylated gene from the InterProScan results to all GO terms associated with all genes from InterProScan. Specifically, a hypergeometric test with a Benjamini–Hochberg false discovery rate (FDR) correction at a familywise error rate of 0.05 was applied after GO terms were converted into generic GO slim terms. All other parameters were set at default values.

Selection Analysis

To assess the rates of molecular evolution within the *N. vespilloides* genome, we used PAML (Yang and Bielawski 2000; Yang 2007) to calculate d*N*, d*S*, and their ratio (ω) and compare these metrics to the beetles *T. castaneum* and *Dendroctonus ponderosae*. We identified a set of 1:1 orthologs between *N. vespilloides, D. ponderosae* and *T. castaneum* using a combination of the BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1997; Camacho et al. 2009), orthAgogue (Ekseth et al. 2014), and mcl (Enright et al. 2002; van Dongen 2008) as well as part of the OrthoMCL (Li et al. 2003) pipeline. In total, 5,584 orthologs between all three species were recovered. Amino acid sequences for each were aligned in PRANK (v100802; Löytynoja and Goldman 2005). Codeml in the PAML package was used to test different models of molecular evolution for each gene. Our interest is in determining which genes show evidence of a differential rate of evolution within *N. vespilloides*. We therefore tested a basic model (model = 0, NSsites = 0, fix_omega = 0) that assumes a single ω across all the entire phylogeny against a branch model (model = 2, NSsites = 0, fix_omega = 0), which assumes one ω for the *N. vespilloides* branch and another ω for the branches to *T. castaneum* and *D. ponderosae*. These models are compared, for each gene, with a likelihood ratio test with 1 degree of freedom. We then adjusted the significance threshold for a gene to show statistically significant different rates of sequence evolution using a Benjamini–Hochberg FDR correction at q of 0.05 (Benjamini and Hochberg 1995). Finally, any estimates of d*S,* d*N* or $\omega > 10$ were discarded. These species are phylogenetically distant (240 Ma) and this increases the likelihood signals of molecular evolution will be lost due to saturation of d*S*.

DNA Methylation Analysis

As the first step to characterize if DNA methylation existed within *N. vespilloides*, we use BLASTp (Altschul et al. 1997) to identify putative DNA methyltransferases. We search our genome with known members of Dnmt families of both vertebrate (*Mus musculus*; 1, 2, 3a, 3b, 3 l) and invertebrate (*T. castaneum, A. mellifera, D. melanogaster*; 1, 2, and 3). After three putative loci were found (one member per Dnmt family), we further characterized the possible functional relationship of the proteins by clustering them with the BLAST query proteins and several more invertebrate species (*Zootermopsis nevadensis* and *Camponotus floridanus*) using ClustalW followed by a neighbor-joining tree with 10,000 bootstraps in CLC Sequence Viewer (v7.5; http://www.clcbio.com) with default settings.

To address whether DNA methylation is present in *N. vespilloides*, we performed methylC-Seq (Lister et al. 2008), whole-genome sequencing of bisulfite-treated genomic DNA, on three biological replicates of whole larvae to create single base resolution of DNA methylation, if present. DNA was extracted from three whole *N. vespilloides* larvae, respectively, using the same protocol as for the Illumina and PacBio sequencing (see above). Due to previous reports that *T. castaneum* contains no DNA cytosine methylation

(Zemach et al. 2010), samples from this species were used as a negative control and DNA was extracted from three biological replicates that each contained at least 15 pooled whole larvae using the same protocol as for *N. vespilloides*. methylC-Seq libraries were prepared according to the protocol of Urich et al. (2015). Deep sequencing was performed using an Illumina NextSeq500 Instrument at the University of Georgia Genomics Facility (supplementary table S5, Supplementary Material online).

Raw fastq files were trimmed for adapters CutAdapt (v1.3) and preprocessed to remove low-quality reads. We aligned quality-controlled reads to the *N. vespilloides* v1.0 and *T. castaneum* v3.0 reference genomes using the method as described in Schmitz et al. (2013). The *T. castaneum* genome and OGS gff (v3.0) were obtained from BeetleBase.org. Lambda sequence (which is fully unmethylated) was used as a control to calculate the efficiency of the sodium bisulfite reaction and the associated nonconversion rate of unmodified cytosines, which ranged from 0.10% to 0.11% (supplementary table S5, Supplementary Material online). Only cytosine sites with a minimum coverage of three reads were used for subsequent analysis. A binomial test coupled with a Benjamini–Hochberg FDR correction at a familywise error rate of 5% was used to determine the methylation status of every cytosine (Benjamini and Hochberg 1995). Weighted methylation levels were calculated as previously described (Schultz et al. 2012).

We next characterized the distribution of methylated cytosines across the *N. vespilloides* genome and gene models. Methylated cytosines and their flanking two bases were extracted out for sequence conservation analysis using the program WebLogo 3.3 (Crooks et al. 2004). To perform the symmetry analysis, both strands of each CpG dinucleotide were required to have a minimum coverage of at least three reads and at least one of the CpG sites was identified as methylated. Upstream regions were defined as 1 kb upstream starting from the translational start site or the transcriptional start sire if a 5′-UTR was annotated. The program bedtools was used to determine the distribution of methylated CpG sites (Quinlan and Hall 2010). We used a two-step process to identify "methylated" and "unmethylated" genes. First, the probability of a methylated CpG site occurring within a gene was determined by totaling all methylated CpG sites within all genes and dividing this value by the total mapped CpG sites within all genes. Second, the methylated CpG sites and mapped CpG sites of each gene were used to determine that gene's methylation status using a binomial test. These results were then corrected for multiple testing using a Benjamini–Hochberg FDR correction at 5%. Only genes with at least five mapped CpG sites were reported. *Nicrophorus vespilloides* replicate 1 was used to compute the exact values and percentages, but all replicates were qualitatively similar (supplementary table S6, Supplementary Material online).

To compare with previous documented signatures of methylation in insects, we calculated $CpG_{O/E}$ ratios for each gene following the method described in Elango et al. (2009), a ratio of the observed level of methylation in genes over expected levels given the GC content of the genes analyzed. Thus, $CpG_{O/E}$ is a normalized measure of depletion of CpG dinucleotides. Following Elango et al. (2009), the $CpG_{O/E}$ for each gene was calculated as

$$CpG_{O/E} = \frac{P_{CpG}}{P_C * P_G},$$

where P_{CpG} is the frequency of CpG dinucleotides, P_C is the frequency of cytosine, and P_G is the frequency of guanine estimated from each gene

Finally, we compared the genes that were methylated in *N. vespilloides* to another insect with a recently characterized active methylation system, *Na. vitripennis* (Wang et al. 2013). For direct comparison, we generated the *Na. vitripennis* results with their previously published data. We downloaded raw reads and mapped them to the published *Na. vitripennis* v1.0 reference genome and OGS v1.2. "Methylated" genes were established with the same protocol as we describe above for *N. vespilloides* to ensure that the comparison was appropriate. We only included single-copy ortholog that existed in both *N. vespilloides* and *Na. vitripennis* genomes in the comparison of the overlap between methylated gene sets.

Results

Genome Sequencing and Assembly

We assembled the genome of *N. vespilloides* by integrating evidence from Illumina short reads, Pacific Bioscience (PacBio) CLRs, and a BioNano Genomics genome map (supplementary table S1, Supplementary Material online). We assembled 195.3 Mb of the *N. vespilloides* genome, which is 95.7% of its predicted size (supplementary fig. S1, Supplementary Material online). The draft genome is contained within 5,858 contigs with an N50 of 102.1 kb and further into 4,664 scaffolds with an N50 of 122.4 kb (Longest scaffold: 1.80 Mb; table 1). The Illumina and PacBio data produced as assembly with a scaffold N50 of 115.4 kb and a longest scaffold of 989 kb. With the addition of the BioNano Genomics genome map, these metrics were increased to 122.4 kb and 1.795 Mb, respectively. The GC content is 32%, consistent with two other beetle genomes, *T. castaneum* at 33% (Tribolium Genome Sequencing Consortium 2008) and *D. ponderosae* at 36% (Keeling et al. 2013).

We assessed how well the protein-coding portion of the genome was assembled using the CEGMA and BUSCO pipelines. Our genome contained 247 complete orthologs (99.6%) and 248 partial orthologs (100%) of the CEGMA proteins. Of the 2,827 *T. castaneum* BUSCO proteins, our genome contained 2,737 (96.8%) as single-copy orthologs

Table 1
Summary Statistics of *Nicrophorus vespilloides* Draft Genome Assembly

Total assembled length (bp)	**195,308,655**
Contigs (*n*)	5,858
Contig N50 (bp)	102,139
Largest contig (bp)	944,646
Scaffolds (*n*)	4,664
Scaffold N50 (bp)	122,407
Largest scaffold (bp)	1,795,199
% GC content	31.85
Predicted gene models (# of loci)	13,526
CEGMA pipeline analysis (% complete/ partial)	99.6/100
Analysis of *N. vespilloides* with *Tribolium castaneum* BUSCO gene set	
Mean % sequence indenity shared	65.95
Mean % of *T. castaneum* gene length found	90
% of *T. castaneum* BUSCO genes found as single-copy orthologs	96.8
BioNano Genomics genome map alignment to in silico maps (%)	31

and 86 (3.1%) as multicopy orthologs. We also mapped the RNA-Seq data used for annotation back to the genome to assess transcriptome coverage. There was an 89.7% mapping rate.

Genome Annotation

We used Maker2 to annotate the protein-coding portion of the genome by integrating ab initio, protein homology, and species-specific RNA-Seq evidence into consensus gene models. We obtained 13,526 predicted gene models. The gene models had an average protein length of 466.7 amino acids and 6.3 exons. Maker2 also predicted 5′-UTRs for 5,813 genes (mean: 512 bp) and 3′-UTRs for 4,549 genes (mean: 980 bp).

We were able to functionally annotate 11,585 gene models (85.6%) against UniProtKB with BLASTp. Restricted to species that had five or more best matches against *N. vespilloides* (encompassing 97.8% of the annotated gene models), the annotated gene models overwhelmingly returned the strongest similarity to other Coleoptera (fig. 2 and supplementary table S2, Supplementary Material online; top three species—*T. castaneum*: 6,969, *D. ponderosae*: 1,368, *Anoplophora glabripennis*: 743; Coleopteran total: 9,210 [79.5%]). Arthropods were the strongest similarity matches for 11,305 (99.7%) gene models (fig. 2). We were also able to identify at least one protein domain in 86.1% of the genes using InterProScan5. Searches against the Pfam database found 9,467 domains from 3,932 unique families. We were also able to assign at least one GO term to 7,492 genes (55.4%). Additionally, we were able to associate KEGG orthology terms with 44.8% of the genes.

Our de novo repeat analysis found that 12.85% of the draft genome is composed of repetitive elements. The top three classifications of repeats were unclassified repetitive elements (6.13%), DNA elements (3.35%), and simple repeats (2.24%). The overall repeat content is lower than that reported for beetles *T. castaneum* (Tribolium Genome Sequencing Consortium 2008) and *D. ponderosae* (Keeling et al. 2013), but higher than the honey bee (Elsik et al. 2014) and the red harvester ant (Smith et al. 2012), all of which have genomes that are of comparable size to *N. vespilloides*. Additionally, when we provided our repeat library to RepeatMasker to mask the *T. castaneum* genome only 1.65% was masked, an outcome consistent when the repeat library of *D. ponderosae* was used for the same task (0.15% of *T. castaneum* masked; Keeling et al. 2013; supplementary table S3, Supplementary Material online).

Orthology of Gene Models

We used OrthoMCL, which clusters proteins based on a reciprocal best BLAST hit strategy, to assign orthology of the *N. vespilloides* proteome against five other insect proteomes chosen either because they are genomic models (*T. castaneum* and *D. melanogaster*) or because they share a social life history (*A. mellifera* and *Na. vitripennis*) or the use of carcasses to breed and as food for offspring (*M. domestica*). Thus, these are simple and limited comparisons but they serve as a first enquiry into the forces that might shape genome evolution.

Our analysis produced 11,929 orthologous groupings with representatives from at least two different lineages. There were 4,928 orthologs groupings that contained at least one protein from each species. Of these, 3,734 groupings were single-copy orthologs among the six insects. There were 153 groupings containing 532 proteins that had proteins from *N. vespilloides* only. The beetles, *N. vespilloides* and *T. castaneum*, were represented in 7,827 groupings and 716 groupings were exclusive to beetles (650 were single-copy ortholog groupings). We then made two specific comparisons of the proteomes of *N. vespilloides*, *T. castaneum*, *A. mellifera*, *Na. vitripennis*, *D. melanogaster,* and *M. domestica* (fig. 3). *Nicrophorus vespilloides* shared 6,465 orthologous groupings with *D. melanogaster*, 6,479 with *M. domestica*, 7,028 with *A. mellifera*, and 6,240 with *Na. vitripennis*. We used a *z*-test to test whether the proportion of shared orthologous groupings was different between our two comparisons (*A. mellifera* vs. *Na. vitripennis* and *D. melanogaster* vs. *M. domestica*). *Nicrophorus vespilloides* shared more orthologous groupings with *A. mellifera* than with *Na. vitripennis* ($z = 9.539$, $P < 0.001$); however, *N. vespilloides* did not share more orthologous groupings exclusively with *A. mellifera* than *T. castaneum* (140 vs. 130, respectively; $z = 0.613$, $P = 0.729$). *Nicrophorus vespilloides* did not share more orthologous

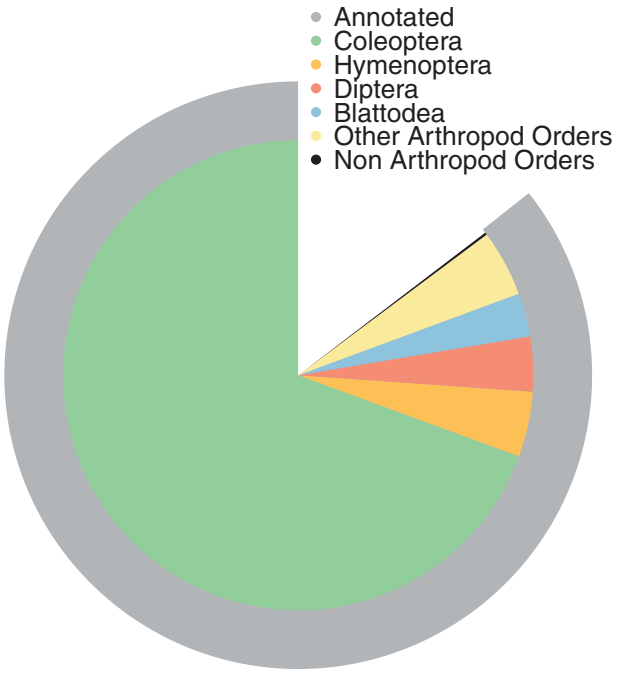


Fig. 2.—A two-ring pie chart showing results of annotation with BLAST against the complete UniProtKB database. First outer ring (gray) shows the proportion of gene models that could be annotated. Second ring (multicolored) shows the proportion of best BLAST hits of the annotations by order for all species with five or more best hits (97.8%). The best BLAST hits were overwhelmingly from other beetles and other Arthropods.

groupings with *M. domestica* than with *D. melanogaster* ($z = -1.427$, $P = 0.156$). However, *N. vespilloides* did share more orthologous groupings exclusively with *M. domestica* than *T. castaneum* (60 vs. 37, respectively; $z = 2.341$, $P = 0.022$).

Gene Family Expansion and Contraction

To investigate whether there had been any gene family expansions or contractions in *N. vespilloides*, we analyzed the results of the OrthoMCL analysis with CAFÉ. There were 269 orthology groupings (or gene families) that showed significant expansion or contraction between the six insect species compared at $P < 0.0001$. Of these groupings 12 showed significant differences within the *N. vespilloides* lineage. There were eight expansions and four contractions (supplementary file S1, Supplementary Material online). The expansions were mostly families of uncharacterized proteins (7/8), whereas the last family was a chymotrypsin protease. There was not an enrichment of any GO term from the expanded gene families. The contracted families had highest similarity to an esterase, a transposase, a cytochrome P450, and an uncharacterized protein in *T. castaneum*. Some of these are also differentially expressed during caring (Parker et al. 2015).

Selection Analysis

Signatures of selection on the protein-coding genes of *N. vespilloides* were investigated by comparing the d*N*/d*S* (ω) ratio to *T. castaneum* and *D. ponderosae* for the 5,584 one-to-one orthologs we detected between these lineages. Twenty-five genes showed signs of differential divergent selection in the *N. vespilloides* lineage after our filtering criteria were applied (see supplementary fig. S2, Supplementary Material online; BH FDR = 0.05 and removal of genes showing d*N*, d*S*, or $\omega > 10$). Two genes show evidence of positive selection $\omega > 1$: Ephrin-B2 (*efn-b2*; $\omega = 1.45$) and NK Homeobox (HOX) 7 (*nk7*; $\omega = 2.16$). *efn-b2* also has a $\omega > 1$ in the other lineages ($\omega = 1.5$), whereas *nk7* shows evidence of strong conservation in the *T. castaneum* and *D. ponderosae* lineages. The median estimates of d*S*, d*N* and ω were higher in the *N. vespilloides* lineage (*N. vespilloides*: 0.0489, *T. castaneum:* 0.0487, and *D. ponderosae*: 0.0487), although not statistically significantly different.

DNA Methylation

We used two approaches to investigate whether the *N. vespilloides* genome has active DNA methylation. First, we looked for the enzymes responsible for methylation in animals (Dnmt1, Dnmt2, and Dnmt3) to determine whether the machinery was present for the establishment and maintenance of DNA methylation. Second, we generated single-base resolution maps of DNA methylation using whole-genome bisulfite sequencing.

