Published online 22 February 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqac013

iLoci: robust evaluation of genome content and organization for provisional and mature genome assemblies

Daniel S. Standage[1], Tim Lai[2] and Volker P. Brendel[1,3,*]

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA, [2]Department of Mathematics, Indiana University, Bloomington, IN 47405, USA and [3]Department of Computer Science, Indiana University, Bloomington, IN 47405, USA

Received October 29, 2021; Revised December 23, 2021; Editorial Decision January 24, 2022; Accepted February 10, 2022

ABSTRACT

We introduce a new framework for genome analyses based on parsing an annotated genome assembly into distinct interval loci (iLoci), available as open-source software as part of the AEGeAn Toolkit (https://github.com/BrendelGroup/AEGeAn). We demonstrate that iLoci provide an alternative coordinate system that is robust to changes in assembly and annotation versions and facilitates granular quality control of genome data. We discuss how statistics computed on iLoci reflect various characteristics of genome content and organization and illustrate how these statistics can be used to establish a baseline for assessment of the completeness and accuracy of the data. We also introduce a well-defined measure of relative genome compactness and compute other iLocus statistics that reveal genome-wide characteristics of gene arrangements in the whole genome context. Given the fast pace of assembly/annotation updates, our AEGeAn Toolkit fills a niche in computational genomics based on deriving persistent and species-specific genome statistics. Gene structure model-centric iLoci provide a precisely defined coordinate system that can be used to store assembly/annotation updates that reflect either stable or changed assessments. Large-scale application of the approach revealed species- and clade-specific genome organization in precisely defined computational terms, promising intriguing forays into the forces of shaping genome structure as more and more genome assemblies are being deposited.

INTRODUCTION

The ready availability of next-generation sequencing (NGS) technologies has resulted in genome data for thousands of species, with no slowing down of data accumulation in sight. Given this volume of data, fast and accurate computational approaches are needed now more than ever to process the initial sequence data into meaningful units of knowledge about the sequenced genomes. The conventional paradigm for such tasks from the early days of genome sequencing is outdated. At that time, one could expect community groups to carefully assemble and annotate the genomes of their expertise, resulting over a period of time in gap-filled assemblies and refined documentation of genome content in terms of protein-coding genes, products of alternative splicing, noncoding RNA (ncRNA) genes, transposable elements, repetitive sequences and so forth. These genomes typically attained the status of 'reference model genomes'. However, the time-consuming and expensive efforts required are impractical for the vast majority of organisms currently being sequenced with NGS technologies.

Out of necessity, the old paradigm has for the most part been replaced by an implicit new standard: genome data are presented as massive short read collections available from databases such as the NCBI Sequence Read Archive (1) and in processed form as sets of assembled and computationally annotated scaffolds. Concomitantly, downstream analyses of these data have to be adjusted to scope and quality limitations intrinsic to the new data production process. First, assembly completeness will vary depending on the degree of read coverage and genome complexity (size and repetitiveness). Typically, assemblies will consist of tens to hundreds of large scaffolds, which in the best case can be ordered into linkage groups that approach pseudo-chromosomes, and additionally of manifold more short scaffolds, typically unplaced relative to any linkage groups. Second, annotation will commonly not have been expertly curated, but rather have resulted from first-pass outputs of annotation work-

*To whom correspondence should be addressed. Email: vbrendel@indiana.edu

flows such as AUGUSTUS (2), MAKER-P (3), BRAKER1 (4) or NCBI Gnomon (5).

The temporary nature of the data is also challenging. As additional sequences can often be acquired cheaply and easily for a species (e.g. genomic DNA reads for libraries of different insert sizes; RNA-seq reads from transcriptome studies under various conditions; or spliced alignments of protein sequences from a newly annotated, closely related species), both the species' genome assembly and its genome annotation may change. However, in the common scenario laid out earlier, the additional analyses will typically come without the community support to carefully sort out and document all the changes. Thus, over a short span of several years, there may be several annotation versions even for a single stable genome assembly, and it becomes difficult to track references to particular genes and genome features. A pertinent example from our experience is provided by the number of concurrent annotations in use for the honey bee (*Apis mellifera*) genome (6–8), including the current much more complete assembly based on long-read sequencing technologies (9).

How then should one compare results of a study on a current genome assembly and annotation version with previous results in the literature that used a prior assembly/annotation pair? How could one derive subsets of just those gene models that are solidly supported by evidence, to the extent that future genome-wide assembly/annotation improvements in all likelihood will not invalidate these current models? How does one disentangle artifacts of incomplete or inaccurate assembly/annotation from genuine species-specific genome features? What statistics should be calculated that capture a (newly sequenced) genome's content and organization and allow meaningful comparison with other genomes?

A solution to the problem must address the dual issues of reproducibility and scalability to accommodate thousands of genomes, each potentially with multiple assemblies and annotations. At the core of a solution must be the ability to distinguish what has changed from what has remained invariant when comparing one assembly/annotation pair to another. Discriminating between solid, reliable annotations and annotations of uncertain quality is also crucial in order to enable separation of technical artifacts from effects of interest rooted in the underlying genome biology. Typical examples of this challenge include annotation of untranslated regions (UTRs) and ncRNA genes or identification of transposable elements: comparing two genome annotations, one would like to know whether differences in UTR lengths or ncRNA gene and transposon content are due to insufficient data for annotation, annotation workflow settings or genome evolution.

Here, we present our AEGeAn (10) (analysis and evaluation of genome annotations) framework and toolset as a practical approach to facilitate comparisons across assemblies, annotations and genomes in view of the described challenges. AEGeAn generalizes our previously published ParsEval software (11) for comparing two sets of annotations for the same genome assembly. The basic idea is to represent a given assembly/annotation pair as a set of distinct units that can be largely independently characterized and updated. We show how the parsing of a genome into such distinct iLoci provides a suitable 'coordinate system' for working with rapidly changing genome assembly/annotation data. Applications to genome project data for various animal and plant species demonstrate how iLoci analyses can provide insights into genome organization and features, as well as assembly and annotation status.

MATERIALS AND METHODS

Toolkit scope and design

Motivated by the challenges of present-day genome data reviewed in the 'Introduction' section, we have developed a computational toolkit for the analysis and evaluation of genome annotations (10). AEGeAn includes functions that address questions of genome content, genome organization and cross-genome comparisons by precisely defined measures. The first range of questions concerning genome content includes: How many genes are annotated for a particular assembly/annotation pair? What can be said about their length, number of exons, nucleotide composition and other characteristics? What proportion of the genome is occupied by these genes? What fraction of genes are protein-coding versus ncRNA genes? How many of the gene models have support from transcript evidence, and how many genes can be identified as likely homologs of genes in other species?

These seemingly simple questions actually require very precise processing of the annotation file to be reproducibly and meaningfully answered. In particular, the handling of alternative transcription as well as overlapping gene models needs to be unambiguously defined.

The second range of questions concerning genome organization includes: How densely or sparsely packed are the genes? Is there clustering of genes, and if so, how large are these clusters, and what types of genes occur in clusters [e.g. (12)]? More generally, how is the intergenic space organized?

Third, all of the above questions are of interest in a comparative genomics context [e.g. (13)]. To what extent are genomes within a clade of species similarly organized? And, maybe even more intriguingly, to what extent is genome organization functionally important?

The design of the toolkit followed bioinformatics software engineering principles that emphasize reproducible, scalable and extensible open-source code that is easy to use and integrates with existing data repositories such as NCBI Genome (14) and other toolkits such as GenomeTools (15,16). Minimal required data input consists of a triplet of files (*G*, *A*, *P*): *G* is a set of one or more genome sequences provided in multi-FASTA formats; *A* is the associated genome annotation provided in GFF3 (17) format; and *P* is the set of annotated protein-coding gene products, in multi-FASTA format. For most sequenced genomes, such files are readily accessible at NCBI Genome (14). For simplicity, genome annotation provided in other formats would have to be converted to GFF3 input using widely available third-party scripts. In most cases, the protein file *P* could be generated from the CDS annotation in the GFF3 file. However, the more general specification of a separate *P* file accounts for nontemplated gene products that may be cited in the annotation file. AEGeAn includes format-checking utilities that flag semantic inconsistencies in the input and

suggest GenomeTools functions to remedy identified problems.

Conceptual definition of interval loci

To address the toolkit design prescriptions, we introduce a precise parsing of an assembly/annotation pair into smaller units, termed interval loci (iLoci), that provide a robust, granular and dynamic strategy for answering the biological questions posed earlier. Each iLocus is intended to capture the local genomic context of a genic or intergenic space, providing an alternative coordinate system to the conventional scaffold-based system, an alternative that is substantially more robust to changes in assemblies and annotations. Conceptually, an iLocus is a genomic interval, the boundaries of which are computed from annotated gene models, with an extension to include probable adjacent *cis*-regulatory regions. The precise procedure for computing iLoci is described in detail in the next section.

iLoci can be distinguished by various characteristics, as summarized in Figure 1. iLoci containing genes are referred to as giLoci, with those encoding protein-coding genes labeled as piLoci and those containing noncoding genes labeled as niLoci. piLoci harboring multiple overlapping gene models are designated complex (ciLoci), while those with a single isolated gene model are designated simple (siLoci). iLoci containing no gene models are designated as intergenic (iiLoci) if they are flanked on both sides by genes, or as incomplete fragments (fiLoci) if they are flanked on at least one side by an end of the corresponding parsed sequence.

To illustrate these concepts, Figure 2 shows the parsing of a hypothetical scaffold into its constituent iLoci. The parsing captures an intuitive and practical decomposition of the genome. The piLoci comprise a nonredundant set of protein-coding genes when reporting gene number or calculating descriptive statistics on gene features. However, more reliable results would be expected from the siLoci, or even better a subset of the siLoci with well-supported gene models. The ciLoci will typically require a whole lot more attention in order to establish whether the overlapping gene models reflect observed transcription or are artifacts of unresolved annotation conflicts.

Operational definition of iLoci

Basic procedure. Computing iLoci for an assembled contig/scaffold/pseudo-chromosome S depends on a set of intervals G (corresponding to gene models annotated on S) and an extension parameter δ (default value: 500). The basic procedure is described in Algorithms 1 and 2. In brief, the COMPUTELOCI algorithm computes a set of intervals L such that any two overlapping elements g_m, $g_n \in G$ are contained within and bounded by the same interval loci $\in L$. Although the algorithm is general, here g_m and g_n refer to gene bodies, defined as the interval from the start to the end of the respective annotated transcription events. The EXTENDINTERVALS algorithm then assesses each pair of adjacent intervals loc_m, $\text{loc}_n \in L$ and determines how far the intervals can be extended toward each other and whether any additional space remains between them for the creation of a third interval: if the number of nucleotides separating the two intervals $\text{dist}(\text{loc}_m, \text{loc}_n) > 3\delta$ nucleotides, then loc_m and loc_n will be extended toward each other by δ nucleotides, each designated as a giLocus, and the remaining space between them will be designated as an iiLocus; if $2\delta < \text{dist}(\text{loc}_m, \text{loc}_n) \leq 3\delta$, then loc_m and loc_n are extended toward each other equally until they meet, with extensions potentially as long as 1.5δ, to prevent recording a short iiLocus of positive length $\leq\delta$; and if $\text{dist}(\text{loc}_m, \text{loc}_n) \leq 2\delta$, loc_m and loc_n will each be extended by δ resulting in slightly overlapping iLoci. The rationale for allowing iLoci boundary overlaps in these cases is to assure that any giLoci selected for inspection will have δ nucleotide flanks around the transcript-based gene annotation. In both cases where $\text{dist}(\text{loc}_m, \text{loc}_n) \leq 3\delta$, the toolkit records a zero-length iLocus (ziLocus) between the adjacent giLoci for consistency and calculation of cumulative statistics described later.

