Similar Ratios of Introns to Intergenic Sequence across Animal Genomes

Warren R. Francis[1,*] and Gert Wörheide[1,2,3]

[1]Department of Earth and Environmental Sciences, Paleontology and Geobiology, Ludwig-Maximilians-Universität München, Munich, Germany

[2]GeoBio-Center, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Bavarian State Collection for Paleontology and Geology, Munich, Germany

*Corresponding author: E-mail: woerheide@lmu.de.

Accepted: June 7, 2017

Abstract

One central goal of genome biology is to understand how the usage of the genome differs between organisms. Our knowledge of genome composition, needed for downstream inferences, is critically dependent on gene annotations, yet problems associated with gene annotation and assembly errors are usually ignored in comparative genomics. Here, we analyze the genomes of 68 species across 12 animal phyla and some single-cell eukaryotes for general trends in genome composition and transcription, taking into account problems of gene annotation. We show that, regardless of genome size, the ratio of introns to intergenic sequence is comparable across essentially all animals, with nearly all deviations dominated by increased intergenic sequence. Genomes of model organisms have ratios much closer to 1:1, suggesting that the majority of published genomes of nonmodel organisms are under-annotated and consequently omit substantial numbers of genes, with likely negative impact on evolutionary interpretations. Finally, our results also indicate that most animals transcribe half or more of their genomes arguing against differences in genome usage between animal groups, and also suggesting that the transcribed portion is more dependent on genome size than previously thought.

Key words: metazoa, comparative genomics, junk DNA, complexity, *C*-value.

Introduction

Understanding why genomes vary greatly in size and how organisms make different use their genomes have been central questions in biology for decades (Thomas 1971). For many bacteria, the majority of the genome is composed of relatively short genes, averaging ~1,000 bp, and coding for proteins. Indeed, the largest bacterial genome (a myxobacterium) that has been sequenced is only 14 megabases, containing an estimated 11,500 genes (Han et al. 2013). However, for eukaryotic organisms, genomes can be over 10,000-fold larger than bacterial genomes due to an increase in the number of genes (tens of thousands compared with a few thousand in most bacteria), expansion of the genes themselves due to the addition of introns, and expansion of the sequence between genes.

As the number of genome projects has grown, massive amounts of data have become available to study how organisms organize and use their genomes. Genome projects vary substantially in quality of assembly and annotation (Guigó et al. 2006; Brent 2008). Unfortunately, the predicted genes are often taken for granted as being correct when these are only hypotheses of gene structure (Vallender 2009). For example, one study found that almost half of the genes in the *Rhesus* monkey genome had a predictable annotation error when compared with the closest human homolog (Zhang et al. 2012). This has profound implications for all downstream analyses, such as studying evolution of orthologous proteins (Altenhoff et al. 2016) and phylogeny based on protein matrices or gene content (Ryan et al. 2013; Pisani et al. 2015). When considered across all genes, systematic errors in genome assembly or annotation would severely skew bulk parameters of a genome.

While issues of assembly are often thought to be technical problems that are resolved before continuing, all subsequent analyses are dependent upon accurate genome assembly and annotation. The absence of a protein family in a particular organism is only meaningful if it is certain that it is absent from the genome and not merely the annotation, therefore it is of utmost importance that all genes are properly represented. Yet

for most genome projects of nonmodel organisms, there are limited methods to determine if the assembly and annotation are sufficient for downstream comparative analyses. Internal metrics can be used, such as the fraction of raw genomic reads or ESTs that map back to the assembly, though this does not tell us if a gene is believable in the context of other animals. Alternatively, counts of "universal" single-copy orthologs have been proposed as a metric of genome completeness (Parra et al. 2007; Simão et al. 2015), though these genes only represent a small subset of all genes (few hundred out of tens of thousands in most animals).

Identification of universal trends in genome organization and transcription may enable better quantitative metrics of genome completeness. Mechanistic models relating to evolution of gene content or coding fractions tended to focus on bacteria or archaea because of the relative ease of annotation. In regards to eukaroytes, some patterns in genome size have been discussed (Lynch and Conery 2003; Daubin and Moran 2004; Lynch 2004). Additionally, a handful of studies have analyzed genome size in connection to other parameters such as indels (Pettersson et al. 2009), transposon content (Kidwell 2002; Elliott and Gregory 2015a; Canapa et al. 2016), average intron length (Deutsch and Long 1999; Zhu et al. 2009) or total intron length (Elliott and Gregory 2015b). Despite these advances, none of these studies have estimated the amount of the genome that is genic (exonic plus intronic, including noncoding) based on independent examination of single genomes and without averaging over a whole kingdom. Additionally, none of them have described a way to account for technical problems in assembly and annotation.

Here, we examine basic trends of genome size and the relationship to annotation quality across animals and some single-celled eukaryotes. We show that assembly and annotation errors are widespread and predictable and that many genomes are likely to be missing many genes. We further show that re-annotation of select species with publicly available tools and transcriptome data improves the annotation. Future users may benefit if databases incorporate more recent data from transcriptome sequencing, and update annotation versions more frequently. Comparison of genomic composition across many animal groups indicated a ratio of introns:intergenic approaching 1:1, suggesting this as a potential parameter to identify genome completeness across metazoans, and potentially other eukaryotes. Finally, this implies that animals transcribe at least half of their genomes whereby small, exon-rich genomes transcribe most of the genome and large genomes transcribe approximately half of the genome.

Materials and Methods

Genomic Data Sources

Data sources and parameters are available in supplementary table S1, Supplementary Material online.

Genomic scaffolds and annotations for *Ciona intestinalis* (Dehal et al. 2002), *Branchiostoma floridae* (Putnam et al. 2008), *Trichoplax adherens* (Srivastava et al. 2008), *Capitella teleta* (Simakov et al. 2013), *Lottia gigantea* (Simakov et al. 2013), *Helobdella robusta* (Simakov et al. 2013), *Saccoglossus kowalevskii* (Simakov et al. 2015), *Monosiga brevicollis* (King et al. 2008), *Emiliania huxleyi* (Read et al. 2013), and *Volvox carteri* (Prochnik et al. 2010) were downloaded from the JGI genome portal.

Genome assemblies and annotations for *Sphaeroforma arctica*, *Capsaspora owczarzaki* (Suga et al. 2013) and *Salpingoeca rosetta* (Fairclough et al. 2013) were downloaded from the Broad Institute.

GFF annotations v2.1 (Fernandez-Valverde et al. 2015) for *Amphimedon queenslandica* were downloaded from the Amphimedon Genome website (http://amphimedon.qcloud.qcif.edu.au/downloads.html), and v1 annotations (Srivastava et al. 2010) and assemblies were downloaded from Ensembl.

For *Nematostella vectensis*, Nemve1 assembly and annotations (Putnam et al. 2007) were downloaded from JGI, and the transcriptome for comparative reannotation was downloaded from http://www.cnidariangenomes.org/ (Moran et al. 2014).

Genome assembly, transcriptome assemblies from Cufflinks and Trinity, and GFF annotations for *Mnemiopsis leidyi* (Ryan et al. 2013) were downloaded from the Mnemiopsis Genome Portal (http://research.nhgri.nih.gov/mnemiopsis/). Assembly and annotations for *Sycon ciliatum* (Fortunato et al. 2014) were downloaded from COMPAGEN. Assembly and annotation for *Botryllus schlosseri* (Voskoboynik et al. 2013) were downloaded from the Botryllus Schlloseri genome project (http://botryllus.stanford.edu/botryllusgenome/). Assembly and annotation for *Exaiptasia pallida* (formerly *Aiptasia sp.*) (Baumgarten et al. 2015) were downloaded from http://reefgenomics.org. Assembly and annotation for *Oikopleura dioica* (Denoeud et al. 2010) were downloaded from Genoscope (http://www.genoscope.cns.fr/externe/GenomeBrowser/Oikopleura/). Assembly and annotation for *Tetrahymena thermophila* were downloaded from the Tetrahymena Genome Database (ciliate.org). Assembly and annotation for *Symbiodinium kawagutii* (Lin et al. 2015) were downloaded from the Dinoflagellate Resources page (web.malab.cn/symka_new/index.jsp).

Assemblies and annotations for *Symbiodinium minutum* (Shoguchi et al. 2013), *Pinctada fucata* (Takeuchi et al. 2012), *Acropora digitifera* (Shinzato et al. 2011), *Lingula anatina* (Luo et al. 2015), *Ptychodera flava* (Simakov et al. 2015), and *Octopus bimaculoides* (Albertin et al. 2015) were downloaded from the OIST Marine Genomics Browser (http://marinegenomics.oist.jp/gallery/).

Builds of *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Canis lupus* (Kirkness 2003), *Monodelphis domestica* (Mikkelsen et al. 2007), *Ornithorhynchus anatinus* (Warren et al. 2008), *Xenopus tropicalis* (Hellsten et al. 2010),

Struthio camelus (Zhang et al. 2014), *Gallus gallus*, *Taeniopygia guttata* (Warren et al. 2010), *Aptenodytes forsteri* (Zhang et al. 2014), *Anas platyrhynchos* (Huang et al. 2013), *Melopsittacus undulatus* (Ganapathy et al. 2014), *Alligator mississippiensis* (Green et al. 2014), *Anolis carolinensis* (Alföldi et al. 2011), *Chrysemys picta bellii* (Shaffer et al. 2013), *Chelonia mydas* (Wang et al. 2013), *Pelodiscus sinensis* (Wang et al. 2013), *Python bivittatus* (Koning et al. 2013), *Salmo salar*, *Danio rerio* (Howe et al. 2013), *Latimeria chalumnae* (Amemiya et al. 2013), *Petromyzon marinus* (Smith et al. 2013), *Callorhinchus milii* (Venkatesh et al. 2014), *Crassostrea gigas* (Zhang et al. 2012), *Dendroctonus ponderosae* (Keeling et al. 2013), *Tribolium castaneum* (Richards et al. 2008), *Bombyx mori* (Mita et al. 2004), *Limulus polyphemus* (Nossa et al. 2014) were downloaded from the NCBI Genome server.

Genome assemblies and annotations of *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), *Drosophila melanogaster*, *Strongylocentrotus purpuratus* (Sodergren et al. 2006), *Daphnia pulex* (Colbourne et al. 2011), *Apis mellifera* (Weinstock et al. 2006), *Ixodes scapularis* (Gulia-Nuss et al. 2016), *Strigamia maritima* (Chipman et al. 2014) were downloaded from Ensembl.

