bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similar ratios of introns to intergenic sequence across animal genomes

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

(1) Department of Earth and Environmental Sciences, Paleontology and Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner Straße 10, 80333 Munich, Germany
(2) GeoBio-Center, Ludwig-Maximilians-Universität München, Munich, Germany
(3) Bavarian State Collection for Paleontology and Geology, Munich, Germany
Keywords: metazoa, comparative genomics, junk DNA, complexity, C-value

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 all animal groups and some single-cell eukaryotes for general trends in genome usage and composition, 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 non-model organisms are underannotated 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.

Author's Summary

Within our anthropocentric genomic framework, many analyses try to define humans, mammals, or vertebrates relative to the so-called "lower" animals. This implicitly posits that vertebrates are complex organisms with large genomes and invertebrates are simple organisms with small genomes. This has the problem that genome size is therefore presumed to correlate with complexity and ignores any unknown complexity of vast numbers of invertebrate groups, many with large genomes. Animals vary widely in genome size, by almost three orders of magnitude, but when sequencing new animal genomes preference is given to those with smaller genomes for reasons of cost. In trying to understand how genomes are used in general, there is an added layer of complication from quality of the assembly and annotation. We have examined genome usage across a wide range of animals and have described ways to account for errors of low-quality annotations. We also show that the genomes of invertebrates and vertebrates are not so different, and that when large-genome invertebrates are considered, the fraction of the genome that is genes appears to be strongly predictable by genome size.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Understanding why genomes vary greatly in size and how organisms make different use their genomes have been central questions in biology for decades [1]. For many bacteria, the majority of the genome is composed of relatively short genes, averaging around 1000bp, 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 [2]. However, in eukaryotic organisms genomes can be over a thousand-fold larger than bacterial genomes, due to an increase in the number of genes (tens of thousands compared to 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 [3, 4]. Unfortunately, the predicted genes are often taken for granted as being correct when these are only hypotheses of gene structure [5]. For example, one study found that almost half of the genes in the *Rhesus* monkey genome had a predictable annotation error when compared to the closest human homolog [6]. This has profound implications for all downstream analyses, such as studying evolution of orthologous proteins [7] and phylogeny based on protein matrices or gene content [8, 9]. 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 non-model 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 [10, 11], 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 usage 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 [12–14]. Additionally, a handful of studies have analyzed genome size in connection to other parameters such as indels [15], transposon content [16–19], average intron length [20, 21] or total intron length [18]. Despite these advances, none of these studies have estimated the amount of the genome that is genic, and 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.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Genomic data sources

Data sources and parameters are available in Supplemental Table 1.

Genomic scaffolds and annotations for *Ciona intestinalis* [22], *Branchiostoma floridae* [23], *Trichoplax adherens* [24], *Capitella teleta* [25], *Lottia gigantea* [25], *Helobdella robusta* [25], *Saccoglossus kowalevskii* [26], *Monosiga brevicollis* [27], *Emiliania huxleyi* [28], and *Volvox carteri* [29] were downloaded from the JGI genome portal.

Genome assemblies and annotations for *Sphaeroforma arctica*, *Capsaspora owczarzaki* [30] and *Salpingoeca rosetta* [31] were downloaded from the Broad Institute.

GFF annotations v2.1 [32] for *Amphimedon queenslandica* were downloaded from the Amphimedon Genome website (http://amphimedon.qcloud.qcif.edu.au/downloads.html), and v1 annotations [33] and assemblies were downloaded from Ensembl.

For *Nematostella vectensis*, Nemve1 assembly and annotations [34] were downloaded from JGI, and the transcriptome for comparative reannotation was downloaded from http://www.cnidariangenomes.org/ [35].

Genome assembly, transcriptome assemblies from Cufflinks and Trinity, and GFF annotations for *Mnemiopsis leidyi* [8] were downloaded from the Mnemiopsis Genome Portal (http://research.nhgri.nih.gov/mnemiopsis/). Assembly and annotations for *Sycon ciliatum* [36] were downloaded from COMPAGEN. Assembly and annotation for *Botryllus schlosseri* [37] were downloaded from the Botryllus Schlloseri genome project (http://botryllus.stanford.edu/botryllusgenome/). Assembly and annotation for *Exaiptasia pallida* (formerly *Aiptasia sp.*) [38] were downloaded from http://reefgenomics.org. Assembly and annotation for *Oikopleura dioica* [39] 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* [40] were downloaded from the Dinoflagellate Resources page (web.malab.cn/symka_new/index.jsp).

Assemblies and annotations for *Symbiodinium minutum* [41], *Pinctada fucata* [42], *Acropora digitifera* [43], *Lingula anatina* [44], *Ptychodera flava* [26], and *Octopus bimaculoides* [45] were downloaded from the OIST Marine Genomics Browser (http://marinegenomics.oist.jp/gallery/).

Builds of *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Canis lupus* [46], *Monodelphis domestica* [47], *Ornithorhynchus anatinus* [48], *Xenopus tropicalis* [49], *Struthio camelus* [50], *Gallus gallus*, *Taeniopygia guttata* [51], *Aptenodytes forsteri* [50], *Anas platyrhynchos* [52], *Melopsittacus undulatus* [53], *Alligator mississippiensis* [54], *Anolis carolinensis* [55], *Chrysemys picta bellii* [56], *Chelonia mydas* [57], *Pelodiscus sinensis* [57], *Python bivittatus* [58], *Salmo salar*, *Danio rerio* [59], *Latimeria chalumnae* [60], *Petromyzon marinus* [61], *Callorhinchus milii* [62], *Crassostrea gigas* [63], *Dendroctonus ponderosae* [64], *Tribolium castaneum* [65], *Bombyx mori* [66], *Limulus polyphemus* [67] were downloaded from the NCBI Genome server.

Genome assemblies and annotations of *Caenorhabditis elegans* [68], *Drosophila melanogaster*, *Strongylocentrotus purpuratus* [69], *Daphnia pulex* [70], *Apis mellifera* [71], *Ixodes scapularis* [72], *Strigamia maritima* [73] 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

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1.8Gb while the reported genome size is 2.7Gb [67], 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. Total bases 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 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 bases or their reverse compliment are transcribed and retained following splicing in some case. Intergenic sequence is defined as the difference between total sequence bases and genic bases, 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 non-coding 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 non-coding 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 called a non-coding RNA. Pseudogene features are not annotated for all species, making it difficult to compare broadly. Additionally, for most non-model 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 were 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∼x), exponential (log(y)∼x), or hyperbolic (y∼1/x) models, and the "predict()" function was used to model curves. The R source code used to generate figures is 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 non-redundant exons for all splice variants were taken into account for determination of averages. Unlike the total base 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 lengths were counted per non-redundant exon for each gene, summed across all genes and divided by the number of non-redundant 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.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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* [24] were reannotated with AUGUSTUS v3.0.3 [74] 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 [75].

The original Monbr1 scaffolds of *M. brevicollis* [27] were reannotated with AUGUSTUS as for *T. adherens*, using the same parameters except trained using the Monbr1 ESTs with the webAugustus Training server [75].

For the hydrozoan *H. magnipapillata*, the original assembly was downloaded from JGI [76] 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 [77] and StringTie v1.0.4 [78] 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 [23] were reannotated using TopHat22 v2.0.13 [77] and StringTie v1.0.4 [78] 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 [77] and StringTie v1.0.4 [78] 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 [45], and were reannotated using TopHat2 v2.0.13 [77] and StringTie v1.0.4 [78] 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 [79, 80].

The smallest animal genome used in this study is that of the larvacean *Oikopleura dioica* (70Mb), while the largest is that of the opossum *Monodelphis domestica* (3598Mb). 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 10Gb (haploid genome size) [81], such that a shark genome of only 1Gb may not be "normal" for sharks.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 non-model 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 [3]. 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 1A 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 60Mb with 10,000 genes and equal fractions of exons, introns, and intergenic sequence, at 20Mb each. For simplicity, all exons are the same size (in this example, 200bp), so an average gene (with ten-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), about 18% of the intronic fraction. This would yield a final annotation where exons are predicted as 16Mb (26.6% of the genome) and introns as 15.5Mb (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 1C). 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 above 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 [82] and heterozygosity [26, 42, 63, 83], repeats often result in breaks in the assembly that could split genes (Fig 2A). 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 2B).