Single copies of all three DNA methyltransferases were in the *N. vespilloides* genome; *T. castaneum* contains only Dnmt1 and Dnmt2 (Kim et al. 2010). The methyltransferases clustered with their putative orthologs (fig. 4*A*). Next, using MethylC-Seq we found direct evidence for DNA cytosine methylation in *N. vespilloides* (mean = 29,224 methylated cytosines) and no evidence for DNA methylation in *T. castaneum* (mean = 29 methylated cytosines), supporting previous reports on the latter (fig. 4*B*; Zemach et al. 2010). Methylation (5′-methlycytosine) in *N. vespilloides* was found within a CpG context exclusively (fig. 4*C*). A small proportion (1.87%) of CpH (H = A, T, or C) was found during the first analysis; however, further analysis of the originally identified CpH methylated sites revealed that greater than 98% of them were artifacts of segregating single nucleotide polymorphisms. Therefore, only strong evidence was found for CpG methylation in the genome. Methylated cytosines in *N. vespilloides* exhibited the typical insect pattern where most mapped reads at a given locus provided support for methylation or not (fig. 4*D*) and a high level of symmetrical methylation on opposing DNA strands (fig. 4*E*). The genome-wide pattern of DNA methylation observed for *N. vespilloides* is also similar to other insects. Most prominently, the majority of methylation was found within genic regions (94.75% of the observed methylation) and further within the exons (62.55 ± 0.26%

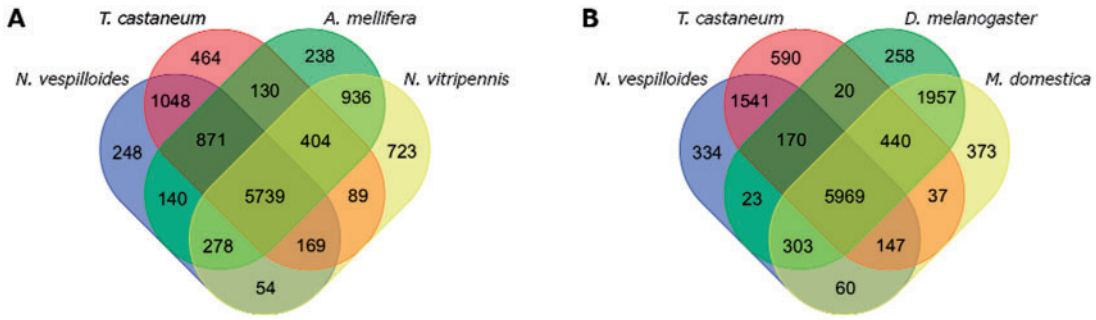


Fig. 3.—Figure shows the results of the OrthoMCL analysis that clustered the proteomes of *N. vespilloides*, *T. castaneum*, *A. mellifera*, *Na. vitripennis*, *D. melanogaster*, and *M. domestica* into orthologous groupings. (*A*) A Venn diagram showing the overlap in the orthologous groupings of the two beetles (*T. castaneum* and *N. vespilloides*) and the two Hymenoptera (*A. mellifera* and *Na. vitripennis*). (*B*) A Venn diagram showing the overlap in orthologous groupings of the two beetles (*T. castaneum* and *N. vespilloides*) and the two Diptera (*D. melanogaster* and *M. domestica*).

of the observed methylation) and much lower levels were found in introns (10.29 ± 0.12% of the observed methylation; fig. 4*F*). All three biological replicates are quantitatively similar in their distribution of methylated CpGs over gene elements (supplementary table S6, Supplementary Material online). We grouped *N. vespilloides* genes as methylated or unmethylated by comparing the level of methylation of an individual gene with the average level of gene methylation found across all genes. We found 2,782 genes that were methylated significantly higher than the null expectation (fig. 4*G* and supplementary file S2, Supplementary Material online). Following this, we performed a GO term enrichment analysis on the GO terms associated with the methylated gene set. We found that nucleic acid binding (GO:0003676), translation factor activity/nucleic acid binding (GO:0008135), and RNA binding (GO:0003723) were significantly enriched molecular function GO terms. Cellular macromolecule metabolic process (GO:0044260), cellular protein metabolic process (GO:0044267), and macromolecule biosynthesis process (GO:043170) were the three most enriched biological process GO terms (see also supplementary table S7, Supplementary Material online). At the level of individual genes, methylation was highest in the exons (fig. 4*H*). Methylation was also observed in the 5′- and 3′-UTRs, with the typical steep decrease in methylation observed at the translational start site. We also observed methylation in the "promoter" region 1 kb upstream from the first annotated gene element. Methylation peaks beginning at the second exon, although this is not a robust trend as methylation levels decrease to the same level of the first exon by the end of the second exon. Transposable elements were methylated to the same level as genomic intergenic background levels (3% vs. 5%, respectively).

Comparing patterns of methylation to other insects, we found that as expected methylated genes had lower $CpG_{O/E}$ values compared with nonmethylated genes (fig. 4*I*). The mean of methylated genes was 0.82, whereas that for nonmethylated genes was 1.13. We further assessed how many of methylated genes overlapped in a Hymenopteran and the burying beetle. We found that there were 4,633 methylated genes in the jewel wasp *Na. vitripennis* and 2,782 *in N. vespilloides*. Of the 1,958 single-copy orthologs that were methylated in *N. vespilloides*, 85% overlapped with methylated genes in the jewel wasp (fig. 4*J*).

Discussion

The ability to detect conserved and novel molecular mechanisms that influence social behavior requires genomic resources from species across different lineages that vary in their level of sociality. Here, we report the draft genome of *N. vespilloides*, a subsocial beetle from the Silphidae. In assessing the genetic changes associated with the evolution of social behavior in insects, the *N. vespilloides* genome provides a useful line of independent evolution, offering data from outside the Hymenoptera, which diverged from Coleoptera approximately 350 Ma (Wiegmann et al. 2009) and at a level between solitary and eusocial. *Nicrophorus vespilloides* has sophisticated and complex parental care (Eggert and Müller 1997; Scott 1998; Trumbo 2012). The highly developed social interactions between parents and offspring place this beetle at the level of "subsocial" on the evolutionary spectrum of social species (Wilson 1971; Costa 2006).

We successfully assembled the *N. vespilloides* genome using Illumina short reads, PacBio CLRs, and a BioNano Genomics genome map. Our assembly quality compares favorably with other recently published insect genomes; especially considering our organism is outbred (Richards and Murali 2015). We found that our genome is similar to other recently sequenced insect genomes, with a comparable number of genes and percentage of genes having a functional annotation (Kim et al. 2010; Wurm et al. 2012; Keeling et al. 2013; Oxley et al. 2014; Wang et al. 2014). Our orthology analysis showed that *N. vespilloides* was as similar to social Hymenoptera or to a Dipteran as it is to the asocial beetle *T. castaneum* with respect to the number of shared gene

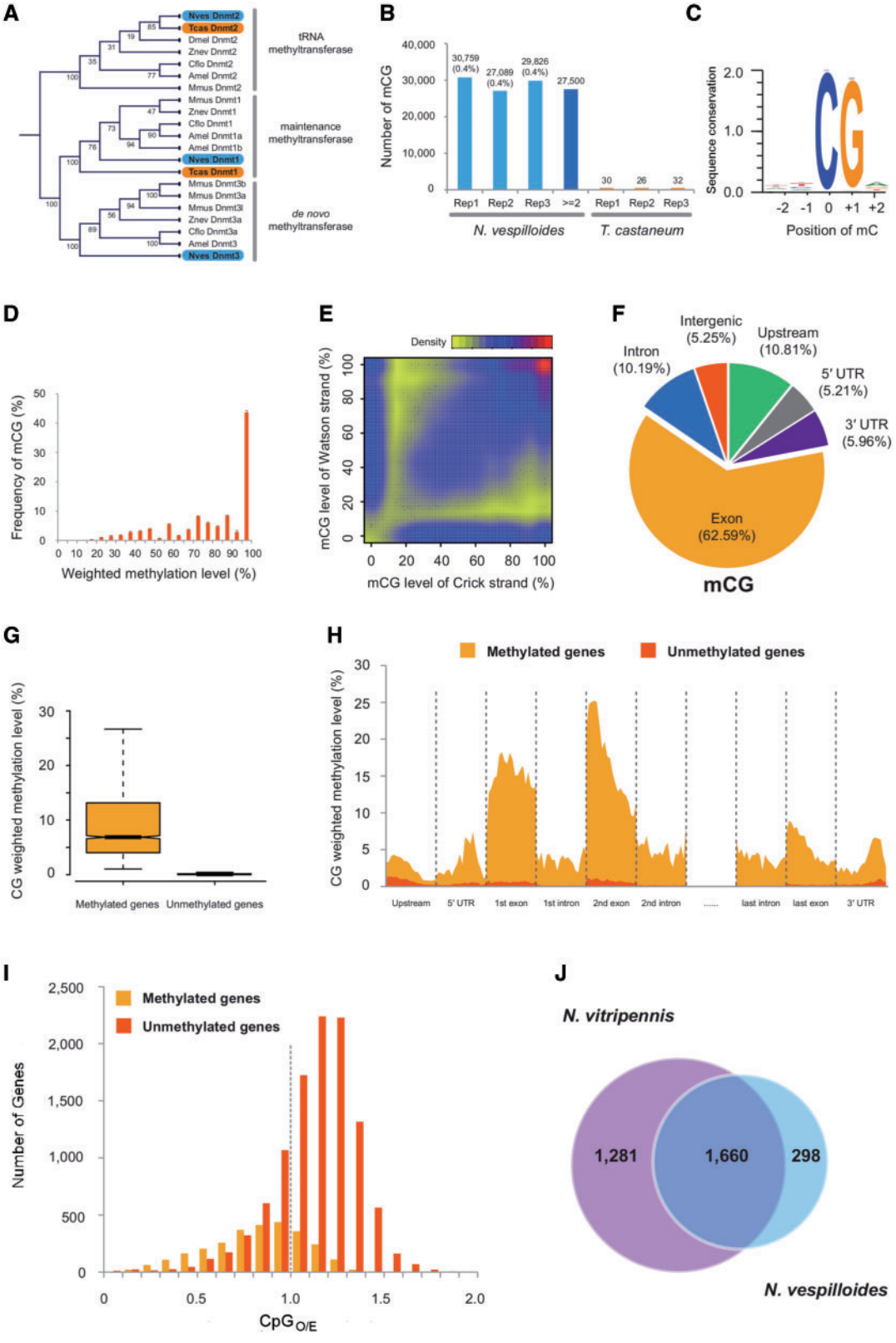


Fig. 4.—Summary of DNA methylation analyses. (*A*) A cladogram showing the relationship of the Dnmt's across several insects and a mammal. Nves, *N. vespilloides*; Tcas, *T. castaneum*; Dmel, *D. melanogaster*; Cflo, *C. floridanus*; Amel, *A. mellifera*; Mmus, *M. musculus*; Znev, *Zootermopsis nevadensis*. (*B*) Number of methylated cytosines in each of the three replicates of *N. vespilloides* and *T. castaneum*. (*C*) A sequence logo of the overwhelming occurrence of methylation in CpG dinucleotide by showing the nucleotide proportions of the two nucleotides both upstream and downstream of the methylated cytosines. (*D*) A histogram of CpGs that are considered methylated versus the proportion of reads that supported their methylation status (weighted methylation level).

families. This is in contrast to the finding of expanded repertoire of immune genes and chemoreceptor genes in *M. domestica* compared with *D. melanogaster* (Scott et al. 2014).

Very few of the *N. vespilloides* genes we examined showed evidence of differential rates of sequence evolution compared with the *T. castaneum* and *D. ponderosae* lineages. Among the genes that did show differential d*N*/d*S* ratios, the majority showed low d*N*/d*S* values consistent with evolutionary conserved amino acid sequence (Yang and Bielawski 2000). We found only two genes with evidence of d*N*/d*S* >1, consistent with positive diversifying selection. NK HOX 7 had an elevated ω in the *N. vespilloides* lineage but is highly conserved in the other lineages. Ephrin-B2 had an elevated ω in all lineages but it is slightly lower in the *N. vespilloides* lineage. Both of these genes are involved in developmental patterning (Dönitz et al. 2015; Dos Santos et al. 2015). Overall, the genes compared show a high degree of conservation. One limitation of this analysis is the approximately 240 Ma of evolutionary distance between *N. vespilloides* and *T. castaneum* (Hunt et al. 2007). Moving forward, it would be interesting to see how robust these results are to other types of analyses of molecular evolution and as more beetle species over a range of phylogenetic distances are available for comparison.

Beetles are typically described as lacking DNA methylation, based on *T. castaneum* (Glastad et al. 2015; Yan et al. 2015), although the sequence for Dmnt3 has been described from transcriptomic data of a dung beetle (*Onthophagus taurus*; Choi et al. 2010) and differential methylation associated with development investigated in this beetle with amplified fragment length polymorphisms (Snell-Rood et al. 2013). In contrast to other beetles with sequenced genomes, we have direct evidence for DNA methylation of the *N. vespilloides* genome and our works show that lacking methylation is not a general feature of Coleoptera. In fact, methylation in *N. vespilloides* looks very similar to most other insects with active systems of methylation. *Nicrophorus vespilloides* has DNA methylation that is restricted to CpG sites at levels similar to honey bees (Lyko et al. 2010) and the jewel wasp *Na. vitripennis* (Wang et al. 2013), the ants *C. floridanus* and *Harpegnathos saltator* (Bonasio et al. 2012), a grasshopper *Schistocerca gregaria* (Falckenhayn et al. 2013), a locust *Locusta migratoria* (Wang et al. 2014), and the silkworm moth *Bombyx mori* (Xiang et al. 2010). Methylation is concentrated within the exons of genes as seen with honey bees (Lyko et al. 2010), ants (Bonasio et al. 2012), the jewel wasp (Wang et al. 2013), but different from a locust (Wang et al. 2014), silkworm moth (Xiang et al. 2010) and termite (Terrapon et al. 2014). Methylation was also found in the UTRs, a pattern also reported in *C. floridanus* and *H. saltator* (Bonasio et al. 2012). Methylation peaks at the beginning of the second exon, a pattern seen in ants (Bonasio et al. 2012) and the jewel wasp (Wang et al. 2013). The methylation status of genes in *N. vespilloides* appears to be evolutionarily conserved compared with jewel wasp, as true for honey bee compared with pea aphid (Hunt et al. 2010) and jewel wasp compared with honey bee (Wang et al. 2013).

It is intriguing that a social beetle, but not a nonsocial beetle, has DNA methylation. Differential DNA methylation has been implicated in the transition between behavioral states in social insects (Lyko et al. 2010; Bonasio et al. 2012; Herb et al. 2012; Terrapon et al. 2014). Because *N. vespilloides* demonstrates dramatic and reversible switches in behavioral states across a breeding cycle, and can have multiple breeding cycles, we hypothesize that DNA methylation is an epigenetic mechanism that influences these behavioral transitions.

Studies of the genetic basis and evolution of complex social behavior have focused on specific genes, with conflicting results. However, these are mostly focused on division of labor in the eusocial Hymenoptera (Zayed and Robinson 2012; Rehan and Toth 2015). The addition of the *N. vespilloides* genome allows us to expand beyond hymenopteran-specific aspects of social behavior, and allows us to begin to address broader categories of social traits. Although there are numerous aspects of the life history of burying beetles that make them unique (Eggert and Müller 1997; Scott 1998), here we have emphasized the value of using *N. vespilloides* as a model for studying family social interactions and social evolution. These beetles are particularly suited for questions of parental care because the phenotype is robust and readily measured, contains diverse subbehaviors that are reliably observed and scored, can vary between males and females in the context in which it is expressed, and is highly replicable (Walling et al. 2008). With the addition of the *N. vespilloides* genome, we have a taxonomically diverse arsenal of phenotypically overlapping organisms to look for phylogenetically independent genomic mechanisms and signatures of evolution, conservation, and novelty.

Fig. 4.—Continued

(*E*) A density plot showing the very high symmetry of methylated CpG sites on opposing strands of DNA. (*F*) A pie chart showing the distribution of methylated CpGs across gene elements. (*G*) A standard box plot of the proportion of reads that supported methylation status (weighted methylation level) with genes grouped by whether they were methylated or not. (*H*) Diagram showing the proportion of reads that supported methylation (weighted methylation level) of methylated and unmethylated genes across each region of a gene model summarized as 20 bins within a region. (*I*) Histograms of methylated and unmethylated genes versus the CpG observed/expected ratio across a gene body. (*J*) Venn diagram illustrating the overlap of methylated genes that had 1:1 orthology between the burying beetle *N. vespilloides* and the jewel wasp *Na. vitripennis*.

Supplementary Material

Supplementary files S1 and S2, figures S1 and S2, tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

For advice or technical assistance, the authors thank the Quantitative Biology Consulting Group at the University of Georgia (especially, Walt Lorenz and Saravanaraj Ayyampalayam), Julie Brown at the University of Georgia's CTEGD Cytometry Shared Resource Center, Jessica Kissinger, the Genomic Services Lab at HudsonAlpha, Roger Nilsen at the Georgia Genomics Facility, the Institute for Genome Science at the University of Maryland, and the staff at the Georgia Advanced Computing Resource Center (especially, Yecheng Huang). They also especially thank Nick Talbot, who first suggested to A.J.M. he develop a genome from short reads in 2007. Three anonymous reviewers provided very helpful suggestions. All genomic data has been submitted to public repositories and can be found using the NCBI BioProject number PRJNA284849. This work was supported by the University of Georgia's Office of the Vice-President for Research to A.J.M. and R.J.S., and a National Science Foundation grant (IOS-1354358) to A.J.M.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bao Z, Eddy SR. 2002. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 12:1269–1276.

Boake CRB, et al. 2002. Genetic tools for studying adaptation and the evolution of behavior. Am Nat. 160:S143–S159.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Meth. 57:289–300.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573.

Bonasio R, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Curr Biol. 22:1755–1764.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Cardoso SD, Teles MC, Oliveira RF. 2015. Neurogenomic mechanisms of social plasticity. J Exp Biol. 218:140–149.

Choi J-H, et al. 2010. Gene discovery in the horned beetle *Onthophagus taurus*. BMC Genomics 11:703.

Costa JT. 2006. The other insect societies. Cambridge (MA): Harvard University Press.

Crooks GE, Hon G, Chandonia, JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

Cunningham CB, Douthit MK, Moore AJ. 2014. Octopaminergic gene expression and flexible social behavior in the subsocial burying beetle *Nicrophorus vespilloides*. Insect Mol Biol. 23:391–404.

Dönitz J, et al. 2015. iBeetle-Base: a database for RNAi phenotypes in the red flour beetle *Tribolium castaneum*. Nucleic Acids Res. 43:D721–D726.

Dos Santos G, et al. 2015. FlyBase: introduction of the *Drosophila melanogaster* release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 43:D690–D697.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z. 2010. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 38:W64–W70.

Eggert A-K, Muüller JK. 1997. Biparental care and social evolution in burying beetles: lessons from the larder. In: Choe JC, Crespi BJ, editors. The evolution of social behavior in insects and arachnids. Cambridge: Cambridge University Press. p. 216–236.

Eggert A-K, Reinking M, Muüller JK. 1998. Parental care improves offspring survival and growth in burying beetles. Anim Behav. 55:97–107.

Ekseth OK, Kuiper M, Mironov V. 2014. OrthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics 30:734–736.

Elango N, Hunt BG, Goodisman MAD, Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. Proc Natl Acad Sci U S A. 106:11206–11211.

Elsik CG, et al. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15:86.

English AC, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7:e47768.

Enright AJ, Dongen SV, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Falckenhayn C, et al. 2013.Characterization of genome methylation patterns in the desert locust *Schistocerca gregaria*. J Exp Biol. 216:1423–1429.

Fischman BJ, Woodard SH, Robinson GE. 2011. Molecular evolutionary analysis of insect societies. Proc Natl Acad Sci U S A. 108:10847–10854.

Gadau J, et al. 2012. The genomic impact of 100 million years of social evolution in seven ant species. Trends Genet. 28:14–21.

Glastad KM, Chau LM, Goodisman MAD. 2015. Epigenetics in social insects. In: Zayed A, Kent CF, editors. Advances in insect physiology. Amsterdam: Elsevier. p. 227–269.

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30:1987–1997.

Hare EE, Johnston JS. 2011. Genome size determination using flow cytometry of propidium iodide-stained nuclei. In: Orgogozo V, Rockman M, editors. Molecular methods for evolutionary genetics. New York: Humana Press. p. 3–12.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Herb BR, et al. 2012. Reversible switching between epigenetic states in honeybee behavioral subcastes. Nat Neurosci. 10:1371–1373.

Hoff KJ, Stanke M. 2015. Current methods for automated annotation of protein-coding genes. Curr Opin Insect Sci. 7:8–14.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.

Hopwood PE, Moore AJ, Royle NJ. 2013. Nutrition during sexual maturation affects competitive ability but not reproductive productivity in burying beetles. Func Ecol. 27:1350–1357.

Hunt BG, Brisson JA, Yi SV, Goodisman MAD. 2010. Functional conservation of DNA methylation in the Pea Aphid and Honey bee. Genome Biol Evol. 2:719–728.

Hunt T, et al. 2007. A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. Science 318:1913–1916.

Hunter S, et al. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D211–D215.