Algorithm 1 Compute giLocus boundaries

```
procedure OVERLAP(loc,G)
    O ← loc
    for g′ ∈ G do
        if g′ overlaps with loc then
            O ← O ∪ g′
            mark g′ as visited
return O
procedure COMPUTELOCI(G,δ)
    L ← ∅
    for interval g ∈ G do
        if g is marked as visited then
            continue
        interval loc ← g
        mark g as visited
        while OVERLAP(loc,G) ⊃ loc do
            loc ← OVERLAP(loc,G)
        L ← L ∪ {loc}
    EXTENDINTERVALS(L,δ)
return L
```

Algorithm 2 Extend giLocus boundaries, identify iiLoci

```
procedure EXTENDINTERVALS(L,δ)
    for adjacent intervals x,y ∈ L do
        if dist(x,y) < 2δ then
            End(x) ← End(x) + δ
            Start(y) ← Start(y) − δ
        else if 2δ < dist(x,y) < 3δ then
            midpoint ← ⌊Average(End(x),Start(y))⌋
            End(x) ← midpoint
            Start(y) ← midpoint + 1
        else
            End(x) ← End(x) + δ
            Start(y) ← Start(y) − δ
            interval iiLocus ← [End(x) + δ + 1, Start(y) − δ − 1]
            L ← L ∪ {iiLocus}
```

Postprocessing to refine iLoci. The iLocus parsing procedure is designed with the canonical case of gene organization in mind: a single gene model flanked on both sides by hundreds or thousands of nucleotides of intergenic space.

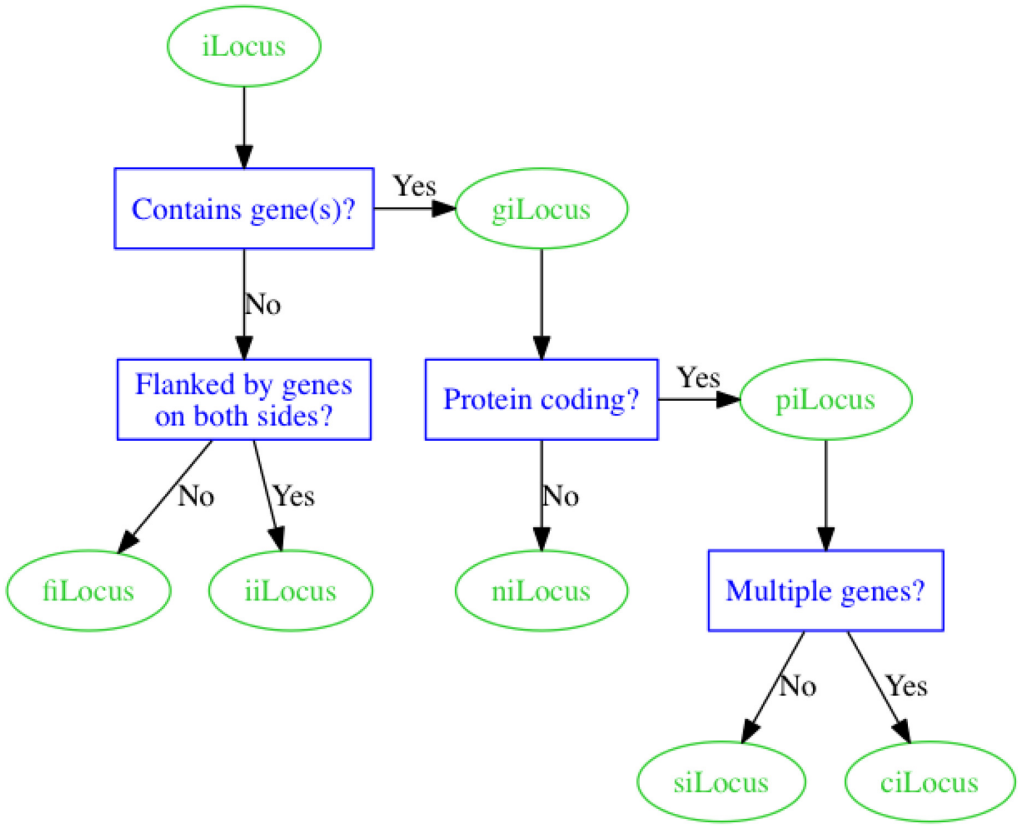

Figure 1. Classification of iLoci. Designation of iLocus types is shown in green, with classification logic described in blue. *Abbreviations*: fiLocus, fragmented intergenic iLocus; iiLocus, complete intergenic iLocus; giLocus, genic iLocus; niLocus, noncoding gene-containing giLocus; piLocus, protein-coding gene-containing giLocus; siLocus, simple piLocus; ciLocus, complex piLocus.

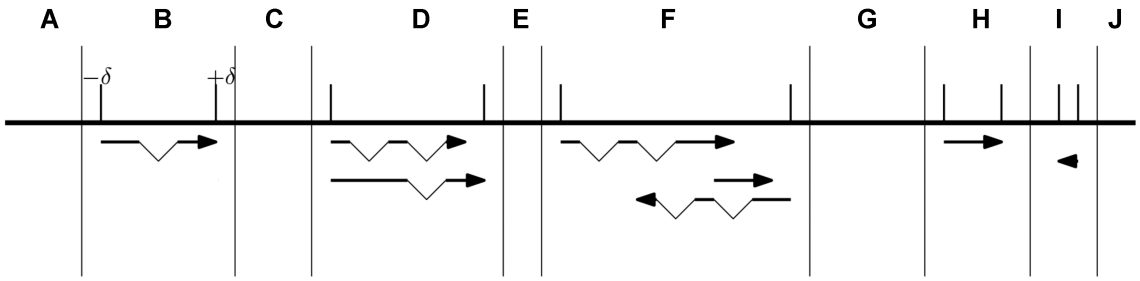

Figure 2. Parsing an annotated genome sequence into iLoci. The letters A to J indicate 10 adjacent iLoci on the genomic sequence (central horizontal line), separated by the long vertical bars. Gene annotations are shown underneath the genome sequence. Exons are schematized by bold horizontal lines and introns by the triangular thin lines connecting exons. Arrows indicate transcriptional direction. iLoci A, C, E, G and J are without gene annotation, with A and J representing potentially incomplete genomic fragments (fiLoci), and C, E and G representing complete intergenic regions (iiLoci). Each siLocus contains annotation for a single gene, which may involve a unique transcript (B, H and I) or multiple alternative transcripts (D). ciLocus F contains three distinct but overlapping genes. The boundaries of the giLoci are derived from the annotation ends, extended in each direction by δ. An exception occurs between giLoci H and I, where the extension would result in an iiLocus shorter than δ: in this case, the bordering giLoci (H and I) are extended toward each other to fill the entire space.

All eukaryotic genomes have exceptions to this case, some to a greater extent than others. The basic parsing procedure can handle some exceptions, such as genes separated by very little intergenic space, but there are additional exceptions that occur frequently enough to merit additional postprocessing and refinement.

The basic procedure places two gene models in the same iLocus if their gene bodies have any overlap. While this is intended to capture gene models that may be conflicting or misannotated and in need of additional attention to resolve coordinates, an unintended consequence is the occasional grouping of genes with a trivial amount of incidental overlap. For example, if two genes—each a few kilobases in length—happen to have 10–20 nucleotides of overlap in their UTRs, they should be separated and handled as distinct loci. In postprocessing, we enable splitting of such trivially overlapping iLoci by introducing two additional parameters: ω, the number of nucleotides that two gene mod-

els must overlap to remain in the same iLocus, and κ indicating whether that overlap is calculated using entire gene bodies ($\kappa = 0$) or just the coding sequences ($\kappa = 1$).

The initial procedure also groups ncRNA genes and protein-coding genes together if they overlap. In postprocessing, ncRNA genes and protein-coding genes are treated separately and will not be grouped in the same iLocus regardless of overlap, although overlapping ncRNA genes are grouped in the same niLocus.

An additional exception occurs when a gene resides completely within a single intron of another gene. These genes are placed in the same iLocus during the initial parsing procedure but are separated into distinct iLoci during postprocessing.

Implementation

In keeping with the conventions implemented by the GenomeTools library (15), most of the core functionality of the AEGeAn Toolkit (10) is implemented by means of *node streams* for sequential processing of genome features that are represented as *feature graphs*. In brief, genome features such as genes, exons, UTRs and coding sequences are represented as nodes in a directed acyclic graph, and parent/child relationships between features, denoted by *ID* and *Parent* attributes in GFF3, are represented as edges in the graph. Each connected component (CC) in the graph, typically corresponding to a gene and its subfeatures, is then processed sequentially by one or more node streams, each designed for a specific annotation processing task. One advantage of this approach is that it leverages streaming algorithms with a low memory footprint, as at most only a small number of CCs need be loaded into memory at any given moment.

The *AgnLocusStream* module in the AEGeAn Toolkit implements a node stream for computing iLocus boundaries. This node stream expects as input gene annotations (CCs with a gene feature as the root node) sorted by genomic position, but it is designed to work with arbitrary feature types. Initially, the node stream will collect a single gene feature from the input and store it in a buffer. Any subsequent gene features that overlap with genes in the current buffer (i.e. the leftmost position of candidate gene is less than or equal to the rightmost position of any gene in the buffer) are accumulated into the buffer. This continues until the node stream encounters a gene that does not overlap with the buffer, initiating two operations: first, the node stream emits a giLocus feature spanning all genes in the buffer; second, the node stream resets the buffer and begins accumulating the next gene or set of genes. A reference to the previously emitted giLocus is also maintained, enabling the refinement of boundaries between giLoci and, when appropriate, the designation of iiLoci, as described in Algorithm 2.

The AEGeAn Toolkit's *AgnLocusRefineStream* module implements a node stream for postprocessing the initial iLocus designations, as described in the previous section. Any genes belonging to the same giLocus that do not overlap by at least ω nucleotides in their gene bodies (or coding sequences if $\kappa = 1$), as well as genes contained completely within the intron of another gene, are split into distinct overlapping giLoci.

More generally, the AEGeAn Toolkit includes a variety of components. Node streams and other core components are implemented in the C language and organized into reusable modules. All core modules are compiled into a single shared object file to facilitate integration with other software by dynamic linking. Finally, a variety of executable programs for annotation processing and analysis composed from these core modules are also provided. In particular, the *LocusPocus* program provides the primary user interface to the *AgnLocusStream* and *AgnLocusRefineStream* modules. A detailed description of command-line usage and program inputs and outputs is provided in the AEGeAn Toolkit's source code distribution.

Genome content statistics

As discussed in the 'Introduction' section, derivation of genome characteristics for comparison across species requires selection of reliable subsets of data for analysis. The precise selection criteria used will depend on the questions being asked, but commonly involve a small set of descriptive statistics [see e.g. (18)] that can easily be computed from the iLocus sequence and/or associated annotation. These include the length and nucleotide composition of the iLocus itself, as well as the count, length and composition of corresponding features such as genes, RNAs, exons, introns and coding sequences. Statistics are computed by invoking the *stats* task of the AEGeAn *fidibus* script (see Additional File S2: wfscripts/run-fidibus-stats.ipynb) and are stored in tab-separated plain text (.tsv) files to facilitate import into popular statistics packages.

