


RESEARCH

Open Access



# Complete vertebrate mitogenomes reveal widespread repeats and gene duplications

Giulio Formenti<sup>1,2,3\*</sup> , Arang Rhie<sup>4</sup>, Jennifer Balacco<sup>1</sup>, Bettina Haase<sup>1</sup>, Jacquelyn Mountcastle<sup>1</sup>, Olivier Fedrigo<sup>1</sup>, Samara Brown<sup>2,3</sup>, Marco Rosario Capodiferro<sup>5</sup>, Farooq O. Al-Ajli<sup>6,7,8</sup>, Roberto Ambrosini<sup>9</sup>, Peter Houde<sup>10</sup>, Sergey Koren<sup>4</sup>, Karen Oliver<sup>11</sup>, Michelle Smith<sup>11</sup>, Jason Skelton<sup>11</sup>, Emma Betteridge<sup>11</sup>, Jale Dolucan<sup>11</sup>, Craig Corton<sup>11</sup>, Iliana Bista<sup>11,12</sup>, James Torrance<sup>11</sup>, Alan Tracey<sup>11</sup>, Jonathan Wood<sup>11</sup>, Marcela Uliano-Silva<sup>11</sup>, Kerstin Howe<sup>11</sup>, Shane McCarthy<sup>11,12</sup>, Sylke Winkler<sup>13</sup>, Woori Kwak<sup>14</sup>, Jonas Korlach<sup>15</sup>, Arkarachai Fungtammasan<sup>16</sup>, Daniel Fordham<sup>17</sup>, Vania Costa<sup>17</sup>, Simon Mayes<sup>17</sup>, Matteo Chiara<sup>18</sup>, David S. Horner<sup>18</sup>, Eugene Myers<sup>13</sup>, Richard Durbin<sup>11,12</sup>, Alessandro Achilli<sup>5</sup>, Edward L. Braun<sup>19</sup>, Adam M. Phillippy<sup>4</sup>, Erich D. Jarvis<sup>1,2,3\*</sup> and The Vertebrate Genomes Project Consortium

\* Correspondence: [gformenti@rockefeller.edu](mailto:gformenti@rockefeller.edu); [ejarvis@rockefeller.edu](mailto:ejarvis@rockefeller.edu)

<sup>1</sup>The Vertebrate Genome Lab, Rockefeller University, New York, NY, USA

Full list of author information is available at the end of the article

## Abstract

**Background:** Modern sequencing technologies should make the assembly of the relatively small mitochondrial genomes an easy undertaking. However, few tools exist that address mitochondrial assembly directly.

**Results:** As part of the Vertebrate Genomes Project (VGP) we develop mitoVGP, a fully automated pipeline for similarity-based identification of mitochondrial reads and de novo assembly of mitochondrial genomes that incorporates both long (> 10 kbp, PacBio or Nanopore) and short (100–300 bp, Illumina) reads. Our pipeline leads to successful complete mitogenome assemblies of 100 vertebrate species of the VGP. We observe that tissue type and library size selection have considerable impact on mitogenome sequencing and assembly. Comparing our assemblies to purportedly complete reference mitogenomes based on short-read sequencing, we identify errors, missing sequences, and incomplete genes in those references, particularly in repetitive regions. Our assemblies also identify novel gene region duplications. The presence of repeats and duplications in over half of the species herein assembled indicates that their occurrence is a principle of mitochondrial structure rather than an exception, shedding new light on mitochondrial genome evolution and organization.

**Conclusions:** Our results indicate that even in the “simple” case of vertebrate mitogenomes the completeness of many currently available reference sequences can be further improved, and caution should be exercised before claiming the complete assembly of a mitogenome, particularly from short reads alone.

**Keywords:** Mitochondrial DNA, Vertebrate, Assembly, Long reads, Sequencing, Duplications, Repeats



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Mitochondria are found in the vast majority of eukaryotic cells [1]. The mitochondrial DNA (mtDNA) can be circular, as in animals, or linear, as in many plant species [2]. In animals, different cell types have varying numbers of mitochondria [3], normally hundreds or thousands, with each mitochondrion usually harboring 1–10 mtDNA copies [4]. In vertebrates, mtDNA varies from 14 to over 20 kbp in size, and albeit gene order can vary [5, 6], its gene content is highly conserved [2]. It usually contains 37 genes, encoding for 2 ribosomal RNAs (rRNAs), 13 proteins, and 22 transfer RNAs (tRNAs). This “mitogenome” generally has short repetitive non-coding sequences, normally within a single control region (CR). However, relatively large repetitive regions, potentially heteroplasmic, have also been reported whose biological significance is still unclear [6, 7].

To date, mtDNA sequences have been generated for hundreds of thousands of specimens in many vertebrate species [8]. With some of its genes having been considered as the universal barcode of metazoa [9], mtDNA is routinely employed at both the population and species levels in phylogeographic [10, 11], phylogenetic [12, 13], and paleogenomics studies [14, 15], among others [16, 17]. The maternal inheritance pattern of mitochondria in vertebrates provides key complementary information to the nuclear DNA (nDNA), helping to reconstruct single maternal lineages without the confounding effects of recombination. The higher mutation rate of mtDNA compared to nDNA [16], coupled with variable levels of conservation of mtDNA regions, can be used to delineate different types of phylogenetic relationships among species [18]. Key mtDNA molecular markers that have been used since the dawn of genetics include cytochrome b (*MT-CYB*), cytochrome c oxidase subunit I (*MT-COI*), NADH dehydrogenase subunit 6 (*MT-ND6*), and 16S rRNA [19–21]. Thanks to its fast evolutionary rate, highly polymorphic nature, the non-coding CR has also been used in both present and ancient DNA studies [22] to resolve the phylogenetic history of closely related species [23, 24].