Jiang N. 2014. ProtExcluder1.1. Available from: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction–Basic
Kajitani R, et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24:1384–1395.
Kapheim KM, et al. 2015. Genomic signatures of evolutionary transitions from solitary to group living. Science 348:1139–1143.
Keeling CI, et al. 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. Genome Biol. 14:R27.
Kilner RM, Hinde CA. 2012. Parent-offspring conflict. In: Royle NJ, Smiseth PT, Kölliker M, editors. The evolution of parental care. Oxford: Oxford University Press. p. 119–132.
Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.
Kim S, et al. 2010. BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. Nucleic Acids Res. 38:D437–D442.
Kocher SD, et al. 2013. The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. Genome Biol. 14:R142.
Koren S, et al. 2012. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol. 30:693–700.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Lewis SE, et al. 2002. Apollo: a sequence annotation editor. Genome Biol. 3:research0082.1–0082.14.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Lister R, et al. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovdky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494–6506.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nuclleic Acids Res. 25:955–964.
Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 102:10557–10562.
Lyko F, et al. 2010. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol. 8(11):e1000506. [Erratum in PLoS Biol. 2011;9(1).]
Magoc T, Salzberg, S. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 17:10–12.
McGlothlin JW, Moore AJ, Wolf JB, Brodie ED 3rd. 2010. Interacting phenotypes and the evolutionary process III. Social evolution. Evolution 64:2558–2574.
Mikheyev AS, Linksvayer TA. 2015. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. eLife 4:e04775.
Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.
Nawrocki EP, et al. 2014. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 43:D130–D137.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935.
Oxley PR, et al. 2014. The genome of the clonal raider ant *Cerapachys biroi*. Curr Biol. 24:451–458.
Parker D, et al. 2015. Transcriptomes of parents identify parenting strategies and sexual conflict in a subsocial beetle. Nat Commun. 6:8449.
Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37:289–297.
Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. Bioinformatics 21:i351–i358.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
Rehan SM, Toth AL. 2015. Climbing the social ladder: the molecular evolution of sociality. Trends Ecol Evol. 30:426–433.
Richards S. 2015. It's more than just stamp collecting: how genome sequencing can unify biological research. Trends Genet. 34:411–421.
Richards S, Murali SC. 2015. Best practices in insect genome sequencing: what works and what doesn't. Curr Opin Insect Sci. 7:1–7.
Richter S. 1993. Phoretic association between the dauerjuveniles of *Rhabditis stammeri* (Rhabditidae) and life history stages of the burying beetle *Nicrophorus vespilloides* (Coleoptera:Silphadae). Nematologica 39:346–355.
Sadd BM, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol. 16:76.
Sanggaard KW, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. Nat Commun. 5:3765.
Schmieder R, Edwards R. 2011a. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6:e17288.
Schmieder R, Edwards R. 2011b. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863–864.
Schmitz RJ, et al. 2013. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res. 23:1663–1674.
Schultz MD, Schmitz RJ, Ecker JR. 2012. "Leveling" the playing field for analyses of single-base resolution DNA methylomes. Trends Genet. 28:583–585.
Scott JG, et al. 2014. Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment. Genome Biol. 15:466.
Scott MP. 1998. The ecology and behavior of burying beetles. Annu Rev Entomol. 43:595–618.
Shelton JM, et al. 2015. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. BMC Genomics 16:734.
Sikes DS, Venables C. 2013. Molecular phylogeny of the burying beetles (Coleoptera: Silphidae: Nicrophorinae). Mol Phylogenat Evol. 69:552–565.
Smiseth PT, Darwell CT, Moore AJ. 2003. Partial begging: an empirical model for the early evolution of offspring signaling. Proc R Soc Lond B Biol Sci. 270:1773–1777.
Smiseth PT, Dawson C, Varley E, Moore AJ. 2005.How do caring parents respond to mate loss? Differential response by males and females. Anim Behav. 69:551–559.
Smiseth PT, Lennox L, Moore AJ. 2007. Interaction between parental care and sibling competition: parents enhance offspring growth and exacerbate sibling competition. Evolution 61:2331–2339.
Smit AFA, Hubley R. 2014. RepeatModeler Open-1.0. 2008–2015. Available from: http://www.repeatmasker.org.
Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. Available from: http://www.repeatmasker.org.
Smith CR, et al. 2012. Draft genome of the red harvester ant *Pogonomyrex barbatus*. Proc Natl Acad Sci U S A. 108:5667–5672.
Snell-Rood EC, Troth A, Moczek AP. 2013. DNA methylation as a mechanism of nutritional plasticity: limited support from horned beetles. J Exp Zool B Mol Dev Evol. 320B:22–34.

Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. 7:S11.
Terrapon N, et al. 2014. Molecular traces of alternative social organization in a termite genome. Nat Commun. 5:3636.
Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.
Trautwein MD, Weigmann BM, Beutel R, Kjer KM, Yeates DK. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. Annu Rev Entomol. 57:449–468.
Tribolium Genome Sequencing Consortium. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452:949–955.
Trumbo ST. 2012. Patterns of parental care in invertebrates. In: Royle NJ, Smiseth PT, Kölliker M, editors. The evolution of parental care. Oxford: Oxford University Press. p. 81–100.
Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. 2015. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc. 10:475–483.
van Dongen S. 2008. Graph clustering via a discrete uncoupling process. J Matrix Anal Appl. 30:121–141.
Walling CA, Stamper CE, Smiseth PT, Moore AJ. 2008. The quantitative genetics of sex differences in parenting. Proc Natl Acad Sci U S A. 105:18430–18435.
Wang S, Lorenzen MD, Beeman RW, Brown SJ. 2008. Analysis of repetitive DNA distribution patterns in the *Tribolium castaneum* genome. Genome Biol. 9:R61.
Wang X, et al. 2013. Function and evolution of DNA methylation in *Nasonia vitripennis*. PLoS Genet. 9:e1003872.
Wang X, et al. 2014. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 5:2957.
Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41:D358–D365.
Wiegmann BM, et al. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7:34.
Wilson EO. 1971. The insect societies. Cambridge: Harvard University Press.
Wurm Y, et al. 2012. The genome of the first ant *Solenopsis invicta*. Proc Natl Acad Sci U S A. 108:5679–5684.
Xiang H, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol. 28:516–520.
Yan H, et al. 2015. DNA methylation in social insects: how epigenetics can control behavior and longevity. Annu Rev Entomol. 60:435–452.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
Yu P, McKinney EC, Kandasamy MM, Albert AL, Meagher RB. 2015. Characterization of brain cell nuclei with decondensed chromatin. Dev Neurobiol. 75:738–756.
Zayed A, Robinson GE. 2012. Understanding the relationship between gene expression and social behavior: lessons from the honey bee. Annu Rev Genet. 46:591–615.
Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916–919.

Associate editor: Ellen Pritham

ECOLOGICAL GENOMICS

The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish

Noah M. Reid,[1] Dina A. Proestou,[2] Bryan W. Clark,[3] Wesley C. Warren,[4] John K. Colbourne,[5] Joseph R. Shaw,[5,6] Sibel I. Karchner,[7,8] Mark E. Hahn,[7,8] Diane Nacci,[9] Marjorie F. Oleksiak,[10] Douglas L. Crawford,[10] Andrew Whitehead[1]*

Atlantic killifish populations have rapidly adapted to normally lethal levels of pollution in four urban estuaries. Through analysis of 384 whole killifish genome sequences and comparative transcriptomics in four pairs of sensitive and tolerant populations, we identify the aryl hydrocarbon receptor–based signaling pathway as a shared target of selection. This suggests evolutionary constraint on adaptive solutions to complex toxicant mixtures at each site. However, distinct molecular variants apparently contribute to adaptive pathway modification among tolerant populations. Selection also targets other toxicity-mediating genes and genes of connected signaling pathways; this indicates complex tolerance phenotypes and potentially compensatory adaptations. Molecular changes are consistent with selection on standing genetic variation. In killifish, high nucleotide diversity has likely been a crucial substrate for selective sweeps to propel rapid adaptation.

The current pace of environmental change may exceed the maximum rate of evolutionary change for many species (*1*), yet little is known of the circumstances and mechanisms through which evolution might rescue species at risk of decline (*2*). The Atlantic killifish, *Fundulus heteroclitus*, is nonmigratory and abundant in U.S. Atlantic coast salt-marsh estuaries (*3*), including sites contaminated with complex mixtures of persistent industrial pollutants (Fig. 1A) that have reached lethal levels in recent decades (*4*). Some killifish populations resident in polluted sites exhibit inherited tolerance to normally lethal levels of these highly toxic pollutants (*5*) (Fig. 1B). To understand the genetics of rapid adaptation to radical environmental change in wild populations, we sequenced complete genomes from 43 to 50 individuals from each of eight populations (Fig. 1A and table S1): four tolerant (T) populations from highly polluted sites, each paired with a nearby reference [sensitive (S)] population. We combined these data with RNA sequencing (RNA-seq) to uncover unique and shared functional pathways and adaptive signatures of selection across populations.

Genomes from T1 and S1 populations were sequenced to 7-fold coverage per individual and the remaining populations, to 0.6-fold coverage (*6*). Genetic variation is strongly partitioned by geography (Fig. 1C); northern populations (T1, S1, T2, S2, T3, and S3) form a cluster distinct from southern populations (T4 and S4), consistent with their known phylogeography (*7*). In tolerant populations, nucleotide diversity is reduced genome-wide, and Tajima's D is shifted positive, relative to sensitive population counterparts (fig. S1); these indicate reduced effective population size in polluted sites. Tolerant-sensitive (T-S) population pairs share the most similar genetic backgrounds, and the fixation index (F_{ST}) is low between them (0.01 to 0.08) (fig. S2). We conclude that tolerant populations are recently and independently derived from local gene pools.

We identified genomic regions that are candidates for pollution tolerance (table S2 and fig. S3) by defining outlier regions as 5-kb windows that fell in the extreme 0.1% tails (for pi and Tajima's D) and 99.9% tails (for F_{ST}) of null distributions simulated from demographic models estimated from the data (*6*). Most outlier regions are small (52 to 69 kb), although a few are up to ~1.8 Mb (fig. S4). For each T-S population pair, signatures of selection are skewed in prevalence toward the tolerant population (fig. S5). Most outliers are specific to a tolerant population (0.5% of 5-kb outlier windows are shared) (fig. S6). However, loci showing the strongest signals of recent selection [highly ranked outliers (*6*)] are shared (Fig. 2A), suggesting convergent evolution for pollution tolerance. Within these shared outliers are key genes involved in the aryl hydrocarbon receptor (AHR) signaling pathway (*AHR2a*, *AHR1a*, *AIP*, and *CYP1A*) (Fig. 2B).

The importance of these outliers is supported by transcriptomics. When sensitive and tolerant populations were raised in a common clean environment for two generations and embryos were challenged with a model toxic pollutant, the polychlorinated biphenyl (PCB) 3,3′,4,4′,5-pentachlorobiphenyl (PCB 126)–tolerant populations exhibit reduced inducibility of AHR-regulated genes (Fig. 2C). The 70 genes up-regulated in response to pollutant challenge in sensitive populations, but not in tolerant populations (table S3), are enriched for those regulated by the AHR signaling pathway ($P < 0.0001$). Impaired AHR signaling is most apparent with the canonical transcriptional targets of AHR (Fig. 2C and table

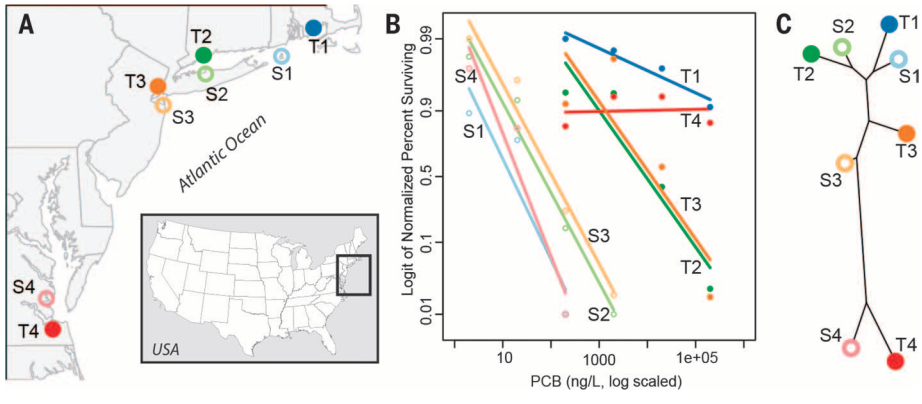


Fig. 1. Focal *F. heteroclitus* populations. (**A**) Locations of pollution-tolerant ("T"; bold tone, filled circles) and sensitive ("S"; pastel tone, open circles) population pairs numbered from north to south. (**B**) Population variation in larval survival (linear regression of logit survival to 7 days post hatch) after two generations reared in a common environment, when challenged with increasing log exposure concentrations of PCB 126. Populations from polluted sites exhibit tolerance to pollutants at concentrations hundreds to thousands of times normally lethal levels. (**C**) Phylogenetic tree, estimated from genome-wide biallelic single-nucleotide polymorphism (SNP) frequencies, showing that genetic differentiation is lowest between T-S population pairs [Phylogeny Inference Package (PHYLIP) Gene Frequencies and Continuous Characters Maximum Likelihood (CONTML) module, bootstrap supports are 100 for all branches].

[1]Department of Environmental Toxicology, University of California, Davis, CA 95616, USA. [2]Agricultural Research Service, U.S. Department of Agriculture, Kingston, RI 02881, USA. [3]Oak Ridge Institute for Science and Education, Office of Research and Development, U.S. Environmental Protection Agency, Narragansett, RI 02882, USA. [4]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA. [5]School of Biosciences, University of Birmingham, Edgbaston B15 2TT, UK. [6]School of Public and Environmental Affairs, Indiana University, Bloomington, IN 47405, USA. [7]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. [8]Boston University Superfund Research Program, Boston University, Boston, MA 02118, USA. [9]Office of Research and Development, U.S. Environmental Protection Agency, Narragansett, RI 02882, USA. [10]Department of Marine Biology and Ecology, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL 33149, USA.
*Corresponding author. Email: awhitehead@ucdavis.edu

S4). Dominant pollutants at T sites include halogenated aromatic hydrocarbons (HAHs) and polycyclic aromatic hydrocarbons (PAHs) that bind AHR and initiate aberrant signaling that causes malformations during development and subsequent embryo and larval lethality, as well as toxicity in adults (*8*). Given that the AHR pathway is repeatedly desensitized in tolerant populations (Fig. 2C) (*9*) and top-ranked outliers contain AHR pathway genes, we conclude that the AHR signaling pathway is likely a key and repeated target of natural selection in tolerant populations. This convergence suggests that adaptive options are constrained to modifications of this signaling pathway that mediates the toxicity of many HAHs and PAHs.

AHR deletions are found in tolerant populations. Four paralogs of *AHR* exist in the *F. heteroclitus* genome (*10*). Knockdown of *AHR2a* is protective of toxicity from many HAHs and PAHs [e.g., (*11*)]. Tandem paralogs *AHR2a* and *AHR1a* are within a highly ranked outlier region in all tolerant populations (Fig. 2A). Note that three tolerant populations have deletions (fig. S7) spanning *AHR2a* and *AHR1a* (Fig. 3A). In T4, a deletion is found in a single haplotypic background (fig. S8) that segregates at high frequency (81%) but is absent in S4 (Fig. 3B). In T4 individuals, RNA-seq data reveal expression of a chimeric transcript (joining exon 10 of *AHR2a* and exon 7 of *AHR1a*). In T1 and T3, different deletions spanning *AHR2a* and *AHR1a* (Fig. 3, A and B) occur in two and one haplotypic backgrounds, respectively (fig. S9). A deletion is present in at least one sensitive population (Fig. 3B), but no deletion was found in T2. Variation in this region is also associated with sensitivity to PCB toxicity in T1 (*12*) and in PCB-adapted tomcod (*13*). We thus conclude that *AHR* genes are likely common loci of selection for multiple genetic variants, including deletions, where a single deletion-associated haplotype has swept in the southern tolerant population.

The strongest signal of selection we observed is in a window that is a shared outlier in all tolerant populations [aryl hydrocarbon receptor–interacting protein (AIP) in Fig. 2A]. In northern tolerant populations, a single large (650-kb) haplotype has swept to high frequency, accompanied by reduced pi. In T4, a different haplotype has swept to high frequency (Fig. 3C). In T1 (sequenced to higher coverage), we detect recombination breakpoints, allowing identification of a core haplotype region (~100 kb) that coincides with peak differentiation (fig. S10), within which we find AIP. Variation near this locus also associates with sensitivity to PCB toxicity in T1 (*12*). AIP regulates cytoplasmic stability and cytoplasmic-nuclear shuttling of the AHR protein and thereby influences AHR signaling and regulates toxicity (*14*).

A key transcriptional target of AHR, the biotransformation gene *CYP1A*, is within a top-ranking outlier region shared by all tolerant populations (Fig. 2A). Genotypes from tolerant populations are highly differentiated from sensitive populations (Fig. 3D) and *CYP1A* single-nucleotide polymorphism (SNP) variants are linked with tolerance (*15*). In northern tolerant populations, *CYP1A* duplications have swept to high frequency, where individuals have up to eight copies of the *CYP1A* gene (Fig. 3E and figs. S7 and S11), and duplicates are present in some sensitive populations. *CYP1A* expression is not increased in northern tolerant populations (embryos) (table S4), as one might expect after duplication. However, because *AHR* knockout in rodents decreases basal *CYP1A* expression (*16*) and AHR signaling is impaired in tolerant killifish, we hypothesize that *CYP1A* duplication has been favored as a compensatory, dosage-compensating adaptation for impaired AHR signaling in northern tolerant fish. In contrast, we find no evidence of duplication in T4 (Fig. 3E), although this region retains a strong signature of selection (Fig. 2A) and is highly differentiated from S4 (Fig. 3D). PAHs primarily contaminate T4, and these chemicals interact

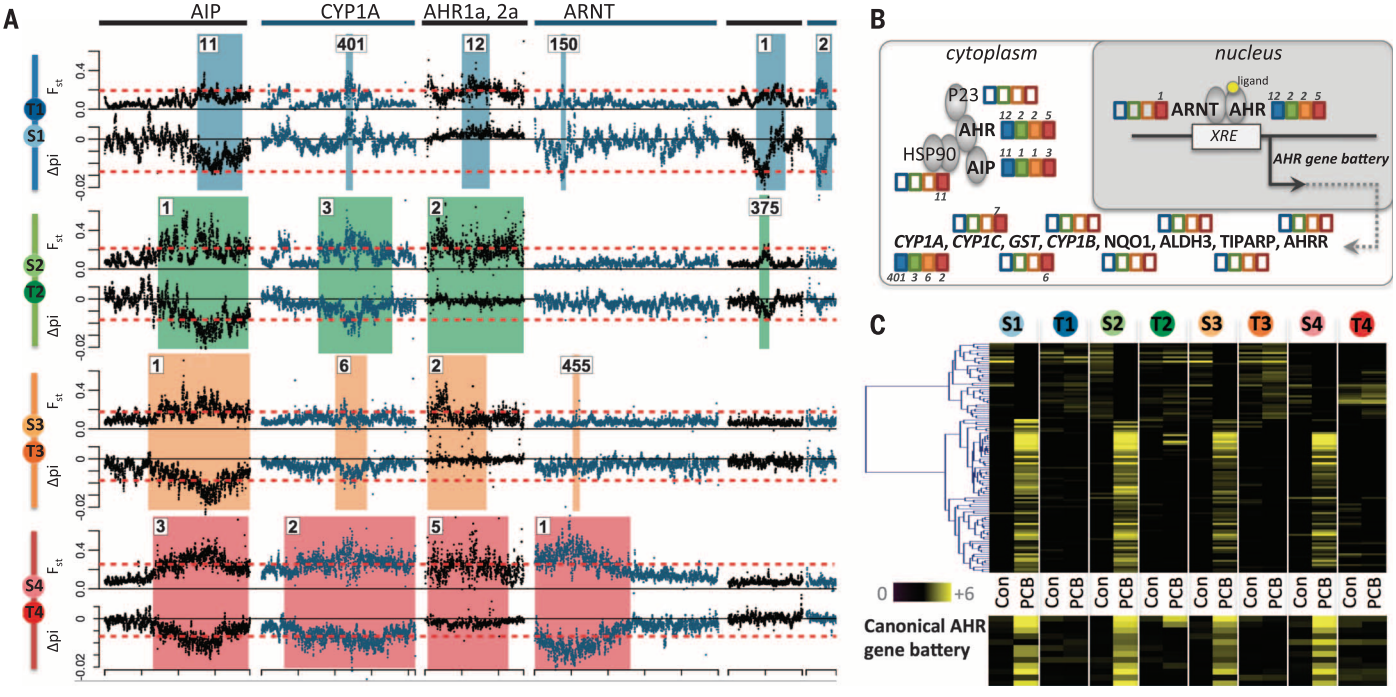


Fig. 2. Patterns of structural and functional genomic divergence. (**A**) Allele frequency differentiation (F_{ST}, top) and nucleotide diversity (pi, bottom) difference (tolerant pi – sensitive pi) for each population pair studied for top-ranking outlier regions (including the top two per pair). Colored panels span the outlier region of each respective population comparison where number indicates outlier rank for each tolerant-sensitive pair. Red dashed lines indicate outlier thresholds. Each tick on *x* axis is at the 500-kb position on the scaffold, and each candidate gene name is indicated (top) for each outlier region. Top outlier regions are not colocalized in the genome (fig. S3). (**B**) Model of key molecules in the AHR signaling pathway, including regulatory genes and transcriptional targets (*AHR* gene battery). Boxes next to genes are color-coded by population pair; filled boxes indicate the gene is within a top-ranking outlier region for that pair, and number indicates ranking of the outlier region as in (A). Top-ranking outlier regions contain AHR pathway genes and tend to be outliers in all population pairs, although some significant outliers are population-specific. (**C**) Gene-expression (of developing embryos) heat map shows up-regulated genes in response to PCB 126 exposure ("PCB"; 200 ng/liter) compared with control exposure ("Con") for sensitive populations, most of which are unresponsive in tolerant populations. The bottom panel highlights genes characterized as transcriptionally activated by ligand-bound AHR (table S4).

differently with *AHR*-induced *CYP1A* than with HAHs, which dominate northern sites (*17*). We propose that different chemical pollutants acting as selective agents may govern the fate of different *CYP1A* variants between HAH- and PAH-polluted sites.

