Metagenome Skimming of Insect Specimen Pools: Potential for Comparative Genomics

Benjamin Linard[1], Alex Crampton-Platt[1,2], Conrad P.D.T. Gillett[1], Martijn J.T.N. Timmermans[1,3] and Alfried P. Vogler[1,3,*]

[1]Department of Life Sciences, Natural History Museum, London, United Kingdom

[2]Department of Genetics, Evolution and Environment, University College London, United Kingdom

[3]Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot, United Kingdom

*Corresponding author: E-mail: a.vogler@imperial.ac.uk.

Accepted: May 9, 2015

Abstract

Metagenomic analyses are challenging in metazoans, but high-copy number and repeat regions can be assembled from low-coverage sequencing by "genome skimming," which is applied here as a new way of characterizing metagenomes obtained in an ecological or taxonomic context. Illumina shotgun sequencing on two pools of Coleoptera (beetles) of approximately 200 species each were assembled into tens of thousands of scaffolds. Repeated low-coverage sequencing recovered similar scaffold sets consistently, although approximately 70% of scaffolds could not be identified against existing genome databases. Identifiable scaffolds included mitochondrial DNA, conserved sequences with hits to expressed sequence tag and protein databases, and known repeat elements of high and low complexity, including numerous copies of rRNA and histone genes. Assemblies of histones captured a diversity of gene order and primary sequence in Coleoptera. Scaffolds with similarity to multiple sites in available coleopteran genome sequences for *Dendroctonus* and *Tribolium* revealed high specificity of scaffolds to either of these genomes, in particular for high-copy number repeats. Numerous "clusters" of scaffolds mapped to the same genomic site revealed intra- and/or intergenomic variation within a metagenome pool. In addition to effect of taxonomic composition of the metagenomes, the number of mapped scaffolds also revealed structural differences between the two reference genomes, although the significance of this striking finding remains unclear. Finally, apparently exogenous sequences were recovered, including potential food plants, fungal pathogens, and bacterial symbionts. The "metagenome skimming" approach is useful for capturing the genomic diversity of poorly studied, species-rich lineages and opens new prospects in environmental genomics.

Key words: environmental genomics, repetitive DNA, histone genes, Coleoptera, bacterial endosymbionts, genome evolution.

Introduction

Environmental genomics provides unparalleled opportunities to understand the diversity of genes and genomes, their evolution, and their interactions with the environment. Thanks to increasingly efficient high-throughput sequencing platforms, many analyses of biological diversity are moving into an era of genomic sequencing (Valentini et al. 2009; Metzker 2010; Sucher et al. 2012; Taberlet, Coissac, Hajibabaei, et al. 2012). In particular, metabarcoding and metagenomic approaches are now widely used to link species diversity and environmental factors. Metabarcoding generally uses polymerase chain reaction (PCR)-based sequencing of short homologous gene regions across species assemblages or environmental samples, to uncover pattern of species distribution and species diversity, which is applied to both prokaryotes and multicellular eukaryotes (Ficetola et al. 2008; Taberlet, Coissac, Pompanon, et al. 2012; Ji et al. 2013). On the other hand, metagenomic approaches are based on shotgun sequencing of DNA from species mixtures, which are used mainly in prokaryotic research, with some recent examples of unicellular eukaryotes (see Raven 2012). Metagenomics already has given insights into the functional diversity of gut, soil, or marine microbial communities (Yang et al. 2013), and it is leading to new applications such as bioprospecting (Lee and Lee 2013) or carbon storage predictions (Trivedi et al. 2013). However, establishing broader links between metagenomics and ecology remains challenging despite novel study designs and innovative statistical modeling (Yoccoz 2012).

In this methodological context, arthropods are a challenging animal phylum for which both species diversity and genomic diversity remain poorly explored. Their presence in every terrestrial and water environment and their essential role in most biomes (Wheeler 1982) highlights the importance of studying their role in the ecosystem. Yet, their great species diversity of possibly 2–20 million species (Zhang 2011; Basset et al. 2012) is poorly represented by the few dozen completed genome sequences currently available. For instance, Coleoptera are the largest Hexapod order with approximately 400,000 described species (May 2010), but only two complete nuclear genomes have been assembled and characterized by 2015 (Friedrich and Muqim 2003; Keeling et al. 2013). This leaves a huge amount of genomic and evolutionary diversity to be explored that ultimately could be linked to specific ecological traits. As it is unlikely that we will be able to obtain full-genome sequences for a large proportion of these species, it may be an attractive option to conduct metagenomic sequencing on numerous representatives of a lineage or entire species assemblages to complement full-genome sequencing. However, it remains unclear what can be learned about genomic and functional diversity in arthropods and other metazoan lineages from shotgun sequencing of mixed species assemblages, and at what cost and effort.

As first step toward the comparative genomics of mixed species assemblages, we conducted low-coverage shotgun sequencing of bulk samples composed of hundreds of coleopteran species. Such shallow genome sequencing, commonly referred to as "genome skimming" (GS), samples the most prevalent DNA elements present in the sample that are assembled readily from the most abundant raw reads (Straub et al. 2012). Several recent botanical studies showed that GS is sufficient to extract a panel of multicopy markers, mainly chloroplast DNA (cpDNA), mitochondrial DNA (mtDNA), rRNA genes, and conserved nuclear repeats, which have been used to build high-quality phylogenetic trees (Bock et al. 2014; Malé et al. 2014). To our knowledge, shallow genomic sequencing and recovery of such markers has not been tested on insect samples, and has not been applied to pools of mixed specimens. (Only mtDNA was exploited in specific cases, see Zhou et al. 2013; Gillett et al. 2014; Tang et al. 2014; Andújar et al. 2015; Crampton-Platt et al. 2015; Gómez-Rodríguez et al. 2015.)

Yet, shallow sequencing from specimen mixtures should equally provide high-copy and repeat regions present within the individual genomes, thus extending the GS approach to "metagenome skimming" (MGS). A challenge arises from the fact that these reads and the resulting assemblies cannot be assigned directly to any of the genomes in the mixtures, which complicates the study of genomic organization and evolutionary diversity of these elements. Yet, the scaffolds obtained from entire species pools that capture the genomic diversity of numerous species collectively may be a powerful approach to assess the conserved and repeated elements present in a set of species and to study their sequence variation and phyletic distribution, as a novel means for comparative genomics. In association with fully sequenced reference genomes, these metagenomic assemblies may be associated with particular clades and functional groups.

Here, we assessed what kind of genomic information can be extracted from low-coverage metagenome sequencing of two specimen pools that were originally generated to address questions about taxonomic (Gillett et al. 2014) and ecological diversity (Crampton-Platt et al. 2015). These existing analyses were performed on the mtDNA fraction of the sequence data only ("mitochondrial metagenomics"; Crampton-Platt et al. 2015), but the much greater nuclear portion of the sequence data was ignored in these studies. It is interrogated here to obtain insights into the genomic diversity of Coleoptera. High-abundance reads producing the scaffolds in MGS are either derived from orthologous loci conserved among multiple genomes, or they are derived from paralogous copies, for example, from repeat elements present in high-copy numbers (*hcn*) within a genome, but they may also arise from a combination of orthologous and paralogous sequences (fig. 1). Short shotgun reads therefore produce a mixture of assembled contigs but their composition may be a largely random outcome of an idiosyncratic assembly process or the chance composition of the pool of reads. As a first step toward the characterization of the metagenomes, we establish if scaffolds are encountered consistently and at what sequencing depth, to identify the recognizable high copy fraction obtained from pools of particular phyletic composition. Next, we attempted to annotate the resulting scaffolds against existing databases, including collections of known repeats, and identify potential conserved coding regions, such as gene families and tandemly repeated genes. Mapping of scaffolds against the two available reference genomes can further provide information on the intragenomic organization and their intergenomic distribution across evolutionary lineages. Vice versa, the number and distribution of scaffolds mapped against full genome sequences can contribute a new approach to comparative genomics, and specifically to the analysis of the repetitive fraction that is notoriously difficult to characterize with standard genome sequencing methods. Finally, the scaffolds may represent the associated fauna and flora, including the microbiome and potential food sources, which provide information on the wider ecosystem in which the specimens partake.

Materials and Methods

Libraries of Bulk Samples Used

The sequencing pools (Gillett et al. 2014; Crampton-Platt A, et al., under review) differ in number of specimens and taxonomic diversity of the included species. The *Weevil* pool (Gillett et al. 2014) includes unique representatives of 173 species of