At present, variation in the mtDNA sequence is usually assessed in two ways: (1) by target-enrichment and sequencing [25, 26] and (2) by de novo assembly from whole-genome sequencing (WGS) with short reads [13]. In both scenarios, overlaps between sequences (Sanger or next-generation sequencing, NGS) have been used to assemble full-length mitogenomes. Despite the quantity and general high quality of mitogenome reconstructions allowed by these methods, complex regions, particularly repetitive regions and segmental duplications such as those sometimes present in the CR [27], have been traditionally challenging to resolve [24, 28]. Theoretically, whenever repeats longer than the reads are present, assemblies are limited within the boundaries of repetitive elements. Another important challenge for mitogenome assembly is posed by the nDNA of mitochondrial origin (NUMT) [29, 30]. NUMTs originate from the partial or even near complete transposition of the mtDNA into the nDNA, potentially leading to multiple copies of the mtDNA sequence scattered throughout it and independently evolving [31, 32]. When probes, PCR primers, or in silico baits designed to match the mtDNA share sequence similarity with NUMTs, they can lead to off-target hybridization and amplification, resulting in the incorrect incorporation of NUMT sequence variation in mitogenome assemblies, ultimately impacting evolutionary analysis [29].

In the last 5 years, novel WGS strategies based on single molecule, long-read sequencing technologies have been successful in improving the quality of the nuclear genome

assemblies, particularly in repetitive regions [33]. Since single-molecule DNA sequencing can currently produce reads of at least 10–20 kbp in size [33], a complete full-length representation of mtDNA genomes can theoretically be obtained in a single read, solving the overlap uncertainty issue of short-read NGS and Sanger-based approaches. However, since single-molecule sequencing technologies have a relatively high error rate, assembly is still required to derive an accurate consensus sequence. Here, we describe a new mitochondrial genome assembly pipeline (mitoVGP) developed in the framework of the Vertebrate Genomes Project (VGP) ([www.vertebrategenomesproject.org](http://www.vertebrategenomesproject.org)) [34, 35]. The VGP aims to generate near complete and error-free reference genome assemblies representing all vertebrate species, and in its Phase 1, it is currently targeting one species for each vertebrate order. Our mitoVGP pipeline complements the nuclear assembly VGP pipeline [35], which has become a standard approach for many species, by combining long reads for structural accuracy and short reads for base calling accuracy. We applied mitoVGP to determine the full mitogenome sequence of 100 vertebrate species, including 33 species for which a reference mtDNA sequence was not previously available. Compared to the published reference mitogenome assemblies, our assemblies filled gaps above 250 bp in size in 25% of cases, added missing repeats, and added genes or gene duplications, all of which led to novel discoveries using these complete assemblies.

## Results

The VGP version 1 assembly pipeline developed for the nuclear genome uses Continuous Long Reads (CLR) and the Pacific Biosciences (PacBio) assembler FALCON to generate contigs [35, 36]. When initially inspecting the contigs, we noted the absence of contigs representing the mitogenome in 75% of species. However, mitogenome sequences could be found in the raw reads. We surmised that the mitogenome reads may have been filtered out due to size cutoffs or depth of coverage using nDNA assembly algorithms [37], reducing the chances for the mitogenome to be represented in the final assembly. Moreover, we found that when present in the final assembly, mtDNA contigs consist of long concatemers of the mtDNA sequence, due to the circular nature of the mitogenome [38]. This issue motivated the development of mitoVGP, a novel bioinformatics pipeline specifically designed to obtain complete and error-free mitogenomes for all vertebrate species. MitoVGP uses bait-reads to fish out the mitogenome long reads from WGS data, assembles complete gapless mitogenome contigs, and polishes for base accuracy using short reads (Additional file 1: Fig. S1; workflow described in the “Methods”).

We evaluated the mitoVGP pipeline using paired PacBio long read and 10x Genomics WGS linked read datasets comprising 125 VGP vertebrate species belonging to 90 families and 59 orders (Additional file 2: Table S1). MitoVGP was able to generate mitogenome assemblies, with no gaps, in 100 cases (80%). The success rate was higher in some groups, such as mammals and amphibians, than in others, such as birds and fishes, albeit these lineage differences did not reach statistical significance (Additional file 3: Table S2; Fisher’s exact test,  $p = 0.12$ ,  $N = 125$ ). To understand if this difference was due to technical issues in the assembly pipeline or intrinsic properties of the raw data, we tested whether the assembly success was associated with the availability of long mtDNA reads. We found that in all cases where long mtDNA reads were

available the assembly was successful, while it failed in all cases where no long mtDNA reads were available ( $\chi^2 = 118.8$ ,  $df = 1$ ,  $p < 2.2 \times 10^{-16}$ , Additional file 1: Fig. S2).