Although AHR pathway genes are among shared outliers, they are also within population-specific outlier regions. Tandem paralogs *AHR1b* and *AHR2b* are within an outlier region in T3 and T4 (fig. S12) so that all four *AHR* paralogs are within outlier regions for one or more tolerant populations. Five additional AHR pathway genes are significant outliers for only T4. Two of these (*ARNT1c* and *HSP90*) (figs. S13 and S14) directly interact with AHR protein, whereas the remaining three (*CYP1C1/1C2*, *GFRP*, and *GSTT1*) (figs. S15 and S16) are PAH biotransformation genes that are also key transcriptional targets of *AHR* (Fig. 2C). The inclusion of PAH biotransformation genes among outliers specific to T4 (primarily polluted with PAHs) likely reflects differences between cellular effects of PAHs and HAHs (*17*).

Other selective targets include genes outside of AHR signaling. Some PAHs, particularly those that are abundant only at T4, cause cardiotoxicity independent of AHR (*18*) through disruption of voltage-gated potassium channels and regulation of intracellular calcium (*19*). Note that two genes whose products form the conductance pore of the voltage-gated potassium channel (*KCNB2* and *KCNC3*) are within top-ranking outlier windows in T4 (figs. S17 and S18). Similarly, ryanodine receptor (RYR) regulates intracellular calcium, and RYR3 is within an outlier window in T4 (fig. S19). We conclude that components of the adaptive phenotype are underpinned by genes that are both related and unrelated to AHR signaling, consistent with complex adaptations to complex chemical mixtures.

Our results also suggest compensatory adaptation associated with the (potential) costs of evolved pollution tolerance. AHR signaling has diverse functions and interacts with multiple pathways, including estrogen and hypoxia signaling, regulation of cell cycle, and immune system function (*20*). Estrogen receptor 2b is within an outlier region in T2 (fig. S20), and estrogen receptor–regulated genes are enriched within outlier gene sets for all tolerant populations ($P < 0.001$) (fig. S21). Estrogen receptor is also inferred as a significant upstream regulator for genes differentially expressed between tolerant and sensitive populations ($P < 0.05$) (e.g., genes in Fig. 2C). Hypoxia-inducible factor 2α is within an outlier window in T3 (fig. S22). Interleukin and cytokine receptors are in outlier windows in T4 (fig. S23). We conclude that some components of the adaptive phenotype in polluted sites may be due to compensation for the altered AHR signaling that underlies the primary pollutant-tolerance phenotype. Selection for compensatory changes may be common following rapid adaptive evolution.

In animal models, single gene (*AHR*) knockout can protect from toxicity of some HAH or PAH compounds [e.g., (*21*)]. However, in wild killifish populations, adaptive genotypes appear complex, including multiple AHR signaling pathway elements and other genes. We suggest that this complexity arises from two primary factors. First, tolerant sites are contaminated with complex mixtures of hydrocarbons. Mixture components may interact in subtly different ways with AHR (*17*), and some exert toxicity through pathways other than AHR (*18*), such that adaptations in multiple pathways are required. Second, because many of the AHR signaling pathway genes identified here as targets of selection interact with multiple regulatory pathways (*20*), changes to their function may have deleterious consequences

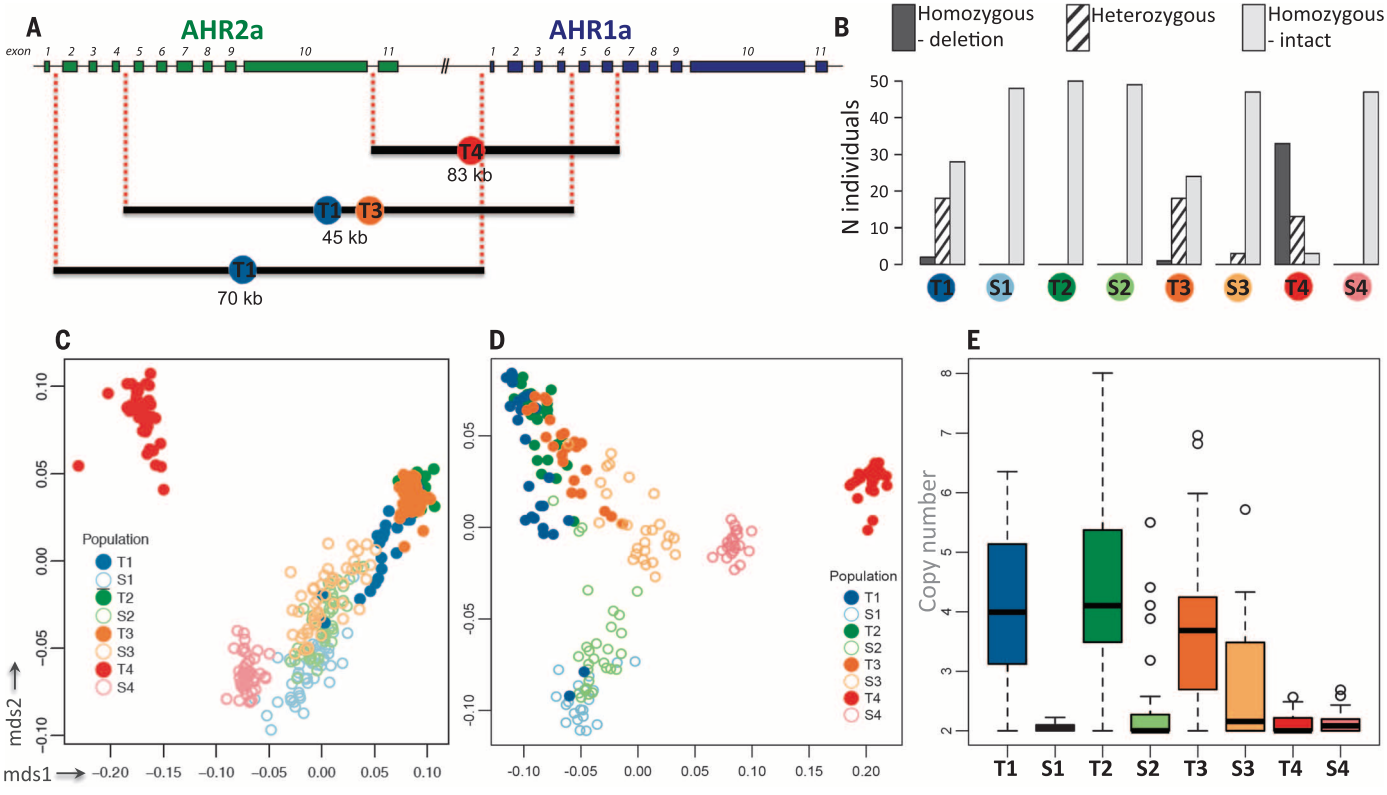


Fig. 3. Patterns of adaptive genetic variation for top-ranking and shared outliers. (**A**) Gene model of *AHR2a* and *AHR1a* (green or blue squares represent exons). Black bars indicate deleted regions present within tolerant populations. (**B**) The number of individuals homozygous for specific deletions (gray bar), heterozygous (hatched gray bar), or homozygous wild type (light bar) within each population. (**C**) Multidimensional scaling (MDS) plot of genotypic variation on the scaffold containing the *AIP* gene. (**D**) MDS plot of genotypic variation on the scaffold containing the *CYP1A* gene. (**E**) Bar plot of copy number of the duplications around *CYP1A*, where boxes, whiskers, and dots represent interquartile range, 1.5× interquartile range, and the remainder, respectively (the background diploid state includes two copies). Although the *CYP1A* region is highly differentiated in all tolerant populations (D), *CYP1A* duplications are found only in northern tolerant populations (E).

that may result in selection for compensatory change. Other changes in these highly altered estuaries may also exert selection pressures [e.g., estrogenic pollutants (*22*), hypoxia, or altered species diversity].

A fundamental question in evolutionary biology pertains to the nature and number of variants recruited by natural selection. The relative contributions of de novo variants, standing variation, and the number of competing beneficial variants depend in part on the strength of selection, its spatial patterning, existing genetic diversity and the beneficial mutation rate. Although modes of evolution can be difficult to distinguish (*23*), our data are revealing. We observe signals of convergence and divergence. Genes in the AHR pathway are repeated targets of selection, even in populations exposed to distinct chemical mixtures and separated by substantial genetic distance. This suggests adaptive constraint. Yet, different variants are often favored in different tolerant populations (e.g., *AHR* and *CYP1A*), some of which are present in sensitive populations, and common variants (e.g., large *AIP* haplotype) have rapidly swept in multiple populations of this low-dispersal fish. This suggests that selection on preexisting variants was important for rapid adaptation in killifish and that multiple molecular targets were available for selective targeting of a common pathway. The prevalence of soft sweeps is predicted to be high during rapid adaptation (*24*).

Evolutionary change relies on genetic variation that may preexist, or arise through new mutation, at a rate that scales by population size. *F. heteroclitus* at present has large population sizes (*3*) and a range of standing genetic variation (nucleotide diversity up to 0.016 for T3 and T4) that places them as one of the most diverse vertebrates (*25*). These factors suggest that Atlantic killifish have been unusually well positioned to evolve the necessary adaptations to survive in radically altered habitats.

REFERENCES AND NOTES

1. A. P. Hendry, T. J. Farrugia, M. T. Kinnison, *Mol. Ecol.* **17**, 20–29 (2008).
2. G. Bell, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120080 (2012).
3. I. Valiela, J. E. Wright, J. M. Teal, S. B. Volkmann, *Mar. Biol.* **40**, 135–144 (1977).
4. D. Nacci *et al.*, *Mar. Biol.* **134**, 9–17 (1999).
5. D. Nacci, D. Champlin, S. Jayaraman, *Estuaries Coasts* **33**, 853–864 (2010).
6. Materials and methods are available as supplementary materials on *Science* Online.
7. D. D. Duvernell, J. B. Lindmeier, K. E. Faust, A. Whitehead, *Mol. Ecol.* **17**, 1344–1360 (2008).
8. R. Pohjanvirta, *The AH Receptor in Biology and Toxicology* (Wiley, Hoboken, NJ, 2012).
9. A. Whitehead, W. Pilcher, D. Champlin, D. Nacci, *Proc. Biol. Sci.* **279**, 427–433 (2012).
10. A. M. Reitzel *et al.*, *BMC Evol. Biol.* **14**, 6 (2014).
11. B. W. Clark, C. W. Matson, D. Jung, R. T. Di Giulio, *Aquat. Toxicol.* **99**, 232–240 (2010).
12. D. Nacci, D. Proestou, D. Champlin, J. Martinson, E. R. Waits, *Mol. Ecol.* **25**, 5467–5482 (2016).
13. I. Wirgin *et al.*, *Science* **331**, 1322–1325 (2011).
14. M. Nukaya *et al.*, *J. Biol. Chem.* **285**, 35599–35605 (2010).
15. D. A. Proestou, P. Flight, D. Champlin, D. Nacci, *BMC Evol. Biol.* **14**, 7 (2014).
16. J. V. Schmidt, G. H. T. Su, J. K. Reddy, M. C. Simon, C. A. Bradfield, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6731–6736 (1996).
17. M. S. Denison, A. A. Soshilov, G. He, D. E. DeGroot, B. Zhao, *Toxicol. Sci.* **124**, 1–22 (2011).
18. J. P. Incardona *et al.*, *Environ. Health Perspect.* **113**, 1755–1762 (2005).
19. F. Brette *et al.*, *Science* **343**, 772–776 (2014).
20. T. V. Beischlag, J. Luis Morales, B. D. Hollingshead, G. H. Perdew, *Crit. Rev. Eukaryot. Gene Expr.* **18**, 207–250 (2008).
21. P. M. Fernandez-Salguero, D. M. Hilbert, S. Rudikoff, J. M. Ward, F. J. Gonzalez, *Toxicol. Appl. Pharmacol.* **140**, 173–179 (1996).
22. S. R. Greytak, A. M. Tarrant, D. Nacci, M. E. Hahn, G. V. Callard, *Aquat. Toxicol.* **99**, 291–299 (2010).
23. J. J. Berg, G. Coop, *Genetics* **201**, 707–725 (2015).
24. B. Wilson, P. Pennings, D. Petrov, *bioRxiv* 10.1101/052993 (2016).
25. E. M. Leffler *et al.*, *PLOS Biol.* **10**, e1001388 (2012).

ACKNOWLEDGMENTS

Sequence data are archived at the National Center for Biotechnology Information (BioProject PRJNA323589). Phylogenetic tree data are archived at Dryad (doi: 10.5061/dryad.68n87). We thank G. Coop, B. Counterman, D. Champlin, I. Kirby, and A. Bertrand for their valuable input. Primary support was from the NSF (collaborative research grants DEB-1265282, DEB-1120512, DEB-1120013, DEB-1120263, DEB-1120333, DEB-1120398 to J.K.C., D.L.C., M.E.H., S.I.K., M.F.O., J.R.S., W.C.W., and A.W.). Further support was provided by the National Institutes of Environmental Health Sciences (1R01ES021934-01 to A.W.; P42ES007381 to M.E.H.; R01ES019324 to J.R.S.), and the National Science Foundation (OCE-1314567 to A.W.). B.W.C. was supported by the Postdoctoral Research Program at the U.S. Environmental Protection Agency (EPA) administered by the Oak Ridge Institute for Science and Education (agreement DW92429801). The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the EPA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/354/6317/suppl/DC1
Materials and Methods
Figs. S1 to S26
Tables S1 to S4
References (*26*–*45*)

6 July 2016; accepted 31 October 2016
10.1126/science.aah4993

The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish

Noah M. Reid, Dina A. Proestou, Bryan W. Clark, Wesley C. Warren, John K. Colbourne, Joseph R. Shaw, Sibel I. Karchner, Mark E. Hahn, Diane Nacci, Marjorie F. Oleksiak, Douglas L. Crawford and Andrew Whitehead (December 8, 2016)
Science **354** (6317), 1305-1308. [doi: 10.1126/science.aah4993]

Editor's Summary

Mapping genetic adaptations to pollution

Many organisms have evolved tolerance to natural and human-generated toxins. Reid *et al.* performed a genomic analysis of killifish, geographically separate and independent populations of which have adapted recently to severe pollution (see the Perspective by Tobler and Culumber). Sequencing multiple sensitive and resistant populations revealed signals of selective sweeps for variants that may confer tolerance to toxins, some of which were shared between resistant populations. Thus, high genetic diversity in killifish seems to allow selection to act on existing variation, driving rapid adaptation to selective forces such as pollution.

Science, this issue p. 1305; see also p. 1232

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Epigenetics and the Evolution of Darwin's Finches

Michael K. Skinner[1,*], Carlos Gurerrero-Bosagna[1,3], M. Muksitul Haque[1], Eric E. Nilsson[1], Jennifer A.H. Koop[2,4], Sarah A. Knutie[2], and Dale H. Clayton[2]

[1]Center for Reproductive Biology, School of Biological Sciences, Washington State University

[2]Department of Biology, University of Utah

[3]Present address: Department of Physics, Biology and Chemistry (IFM), Linköping University, Sweden

[4]Present address: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ

*Corresponding author: E-mail: skinner@wsu.edu.

Accepted: July 18, 2014

Data deposition: All DMR and CNV genomic data obtained in this study have been deposited in the NCBI public GEO database under the accession (GEO #: GSE58334).

Abstract

The prevailing theory for the molecular basis of evolution involves genetic mutations that ultimately generate the heritable phenotypic variation on which natural selection acts. However, epigenetic transgenerational inheritance of phenotypic variation may also play an important role in evolutionary change. A growing number of studies have demonstrated the presence of epigenetic inheritance in a variety of different organisms that can persist for hundreds of generations. The possibility that epigenetic changes can accumulate over longer periods of evolutionary time has seldom been tested empirically. This study was designed to compare epigenetic changes among several closely related species of Darwin's finches, a well-known example of adaptive radiation. Erythrocyte DNA was obtained from five species of sympatric Darwin's finches that vary in phylogenetic relatedness. Genome-wide alterations in genetic mutations using copy number variation (CNV) were compared with epigenetic alterations associated with differential DNA methylation regions (epimutations). Epimutations were more common than genetic CNV mutations among the five species; furthermore, the number of epimutations increased monotonically with phylogenetic distance. Interestingly, the number of genetic CNV mutations did not consistently increase with phylogenetic distance. The number, chromosomal locations, regional clustering, and lack of overlap of epimutations and genetic mutations suggest that epigenetic changes are distinct and that they correlate with the evolutionary history of Darwin's finches. The potential functional significance of the epimutations was explored by comparing their locations on the genome to the location of evolutionarily important genes and cellular pathways in birds. Specific epimutations were associated with genes related to the bone morphogenic protein, toll receptor, and melanogenesis signaling pathways. Species-specific epimutations were significantly overrepresented in these pathways. As environmental factors are known to result in heritable changes in the epigenome, it is possible that epigenetic changes contribute to the molecular basis of the evolution of Darwin's finches.

Key words: epimutations, DNA methylation, copy number variation, phylogeny, adaptive radiation, BMP, toll, melanogenesis.

Introduction

Epigenetic change has been postulated to play a role in the ecology and evolution of natural populations (Richards et al. 2010; Holeski et al. 2012; Liebl et al. 2013). Epigenetic changes are broadly defined as "molecular processes around DNA that regulate genome activity independent of DNA sequence and are mitotically stable" (Skinner et al. 2010). Some epigenetic processes are also meiotically stable and are transmitted through the germline (Anway et al. 2005; Jirtle and Skinner 2007). These epigenetic mechanisms, such as DNA methylation, can become programmed (e.g., imprinted) and inherited over generations with potential evolutionary impacts. Environmental factors have been shown to promote the epigenetic transgenerational inheritance of phenotypic variants (Skinner et al. 2010). In recent years, the importance of environmental cues in the induction of such variation has been widely acknowledged (Bonduriansky 2012). Thus, like genetic change (Greenspan 2009), epigenetic change may also play an important role in evolution (Guerrero-Bosagna et al. 2005; Day and Bonduriansky 2011; Geoghegan and Spencer 2012, 2013a, 2013b, 2013c; Klironomos et al. 2013).