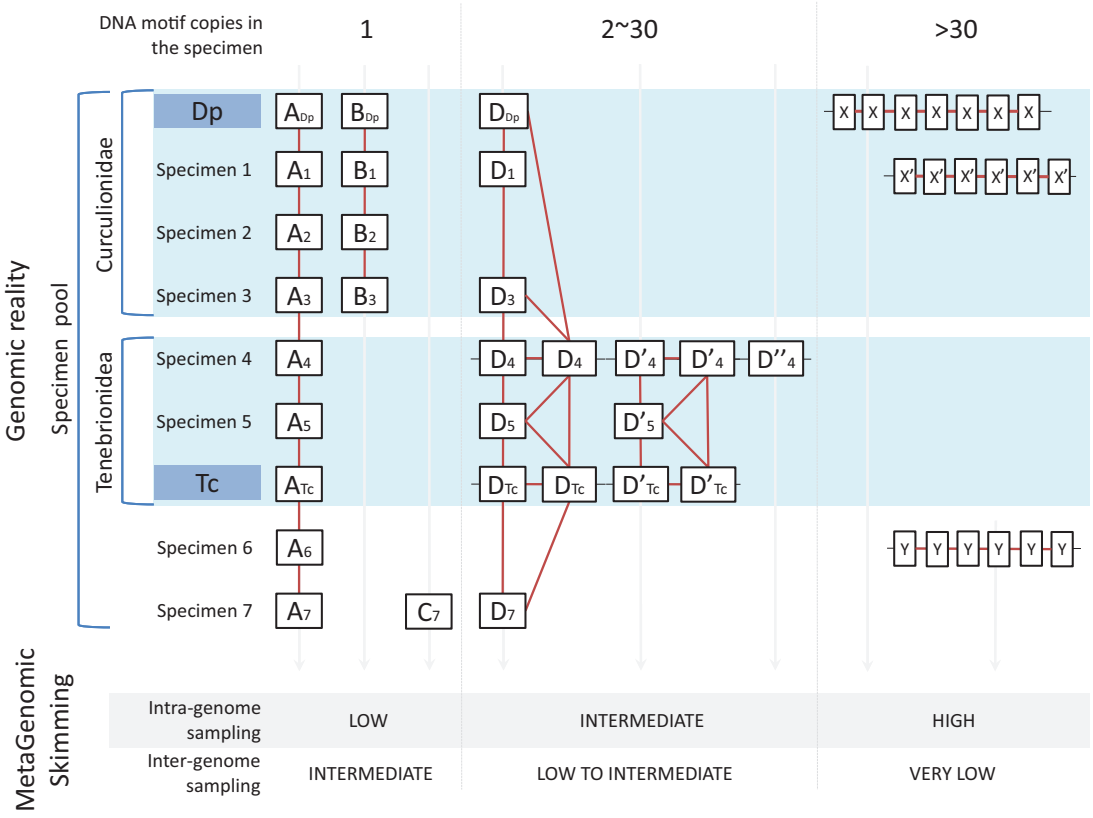

Fig. 1.—Hypothetical scenarios of scaffold formation from low-coverage DNA sequencing of specimen pools. The figure represents specimens in the superfamilies Tenebrionoidea, Curculionidea, and other coleopteran superfamilies represented by two reference genomes for *Tc* and *Dp*. Eight scenarios of scaffold formation (*A*, *B*, *C*, *D*, *D*′, *D*″, *x*, and *y*) are depicted along gray vertical arrows and represent the aggregation of similar DNA motifs (white boxes) into a single scaffold (red lines). The horizontal axis from left to right represent an increasing intragenomic copy number of a locus, and the vertical axis represent the greater phylogenetic distance of taxa. The first three scenarios (*A*, *B*, *C*) represent single copy motifs. *A* and *B* are phylogenetically conserved and their presence across specimens will increase the rate of recovery. Their homology to the reference genomes depends on phylogenetic conservation and the distance from available reference genomes (scenario *A* vs. scenario *B*). These simple scenarios are overlain on the effects of copy number and variation among paralogs. Scenario *D* represents several copies of the same DNA motif present in different genome locations and similar enough to be aggregated into the same scaffold. Motifs *D*′ and *D*″ are homologous but less similar and will be aggregated into two other scaffolds. The sampling probability of these motifs is increased by higher copy number and wider conservation over the specimens. The probability to generate a scaffold is decreasing from *D*, *D*′ to *D*″. Copy number information is partially lost during their scaffold aggregation process. Finally, high-copy number genomic repeats (scenarios *x*, *x*′, and *y*), may produce scaffolds even if they are limited to a single genome in the mixture. Repeats *x*′ is aggregated into a single scaffold and can be identified by similarity to repeat *x*, present in the closely related *Dp* genome. The repetitive and taxonomic nature of *y* cannot be deduced as no closely related reference genome is available to observe a similar motif. The bottom of the figure depicts the probability that a particular kind of locus is assembled from shotgun reads derived from within and among genomes.

weevils (Curculionoidea) in 7 families and 15 subfamilies of the largest family, Curculionidae. The *Canopy* sample (Crampton-Platt et al. 2015) contains DNA extractions from 477 specimens in greater than 200 morphospecies from canopy fogging and includes a representative set of most major lineages of Coleoptera from 3 suborders and 14 superfamilies, including Curculionoidea. Specimen identification and abundance are described in supplementary file S1, Supplementary Material online. Hence, the *Canopy* sample covers a much wider phyletic spectrum than *Weevil*, which is taxonomically nested within the former. This is reflected in the genetic distances based on existing COI data for both pools, which showed a much smaller range of divergences in *Weevil* than *Canopy* (supplementary table S1, Supplementary Material online). Each pool was generated by combining total genomic DNA extracted from individual specimens. For the *Weevil* pool, DNA for each individual (species) was added in equal amounts where possible, whereas for the *Canopy* pool equal volumes from each extraction were added, which resulted in greatly different concentrations due to the different body sizes and relative abundances of species. The Illumina libraries were generated with TruSeq kits (v. 2.0, PE) and insert sizes of 850 bp (*Canopy_Long*) and 480 bp (*Canopy_Short*), respectively, or with the

Illumina Nextera (Bentley et al. 2008) library kit (*Canopy_Next*). The *Weevil* library was generated with TruSeq and had an average insert size of 790 bp. Each library was sequenced on a full flow cell of an Illumina MiSeq sequencer and 2 × 250 bp paired-end reads. The entire set of scaffolds assembled independently in the three *Canopy* libraries will be referred as *Canopy_merged* (*Canopy_Long* + *Canopy_Short* + *Canopy_Next* scaffolds) to discuss the resampling of a same insect pool.

Sequence Quality Control and Assembly

Remaining library adapter sequences were removed with Trimmomatic using default parameters (Lohse et al. 2012). All libraries were assembled with version 7.0 of the Celera package using default settings (Miller et al. 2010). Read quality control was part of the Celera pipeline and was based on default parameters. Assemblies required 128 Gb RAM and two 2.40 GHz Intel Xeon cores. Assembly time was up to 6 days (*Weevil* library).

Coleopteran Complete Genomes

The genomes of *Tribolium castaneum* (*Tc*) (Friedrich and Muqim 2003) and *Dendroctonus ponderosae* (*Dp*) (Keeling et al. 2013) were retrieved for comparative analyses. We used the latest assembly of *Tc* (Tcas 3.0; NCBI project accession AAJJ00000000), considering the ten main chromosome linkage groups (accessions CM000276–CM000285), the unplaced scaffolds (accessions DS497665–DS497969) and unplaced singletons (GG694051–GG695898). Similarly, we downloaded the *Dp* draft genome with the identifier GCA_000355655.1. For both species, gene and intron/exon coordinates were retrieved through the annotations of the corresponding entries.

Taxonomic Assignation

Scaffolds were submitted to a custom Java pipeline retrieving their closest homologs in several NCBI databases. First, low complexity regions (simple sequence repeats, SSRs), were de novo predicted and masked with RepeatMasker using default settings (Tarailo-Graovac and Chen 2009) and any scaffolds with greater than 50% masked were discarded to avoid BLAST search slowdown. Then, each scaffold was categorized based on the best-hit sequence identified by discontiguous-megablast in the NCBI complete genome database (ftp.ncbi.nlm.nih.gov/blast/db/other_genomic; 2 September 2013; sequence identity greater than 70%, alignment length >150 bp, E<1e−9), to either "Hexapoda" (best hit to a Hexapoda genome), "non-Hexapoda" (best hit to a genome other than Hexapoda), and "unidentified" (no hit at the selected threshold). Such classification is strongly biased by the phylogenetic content of the database, for example, only two complete beetle genomes were available but dozens of Diptera and Hymenoptera genomes. Consequently, supplementary BLAST identifications were performed with the nt (nucleotides) and EST (expressed sequence tags) databases (Benson et al. 2013). Non-Hexapoda or unidentified scaffolds with better hits in these databases were reassigned. Mitochondrial scaffolds were identified against a reference database (ftp://ftp.ncbi.nlm.nih.gov/genomes/MITOCHONDRIA/; 4 September 2014). A final category, "potential symbiont," included scaffolds with similarities to genomes of bacterial genera with known symbiotic interactions (Werren et al. 2008; Duron and Hurst 2013). Scaffolds with best hit to any species of the bacterial genera *Wolbachia, Rickettsia, Spiroplasma, Arsenophonus, Cardinium, Hamiltonella, Blattabacterium, Midichloria*, or *Rickettsiella* belonged to this category. For all categories, potential coding regions were identified by BLASTx to NCBI's RefseqP protein database (amino acid sequence identity >40%).

Hexapoda Repeats Identifications

Hexapod retroelements, satellites and DNA transposons were identified via a RepeatMasker analysis using the Hexapoda repeat definitions of RepBase (Jurka et al. 2005) and DFam (Wheeler et al. 2013) (April 22, 2013 updates). Identifications were complemented by a comparative approach aligning all Hexapoda scaffolds to the *Tc* and *Dp* genomes. Scaffolds with BLAST hits matching ≤30 regions and >30 regions of the *Tc* or *Dp* reference genomes were designated as "low-copy number" (*LCN*) and "*HCN*" repeats, respectively (dc-megablast, E-value threshold 1e-9, alignment >150 bp). *hcn* repeats were also annotated as "similar to a transcript" when they matched a NCBI EST database transcript.

rRNA and Histone Gene Identifications

Histone-related scaffolds were extracted via protein BLAST (BLASTp) to *Drosophila melanogaster* histones and open reading frames (ORFs) greater than 250 bp in the six frames of each extracted scaffold were aligned to the NCBI Conserved Domain Database (CDD, v3.11) via CD-search (Marchler-Bauer et al. 2011). All ORFs matching a histone domain signature at E < 1e-9 were then annotated with the corresponding histone gene annotation. rRNA-containing scaffolds were identified with discontiguous-megablast alignments to the content of the SILVA database (Quast et al. 2013; E > 1e-9, alignment length >150 bp and sequence identity >85%). The database contained both curated large subunit (LSU) and small subunit (SSU) rRNA genes, representing about 3,000,000 bacteria, 150,000 archaea, and 250,000 eukaryote sequences.

MGS Characterization against Full Reference Genomes

The proportion of a reference genome contained in a given metagenomic scaffold library was assessed by the global metagenomic coverage (GMC), which is defined as the proportion of nucleotides in a reference genome that is matched by nucleotides in the scaffolds of a given library. GMC values

are denoted GMC_{ref}^{pool} ("GMC of the reference genome *ref* by the library *pool*"). The metric is formalized as follows. In a reference genome G of size S_G, a base *b* is marked as "covered" by a homologous scaffold base ($b_{covered}$) when it is located within the bases of a BLAST high-scoring segment pair (HSP) at the selected threshold. Then, the library GMC of the genome G by the library *lib* (GMC_G^{lib}) is defined by:

$$GMC_G^{lib}(\%) = \frac{\sum_{i=1}^{S} b_{covered}}{SG} \times 100$$

To measure the proportion of scaffolds with sequence similarities to these genomes for calculating the GMC, we recorded all regions in a pool library for which BLASTn produced HSPs with the reference genome under sequence identity >70%, alignment length >150 bp, and E < 1e-9. The scaffolds found to produce BLASTn hits of the same chromosomal region under the criteria described above were grouped into "clusters" of putatively homologous sequences. A cluster is defined by three parameters: the first and last coordinates of the reference genome region that is homologous to at least one library scaffold, and the set of library scaffolds that are homologous to this same region. Two clusters were differentiated when at least ten reference genome bases separated the last coordinate of cluster *n* and the first coordinate of cluster *n* + 1. The scaffold content of each cluster was aligned with Mafft 6.864 b, using the L-iNSI method and default parameters (Katoh et al. 2005). Alignment graphics were generated with UGENE (Okonechnikov et al. 2012).