Calculation of Exonic and Genic Sequence

For all analyses, we used the total number of bases in the downloaded assembly as the total genome size, bearing in mind that this may result in a systematic underestimation of total genome size as repeated regions may be omitted from assemblies. For example, the horseshoe crab *L. polyphemus* has a scaffold assembly of 1.8 Gb while the reported genome size is 2.7 Gb (Nossa et al. 2014), a difference of almost a gigabase.

If GFF format files were available for download with a genome project, or on databases (Ensembl or NCBI), those were used preferentially. The analysis procedure is described in figure 1. Total base pairs of exon, intron, intergenic, and gaps were counted from each GFF file and genomic contigs (or scaffolds) with a custom Python script (gtfstats.py, available at bitbucket.org/wrf/sequences). For calculations of exonic or genic bases, the script converts all gene and exon annotations to intervals and ignores the strand. Here, gene (or genic) is defined as transcribed bases that are either exon or intron, regardless of coding potential. All overlapping exon intervals are merged, meaning that alternative splice sites, or exons on the opposite strand, are treated as a single interval for bulk calculations. The same is done for genes or transcripts, whichever is available. Introns are calculated as the difference of the genic set and the exonic set, as introns are typically not defined as separate features in normal GFF files. This means that any sequence that is an exon on one strand and an intron on the other is treated for these calculations as an exon, meaning those base or their reverse complement (hence base pairs) are transcribed and retained following splicing in some case (fig. 1*D* and *E*). Intergenic sequence is defined as the difference between total sequence base pairs and genic base pairs, and gaps are defined as any repeats of 'N's longer than one base.

If exons are not specified, then coding sequences (CDS) are used instead if they are available, such as for AUGUSTUS predictions. Additional noncoding features such as "microRNA," "tRNA," "ncRNA" are included for gene and exon calculations if they were in the standard GFF3 format. Some genomes made use of mapped RNAseq data, which implicitly included all noncoding RNAs as well. Some annotations had to determine the gene ID from the exons. For example, most of the GTF files from the earlier JGI genomes had only exons annotated, without individual features for genes or mRNAs, so the gene was then defined as all of the exons with the same feature ID even though a specific gene feature was undefined.

Exons defined as part of a "pseudogene," or genes defined as pseudogenes, were also excluded from all counts. We justify this because pseudogenes are subject to problems of definitions and population sampling bias. Pseudogenes are defined as having the appearance or structure of normal protein coding genes, independent of transcriptional potential, but that would be unable to produce a functional protein, perhaps through nonsense mutations. Therefore, a pseudogene that is transcribed and cannot code for a protein should be annotated as a "transcribed pseudogene," though potentially could be a noncoding RNA. Pseudogene features are not annotated for all species, making it difficult to compare broadly. Additionally, for most nonmodel species, the genomes are generally based upon a single individual rather than a reference for a population based on a large number of individuals. Therefore, if that single individual was homozygous for a nonsense mutation but other individuals in the population were not, that gene should not be a pseudogene.

All downstream correlation calculations and graphs were done in R. Regression was calculated using the "lm()" function, for linear ($y \sim x$), exponential ($\log(y) \sim x$), or hyperbolic ($y \sim 1/x$) models, and the "predict()" function was used to model curves. The raw data table and the R source code used to generate figures are available at bitbucket.org/wrf/genome-reannotations.

Calculation of Average Exon and Intron Length

The same script (gtfstats.py, available at bitbucket.org/wrf/sequences) also calculated the average exon and intron length, though these were analyzed separately. All nonredundant exons for all splice variants were taken into account for determination of averages. Unlike the total base pair calculations, genes are separated by strand. Identical exons of splice variants were treated as one exon and counted once, however, alternative boundaries were treated as a separate exons. Retained introns are treated as exons, not introns. Exon

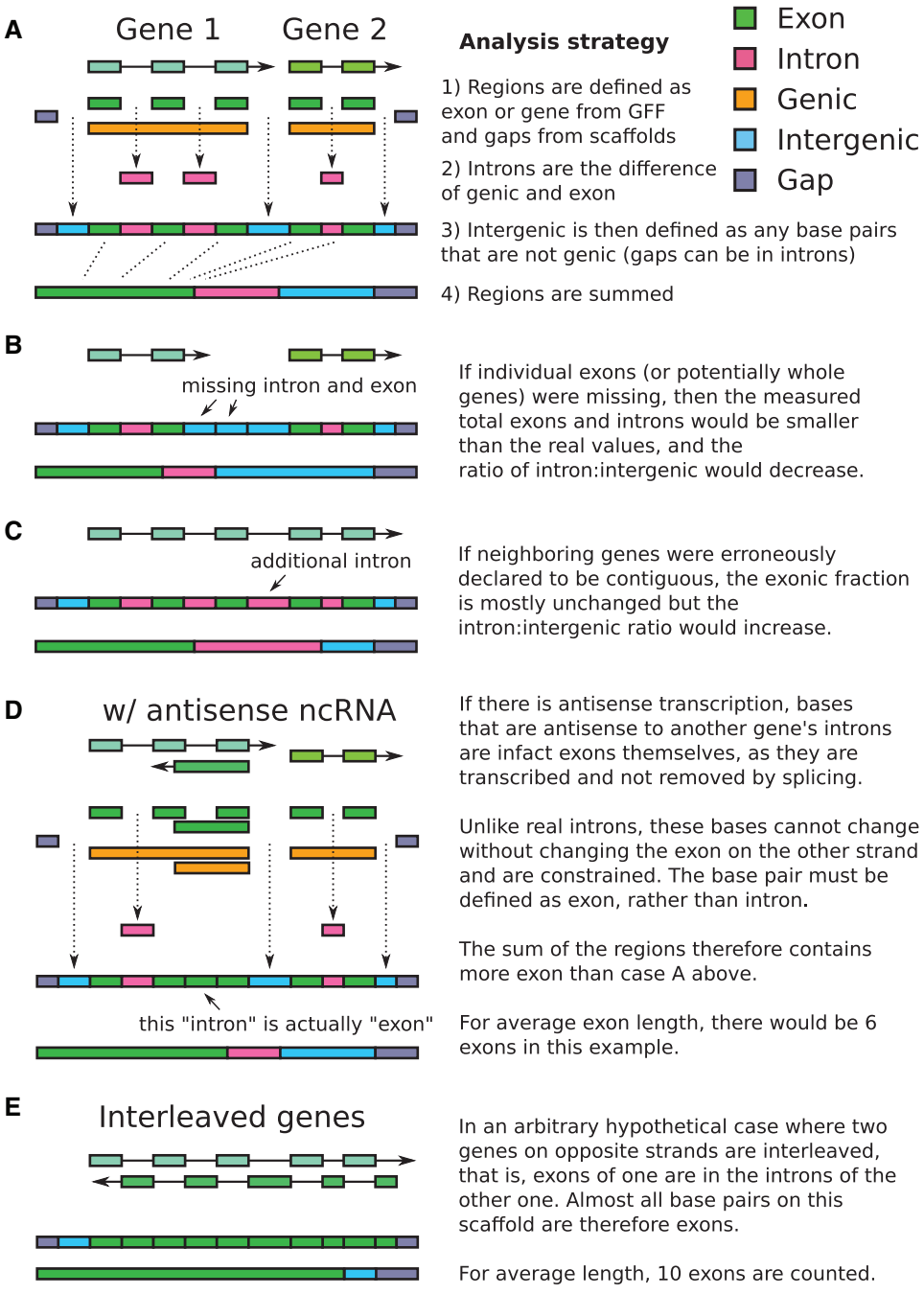

Fig. 1.—Schematic of analysis, misannotations and the effects on coding fraction. (*A*) In a normal case, two hypothetical genes on the same strand are identified. The exons and introns are defined, and the total lengths of those features are summed and displayed in the bars below. Because real genome assemblies can often contain gaps, sample gaps are also shown at the edges of the segment. (*B*) Case of missing exon or gene annotations, where the intron:intergenic decreases. (*C*) Case of falsely fused genes, where the intron:intergenic ratio would increase. (*D*) Case of antisense transcription, where base pairs that are intron on the sense strand and exon on the antisense strand are necessarily defined as exon. (*E*) Any arbitrary, interleaved genes, or any exons inside of introns, must as well be counted as exon.

lengths were counted per nonredundant exon for each gene, summed across all genes and divided by the number of nonredundant exons across all genes. The sum of exon lengths for the average length calculation does include redundant bases from antisense transcripts or splice variants, meaning bases from antisense transcripts and alternative-boundary splice variants can be double-counted. Introns were calculated as the space between exons, calculated by gene.

Reannotation of Select Species

Due to unexpectedly high or low gene content, six genomes were selected for reannotation.

The original Triad1 scaffolds of *T. adherens* (Srivastava et al. 2008) were reannotated with AUGUSTUS v3.0.3 (Stanke et al. 2008) with the following options: -strand = both –genemodel = atleastone –sample = 100 –keep_viterbi = true –alternatives-from-sampling = true –minexonintronprob = 0.2 –minmeanexonintronprob = 0.5 –maxtracks = 2. Species training was generated using the Triad1 ESTs with the webAugustus Training server (Hoff and Stanke 2013).

The original Monbr1 scaffolds of *M. brevicollis* (King et al. 2008) were reannotated with AUGUSTUS as for *T. adherens*, using the same parameters except trained using the Monbr1 ESTs with the webAugustus Training server (Hoff and Stanke 2013).

For the hydrozoan *H. magnipapillata*, the original assembly was downloaded from JGI (Chapman et al. 2010) and a new scaffold assembly was downloaded from the FTP of Rob Steele at UC Irvine (at https://webfiles.uci.edu/resteele/public). For both cases, the scaffolds were reannotated using TopHat22 v2.0.13 (Kim et al. 2013) and StringTie v1.0.4 (Pertea et al. 2015) with default options by mapping the reads from two paired-end RNAseq libraries, NCBI Short Read Archive accessions SRR922615 and SRR1024340, derived from whole adult animals.

For the lancelet *B. floridae*, the Brafl1 scaffolds (Putnam et al. 2008) were reannotated using TopHat22 v2.0.13 (Kim et al. 2013) and StringTie v1.0.4 (Pertea et al. 2015) with default options by mapping the reads from the paired-end RNAseq library, NCBI SRA accession SRR923751, from the adult body.

For the lamprey *P. marinus*, we were unable to find any annotation as GFF or GTF, so we generated one using TopHat2 v2.0.13 (Kim et al. 2013) and StringTie v1.0.4 (Pertea et al. 2015) based on the Pmarinus-v7 scaffolds from NCBI and the 16 single-end Illumina libraries from NCBI BioProject PRJNA50489.