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. 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 3A)

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

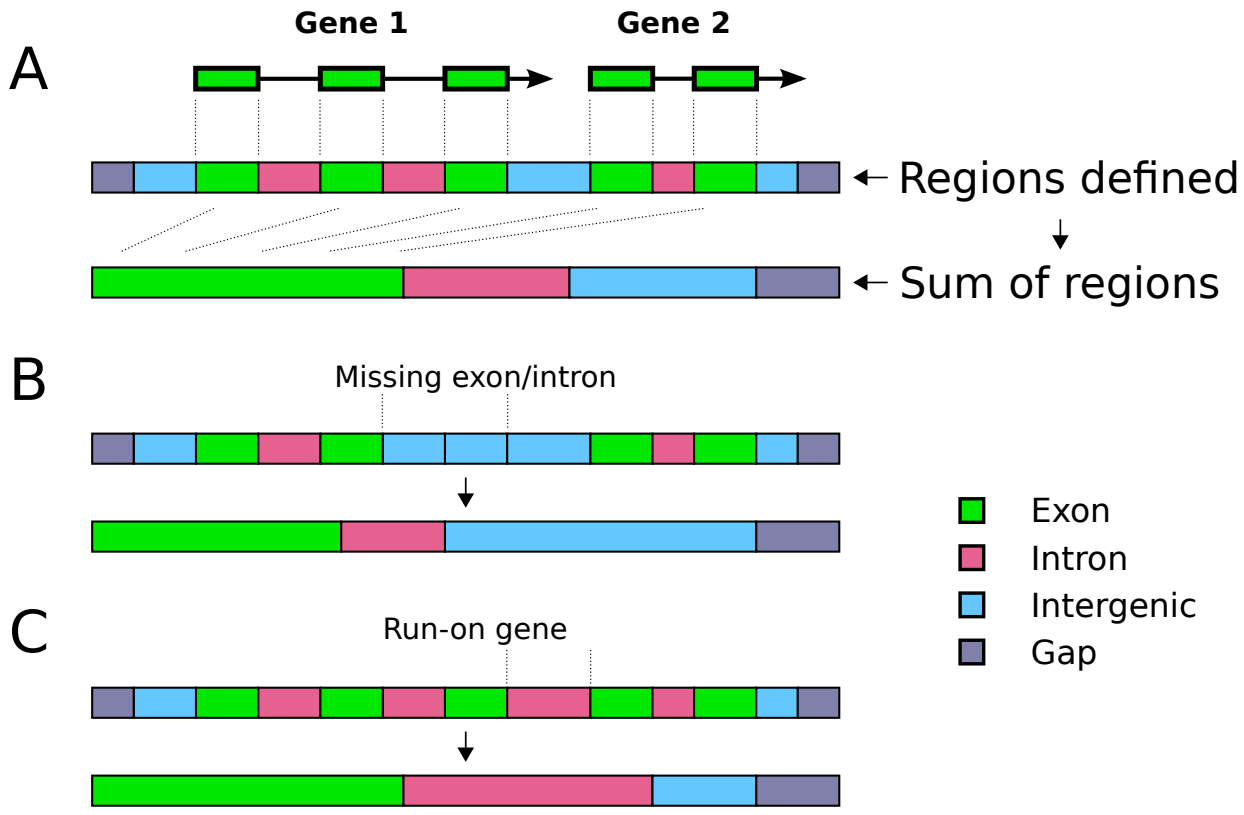

Figure 1: **Schematic of misannotations and the effects on coding fraction analyses** (A) In a normal case, two hypothetical genes on the same strand are identified. The exons and introns are defined, and the total lengths of those features are summed and displayed in the bars below. Because real genome assemblies can often contain gaps, sample gaps are also shown at the edges of the segment. (B) If individual exons (or potentially whole genes) were missing, then the measured total exons and introns would be smaller than the real values, and the ratio of intron:intergenic would decrease. (C) If neighboring genes were erroneously declared to be contiguous, the exonic fraction is mostly unchanged but the intron:intergenic ratio would increase.

(also see schematics in [83] and [84]). 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 3C). 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 3D). 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 3E) 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 3F), but also falsely indicate that splice variation is more prevalent for this gene.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

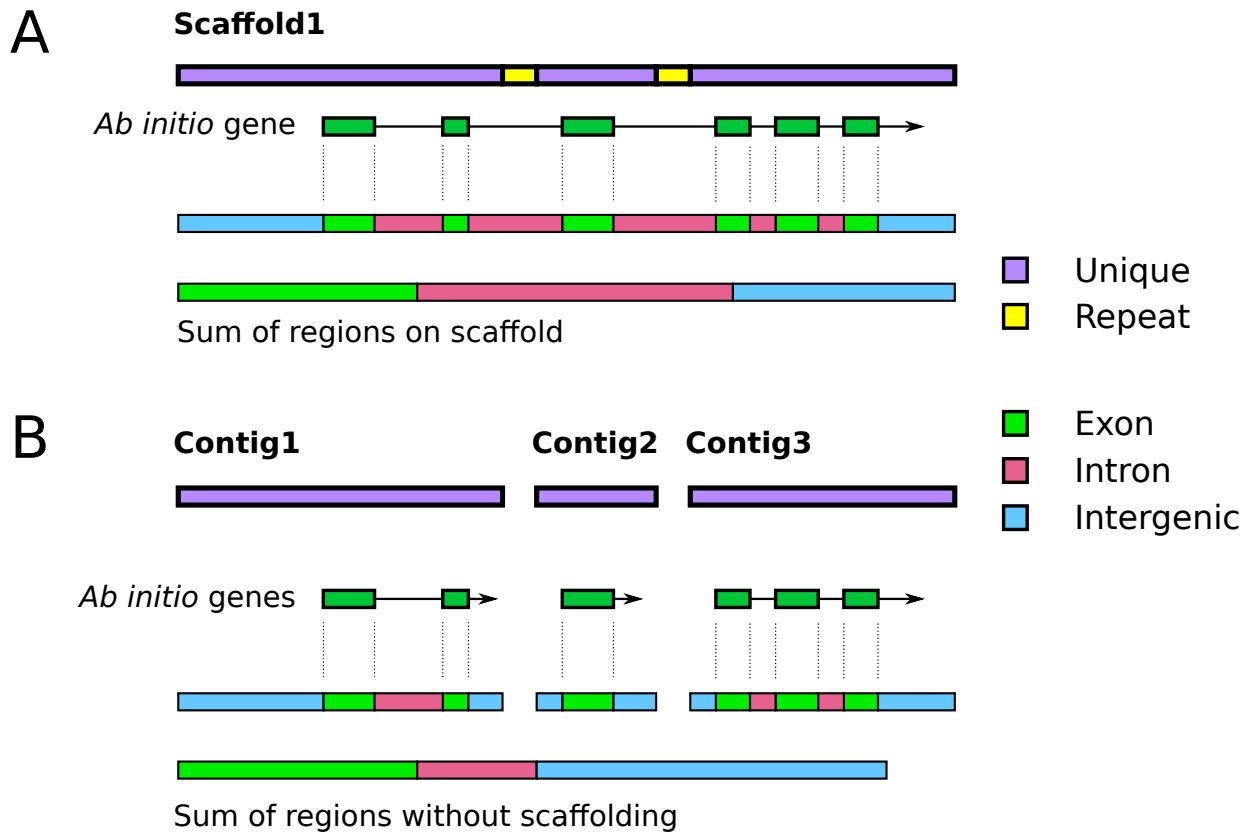

Figure 2: **Schematic of the effects of scaffolding and repeats on genic fraction analyses** (A) For a hypothetical scaffold in a genome assembly, two identical repeats are found within introns. The gene is correctly predicted to span the two repeats and the regions are define below as in Fig 1. (B) 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.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

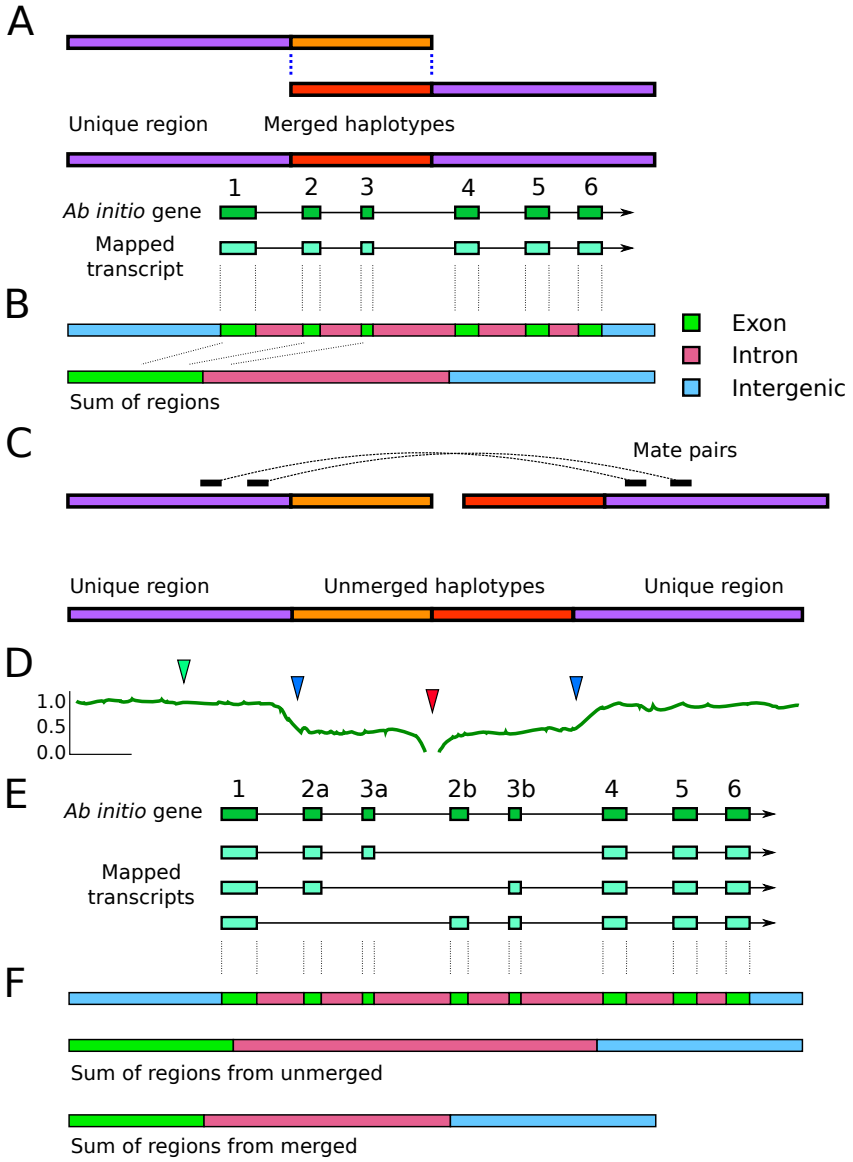

Figure 3: **Schematic of misassembly and the effects on genic fraction analyses** (A) During assembly, regions that are heterozygous (differing by SNPs or indels) are combined to make a single reference contig. When genes are predicting that this locus, or when assembled transcripts are aligned to the genome, the correct exon structure is found. (B) Regions are defined as exon, intron, or intergenic, as in Fig 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 (1x) 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 to the same genomic locus where the haplotypes were correctly merged.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 14Mb, over twofold 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.