In order for inherited epigenetic changes to play a significant role in microevolution, they must persist for tens of generations, or longer (Slatkin 2009). It is conceivable that epigenetic changes may also accumulate over longer periods of evolutionary time, contributing to processes such as adaptive radiation (Rebollo et al. 2010; Flatscher et al. 2012). This hypothesis assumes that epigenetic changes persist over thousands of generations. An initial step in testing this hypothesis would be to compare epigenetic differences among closely related species, and whether such changes accumulate over short spans of macroevolutionary time. For example, do epigenetic changes accumulate with phylogenetic distance? Addressing this question was the primary goal of this study.

The study was designed to explore the relationship between epigenetic changes and the evolutionary history of several species of Darwin's finches in the Galapagos Islands. This group of birds has been central to work on a variety of important topics in evolutionary biology, including adaptive radiation, character displacement, rapid evolution, hybridization between species, evolutionary developmental mechanisms, and the effect of invasive pathogens and parasites (Grant and Grant 2008; Huber et al. 2010; Donohue 2011). The adaptive radiation of Darwin's finches over a period of 2–3 Myr resulted in 14 extant species that fill distinct ecological niches. These species show striking variation in body size and the size and shape of their beaks (Grant and Grant 2008). Darwin's finches were selected for study because they are a well-studied example of the evolution of closely related species into different ecological niches (Grant and Grant 2008; Donohue 2011).

Natural selection is a process in which environmental factors influence the survival and reproductive success of individuals bearing different phenotypes. Only selection on phenotypic traits with a heritable basis can lead to evolutionary change (Endler 1986). Observations indicate that epigenetic mechanisms have a role in influencing genomic variability (Huttley 2004; Ying and Huttley 2011). As epigenetic changes are also influenced by environmental factors, and can be heritable across generations (Skinner et al. 2010), they provide another molecular mechanism that can influence evolutionary change. Although Lamarck (1802) proposed that environmental factors can influence inheritance directly, his mechanism has not been widely recognized as a component of modern evolutionary theory (Day and Bonduriansky 2011). Recent work in epigenetics shows that epigenetic changes can, in fact, increase the heritable phenotypic variation available to natural selection (Holeski et al. 2012; Liebl et al. 2013). Thus, epigenetics appears to provide a molecular mechanism that can increase phenotypic variation on which selection acts (Skinner 2011). The integration of genetic and epigenetic mechanisms has the potential to significantly expand our understanding of the origins of phenotypic variation and how environment can influence evolution.

For example, Crews et al. (2007) investigated the ability of an environmental factor (toxicant) to promote the epigenetic transgenerational inheritance of alterations in the mate preferences of rats, with consequences for sexual selection. An F0 generation gestating female rat was exposed to the agricultural fungicide vinclozolin transiently. A dramatic alteration in the mate preferences of the F3 generation was observed (Crews et al. 2007) along with epigenetic alterations (termed epimutations) in the germline (sperm) (Guerrero-Bosagna et al. 2010). Transgenerational transcriptome changes in brain regions correlated with these alterations in mate preference behavior were also observed (Skinner et al. 2008, 2014). Thus, an environmental factor that altered mate preference was found to promote a transgenerational alteration in the sperm epigenome in an imprinted-like manner that was inherited for multiple generations (Crews et al. 2007; Skinner et al. 2010). Studies such as these suggest that environmental epigenetics may play a role in evolutionary changes through processes, such as sexual selection.

Recent reviews suggest a pervasive role for epigenetics in evolution (Rebollo et al. 2010; Day and Bonduriansky 2011; Kuzawa and Thayer 2011; Flatscher et al. 2012; Klironomos et al. 2013). The primary goal of this study was to test whether epigenetic changes accumulate over the long periods of evolutionary time required for speciation with adaptive radiation. Genome wide analyses were used to investigate changes in genetic and epigenetic variation among five species of Darwin's finches. The measure of genetic variation was copy number variation (CNV), which has been shown to provide useful and stable genetic markers with potentially more phenotypic functional links than point mutations such as single nucleotide polymorphisms (SNPs) (Lupski 2007; Sudmant et al. 2013). CNVs involve an increase or decrease in the number of copies of a repeat element at a specific genomic location. Recently, CNV changes in primates and other species have been shown to be very useful genetic measures for comparing evolutionary events (Nozawa et al. 2007; Gazave et al. 2011; Poptsova et al. 2013). CNV changes are involved in gene duplication and deletion phenomena, as well as repeat element phenomenon such as translocation events and can be influenced by DNA methylation (Skinner et al. 2010; Macia et al. 2011; Tang et al. 2012). The measure of epigenetic variation used was differential DNA methylation sites, which are known to be stable and heritable (Skinner et al. 2010). Comparing data for both genetic mutations (i.e., CNV) and epimutations (i.e., DNA methylation) allowed the relative magnitudes of these sources of variation to be compared across the five species included in the study.

Materials and Methods

Finch Field Work and Collection of Blood

Blood samples were collected from birds captured January–April 2009 at El Garrapatero, a lowland arid site on Santa Cruz Island, Galapagos Archipelago, Ecuador (Koop et al. 2011).

Birds were captured with mist nests and banded with numbered Monel bands to track recaptures. Birds were identified, aged, and sexed using size and plumage characteristics. A small blood sample (90 μl) from each bird was collected in a microcapillary tube through brachial venipuncture. Samples were stored on wet ice in the field, then erythrocytes purified by centrifugation and cells stored in a −20 °C freezer at a field station. Following the field season, samples were placed in a −80 °C freezer for longer term storage. All procedures were approved by the University of Utah Institutional Animal Care and Use Committee (protocol #07-08004) and by the Galápagos National Park (PC-04-10: #0054411).

DNA Processing

Erythrocyte DNA was isolated with DNAeasy Blood and Tissue Kit (Qiagen, Valencia, CA) and then stored at −80 °C prior to analysis. DNA was sonicated following a previously described protocol (without protease inhibitors) (Tateno et al. 2000) and then purified using a series of washes and centrifugations (Ward et al. 1999) from variable number of animals per species analyzed. The same concentrations of DNA from individual blood samples were then used to produce pools of DNA material. Two DNA pools were produced in total per species, each one containing the same amount of DNA from different animals. The number of individuals used per pool is shown in supplementary table S6, Supplementary Material online. These DNA pools were then used for chromosomal genomic hybridization (CGH) arrays or chromatin immunoprecipitation of methylated DNA fragments (MeDIP).

CNV Analysis

The array used for the CNV analysis was a CGH custom design by Roche Nimblegen that consisted of a whole-genome tiling array of zebra finch (*Taeniopygia guttata*) with 720,000 probes per array. The probe size ranged from 50 to 75 mer in length with median probe spacing of 1,395 bp. Two different comparative (CNV vs. CNV) hybridization experiments were performed (two subarrays) for each species in query (*Geospiza fuliginosa* [FUL], *G. scandens* [SCA], *Camarhynchus parvulus* [PAR], and *Platyspiza crassirostris* [CRA]) versus control *G. fortis* (FOR), with each subarray including hybridizations from DNA pools from these different species. Two DNA pools were built for each species (supplementary table S6, Supplementary Material online). For one subarray of each species, DNA samples from the experimental groups were labeled with Cy5 and DNA samples from the control lineage were labeled with Cy3. For the other subarray of each species, a dye swap was performed so that DNA samples from the experimental groups were labeled with Cy3 and DNA samples from the control lineage were labeled with Cy5.

For the CNV experiment raw data from the Cy3 and Cy5 channels were imported into R (R Development Core Team 2010), checked for quality, and converted to *MA* values ($M = \text{Cy5} - \text{Cy3}$; $A = [\text{Cy5} + \text{Cy3}]/2$). Within array and between array normalizations were performed as previously described (Manikkam et al. 2012). Following normalization, the average value of each probe was calculated and three different CNV algorithms were used on each of these probes including circular binary segmentation from the DNA copy (Olshen et al. 2004), CGHseg (Picard et al. 2005) and cghFlasso (Tibshirani and Wang 2008). These three algorithms were used with the default parameters. The average values from the output of these algorithms were obtained. A threshold of 0.04 as a cutoff was used on the summary (average of the log-ratio from the three algorithms) where gains are probes above the positive threshold and losses are probes below the negative threshold. Consecutive probes (≥ 3) of gains and losses were used to identify separate CNV regions. A cutoff of three-probe minimum was used and those regions were considered a valid CNV. The statistically significant CNVs were identified and *P* values associated with each region presented. A cutoff of $P < 10^{-5}$ was used to select the final regions of gains and losses.

Differential DNA Methylation Regions Analysis

MeDIP was performed as previously described (Guerrero-Bosagna et al. 2010) as follows: 6 μg of genomic DNA was subjected to series of three 20-pulse sonications at 20% amplitude and the appropriate fragment size (200–1,000 ng) was verified through 2% agarose gels; the sonicated genomic DNA was resuspended in 350 μl TE buffer and denatured for 10 min at 95 °C and then immediately placed on ice for 5 min; 100 μl of 5× IP buffer (50 mM Na-phosphate pH 7, 700 mM NaCl (PBS), 0.25% Triton X-100) was added to the sonicated and denatured DNA. An overnight incubation of the DNA was performed with 5 μg of antibody anti-5-methylCytidine monoclonal from Diagenode (Denville, NJ) at 4 °C on a rotating platform. Protein A/G beads from Santa Cruz were prewashed on PBS–BSA (bovine serum albumin) 0.1% and resuspended in 40 μl 1× IP (immunoprecipitation) buffer. Beads were then added to the DNA-antibody complex and incubated 2 h at 4 °C on a rotating platform. Beads bound to DNA-antibody complex were washed three times with 1 ml 1× IP buffer; washes included incubation for 5 min at 4 °C on a rotating platform and then centrifugation at 6,000 rpm for 2 min. Beads DNA-antibody complex were then resuspended in 250 μl digestion buffer (50 mM Tris–HCl pH 8, 10 mM ethylenediaminetetraacetic acid, 0.5% SDS (sodium dodecyl sulfate) and 3.5 μl of proteinase K (20 mg/ml) was added to each sample and then incubated overnight at 55 °C on a rotating platform. DNA purification was performed first with phenol and then with chloroform:isoamyl alcohol. Two washes were then performed with 70% ethanol, 1 M NaCl, and glycogen. MeDIP-selected DNA was then resuspended in 30 μl TE buffer.

The array used for the differential methylation analysis was a DNA-methylated custom array by Roche Nimblegen that

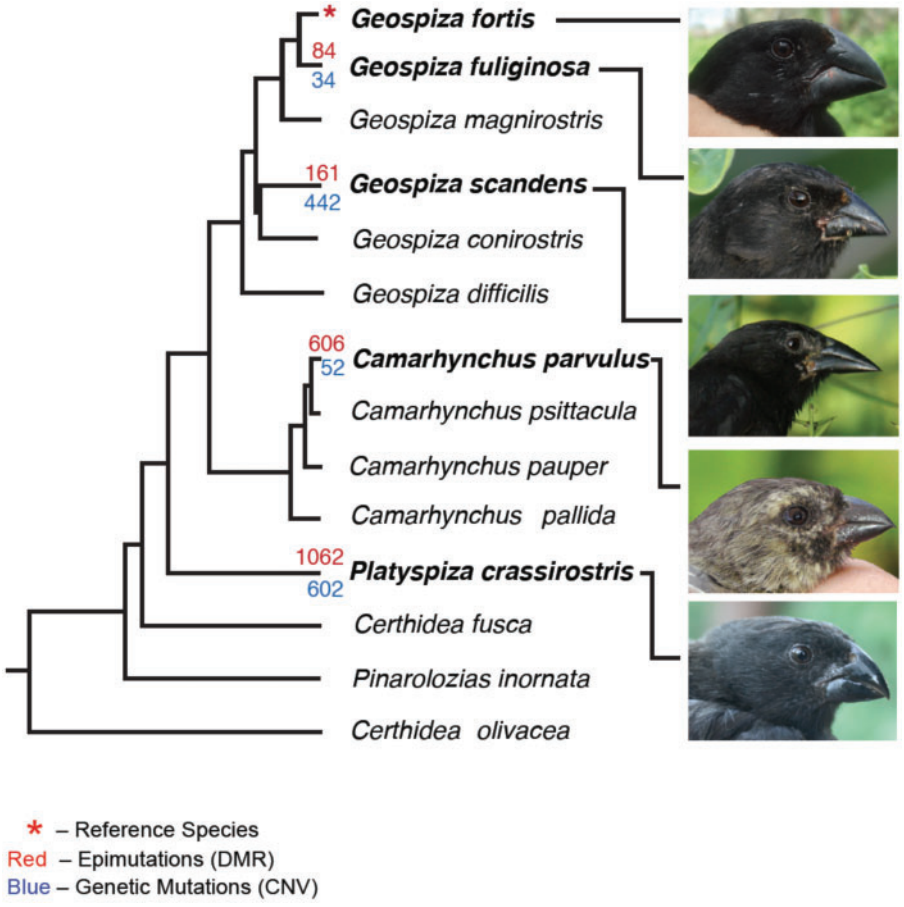


Fig. 1.—Number of epimutations and genetic mutations in relation to the phylogenetic relationships of five species of Darwin's finches. Photographs (by J.A.H.K. or S.A.K.) show variation in bill size and shape. Numbers on branches are the number of differences (three or more probes; table 1) in epimutations (DMR; in red) .and genetic mutations (CNV; in blue) for each of four species, compared with a single reference species FOR (asterisk). The phylogram is based on allele length variation at 16 polymorphic microsatellite loci (from Petren et al. 1999). The topology of the tree is similar to that proposed by Lack (1947) on the basis of morphological traits.

consisted of a whole-genome tiling array of zebra finch (*Taeniopygia guttata*) made of four 2.1M and one 3x720k array with 8,539,570 probes per array. Probe sizes were 50–75 mer in length and median probe spacing was 200 bp. Two different comparative (MeDIP vs. MeDIP) hybridization experiments were performed (two subarrays) for each experimental species (FUL, SCA, PAR, CRA) versus control FOR, with each subarray including hybridizations from MeDIP DNA from DNA pools from these different species (supplementary table S6, Supplementary Material online). For one subarray of each species, MeDIP DNA samples from the experimental groups were labeled with Cy5 and MeDIP DNA samples from the control lineage were labeled with Cy3. For the other subarray of each species, a dye swap was performed so that MeDIP DNA samples from the experimental groups were labeled with Cy3 and MeDIP DNA samples from the control lineage were labeled with Cy5.

For each comparative hybridization experiment, raw data from both the Cy3 and Cy5 channels were imported into R, checked for quality, and converted into *MA* values. The normalization procedure is as previously described (Guerrero-Bosagna et al. 2010). Following normalization each adjacent $\geq$3 probe set value represents the median intensity difference between FUL, SCA, PAR and CRA and control FOR of a 600-bp window. Significance was assigned to probe differences between experimental species samples and reference FOR samples by calculating the median value of the intensity differences as compared with a normal distribution scaled to the experimental mean and standard deviation of the normalized data. A *Z* score and *P* value were computed for each probe from that distribution. The statistically significant differential DNA methylation regions (DMR) were identified and *P* values associated with each region represented, as previously described (Guerrero-Bosagna et al. 2010).

A

Differential DNA Methylation Regions (DMR) (Epimutations)							
All probes ($p<10^{-5}$)				3 or more probes ($p<10^{-5}$)			
	(Up)	(Down)	Total		(Up)	(Down)	Total
FUL	116	398	514	FUL	76	8	84
SCA	211	679	890	SCA	17	144	161
PAR	191	1438	1629	PAR	28	578	606
CRA	361	2406	2767	CRA	61	1001	1062
Total	**Up**	**Down**	**Total Sites**	**Total**	**Up**	**Down**	**Total Sites**
	879	4921	5800		182	1731	1913

Copy Number Variation (CNV)							
All probes ($p<10^{-5}$)				3 or more probes ($p<10^{-5}$)			
	Gains	Loss	Total		Gains	Loss	Total
FUL	59	12	71	FUL	28	6	34
SCA	567	22	589	SCA	440	2	442
PAR	78	217	295	PAR	15	37	52
CRA	621	194	815	CRA	541	61	602
Total	**Gains**	**Loss**	**Total Sites**	**Total**	**Gains**	**Loss**	**Total Sites**
	1325	445	1770		1024	106	1130

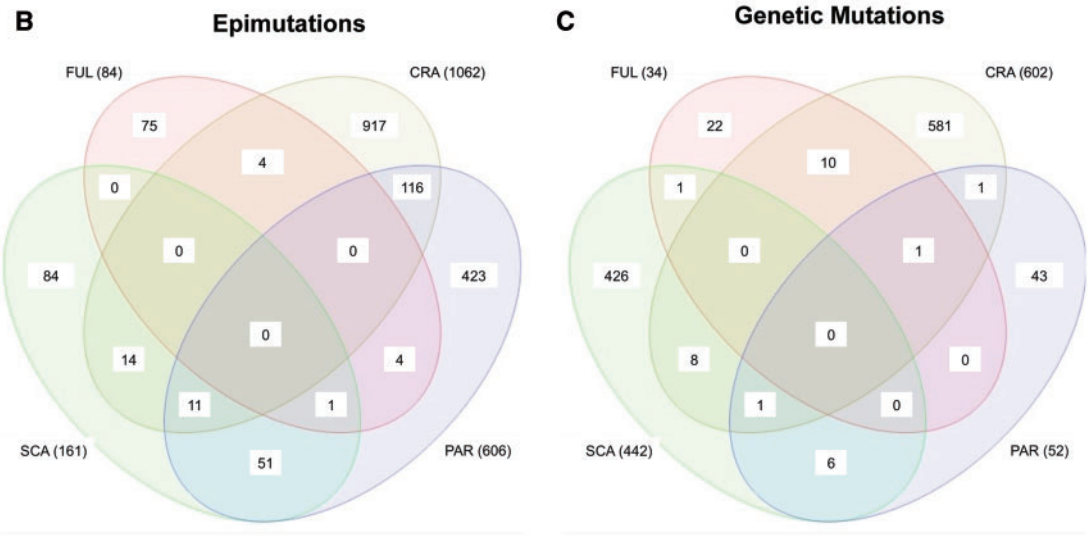


Fig. 2.—Number of epimutations and genetic mutations associated with Darwin's finches. The number of differential DMR epimutations and CNV genetic mutations (*A*). DMR and CNV that differ significantly ($P < 10^{-5}$) from the reference species (FOR) are presented for all oligonucleotide probes, compared with peaks of three or more adjacent probes. The epimutations with an increase (Up) or decrease (Down) in DNA methylation are indicated. Those genetic mutations with an increase (Gain) or decrease (Loss) in CNV are indicated. Venn diagrams for epimutations (*B*) and genetic mutations (*C*) show overlaps between epimutations (DMR) and genetic mutations (CNV) among species. The species and total number of sites compared are listed on the outside of each colored elliptical.

Additional Bioinformatics and Statistics

The July 2008 assembly of the zebra finch genome (taeGut1, WUSTL v3.2.4) produced by the Genome Sequencing Center at the Washington University in St Louis (WUSTL) School of Medicine was retrieved (WUSTL 2008). A seed file was constructed and a BSgenome package was forged for using the Finch DNA sequence in the R code (Herve Pages BSgenome: Infrastructure for Biostrings-based genome data packages. R

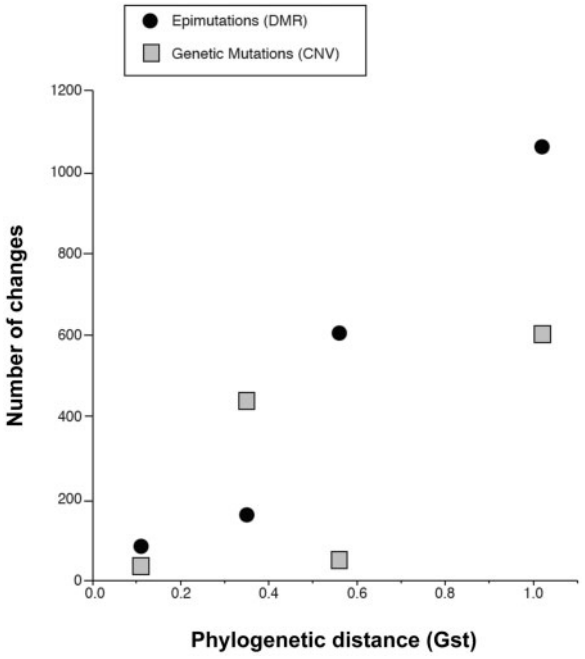


Fig. 3.—Phylogenetic distance is correlated with epigenetic changes, but not genetic changes. Branch lengths in figure 1 were used as measures of phylogenetic distance. The number of epimutations increased with phylogenetic distance (Spearman Rho = 1.0, $P < 0.0001$). In contrast, the number of genetic mutations did not increase with phylogenetic distance (Spearman Rho = 0.8, $P = 0.2$).

package version 1.24.0). This sequence was used to design the custom tiling arrays and to perform the bioinformatics.