For the octopus *O. bimaculoides*, scaffolds were downloaded from the OIST Marine Genomics platform (Albertin et al. 2015), and were reannotated using TopHat2 v2.0.13 (Kim et al. 2013) and StringTie v1.0.4 (Pertea et al. 2015) with default options by mapping 19 paired-end RNAseq libraries from NCBI BioProject PRJNA285380.

All reannotations are available for download as GTF or GFF files (see https://bitbucket.org/wrf/genome-reannotations/downloads).

Results

Overview and Organization of Data

A total of 68 genomes were analyzed, with 59 selected across all major metazoan groups and nine genomes of single-celled eukaryotes. For each group, only select species were taken to avoid having a single group dominate the analysis. For example, over 100 mammalian genomes are available though only six were used including three model organisms (human, mouse, dog), opossum and platypus (for the non-Eutherian clades, marsupial and monotreme, respectively) and the chimp, to compare directly to the human annotation. In general, parasites were excluded because they often have unusual biology, such as the single-celled eukaryote *T. brucei*, which is known for its unusual RNA processing (Siegel et al. 2010; Preußer et al. 2012).

Generally, we refer to small and large genomes as those below and above 500 Mb, respectively. The smallest animal genome used in this study is that of the larvacean *Oikopleura dioica* (70 Mb), while the largest is that of the opossum *Monodelphis domestica* (3,598 Mb). This range incorporates an existing selection bias, as some of the public genome sequencing projects selected the animal of their clade based on their known small genomes. Two examples of this are the shark *C. milii* and the pufferfish *T. rubripes*. Yet it must be considered that in terms of genomes, they may not be representative of their clades; many other shark genomes are estimated to be over 10 Gb (haploid genome size) (Hardie and Hebert 2004), such that a shark genome of only 1 Gb may not be "normal" for sharks.

Additionally, not all of the species in the sample were sequenced or annotated with the same method, making direct comparison more challenging. For instance, some of the earlier genomes (such as *Branchiostoma floridae* and *Trichoplax adherens*) were annotated only with Sanger ESTs (order of tens of Mb), which were used to train gene prediction algorithms. Because not all genes have features easily captured by the EST training, several different results are expected: some genes are split because internal exons are not properly found or may have misassemblies in the draft genomes; adjacent genes on the same strand are fused; or genes are omitted entirely.

Connection between Annotation and Understanding of Genomes

Genome projects of nonmodel species usually report protein coding regions of a genome. Broadly, there are two methods of doing this, comparison to other proteins from other genomes and by aligning mRNA from ESTs or RNAseq (Brent 2008). In practice, improvements in methods have made it relatively easy to directly predict proteins from the genome sequence. However, untranslated regions (UTRs) are difficult to predict and often require evidence from ESTs or transcriptome sequencing for accurate predictions, and this has implications for our measurements of total exons in each genome. This means that even in a "perfect" genome where all coding genes are correctly predicted by an annotation program (perhaps based on similarity to a related species) that the

precise positions and amount of UTR may still be unknown, resulting in an underestimation of the amount of exonic sequence (fig. 1*A* and *B*). Because of this, the reliance on coding genes is likely to underestimate the usable fraction of the genome.

To illustrate this, one may consider a hypothetical eukaryotic genome of 60 Mb with 10,000 genes and equal fractions of exons, introns, and intergenic sequence, at 20 Mb each. For simplicity, all exons are the same size (in this example, 200 bp), so an average gene (with 10-exons) may contain one exon for the 5′-UTR, and one for the 3′-UTR, and the remaining eight exons are coding. Based on the above annotation scheme, 20% of the exonic fraction (those containing the 5′- and 3′-UTRs) is missing in the final annotation. Two introns per gene are also missing (the first and last introns), ~18% of the intronic fraction. This would yield a final annotation where exons are predicted as 16 Mb (26.6% of the genome) and introns as 15.5 Mb (25.9% of the genome). This would also indicate that 52.6% of the genome is genes, a substantial underestimation from the actual value of 66.6%.

However, other systematic errors can result in an overestimation of the genic fraction. If we consider multiple genes on the same strand, in a head-to-tail arrangement, and recall that UTRs are often not predicted, then an exon containing the stop codon with a 3′-UTR may be omitted and the predicted gene may continue into the next gene (fig. 1*C*). If it is assumed that the majority of coding exons are correctly predicted, then if such predictions were made systematically one may expect that the measured amount of exons does not deviate much from the true exonic fraction. However, because introns are defined as the removed sequence between exons of the same gene, then the sequence between the two genes that should have been defined as intergenic will instead be defined as intronic, thus raising the intron:intergenic ratio >1.

The above problems assume that the genomic assembly is nonetheless correct, yet the annotation is directly affected by assembly problems as well. Of the two main sources of problems, repeats (Treangen and Salzberg 2012) and heterozygosity (Takeuchi et al. 2012; Zhang et al. 2012; Kajitani et al. 2014; Simakov et al. 2015), repeats often result in breaks in the assembly that could split genes (fig. 2*A*). Genes that are split at contig boundaries are likely to have exons missing (or on other scaffolds) and thus the sequence that should be defined as introns would be instead defined as intergenic (fig. 2*B*).

For normal diploid genomes (wild strains, not inbred lab strains), heterozygosity is not uniform across the genome. Some regions are identical between the two haplotypes (hence are homozygous alleles or loci), while others may vary by SNPs, short indels, or copy numbers of repeats, exons, or even genes. For sequences that are identical between both haplotypes, the contigs are generally kept as is, while a more complex decision must be made for the heterozygous loci.

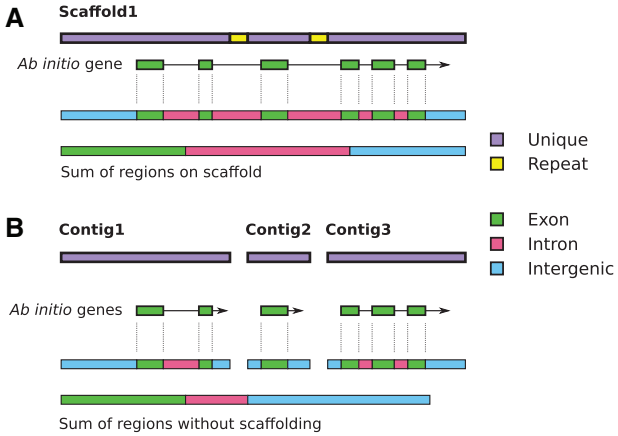

Fig. 2.—Schematic of the effects of scaffolding and repeats on genic fraction analyses. (*A*) For a hypothetical scaffold in a genome assembly, two identical repeats are found within introns. The gene is correctly predicted to span the two repeats and the regions are defined below as in figure 1. (*B*) For the case without scaffolding, or where the assembler breaks the assembly at repeats (or other high coverage regions), three contigs are generated. Note that the numbers are arbitrary, and in a real assembly they are unlikely to be in order. When annotated, all of the exons are correctly found, but the connections between them are missing for the single exon on Contig 2, resulting in a loss of intronic sequence. The final measured amount of exons is comparable, but the intron:intergenic ratio would decrease.

During normal genome assembly, the assembler evaluates the coverage at each "bubble" (where the de Brujin graph has two paths out of a node, and both paths merge again at the next node) and ultimately has to retain one of the paths at the exclusion of the other (fig. 3*A*) (also see schematics in Kajitani et al. 2014 and Bankevich et al. 2012). This merging is the essential process that creates the reference genome, even though that reference is an arbitrary merge of the two haplotypes. Therefore, it must be kept in mind that predicted genes or proteins in reference genomes may not be identical to either haplotype.

Regions with relatively high heterozygosity may fail to be merged in this way, leaving contigs of both haplotypes in the assembly (fig. 3*C*). During subsequent scaffolding steps, contigs of separate haplotypes can be fused head-to-tail if mate pairs are bridging the unique regions. Because this head-to-tail joining is an artifact, no reads should map at the junction point, resulting in a region of zero coverage at the junction and flanked by regions where coverage is half of the expected value (fig. 3*D*). One additional feature may reveal this artifact: exons in the unmerged sections may be individually annotated but mapped ESTs or de novo assembled transcripts may show a staggered exon pattern (fig. 3*E*) because transcripts can only map to one of the two possible exons (2a or 2b, 3a or 3b). This may increase the ratio of intron:intergenic sequence (fig. 3*F*), but also falsely indicate that splice variation is more prevalent for this gene.