Results

Number of Scaffolds and Similarity of Libraries

The Truseq libraries (*Weevil, Canopy_Long, Canopy_Short*) produced 17.3–23.9 M reads pairs and the Nextera library (*Canopy_Next*) produced 7.3 M reads. Following trimming, 30% of reads were discarded in the three *Canopy* libraries and 5% in the *Weevil* library (table 1). Assembly of the four Illumina libraries each produced between 20,000 and nearly 100,000 contigs and numbers were only slightly lower for (noncontiguous) scaffolds (table 1). Using the same DNA pool, both TruSeq libraries resulted in more than twice the number of reads as the Nextera library, and *Canopy_Long* assembled almost twice as many contigs and scaffolds as *Canopy_Short* and over three times as many as *Canopy_Next*. The *Weevil* pool produced the largest number of scaffolds despite containing the second lowest number of reads, whereby long insert size and greater homogeneity of read numbers from equimolar DNA samples may have aided the assembly. We determined intersections of library contents with pairwise alignments of the scaffolds (fig. 2*A*). The scaffolds of the three *Canopy* libraries were aligned with a stringent threshold of sequence identity >90%, E < 1e-18, alignment length >250 bp. In total, 19,297 scaffolds were shared by at least two *Canopy* libraries, and the tripartite intersection showed a core of 6,940 scaffolds (11–35% of the libraries) that was consistently recovered despite the low-coverage sequencing (fig. 2*A*, left). We performed a similar pairwise alignment between the *Weevil* library and the scaffold collection of all *Canopy* libraries (*Canopy_merged*), with a slightly lower threshold (sequence identity >80%, E < 1e-18, alignment length >250 bp) to recover potential homologs among different species (fig. 2*A*, right). A total of 5,174 scaffolds were shared by both samples (5.8% of *Weevil* scaffolds; 4.7% of *Canopy* scaffolds), showing that thousands of similar scaffolds can also be recovered between pools of different species composition.

Phyletic Composition

The scaffolds were assigned to four categories based on their top-hits in sequence databases (see Materials and Methods, and supplementary file S2, Supplementary Material online): Hexapoda, non-Hexapoda, unidentified and potential symbionts. Between 15% (*Weevil*) and 23% (*Canopy_Next*) of scaffolds showed similarity to a known Hexapoda sequence (fig. 2*B*). Despite variation in sequencing procedures, the libraries exhibited fairly similar identification profiles and mainly differed in the absolute number of scaffolds in each category. Two-thirds of Hexapoda identifications resulted from sequence homology with complete genomes and one-third was based on shorter nucleotide fragments and transcriptome data form the NCBI nt and EST databases (supplementary file S2, Supplementary Material online). The intersection of the Hexapoda scaffolds among the three *Canopy* libraries (fig. 2*C*, left) showed that a core of 11–36% of scaffolds was found consistently, and the proportion was similar to the intersection of all scaffolds (not limited to Hexapoda sequences) (fig. 2*A*, left). However, for the Hexapoda scaffolds the proportion shared was much greater for the *Weevil* (17%) than for the *Canopy* libraries (7%; fig. 2*C*, right), whereas this asymmetry was much weaker when comparing all scaffolds (fig. 2*A*, right). The asymmetry may be expected, because the *Weevil* sample (only Curculionoidea) fully overlaps at the clade level with the *Canopy* pool, of which about one-third of species were Curculionoidea, but not vice versa.

Non-Hexapoda Scaffolds

The portion of scaffolds identified as non-Hexapoda comprised between 0.6% (*Canopy_Long*) and 8.2% (*Weevil*). In *Canopy*, we identified 348 bacterial scaffolds and 429 eukaryote (non-Hexapoda) scaffolds (supplementary file S3, Supplementary Material online). Plant DNA was represented by dozens of long chloroplast scaffolds (up to 5 kb), of which 21 have a high similarity (up to 97%) to the genus *Theobroma* (cocoa), 8 other scaffolds were highly similar (up to 98%) to the genus *Gossypium* (cotton) and dozens of additional fragments showed a lower similarity to other known chloroplasts.

Table 1
Sample Composition, Sequencing Procedures, and Assembly Results for the Four Libraries Used in This Study

Sample	Content			Sequencing							Assembly	
	Specimens	Morphospecies	Superfamilies	Library Identifier	Library Preparation Kit	Mean Insert Size (bp)	Sequencing Platform	Read Length (pb)	Read Pairs	Read Pair after Trimming	Contigs	Scaffolds
Weevil beetles	173	173	1	*Weevil*	Illumina TruSeq	790 200	MiSeq	250	18,341,901	17,389,929	95,233	90,392
Canopy beetles	480	212	14	*Canopy_Long*	Illumina TruSeq	850 215	MiSeq	250	33,796,432	23,922,520	68,420	65,412
				Canopy_Short	Illumina TruSeq	480 120	MiSeq	250	33,992,316	21,402,938	35,758	33,054
				Canopy_Next	Illumina Nextera	650 325	MiSeq	250	15,426,678	7,292,986	20,876	20,121

A 7 kb scaffold had high homology to an unannotated genomic region of nematodes. Fungal identifications were rare but particularly interesting in the context of insect biology. For instance, we uncovered a 1 kb fungal scaffold 99.5% similar to *Glarea lozoyensis*, a fungal species linked to insect pathogenicity (Chen et al. 2013) and a 2 kb scaffold homologous to *Phaeosphaeria nodorum* (91% identity), another insect pathogen previously studied for its antibiotic synthesis pathway (Amnuaykanjanasin et al. 2009; supplementary file S3, Supplementary Material online). Bacterial scaffolds included potential symbiotic species in all libraries (fig. 3). In *Canopy_Long*, 108 scaffolds were homologous to the genera *Wolbachia* (42%), *Rickettsia* (50%), *Spiroplasma*, *Hamiltonella*, and *Blattabacterium* (3%). The *Weevil* pool, while comprising a similar proportion of Hexapoda scaffolds, contained more non-Hexapoda and potential symbionts scaffolds (fig. 3). In the latter, 89% were homologous to the genus *Wolbachia* (1,722 scaffolds). We noticed a tendency for lower representation of taxa with fewer available bacterial reference genomes or nucleotide sequences, thus sequences from these genera may be present in the samples but cannot be detected with the current database (fig. 3).

Repeats and Hexapoda Scaffolds

In addition to mitochondrial genomes, nuclear repeat regions were expected to dominate in MGS (fig. 1, scenarios *x* and *y*). First, using RepeatMasker (see Materials and Methods, and supplementary file S4, Supplementary Material online), low complexity repeats (microsatellites, SSRs) constituted between 2.49% (*Canopy_Next*) and 6.08% (*Canopy_Long*) of all scaffolds (fig. 4*A*) and their number was correlated with the number of sequenced reads. The combined fraction of retroelements, DNA transposons, and small RNAs in *Weevil*, *Canopy_Long*, *Canopy_Short*, and *Canopy_Next* constituted 18.6%, 15.7%, 16.1%, and 21.8% of all scaffolds, respectively (fig. 4*A*). Despite its smaller size, *Canopy_Next* showed high proportions of repeats in all non-SSRs categories.

When aligned to the *Tc* and *Dp* reference genomes (see Materials and Methods) between 22.6% (*Canopy_merged*) and 32.3% (*Weevil*) of the Hexapoda scaffolds matched one or more regions of the two reference genomes (table 2). In addition, both *Canopy* and *Weevil* contained numerous scaffolds related to protein sequences, as identified by their similarity to NCBI's RefseqP (protein sequences; table 2). The uncovered genes (5.5% and 9.7% of total Hexapoda scaffolds based on amino acid sequence identity >40%) were mostly lacking functional annotations, and the annotated fraction was mainly related to universal components of metabolism, development, and basic signaling pathways (supplementary file S5, Supplementary Material online). Even so, we uncovered some insect-specific functions such as *Canopy_Long* scaffolds aligned to an exon of the chemosensory protein 1 gene of *Tc*, a gene known to play an essential role in insect communication and development (Forêt et al. 2007). A surprisingly high proportion of scaffolds showed homology to EST databases; 21.3% in *Canopy* and 29.3% in *Weevil* (table 2). A noticeable proportion of the latter exhibited high similarity (>90%) to the *Dp* transcriptome sequences generated by Chan et al. (2012) (data not shown).

Many Hexapoda scaffolds had multiple mapping sides, which identified them as dispersed repeat regions. An example of Hexapoda repeat is illustrated in figure 4*B*. This conserved 200 bp motif showed high similarity to a single scaffold from the *Canopy_Long* library and was found in 461 *Dp* genome regions and in 9 different mRNAs in the *Dp* transcriptome. A core fragment of this repeat was similar (85% identity) to an mRNA of *Hypothenemus hamperi*, a member of the same weevil subfamily (Scolytinae). Despite the similarity to an mRNA, many indels among copies suggest that this *hcn* repeat is not a coding region.