A series of factors can theoretically affect the relative abundance of mtDNA reads, including taxonomic differences, tissue type, DNA extraction method, and library preparation protocols, particularly key steps such as size selection. We therefore fitted a linear model including all these factors (“Methods,” Additional file 4: Table S3). We found that the availability of mtDNA reads, and thus assembly success (Additional file 5: Table S4), varied significantly by *tissue type* (Additional file 1: Fig. S3), which explained the largest fraction of the total variance (25.3%). Mitogenome assembly using DNA derived from muscle was successful in 100% of the cases ( $N = 31$ ). By contrast, blood ( $N = 36$ ) and liver ( $N = 19$ ) were used successfully only in 63.9% and 63.2% of the cases, respectively. The fraction of the variance in the number of mtDNA reads explained by *taxonomic group* (12.5%) was also significant, potentially as a consequence of real biological differences in mitochondria copy number among cells of different lineages. However, we noted that in our dataset *tissue type* and *taxonomic group* covaried, as seen in the preferential usage of some tissue types for specific taxonomic groups, such as nucleated blood for birds (Additional file 6: Table S5,  $\chi^2 = 302.8$ , simulated  $p = 9.999 \times 10^{-5}$ ). DNA extraction protocols (explained variance 1.2%) and library prep kits (1.1%) were not significantly associated with mtDNA read availability, but the library fragmentation approach (explained variance 6.9%) and size selection (11.7%) had significant effects. In particular, fragmentation with Megaruptor significantly preserved more mtDNA fragments than needle shearing ( $t = -2.767$ ,  $p = 0.018$ ) or no fragmentation ( $t = -3.387$ ,  $p = 0.003$ ), potentially because Megaruptor fragmentation yields a tight fragment size distribution whereas needles may over-shear the DNA and samples with low DNA integrity skipped fragmentation. While the relationship between the size selection cutoff used for the library preparation and assembly success is not monotonic, a significantly higher proportion of complete mitochondrial genomes were reconstructed from libraries with a size cutoff less than 20 kbp ( $\chi^2 = 16.6$ , simulated  $p = 3 \times 10^{-4}$ , Additional file 7: Table S6). This is consistent with the mitogenome size usually being below 20 kbp. Library preparation protocols involving a higher cutoff would deplete the library of mtDNA. We also considered whether total sequence data impacted the availability of mtDNA reads. VGP datasets were all generated at approximately 60X CLR coverage of the estimated nuclear genome size. Coverage being equal, larger genomes (e.g., those of some amphibians > 4 Gbp) would generate considerably more raw data than smaller genomes (e.g., birds ~ 1 Gbp). However, theoretically (see also Additional file 8: Supplementary Note 1) and in practice (explained variance 0.3%), this did not translate into a substantial increase in the number of mtDNA reads.

Overall, the major role of read availability in assembly success, driven by the factors described above, supports the robustness and the unbiased nature of the mitoVGP pipeline in generating mitogenome assemblies in any genomic context.

#### **MitoVGP assemblies are more accurate and complete**

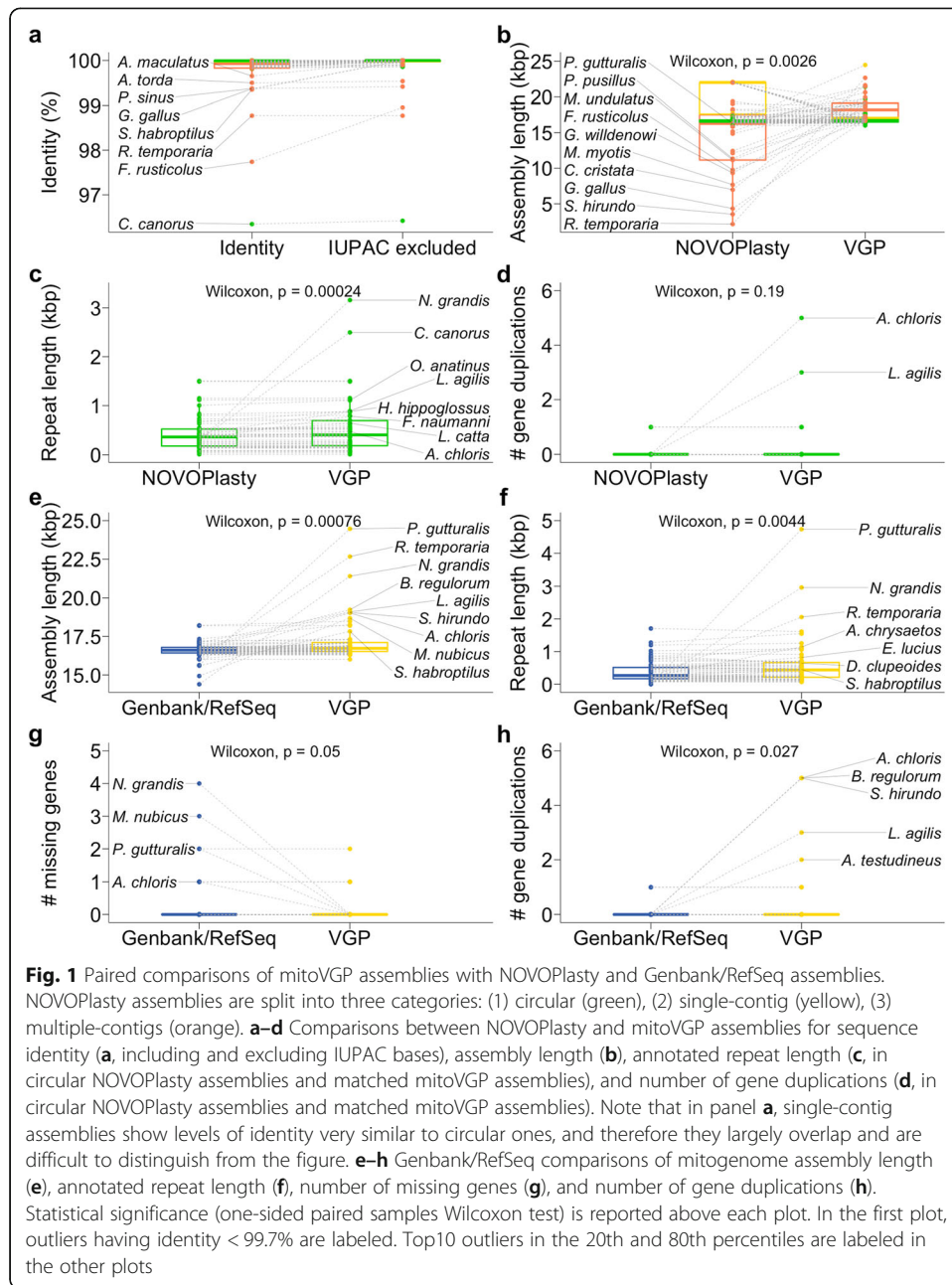
The mitoVGP pipeline was developed under the assumption that a combination of long- and short-read datasets could provide a better representation of the mitogenome. *K*-mer-based QV estimates suggest a high base calling accuracy of mitoVGP assemblies