Additional characteristics for comparison and filtering may not always be directly accessible from the iLocus sequence or annotation but derive from computation using external data sources. Such values can then be attached to an iLocus annotation using key–value pairs in GFF3's attribute column. For example, gene model quality can be measured with statistics such as Maker's *annotation edit distance* (19) or the GAEVAL *integrity score* (20), and homology status can be determined via reciprocal BLAST searches or clustering of iLocus protein products.

Descriptive statistics are reported only for a single annotated transcript at each iLocus to ensure that aggregate statistics are not biased by redundancy in the data resulting from genes with many annotated isoforms, for example. The reported transcript is selected according to the amino acid length of its translation product: the transcript with the longest product is reported. In cases where multiple transcripts have translation products of identical length, the transcript with the lexicographically smallest ID attribute is reported, ensuring reproducible and deterministic reporting.

Cumulative lengths of different iLocus types are calculated after proper accounting of any iLocus overlaps to ensure each nucleotide in the genome is counted only once (see Additional File S2: wfscripts/make-Tables1-3.sh). When reported as a fraction of the entire genome, the genomic space occupied for different iLocus categories is calculated as a fraction of *effective genome size*, defined as the total number

of nucleotides in the genome that do not reside within fiLoci (see Figure 1). This will mitigate potentially confounding inflation of genome size by many short unannotated sequences or sequence fragments.

Genome organization statistics

Beyond genome content, the iLocus framework also allows systematic study of different aspects of genome organization. Here, we focus on gene orientation and spacing: are there species-specific patterns of gene arrangements, and how do natural genomes differ in these respects from statistical expectation [e.g. (21)]? Because of the flexible design of the code base described in the 'Implementation' section, these questions can easily be generalized and extended, for example with respect to selection of subtypes of genic loci.

To study gene orientation, the *LocusPocus* program reports for each iiLocus (see Figure 1) the transcriptional orientation of the flanking genic iLoci as FF, RR, RF or FR, corresponding to forward, reverse, outward and inward orientations, respectively. For example, FF indicates that both flanking genes are transcribed on the top strand relative to the given assembly and annotation. In the case that an iiLocus is flanked by one or more ciLoci (see Figure 1), the orientation of the gene models directly flanking the intergenic space is reported. Differences in occurrence numbers and lengths of outward and inward iiLoci are determined for possible interpretation in terms of promoter architecture: outward orientation for a short iiLocus might correspond to a bidirectional promoter. One could also identify the longest stretches of genes all on the forward strand, all on the reverse strand or periodically alternating between strands to probe the extent of colinear transcription.

Long iiLoci are flagged as regions for annotation review. More generally, for each giLocus (see Figure 1), the lengths of the flanking iiLoci are reported. In cases where a giLocus abuts or overlaps with another giLocus, the corresponding iiLocus length is set to zero, and the number of overlapping nucleotides is recorded. The software tracks these cases as ziLoci. The iiLocus lengths are used in two different ways to reveal gene spacing characteristics. First, the distribution of aggregate lengths of n adjacent iiLoci shows the mode of typical gene spacings as well as outliers. Second, overlapping or abutting giLoci are collapsed into merged iLoci (miLoci) during postprocessing and represent gene clusters; the resulting ziLoci are reflected in statistics that measure the characteristics of N adjacent iiLoci or of all iiLoci in aggregate.

To evaluate observed gene spacing patterns with statistical expectation, we implemented a procedure to generate randomized gene arrangements relative to a given input genome annotation. First, iLoci are computed with $\delta = 0$ to identify the precise boundaries of annotated genic regions. Next, giLoci are removed from the sequence and the remaining iiLoci are concatenated. Then, new positions are randomly selected from a uniform distribution for reinserting the giLoci in shuffled order into the sequence. As each giLocus is re-inserted, the genomic sequence is expanded, and all downstream re-insertion site positions are adjusted accordingly. Re-running the iLocus parsing procedure and computing neighbor statistics on these random arrangements provides a baseline for comparison, revealing how genome annotations as observed differ (at the genome scale) from what could be expected from a completely random arrangement of genes.

Comparing assembly/annotation pairs: iLocus stability

Given two assembly/annotation versions A and B for the same genome, the question arises how the A iLoci map onto the iLoci set calculated for B. Let us assume that B is a later, improved version of A. Two cases can be distinguished. In the first case, the genome assembly is the same for A and B, but the annotation has changed, for example by inclusion of newer experimental data that led to annotation of noncoding genes and novel splice forms or rejection of previous hypothetical gene structure models. In the second case, both the genome assembly and the annotation have changed, the former presumably due to additional genomic sequencing data that led to a less fragmented assembly. The mapping of iLoci may include several possibilities: (i) an A iLocus maps essentially unchanged to a B iLocus (although its genomic sequence identifier and coordinates may be different in a new assembly); (ii) an old iLocus may not map at all to the B set; (iii) set B may include novel iLoci; and (iv) there may be partial mapping of iLoci, for example when a novel noncoding gene annotation breaks up genomic space that had previously been annotated as intergenic space.

Mapping of the iLoci may involve sequence alignments spanning considerable gaps, as would be the case when a newer assembly provides gap filling compared to the older assembly. Thus, we chose the LASTZ pairwise aligner (22) and evaluate results based on the overall quality and length of the maximal chain of high-scoring segment pairs. Specifically, query iLoci sequences were matched against target iLoci sequence sets with LASTZ parameters *–ambiguous=iupac –filter=identity:95 –chain* (Additional File S2: comparisons/run*.sh). The output of LASTZ was processed as follows to provide a classification of a query locus *qlocus* of length *qlength* based on any chained matches (chain length *clength*) against a subject locus *slocus* of length *slength*:

- *qlocus* is without hits;
- *qlocus* has no qualifying hits and is designated as unmapped;
- *qlocus* matches *slocus* such that $clength/qlength \geq 0.9$ and $clength/slength \geq 0.9$ and is designated as conserved;
- *qlocus* matches *slocus* such that $clength/qlength \geq 0.9$ and $clength/slength < 0.9$ and is designated as contained;
- *qlocus* matches *slocus* such that $clength/qlength < 0.9$ and $clength/slength \geq 0.9$ and is designated as anchored;
- *qlocus* is designated redefined if there are subject loci with respect to which it is contained and others with respect to which it is anchored.

Cases in which a query iLocus is conserved with respect to multiple subject iLoci may occur when the assemblies

Table 1. iLocus content of genomes from 10 model organisms and 4 additional species

Species	Mb[a]	#Seq[b]	fiLoci	iiLoci	niLoci	siLoci	ciLoci
Saccharomyces cerevisiae	12.1	16	11	274	393	5777	101
Caenorhabditis elegans	100.3	6	3	6152	19 843	19 373	359
Chlamydomonas reinhardtii	120.2	1556	1487	6245	0	14 253	42
Medicago truncatula	429.6	40	64	28 175	5929	31 350	142
Anopheles gambiae	265.0	8089	8037	7726	644	11 986	318
Drosophila melanogaster	143.7	1869	1874	3452	3356	12 626	650
Xenopus tropicalis	1437.5	7727	8004	18 580	5181	21 454	403
Danio rerio	1371.7	1060	1295	23 979	15 208	25 392	481
Mus musculus	2725.5	21	42	23 771	13 342	21 117	616
Homo sapiens	3088.3	24	48	22 242	16 584	18 947	927
Volvox carteri	137.7	1251	1198	7790	0	14 346	44
Polistes dominula	208.0	1483	1665	3969	1049	9715	282
Daphnia pulex	197.3	5191	4759	13 052	0	30 454	160
Manacus vitellinus	1072.3	3619	3760	11 319	1999	13 096	193

[a]Total number of nucleotides in the genome assembly.
[b]Total number of assembled (pseudo-)chromosomes plus any unplaced genomic scaffolds.

contain duplicated genes and are noted in the LASTZ parsing script output.

Comparing genome content and organization between related genomes: homologous iLoci

Given a set of annotated genome assemblies for a clade of related species, we compute homologous iLoci (hiLoci) via a protein clustering procedure. For each species, a representative protein sequence is selected for each piLocus (see Figure 1), as described in the 'Genome content statistics' section. The distinct protein complements from all species are then combined, and the aggregate collection of protein sequences is clustered using cd-hit (23).

In brief, cd-hit processes proteins iteratively from longest to shortest. The first protein is assigned to a cluster by itself and is designated the *representative sequence* of the cluster. Each subsequent protein is compared to all previous clusters: If the alignment of the protein to a cluster's representative sequence satisfies the specified sequence identity, length similarity and alignment coverage criteria, it is added to that cluster, and the program advances to the next protein; if a protein cannot be added to any cluster by user-specified clustering criteria, it is placed in a new cluster by itself and designated the representative sequence of that cluster.

Following the clustering procedure, a data structure designated as hiLocus is created for each protein cluster, and the piLoci corresponding to the proteins in that cluster are assigned to that hiLocus. The hiLocus thus provides a link between piLoci from related species and a relative measure of how well conserved the corresponding protein is within the given clade.

This protein clustering procedure is invoked using the *cluster* task of the AEGeAn *fidibus* script. The default parameters are as follows: sequence identity ≥50%; length difference ≤50%; alignment coverage for longer sequence ≥60%; and alignment coverage for shorter sequence ≥60%. On the command line, these parameters are specified as `-c 0.50 -s 0.50 -aL 0.60 -aS 0.60`. The default values can be overridden, and additional criteria can be set by the user.

Data sets analyzed

We retrieved RefSeq genome assemblies and corresponding annotations for 10 model organisms (as listed in Table 1) to illustrate the utility of iLoci for providing a descriptive overview of genome composition and organization. Species were selected to provide a broad sampling of eukaryotic diversity, with a preference for robust model organisms with mature chromosome-level genome assemblies and extensive community-supported annotation. For each species, we computed iLoci and associated feature statistics, including length, nucleotide composition, exon count and *effective length*, using standard fidibus build tasks as described before.

Using iLocus summaries of these 10 model organisms as a baseline for comparison, we characterized the genome content and organization of 4 additional species of interest that serve as important experimental models for evolutionary and ecological studies: the microcrustacean *Daphnia pulex*, the primitively eusocial paper wasp *Polistes dominula*, the green alga *Volvox carteri* and the suboscine passerine bird *Manacus vitellinus*. These four genomes were processed using the same procedure as the 10 model organisms. Precise configurations and commands run for all analyses are available in Additional File S1 and at https://github.com/BrendelGroup/iLoci_SLB22NARGB. The complete data work presented in this paper is available for download at https://BrendelGroup.org/research/publications.php.

Finally, we retrieved and processed, in the same manner as mentioned earlier, large collections of genomes from NCBI RefSeq branches and computed branch averages of all statistics of interests. We report on these statistics as another baseline for genome evaluation in taxonomic evolutionary context.