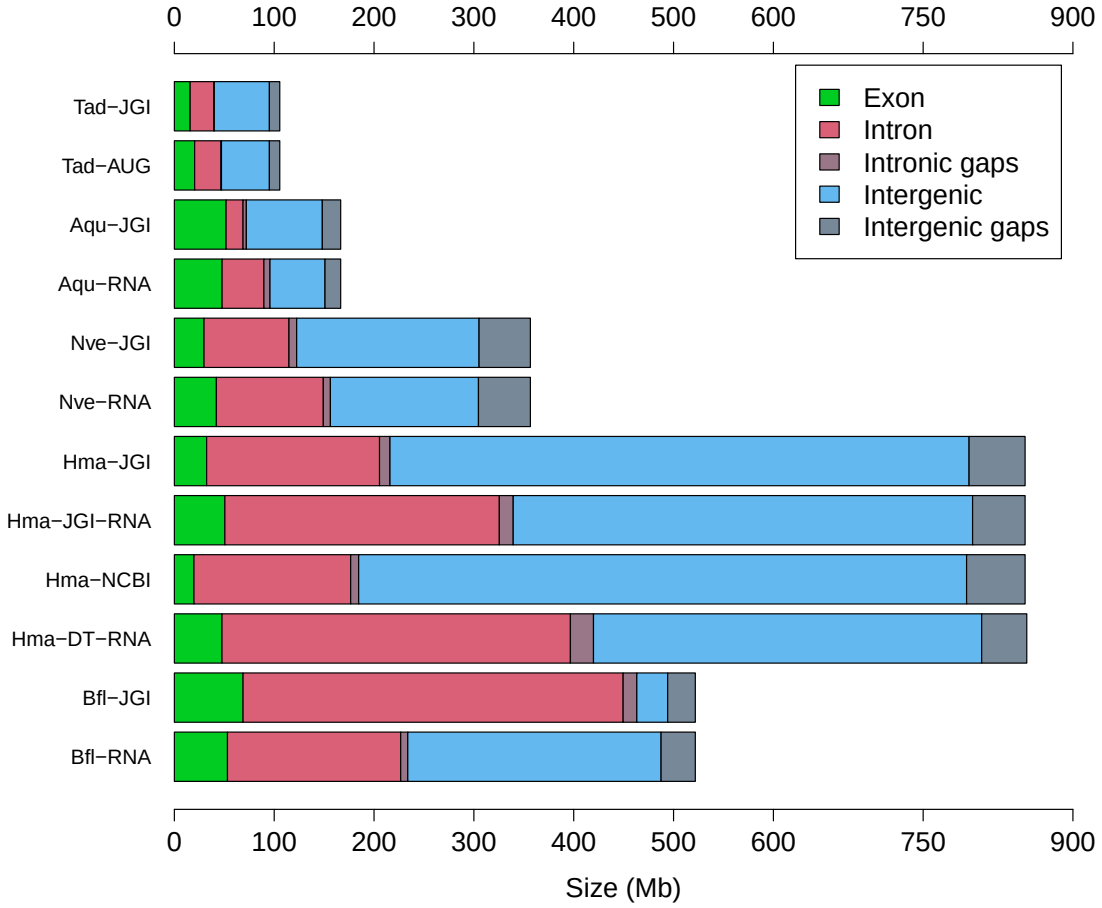

Figure 4: **Proportions of exons, introns, and intergenic sequences** Barplot showing the summed proportions of genomes composed of exons (green), introns (red) and intergenic sequences (blue). The reannotation for *O. bimaculoides* was not shown for clarity, as this genome is substantially larger than the others. Abbreviations are as follows: Tad:*T. adherens*, Aqu:*A. queenslandica*, Nve:*N. vectensis*, Hma:*H. magnipapillata*, Bfl:*B. 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*.

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

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 to the 9,196 genes in Monbr1 "best models".

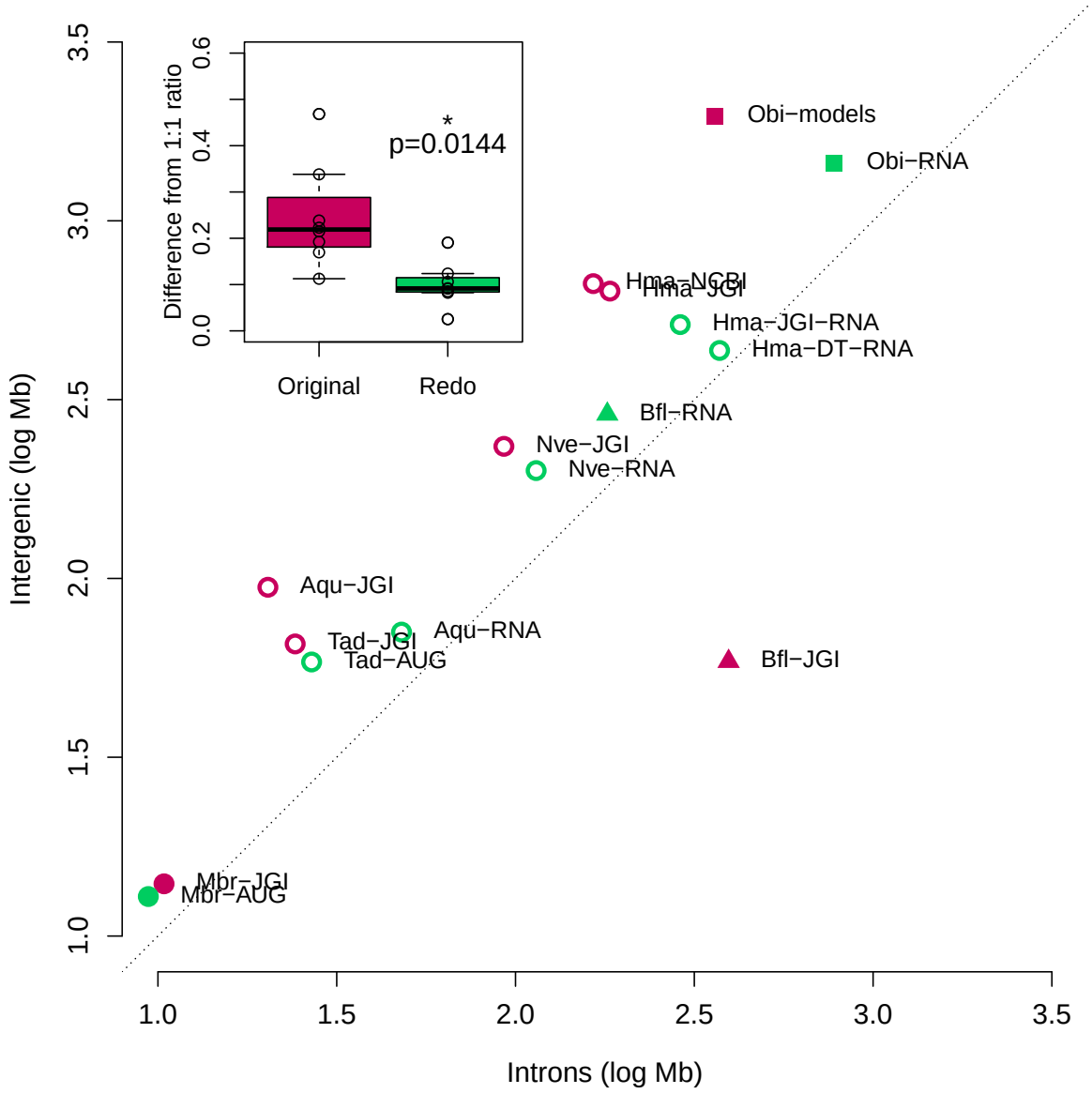

Figure 5: **Improvements from reannotation** Log-scale plot of total intronic size versus total intergenic size where original annotations from the published genomes are shown in red and reannotations are shown in green. The dotted line shows a ratio of 1:1 as a reference. Abbreviations are as in Fig 4, with the addition of Mbr:*M. brevicollis* from the original JGI annotation and the redo with AUGUSTUS, and Obi:*O. 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).