(Additional file 2: Table S1, column AF), with 58 of the assemblies having no false mtDNA *k*-mers (i.e., false *k*-mers only found in the assembly and not in the high-coverage fraction of the raw data, see “Methods”), and the remaining with average QV 41.25 (approximately 1 base calling error per assembly). We confirmed the general high accuracy of mitoVGP assemblies generating orthogonal Nanopore long-read datasets on a subset of four VGP species using the same sample tissue. Overall, the two datasets for each species generated identical or near identical assemblies (Additional file 9: Table S7).

We decided to benchmark the mitoVGP pipeline in the context of currently available mitogenome assembly tools using our VGP datasets. Unfortunately, the only alternative long-read organelle genome assembler to our knowledge, *Organelle\_PBA* [39], is no longer maintained nor functional (Aureliano Bombarely, personal communication). Therefore, we focussed on NOVOPlasty, a popular WGS short-read mitogenome assembler [40]. Using NOVOPlasty on the short reads of all 125 VGP species, 70 assembled in a single contig, of which 61 were labeled as circular by NOVOPlasty, 48 assembled as multiple contigs, and the assembly failed in 7 cases (Additional file 1: Fig. S4a). NOVOPlasty did succeed in 24 cases where mitoVGP could not because of the absence of long mtDNA reads (15 assembled in multiple contigs, one in a single contig, 8 labeled as circular). When we compared the circular NOVOPlasty assemblies to their mitoVGP counterparts, we found the average level of identity in the alignable regions to be 99.913% (99.918% if IUPAC ambiguous base calls introduced by NOVOPlasty are not considered, Fig. 1a, green). There were 15 mitoVGP assemblies that were substantially larger than their NOVOPlasty counterparts, while all other assemblies had length differences < 5 bp (Fig. 1b, green). Interestingly, single-contig NOVOPlasty assemblies that were not labeled as circular ( $N = 8$ ) were larger than their corresponding mitoVGP assemblies because they had failed to circularize, leaving large overlapping ends (Fig. 1b, yellow). Despite this, sequence identity levels were still remarkable: 99.978% (99.984% disregarding IUPAC bases in NOVOPlasty assemblies, Fig. 1a, yellow). In multiple-contig NOVOPlasty assemblies, the largest contig had an average identity of 99.764% (99.888% removing IUPAC bases, Fig. 1a, orange) with the respective mitoVGP assembly, and the average difference in length was 3867 bp (Fig. 1b, orange), supporting the fragmented nature of some of these NOVOPlasty assemblies. Fragmented NOVOPlasty assemblies correspond to larger than average mitoVGP assemblies, suggesting that more complex mitogenomes were more problematic for NOVOPlasty (Fig. 1b, orange). The general high base-level identity observed is supportive of the high base calling accuracy of mitoVGP assemblies based on the combination of long and short reads. However, significant length divergence, particularly in the comparison with circular NOVOPlasty assemblies (one-sided paired samples Wilcoxon test,  $p = 0.0002$ ), was due to the presence of few large indels. These large indels originate from an underrepresentation of repeats (Fig. 1c) and gene duplications (Fig. 1d) in the NOVOPlasty assemblies.