Classifying hiLoci from a clade of nine chlorophyte species. To investigate the extent of gene conservation in the green algae (phylum: Chlorophyta), we collected and processed data for nine chlorophyte species (*Auxenochlorella protothecoides*, *Chlamydomonas reinhardtii*, *Chlorella variabilis*, *Coccomyxa subellipsoidea*, *Micromonas commoda*, *Micromonas pusilla*, *Ostreococcus lucimarinus*, *Ostreococcus tauri* and *V. carteri*), as well as four land plants (*Ara-*

bidopsis thaliana, *Brachypodium distachyon*, *Medicago truncatula* and *Oryza sativa*) as an outgroup. Retrieval of annotations and sequences and calculation of hiLoci were invoked using standard procedures as described in the 'Materials and Methods' sections (and Additional File S2: README_explore-Chlorophyta.md). Following the protein clustering procedure, each hiLocus was assigned a preliminary classification: *highly conserved* if it had a representative from each of the nine chlorophyte genomes; *conserved* if it had a representative from at least four chlorophyte genomes; *matched* if it had a representative from at least two genomes (including the outgroups); and *unmatched* if it had a representative from only a single genome.

hiLoci initially classified as *unmatched* were subjected to additional screening to distinguish conserved proteins lacking a nearly full-length match (due to incomplete or incorrect annotation, or true evolutionary divergence) from orphan proteins without any reliable match. hiLoci with a BLASTP match against another chlorophyte species (`-evalue 1e-10`) were reclassified as *matched*, while those lacking a match were reclassified as *orphan*.

RESULTS

iLoci provide an informative decomposition of genome content

We computed iLoci for 10 model organisms representing a wide range of eukaryotic diversity and provide a summary of each genome and its iLocus complement in Table 1 (for workflow commands, see Additional File S2: README_refr-genome-summary.md). The genome assembly sizes in this sampling of eukaryotes span two orders of magnitude, ranging from 12.1 Mb in *Saccharomyces cerevisiae* to over 3 Gb in *Homo sapiens*. Several genomes are represented exclusively by chromosome sequences, some exclusively by unplaced genomic scaffolds and some by a combination of both. The number of fiLoci (see Figure 1), with a strict upper bound of twice the number of assembled sequences, is informative primarily with respect to assembly status. For most of these genomes, the observed number of fiLoci is close to half of the upper limit. There are two reasons for why the observed number of fiLoci can be lower than the upper limit: (i) the presence of gene annotations near the end of a genomic sequence (within $2 \times \delta$, in which case no fiLocus is recorded) and (ii) the inclusion of unannotated (short) scaffolds in the genome sequence set (which results in one fiLocus per unannotated scaffold spanning the entire sequence). Here, for example, the numbers for *S. cerevisiae* are consistent with a compact genome, the numbers for *C. reinhardtii* are consistent with a fragmented genome assembly including many unannotated scaffolds and the numbers for mouse and human are consistent with complete genomes.

iiLoci correspond to intergenic DNA and are reflective of genome organization. There can be at most $n - m$ iiLoci in a genome with n genes and m annotated sequences, but closely spaced genes will reduce the number of observed iiLoci, as will the presence of unannotated scaffolds.

The abundance of piLoci in each genome (representing distinct protein-coding regions) spans just a single order of magnitude, from 5878 piLoci in *S. cerevisiae* to 31 492 in *M. truncatula* (Table 2). The total space occupied by piLoci, however, spans two orders of magnitude, similar to genome size. This is explained by a distinct contrast in length of simple iLoci between vertebrates and the other species (Additional File S1: Supplementary Figure S1; for workflow commands, see Additional File S2: notebooks/make-SF1.ipynb), the compound result of increases in both intron abundance and length (Additional File S2: notebooks/make-SF5c-SF8.ipynb). We note that while the protein-coding gene portion of the human genome is commonly reported as 2–4%, this refers only to protein-coding exons. The inclusion of introns and UTRs places the protein-coding gene fraction of the genome at ∼40% for both human and mouse. ciLoci (see Figure 1) are present in dozens to hundreds in most genomes, accounting for only a small proportion of genes.

iLoci reflect patterns of genome organization

Gene clustering is abundant in eukaryotic genomes. There are well-described examples of gene clusters in eukaryotic genomes, such as those associated with *Hox* genes (24). *Hox* clusters are composed of functionally related developmental genes with a conserved colinear arrangement, a common direction of transcription, occurring in close proximity in the genome. More generally, gene clusters described in the literature need not be comprised only of genes that are directly adjacent but are loosely defined as sets of genes of a common function situated much closer to each other than would be expected by chance (25). However, the spatial distribution of genes in general, the extent to which genes are tightly packed throughout the entire genome and the characteristics of these gene-dense regions have not been extensively studied in eukaryotes. miLoci (see Figure 1) are not precisely equivalent to gene clusters, but they do provide utility that gene clusters—as conventionally defined—do not: a well-defined unit of analysis for investigating the spatial distribution of genes genome-wide. Using miLoci, we surveyed genome organization in the selected 10 model organisms.

Genes cluster together frequently in eukaryotic genomes. The most frequent groupings involve a small number (2–4) of genes (see Table 3), but all genomes include larger clusters involving dozens or even hundreds of tightly packed genes. The budding yeast *S. cerevisiae* is an extreme example, populated almost entirely by just 294 miLoci encompassing all but 176 genes in the entire genome. *Caenorhabditis elegans* and *Drosophila melanogaster* also bear signatures of a higher overall level of genome compactness, with larger numbers (and overall proportion) of genes merged into miLoci and a larger proportion of genomic sequence occupied by miLoci.

In general, clustered genes do not differ substantially in length or nucleotide composition from spaced out genes. However, especially among large miLoci, clustered genes are often functionally related. The longest miLoci in the human genome include a cluster of 22 snoRNA genes on chromosome 14, a cluster of 19 genes from AP2A1 to NUP62CL on chromosome 19 and a cluster of keratin-associated proteins on chromosome 21, while in mouse the longest miLocus is comprised of 76 microRNA genes, on chromosome 2.

Table 2. Summary of piLoci from genomes of 10 model organisms and 4 additional species

Species	piLoci	Occupancy[a]	Single-exon piLoci
Saccharomyces cerevisiae	5878	11.4 Mb (94.6%)	5613 (95.5%)
Caenorhabditis elegans	19 732	75.7 Mb (75.5%)	685 (3.5%)
Chlamydomonas reinhardtii	14 295	74.1 Mb (68.2%)	1127 (7.9%)
Medicago truncatula	31 492	158.6 Mb (37.0%)	5701 (18.1%)
Anopheles gambiae	12 304	83.8 Mb (35.9%)	1154 (9.4%)
Drosophila melanogaster	13 276	95.4 Mb (70.0%)	2100 (15.8%)
Xenopus tropicalis	21 857	687.2 Mb (50.4%)	1344 (6.2%)
Danio rerio	25 873	828.8 Mb (61.2%)	1095 (4.3%)
Mus musculus	21 733	1034.4 Mb (38.9%)	2370 (11.0%)
Homo sapiens	19 874	1240.1 Mb (41.2%)	1270 (6.5%)
Volvox carteri	14 390	89.2 Mb (69.1%)	1086 (7.5%)
Polistes dominula	9997	137.7 Mb (72.4%)	405 (4.1%)
Daphnia pulex	30 614	89.2 Mb (54.3%)	5053 (16.5%)
Manacus vitellinus	13 289	461.4 Mb (44.6%)	529 (4.0%)

[a]Total number of nucleotides occupied by piLoci and the corresponding fraction of effective genome size.

In the nonmammal vertebrates, the longest miLoci consist exclusively of long stretches of hundreds of tRNA gene annotations. As tRNA-derived SINE transposons are known to be abundant in at least one of these species (26), and no annotations for such transposons appear to be included in the RefSeq annotation, it is likely these miLoci capture large clusters of misannotated repetitive elements. The latest annotation of *M. truncatula* includes several rRNA gene clusters, identified in the miLoci list by our default parameter $\delta = 500$.

The distribution of miLoci along the chromosome is mostly uniform for compact genomes such as *Drosophila* and *C. elegans*. For less compact genomes, we observe variation in the uniformity of miLocus distribution. For example, in *Medicago*, miLoci appear to be more frequent at the chromosome ends, while in vertebrate species a depletion of miLoci in pericentromeric regions is most obvious (Additional File S2: notebooks/explore-miLoci.ipynb).

The spacing of genes over longer ranges is revealed by distributions of aggregate lengths of r adjacent intergenic iLoci [*r-scans* (27)]. Long-range spacing of genes varies considerably in eukaryotes, with some species exhibiting homogeneous gene spacing over relatively short spans (spans of 5–10 genes in *C. elegans* and *M. truncatula*), and others showing heterogeneous spacing even over long spans (spans of >30 genes in *Mus musculus*; see Additional File S1: Supplementary Figure S2; for workflow commands, see Additional File S2: notebooks/make-SF2.ipynb).

Gene orientation. The iLocus framework provides a convenient approach to analyzing the strand locations of genes. We categorize the iiLoci (see Figure 1) based on the length and orientation of the flanking genic iLoci, as described in the 'Genome organization statistics' section. The stacked bar plots showing the distribution of iiLocus length, grouped by orientation, are given for the 10 model organisms in Additional File S1: Supplementary Figure S3 (generated by Additional File S2: notebooks/make-SF3.ipynb) and for the randomized gene positioning control in Supplementary Figure S4 (generated by Additional File S2: notebooks/make-SF4.ipynb). Note that for this study, iLoci were determined with $\delta = 0$ to allow investigation of short intergenic regions (for workflow commands, see Additional File S2: wfscripts/run-explore-gene-orientation.sh).

Comparing the two sets of figures, it is clear that the iiLocus orientation types do not occur in random proportions in the natural genomes. However, the patterns of deviation depending on iiLocus length are different between species. *Anopheles* and *Drosophila* show the most even pattern across all length bins. Mouse and human show an intriguing preponderance of the outward (RF) orientation type for short iiLoci but relative avoidance of the type for longer iiLoci. Zebrafish (*Danio rerio*) seems to favor the colinear types FF and RR until the longest iiLocus length bins. *Caenorhabditis elegans* shows a preference for FF in the same length ranges. Lastly, *M. truncatula* has high numbers of inward (FR) orientation types for iiLocus lengths up to around 1 kb. Detailed interpretations of these differences would involve exploration of gene types and chromosomal location, but here we simply emphasize the readily availability of these genome organization data within the iLocus framework.

Compactness of eukaryotic genomes varies widely. We further explore the notion of compactness of a genome by two complementary measures calculated on the constituent chromosome or long scaffold sequences: ϕ, defined as the fraction of giLoci in the sequence merged into miLoci; and σ, defined as the proportion of the sequence occupied by miLoci. Distinct quadrants in the plot reflect characteristic overall genome organization. Low values of ϕ associated with low values of σ (lower left) correspond to genes as 'islands' in an 'ocean' of intergenic (presumably repetitive) DNA. High values of ϕ associated with low values of σ (lower right) correspond to 'archipelagos' of genes, and high values of ϕ associated with high values of σ (upper right) correspond to 'compact' (or 'continental') genome organization.

Let the average iiLocus length be ρ times the average giLocus length (g), and let m and n be the number of giLoci and iiLoci, respectively. Then, $\sigma = \phi mg/(mg + n\rho g)$, and if ϕ is small, then $n \approx (1 - \phi)m$, and the following approxi-

mation holds:

$$\sigma \approx \frac{\phi}{1+(1-\phi)\rho}. \quad (1)$$

When ϕ is close to 1, then also $\sigma \approx \phi$, unless the genome had very distinct densely packed multigenic regions separate from substantial nongenic regions. Thus, major deviations from the expected curve are revealing of extreme genome organization, as discussed earlier. Figure 3A gives the curves for ρ equal to 0.1, 1, 2, 4 and 8 (produced by Additional File S2: notebooks/make-F3a-SF6-SF7.ipynb; for workflow commands, see Additional File S2: README_refr-genome-compactness.md and Additional File S2: wfscripts/run-explore-compactness-refr.ipynb).