The chromosomal location of CNV and DMR clusters used an R-code developed to find chromosomal locations of clusters (Skinner et al. 2012). A 2-Mb sliding window with 50,000 base intervals was used to find the associated CNV and DMR in each window. A *Z*-test statistical analysis with $P < 0.05$ was used on these windows to find the ones with overrepresented CNV and DMR were merged together to form clusters. A typical cluster region averaged approximately 3 Mb in size.

The DMR and CNV association with specific zebra finch genes and genome locations used the Gene NCBI database for zebra finch gene locations and correlated the epimutations associated (overlapped) with the genes. The three adjacent probes constituted approximately a 200-bp homology search. The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway associations were identified as previously described (Skinner et al. 2012). Statistically significant overrepresentation uses a Fisher's exact analysis.

Spearman Rank correlation coefficients were used to test for a relationship between phylogenetic distance and epigenetic and genetic changes (Whitlock and Schluter 2009).

Results

Phylogenetic relationships of the five finch species in this study are shown in figure 1. The taxa chosen for this study included: Two species of ground finches, FOR and FUL, which have crushing beaks with relatively deep bases; the cactus finch SCA, which has a long thin beak used for probing flowers; the small tree finch PAR, which has curved mandibles used for applying force at the tips; and the vegetarian finch CRA, which has a relatively short stubby bill used for crushing food along its entire length (Grant and Grant 2008; Donohue 2011; Rands et al. 2013). FOR was selected as a reference species for comparing genetic and epigenetic alterations among the remaining four species. Branch lengths in figure 1 were used as measures of phylogenetic distance.

The experimental design used purified erythrocytes from the different species. Although DNA sequences are the same for all cell types of an organism, the epigenome is distinct for each cell type, providing a molecular mechanism for the genome activity and functions that differ among different cell types (Skinner et al. 2010). Therefore, to investigate the overall epigenome requires a purified cell type. As birds have erythrocytes (red blood cells) that contain nuclei, samples of purified erythrocytes were collected from each of the Darwin's finch species to obtain DNA for molecular analysis.

The epigenetic alterations termed epimutations were assessed through the identification of differential DMR. The DMR were identified with the use of MeDIP with a methyl cytosine antibody, followed by a genome wide tiling array (Chip) for an MeDIP-Chip protocol (Guerrero-Bosagna et al. 2010). Although other epigenetic processes such as histone modifications, chromatin structure, and noncoding RNA are also important, DNA methylation is the best known epigenetic process associated with germline-mediated heritability and environmental manipulations (Skinner et al. 2010). Genetic variation was assessed using CNVs (i.e., amplifications and deletions of repeat elements) in the DNA using a CGH protocol (Pinkel and Albertson 2005; Gazave et al. 2011).

The reference genome used for the analysis was that of the zebra finch (*Taeniopygia guttata*) (Clayton et al. 2009), which had a preliminary estimate of greater than 83% similarity with a partial shotgun sequence of a Darwin's finch genome (Rands et al. 2013). This study actually suggests a much higher degree of identity. The zebra finch genome was tiled in a genome wide array with a 200-bp resolution and for a CGH array with a 1,500-bp resolution. These arrays were used in a competitive hybridization protocol between FOR (reference species) and the other four species (Guerrero-Bosagna et al. 2010). Differential hybridization using two different fluorescent DNA labeling tags identified the CNV with CGH using genomic DNA and the epimutation DMR with a MeDIP-Chip protocol. A statistical significance threshold of $P < 10^{-5}$ was set for the CNV or epimutation to be identified as a gain or loss compared with the reference species (fig. 2 and supplementary tables S1 and S2, Supplementary Material online). The data for all probes (oligonucleotides on the arrays) are presented. However, the criteria used to identify the CNV and DMR required the involvement of three or more adjacent

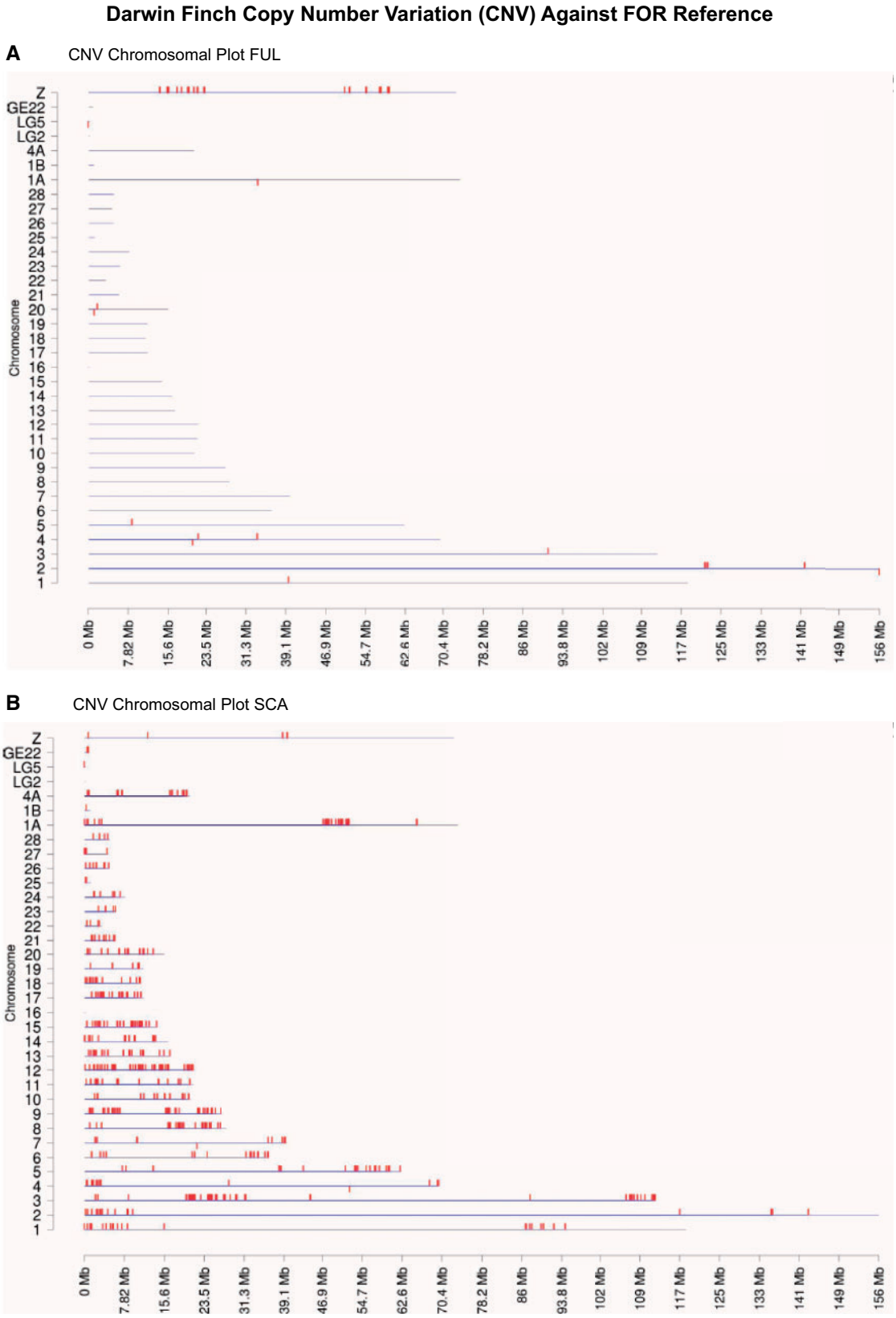


Fig. 4.—Chromosomal locations of the CNVs for each species. The chromosome number and size are presented in reference to the zebra finch genome. The chromosomal location of each CNV is marked with a red tick for FUL (*A*), SCA (*B*), PAR (*C*), and CRA (*D*).

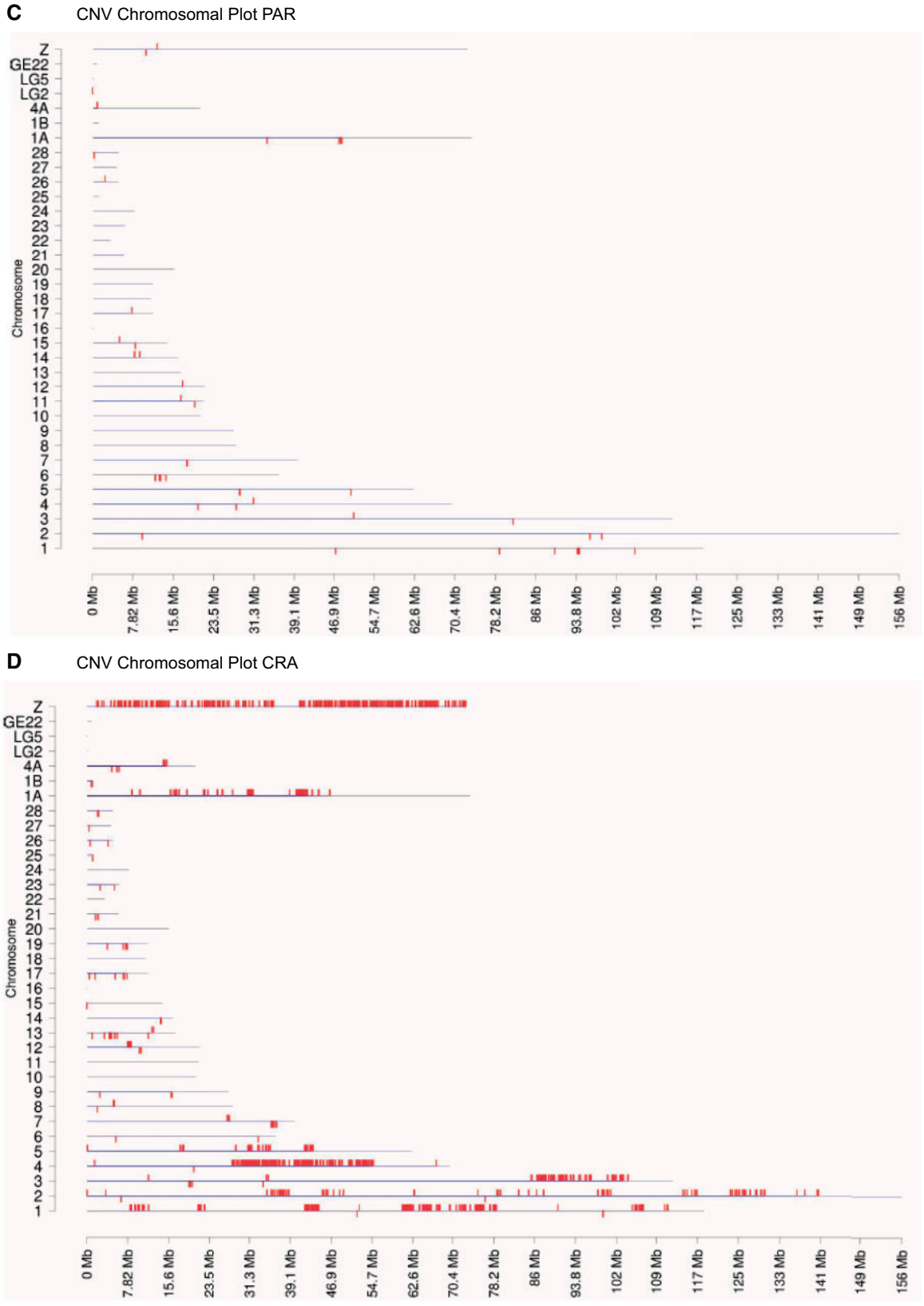


Fig. 4.—Continued.

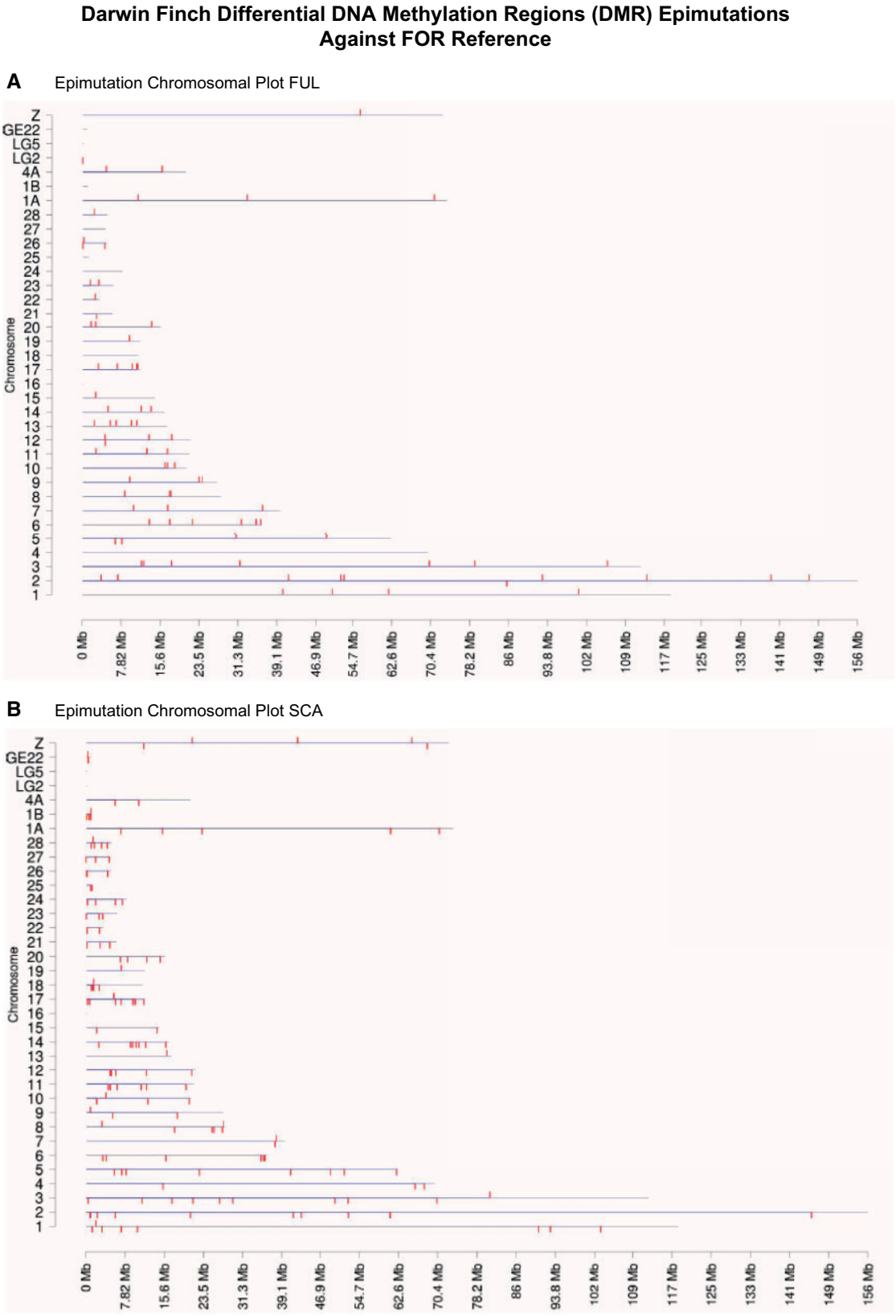
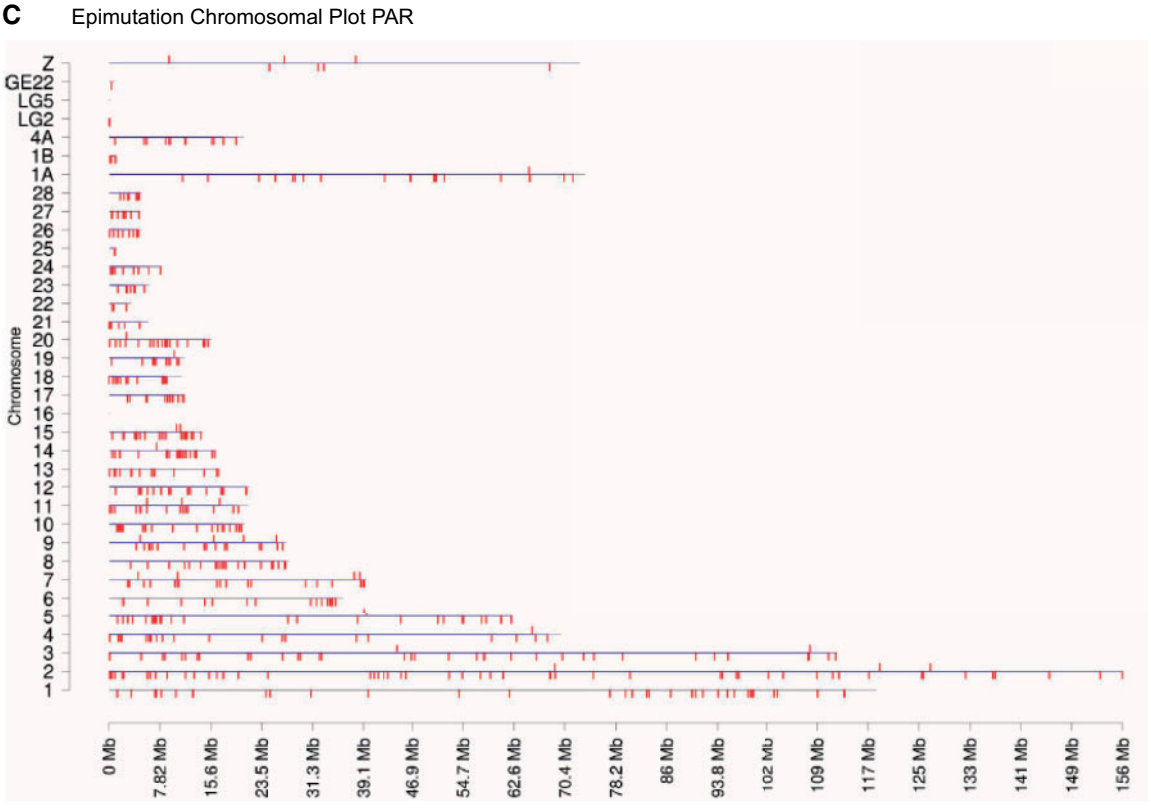


Fig. 5.—Chromosomal locations of the epimutations for each species. The chromosome number and size are presented in reference to the zebra finch genome. The chromosomal location of each DMR is marked with a red tick for FUL (*A*), SCA (*B*), PAR (*C*), and CRA (*D*).