We compared the length of mitoVGP assemblies to their reciprocal Genbank/RefSeq counterparts, when available (Additional file 1: Fig. S4b, Additional file 2: Table S1,  $N = 66$ , RefSeq = 34, non-RefSeq = 32). MitoVGP assemblies were on average significantly longer (Fig. 1e). Similar to the NOVOPlasty comparison, this difference was mostly driven by mitoVGP assemblies having a significantly higher repeat content (Fig. 1f), particularly when they were longer than their Genbank/RefSeq counterparts





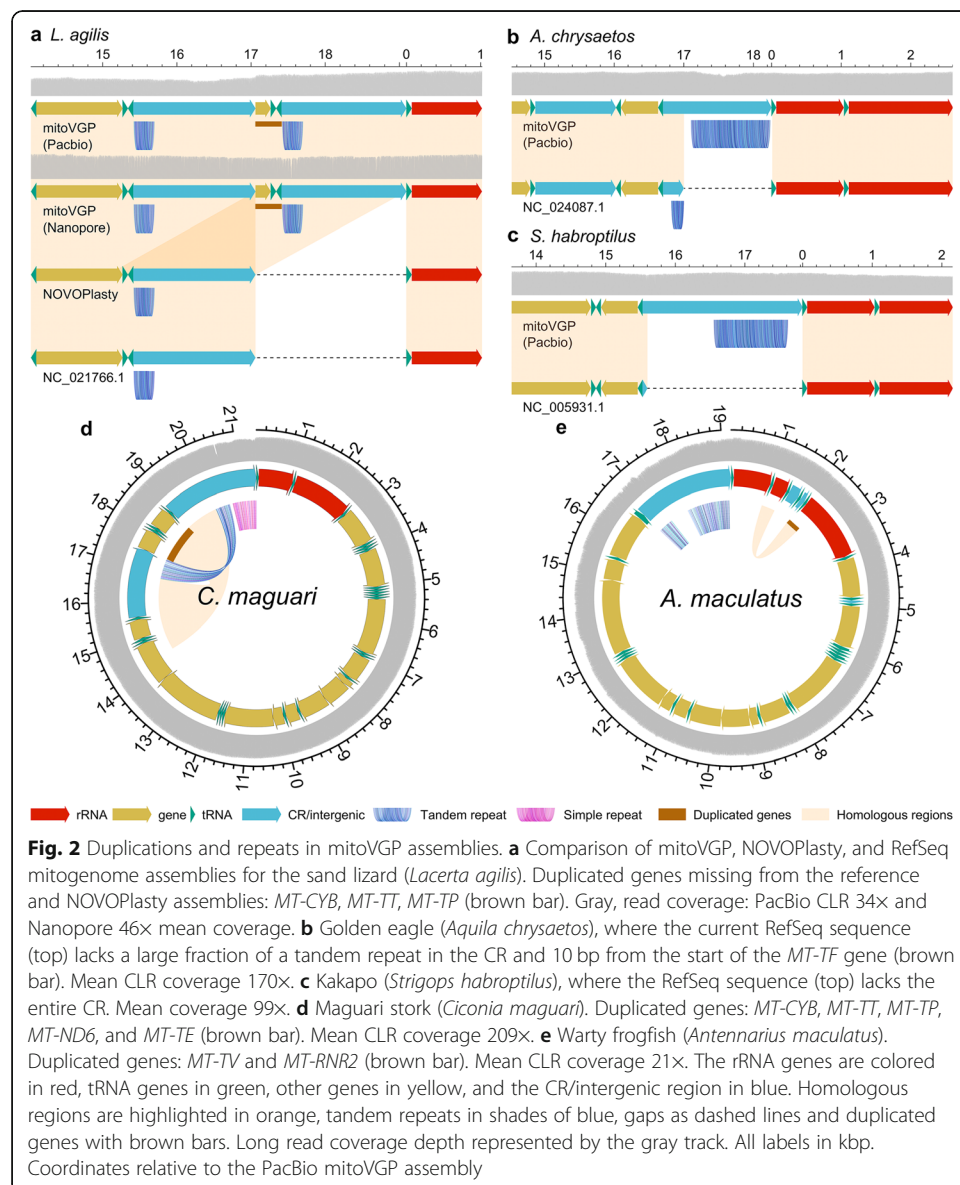
**Fig. 1** Paired comparisons of mitoVGP assemblies with NOVOPlasty and Genbank/RefSeq assemblies. NOVOPlasty assemblies are split into three categories: (1) circular (green), (2) single-contig (yellow), (3) multiple-contigs (orange). **a–d** Comparisons between NOVOPlasty and mitoVGP assemblies for sequence identity (**a**, including and excluding IUPAC bases), assembly length (**b**), annotated repeat length (**c**, in circular NOVOPlasty assemblies and matched mitoVGP assemblies), and number of gene duplications (**d**, in circular NOVOPlasty assemblies and matched mitoVGP assemblies). Note that in panel **a**, single-contig assemblies show levels of identity very similar to circular ones, and therefore they largely overlap and are difficult to distinguish from the figure. **e–h** Genbank/RefSeq comparisons of mitogenome assembly length (**e**), annotated repeat length (**f**), number of missing genes (**g**), and number of gene duplications (**h**). Statistical significance (one-sided paired samples Wilcoxon test) is reported above each plot. In the first plot, outliers having identity < 99.7% are labeled. Top10 outliers in the 20th and 80th percentiles are labeled in the other plots

(Spearman’s  $\rho$  correlation = 0.83, Additional file 1: Fig. S5), but only a marginally significant difference in GC content, with Genbank/RefSeq assemblies having slightly higher GC content (two-sided paired samples Wilcoxon test,  $p = 0.04$ , Additional file 1: Fig. S6). MitoVGP assemblies also tended to have fewer missing genes (Fig. 1 g), as well as a significantly higher representation of gene duplications (Fig. 1 h).