Highly Sampled Gene Families

Gene families recovered from the metagenomes correspond to multicopy and highly conserved elements (fig. 1, scenario *D*) and represent some of the *Icn* repeats mapped on the reference genomes (fig. 5*C*, red bars). Tandemly repeated rRNA genes were present in hundreds of scaffolds, with twice as many matches to the LSU than to the 2-fold shorter SSU gene (fig. 5). In all libraries, more than 80% of rRNA-containing scaffolds were homologous to Hexapoda, but other phyla were represented also (fig. 5). The *Weevil* and

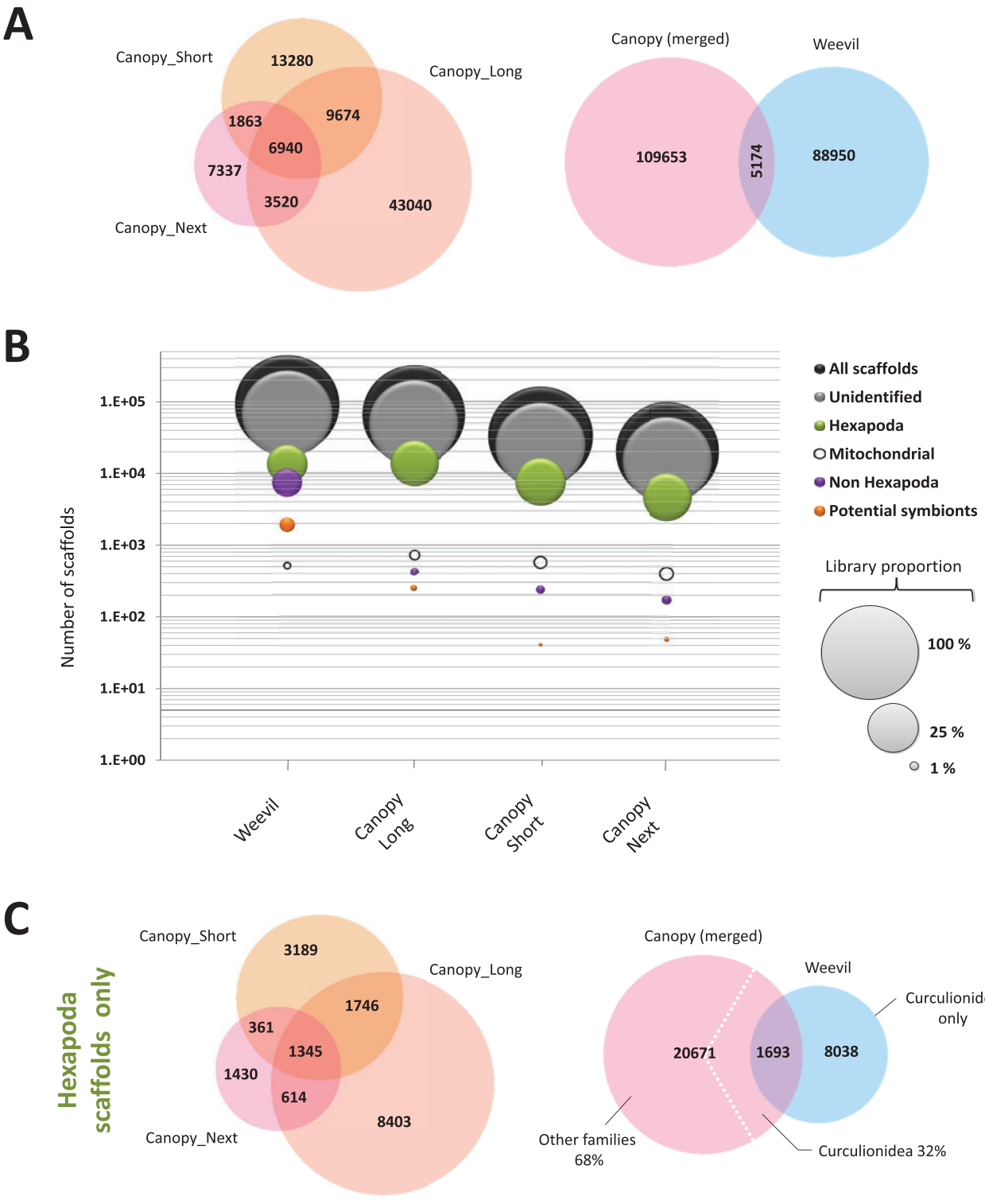

Fig. 2.—Phylogenetic content of the metagenomes. (*A*) Circle intersections of scaffold overlap among the different libraries. (*B*) Classification of scaffolds based on best hits in genome databases. See text for details of the five categories. The *y* axis represents the absolute number of scaffolds. Circle size represents category proportion relative to the total number of scaffolds in a library. (*C*) Circle intersections of scaffold overlap among the different libraries for scaffolds assigned to Hexapoda.

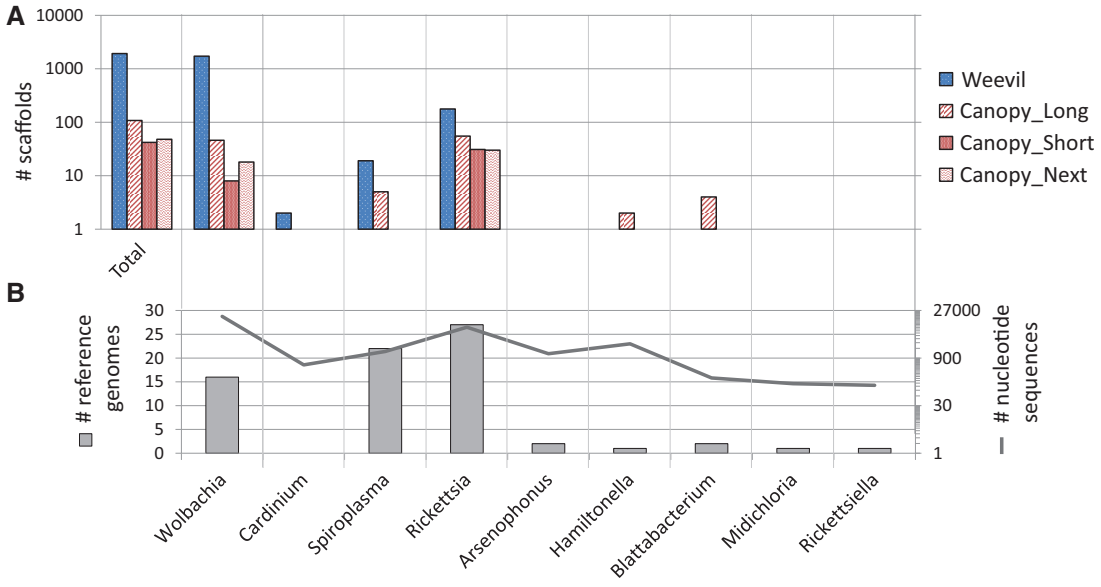

Fig. 3.—Bacterial symbiont profile for the *Canopy* and *Weevil* samples. (*A*) Number of bacterial scaffolds (number of scaffolds) classified by their homology to different bacterial symbiont genera (given on *x* axis) for each of the four libraries (*Weevil*: blue bars; *Canopy* libraries: red bars). (*B*) Number of complete genomes (number of reference genomes) and nucleotides entries (number of nucleotide sequences) currently available in NCBI Chromosome ("other_genomic") and nulceotide ("nt") databases, respectively, for each symbiont genus (*x* axis).

Canopy_Long libraries contained 2.3 and 0.95 rRNA scaffolds per morphospecies on average.

The histone family represented the most frequently sampled protein-coding scaffolds, including 100 scaffolds of greater than 95% amino acid level similarity to *Drosophila* histones in the *Weevil* library. Figure 6 depicts 7 of the 18 *Weevil* scaffolds that had the same *Drosophila* histone best-hit in RefseqP. The scaffolds range from 2 to 7 kb in length and contain between 2 and 5 ORFs confirmed by the presence of histone domain signatures. The observed histone quintet (composed of four core histones H2A, H2B, H3, H4, and the H1 linker histone prevalent in insects) had a structure similar to that described in the boll weevil *Anthonomus grandis* (Roehrdanz et al. 2010). The seven scaffolds were used to generate an alignment centered on the first 100 bp of the H2B ORFs. The alignment consensus profile was typical for an exonic region, with the third codon base being less conserved between scaffolds, while there was additional intrascaffold variation on a small number of polymorphic sites (fig. 6, bottom). The fifth scaffold differed from the others by a duplication of the H2A gene and a 3 bp insertion in H2B.

Characterizing Metagenomes against Complete Reference Genomes

Mapping against the *Tc* and *Dp* reference genomes established the degree of repetitiveness and a given scaffold. *hcn* scaffolds (>30 mapping sites in a reference genome) represented between 3.9% (against *Tc*) and 6.5% (against *Dp*) of the *Canopy* scaffolds and 2.0% (against *Tc*) and 11.0% (against *Dp*) of the *Weevil* scaffolds. There were many more *lcn* scaffolds (≤30 mapping sites) and the number of hits was greater against *Dp* than *Tc* (table 2). An interesting finding was that in *Weevil* the ratio of *lcn* to *hcn* scaffolds was 2.2 × on the *Dp* genome but it was increased to 5.5 × for the heterologous *Tc* genome, that is, there was a relative loss of *hcn* over the *lcn* scaffolds for the distantly related reference genome, suggesting that the *hcn* repeats are more taxon specific. In accordance with this interpretation, the pattern was not observed in *Canopy*, where *lcn* scaffolds dominated only by 1.5- to 1.7-fold over *hcn* scaffolds, presumably because of the greater taxonomic diversity of this metagenome. The general taxon specificity of scaffolds was confirmed by the fact that less than 0.8% of all scaffolds were homologous to both the *Tc* and *Dp* genomes.

The effect of different phyletic composition of the *Canopy* and *Weevil* metagenomes was further tested with genome coverage metrics against the two available Coleoptera genomes (see Materials and Methods). Figure 7 details the covered bases ($b_{covered}$) and the GMC values (coverage normalized for genome size, "+" symbol). Chromosome reconstruction was available only for *Tc*. Up to 5% of the *Tc* genome was covered by the *hcn* scaffolds of the *Canopy (merged)* metagenome (fig. 7*A*), but when excluding *hcn* scaffolds, the $GMC_{Tc}^{Canopy\_merged}$ of various chromosomes was always less than 0.5%. This effect was particularly striking in chromosomes 3, 6, and 10, whereby the latter two are small chromosomes highly enriched in *hcn* sequences conserved between the *Canopy* metagenome and *Tc*.