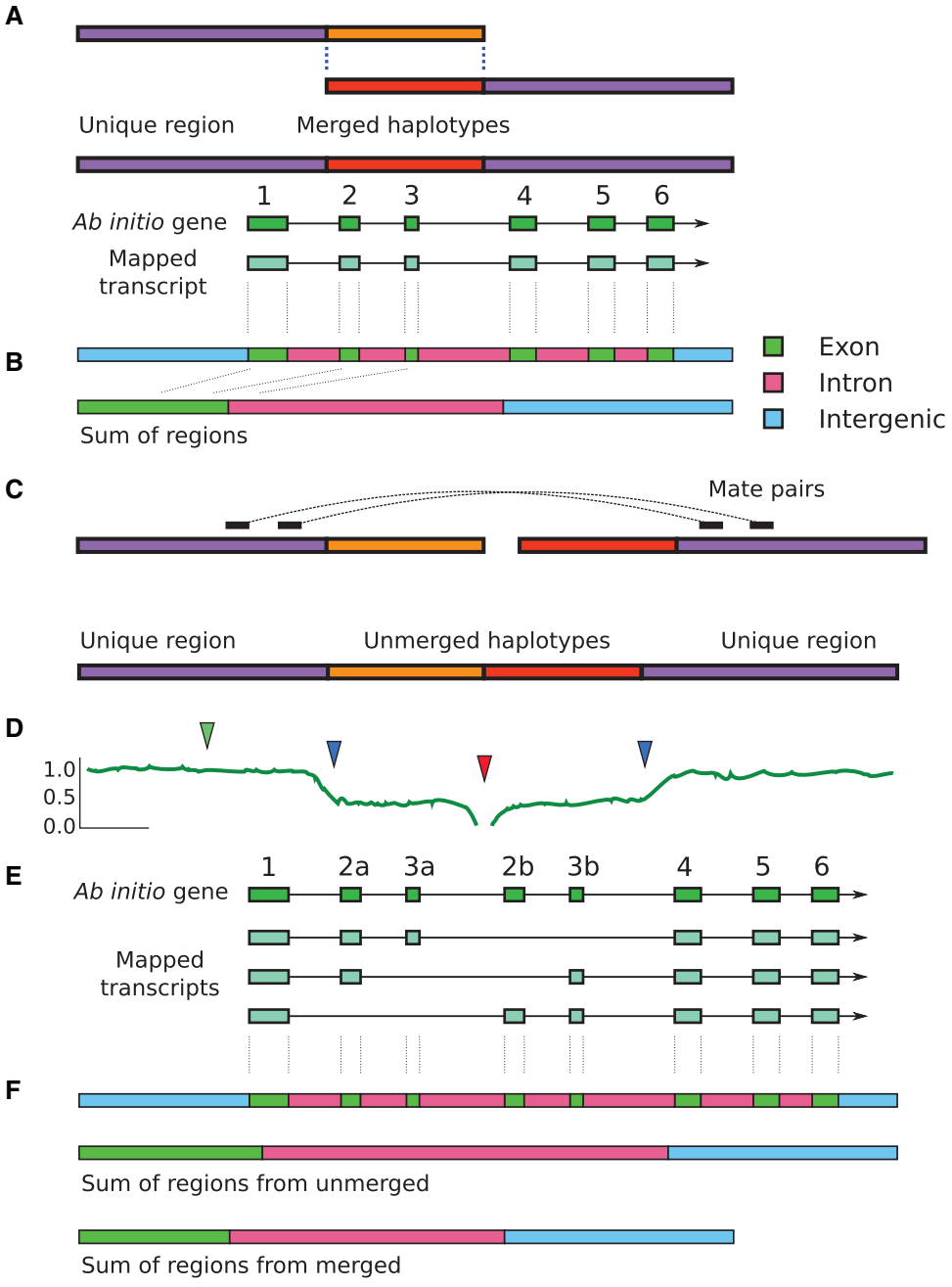

Fig. 3.—Schematic of misassembly and the effects on genic fraction analyses. (*A*) During assembly, regions that are heterozygous (differing by SNPs or indels) are combined to make a single reference contig. When genes are predicting that this locus, or when assembled transcripts are aligned to the genome, the correct exon structure is found. (*B*) Regions are defined as exon, intron, or intergenic, as in figure 1. (*C*) Reference genomes are a mix of the maternal and paternal haplotypes, but not uniformly. Rather than being merged into a single sequence, highly heterozygous regions may be assembled as different contigs that get erroneously fused during scaffolding steps. Mate pairs that bridge the two purple unique regions will instead result in a head-to-tail joining of the two unmerged haplotype sequences. (*D*) Hypothetical plot of read coverage across the contig. The green arrow shows a region of normal coverage (1×) while the blue arrows show sites where coverage is reduced because reads for each haplotype map separately. At the fusion point between the two haplotypes (red arrow), no reads will map since the sequence is an artifact, or is represented by a gap. (*E*) Mapped transcripts (or ESTs) or transcripts derived from mapped RNAseq reads (such as by Cufflinks or StringTie) may only be mapped to one of the two haplotypes, thereby producing a staggered exon structure. A mapped transcript can only align to either exon 2a or 2b, but not both, likewise for 3a or 3b, yet all other exons are unique and would align correctly. Genes predicted ab initio may annotate both sets of exons (2a/3a and 2b/3b), which may result in a duplication in some part of the protein, or a premature stop codon if 3a and 2b are out of phase. (*F*) For this hypothetical case, the sum of the regions would appear to have increased total exon size and the total intron size compared with the same genomic locus where the haplotypes were correctly merged.

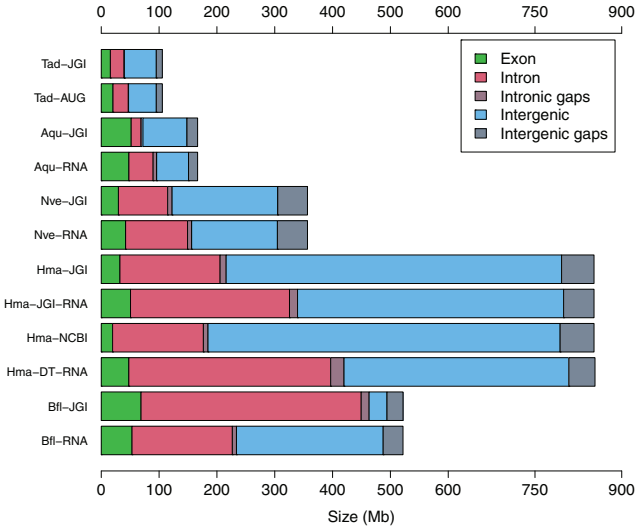

Fig. 4.—Proportions of exons, introns, and intergenic sequences. Barplot showing the summed proportions of genomes composed of exons (green), introns (red) and intergenic sequences (blue). The reannotation for *Octopus bimaculoides* was not shown for clarity, as this genome is substantially larger than the others. Abbreviations are as follows: Tad: *Trichoplax adherens*, Aqu: *Amphimedon queenslandica*, Nve: *Nematostella vectensis*, Hma: *Hydra magnipapillata*, Bfl: *Branchiostoma floridae*. JGI refers to the original annotations for each species downloaded from the JGI Genome Portal. RNA refers to reannotation (see Methods) with RNAseq. Hma-NCBI is the NCBI GNOMON annotation of *H. magnipapillata*. Hma-DT-RNA is the Dovetail reassembly of *H. magnipapillata* annotated with RNAseq. AUG is the reannotation using AUGUSTUS for *T. adherens*.

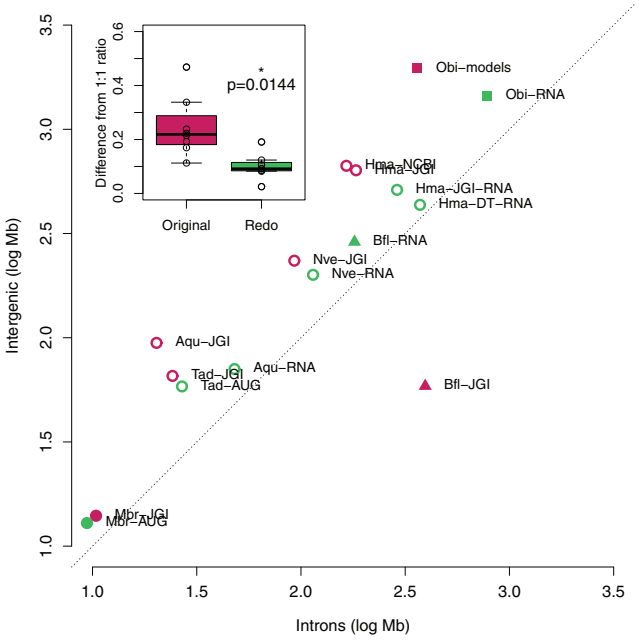

Fig. 5.—Improvements from reannotation. Log-scale plot of total intronic size versus total intergenic size where original annotations from the published genomes are shown in red and reannotations are shown in green. The dotted line shows a ratio of 1:1 as a reference. Abbreviations are as in figure 4, with the addition of Mbr: *Monosiga brevicollis* from the original JGI annotation and the redo with AUGUSTUS, and Obi: *Octopus bimaculoides* from the published gene models and the reannotation with Tophat/StringTie. The inset graph shows box plot of difference of the intron:intergenic ratio to 1, showing the reannotated genomes (green) are significantly closer than the original version (paired two-end *t*-test, *P* value: 0.0144).

Reannotation and Changes following RNAseq Reannotation

Keeping in mind the above error sources, some of the genomes used in our study had obvious problems of too much or too little genic content that would confound our analyses. For instance, the total amount of exons in the JGI annotation of *T. adherens* (Triad1) was only 14 Mb, over 2-fold lower than the related species, the placozoan *H. hongkongensis*, and thus it was expected to contain many more or longer genes than were present in the original Triad1 annotation. Because of this, we remade a gene annotation for five of the species (see Methods) and used two additional publicly available annotations for *N. vectensis* and *A. queenslandica*. For most species, the reannotation dramatically increased the total amount of exons as well as the total bases of genes (fig. 4). The only exception was *B. floridae*, where the original annotation had predicted 90% of the genome as genes, while the reannotation had annotated only 44.8% as genes.

We then compared the ratio of intron:intergenic sequence across seven of the reannotated species (fig. 5). Across these species, reannotation significantly shifted the ratio of intron:intergenic sequence, approaching a 1:1 ratio (difference from 1:1 ratio, paired two-end *t*-test, *P* value: 0.014). For *M. brevicollis*, the genome is very small and the majority is exons, so the reannotation was likely to change gene boundaries (separating run-on genes) rather than defining many new genes; our reannotation contains 10,864 genes compared with the 9,196 genes in Monbr1 "best models."

Basic Trends Related to Genome Size

We observed linear correlations of total genome size to both total intronic size and intergenic size (fig. 6) (*P* value: $< 10^{-37}$ for both parameters). A much weaker correlation is observed for exons (*R*-squared: 0.3856, *P* value: 10^{-8}). Because the total amount of exons in the largest genomes can be several times greater than the total size of the smallest genomes used in the study, a correlation is likely to be observed. Thus, the total amount of exons is necessarily affected by total genome size, even if this is not strongly correlated.

Average Intron and Exon Length

The average length of introns linearly scales with the total genome size (fig. 7), in agreement with another study

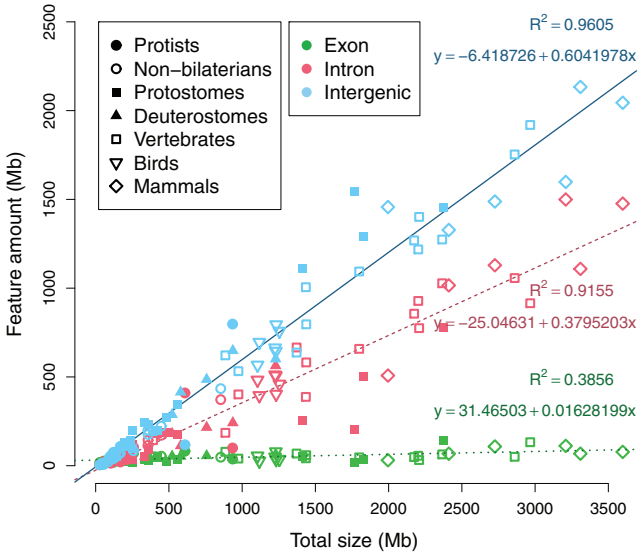

Fig. 6.—Comparison of features to total genome size. The sums of exons, introns, and intergenic regions are plotted against total genome size. Linear coefficients of determination of the three features are displayed by their respective lines. For legend symbols, Deuterostomes refers to all invertebrate deuterostomes, Vertebrates excludes Birds and Mammals.