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

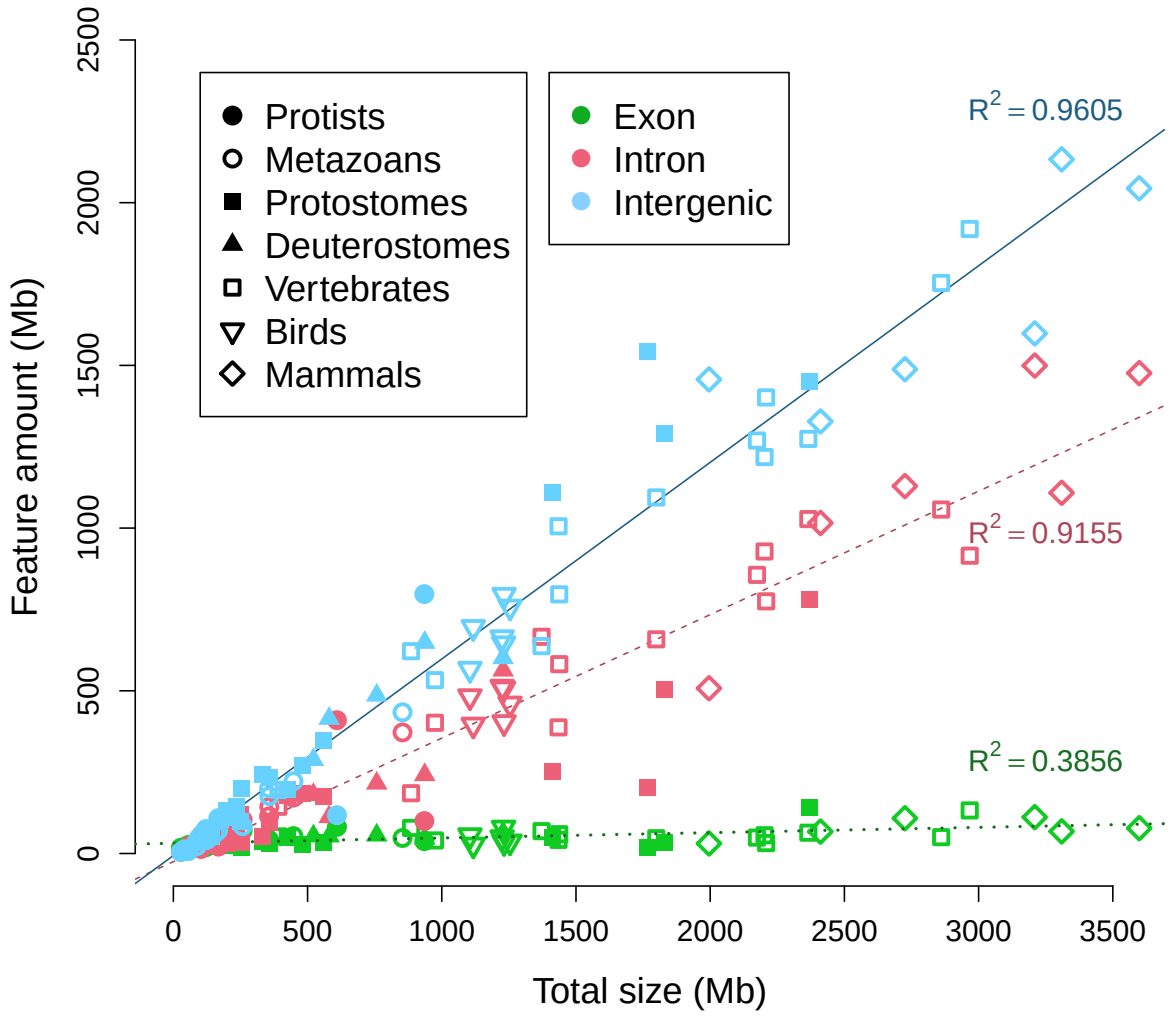

Figure 6: **Comparison of features to total genome size** The sums of exons, introns, and intergenic regions are plotted against total genome size. Linear coefficients of determination of the three features are displayed by their respective lines. For legend symbols, Deuterostomes refers to all invertebrate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

Average intron and exon length

The average length of introns linearly scales with the total genome size (Fig 7), in agreement with another study [18]. However, the average exon length is clearly constrained across animals relative to total genome size, and this may be related to interactions with nucleosomes [85]. Most species have an average exon

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

length between 200 and 300 bases (mean of 263bp), higher than values reported from previous surveys of exon length [21, 86]. 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 [21], respectively, absence of five-prime UTRs may result in an underestimation of the average exon length for that species.

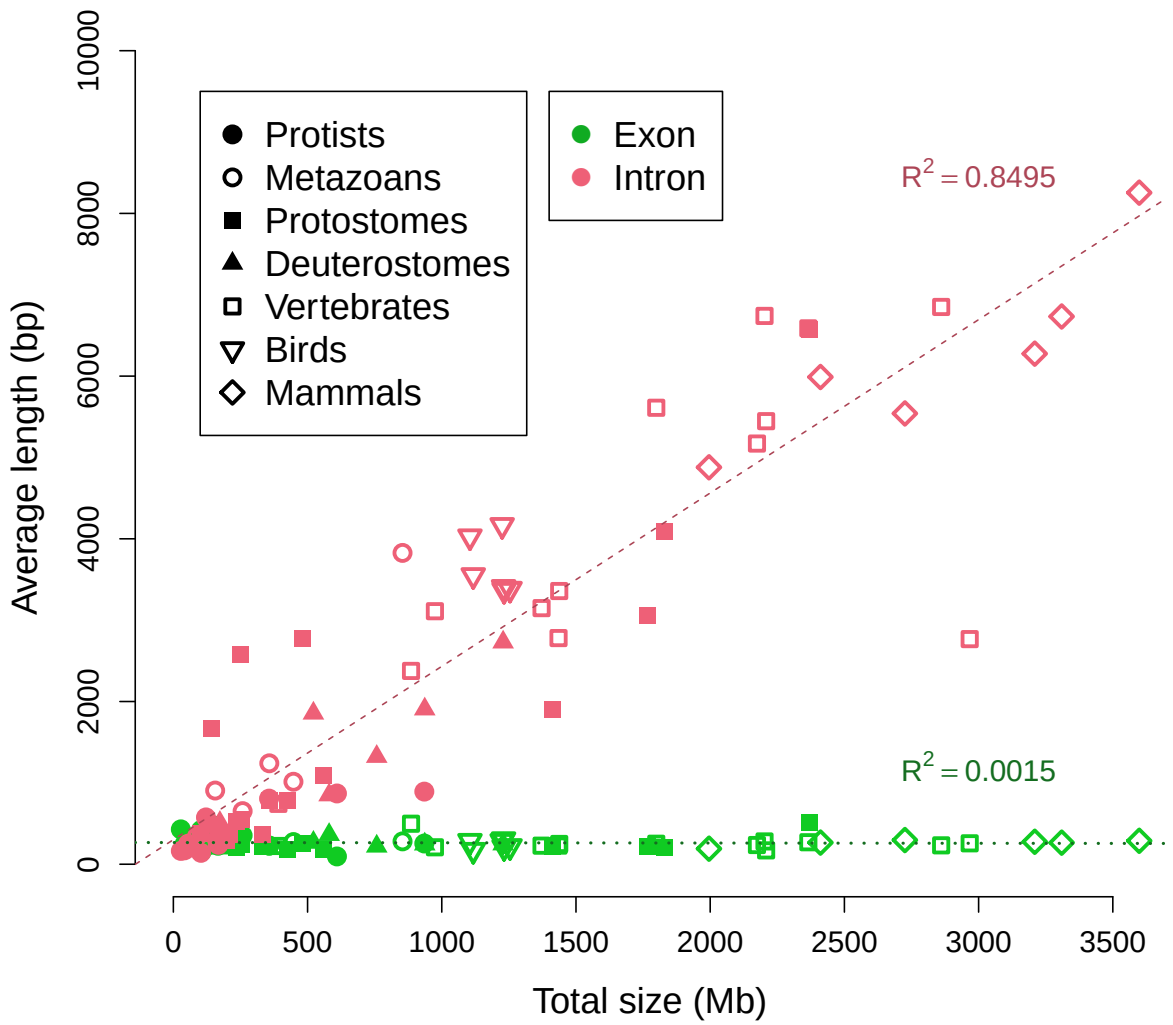

Figure 7: **Average length of exons and introns** Plot of the average length of exons (green) and introns (pink) as a function of total genome size across all species in this study. Linear coefficients of determination are displayed next to the green (dotted) and red (dashed) linear fit lines, for exons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

Nature of the exonic fraction

The total amount of exons is not strongly correlated 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 to 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

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [18]. 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.