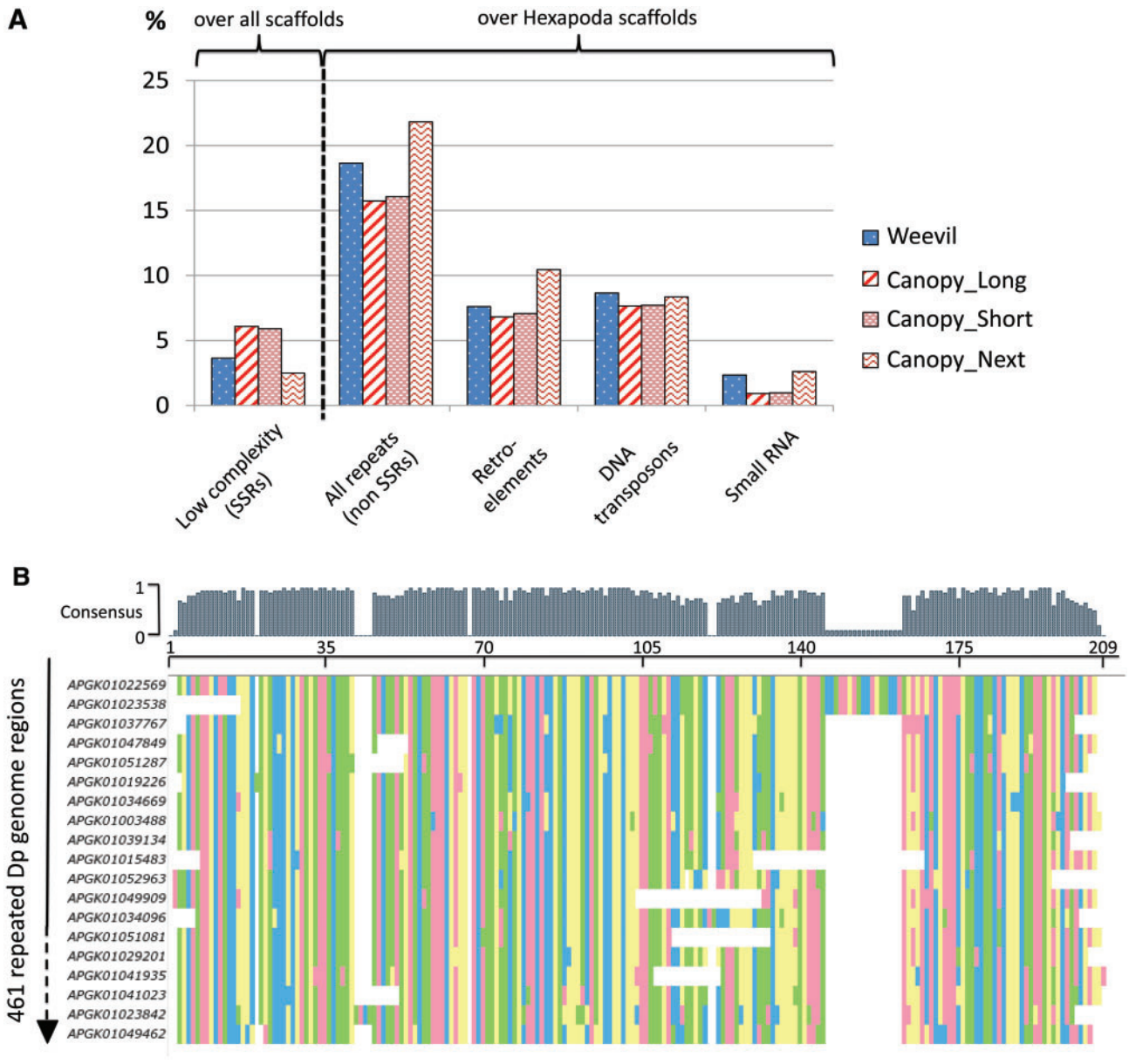

Fig. 4.—Genomic repeats inventory. (*A*) Proportion of base pairs (%) identified as genomic repeats by RepeatMasker, plotted separately for short simple repeats (SSR) all non-SSR repeats which include retroelements, DNA retrotransposons and small RNAs. Low complexity repeats are presented for all scaffolds. Other repeats are given only for the Hexapoda scaffolds, as the proportions represent only previously known repeats having a definition in RepBase. (*B*) Part of an aligned matrix covering 461 copies of the *Dp* reference genome identified by their similarity to a paralogous scaffold (scf7180004984182; *Canopy*_Long). The corresponding regions are aligned, showing a similar 200 bp motif (A: yellow; T: red; C: green; G: blue; gap: white).

Comparison of the GMCs produced by the *lcn* component of each library on the two reference genomes (fig. 7*B*) showed that: 1) genome coverage was always higher for the *Dp* than *Tc* genome (fig. 7*B*, panel 2 compared with panel 1); 2) GMCs were greatly increased when merging the three *Canopy* libraries; 3) as expected from its taxonomic composition, the *Weevil* metagenome showed higher coverage of the *Dp* than *Tc* genome, with a GMC_{Dp}^{Weevil} (panel 2) five times higher than GMC_{Tc}^{Weevil} (panel 1); 4) in contrast, *Canopy_Long* showed greater coverage of the *Tc* genome than *Weevil* (panel 1), as expected from the presence of Tenebrionoidea in the former library; and 5) however, both the *Weevil* and *Canopy*_Long libraries had a fairly similar coverage of the *Dp* genome (panel 2, $GMC_{Dp}^{Weevil} \approx GMC_{Dp}^{Canopy\_Long}$), as both metagenomes contained members of Curculionoidea.

Finally, we investigated the *lcn* scaffolds that mapped to the same reference genome site forming clusters of scaffolds

Table 2

Number of Identified Hexapoda Scaffolds and Their Characterization by Mapping to *Tc* and *Dp* Reference Genomes and Other Databases

Number of Hexapoda Scaffold in Library	*Canopy (merged)*		*Weevil*	
	26,002 (100%)		13,482 (100%)	
Mapped to…				
Tc or *Dp*	5,876 (22.6)		4,355 (32.3)	
A protein (refseqP database)	1,433 (5.5)		1,308 (9.7)	
A transcript (EST database)	5,538 (21.3)		3,950 (29.3)	
Identified as repeats (per reference)	*Tc*	*Dp*	*Tc*	*Dp*
Low-copy number (≤30)	2,494 (9.6)	3,835 (14.7)	1,514 (11.2)	3,379 (25.1)
└ In *Tc* and *Dp*	758 (2.9)		690 (5.1)	
High-copy number (>30)	527 (3.9)	1,682 (6.5)	273 (2.0)	1,477 (11.0)
└ Similar to a transcript	77 (0.3)	405 (1.6)	111 (0.8)	773 (5.7)
└ In *Tc* and *Dp*	220 (0.8)		97 (0.7)	
└ In *Tc* and *Dp*, similar to a transcript	24 (<0.1)		36 (0.2)	

Note.—Low-copy number and high-copy number scaffolds are those with hits to fewer and more than 30 reference genome regions respectively.

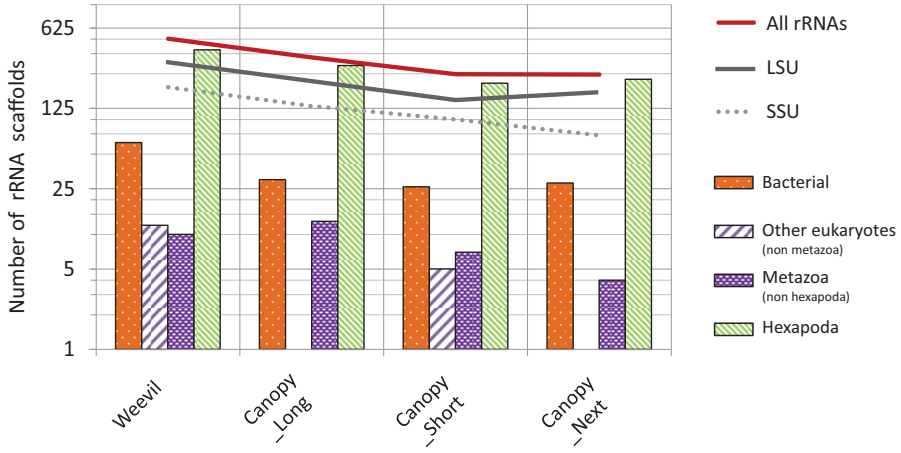

Fig. 5.—Inventory of rRNA gene scaffolds. Scaffolds identified to contain rRNA genes are classified into different life domains by using their closest rRNA homolog. The number of scaffolds holding a rRNA gene is reported for each library and each life domain (orange dots: bacterial rRNAs; green bars: Hexapoda rRNAs; purple dots: Metazoa; purple bars: other eukaryotes). For each library, the global number of rRNA scaffolds is also reported (red line) as well as the number of SSU and LSU rRNAs.

with highest similarity to the same reference genome region (but sufficiently divergent not to be integrated into a single assembled scaffold). Clusters typically contained less than 5 scaffolds but others were more numerous and in some cases accumulated high sequence variation. Clusters mapped primarily to intergenic regions but not exclusively so (fig. 7*C*). The detailed results are reported in supplementary table S6, Supplementary Material online. There were generally more clusters composed of a given number of scaffolds (e.g., clusters of >5, >10 or >20 copies) against the *Dp* compared with the *Tc* reference genome. Cluster sizes were also larger for the *Canopy* compared with *Weevil* metagenomes, presumably as higher taxonomic diversity contributes to a greater diversity of scaffolds corresponding to a given gene region. Finally, the cluster sizes increase greatly in the *Canopy_Merged* over the single *Canopy* libraries, and this gain was higher than expected from simply resampling the *Canopy* DNA pools three times (supplementary fig. S6, Supplementary Material online).