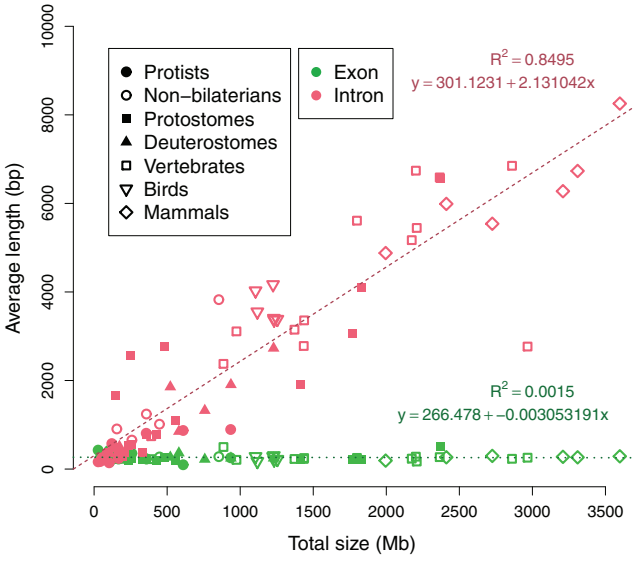

Fig. 7.—Average length of exons and introns. Plot of the average length of exons (green) and introns (pink) as a function of total genome size across all species in this study. Linear coefficients of determination are displayed next to the green (dotted) and red (dashed) linear fit lines, for exons and introns, respectively.

(Elliott and Gregory 2015). However, the average exon length is clearly constrained across animals relative to total genome size, and this may be related to interactions with nucleosomes (Tilgner et al. 2009). Most species have an average exon length between 200 and 300 bases (mean of 263 bp), higher than values reported from previous surveys of exon length (Sakharkar et al. 2004; Zhu et al. 2009). It must be stated that the average values presented here should not be taken as final, because variations in format of the annotations and quality of the genomes will affect the values. Since many genomes are only annotated with ab initio gene predictions, UTR exons may be missing from the annotation and all downstream calculations. Given that the first exon and intron tend to be longer than other exons and introns(Zhu et al. 2009), respectively, absence of five-prime UTRs may result in an underestimation of the average exon length for that species.

Nature of the Exonic Fraction

Unlike introns or intergenic sequence, the total amount of exons does not show a strong linear correlation with total genome size (as seen in fig. 6). However, there is a hyperbolic correlation of the relative fraction of exons (megabases of exons divided by total megabases) compared with total genome size (fig. 8). The smallest genomes are dominated by exons, while the largest genomes are dominated by introns and intergenic regions. This implies a relatively fixed pool of exons or coding space that becomes spread over the genome as the total size increases. The hyperbolic trend resembled the observed hyperbolic relationship between total genome size and coding proportion (Elliott and Gregory 2015). As coding exons are a subset of total exons, measurements of total exons may be a reasonable approximation of coding sequence, but not necessarily vice versa.

Ratio of Introns to Intergenic

Because both intronic and intergenic fractions displayed a linear correlation to total genome size (fig. 6), we next examined the connection between the two fractions. While many species have a ratio of introns:intergenic approaching 1:1 (*R*-squared: 0.8286, *P* value: 5.6×10^{-27}), the majority of genomes are composed of sequence annotated as intergenic regions (fig. 9).

Because of the potential issue of gene annotation accuracy, we tested the linear correlation of introns:intergenic sequence for seven model organisms likely to have accurate annotations. A better linear fit was observed when restricted to the model organisms (*R*-squared: 0.9931, *P* value $= 1.3\times10^{-6}$), suggesting that deviations from the 1:1 ratio of intron:intergenic sequence are due to missing annotations, rather than biological differences. Genomes of model organisms are significantly closer to the reference line (two-tailed *t*-test, *P* value: $< 10^{-7}$ for either absolute distance from 1:1 reference or absolute difference of intron:intergenic ratio to 1), suggesting that the better annotations of model organisms predict a ratio of 1:1 of intron:intergenic sequence. Overall, the comparison of genomes of model to nonmodel organisms is compatible with the hypothesis that the predicted amount of the genome that is transcribed varies more by annotation quality than biological differences.

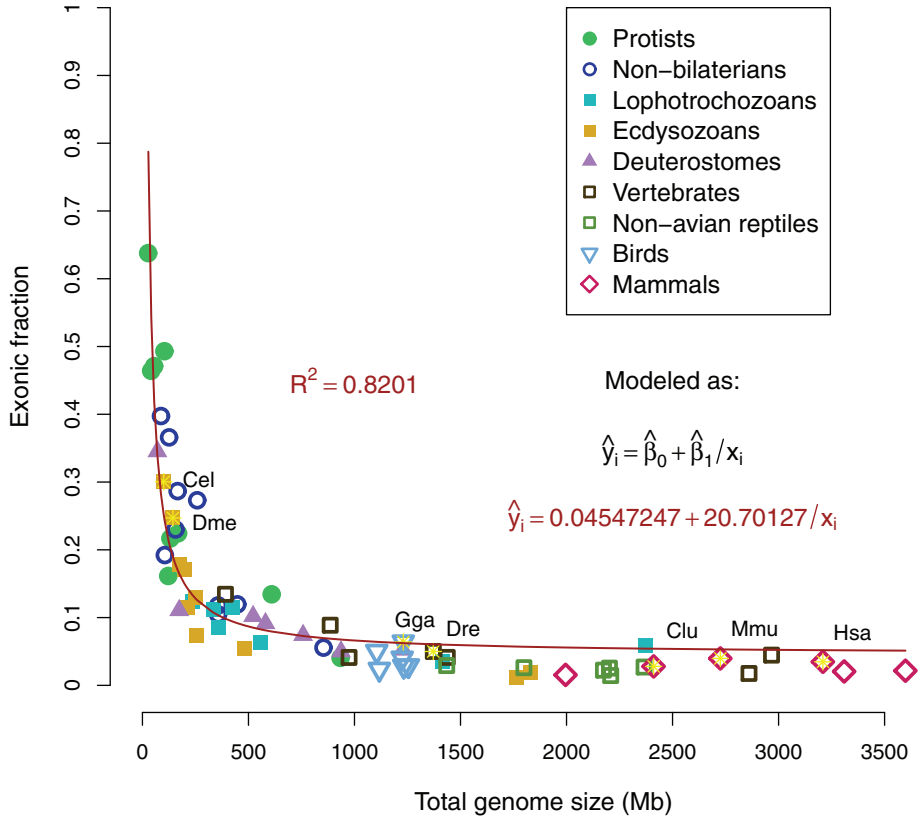

Fig. 8.—Exonic fraction compared with total genome size. Relative fraction of the genome that is defined as exons compared as a function of total size. Coefficients of determination of a hyperbolic model are displayed. Seven model organisms (human, mouse, dog, chicken, zebrafish, fruit fly and nematode) are indicated by three-letter abbreviations. The formula for the fitted model is displayed in red.

We then examined if there is a difference between genomes of vertebrates and invertebrates. No significance difference is observed between the two model invertebrates and five vertebrates (two-tailed *t*-test, *P* value: 0.99). Among all species in the study, significant differences are tenuous and highly dependent on the species selected (fig. 10). For example, chordates against nonchordates is not significant (*P* value: 0.128) while vertebrates against invertebrates is significant (*P* value: 0.008). However, the observed significance appears to be an artifact of the abundance of low-quality genomes of protostomes, since comparison of vertebrates against nonbilaterians is not significant (*P* value: 0.83). This difference is most simply explained by the similarity between vertebrate groups. That is to say, annotation of a new mammalian genome is facilitated by existing knowledge of gene structures in other mammals, rather than true differences in genome organization.

Several genomes are below the 1:1 reference line, indicating slightly more introns than intergenic, such as the choanoflagellate *S. rosetta*, the honeybee *A. mellifera*, the anemone *E. pallida*, and placozoan *Hoilungia hongkongensis*. For *A. mellifera*, it was noted that improvements in versions of the genome also included better placement of repetitive intergenic sequences (Weinstock et al. 2006), suggesting that the relative surplus of introns is merely due to the absence of some intergenic sequences in the final assembly. As for *E. pallida* and *H. hongkongensis*, these species stand out as having relatively high heterozygosity, 0.4% (Bellis et al. 2016) and 1.8% (manuscript in preparation), respectively. Although these values are lower than the observed heterozygosity in many other invertebrates (Leffler et al. 2012), some highly heterozygous sequences may have caused assembly problems during scaffolding (as proposed in fig. 3).

Evolution of the Genic Fraction

The amount of the genome that is composed of genes was highly variable across the genomes in our study, ranging from 12.5% up to 87.1% of the genome. Unlike the exonic fraction, the relationship of the fraction of the genome that is genes to the total size is less obvious (fig. 11), in part because this parameter is most subject to gene annotation accuracy. The fraction of the genome that is exons (and perhaps coding) appeared relatively fixed (fig. 8), yet the intron size was linearly correlated to the total size (fig. 6), therefore the fraction that is genes (exons and introns combined) was expected to be a combination of the two trends. Three correlation models