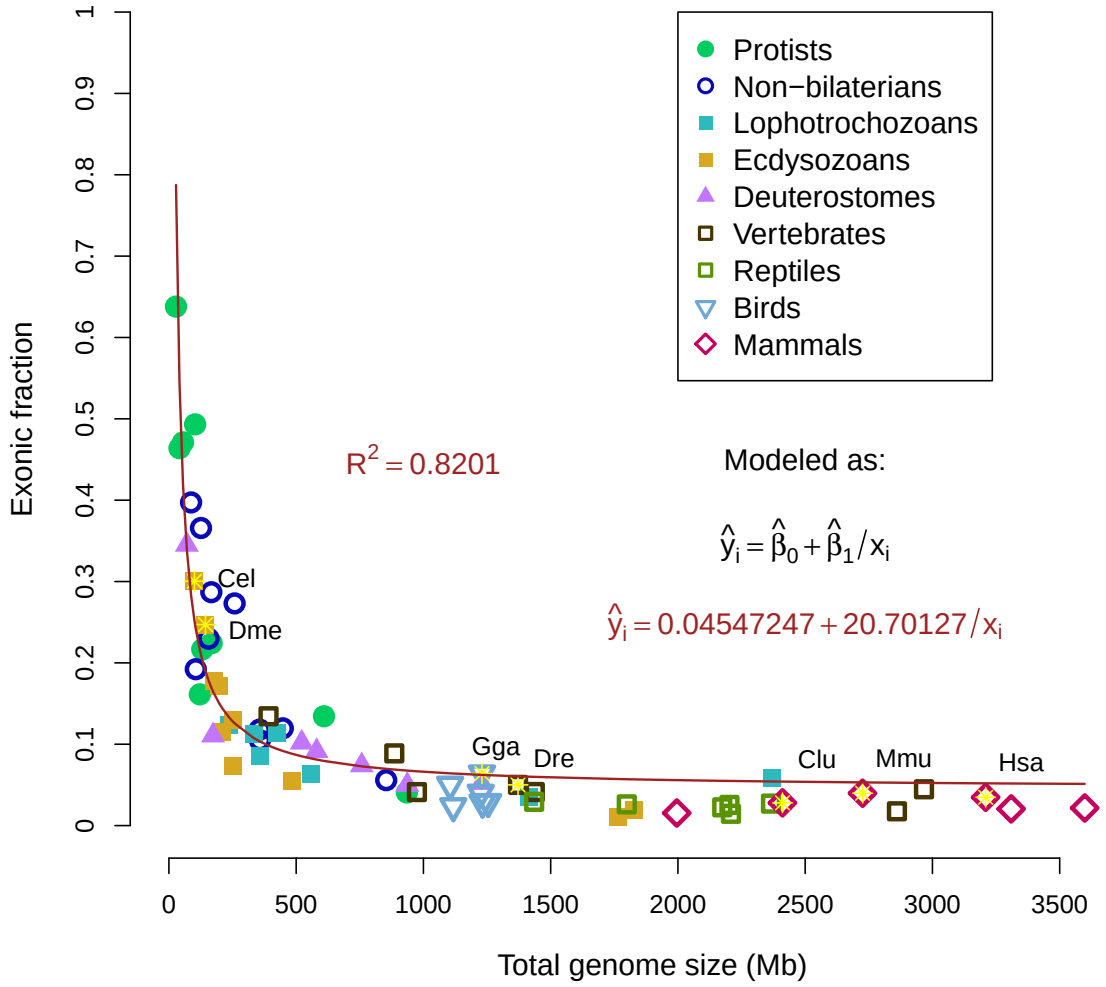

Figure 8: **Exonic fraction compared to total genome size** Relative fraction of the genome that is defined as exons compared as a function of total size. Coefficients of determination of a hyperbolic model is displayed. Seven model organisms (human, mouse, dog, chicken, zebrafish, fruit fly and nematode)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

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).

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

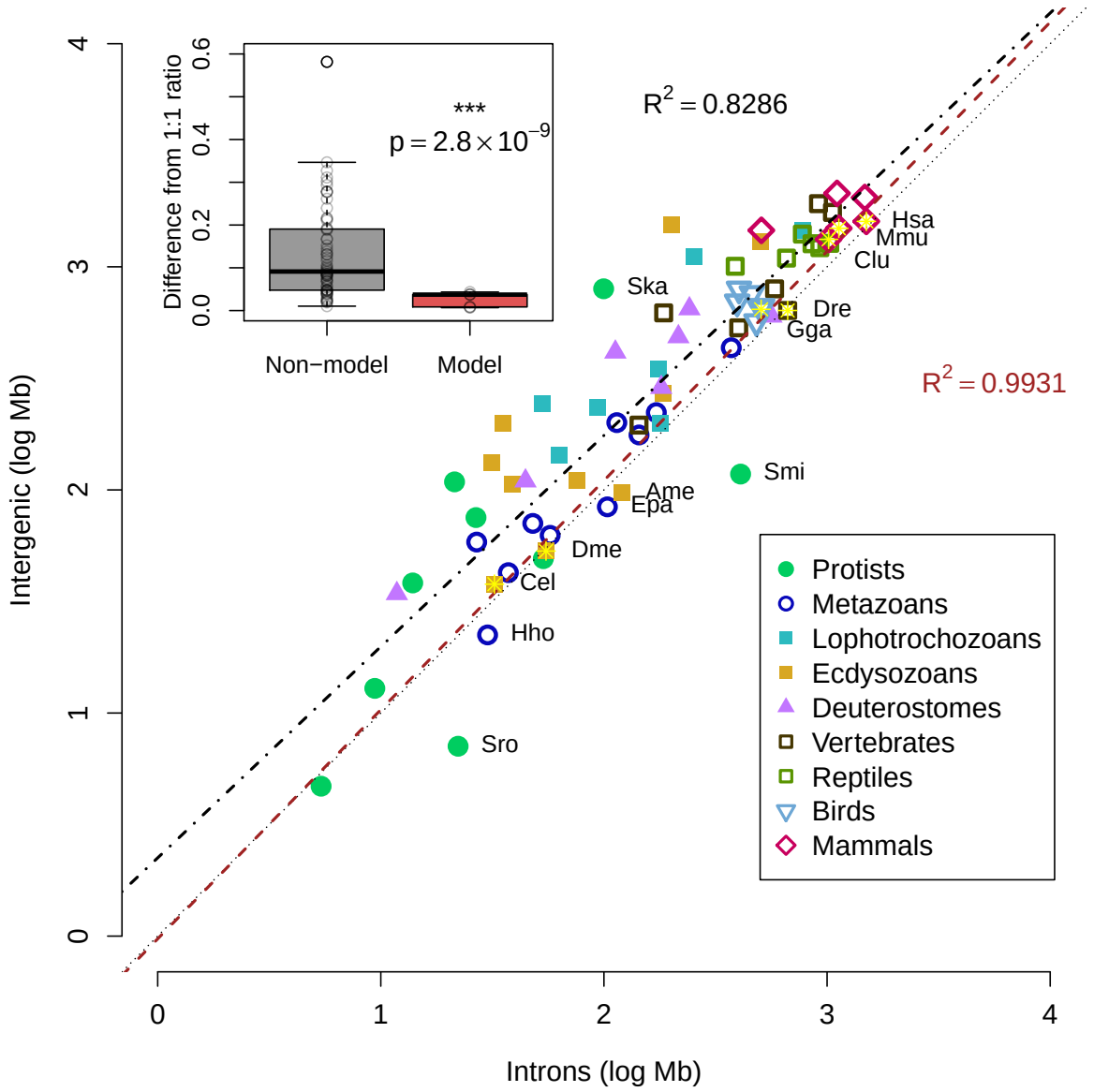

Figure 9: **Comparing intronic and intergenic fractions** Log-scale plot of total intronic size versus total intergenic size. The dotted line shows a ratio of 1:1 as a reference, although most genomes are above this line. Seven model organisms (as in Fig 8) are indicated by three-letter codes with yellow 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:*S. kawagutii*, Smi:*S. minutum*) and select species with ratios of intron:intergenic greater than 1, choanoflagellate *S. rosetta* (Sro), honeybee *A. mellifera* (Ame), anemone *E. pallida* (Epa), and placozoan *H. 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 to the rest of the genomes (paired two-end t-test, p-value: $2.8 * 10^{-9}$).

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

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fit was observed when restricted to the model organisms (R-squared: 0.9931, p-value=$1.3 * 10^{-6}$), 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 non-model organisms is compatible with the hypothesis that the predicted amount of the genome that is transcribed varies more by annotation quality than biological differences.