### Novel duplications, repeats, and heteroplasmy

We analyzed the three top outliers in terms of mitogenome assembly length differences found in birds in the mitoVGP vs Genbank/RefSeq comparison (Fig. 1e): the yellow-

throated sandgrouse (*Pterocles gutturalis*, reference generated using Illumina WGS [41]), the great potoo (*Nyctibius grandis*, long-range PCR and direct Sanger sequencing [42]), and the rifleman (*Acanthisitta chloris*, long-range PCR and direct Sanger sequencing [43]). These are indeed labeled as partial in Genbank, and lack most of the CR where repeats and gene duplications typically occur. In contrast, several other outliers have Genbank/RefSeq assemblies labeled as complete but are still shorter than their mitoVGP assembly counterparts. These include the sand lizard (*Lacerta agilis*), where in the mitoVGP assemblies (both Pacbio and Nanopore versions) we found a 1977-bp-long repetitive duplication involving part of the origin of replication that comprises a tandem repeat (repeat unit = 36 bp, 199 bp long), the terminal portion of *MT-CYB* gene, and *MT-TT* and *MT-TP* genes (Fig. 2a). Coverage profiles and repeat-spanning reads supported the presence and extent of the duplication. In particular, out of 145 Nanopore reads of average length 5758 bp, 11 were above 16 kbp and fully spanned the



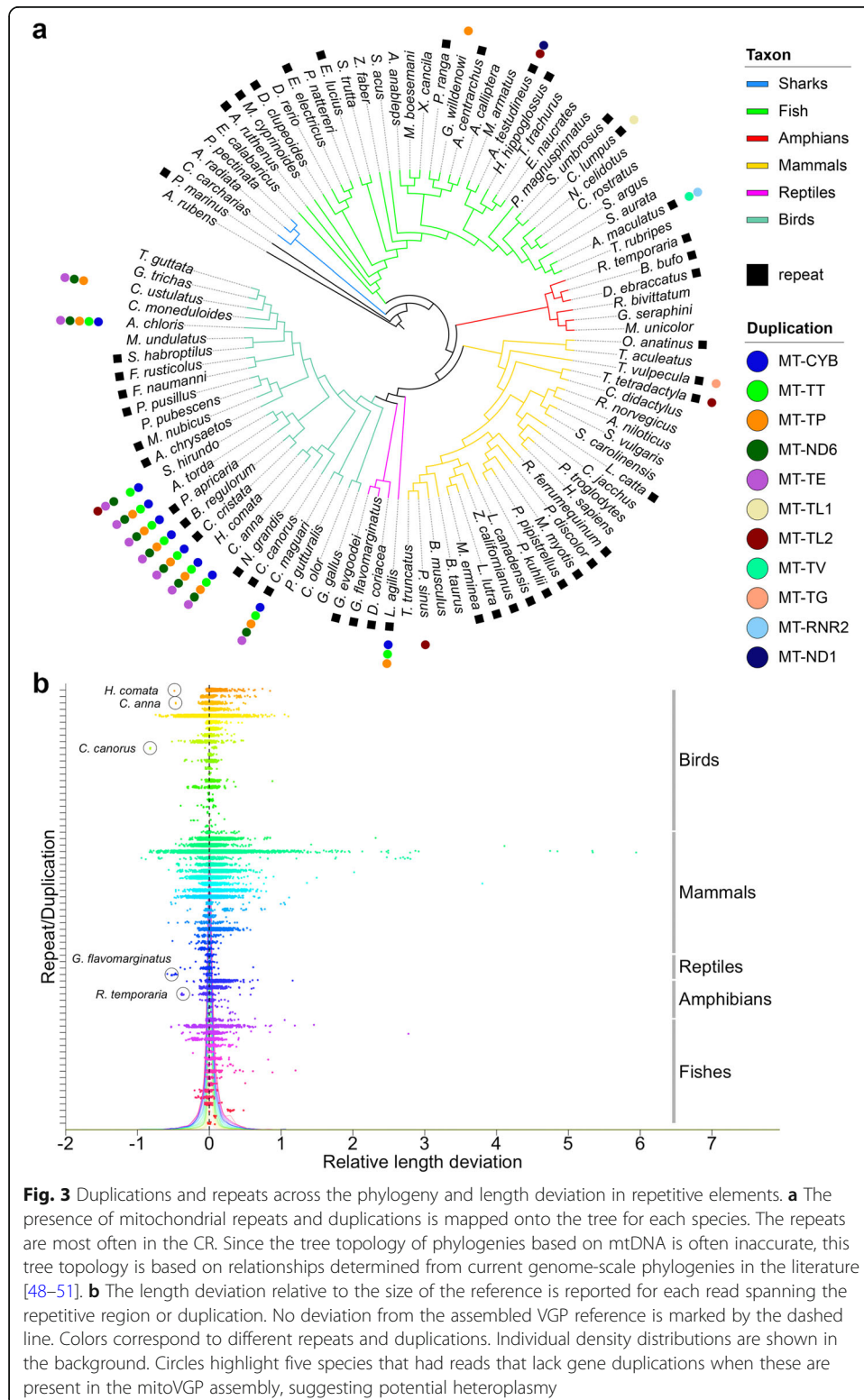
repeat; similarly, of 36 PacBio reads of average length 8153 bp, 6 were over 16 kbp long and spanned the repeat. Due to the nature of PacBio libraries, the same molecule can potentially be read multiple times while the polymerase enzyme passes over the circular SMRTbell. We found at least 2 full-pass reads covering the same mtDNA molecule twice (forward and reverse strand), and the duplication was still present. This duplication was completely absent in our circular NOVOPlasty assembly (which is however 100% identical to the PacBio assembly for the rest of the sequence) as well as in the current RefSeq reference [44] (Fig. 2a), strengthening the notion that short-read assemblies inevitably fail to represent gene duplications and repeats.