Empirical (ϕ, σ) values calculated for continuous genome sequences of at least 1 Mb for the 10 model species reveal a wide range of genome compactness across eukaryotes, yet show remarkable consistency within species (Figure 3A) and even within clades and branches (as confirmed by sampling of additional species; see Additional File S1: Supplementary Figure S5A–D and Figure 3B; for workflow commands, see Additional File S2: notebooks/make-F3b.ipynb and Additional File S2: wfscripts/run-explore-compactness-othg.sh and taxa/README.md). Genome compactness scales roughly and inversely with genome size, at least across major clade divisions and levels of organismal complexity. Within Chlorophyta, compactness scales almost perfectly with genome size, although this trend is not maintained in clades characterized by larger genome sizes. Consistent with previously described observations, sequences from *S. cerevisiae* are the most compact of all 10 model organisms analyzed. Alternatively, very few sequences show extremely low levels of overall compactness: only 6 sequences of the 10 model organisms have $\phi < 0.2$ and $\sigma < 0.2$, 2 of which correspond to mammalian sex chromosomes, with the other 4 corresponding to unplaced scaffolds from *Xenopus tropicalis*. This trend continues with most other genomes and annotations from RefSeq, with only a few genome averages below both thresholds (Figure 3B). Likewise, very few sequences are dominated by an 'archipelago'-type organization (high ϕ and low σ). Those with $\phi > 0.7$ and $\sigma < 0.3$ are annotated almost exclusively with long stretches of dozens or hundreds of tRNA gene annotations in *X. tropicalis* and *D. rerio*.

Adjusting the value of the δ parameter used in the initial iLocus parsing procedure can have a moderate effect on (ϕ, σ) measures of genome compactness. As expected, reducing the length of δ extensions results in a decrease in reported genome compactness, while increased values of δ result in reports of higher genome compactness (Additional File S1: Supplementary Figure S6). However, relative compactness between different genomes appears robust to changes in the δ parameter.

Gene clustering occurs more frequently than expected by chance. To investigate whether gene clustering occurs more frequently than expected by chance, we computed random arrangements of genes on each long (≥1 Mb) chromosome or scaffold sequence and recomputed iLoci and associated summary statistics for comparison with the observed annotation.

Random positioning of genes results in decreased levels of gene clustering across all species as reflected by several measures: a decrease in the number of miLoci; a decrease in the space occupied by miLoci; a decrease in the number of genes per miLocus; and an increase in the number of singleton genes not associated with miLoci (Additional File S1: Supplementary Table S1). Signatures of genome compactness are also influenced by random arrangement of genes, reflecting less compactness relative to the actual annotated positioning of genes. The (ϕ, σ) statistics calculated on long genomic sequences are consistently lower for random arrangements than actual arrangements for all model species (Additional File S1: Supplementary Figure S7), with the exception of the extremely compact *S. cerevisiae* genome.

'LocusPocus Fidibus': an incantation for any genome

Evaluating new genome assemblies and annotations is a common and critical challenge in contemporary biology but is hampered by limited community bioinformatics support and the lack of precise standards for systematic comparisons of genome content and arrangement. Having explored the range of genome composition and organization in eukaryotic model organisms, we turn now to the question of newly sequenced genomes: How does the new genome fit into the broader universe of eukaryotic organisms, and more interestingly, how does the new genome compare to genomes from closely related species?

iLoci address these challenges both by offering a well-defined 'common currency' for comparisons of genome content and organization and by providing associated software tools to facilitate analysis and re-analysis of old and new data alike. The *LocusPocus* and *Fidibus* programs are designed for painless adoption by researchers with minimal bioinformatics expertise and require only a small number of standard input files. In return, they produce a wealth of descriptive statistics not only on iLoci but also on their constituent genes, transcripts and associated features.

With baseline expectations about eukaryotic genome content and organization established by iLocus analysis of large numbers of genomes from RefSeq, including 10 model organism genomes, we now demonstrate how these tools can be applied to evaluate genomes of particular species of interest.

Volvox carteri. The green algae (phylum: Chlorophyta) diverged from land plants an estimated 1 billion years ago (28) and encompass a diverse set of organisms ubiquitous in marine and soil environments. Chlorophytes exhibit substantial variation in physical stature, genome size and cellular complexity, and include many important systems for study of the evolution of multicellularity and photosynthesis. The publication of the *V. carteri* genome (29) reported over 5000 protein families conserved between *Volvox* (a multicellular alga) and *C. reinhardtii* (a unicellular relative), accounting for over a third of both species' respective proteomes.

The genome content of *Volvox* is very similar to that of *Chlamydomonas* across a variety of iLocus measures. Characteristics of protein-coding regions in particular (summarized in Tables 1 and 2) show striking similarity: piLoci (see Figure 1) account for 89.2 Mb (69.1%) of the *Volvox*

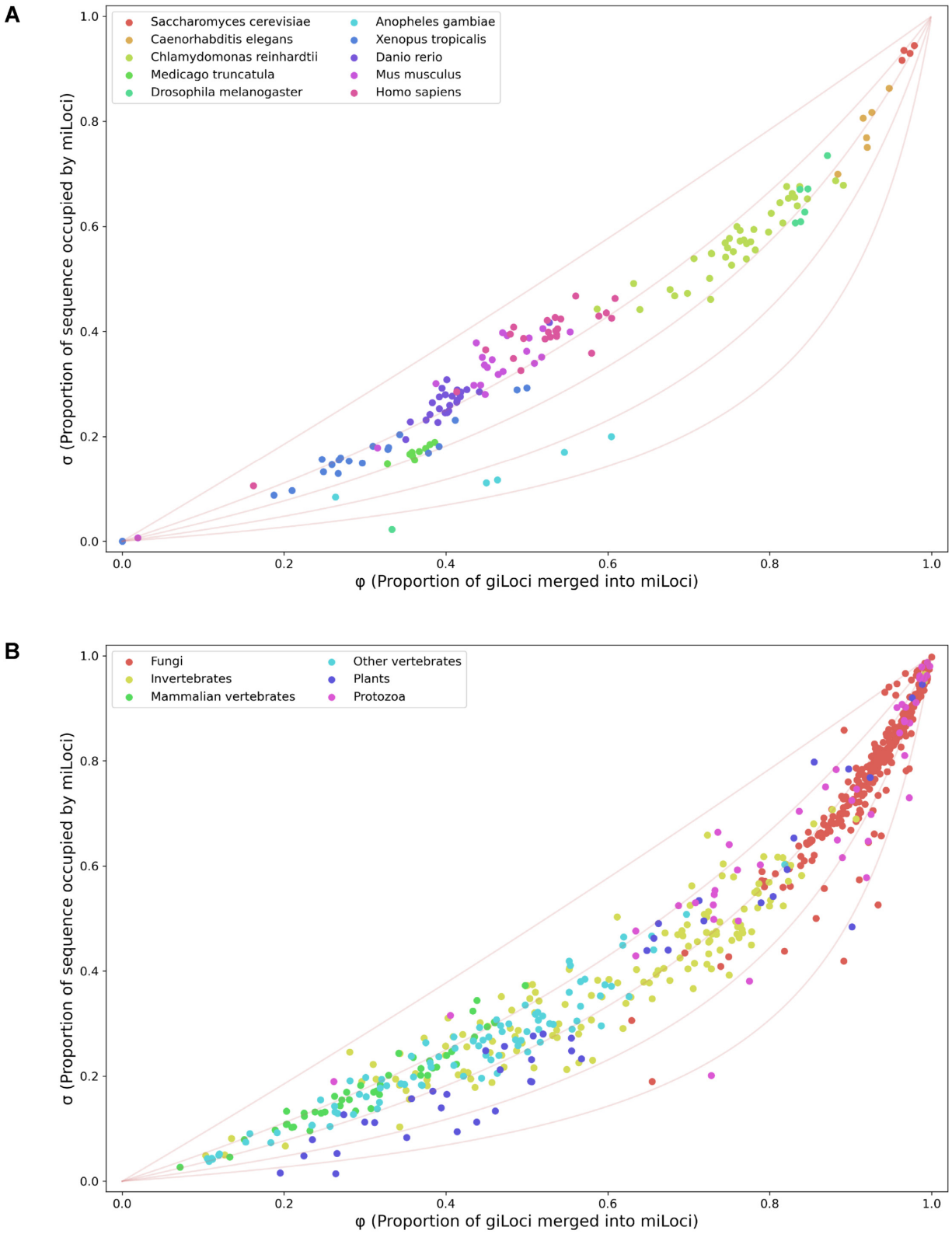

Figure 3. Genome compactness. **(A)** Ten reference genomes. The curves correspond from top to bottom to the theoretical ϕ, σ functions (Equation 1) for ρ equal to 0.1, 1, 2, 4 and 8, respectively. Each data point corresponds to a sequence of length at least 1 Mb. Short giLoci (lower 5%) and long iiLoci (top 5%) were removed for each genome prior to calculation. **(B)** Compactness as a persistent genome characteristic. Centroids of (ϕ, σ) values calculated as in panel (A) for different genomes from the indicated taxonomic groups.

Table 3. Summary of miLoci from genomes of 10 model organisms and 4 additional species

Species	miLoci	Occupancy[a]	Median no. of genes[b]	Singletons[c]
Saccharomyces cerevisiae	294	11.1 Mb (92.0%)	12	176 (2.8%)
Caenorhabditis elegans	4496	74.1 Mb (73.9%)	5	2425 (6.1%)
Chlamydomonas reinhardtii	3029	54.5 Mb (50.2%)	3	3796 (26.6%)
Medicago truncatula	5715	61.1 Mb (14.3%)	2	22 657 (60.5%)
Anopheles gambiae	2036	26.2 Mb (11.2%)	2	6521 (50.4%)
Drosophila melanogaster	2155	75.2 Mb (55.2%)	4	2626 (15.8%)
Xenopus tropicalis	3224	174.0 Mb (12.8%)	2	17 698 (65.5%)
Danio rerio	5843	301.6 Mb (22.3%)	2	19 348 (47.1%)
Mus musculus	6039	661.9 Mb (24.9%)	2	18 558 (52.9%)
Homo sapiens	6790	932.8 Mb (31.0%)	2	16 668 (45.7%)
Volvox carteri	3229	57.0 Mb (44.1%)	2	5256 (36.5%)
Polistes dominula	2085	74.1 Mb (38.9%)	3	2870 (26.0%)
Daphnia pulex	6252	58.4 Mb (35.6%)	3	9934 (32.4%)
Manacus vitellinus	2156	118.5 Mb (11.5%)	2	10 061 (65.8%)

[a]Total number of nucleotides occupied by miLoci and, in parentheses, the corresponding fraction of effective genome size.
[b]Median gene count per miLocus.
[c]Total number of giLoci not contained in miLoci and, in parentheses, corresponding fraction of all giLoci.

genome [compared to 74.1 Mb (68.2%) of the *Chlamydomonas* genome], and both genomes harbor a similar number of single-exon piLoci (1086 versus 1127, respectively) and very few complex iLoci (44 and 42, respectively). With respect to genome organization, *Volvox* and *Chlamydomonas* contain comparable numbers of merged iLoci (3229 and 3029, respectively; Table 3) and exhibit a remarkably similar level of gene density. The (ϕ, σ) values measuring genome compactness of the two species fall within a nearly identical range, with *Volvox* shifted to slightly lower values (Additional File S1: Supplementary Figure S5A, produced by Additional File S2: notebooks/make-F4a-F4b-SF5a.ipynb). These observations are consistent with the claims that, despite an estimated 50–200 million years of divergence and major differences in cellular complexity, the genomes of *Volvox* and *Chlamydomonas* are impressively similar (29).