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 (Figure 10). For example, chordates against non-chordates 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 non-bilaterians 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 [71], 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% [87] and 1.8% (manuscript in preparation), respectively. Although these values are lower than the observed heterozygosity in many other invertebrates [88], 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 fraction that is intron 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 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 500Mb 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 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 around 50% in large genomes.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

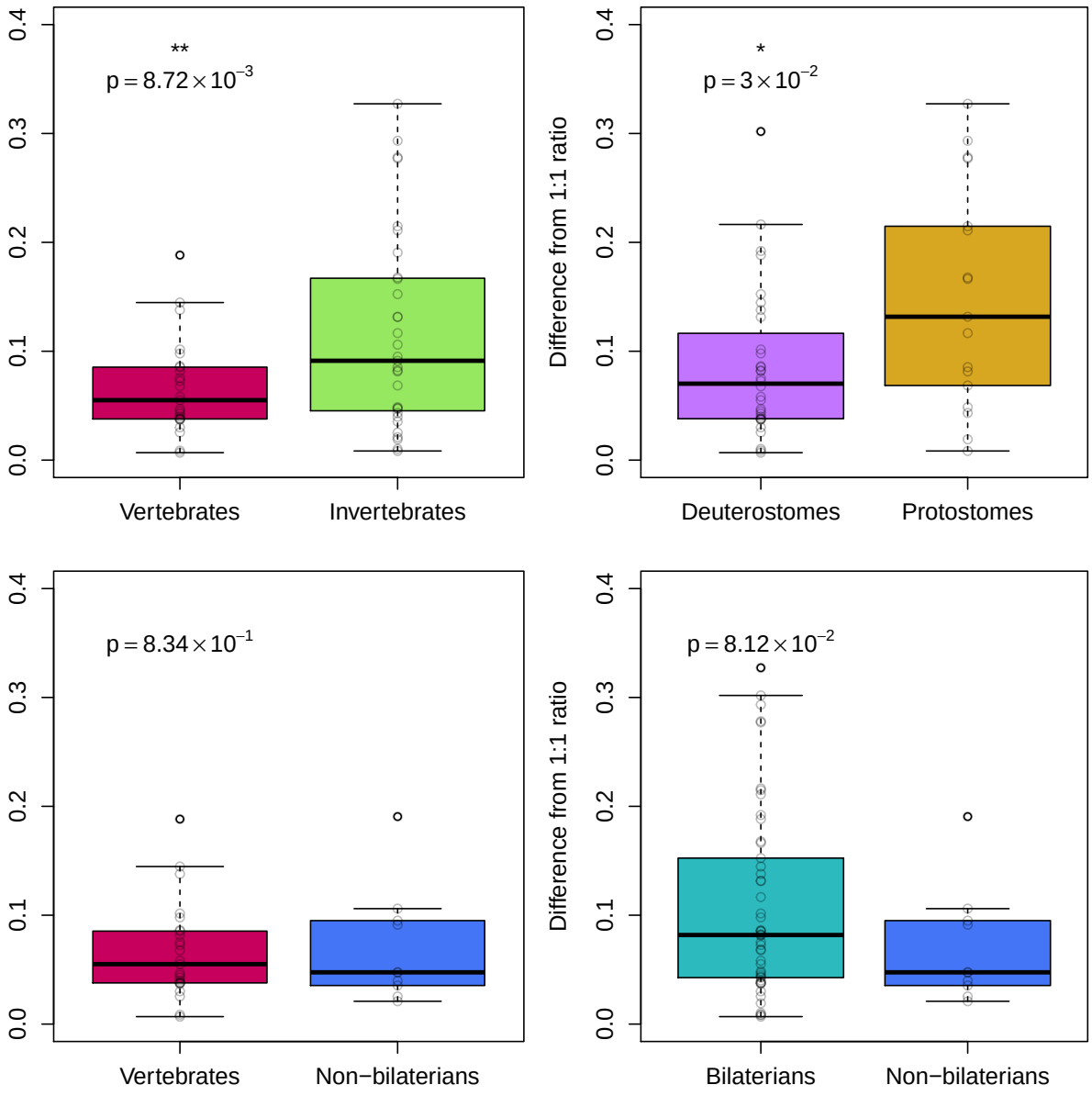

Figure 10: **Comparing intron-intergenic ratios among animal groups** Difference of the intron:intergenic ratio to 1 across four pairs of animal groups. Invertebrates (green) includes all non-bilaterian taxa. Deuterostomes and protostomes are both assumed to be monophyletic.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

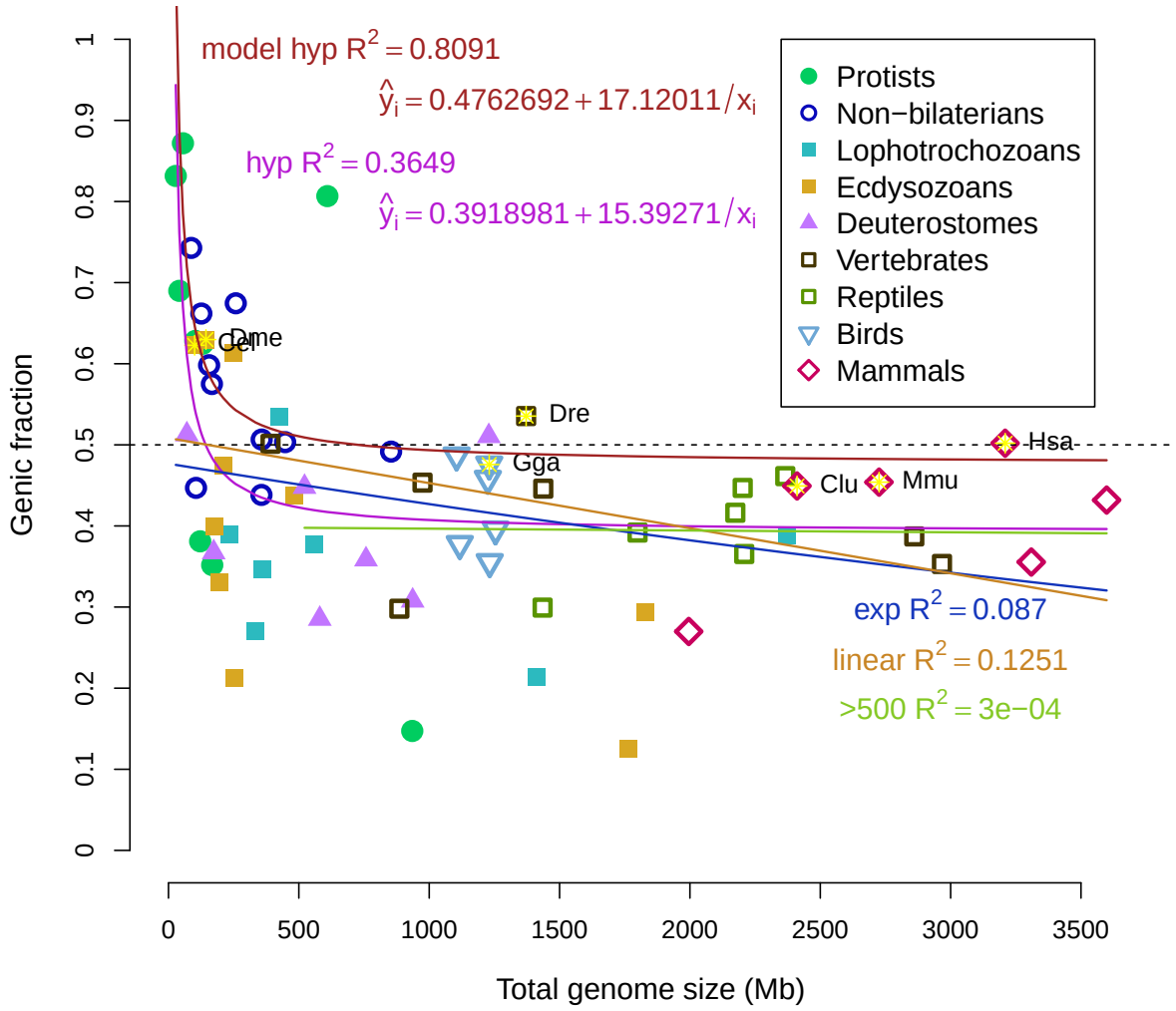

Figure 11: **Genic fraction compared to total genome size** Relative fraction of the genome that is defined as genes compared as a function of total size. A number of correlative models (hyperbolic in purple, exponential in blue, linear in orange) were tested and coefficients are displayed. Linear correlation is expected to be zero if genic and intergenic fractions "expand" indifferently after a certain size, which appears to be around 500Mb. Linear correlation including only genomes larger than 500Mb 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.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Diagnostic relationship of introns to intergenic sequence

An increasing number of genomes of any non-model organisms are sequenced to answer evolutionary questions. For example, genomes of taxa from all four non-bilaterian groups were recently sequenced to understand how similar these genomes are to humans [8,24,33,34], 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 non-model organisms is that there is little *a priori* information about gene structure or content. It 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 [23], 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 [89]. 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 [90], 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 [41]. 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* [40] 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 [91–93]. One study [18] 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

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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 9.3Mb of introns and 10.1Mb of intergenic sequence (a ratio of 0.92) compared to 19.3Mb of exons.

Therefore, model animals (and probably all animals) transcribe at least half of the genome, where species with smaller genomes (exon-rich) transcribe more than half (Figure 11). There does not appear to be a significant difference in the genic fraction based on animal group (Figure 10), that is, all animals appear to follow this rule. One study had shown that some larger metazoan genomes were depleted in genes [94], 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 50kb and found that 80% of the human genome was lacking any gene [94], 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 non-models 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 non-coding ones, we could not determine coding fractions for all species. Even for putative non-coding transcripts, some may be coding [95–97], 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 around 500Mb, after that point, introns and intergenic regions are expected to expand, on average, equally across the genome resulting in approximately 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 [98]. If the last common ancestor of all metazoans had a relatively small genome (under 100Mb, 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 Fig 9 and Fig 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 (around 500Mb), genic and intergenic sequences expand almost equally, where 50%

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 500Mb to 1.5Gb [17]. Larger genomes appeared to be flooded by transposable elements of a single type. Thus, above 500Mb, it can be predicted that select transposable elements become prevalent and multiply throughout the genome, but on 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" [1,99], 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* [100], 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" [91, 101]).