In the case of the golden eagle (*Aquila chrysaetos*), the 921-bp CR had a tandem repeat (repeat unit = 49 bp, ~ 782 bp long), which was essentially absent from the RefSeq reference (generated using Illumina WGS [45]) along with the first 10 bp encoding the *MT-TF* gene (Fig. 2b). Aside from the missing repeat, the remainder of the two sequences were 99.73% identical, the differences most likely attributable to individual variation. The high level of similarity is supportive of the overall quality of the mitoVGP assembly, which is also confirmed by its Q44.30 base call accuracy and 100% identity to the NOVOPlasty assembly in non-repetitive regions. In the kakapo (*Strigops habroptilus*), an entire 2.3-kbp CR region, including a ~ 925-bp-long repeat (repeat unit = 84 bp), was also missing from the RefSeq sequence (long-range PCR and direct Sanger sequencing [35, 43], Fig. 2c). In the case of the common tern (*Sterna hirundo*), the current RefSeq sequence (PCR and Sanger sequencing [46]) was missing a duplication in the CR involving the *MT-CYB*, *MT-TT*, *MT-TL2*, *MT-ND6*, and *MT-TE* genes, as well as a substantial fraction of a tandem repeat. In the case of the Indo-Pacific tarpon (*Megalops cyprinoides*), a ~ 650-bp tandem repeat was represented in the RefSeq reference (long-range PCR and direct Sanger sequencing [47]) for two thirds of its length, and the sequence downstream was completely absent, for ~ 500 bp. The two sequences shared 99.89% identity, with mitoVGP assembly showing no base calling errors; our mitoVGP assembly was 100% identical to our NOVOPlasty assembly in the homologous regions, but the latter lacked 165 bp of the repeat, compatible with short reads falling short in covering the 650 bp repeat.

A total of 33 mitoVGP assemblies did not have a Genbank/RefSeq representative. Among these assemblies, 8 showed duplications and/or large repetitive elements. These include 5 birds belonging to 5 separate orders that have highly similar, but not identical patterns of duplicated genes in the CR (*MT-CYB*, *MT-TT*, *MT-TP*, *MT-ND6* and *MT-TE*): the Anna's hummingbird (*Calypte anna*) [35], red-legged seriema (*Cariama cristata*), Swainson's thrush (*Catharus ustulatus*), maguari stork (*Ciconia maguari*), whiskered treeswift (*Hemiprocne comata*), and European golden plover (*Pluvialis apricaria*). For example, in the maguari stork (*Ciconia maguari*, Fig. 2d) part of the CR itself is duplicated, and there is a simple repeat (CAA/CAAA, 842 bp long) and two nearly identical tandem repeats (repeat unit = 70 bp, ~ 523 bp long and ~ 719 bp long respectively), leading to a 21.4-kbp-long mitogenome (assembly Q41.41). Another example of newly identified gene duplications is represented by the warty frogfish (*Antennarius maculatus*, Fig. 2e). Here *MT-TV* and *MT-RNR2* genes are partially duplicated while the CR shows two distinct repetitive elements, a short tandem repeat (repeat unit = 14 bp, ATAACATACATTAT / ATAGTATACATTAT, 1330 bp long) and a longer tandem repeat (repeat unit = 54 bp, ~ 479 bp-long), leading to a mitogenome size of 19.2 kbp with no base calling errors.



Overall, over 50% of mitoVGP assemblies (52/100) contained 1 to 4 repeats ( $N = 45$  species, for a total of 68 repetitive regions; Fig. 3a, Additional file 10: Table S8) and/or gene duplications ( $N = 18$  species; Fig. 3a, Additional file 2: Table S1, columns AJ-AQ).