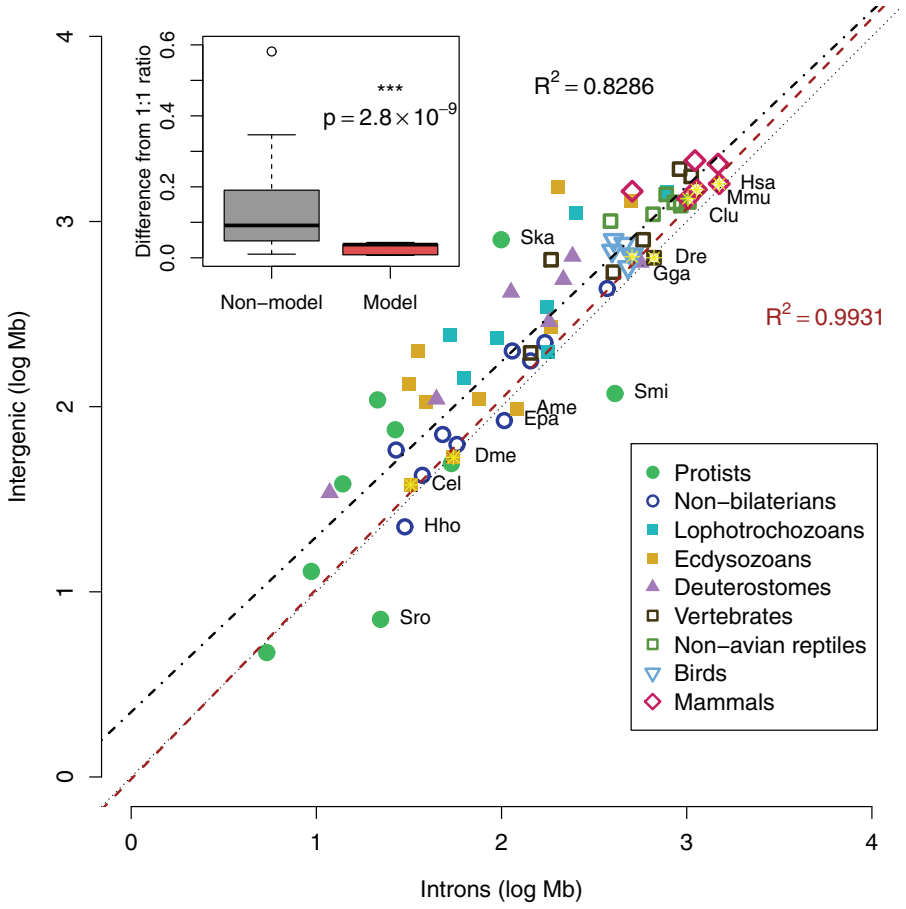

Fig. 9.—Comparing intronic and intergenic fractions. Log-scale plot of total intronic size versus total intergenic size. The dotted line shows a ratio of 1:1 as a reference, although most genomes are above this line. Seven model organisms (as in fig. 8) are indicated by three-letter codes with yellow stars. Black dashed line displays the linear fit of all species in the study (*R*-squared: 0.8286, *P* value: 5.6×10^{-27}), while the red line displays the linear fit for only the seven model organisms (*R*-squared: 0.9931, *P* value: 1.3×10^{-6}). Names are displayed for model species, two dinoflagellates (Ska: *Symbiodinium kawagutii*, Smi: *Symbiodinium minutum*) and select species with ratios of intron:intergenic >1, choanoflagellate *Salpingoeca rosetta* (Sro), honeybee *Apis mellifera* (Ame), anemone *Exaiptasia pallida* (Epa), and placozoan *Hoilungia hongkongensis* (Hho). All other species names are omitted for clarity. The inset graph shows box plot of difference of the intron:intergenic ratio to 1, showing the model organisms (red) have significantly different ratios compared with the rest of the genomes (paired two-end *t*-test, *P* value: 2.8×10^{-9}).

were tested: hyperbolic (double-log), exponential (single-log), and linear. Of these, the hyperbolic model fit best (*R*-square: 0.3649, *P* value: $< 10^{-8}$), and no correlation was found for the other models. Restricting the linear model to only genomes larger than 500 Mb found essentially no correlation (*R*-squared: 2.5×10^{-4}), suggesting that the genic fraction is unrelated to total genome size in large genomes but not in small genomes.

Again, the importance of gene annotation accuracy cannot be ignored and needs to be emphasized. When restricting to the seven model organisms, the range of values is narrower, from 44.9% to 62.9%. The same three correlation models were applied to the genomes of model organisms, again finding that the hyperbolic model best explained the variation in the genic fraction of model organisms (hyperbolic *R*-squared: 0.8091, *P* value = 0.0058; exponential *R*-squared = 0.6709; linear *R*-squared = 0.6835). Rather than simply having no correlation to total size, these results suggest that the genic fraction is fixed at ~50% in large genomes.

Discussion

Diagnostic Relationship of Introns to Intergenic Sequence

An increasing number of genomes of any nonmodel organisms are sequenced to answer evolutionary questions. For example, genomes of taxa from all four nonbilaterian groups were recently sequenced to understand how similar these genomes are to humans (Putnam et al. 2007; Srivastava et al. 2008, 2010; Ryan et al. 2013), and found that we share much more in terms of genes with these groups than had been previously thought. Yet, one of the main challenges in studying the genomes of nonmodel organisms is that there is little *a priori* information about gene structure or content. It

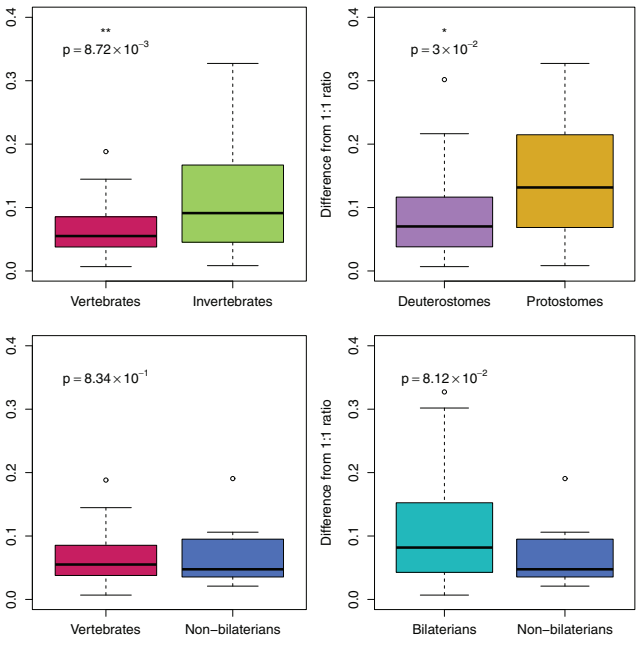

Fig. 10.—Comparing intron-intergenic ratios among animal groups. Difference of the intron:intergenic ratio to 1 across four pairs of animal groups. Invertebrates (green) include all nonbilaterian taxa. Deuterostomes and protostomes are both assumed to be monophyletic.

would be expected that finding orthologs of human genes is relatively easy, but does not inform us about other genes that differ from humans. How should we know when we have found all of the genes? Our results provide some guidance here and suggest that there is a constant ratio of introns to intergenic sequence in all animals. This relationship holds even for animals with small genomes, such as the model organisms *D. melanogaster* and *C. elegans*, suggesting that organisms with small genomes and many currently sequenced invertebrates are subject to the same forces as organisms with large genomes.

Unusual Cases of Genomes

Based on our model, the majority of genomes appear to be underannotated, in that substantial portions of the genome are not predicted to be transcribed when in fact many probably are. However, only two species, the lancelet *B. floridae* and the dinoflagellate *S. minutum*, display a dramatic trend in the opposite way, that is, the majority of the genome is annotated as genic (being primarily introns).

For the lancelet *B. floridae*, the original JGI gene models had annotated almost 90% of the genome as genes (Putnam et al. 2008), the majority (85%) of that sequence being introns. Our reannotation of this genome displays the opposite trend, where more of the genome is intergenic than intronic. The original JGI annotations did not include any validation of the predicted genes, as predictions were made using mapped ESTs only as inputs for the gene model training. From this, we consider it more likely that the RNAseq-based transcripts more accurately resemble the true gene structures, albeit missing some genes. In addition, other evidence suggests that the *B. floridae* annotations may have been unusual or erroneous (Bányai and Patthy 2016). A study of domain combinations found that *B. floridae* had by far more fusions than any other species (across all eukaryotes) and had to be excluded from the analysis (Zmasek and Godzik 2012), precisely the expected result if the majority of genes were erroneously fused.

The only other species have a much larger ratio of intron to intergenic was the dinoflagellate *S. minutum*. It was described that its genome contained many long stretches of genes on the same strand, sometimes continuing for hundreds of kilobases (Shoguchi et al. 2013). The authors also note that the de novo assembled transcriptome appears to contain transcripts spanning multiple genes and containing multiple open reading frames, indicating the possibility that dinoflagellate symbionts can make cistronic transcripts. This species is not an animal, so it should not be assumed that animal modes of transcription are conserved across all eukaryotes. However, it should be noted that a recently published genome of another symbiotic dinoflagellate species *S. kawagutii* (Lin et al. 2015) does not display the same pattern, and instead appears to have a much greater fraction of intergenic regions than introns.

Genome Composition across Metazoa

Previous studies have discussed problems with trying to relate the number of genes to the size of the genome (Hahn and Wray 2002; Gregory 2005; Denton et al. 2014). One study (Elliott and Gregory 2015) found a weak positive correlation between genome size and number of genes. This parallels our finding that total exonic sequence is weakly correlated to total genome size (fig. 6). However, this measurement can be problematic if the genome assembly is highly fragmented, containing a large number of short contigs or scaffolds. In such cases, gene number is unlikely to correlate to genome size for the same reason as the difficulties in predicting the genic fraction, that is, it is strongly affected by gene annotation errors. In our schematic (fig. 2), a gene that is split up onto three contigs would therefore be counted as three genes, albeit short ones. If this occurs on a genome-wide scale, the count of genes will be inaccurate. Parts of genes would be individually annotated as genes, increasing the total gene number without much change to the total number of exonic bases.