Discussion

MGS of Pools

This study takes the idea of genome skimming (Straub et al. 2012) further to examine the complexity of a low-coverage metagenome based on hundreds of metazoan specimens

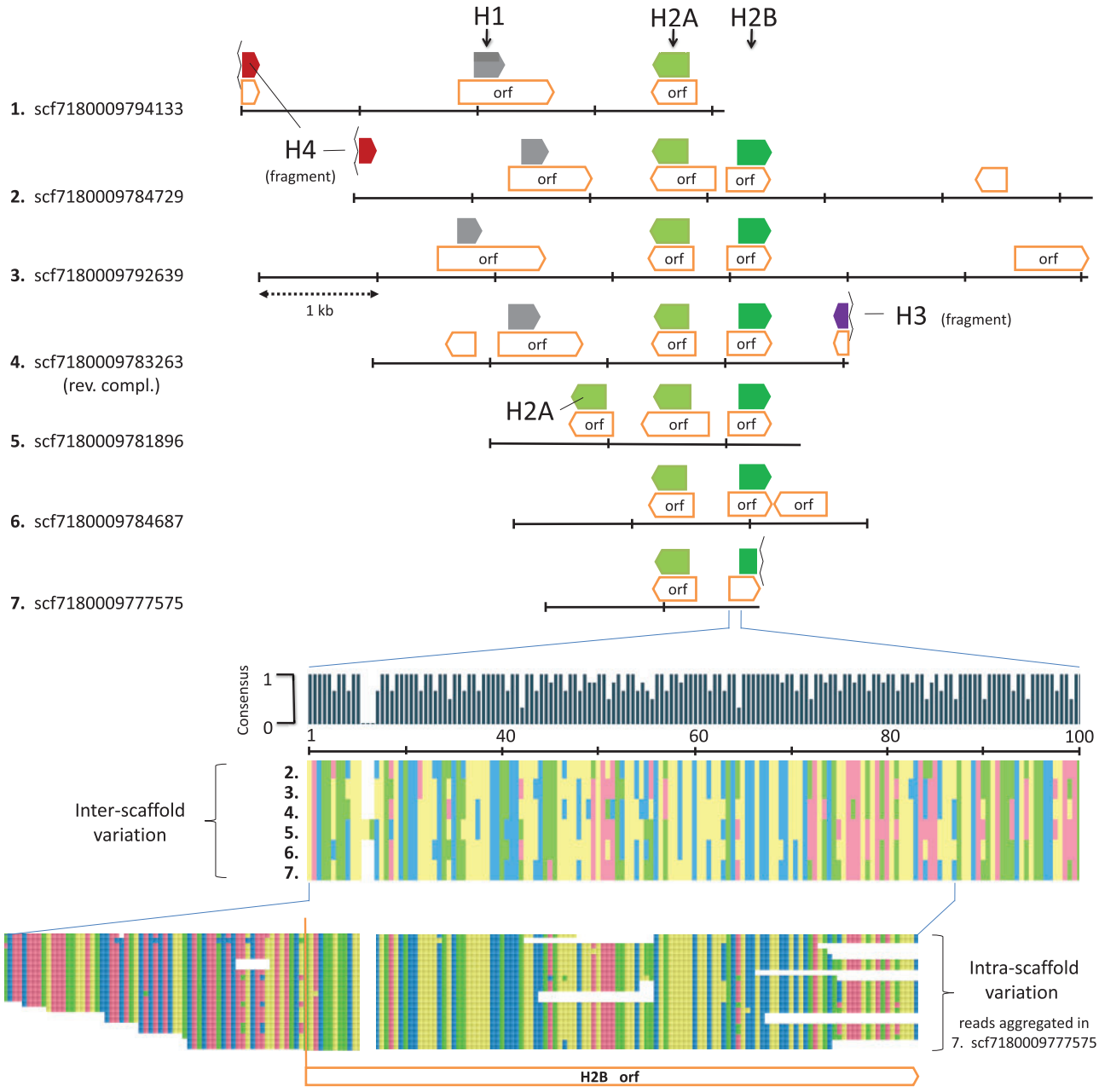

Fig. 6.—Variation in *Weevil* histone clusters. The top of the figure gives seven scaffolds homologous to *D. melanogaster* GenBank entry NP_001027366.1 containing the H2A gene. Histone domain signatures and their orientation are represented with colored boxes (H1: gray; H2A: light green; H2B: dark green; H3: purple; H4: red). The inter- and intrascaffold base variation are detailed at the bottom showing the alignment of the H2B ORFs and the alignment of the reads aggregated to generate the seventh scaffold (scf7180009777575, *Weevil*) (A: yellow; T: red; C: green; G: blue; gap: white).

without multiplexing. We confirm that this approach generates many thousands of scaffolds assembled from the *hcn* regions in the metagenomes and that these scaffolds were detected consistently in repeated sequencing of a given pool, while a large set is also detected by sequencing different, but phylogenetically overlapping pools (fig. 2).

The main advantage of pooled metagenomic sequencing is the large number of species that can be assayed for their multicopy elements in a single sequencing run, which would have been prohibitive for each individual species with current sequencing technology. However, the assembly process may lose information about taxonomic origin, genome localization, and copy number during scaffold aggregation, and slight sequence variation may not be preserved when integrated into a single scaffold. Equally, while higher read numbers increase the probability of scaffold formation, the "real" source of

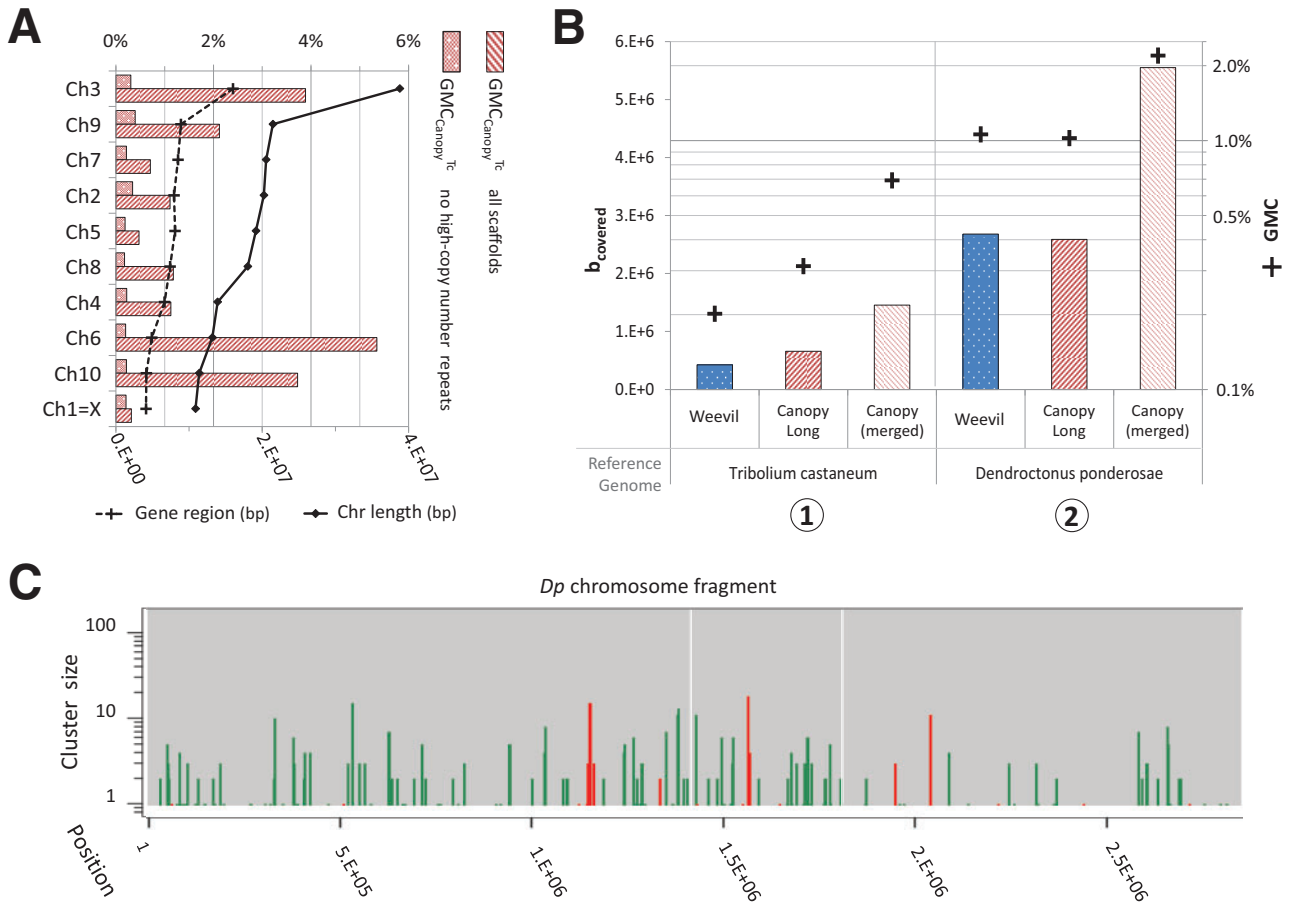

Fig. 7.—Genome coverage estimations. GMC values are reported for the *Weevil* and *Canopy* metagenomes mapped on *Tc* and *Dp*. (*A*) GMC_{Canopy}^{Tc} is detailed for the ten *Tc* chromosomes. Bars with parallel lines: GMC value based on both high-copy number and low-copy number repeats (top *x* axis). Bars with dots: GMC value based on low-copy number repeats only. Also given is the size for each chromosome (bottom *x* axis). (*B*) GMCs calculated for the different *Dp*/*Tc* and *Canopy*/*Weevil* combinations (*x* axis). Right *y* axis: GMC normalizations. Left *y* axis: Genome base covered by low-copy number scaffolds ($b_{covered}$, vertical bars). (*C*) A typical cluster profile is illustrated for the chromosome scaffold KB741028.1 (*Dp* genome). The *y* axis presents the cluster size, as the number of metagenomic scaffolds similar to the same reference genome region, along the linear genome assembly. Green bars correspond to intergenic regions, and red bars represent regions annotated as genes. The gray background indicates a region for which the sequence is known (bases ATCG in the genome assembly), and white background represents unknown bases (N in assembly).

reads is lost, for example, the assembler may collapse multi-copy loci within a genome or sequences that are conserved between species into a single scaffold. These issues may appear to undermine the MGS approach, but we obtain valuable information on the genomic and evolutionary diversity in Coleoptera through a comparative approach that uses two phylogenetically nested metagenomes and their alignment to two fully sequenced reference genomes representing different clades. An example of the products of this assembly process are the histone gene scaffolds, which clearly show variation in primary sequence and gene organization and which probably are a good reflection of the diversity in this gene family across lineages of Coleoptera. Individual scaffolds may also be composed of multiple, slightly different primary sequences, as is evident from variation in individual reads mapped back on the contigs (fig. 6, bottom), which again reflects genomic variation at a lower taxonomic level. Overall, whereas the taxonomic resolution is limited, the power of the analysis comes from the possibility to compare the diversity of scaffolds across metagenomes and in relation to existing reference genomes.