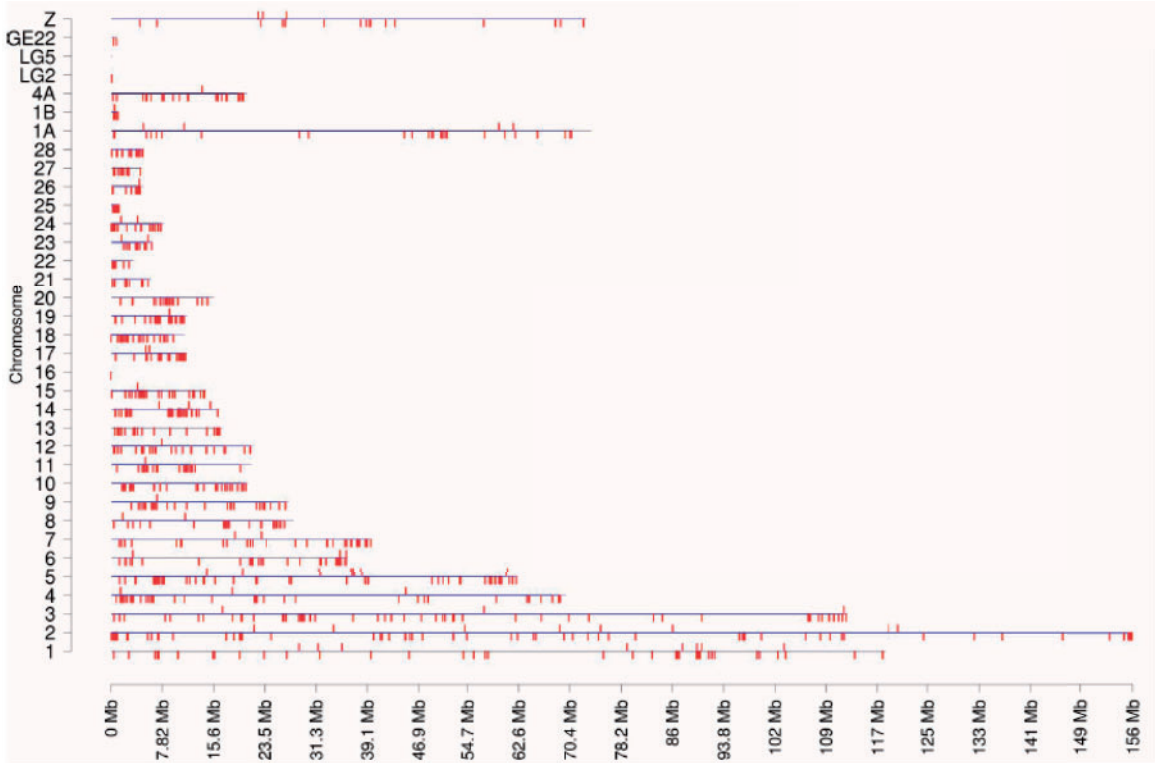


Fig. 5.—Continued.

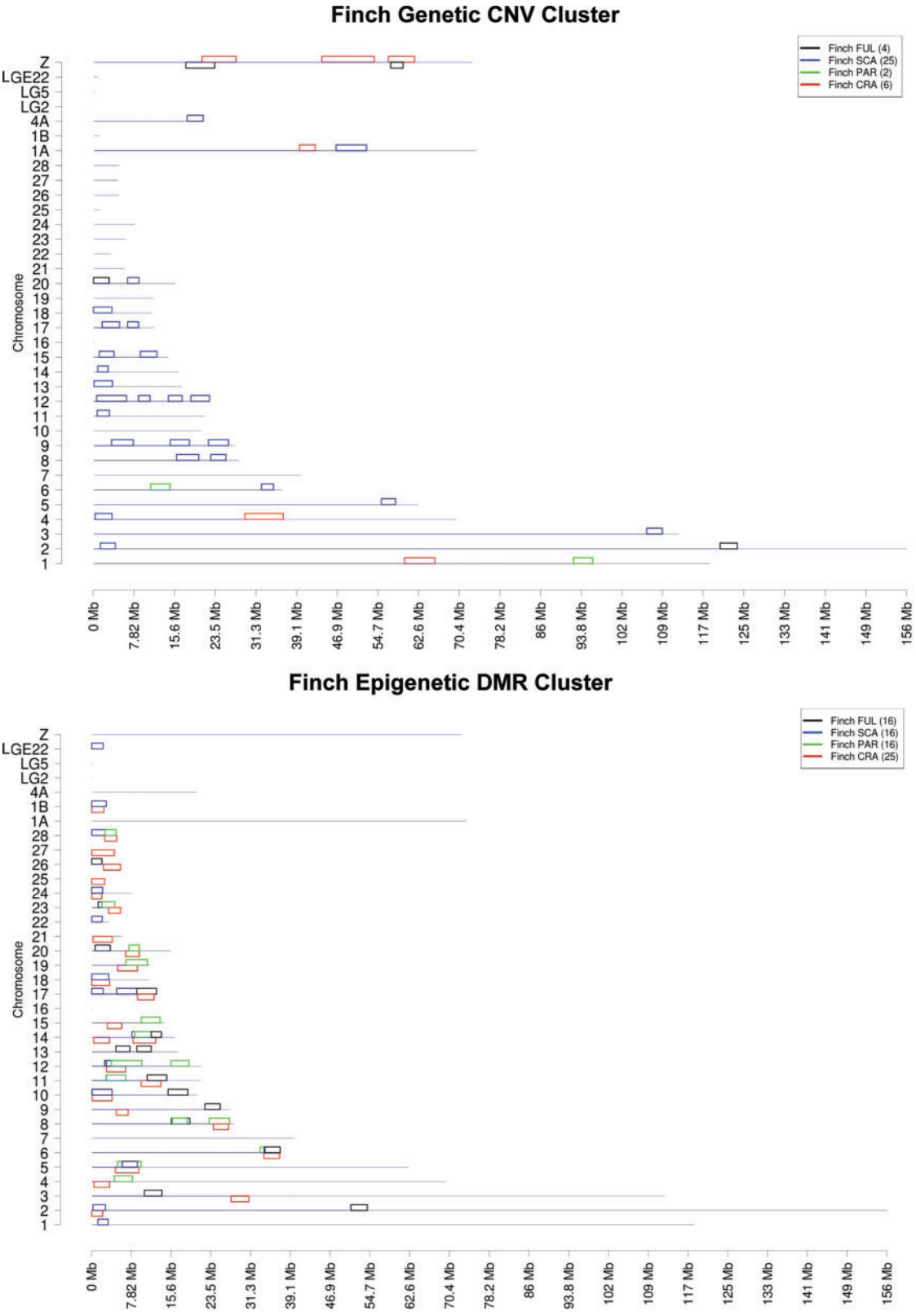


Fig. 6.—Chromosomal locations for clusters of CNV and DMR. The chromosome number and size are presented in reference to the zebra finch genome. The chromosomal location of statistically significant ($P < 10^{-5}$) overrepresented clusters of CNV (*A*) and DMR (*B*). The legend shows species and total number of clusters.

probes in the genome sequence having significant differential hybridization. These selection criteria reduce the number of false positives and provide a more reliable comparison (fig. 2). Therefore, the data presented used stringent criteria and represent the most reproducible epimutations and genetic CNV mutations among all three different experiments.

The increases or decreases in DNA methylation for the DMR are presented, along with the total number of epimutations in figure 2. The majority of epimutations for all species but FUL involves a decrease in DNA methylation (fig. 2*A*). The gains or losses in CNV are also presented, along with the total number of genetic alterations. The majority of genetic mutations for all species but PAR involves an increase in CNV number. Interestingly, the number of epimutations observed was generally higher, using the criteria selected, than the number of genetic alterations (fig. 2). However, the overall magnitude of epigenetic change was comparable to that of genetic change. Data for the five different species are shown in figure 1 for both epimutations (red) and genetic alterations (blue). The number of epimutations was significantly correlated with phylogenetic distance, whereas the number of genetic mutations was not (fig. 3).

The chromosomal locations of the CNV for the different finch species are shown in figure 4. CNVs were found on most chromosomes, with FUL having the least and CRA having the most. The chromosomal locations of the DMR epimutations for the different finch species are shown in figure 5. All chromosomes were found to have epimutations, with CRA having the highest number. These chromosomal plots suggested that some of the species might have clusters of CNV and/or DMR on some of the chromosomes (figs. 3 and 4). Therefore, a cluster analysis previously described (Skinner et al. 2012) was used to examine 50-kb regions throughout the genome to test for statistically significant ($P < 10^{-5}$) overrepresentation of CNV or DMR (fig. 6). Clusters, which have an average size of 3 Mb, are shown as species-specific boxes for CNV (fig. 6*A*) and for DMR (fig. 6*B*). Cluster characteristics and overlap are presented in supplementary table S3, Supplementary Material online. Clusters were obtained for all species, with a higher number of DMR clusters than CNV clusters. The highest number of CNV clusters was in SCA, with more than a 4-fold increase over CRA (fig. 6). Therefore, in addition to having more CNV than expected (assuming an increasing number with phylogenetic distance), SCA showed more CNV clusters than other species (fig. 2). Genome instability in these cluster regions may influence the increased numbers of CNV in SCA, which increases the presence of CNV clusters. In contrast, SCA did not show more DMR numbers or clusters than expected, assuming an increasing number with phylogenetic distance. Epimutation cluster overlap was more common among species (fig. 6 and table 1), suggesting that specific regions of the chromosomes were more susceptible to epigenetic alterations. Altered DNA methylation states have been experimentally shown to be stable for hundreds of generations (Cubas et al. 1999; Akimoto et al. 2007; Skinner et al. 2010).

Table 1
Cluster Overlap between Species

CNVs				
	CNV			
	FUL	**SCA**	**PAR**	**CRA**
FUL	4	0	0	2
SCA	0	25	0	0
PAR	0	0	2	0
CRA	2	0	0	6
Epimutations				
	DMR			
	FUL	**SCA**	**PAR**	**CRA**
FUL	16	5	6	7
SCA	5	16	8	11
PAR	6	8	16	11
CRA	7	11	11	25

NOTE.—The overlap of CNV or DMR clusters between species is presented for the CNVs and epimutations.

The potential overlaps in specific CNV or DMR sites among species were examined. The overlap in genetic mutations among the four species is shown in a Venn diagram in figure 2*C*, whereas the overlap in epimutations is shown in figure 2*B*. No overlap in specific CNV or DMR sites was observed among all species, and less than 10% overlap was generally observed between any two species. Interestingly, the CNV overlap between FUL and CRA was higher than for the other species (fig. 2*C*). Generally, genetic and epigenetic alterations were distinct between species, with the majority being species specific. The epimutations showed more overlap between species than the genetic CNV mutations (fig. 2*B* and table 1). In considering within species overlap between the CNV and epimutations, less than 3% had common genomic locations. Therefore, the epimutations do not appear to be linked to the genetic CNV mutations, but are distinct.

The final analysis examined the potential functional significance of the epimutations by examining DMR and genes known to be associated with avian evolution. Several gene families and cellular signaling pathways have previously been shown to be involved in bird evolution, including the bone morphogenic protein (BMP) family and pathway (Abzhanov et al. 2004; Badyaev et al. 2008), the toll receptor family and signaling pathway (Alcaide and Edwards 2011), and the melanins family and pathway (Mundy 2005). All the genes associated with these signaling pathways were localized on the finch genome and compared with the genomic locations of the epimutations and CNV. Epimutation-associated genes within the BMP pathway (fig. 7), toll pathway (fig. 8), and

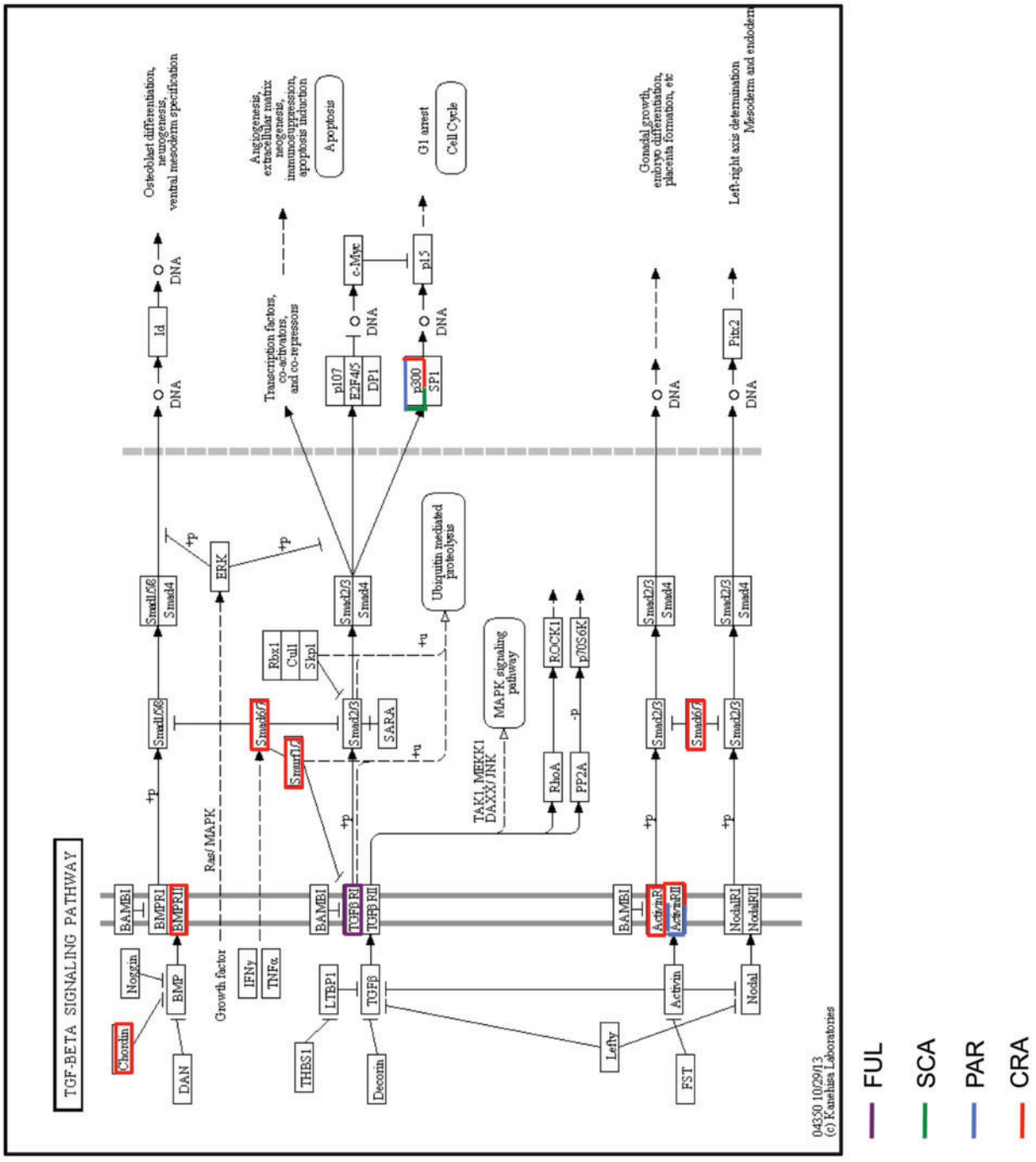


Fig. 7.—Epimutation-associated genes and correlated BMP pathway. The genes having associated epimutations in the signaling pathway presented for the different species are identified as FUL (purple), SCA (green), PAR (blue), and CRA (red) colored boxed genes.

melanin's pathway (fig. 9) are shown. Epimutations were overrepresented in all of these pathways (Fisher's exact test: BMP/TGFbeta (transforming growth factor) pathway, $P < 1 \times 10^{-6}$; toll pathway, $P < 5.7 \times 10^{-4}$; melanogenesis pathway, $P < 2.5 \times 10^{-13}$). Interestingly, the BMP pathway involved in beak development and shape had a statistically significant overrepresentation of CRA-associated epimutations when examined independently ($P < 2.7 \times 10^{-5}$) (fig. 7). In addition, the toll receptor pathway involved in immune response had a statistically significant overrepresentation of PAR-associated epimutations when examined independently ($P < 7.7 \times 10^{-4}$) (fig. 8). The melanogenesis pathway involved in color had a mixture of epimutations from most of the species when examined independently ($P < 7 \times 10^{-5}$) (fig. 9).

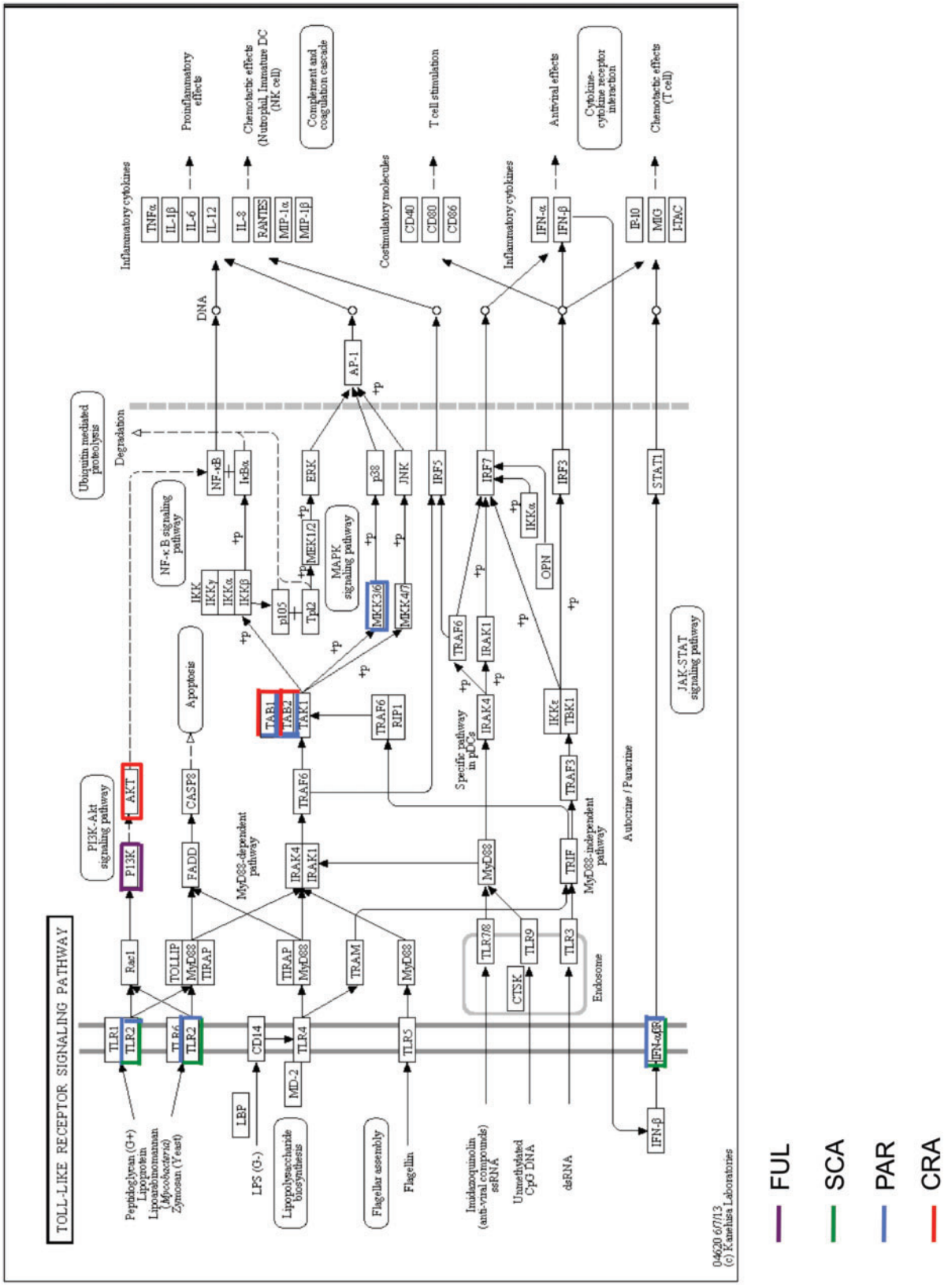


Fig. 8.—Epimutation-associated genes and correlated toll receptor pathway. The genes having associated epimutations in the signaling pathway presented for the different species are identified as FUL (purple), SCA (green), PAR (blue), and CRA (red) colored boxed genes.