With several representative chlorophyte genomes now available from RefSeq (30), we leveraged iLoci to characterize the extent of gene conservation in *Volvox* relative to the entire phylum. piLoci from all nine species were grouped together as hiLoci based on a clustering of their protein products, and the relative conservation status of each hiLocus was determined (see the 'Materials and Methods' section). Figure 4 presents a breakdown of all nine genomes according to iLocus type and conservation status, showing both the number of iLoci in each category and the proportion of the genome occupied by iLoci from each category (figure produced by Additional File S2: notebooks/make-F4a-F4b-SF5a.ipynb). Counts and aggregate space occupied by intergenic regions and assembly fragments (iiLoci and fiLoci, respectively) reflect the diversity of genome size and gene density across Chlorophyta, ranging from 10–25 Mb genomes almost completely devoted to protein-coding genes (in *Micromonas* and *Ostreococcus*) to genomes well over 100 Mb in size with abundant intergenic space (in *Volvox* and *Chlamydomonas*).

A small number of piLoci from each genome are designated as *orphans*, indicating no reliable protein match in any other species, while the majority are designated as *matched*, having at least one match in another species. The designations *conserved* and *highly conserved* were applied only to hiLoci whose protein products are well conserved throughout the phylum (*conserved*: conserved in at least four species; *highly conserved*: conserved in all nine species) and differ in amino acid length by no more than a factor of 2 within a hiLocus. Given these stricter criteria, we observe on the order of 100 *highly conserved* piLoci and 1000 *conserved* piLoci in each species. A total of 3130 *Volvox* piLoci and 3261 *Chlamydomonas* piLoci were grouped into 2928 common hiLoci, 2803 of which contain a single ortholog from both species.

Highly conserved piLoci are associated with a variety of cellular components and processes, most prominently proteins related to ribosomes and kinase/phosphatase activity. The vast majority of orphan piLoci are annotated as 'predicted' or 'hypothetical proteins'. Among the handful with functional annotations, flagellar-associated proteins are prominent in *C. reinhardtii* orphans, while Jordan transposition proteins are prominent in *V. carteri* orphans.

Polistes dominula. The paper wasp *P. dominula* is an important model for the study of social behavior and evolution and was one of the first species of the family Vespidae to have its genome sequenced (31). The *Polistes* genome is intermediate across many measures relative to the survey of 10 reference genomes, in particular the 2 insect genomes (the fruit fly *D. melanogaster* and the mosquito *Anopheles gambiae*). *Polistes* contains 3969 intergenic iLoci occupying 48.8 Mb (23.4%) of the genome, compared to 3452 intergenic iLoci occupying 35.5 Mb (24.7%) of the *Drosophila* genome and 7726 intergenic iLoci occupying 149.2 Mb (56.3%) of the *Anopheles* genome (Table 1 and Additional File S2: notebooks/make-SF5c-SF8.ipynb). *Polistes* is distinct from the other insects, however, in that both simple and complex iLoci are less abundant in its genome, and yet collectively they account for a larger proportion of the genome and a larger amount of absolute space (Tables 1 and 2). Similar results are observed when compared against invertebrates: raw counts of iLoci are comparable across each category, with a decreased number of protein-coding iLoci, yet the

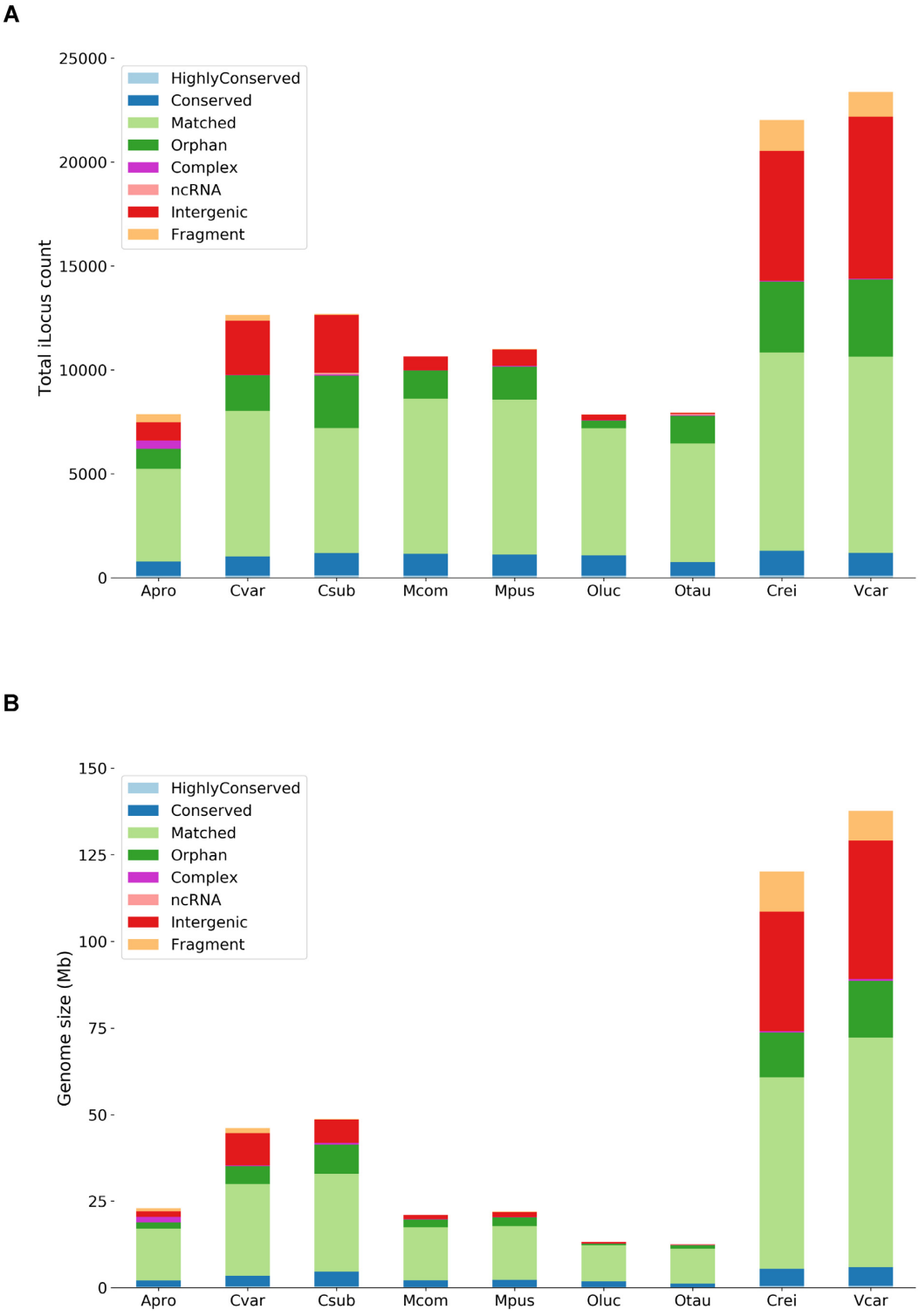

Figure 4. Breakdown of piLocus conservation status across chlorophyte species.

relative occupancy of these iLoci is greater (Additional File S1: Supplementary Table S2).

In terms of gene organization, the *Polistes* genome harbors 2085 merged iLoci, compared to 2155 in *Drosophila*, 2036 in *Anopheles* (Table 3) and a median of 2298 merged iLoci in invertebrates (Additional File S1: Supplementary Table S3). The (ϕ, σ) statistics computed on merged iLoci reveal an intermediate level of genome compactness (Additional File S1: Supplementary Figure S5B, produced by Additional File S2: notebooks/make-SF5b.ipynb). The increased dispersion of (ϕ, σ) values per sequence in *Polistes* is reduced in the longest genomic scaffolds, likely reflecting local fluctuations in genome organization that are evened out in the pseudo-chromosome-level assemblies for *Drosophila* and *Anopheles*.

Daphnia pulex. The water flea *D. pulex* is a species of ecological and evolutionary interest and was the first crustacean genome to be sequenced (32). Like *P. dominula*, characteristics of genome content and organization in *D. pulex* are intermediate relative to the two arthropods surveyed. The most striking feature of the *Daphnia* genome is the large number of annotated genes and large fraction of single-exon protein-coding iLoci (see Table 2). *Daphnia* contains 30 614 piLoci, more than twice the number in *Drosophila* and *Anopheles*, and the median in all invertebrates (Additional File S1: Supplementary Table S2) and second overall only to *Medicago*. However, the space occupied by these piLoci—89.2 Mb (54.3%) of the genome—is around average with respect to the species surveyed.

The amount of intergenic space in the *D. pulex* genome is also moderate—intergenic iLoci account for 75.1 Mb (38.0%) of the *D. pulex* genome, compared to 35.5 Mb (24.7%) in *Drosophila* and 149.1 Mb (56.3%) in *Anopheles*. However, the abundance of piLoci punctuating the intergenic space results in an elevated number of (shorter) intergenic iLoci (13 052 in contrast to 3452 in *Drosophila* and 7726 in *Anopheles*) (see Table 1).

Claims regarding the relative compactness of the *Daphnia* genome, based primarily on average lengths of gene spans and introns, are not supported by our analysis (32). Intergenic iLocus lengths are on average shorter in *Daphnia* compared to *Drosophila* and *Anopheles* (Additional File S1: Supplementary Figure S8A, produced by Additional File S2: notebooks/make-SF5c-SF8.ipynb). We confirm that genes are on average shorter in *Daphnia* than in *Drosophila* (Additional File S1: Supplementary Figure S8B), despite a larger number of exons per gene (Additional File S1: Supplementary Figure S8C). However, this appears to be influenced more by reduced exon length rather than by reduced intron length, as originally claimed. Median exon length is substantially shorter in *Daphnia* (154 bp versus 248 bp in *Anopheles* and 289 bp in *Drosophila*, respectively; see Additional File S1: Supplementary Figure S8D). In contrast, median intron length of simple iLoci is almost indistinguishable between *Daphnia* and *Drosophila* (75 and 70 bp, respectively), and shorter than for *Anopheles* (91 bp) (see Additional File S1: Supplementary Figure S8E).

Further, although we observe consistently higher (ϕ, σ) values for *Daphnia* than for *Anopheles*, relative to *Drosophila* the values are consistently lower, reflective of a smaller fraction of tightly packed genes and a smaller proportion of the genome sequence occupied by such gene clusters (Additional File S1: Supplementary Figure S5C, produced by Additional File S2: notebooks/make-SF5c-SF8.ipynb). Thus, across multiple quantitative measures, *D. pulex* is characterized by a moderate level of genome compactness relative to other arthropods and eukaryotes in general.

Manacus vitellinus. A widespread effort to collect and sequence avian genomes was undertaken in 2014, spanning most orders of bird species, including 38 new genome assemblies (33). As a representative species, we chose the golden-collared manakin (*M. vitellinus*) with the latest NCBI assembly/annotation available from July 2019 (34).

The current genome assembly is still highly discontiguous given the large number of sequences and fragmented intergenic iLoci for *M. vitellinus* (Table 1). Relatively few protein-coding iLoci (13 289) occupy 44.6% of the genome space (Table 2), a value closer to the mammalian average than the average of other vertebrates (Additional File S1: Supplementary Table S2). Notable is the small number of single-exon protein-coding iLoci (Additional File S1: Supplementary Table S2).