The high proportion of unidentified scaffolds suggests that most of the genomic diversity in Coleoptera, including the conserved and *hcn* components, remains unknown. The high lineage specificity of scaffolds captured here suggests that this large unidentified portion is derived from lineages that are not represented by available reference sequences. Lineage specificity applies equally to those identified as Hexapoda and to all others, as the proportions of scaffolds shared among libraries were similar for the Hexapoda and

unidentified scaffolds (fig. 2), that is, the unidentified fraction was not generally more clade specific, but probably was more distant from an available reference genome, which currently include a limited portion of the Coleoptera tree in the infra-order Cucujiformia within one of the four recognized suborders.

Approximately 20% of all scaffolds matched known repeat sequences deposited in the RepBase database that includes sequences derived from the *Tc* genome (fig. 4*A*). Genomic repeats are copiously sampled during GS (Straub et al. 2012) and scaffolds captured here through MGS, but those remaining unidentified are probably repeats specific to particular lineages of Coleoptera without available reference sequences. In support of this interpretation, many more repeats are identified in *Weevil* when mapped to the closely related *Dp* genome compared with *Tc* (table 2), and while around 20% of the identified Hexapoda DNA contained known repeat sequences in the database (fig. 4*A*), this is a low estimate as only 93 coleopteran repeat definitions were available in the repeat databases interrogated by RepeatMasker (compared with >5,000 for Diptera). Interestingly, the scaffolds built here appear to maintain some correlation between sequencing depth and low complexity repeats (SSRs) sampling, whereas other categories of identifiable repeats are obtained already at lower sequencing depth (see *Canopy_Next* in fig. 4*A*).

Potential for Diversity Estimates and Phylogenomics

Previous GS studies demonstrated that, in accordance with their abundance in genomes, cpDNA, mtDNA, and nuclear rRNA cistrons are sufficiently sampled at low sequencing depth to be de novo assembled and used for phylogenomics (Straub et al. 2012; Bock et al. 2014; Malé et al. 2014). The MGS approach with mixed-species samples reveals similar prevalence of these loci. We do not discuss here the great potential of mtDNA scaffolds as they are described in dedicated studies (Gillett et al. 2014, Crampton-Platt A, et al., under review). rRNA cistrons and histone clusters are additional multicopy nuclear regions with high chance of being sampled. LSUs scaffolds were more numerous than SSUs either due to their longer gene sequence that increases the chance of detection, or due to the greater variability that creates more separate scaffolds rather than being subsumed in a single scaffold. In all metagenomes, the sampling of hundreds of rRNA scaffolds (between 0.95 and 2.3 per morphospecies, fig. 5) suggests that the taxonomic diversity of a sample can be estimated from these genes. For some samples their exploitation may be tedious as the assemblies from mixed reads of low divergence may produce chimeric scaffolds, in particular for conserved, short contigs (Mavromatis et al. 2007). Therefore, the long histone gene cluster is an interesting alternative. Histone clusters are present in hundreds of copies per genome and appear to be sampled in sufficient depth by MGS to generate numerous scaffolds giving new insights into its gene cluster diversity. The few annotated histone clusters (fig. 6) already highlight gene duplications and varying intergenic distances that may contribute to phylogenetic resolution. Here, these features cannot be linked to particular species or clades if a variant is shared among genomes in the pool but the data may be useful for the analysis of duplicated and repetitive regions that are notoriously difficult to assemble from standard genome sequencing. For example, neither the published *Tc* nor *Dp* genome sequences provide a scaffold holding the complete histone gene region. Where needed, library multiplexing could also answer the problem of read (and scaffold) association with the pooled specimens and testing for the phyletic extent of particular features.

Insights into Genome Evolution

The MGS approach provides new insights into genome evolution, by characterizing thousands of scaffolds with regard to their phyletic distribution and levels of variation. We found that the number of newly predicted repeats depends both on the species pool and the sequencing depth. We also found great differences between the reference genomes when they were used to diagnose the composition of the metagenomes. The *Tc* genome produced many fewer matching scaffolds than *Dp*, even after taking into account differences in the size of the genomes that are the template for mapping of the metagenomes. Although the genome assembly of *Tc* is smaller than *Dp,* a higher proportion of *Dp* is covered after GMC normalization (fig. 7*C*). The differences in genome coverage are moderated by the taxonomic composition of the species pools from which metagenomes were obtained. As expected, the *Weevil* library containing exclusively specimens of Curculionoidea (weevils) has fewer scaffolds matching the *Tc* (Tenebrionoidea) than in the *Dp* (Curculionoidea) reference genome (table 2), resulting in a larger coverage for the latter (fig. 7*B*). For the *Canopy* library, scaffolds matching to these genomes are compatible with the pool composition (23% of Cuculionoidea and 12% of Tenebrionoidea, supplementary file S1, Supplementary Material online) but the *Tc* still produced fewer matches (table 2) and a similarly high coverage of Dp (fig. 7*B*), despite containing many species of Tenebrionoidea. Although the stochastic nature of the assembly also has an effect in the sampling, these results suggest differences between both reference genomes. Repeat elements are highly variable among metazoan genomes, for example, the proportions of transposable elements ranges from 3% to 45% (Wicker et al. 2007), and possibly a greater prevalence of repetitive elements in larger genomes could account for the higher base coverage in *Dp*.

The results confirmed the taxonomically restricted distribution of repeat elements, as the fraction common to both coleopteran superfamilies is small (table 2). This is in agreement

with the conclusions from the *Dp* genome study that find only 0.15% of its approximately 3,000 novel repeats to be homologous between the two coleopteran genomes (Keeling et al. 2013). Nevertheless, *lcn* scaffold aggregation appears to be due to broader phylogenetic distributions across multiple loci (fig. 1, MGS scenarios *A*, *B*, *D*), as evident from a genome coverage moderated by the taxonomic composition of the metagenome pools (fig. 7*B*). This is in contrast to the higher taxonomic specificity evident for the *hcn* scaffolds (>30 mapping sites) (table 2). Many *hcn* scaffolds probably were not captured by the mapping exercise to the two reference genomes but the recovery of approximately 1,500 *hcn* repeats in both libraries indicates a significant sampling of the repeat repertory. Elements present in *hcn* may have different evolutionary dynamics that promote divergence. They may have effects on the entire genomes, for example, exerting isolation mechanisms or exerting functional constraints that restrict recombination and promote genome-wide rapid divergence. Some high-copy repeats are derived from fast-evolving elements, such as retroviruses, that integrate into genomes. Interestingly, clade-specificity is even stronger when only *hcn* repeats associated to transcribed regions are considered, suggesting that repetitive regions constitute related regulatory sequences present in multiple, possibly coregulated, genes. Considering the importance of higher eukaryotes RNA cis (promoters, enhancers, riboswitches, etc.) or trans (miRNA) regulatory regions in higher eukaryotes, in particular their role in phenotypic evolution (Wittkopp and Kalay 2012), MGS may be an interesting alternative to extract the variation associated to regulatory elements of specific lineages.

The mapped clusters also revealed differences in cluster size (scaffolds matched to a reference genome site) for both metagenomes, which was consistently greater when mapped against *Dp* than *Tc* (supplementary file S6, Supplementary Material online). Hence, not only is the number of sites mapped greater in *Dp* but also there are more scaffolds mapping to any one of these sites (on average). Moreover, resampling the DNA pools by combining the three *Canopy* libraries (*Canopy_Merged*) adds more scaffolds (on average) that can be mapped to each site (supplementary file S6, Supplementary Material online). The nature of these additional variants was not analyzed further, but considering the hypothetical scaffold aggregation of *lcn* elements (fig. 1), deeper sequencing presumably results in higher degree of completion of homologous elements for more taxa in the pool. Here, this appears true for multilocus genes families, as shown by the extensive rRNA and histone scaffolds recovery, two tangible *lcn* repeat examples (figs. 5 and 6). The reason for identifying a greater scaffold diversity (larger clusters) with *Dp* compared with *Tc* remains unclear without more detailed study of individual clusters but it appears that in the species represented by our metagenomes, the *Dp* genome is a better reflection of the wider pool (pan-genome) of Coleoptera. Some of these conserved regions have a wide taxonomic distribution across virtually all species in the pool, such as the histone or rRNA genes, and they are captured with any of the reference genomes. Others are likely more limited to specific clades and their detection depends on the composition of the reference genomes and the (presumably ancestral) presence of similar elements in clades related to these references and the phyletic extent of such homologs. More complete reference genomes are needed to confirm these trends and, in the case of the *Canopy* sample, to disentangle the influence of abundance and biomass of species and clades.

Skimming Environmental Interactions

The detection of non-Hexapod DNA (up to 8% of scaffolds) may reflect the interactions of the sequenced specimens with their environment. Our sequencing libraries were obtained by extracting DNA from isolated specimens. Hence, non-Hexapoda reads mostly constitute either internal fauna, ingested material or genes recently acquired by horizontal gene transfer (increasingly observed in insects; see (Nakabachi 2015). For instance, plant DNA decay is slow enough in Coleoptera guts for PCR amplification of chloroplasts up to 72 h after feeding (Wallinger et al. 2013). Short cpDNA scaffolds are insufficient for unequivocal species-level identification, in particular in areas where reference data are lacking, although the presence of *Theobroma cacao* in *Canopy* is plausible as smallholder plantations are common in the sampling region. Performing a read-level analysis (rather than scaffolds analyzed here) would probably provide more refined clues on diet. Combined with models of food decay, low-coverage sequencing of species pools can be a powerful tool for determining herbivory and predation levels in the ecosystem (Wallinger et al. 2013; Paula et al. 2014). The identification of hundreds of bacterial scaffolds related to symbionts of insects (2% of the total scaffolds) is promising for studies of the associated microbiome. The species composition and genetic diversity of primary symbionts are a major source of variation in ecosystems (Ferrari and Vavre 2011), as they manipulate plant physiology in favor of interactions with insects (Frago et al. 2012), and understanding this tight interaction opens new perspectives, such as nonchemical pest control (McMeniman et al. 2009). Already, the *Weevil* and *Canopy* samples show different symbiont profiles. Considering that our identification is limited to the known symbiont genomes, the systematic search for bacterial symbiont populations should be an essential part of insect MGS analyses.