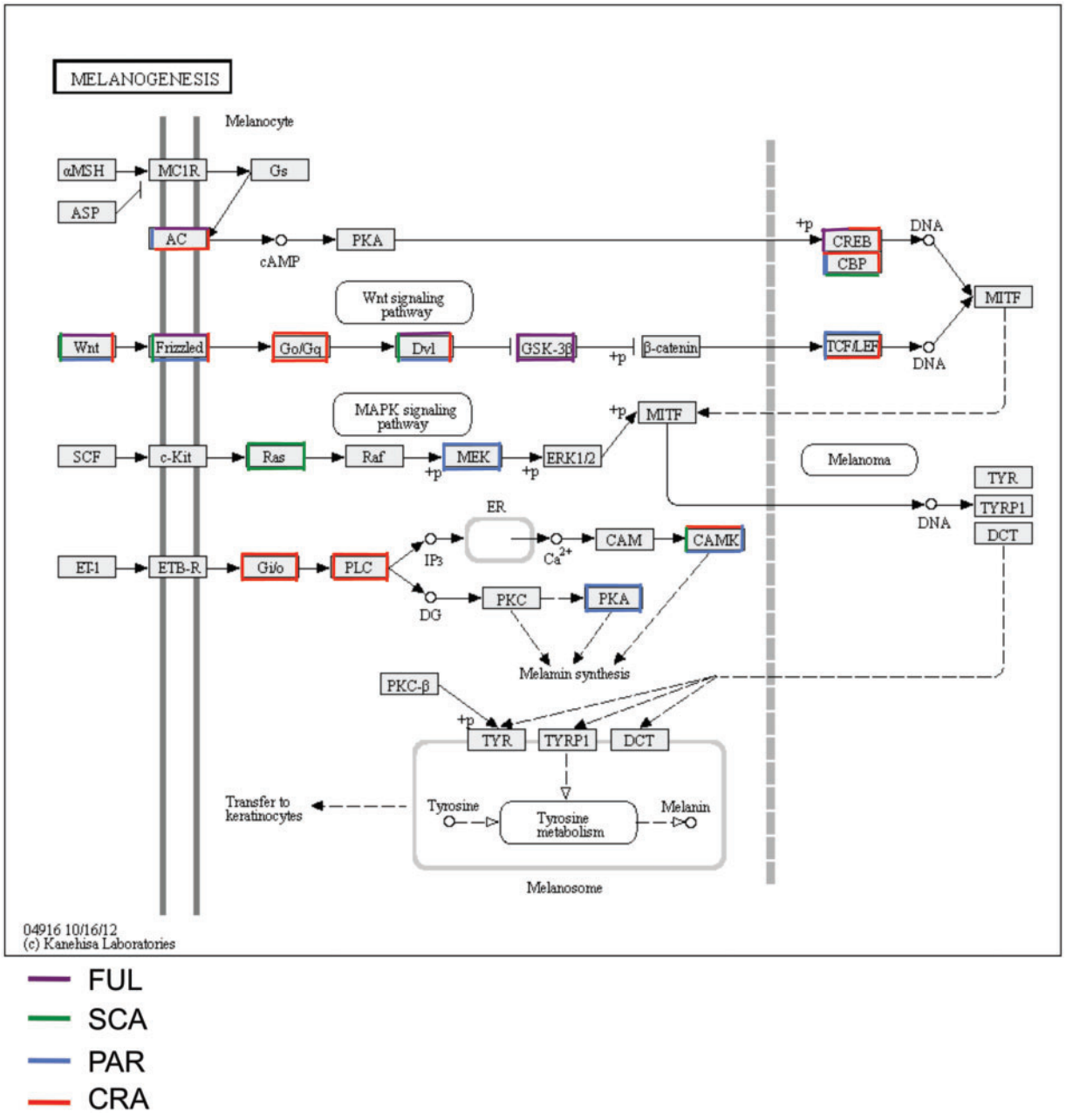


Fig. 9.—Epimutation-associated genes and correlated melanogenesis pathway. The genes having associated epimutations in the signaling pathway presented for the different species are identified as FUL (purple), SCA (green), PAR (blue), and CRA (red) colored boxed genes.

In addition to the pathway-specific genes, the total number of epimutations and CNV associated with genes are presented in table 2, with full lists in supplementary tables S4 and S5, Supplementary Material online. The epimutations and CNV for single probe and ≥3 probe identification are presented in table 2. Observations indicate that approximately half of the epimutations and CNV identified were associated with genes. Therefore, a high percentage of the epimutations and CNV identified were associated with genes and were statistically overrepresented in several gene pathways previously shown to be involved in particular aspects of avian evolution. Although this gene association analysis demonstrates that epimutations correlate with genes and important pathways, the functional or causal link to specific evolutionary processes remains to be investigated.

Discussion

This study provides one of the first genome-wide comparisons of genetic and epigenetic mutations among related species of

Table 2
Epimutation and CNV Gene Associations

CNVs				
	Total CNV 1+ Probes	Total CNV 3+ Probes	CNV Association with 14K Genes 1+ Probes	CNV Association with 14K Genes 3+ Probes
FUL	71	34	40	24
SCA	589	442	363	350
PAR	295	52	136	37
CRA	815	602	437	345
Epimutations				
	Total Epimutations 1+ Probes	Total Epimutations 3+ Probes	Epimutation Association with 14K Genes 1+ Probes	Epimutation Association with 14K Genes 3+ Probes
FUL	514	84	295	48
SCA	890	161	558	115
PAR	1,629	606	996	407
CRA	2,767	1,062	1,611	639

NOTE.—The 14,000 zebra finch genes annotated having epimutation or CNV associations are presented for the total number of associations (overlaps) for both regions identified with single (1+ probes) and adjacent (3+ probes) data sets.

organisms. There were relatively more epimutations than genetic CNV mutations among the five species of Darwin's finches, which suggests that epimutations are a major component of genome variation during evolutionary change. There was also a statistically significant correlation between the number of epigenetic differences and phylogenetic distance between finches (figs. 1 and 3), indicating that the number of epigenetic changes continues to accumulate over long periods of evolutionary time (2–3 Myr). In contrast, there was no significant relationship between the number of genetic CNV changes and phylogenetic distance.

The zebra finch genome was used as a reference for this study because a complete Darwin's finch genome is not yet available. The zebra finch genome showed hybridization with all probes on the array for each of the Darwin's finch species, suggesting that the genomes appear to be extremely similar. Loss of heterozygosity (absence of genomic regions, resulting in lack of probe hybridization) was not identified in any of the analyses. This suggests a high level of conservation and identity between the species' genomes. In the event the Darwin's finch genome has additional DNA sequence that is not present in the zebra finch genome, we would not have detected this DNA. Therefore, our data may be an underestimate of the Darwin's finch genome. Another technical limitation of our study was that we only considered genetic CNV (amplifications and deletions of repeat elements), but not other genetic variants such as point mutations or translocations. Although CNV frequency is higher than other mutations (e.g., SNPs) and stable in the genome (Gazave et al. 2011), this study's focus on CNV should kept in mind. The epimutations examined are differential DMR that have previously been shown to be frequent and transgenerationally stable (Anway et al. 2005; Guerrero-Bosagna et al. 2010; Skinner et al. 2010). Although other epigenetic processes such as histone modification, altered chromatin structure, and noncoding RNA may also be important, DNA methylation is the most established heritable epigenetic mark. This aspect of the experimental design should be kept in mind.

Among the five species of finches there were fewer genetic mutations (CNV) than epigenetic mutations. However, the cactus finch SCA showed a surprisingly large number of genetic CNV mutations than expected when compared with the reference species (FOR). The SCA mutations also clustered to similar locations on the genome to a greater extent than in the other species (fig. 6*A*). The reason for the disproportionately large number of CNV in the SCA comparison is unclear.

In contrast to the genetic mutation (CNV) analysis, the number of epimutations increased monotonically with phylogenetic distance (figs. 1 and 3). Overlap of specific epigenetic sites among species was minimal, including those for SCA (fig. 2*B*). An interesting possibility is that the epigenome may alter genome stability and generate genetic variation within species. A similar phenomenon has been shown for cancer, in which epigenetic alterations may precede genetic changes and alter genomic stability (Feinberg 2004). A decrease in the DNA methylation of specific repeat elements has previously been shown to correlate with an increase in CNV (Macia et al. 2011; Tang et al. 2012). Therefore, environmentally induced abnormal epigenetic shifts may influence genetic

mutations, such that a combination of epigenetics and genetics promotes phenotypic variation. Our observations demonstrate a relationship between the number of epigenetic changes and phylogenetic distance.

A comparison of the positions of epimutations and known gene families was also carried out. These gene families included those involved in the BMP pathway, which is related to beak shape (Badyaev et al. 2008), the toll receptor pathway, which is involved in immunological function (Alcaide and Edwards 2011), and the melanogenesis pathway, which affects color (Mundy 2005). Genes in all three of these families and signaling pathways were found to have species-specific epimutations (figs. 7–9). Future studies should focus on the causal relationship between epigenetic alterations and phenotypic traits.

Genetic mutations are postulated to provide much of the variation upon which natural selection acts (Gazave et al. 2011; Stoltzfus 2012). However, genetic changes alone are limited in their ability to explain phenomena ranging from the molecular basis of disease etiology to aspects of evolution (Skinner et al. 2010; Day and Bonduriansky 2011; Longo et al. 2012; Klironomos et al. 2013). Therefore, genetic mutations may not be the only molecular factors to consider (Richards 2006, 2009). Indeed, epigenetic and genetic changes may jointly regulate genome activity and evolution, as recent evolutionary biology modeling suggests (Day and Bonduriansky 2011; Klironomos et al. 2013). This integration of genetics and epigenetics may improve our understanding of the molecular control of many aspects of biology, including evolution.

Supplementary Material

Supplementary tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe. oxfordjournals.org/).

Acknowledgments

The authors acknowledge the advice and critical reviews of Dr Jeb Owen (WSU), Dr Hubert Schwabl (WSU), Dr David Crews (U Texas Austin), Dr Kevin P. Johnson (U Illinois) and Dr Sarah Bush (Utah). They thank Ms Sean Leonard, Ms Shelby Weeks, Dr C. Le Bohec, Mr O. Tiselma and Mr R. Clayton for technical assistance and Ms Heather Johnson for assistance in preparation of the manuscript. The research was supported by the National Institute of Health grants to M.K.S. and National Science Foundation grants to D.H.C. M.K.S. conceived the study. M.K.S. and D.H.C. designed the study. C.G.B., M.M.H., E.E.N., J.A.H.K., S.A.K., and D.H.C. performed the experiments and acquired the data. All authors analyzed the data. M.K.S. and D.H.C. wrote the manuscript. All authors edited and approved the manuscript. The authors declare no competing financial interests.

Literature Cited

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science 305(5689):1462–1465.

Akimoto K, et al. 2007. Epigenetic inheritance in rice plants. Ann Bot. 100(2):205–217.

Alcaide M, Edwards SV. 2011. Molecular evolution of the toll-like receptor multigene family in birds. Mol Biol Evol. 28(5):1703–1715.

Anway MD, Cupp AS, Uzumcu M, Skinner MK. 2005. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science 308(5727):1466–1469.

Badyaev AV, Young RL, Oh KP, Addison C. 2008. Evolution on a local scale: developmental, functional, and genetic bases of divergence in bill form and associated changes in song structure between adjacent habitats. Evolution 62(8):1951–1964.

Bonduriansky R. 2012. Rethinking heredity, again. Trends Ecol Evol. 27(6): 330–336.

Clayton DF, Balakrishnan CN, London SE. 2009. Integrating genomes, brain and behavior in the study of songbirds. Curr Biol. 19(18): R865–R873.

Crews D, et al. 2007. Transgenerational epigenetic imprints on mate preference. Proc Natl Acad Sci U S A. 104(14):5942–5946.

Cubas P, Vincent C, Coen E. 1999. An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401(6749):157–161.

Day T, Bonduriansky R. 2011. A unified approach to the evolutionary consequences of genetic and nongenetic inheritance. Am Nat. 178(2):E18–E36.

Donohue K. 2011. Darwin's finches: readings in the evolution of a scientific paradigm. Chicago (IL): University of Chicago Press, p. 492.

Endler J. 1986. Natural selection in the wild. Princeton (NJ): Princeton University Press.

Feinberg AP. 2004. The epigenetics of cancer etiology. Semin Cancer Biol. 14(6):427–432.

Flatscher R, Frajman B, Schönswetter P, Paun O. 2012. Environmental heterogeneity and phenotypic divergence: can heritable epigenetic variation aid speciation? Genet Res Int. 2012:698421.

Gazave E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res. 21(10):1626–1639.

Geoghegan JL, Spencer HG. 2012. Population-epigenetic models of selection. Theor Popul Biol. 81(3):232–242.

Geoghegan JL, Spencer HG. 2013a. Exploring epiallele stability in a population-epigenetic model. Theor Popul Biol. 83:136–144.

Geoghegan JL, Spencer HG. 2013b. The adaptive invasion of epialleles in a heterogeneous environment. Theor Popul Biol. 88:1–8.

Geoghegan JL, Spencer HG. 2013c. The evolutionary potential of paramutation: a population-epigenetic model. Theor Popul Biol. 88:9–19.

Grant P, Grant R. 2008. How and why species multiply: the radiation of Darwin's finches. Princeton (NJ): Princeton University Press.

Greenspan RJ. 2009. Selection, gene interaction, and flexible gene networks. Cold Spring Harb Symp Quant Biol. 74:131–138.

Guerrero-Bosagna C, Sabat P, Valladares L. 2005. Environmental signaling and evolutionary change: can exposure of pregnant mammals to environmental estrogens lead to epigenetically induced evolutionary changes in embryos? Evol Dev. 7(4):341–350.

Guerrero-Bosagna C, Settles M, Lucker B, Skinner MK. 2010. Epigenetic transgenerational actions of vinclozolin on promoter regions of the sperm epigenome. PLoS One 5(9):e13100.

Holeski LM, Jander G, Agrawal AA. 2012. Transgenerational defense induction and epigenetic inheritance in plants. Trends Ecol Evol. 27: 618–626.

Huber SK, et al. 2010. Ecoimmunity in Darwin's finches: invasive parasites trigger acquired immunity in the medium ground finch (*Geospiza fortis*). PLoS One 5(1):e8605.

Huttley GA. 2004. Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. Mol Biol Evol. 21(9):1760–1768.
Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 8(4):253–262.
Klironomos FD, Berg J, Collins S. 2013. How epigenetic mutations can affect genetic evolution: model and mechanism. Bioessays 35(6): 571–578.
Koop JA, Huber SK, Laverty SM, Clayton DH. 2011. Experimental demonstration of the fitness consequences of an introduced parasite of Darwin's finches. PLoS One 6(5):e19706.
Kuzawa CW, Thayer ZM. 2011. Timescales of human adaptation: the role of epigenetic processes. Epigenomics 3(2):221–234.
Lack D. 1947. Darwin's finches. Cambridge University Press.
Lamarck JB. 1802. Recherches sur l'organisation des corps vivans. Paris: Chez L'auteur, Maillard.
Liebl AL, Schrey AW, Richards CL, Martin LB. 2013. Patterns of DNA methylation throughout a range expansion of an introduced songbird. Integr Comp Biol. 53(2):351–358.
Longo G, Miquel PA, Sonnenschein C, Soto AM. 2012. Is information a proper observable for biological organization? Prog Biophys Mol Biol. 109(3):108–114.
Lupski JR. 2007. An evolution revolution provides further revelation. Bioessays 29(12):1182–1184.
Macia A, et al. 2011. Epigenetic control of retrotransposon expression in human embryonic stem cells. Mol Cell Biol. 31(2):300–316.
Manikkam M, Guerrero-Bosagna C, Tracey R, Haque MM, Skinner MK. 2012. Transgenerational actions of environmental compounds on reproductive disease and epigenetic biomarkers of ancestral exposures. PLoS One 7(2):e31901.
Mundy NI. 2005. A window on the genetics of evolution: MC1R and plumage colouration in birds. Proc Biol Sci. 272(1573):1633–1640.
Nozawa M, Kawahara Y, Nei M. 2007. Genomic drift and copy number variation of sensory receptor genes in humans. Proc Natl Acad Sci U S A. 104(51):20421–20426.
Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5(4):557–572.
Petren K, Grand BR, Grant PR. 1999. A phylogeny of Darwin's finches based on microsatellite DNA lenght variation. Proc R Soc Lond B. 266(1417):321–329.
Picard F, Robin S, Lavielle M, Vaisse C, Daudin J-J. 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 6:27.
Pinkel D, Albertson DG. 2005. Comparative genomic hybridization. Annu Rev Genomics Hum Genet. 6:331–354.
Poptsova M, Banerjee S, Gokcumen O, Rubin MA, Demichelis F. 2013. Impact of constitutional copy number variants on biological pathway evolution. BMC Evol Biol. 13:19.
R Development Core Team. 2010. R: a language for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org.
Rands CM, et al. 2013. Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. BMC Genomics 14:95.
Rebollo R, Horard B, Hubert B, Vieira C. 2010. Jumping genes and epigenetics: towards new species. Gene 454(1–2):1–7.
Richards CL, Bossdorf O, Pigliucci M. 2010. What role does heritable epigenetic variation play in phenotypic evolution? BioScience 60: 232–237.
Richards EJ. 2006. Inherited epigenetic variation—revisiting soft inheritance. Nat Rev Genet. 7(5):395–401.
Richards EJ. 2009. Quantitative epigenetics: DNA sequence variation need not apply. Genes Dev. 23(14):1601–1605.
Skinner MK. 2011. Environmental epigenetic transgenerational inheritance and somatic epigenetic mitotic stability. Epigenetics 6(7): 838–842.
Skinner MK, Anway MD, Savenkova MI, Gore AC, Crews D. 2008. Transgenerational epigenetic programming of the brain transcriptome and anxiety behavior. PLoS One 3(11):e3745.
Skinner MK, Manikkam M, Guerrero-Bosagna C. 2010. Epigenetic transgenerational actions of environmental factors in disease etiology. Trends Endocrinol Metab. 21(4):214–222.
Skinner MK, Mohan M, Haque MM, Zhang B, Savenkova MI. 2012. Epigenetic transgenerational inheritance of somatic transcriptomes and epigenetic control regions. Genome Biol. 13(10):R91.
Skinner MK, Savenkova MI, Zhang B, Gore AC, Crews D. 2014. Gene bionetworks involved in epigenetic transgenerational inheritance of altered mate preference: environmental epigenetics and evolutionary biology. BMC Genomics 15:377.
Slatkin M. 2009. Epigenetic inheritance and the missing heritability problem. Genetics 182(3):845–850.
Stoltzfus A. 2012. Constructive neutral evolution: exploring evolutionary theory's curious disconnect. Biol Direct. 7:35.
Sudmant PH, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. Genome Res. 23:1373–1382.
Tang MH, et al. 2012. Major chromosomal breakpoint intervals in breast cancer co-localize with differentially methylated regions. Front Oncol. 2:197.
Tateno H, Kimura Y, Yanagimachi R. 2000. Sonication per se is not as deleterious to sperm chromosomes as previously inferred. Biol Reprod. 63(1):341–346.
Tibshirani R, Wang P. 2008. Spatial smoothing and hot spot detection for CGH data using the fused lasso. Biostatistics 9(1):18–29.
Ward WS, Kimura Y, Yanagimachi R. 1999. An intact sperm nuclear matrix may be necessary for the mouse paternal genome to participate in embryonic development. Biol Reprod. 60(3):702–706.
Whitlock MC, Schluter D. 2009. The analysis of biological data. Greenwood Village (CO): Roberts and Company Publishers.
WUSTL. 2008. Jul. 2008 assembly of the zebra finch genome (taeGut1, WUSTL v3.2.4), as well as repeat annotations and GenBank sequences, Database Provider, NCBI.
Ying H, Huttley G. 2011. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. Genome Biol Evol. 3:938–949.

Associate editor: Bill Martin