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 [102]. Vertebrates and specifically mammals tend to splice transcripts more than invertebrates (meaning models fruit fly and nematode) [103, 104]. One study reported a good correlation (R-squared of 0.80) of splicing to organismic complexity measured by cell types [105], 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 [106, 107], 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 [45]. 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 Figure 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 [105]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [92], 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 around 600Mb [108]. Thus, there may not be any "large" genomes in this phylum, and likewise for other invertebrate groups. Compared to birds, however, where the smallest genome identified to date is from the black-chinned hummingbird (estimated 910Mb) [109], 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 non-coding RNAs, or lincRNAs, presents a paradox of definitions. Non-coding 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 [110], transcription was found of intergenic regions accounting for almost another 20% of the genomes of human and mouse, depending on the analysis [111, 112]. 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 non-functional "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 [113]. 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 all animals 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 a universal feature of animal genomes, and indicates that most genomes could benefit from new annotations. Previous findings of genomic differences between animal groups are likely to result from a sampling bias, rather than biological differences. Future sequencing of more high-quality genomes from animals may reveal unanticipated sources of complexity and gene regulation with implications for the evolution of animals.

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.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Thomas CA. The Genetic Organization of Chromosomes. Annual review of genetics. 1971;5(1):237–256. doi:10.1146/annurev.ge.05.120171.001321.

[2] Han K, Li Zf, Peng R, Zhu Lp, Zhou T, Wang Lg, et al. Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Scientific reports. 2013;3:2101. doi:10.1038/srep02101.

[3] Brent MR. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nature reviews Genetics. 2008;9(1):62–73. doi:10.1038/nrg2220.

[4] Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, et al. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome biology. 2006;7 Suppl 1(Suppl 1):S2.1–31. doi:10.1186/gb-2006-7-s1-s2.

[5] Vallender EJ. Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. Methods. 2009;49(1):50–55. doi:10.1016/j.ymeth.2009.05.010.

[6] Zhang X, Goodsell J, Norgren RB. Limitations of the rhesus macaque draft genome assembly and annotation. BMC genomics. 2012;13(1):206. doi:10.1186/1471-2164-13-206.

[7] Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, et al. Standardized benchmarking in the quest for orthologs. Nature Methods. 2016;13(5):425–430. doi:10.1038/nmeth.3830.

[8] Ryan JF, Pang K, Schnitzler CE, a D Nguyen Ad, Moreland RT, Simmons DK, et al. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. Science. 2013;342(6164):1242592–1242592. doi:10.1126/science.1242592.

[9] Pisani D, Pett W, Dohrmann M, Feuda R, Rota-Stabelli O, Philippe H, et al. Genomic data do not support comb jellies as the sister group to all other animals. Proceedings of the National Academy of Sciences. 2015;112(50):201518127. doi:10.1073/pnas.1518127112.

[10] Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics (Oxford, England). 2007;23(9):1061–7. doi:10.1093/bioinformatics/btm071.

[11] Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV. BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. Genome analysis. 2015;31(June):9–10. doi:10.1093/bioinformatics/btv351.

[12] Lynch M, Conery JS. The origins of genome complexity. Science (New York, NY). 2003;302(5649):1401–4. doi:10.1126/science.1089370.

[13] Lynch M. Response to Comment on "The Origins of Genome Complexity". Science. 2004;306(5698):978b–978b. doi:10.1126/science.1100559.

[14] Daubin V, Moran Na. Comment on "The origins of genome complexity". Science (New York, NY). 2004;306(5698):978; author reply 978. doi:10.1126/science.1098469.

[15] Pettersson ME, Kurland CG, Berg OG. Deletion rate evolution and its effect on genome size and coding density. Molecular Biology and Evolution. 2009;26(6):1421–1430. doi:10.1093/molbev/msp054.

[16] Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 2002;115(1):49–63. doi:10.1023/A:1016072014259.

[17] Elliott TA, Gregory TR. Do larger genomes contain more diverse transposable elements? BMC evolutionary biology. 2015;15(1):69. doi:10.1186/s12862-015-0339-8.

[18] Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Phil Trans R Soc B. 2015;370(1678):20140331. doi:10.1098/rstb.2014.0331.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[19] Canapa A, Barucca M, Biscotti MA, Forconi M, Olmo E. Transposons, Genome Size, and Evolutionary Insights in Animals. Cytogenetic and Genome Research. 2016; p. 217–239. doi:10.1159/000444429.

[20] Deutsch M, Long M. Intron-exon structures of eukaryotic model organisms. Nucleic acids research. 1999;27(15):3219–28.

[21] Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC genomics. 2009;10(1):47. doi:10.1186/1471-2164-10-47.

[22] Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, et al. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science (New York, NY). 2002;298(5601):2157–2167. doi:10.1126/science.1080049.

[23] Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. The amphioxus genome and the evolution of the chordate karyotype. Nature. 2008;453(7198):1064–71. doi:10.1038/nature06967.

[24] Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, et al. The Trichoplax genome and the nature of placozoans. Nature. 2008;454(7207):955–60. doi:10.1038/nature07191.

[25] Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, et al. Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493(7433):526–31. doi:10.1038/nature11696.

[26] Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, et al. Hemichordate genomes and deuterostome origins. Nature. 2015; p. 1–19. doi:10.1038/nature16150.

[27] King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, et al. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature. 2008;451(7180):783–8. doi:10.1038/nature06617.

[28] Read Ba, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature. 2013; p. 9–13. doi:10.1038/nature12221.

[29] Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science (New York, NY). 2010;329(5988):223–6. doi:10.1126/science.1188800.

[30] Suga H, Chen Z, de Mendoza A, Sebé-Pedrós A, Brown MW, Kramer E, et al. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nature communications. 2013;4:2325. doi:10.1038/ncomms3325.

[31] Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, et al. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. Genome biology. 2013;14(2):R15. doi:10.1186/gb-2013-14-2-r15.

[32] Fernandez-Valverde SL, Calcino AD, Degnan BM. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge Amphimedon queenslandica. BMC Genomics. 2015;16(1):1–11. doi:10.1186/s12864-015-1588-z.

[33] Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEa, Mitros T, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. Nature. 2010;466(7307):720–6. doi:10.1038/nature09201.

[34] Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science (New York, NY). 2007;317(5834):86–94. doi:10.1126/science.1139158.

[35] Moran Y, Fredman D, Praher D, Li XZ, Wee LM, Rentzsch F, et al. Cnidarian microRNAs frequently regulate targets by cleavage. Genome Research. 2014;24(4):651–663. doi:10.1101/gr.162503.113.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[36] Fortunato SaV, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DEK, et al. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. Nature. 2014;514(7524):620–623. doi:10.1038/nature13881.

[37] Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, Koh W, et al. The genome sequence of the colonial chordate, Botryllus schlosseri. eLife. 2013;2:e00569. doi:10.7554/eLife.00569.

[38] Baumgarten S, Simakov O, Esherick LY, Liew YJ, Lehnert EM, Michell CT, et al. The genome of Aiptasia , a sea anemone model for coral symbiosis. Proceedings of the National Academy of Sciences. 2015; p. 201513318. doi:10.1073/pnas.1513318112.

[39] Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, et al. Plasticity of Animal Genome Architecture Unmasked by Rapid Evolution of a Pelagic Tunicate. Science. 2010;1381(2010). doi:10.1126/science.1194167.

[40] Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, et al. The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. Science. 2015;350(6261):691–694. doi:10.1126/science.aad0408.

[41] Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, et al. Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. Current biology : CB. 2013;23:1399–1408. doi:10.1016/j.cub.2013.05.062.

[42] Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, et al. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA research : an international journal for rapid publication of reports on genes and genomes. 2012;19(2):117–30. doi:10.1093/dnares/dss005.

[43] Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, et al. Using the Acropora digitifera genome to understand coral responses to environmental change. Nature. 2011;476(7360):320–3. doi:10.1038/nature10249.

[44] Luo YJ, Takeuchi T, Koyanagi R, Yamada L, Kanda M, Khalturina M, et al. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. Nature Communications. 2015;6:1–10. doi:10.1038/ncomms9301.

[45] Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;doi:10.1038/nature14668.

[46] Kirkness EF. The Dog Genome: Survey Sequencing and Comparative Analysis. Science. 2003;301(5641):1898–1903. doi:10.1126/science.1086432.

[47] Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, et al. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 2007;447(7141):167–77. doi:10.1038/nature05805.

[48] Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008;453(7192):175–183. doi:10.1038/nature06936.

[49] Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, et al. The genome of the Western clawed frog Xenopus tropicalis. Science (New York, NY). 2010;328(5978):633–6. doi:10.1126/science.1183670.

[50] Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346(6215):1311–1320. doi:10.1126/science.1251385.

[51] Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, et al. The genome of a songbird. Nature. 2010;464(7289):757–62. doi:10.1038/nature08819.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[52] Huang Y, Li Y, Burt DW, Chen H, Zhang Y, Qian W, et al. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. Nature genetics. 2013;45(7):776–83. doi:10.1038/ng.2657.