The merged iLocus count (2156) and genome occupancy (11.5%) are considerably lower compared to human and mouse and also low relative to vertebrate averages (Additional File S1: Supplementary Table S3), while the proportion of gene-harboring iLoci containing only a single gene is large at 65.8% (Table 3). Correspondingly, the (ϕ, σ) statistics computed on merged iLoci confirm a low level of genome compactness (see Additional File S1: Supplementary Figure S5D, produced by Additional File S2: notebooks/make-SF5d.ipynb).

More complete sequencing and annotation would seem necessary in order to distinguish avian-specific genome organization from effects of scope and approach by the avian genome sequencing effort (33), as the currently available assemblies contain multiple long, isolated gene structure models, sometimes even spanning an entire assembly scaffold (suggestive of incomplete presumed intergenic space sequencing).

iLoci provide a robust representation of the genome

Improvements in genome assemblies come at the expense of disrupting the sequence-based coordinate system typically used for annotating the location of genome features. Parsing an annotated genome into iLoci provides an alternative representation of the genome that is robust to assembly and annotation updates. We illustrate this use case with two model organism examples: (i) comparing two annotation versions on the same *A. thaliana* assembly and (2) updated annotation on more complete genome assembly of the honey bee *A. mellifera* compared to the original assembly and an earlier community annotation. For (i), the 2005 TAIR6 release was the first annotation of the *A. thaliana* genome managed by The *Arabidopsis* Information Resource (TAIR) (35), while the 2010 TAIR10 release integrates TAIR's latest

improvements to both the reference genome assembly and annotation using EST data from Sanger platforms (36).

For both species, we computed iLoci for each assembly/annotation version and determined *iLocus stability* as described in the 'Materials and Methods' section (Additional File S1: Supplementary Table S6). Figure 5 and Additional File S1: Supplementary Table S7 provide a breakdown of conservation by iLocus type.

The most obvious observation is that in both case studies the numbers of iLoci are fairly stable, except for easily explained changes. Thus, for the TAIR6 to TAIR10 comparison, new developments in publicly available RNA-seq data and ncRNAs presented an opportunity to update the genome annotation, culminating in Araport11 (37), which has since been incorporated into the TAIR10 labeled annotation used here. As a result, we see a significant increase in ncRNA annotations and a modest increase in protein-coding genes (see Additional File S1: Supplementary Table S5). Improvements in protein-coding gene annotations can be credited to incorporation of augmented depth of RNA-seq data identifying novel transcript and splicing isoforms (37).

Figure 5 shows that very few siLoci are unique to TAIR6, indicating stability over many years of annotation updates (figure produced by Additional File S2: notebooks/make-F5-F6.ipynb). Nonconserved simple iLoci are mostly contained, i.e. embedded in longer iLoci in the current annotation. In contrast, nonconserved intergenic iLoci are mostly anchored; i.e. the original iLocus annotation was mapped to a shorter new iLocus. TAIR6-unique gene models not transferred to TAIR10 tend to be short (Figure 6).

For *A. mellifera*, the Honey Bee Genome Sequencing Consortium's assembly version Amel_2.0 and Official Gene Set 1 (OGSv1.0) were preliminary data resources in use prior to the initially published description of the honey bee genome in 2006 (6), while assembly Amel_4.5 (corresponding to NCBI release 102) and OGSv3.2 represent the consortium's latest improvements to the genome and corresponding annotation as of 2014 (7,8). Release 103, still labeled Amel_4.5 (38), features some small differences from release 102, such as a slight increase in the number of protein-coding genes, likely a result of newer gene annotation software. Release 104 (HAv3.1) is NCBI's latest genome entry for *A. mellifera*, describing a new assembly derived from novel DNA sequencing technologies and, consequently, updated and revised annotations compared to Amel_4.5. Unlike in the previous case, both annotations for *A. mellifera* were performed by the NCBI Eukaryotic Genome Annotation Pipeline, an automated pipeline for gene annotation, as part of (30).

A large number of annotation 4.5 simple iLoci are unmapped to assembly/annotation HAv3.1 (Figure 5). Figure 6 shows that these are largely shorter gene structure models (and thus probably explained by gene model prediction algorithm parameter choices).

The main insight from these case studies is that the majority of iLoci can be faithfully mapped from one assembly/annotation pair to another. Practically, this suggests that iLoci identifiers can be used as database keys that point to entries containing both gene information and genome context information that will remain largely stable as assembly and annotation gaps are being filled.

DISCUSSION

Within the context of annotating a new genome, iLoci provide a quick and convenient solution for leveraging genomes of related model organisms to establish baseline expectations about genome composition and organization for the organism of interest. Similarities to genomes of related species across a broad range of measures give one confidence in the quality of the genome assembly and annotation. In contrast, any stark differences should point to specific genomic features that warrant additional investigation to distinguish the effects of annotation from real differences in genome biology. Considerable effort has been devoted to making such comparisons as easy as possible: relevant software is freely available as open-source code, is engineered with a focus on resource efficiency (enabling it to run easily on laptop or desktop computers) and works with a small number of standard input files. In short, iLoci provide a 'common currency' for evaluating new data sets and re-evaluating previously published data sets alike.

Additional applications of iLoci in the annotation and analysis of novel genomes are numerous. Leveraging iLoci with strong support from expression and homology evidence to train species-specific gene prediction models can yield improvements in subsequent annotation efforts. The longest regions of the genome annotated as intergenic can yield insight into the proliferation of transposable and other repetitive elements and ncRNA genes, or alternatively characteristics of regions where annotation workflows fail to predict genes. The largest regions of high gene density, as represented by merged iLoci, provide an excellent starting place for investigating the clustering of functionally related genes, whereas merged iLoci containing two genes are candidates for genome-wide analysis of tandem gene duplication.

iLoci also facilitate analysis of genome organization at multiple scales. At the scale of whole chromosomes (or large fractions thereof), iLoci provide a well-defined measure of genome compactness that can be compared across annotations, assemblies and species. At a slightly smaller scale, iLoci can be leveraged to investigate large-scale changes in genome organization along the length of the chromosome, with possible interpretation in terms of transposon activity and other dynamic mechanisms of genome expansion and contraction. At the scale of individual genes, iLoci capture local aspects of genome organization, furnishing insight into gene spacing and orientation for specific genes of interest. Insight gained from analysis of genome organization at these various scales also lays a foundation for more detailed modeling of genome architecture, and perhaps even simulation of genome evolutionary dynamics. Simulating transposon activity, gene duplication and genome rearrangements at various rates and observing the effect these have on signatures of large-scale genome organization provided by iLoci could yield insight into the dominant mechanisms driving the evolution of genome architecture in particular species or clades of interest.

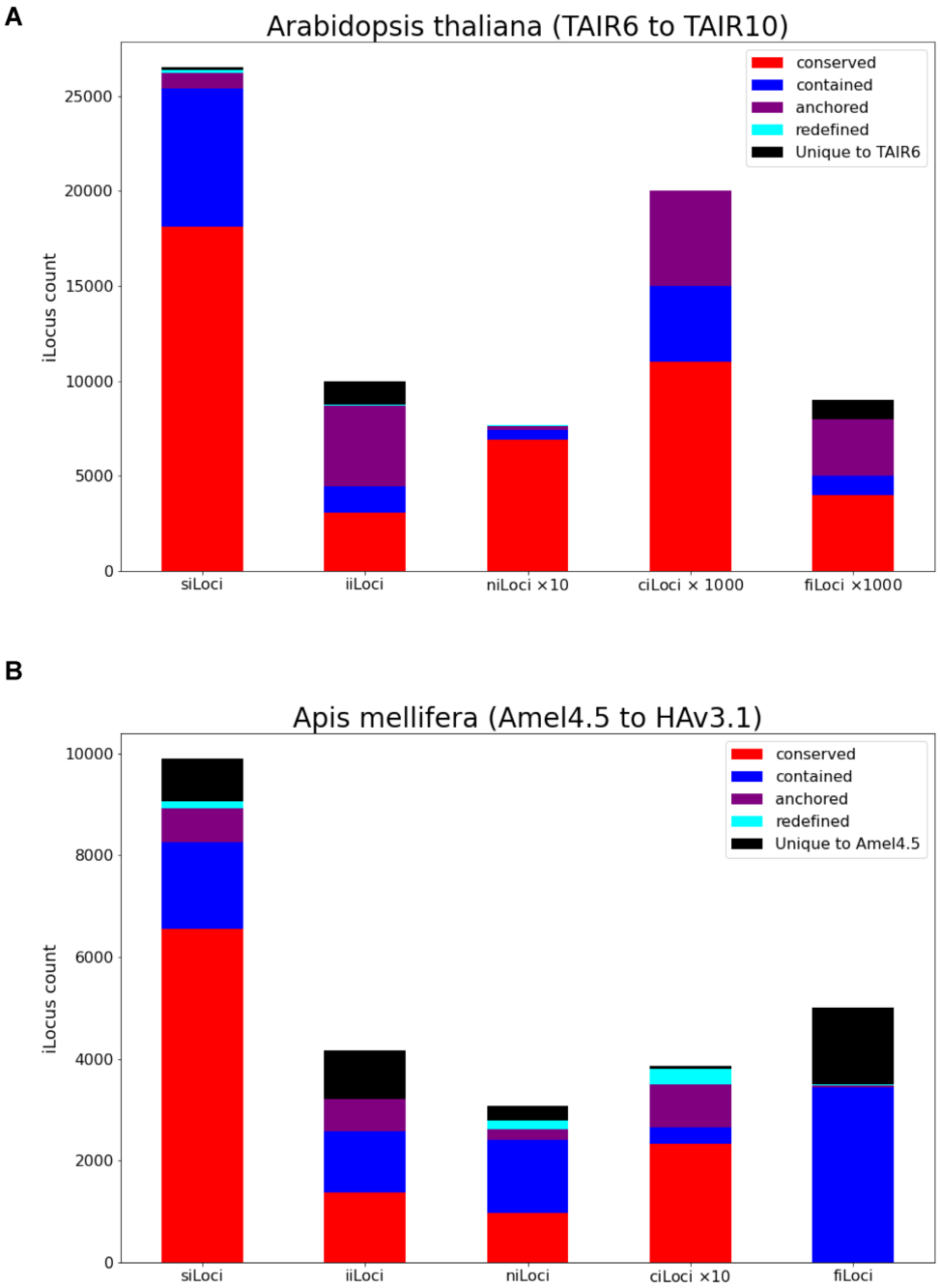

Figure 5. Breakdown of conservation per iLocus type. Note that the numbers of some iLocus types have been multiplied as indicated on the *x*-axis to allow visualization on the same plot.

CONCLUSIONS

Parsing annotated genome sequences into iLoci and then using these iLoci as a new coordinate system provides a robust and reproducible framework for investigating a variety of questions about genome content, architecture and evolution. iLocus annotation might include contextual information for gene models in the form of up- and downstream regulatory sequences. iLoci containing overlapping gene models can easily be identified for scrutiny seeking to distinguish gene model prediction errors from true compact gene organization that would likely be missed if analysis were performed at the level of individual genes. iLoci also provide stability across different versions of an annotated genome assembly, preserving gene models or intergenic regions for which local genomic context remained invariant to assembly and annotation updates. Finally, iLoci provide a way to break down the entire genome into distinct blocks that can be filtered based on their composition, gene content, conservation or a variety of other characteristics of interest, thus providing finely tuned data sets for analyses or training and testing of predictive models.