Toward Integrative Analysis of Specimen Pools

The metagenomes were based on a complex mixture of specimens varying by their genome structure, phyletic composition and, in the case of the *Canopy* sample, by variable biomass and intraspecific variation. These factors determine the

abundance and sampling probability of DNA motifs in the pool and hence the MGS outcome. The chance for assembling a scaffold depends on the copy number of a sequence in the genome and its evolutionary conservation, given the phylogenetic proximity of species in the pool. We are only beginning to disentangle these complex sets of parameters, and more targeted sampling would be needed to confirm these conclusions (the current pools were initially designed for answering different questions). The results demonstrate that MGS of insect specimen pools extracts data relevant for phylogenomics, ecology or simply to explore the black box of arthropod genome evolution through their multicopy and most conserved elements. Future analyses will further dissect intragenomic and intergenomic components of the skimmed metagenome based on additional reference genomes (e.g., i5K Consortium 2013; Misof et al. 2014) that will refine conclusions on the phylogenetic depth of scaffolds and place many unidentified scaffolds into a genomic context. Eventually, the comparative genomic approach on different pools could correlate the sampling of specific DNA motifs to the presence of particular insect clades based on numerous diagnostic repeats, including species or population-specific sequences (Grasela and McIntosh 2003) and satellite DNA of related genera (Bruvo-Madarić et al. 2007). In botany, rRNA genes sampled through GS in particular were used previously to determine levels of intragenus variation (Weitemier et al. 2015) or to identify diagnostic polymorphisms uncovering geographical patterns of hybrid formation (Bock et al. 2014). When carefully designed, MGS has the potential to provide new metagenome-based community traits, such as the ones explored today in the microbial world (Barberán et al. 2012; Zarraonaindia et al. 2013). Finally, MGS shows potential for understanding of complex ecosystems such as insect–symbiont interactions (Duron and Hurst 2013). When combined with environmental variables, data mining approaches (clustering, classification, etc.) could ultimately be used on the large panel of information extracted from different pool communities by MGS, generating new hypotheses that link genomic and ecological perspectives of community composition (Peng et al. 2008; Huttenhower and Hofmann 2010).

Supplementary Material

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

Acknowledgments

The authors thank Carmelo Andujar, Paula Arribas Blázquez, Kirsten Miller, Rosli Kasah, Steven Dodsworth, and Chris Barton for their data contribution and many fruitful discussions. They also thank Peter Foster for help with server computing and computational support. The work was developed within the NHM Biodiversity Initiative. Additional funding was from The Leverhulme Trust (F/969/P to A.P.V.); an NHM/UCL PhD studentship (to A.C.P.); and a NERC Postdoctoral Fellowship (NE/I021578/1 to M.J.T.N.T.).

Literature Cited

Amnuaykanjanasin A, Phonghanpot S, Sengpanich N, Cheevadhanarak S, Tanticharoen M. 2009. Insect-specific polyketide synthases (PKSs), potential PKS-nonribosomal peptide synthetase hybrids, and novel PKS clades in tropical fungi. Appl Environ Microbiol. 75:3721–3732.

Andújar C, et al. 2015. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. Mol Ecol.

Barberán A, Fernández-Guerra A, Bohannan BJM, Casamayor EO. 2012. Exploration of community traits as ecological markers in microbial metagenomes. Mol Ecol. 21:1909–1917.

Basset Y, et al. 2012. Arthropod diversity in a tropical forest. Science 338:1481–1484.

Benson DA, et al. 2013. GenBank. Nucleic Acids Res. 41:36–42.

Bentley DR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Bock DG, Kane NC, Ebert DP, Rieseberg LH. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytol. 201:1021–1030.

Bruvo-Madarić B, Plohl M, Ugarković D. 2007. Wide distribution of related satellite DNA families within the genus *Pimelia* (Tenebrionidae). Genetica 130:35–42.

Chan SK, et al. 2012. Transcriptome and full-length cDNA resources for the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major insect pest of pine forests. Insect Biochem Mol Biol. 42:525–536.

Chen L, et al. 2013. Genomics-driven discovery of the pneumocandin biosynthetic gene cluster in the fungus *Glarea lozoyensis*. BMC Genomics 14:339.

Crampton-Platt A, Timmermans MJTN, Gimmel ML, Narayanan Kutty S, Cockerill TD, Vun Khen C, Vogler AP. 2015. Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. Mol Biol Evol. Advance Access publication May 8, 2015, doi:10.1093/molbev/msv111.

Duron O, Hurst GD. 2013. Arthropods and inherited bacteria: from counting the symbionts to understanding how symbionts count. BMC Biol. 11:45.

Ferrari J, Vavre F. 2011. Bacterial symbionts in insects or the story of communities affecting communities. Philos Trans R Soc Lond B Biol Sci. 366:1389–1400.

Ficetola GF, Miaud C, Pompanon F, Taberlet P. 2008. Species detection using environmental DNA from water samples. Biol Lett. 4:423–425.

Forêt S, Wanner KW, Maleszka R. 2007. Chemosensory proteins in the honey bee: insights from the annotated genome, comparative analyses and expressional profiling. Insect Biochem Mol Biol. 37:19–28.

Frago E, Dicke M, Godfray HCJ. 2012. Insect symbionts as hidden players in insect–plant interactions. Trends Ecol Evol. 27:705–711.

Friedrich M, Muqim N. 2003. Sequence and phylogenetic analysis of the complete mitochondrial genome of the flour beetle *Tribolium castaneum*. Mol Phylogenet Evol. 26:502–512.

Gillett CPDT, et al. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). Mol Biol Evol. 31:2223–2237.

Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJTN, Baselga A, Vogler AP. 2015. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. Methods Ecol Evol.

Grasela JJ, McIntosh AH. 2003. Application of inter-simple sequence repeats to insect cell lines: identification at the clonal and tissue-specific level. In Vitro Cell Dev Biol Anim. 39:353–363.
Huttenhower C, Hofmann O. 2010. A quick guide to large-scale genomic data mining. PLoS Comput Biol. 6:e1000779.
i5K Consortium. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 104:595–600.
Ji Y, et al. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett. 1245–1257.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.
Keeling CI, et al. 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. Genome Biol. 14:R27.
Lee MH, Lee S-W. 2013. Bioprospecting potential of the soil metagenome: novel enzymes and bioactivities. Genomics Inform. 11:114–120.
Lohse M, et al. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 40:W622–W627.
Malé PJG, et al. 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour. 14:966–975.
Marchler-Bauer A, et al. 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. 39:D225–D229.
Mavromatis K, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 4:495–500.
May RM. 2010. Ecology Tropical arthropod species, more or less? Science 329:41–42.
McMeniman CJ, et al. 2009. Stable introduction of a life-shortening Wolbachia infection into the mosquito *Aedes aegypti*. Science 323:141–144.
Metzker ML. 2010. Sequencing technologies—the next generation. Nat Rev Genet. 11:31–46.
Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95:315–327.
Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.
Nakabachi A. 2015. Horizontal gene transfers in insects. Curr Opin Insect Sci. 24–29.
Okonechnikov K, Golosova O, Fursov M. 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166–1167.
Paula DP, et al. 2014. Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. Mol Ecol Resour.
Peng Y, Kou G, Shi Y, Chen Z. 2008. A descriptive framework for the field of data mining and knowledge discovery. Int J Inf Technol Decis Mak. 7:639–682.
Quast C, et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41:D590–D596.
Raven JA. 2012. Algal biogeography: metagenomics shows distribution of a picoplanktonic pelagophyte. Curr Biol. 22:R682–R683.
Roehrdanz R, Heilmann L, Senechal P, Sears S, Evenson P. 2010. Histone and ribosomal RNA repetitive gene clusters of the boll weevil are linked in a tandem array. Insect Mol Biol. 19:463–471.
Straub SCK, et al. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot. 99:349–364.
Sucher NJ, Hennell JR, Carles MC. 2012. DNA fingerprinting, DNA barcoding, and next generation sequencing technology in plants. Methods Mol Biol. 862:13–22.
Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. 2012. Environmental DNA. Mol Ecol. 21:1789–1793.
Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol. 21:2045–2050.
Tang M, et al. 2014. Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res. 42:e166.
Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. Chapter 4:Unit 4.10.
Trivedi P, Anderson IC, Singh BK. 2013. Microbial modulators of soil carbon storage: integrating genomic and metabolic knowledge for global prediction. Trends Microbiol. 641–651.
Valentini A, et al. 2009. New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. Mol Ecol Resour. 9:51–60.
Wallinger C, et al. 2013. The effect of plant identity and the level of plant decay on molecular gut content analysis in a herbivorous soil insect. Mol Ecol Resour. 13:75–83.
Weitemier K, Straub SCK, Fishbein M, Liston A. 2015. Intragenomic polymorphisms among high-copy loci: a genus-wide study of nuclear ribosomal DNA in Asclepias (Apocynaceae). PeerJ. 3:e718.
Werren JH, Baldo L, Clark ME. 2008. Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol. 6:741–751.
Wheeler Q. 1982. The Biology of the Coleoptera. by R. A. Crowson. Review by: Quentin Wheeler. Syst Zool. 31:342–345.
Wheeler TJ, et al. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucl. Acids Res. (1 January 2013) 41 (D1): D70-D82 first published online November 30, 2012 doi:10.1093/nar/gks1265
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 13:59–69.
Yang Y, et al. 2013. The microbial gene diversity along an elevation gradient of the Tibetan grassland. ISME J. 1–11.
Yoccoz NG. 2012. The future of environmental DNA in ecology. Mol Ecol. 21:2031–2038.
Zarraonaindia I, Smith DP, Gilbert JA. 2013. Beyond the genome: community-level analysis of the microbial world. Biol Philos. 28:261–282.
Zhang Z. 2011. Animal biodiversity: an introduction to higher-level classification and taxonomic richness. Zootaxa 12:7–12.
Zhou X, et al. 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. Gigascience 2:4..

Associate editor: Gunter Wagner