Rather than relying on counts of genes or determining coding sequence, we instead examined sequence that is annotated as exons. We found that while a weak positive correlation is observed between total exonic bases and genome size, most of the difference in size is related to introns and intergenic sequence. The amount of the genome that is composed of introns is linearly related to the total genome

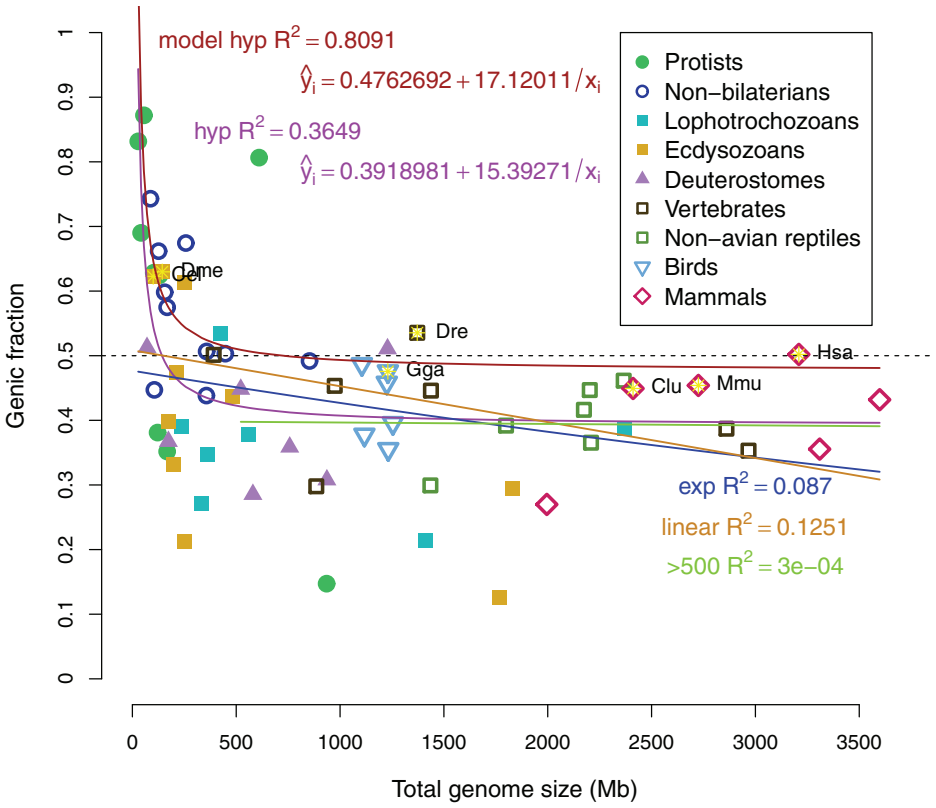

Fig. 11.—Genic fraction compared with total genome size. Relative fraction of the genome that is defined as genes compared as a function of total size. A number of correlative models (hyperbolic in purple, exponential in blue, linear in orange) were tested and coefficients are displayed. Linear correlation is expected to be zero if genic and intergenic fractions "expand" indifferently after a certain size, which appears to be ~500 Mb. Linear correlation including only genomes larger than 500 Mb is also displayed as the green line. Seven model organisms (as in fig. 8) are indicated by three-letter codes and yellow stars. The hyperbolic correlation model for the seven model organisms is shown in red. The formulae for the fitted models are displayed in red and purple, for model organisms and all organisms, respectively.

size (fig. 6). Also considering the measured linear correlation of intergenic sequence to total size, it is not surprising that most species have roughly a 1:1 ratio of introns:intergenic sequence (fig. 9). This appears to be the case regardless of genome size or the total exonic sequence. For instance, the genome of the choanoflagellate *M. brevicollis* has 9.3 Mb of introns and 10.1 Mb of intergenic sequence (a ratio of 0.92) compared with 19.3 Mb of exons.

Therefore, model animals (and probably all animals) transcribe nearly half of the genome, where species with smaller genomes (exon-rich) transcribe more than half (fig. 11). There does not appear to be a significant difference in the genic fraction based on animal group (fig. 10), that is, all animals appear to follow this rule. One study had shown that some larger metazoan genomes were depleted in genes (Fernandez-Valverde and Degnan 2016), yet this study made use of a small number of species for comparison and included several chordates known for their very small genomes, the tunicate *C. intestinalis* and the pufferfish *T. rubripes*. The authors examined windows of 50 kb and found that 80% of the human genome was lacking any gene (Fernandez-Valverde and Degnan 2016), though it is unclear if this analysis was restricted to protein coding genes. However, we found that 50.2% of the human genome is composed of genes (93% of that is introns).

While genomes of the model organisms and many nonmodels organisms appear to follow the hyperbolic relationship of genic fraction to size, nonetheless, a large number of the genomes in this study appear to be composed of much less than 50% genes. That observation is best explained by the hypothesis that many genomes are missing genes. These missing genes may or may not be coding, though perhaps missing gene content is made of lineage-specific proteins. Because annotation of the genome by RNAseq per se cannot distinguish coding genes from noncoding ones, we could not determine coding fractions for all species. Even for putative noncoding transcripts, some may be coding (Wilson and Masel 2011; Slavoff et al. 2012; Guttman et al. 2013), thus protein sequencing may reveal the true nature of these transcripts.

Evolution of Genomes

The genic fraction has a hyperbolic relationship to the total genome size. The modeled curve flattens ~500 Mb, after that

point, introns and intergenic regions are expected to expand, on an average, equally across the genome resulting in ~50% of the genome as genes (the majority of that being introns) and the other 50% as intergenic sequence. It should be noted that larger genomes still have more exonic bases than small genomes, though the difference in total genome size across animals is mostly from introns or intergenic sequence.

It has been theorized that changes in genome size are a balance between short deletions and long insertions (Petrov 2002). If the last common ancestor of all metazoans had a relatively small genome (under 100 Mb, resembling some single-cell eukaryotes in our study), then the majority of modern animals have undergone dramatic expansion of their genomes, meaning dominated by insertions or duplications. How does this expansion occur and does it favor a novel origin of introns or expansion of intergenic sequences? Following the trend in figures 9 and 11, it appears that small genomes are dominated by genes because they are mostly exons, and both genes and intergenic sequences are expanded in equally as the genomes enlarge. Mechanistically, these insertions are likely to be mediated by transposable elements or replication errors. As small genomes become invaded by transposable elements (perhaps following some genomic stress like genome duplication), introns appear and expand at roughly the same rate as intergenic sequences producing a 1:1 ratio of intron:intergenic across all species (fig. 9).

Above a certain size (~500 Mb), genic and intergenic sequences expand almost equally, where 50% of the genome is genic; exons comprise an almost negligible fraction of the genome, which is otherwise composed of approximately equal fractions of introns and intergenic sequences. This might be explained by changes in diversity of transposable elements, as the highest diversity was found in genomes ranging from 500 Mb to 1.5 Gb (Elliott and Gregory 2015). Larger genomes appeared to be flooded by transposable elements of a single type. Thus, above 500 Mb, it can be predicted that select transposable elements become prevalent and multiply throughout the genome, but on an average end up expanding introns and intergenic sequences equally.

Relationship to Phenotypic Complexity

The size of the genome can vary greatly even for closely related organisms. This has been called the "c-value paradox" (Thomas 1971; Moore 1984), based on the observation that although the many organisms have larger genomes relative to similar species (bigger "c-value"), this measurement does not equate with more or less complex organisms in a straightforward way. A classic example of this is frog genus *Xenopus*, where the genome of the species *X. laevis* is almost twice as large as the species *X. tropicalis* (Thiébaud and Fischberg 1977), though the animal is not twice as "complex." Similar observations have been made that the number of genes appears unrelated to the size of the genome and the complexity (sometimes called the "g-value paradox"; Hahn and Wray 2002; Schad et al. 2011).

If neither genome size nor gene number are clearly related to complexity, then what is? Another relationship has been proposed between the usage of alternative splice variants and organismic complexity because variation in splicing can increase the number of potential proteins from an overall fixed pool of exons (Nilsen and Graveley 2010). Vertebrates and specifically mammals tend to splice transcripts more than invertebrates (meaning models fruit fly and nematode) (Brett et al. 2002; Kim et al. 2004). One study reported a good correlation (*R*-squared of 0.80) of splicing to organismic complexity measured by cell types (Chen et al. 2014), but also reported that this trend effectively disappeared when correcting for sequencing depth, using the number of ESTs available as a proxy for annotation quality. The largest invertebrate genome used in that study was the deer tick *I. scapularis*, which did have a measured number of cell types but unfortunately could not be analyzed further, leaving the bulk of the analysis weighted heavily by mammals and small-genome insects.

However, other studies report that alternative splicing is more frequent when the surrounding introns are long (Kim et al. 2007; Pickrell et al. 2010), suggesting that organisms with large genomes (and therefore larger introns) might be predisposed to splice. This could suggest that some of the invertebrates in our study may have more complex splicing patterns than are annotated in the current genome versions. For the largest invertebrate genome in our study, the octopus *O. bimaculoides*, only 14.8% of loci appeared to have alternative splice variants (Albertin et al. 2015). In our reannotation we found only 6.4% of all loci have any type of splice variant. However, the majority of predicted transcripts (75%) are single exon loci, and possibly many genes are fragmented across multiple contigs (as in fig. 2). When restricted to loci with multiple exons (15% of total loci), 41% have more than one variant. These data from *O. bimaculoides* suggested that overall patterns in splicing do not display a reliable connection to organismic complexity when complexity is generalized across animal groups. However, without proper measurements of cell types from the octopus, it cannot be assumed that the number of cell types resembles the value for the fruit fly, which was implicit in other studies given that all protostomes were effectively represented by insects (Chen et al. 2014). Thus, it could be the case that the octopus, with a large genome, has a large number of cell types and many genes are spliced, all in agreement with the splicing-complexity hypothesis.

It is a challenge to separate these observations from biases in sequencing depth (of transcripts or ESTs) and data availability. In our study, we could only make use of five invertebrates with relatively large genomes, the cnidarian *H. magnipapillata*, the pearl oyster *P. fucata*, the horseshoe crab *L. polyphemus*, the deer tick *I. scapularis*, and the octopus *O. bimaculoides*. On the

other hand, NCBI has over 100 genomes of mammals available for download. Alternatively, the repertoire of splice factors or the genes that are most spliced may be of greater importance than just splicing in general. Our understanding is likely to be improved with more deeply sequenced transcriptomes from large-genome invertebrates.

Limitations

Because we were making use of mostly public data, our analyses were subject to both technical and biological limitations. There are a small number of taxa with sequenced genomes from many invertebrate groups. Because the majority of sequenced vertebrate genomes are large and the majority of sequenced invertebrate genomes are small (Gregory 2005), the axis of simple invertebrate to complex vertebrate is synonymous with small to large genomes, and thus the prevalence of splicing in large-genome animals may be a consequence of the size of the genome and complexity may be only correlated. This issue is not simple to resolve, as there may not be members in all animal groups with both small and large genomes. For instance, a survey of genome sizes across Porifera stated that the largest genome out of the 70 species sampled was ~600 Mb (Jeffery et al. 2013). Thus, there may not be any "large" genomes in this phylum, and likewise for other invertebrate groups. Compared with birds, however, where the smallest genome identified to date is from the black-chinned hummingbird (estimated 910 Mb) (Gregory et al. 2009), perhaps no bird will be found that has a "small" genome.