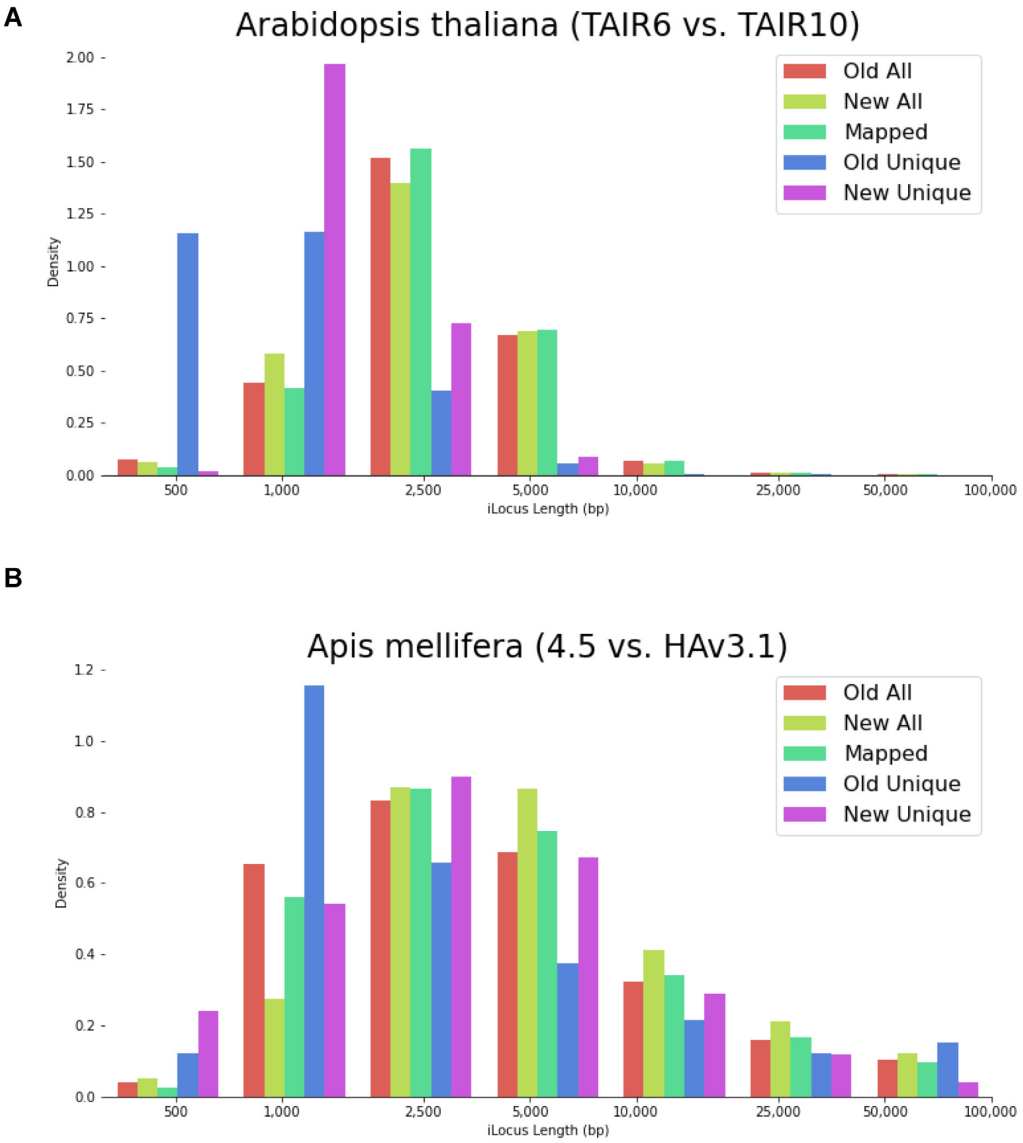

Figure 6. Breakdown of conservation per iLocus length bin.

SUPPLEMENTARY DATA

Supplementary data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors are grateful for use of the Extreme Science and Engineering Discovery Environment (XSEDE) Jetstream resource at Indiana University and the Texas Advanced Computing Center through allocation TG-BIO160012 (Computational Genomics) to V.B. XSEDE is supported by National Science Foundation grant number ACI-1548562. The authors would like to thank their colleagues Robert Policastro and Haixu Tang for helpful comments on an earlier version of the manuscript as well as an anonymous reviewer for very constructive evaluation.
Author contributions: D.S.S. implemented the AEGeAn Toolkit, contributed to the initial study design, performed the initial studies of genome content, genome compactness and iLocus stability, wrote early drafts of the manuscript and edited the final version of the manuscript. T.L. updated and expanded the original data analysis and Python scripts. In addition, he designed and implemented the procedure to retrieve and analyze genome compactness characteristics for large sets of branch-specific genomes. V.P.B. conceived the concept of iLoci, designed the study, finalized the code and code repositories, and wrote drafts and most of the final version of the manuscript.

FUNDING

No external funding.
Conflict of interest statement. None declared.

REFERENCES

1. Sequence Read Archive (2022) http://www.ncbi.nlm.nih.gov/sra, (11 February 2022, date last accessed).

2. Stanke,M., Diekhans,M., Baertsch,R. and Haussler,D. (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics*, **24**, 637–644.
3. Campbell,M.S., Law,M., Holt,C., Stein,J.C., Moghe,G.D., Hufnagel,D.E., Lei,J., Achawanantakun,R., Jiao,D., Lawrence,C.J. *et al.* (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.*, **164**, 513–524.
4. Hoff,K.J., Lange,S., Lomsadze,A., Borodovsky,M. and Stanke,M. (2016) BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.
5. Souvorov,A., Kapustin,Y., Kiryutin,B., Chetvernin,V., Tatusova,T. and Lipman,D. (2010) Gnomon—NCBI eukaryotic gene prediction tool. http://www.ncbi.nlm.nih.gov/RefSeq/Gnomon-description.pdf, (11 February 2022, date last accessed).
6. Elsik,C., Mackey,A., Reese,J., Milshina,N., Roos,D. and Weinstock,G. (2007) Creating a honey bee consensus gene set. *Genome Biol.*, **8**, R13.
7. Elsik,C., Worley,K., Bennett,A., Beye,M., Camara,F., Childers,C., de Graaf,D., Debyser,G., Deng,J., Devreese,B. *et al.* (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics*, **15**, 86.
8. NCBI *Apis mellifera* Annotation Release 102 (2022) http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Apis_mellifera/102/, (11 February 2022, date last accessed).
9. Wallberg,A., Bunikis,I., Pettersson,O.V., Mosbech,M.-B., Childers,A.K., Evans,J.D., Mikheyev,A.S., Robertson,H.M., Robinson,G.E. and Webster,M.T. (2019) A hybrid *de novo* genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*, **20**, 275.
10. Standage,D.S. (2022) The AEGeAn Toolkit: analysis and evaluation of genome annotations. http://brendelgroup.github.io/AEGeAn/, (11 February 2022, date last accessed).
11. Standage,D. and Brendel,V. (2012) ParsEval: parallel comparison and analysis of gene structure annotations. *BMC Bioinformatics*, **13**, 187.
12. Riley,M.C., Clare,A. and King,R.D. (2007) Locational distribution of gene functional classes in *Arabidopsis thaliana*. *BMC Bioinformatics*, **8**, 112.
13. Koonin,E.V. (2009) Evolution of genome architecture. *Int. J. Biochem. Cell Biol.*, **41**, 298–306.
14. NCBI Genome (2022) http://www.ncbi.nlm.nih.gov/genome/, (11 February 2022, date last accessed).
15. Gremme,G., Steinbiss,S. and Kurtz,S. (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **10**, 645–656.
16. GenomeTools website (2022) http://genometools.org/, (11 February 2022, date last accessed).
17. Stein,L. (2022) GFF3 specification. The Sequence Ontology Project. http://www.sequenceontology.org/gff3.shtml, (11 February 2022, date last accessed).
18. Wilbrandt,J., Misof,B., Panfilio,K.A. and Niehuis,O. (2019) Repertoire-wide gene structure analyses: a case study comparing automatically predicted and manually annotated gene models. *BMC Genomics*, **20**, 753.
19. Eilbeck,K., Moore,B., Holt,C. and Yandell,M. (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, **10**, 67.
20. GAEVAL: A Tool for Gene Annotation Evaluation (2022) http://www.plantgdb.org/GAEVAL/docs/index.html, (11 February 2022, date last accessed).
21. Franck,E., Hulsen,T., Huynen,M.A., de Jong,W.W., Lubsen,N.H. and Madsen,O. (2008) Evolution of closely linked gene pairs in vertebrate genomes. *Mol. Biol. Evol.*, **25**, 1909–1921.
22. Harris,R. (2007) Improved pairwise alignment of genomic DNA. Ph.D. thesis, The Pennsylvania State University, https://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf, (11 February 2022, date last accessed).
23. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
24. Pascual-Anaya,J., D'Aniello,S., Kuratani,S. and Garcia-Fernàndez,J. (2013) Evolution of *Hox* gene clusters in deuterostomes. *BMC Dev. Biol.*, **13**, 26.
25. Yi,G., Sze,S.-H. and Thon,M.R. (2007) Identifying clusters of functionally related genes in genomes. *Bioinformatics*, **23**, 1053–1060.
26. Howe,K., Clark,M.D., Torroja,C.F., Torrance,J., Berthelot,C., Muffato,M., Collins,J.E., Humphray,S., McLaren,K., Matthews,L. *et al.* (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, **496**, 498–503.
27. Karlin,S. and Brendel,V. (1992) Chance and statistical significance in protein and DNA sequence analysis. *Science*, **257**, 39–49.
28. Merchant,S.S., Prochnik,S.E., Vallon,O., Harris,E.H., Karpowicz,S.J., Witman,G.B., Terry,A., Salamov,A., Fritz-Laylin,L.K., Maréchal-Drouard,L. *et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.
29. Prochnik,S.E., Umen,J., Nedelcu,A.M., Hallmann,A., Miller,S.M., Nishii,I., Ferris,P., Kuo,A., Mitros,T., Fritz-Laylin,L.K. *et al.* (2010) Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science*, **329**, 223–226.
30. RefSeq: NCBI Reference Sequence Database (2022) http://www.ncbi.nlm.nih.gov/refseq/, (11 February 2022, date last accessed).
31. Standage,D.S., Berens,A.J., Glastad,K.M., Severin,A.J., Brendel,V.P. and Toth,A.L. (2016) Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. *Mol. Ecol.*, **25**, 1769–1784.
32. Colbourne,J.K., Pfrender,M.E., Gilbert,D., Thomas,W.K., Tucker,A., Oakley,T.H., Tokishita,S., Aerts,A., Arnold,G.J., Basu,M.K. *et al.* (2011) The ecoresponsive genome of *Daphnia pulex*. *Science*, **331**, 555–561.
33. Zhang,G., Li,B., Li,C., Gilbert,M., Jarvis,E. and Wang,J. (2014) Comparative genomic data of the Avian Phylogenomics Project. *Gigascience*, **3**, 789–804.
34. NCBI Manacus vitellinus Annotation Release 103 (2022) https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Manacus_vitellinus/103/, (11 February 2022, date last accessed).
35. The *Arabidopsis* Information Resource (2022) http://www.arabidopsis.org, (11 February 2022, date last accessed).
36. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
37. Cheng,C.-Y., Krishnakumar,V., Chan,A.P., Thibaud-Nissen,F., Schobel,S. and Town,C.D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.*, **89**, 789–804.
38. NCBI Apis mellifera Annotation Release 103 (2022) https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Apis_mellifera/103/, (11 February 2022, date last accessed).