[53] Ganapathy G, Howard JT, Ward JM, Li J, Li B, Li Y, et al. High-coverage sequencing and annotated assemblies of the budgerigar genome. GigaScience. 2014;3:11. doi:10.1186/2047-217X-3-11.

[54] Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. Science. 2014;346(6215):1254449–1254449. doi:10.1126/science.1254449.

[55] Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 2011;477(7366):587–91. doi:10.1038/nature10390.

[56] Shaffer HB, Minx P, Warren DE, Shedlock AM, Thomson RC, Valenzuela N, et al. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome biology. 2013;14(3):R28. doi:10.1186/gb-2013-14-3-r28.

[57] Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, et al. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nature Genetics. 2013;45(6):701–706. doi:10.1038/ng.2615.

[58] Koning APJD, Hall KT, Card DC, Drew R, Fujita MK, Ruggiero RP, et al. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. Proceedings of the National Academy of Sciences. 2013;110(51):20645–20650. doi:10.1073/pnas.1324475110.

[59] Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496(7446):498–503. doi:10.1038/nature12111.

[60] Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, et al. The African coelacanth genome provides insights into tetrapod evolution. Nature. 2013;496(7445):311–316. doi:10.1038/nature12027.

[61] Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, et al. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nature genetics. 2013;45(4):415–21, 421e1–2. doi:10.1038/ng.2568.

[62] Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, et al. Elephant shark genome provides unique insights into gnathostome evolution. Nature. 2014;505(7482):174–179. doi:10.1038/nature12826.

[63] Zhang GG, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490(7418):49–54. doi:10.1038/nature11413.

[64] Keeling CI, Yuen MM, Liao NY, Roderick Docking T, Chan SK, Taylor Ga, et al. Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. Genome biology. 2013;14(3):R27. doi:10.1186/gb-2013-14-3-r27.

[65] Richards S, Gibbs Ra, Weinstock GM, Brown SJ, Denell RE, Beeman RW, et al. The genome of the model beetle and pest Tribolium castaneum. Nature. 2008;452(7190):949–55. doi:10.1038/nature06784.

[66] Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, et al. The genome sequence of silkworm, Bombyx mori. DNA research. 2004;11:27–35.

[67] Nossa CW, Havlak P, Yue JX, Lv J, Vincent KY, Brockmann HJ, et al. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. GigaScience. 2014;3:9. doi:10.1186/2047-217X-3-9.

[68] The C elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 1998;282(5396):2012–2018. doi:10.1126/science.282.5396.2012.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[69] Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, et al. The genome of the sea urchin Strongylocentrotus purpuratus. Science (New York, NY). 2006;314(5801):941–52. doi:10.1126/science.1133609.

[70] Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, et al. The ecoresponsive genome of Daphnia pulex. Science (New York, NY). 2011;331(6017):555–61. doi:10.1126/science.1197761.

[71] Weinstock GM, Robinson GE, Gibbs Ra, Worley KC, Evans JD, Maleszka R, et al. Insights into social insects from the genome of the honeybee Apis mellifera. Nature. 2006;443(7114):931–949. doi:10.1038/nature05260.

[72] Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, et al. Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nature Communications. 2016;7(May 2015):10507. doi:10.1038/ncomms10507.

[73] Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, et al. The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. PLoS Biology. 2014;12(11). doi:10.1371/journal.pbio.1002005.

[74] Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–644. doi:10.1093/bioinformatics/btn013.

[75] Hoff KJ, Stanke M. WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Research. 2013;41(W1):W123–W128. doi:10.1093/nar/gkt418.

[76] Chapman Ja, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, et al. The dynamic genome of Hydra. Nature. 2010;464(7288):592–6. doi:10.1038/nature08830.

[77] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36. doi:10.1186/gb-2013-14-4-r36.

[78] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;33(3). doi:10.1038/nbt.3122.

[79] Preußer C, Jaé N, Bindereif A. MRNA splicing in trypanosomes. International Journal of Medical Microbiology. 2012;302(4-5):221–224. doi:10.1016/j.ijmm.2012.07.004.

[80] Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GAM. Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. Nucleic Acids Research. 2010;38(15):4946–4957. doi:10.1093/nar/gkq237.

[81] Hardie DC, Hebert PD. Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences. 2004;61(9):1636–1646. doi:10.1139/f04-106.

[82] Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. Nature Reviews Genetics. 2012;13(1):36–46. doi:10.1038/nrg3117.

[83] Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research. 2014;24(8):1384–1395. doi:10.1101/gr.170720.113.

[84] Bankevich A, Nurk S, Antipov D, Gurevich Aa, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19(5):455–477. doi:10.1089/cmb.2012.0021.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[85] Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcárcel J, et al. Nucleosome positioning as a determinant of exon recognition. Nature structural & molecular biology. 2009;16(9):996–1001. doi:10.1038/nsmb.1658.

[86] Sakharkar MK, Chow VTK, Kangueane P. Distributions of exons and introns in the human genome. In silico biology. 2004;4(4):387–93.

[87] Bellis ES, Howe DK, Denver DR. Genome-wide polymorphism and signatures of selection in the symbiotic sea anemone Aiptasia. BMC Genomics. 2016;17:160. doi:10.1186/s12864-016-2488-6.

[88] Leffler EM, Bullaughey K, Matute DR, Meyer WK, S??gurel L, Venkat A, et al. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? PLoS Biology. 2012;10(9). doi:10.1371/journal.pbio.1001388.

[89] Bányai L, Patthy L. Putative extremely high rate of proteome innovation in lancelets might be explained by high rate of gene prediction errors. Scientific Reports. 2016;6(April):30700. doi:10.1038/srep30700.

[90] Zmasek CM, Godzik A. This Déjà Vu Feeling-Analysis of Multidomain Protein Evolution in Eukaryotic Genomes. PLoS Computational Biology. 2012;8(11). doi:10.1371/journal.pcbi.1002701.

[91] Hahn MW, Wray GA. The g-value paradox. Evolution and Development. 2002;4(2):73–75. doi:10.1046/j.1525-142X.2002.01069.x.

[92] Gregory TR. Synergy between sequence and size in large-scale genomics. Nature reviews Genetics. 2005;6(9):699–708. doi:10.1038/nrg1674.

[93] Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Computational Biology. 2014;10(12). doi:10.1371/journal.pcbi.1003998.

[94] Fernandez-Valverde SL, Degnan BM. Bilaterian-like promoters in the highly compact Amphimedon queenslandica genome. Scientific Reports. 2016;6(February):22496. doi:10.1038/srep22496.

[95] Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. Genome biology and evolution. 2011;3:1245–52. doi:10.1093/gbe/evr099.

[96] Slavoff Sa, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nature chemical biology. 2012;9(1):59–64. doi:10.1038/nchembio.1120.

[97] Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013;154(1):240–251. doi:10.1016/j.cell.2013.06.009.

[98] Petrov D. Mutational Equilibrium Model of Genome Size Evolution. Theoretical Population Biology. 2002;61(4):531–544. doi:10.1006/tpbi.2002.1605.

[99] Moore G. The C-Value Paradox. BioScience. 1984;34(7):425–429. doi:10.2307/1309631.

[100] Thiébaud CH, Fischberg M. DNA content in the genus Xenopus. Chromosoma. 1977;59(3):253–7.

[101] Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome biology. 2011;12(12):R120. doi:10.1186/gb-2011-12-12-r120.

[102] Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463(January). doi:10.1038/nature08909.

[103] Brett D, Pospisil H, Valcárcel J, Reich J, Bork P. Alternative splicing and genome complexity. Nature genetics. 2002;30(1):29–30. doi:10.1038/ng803.

bioRxiv preprint doi: https://doi.org/10.1101/068627; this version posted January 2, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[104] Kim H, Klein R, Majewski J, Ott J. Estimating rates of alternative splicing in mammals and invertebrates. Nature genetics. 2004;36(9):915–6; author reply 916–7. doi:10.1038/ng0904-915.

[105] Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO. Correcting for differential transcript coverage reveals a strong relationship between alternative splicing and organism complexity. Molecular biology and evolution. 2014;31(6):1402–13. doi:10.1093/molbev/msu083.

[106] Kim E, Magen A, Ast G. Different levels of alternative splicing among eukaryotes. Nucleic Acids Research. 2007;35(1):125–131. doi:10.1093/nar/gkl924.

[107] Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genetics. 2010;6(12):1–11. doi:10.1371/journal.pgen.1001236.

[108] Jeffery NW, Jardine CB, Gregory TR. A first exploration of genome size diversity in sponges. Genome. 2013;56(8):451–6. doi:10.1139/gen-2012-0122.

[109] Gregory TR, Andrews CB, McGuire JA, Witt CC. The smallest avian genomes are found in hummingbirds. Proceedings Biological sciences / The Royal Society. 2009;276(1674):3753–3757. doi:10.1098/rspb.2009.1004.

[110] Consortium ME. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515(7527):355–364. doi:10.1038/nature13992.

[111] van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most "dark matter" transcripts are associated with known genes. PLoS Biology. 2010;8(5). doi:10.1371/journal.pbio.1000371.

[112] Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. PLoS Biology. 2011;9(7):5–10. doi:10.1371/journal.pbio.1000625.

[113] Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. Megabase deletions of gene deserts result in viable mice. Nature. 2004;431(7011):988–93. doi:10.1038/nature03022.