Our use of public genome annotations was limited in part from difficulties in defining elements. Much like definitions of transcribed pseudogenes, the identification of long-intergenic noncoding RNAs, or lincRNAs, presents a paradox of definitions. Noncoding RNAs with known functions are arguably genes, such as the X-inactivation transcript Xist, thus any functional transcribed intergenic RNA is by definition not intergenic; it is genic. This distinction rests upon discovery of a function of these putative RNAs. In the context of the ENCODE project or MouseENCODE (Consortium 2014), transcription was found of intergenic regions accounting for almost another 20% of the genomes of human and mouse, depending on the analysis (van Bakel et al. 2010; Clark et al. 2011). If this were all functional, then the genic fraction of the genome would be far above 50% for large genomes and the ratio of intron:intergenic sequence would not be expected to be close to 1:1. Alternatively, if most of these intergenic transcripts are nonfunctional "noise," then our results are supported as presented. Therefore, consideration of the importance or genic quality rests upon the distinction between functional RNAs and noisy transcription. Existing data are not adequate to identify functions, but several experiments may improve our understanding. Conceptually, the most straightforward approach is knocking out regions of transcribed "gene deserts" in mouse or human cells, but on a larger scale than a previous study (Nóbrega et al. 2004). Additionally, better models of transcriptional noise or random transcription may inform whether or not the observed transcriptional patterns from the ENCODE project are consistent with noise.

Conclusion

We have shown that a set of animals from 12 phyla transcribe at least half of their genomes in a size-dependent fashion. For large genomes, the amount of exons is almost negligible, where introns account for most of the genic sequence. In such cases, genic sequence is almost equal to the amount of intergenic sequence. Whereas for small genomes, exons can be a major fraction of the genome, resulting in the appearance of gene-dense genomes. This parity between introns and intergenic sequence is likely a universal feature of animal genomes, though this may be tested with addition of many more animal taxa from other phyla that do not have sequenced members. Previous findings of genomic differences between animal groups are likely to result from a sampling bias, rather than biological differences. Future improvements in assembly and annotation of animal genomes may reveal unanticipated sources of complexity and gene regulation with implications for the evolution of animals.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

W.R.F would like to thank M. Eitel for helpful comments on the manuscript. This work was supported by a LMUexcellent grant (Project MODELSPONGE) to G.W. as part of the German Excellence Initiative. The authors declare no competing interests.

Literature Cited

Albertin CB, et al. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524:220–224.

Alföldi J, et al. 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477(7366):587–591.

Altenhoff AM, et al. 2016. Standardized benchmarking in the quest for orthologs. Nat Methods 13(5):425–430.

Amemiya CT, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496(7445):311–316.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bányai L, Patthy L. 2016. Putative extremely high rate of proteome innovation in lancelets might be explained by high rate of gene prediction errors. Sci Rep. 6(April):30700.

Baumgarten S, et al. 2015. The genome of *Aiptasia*, a sea anemone model for coral symbiosis. Proc Natl Acad Sci. USA. 112(38):11893–11898.
Bellis ES, Howe DK, Denver DR. 2016. Genome-wide polymorphism and signatures of selection in the symbiotic sea anemone *Aiptasia*. BMC Genomics 17:160.
Brent MR. 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet. 9(1):62–73.
Brett D, Pospisil H, Valcárcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. Nat Genet. 30(1):29–30.
Canapa A, Barucca M, Biscotti MA, Forconi M, Olmo E. 2016. Transposons, genome size, and evolutionary insights in animals. Cytogenet Genome Res. 217–239.
Chapman JA, et al. 2010. The dynamic genome of hydra. Nature 464(7288):592–596.
Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. 2014. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Mol Biol Evol. 31(6):1402–1413.
Chipman AD, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. PLoS Biol. 12(11).
Clark MB, et al. 2011. The reality of pervasive transcription. PLoS Biol. 9(7):5–10.
Colbourne JK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331(6017):555–561.
Consortium ME. 2014. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515(7527):355–364.
Daubin V, Moran N. 2004. Comment on "The origins of genome complexity". Science 306(5698):978.
Dehal P, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298(5601):2157–2167.
Denoeud F, et al. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science (2010):1381.
Denton JF, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 10(12).
Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. 27(15):3219–3228.
Elliott TA, Gregory TR. 2015a. Do larger genomes contain more diverse transposable elements? BMC Evolutionary Biol. 15(1):69.
Elliott TA, Gregory TR. 2015b. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B 370(1678):20140331.
Fairclough SR, et al. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. Genome Biol. 14(2):R15.
Fernandez-Valverde SL, Calcino AD, Degnan BM. 2015. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*. BMC Genomics 16(1):1–11.
Fernandez-Valverde SL, Degnan BM. 2016. Bilaterian-like promoters in the highly compact *Amphimedon queenslandica* genome. Sci Rep. 6:22496.
Fortunato SV, et al. 2014. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature 514(7524):620–623.
Ganapathy G, et al. 2014. High-coverage sequencing and annotated assemblies of the budgerigar genome. GigaScience 3:11.
Green RE, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science 346(6215):1254449–1254449.
Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. Nat Rev Genet. 6(9):699–708.
Gregory TR, Andrews CB, McGuire JA, Witt CC. 2009. The smallest avian genomes are found in hummingbirds. Proc Biol Sci. 276(1674):3753–3757.
Guigó R, et al. 2006. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 7(Suppl 1):S2.1–31.
Gulia-Nuss M, et al. 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun. 7:10507.
Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 154(1):240–251.
Hahn MW, Wray GA. 2002. The g-value paradox. Evol Dev. 4(2):73–75.
Han K, et al. 2013. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. Sci Rep. 3:2101.
Hardie DC, Hebert PD. 2004. Genome-size evolution in fishes. Can J Fish Aquat Sci. 61(9):1636–1646.
Hellsten U, et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. Science 328(5978):633–636.
Hoff KJ, Stanke M. 2013. WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 41:W1):W123–W128.
Howe K, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496(7446):498–503.
Huang Y, et al. 2013. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. Nat Genet. 45(7):776–783.
Jeffery NW, Jardine CB, Gregory TR. 2013. A first exploration of genome size diversity in sponges. Genome 56(8):451–456.
Kajitani R, et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24(8):1384–1395.
Keeling CI, et al. 2013. Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. Genome Biol. 14(3):R27.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115(1):49–63.
Kim D, et al. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14(4):R36.
Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35(1):125–131.
Kim H, Klein R, Majewski J, Ott J. 2004. Estimating rates of alternative splicing in mammals and invertebrates. Nat Genet. 36(9):915–916.
King N, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451(7180):783–788.
Kirkness EF. 2003. The dog genome: survey sequencing and comparative analysis. Science 301(5641):1898–1903.
Koning APJD, et al. 2013. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proc Natl Acad Sci. 110(51):20645–20650.
Leffler EM, et al. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10(9).
Lin S, et al. 2015. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. Science 350(6261):691–694.
Luo YJ, et al. 2015. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. Nat Commun. 6:1–10.
Lynch M. 2004. Response to comment on "The Origins of Genome Complexity". Science 306(5698):978b–978b.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302(5649):1401–1404.
Mikkelsen TS, et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature 447(7141):167–177.
Mita K, et al. 2004. The genome sequence of silkworm, *Bombyx mori*. DNA Res. 11:27–35.
Moore G. 1984. The C-value paradox. BioScience 34(7):425–429.

Moran Y, et al. 2014. Cnidarian microRNAs frequently regulate targets by cleavage. Genome Res. 24(4):651–663.
Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature 463(January).
Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. 2004. Megabase deletions of gene deserts result in viable mice. Nature 431(7011):988–993.
Nossa CW, et al. 2014. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. GigaScience 3:9.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067.
Pertea M, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33(3).
Petrov D. 2002. Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61(4):531–544.
Pettersson ME, Kurland CG, Berg OG. 2009. Deletion rate evolution and its effect on genome size and coding density. Mol Biol Evol. 26(6):1421–1430.
Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet. 6(12):1–11.
Pisani D, et al. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc Natl Acad Sci. 112(50):201518127.
Preußer C, Jaé N, Bindereif A. 2012. MRNA splicing in trypanosomes. Int J Med Microbiol. 302(4-5):221–224.
Prochnik SE, et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science 329(5988):223–226.
Putnam NH, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317(5834):86–94.
Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453(7198):1064–1071.
Read B, et al. 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 9–13.
Richards S, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452(7190):949–955.
Ryan JF, et al. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342(6164):1242592–1242592.
Sakharkar MK, Chow VTK, Kangueane P. 2004. Distributions of exons and introns in the human genome. In Silico Biol. 4(4):387–393.
Schad E, Tompa P, Hegyi H. 2011. The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 12(12):R120.
Shaffer HB, et al. 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol. 14(3):R28.
Shinzato C, et al. 2011. Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature 476(7360):320–323.
Shoguchi E, et al. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. Curr Biol. 23:1399–1408.
Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GAM. 2010. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res. 38(15):4946–4957.
Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493(7433):526–531.
Simakov O, et al. 2015. Hemichordate genomes and deuterostome origins. Nature 1–19.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Genome Anal. 31(June):9–10.
Slavoff SA, et al. 2012. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol. 9(1):59–64.
Smith JJ, et al. 2013. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet. 45(4):415–421. 421e1–2.
Sodergren E, et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314(5801):941–952.
Srivastava M, et al. 2008. The Trichoplax genome and the nature of placozoans. Nature 454(7207):955–960.
Srivastava M, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466(7307):720–726.
Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24(5):637–644.
Suga H, et al. 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun. 4:2325.
Takeuchi T, et al. 2012. Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Res. 19(2):117–130.
The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282(5396):2012–2018.
Thiébaud CH, Fischberg M. 1977. DNA content in the genus Xenopus. Chromosoma 59(3):253–257.
Thomas CA. 1971. The genetic organization of chromosomes. Annu Rev Genet. 5(1):237–256.
Tilgner H, et al. 2009. Nucleosome positioning as a determinant of exon recognition. Nat Struct Mol Biol. 16(9):996–1001.
Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13(1):36–46.
Vallender EJ. 2009. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. Methods 49(1):50–55.
van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. PLoS Biol. 8(5).
Venkatesh B, et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. Nature 505(7482):174–179.
Voskoboynik A, et al. 2013. The genome sequence of the colonial chordate, *Botryllus schlosseri*. eLife 2:e00569.
Wang Z, et al. 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet. 45(6):701–706.
Warren WC, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453(7192):175–183.
Warren WC, et al. 2010. The genome of a songbird. Nature 464(7289):757–762.
Weinstock GM, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443(7114):931–949.
Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol. 3:1245–1252.
Zhang G, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346(6215):1311–1320.
Zhang GG, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490(7418):49–54.
Zhang X, Goodsell J, Norgren RB. 2012. Limitations of the rhesus macaque draft genome assembly and annotation. BMC Genomics 13(1):206.
Zhu L, et al. 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics 10(1):47.
Zmasek CM, Godzik A. 2012. This Déjà Vu feeling-analysis of multidomain protein evolution in eukaryotic genomes. PLoS Comput Biol. 8(11).

Associate editor: Dan Graur