**Fig. 3** Duplications and repeats across the phylogeny and length deviation in repetitive elements. **a** The presence of mitochondrial repeats and duplications is mapped onto the tree for each species. The repeats are most often in the CR. Since the tree topology of phylogenies based on mtDNA is often inaccurate, this tree topology is based on relationships determined from current genome-scale phylogenies in the literature [48–51]. **b** The length deviation relative to the size of the reference is reported for each read spanning the repetitive region or duplication. No deviation from the assembled VGP reference is marked by the dashed line. Colors correspond to different repeats and duplications. Individual density distributions are shown in the background. Circles highlight five species that had reads that lack gene duplications when these are present in the mitoVGP assembly, suggesting potential heteroplasmy

Repeats, mostly in the CR, were detected in nearly half of the members of each tetrapod group, with the highest proportions being found in birds and reptiles (51.6%, 16/31; average length = 725 bp, sd = 592 bp), with all four reptiles (1 lizard and 3 turtles) showing one or more repeats ( $N = 4$ ); followed by amphibians (50%, 3/6 species; average length = 826; sd = 399 bp), and mammals (46.2%, 12/26; average length 317 bp, sd = 128 bp). A smaller proportion of fish species had repeats (37.5%, 12/32; average length = 550 bp; sd = 352). We classified the repeat patterns, and found diversity in repeat lengths and sequences, but with a distribution to lower GC content, ranging 0–61% (Fig. S7a-c; Additional file 10: Table S8). We clustered the repeats by  $k$ -mer-sequence similarity and found specific types with generally higher GC content shared by mammals (Fig. S7d, yellow) and another by birds (Fig. S7d, light green), with some low-to-high similarity within the clusters. The fish did not have any specific clustering, and multiple species appeared to have nearly unique sequences (Fig. S7d). When multiple repeats were present in the same species, these were usually either almost identical if involved in the same duplication event, or shared little or no sequence similarity (Additional file 1: Fig. S7d). Taken together, our results suggest that these vertebrate mitochondrial repeats are hypervariable and thus are generally of recent origin, in line with other reports of extreme variation in patterns of tandem repeats in the CR even within the same species [52].

Duplications were in the highest proportion in birds (37%, 10/27), followed by reptiles (25%, 1/4), fish (12.5%, 4/32), and the least in mammals (11.5%, 3/26) (Fig. 3a). According to the phylogeny and given the high variability among species, some repeats and duplications within and across vertebrate lineages can be interpreted to have diverged from a common ancestor or have independently converged (Fig. 3a). For example, since most birds sequenced have repeats in the CR, we can infer that it was likely the ancestral state, but then lost in some lineages and widely diverged in others. For duplications, similar to what has been claimed for Passeriformes [27], duplications in or near the control region were present in several monophyletic clades that represent basal branches, spanning the Caprimulgiformes (e.g., hummingbird, *Calypte anna*) to Charadriiformes (e.g., razorbill, *Alca torda*), suggesting that this duplication was ancestral among them, and lost in close relatives (e.g., great potoo, *Nictus grandis*; Fig. 3a). These and other hypotheses on the phylogenetic history of mitochondrial repeats and duplications will be more quantitatively resolved once the remaining ordinal level genomes of the VGP Phase 1 are completed.

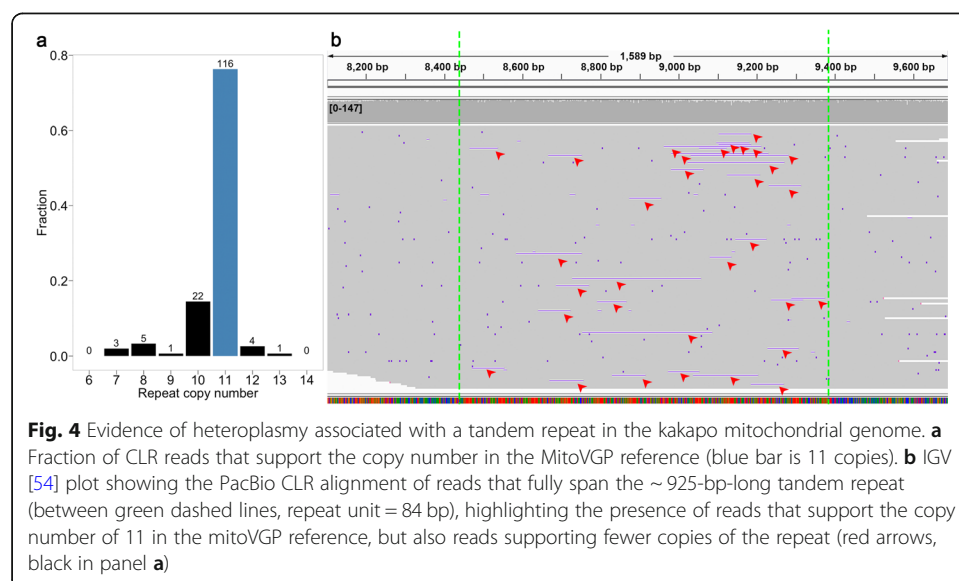
Long-read platforms offer the additional advantage of single molecule-resolution without amplification bias. To assess the presence and degree of heteroplasmy and make sure that our assemblies represented the most frequent allele, we measured individual read length variation in repetitive elements and duplications. The analysis revealed that the read length difference from the assembled mitoVGP reference allele is centered around zero (Fig. 3b, deviation relative to the reference; Fig. S8, deviation in absolute kbp). Some minor deviations from zero could be explained by the high indel error rate of PacBio reads. Positive deviations were significantly favored ( $t = 29.631$ ,  $df = 38,374$ ,  $p < 2.2 \times 10^{-16}$ ), compatible with the higher number of insertions over deletions in PacBio indel errors [53], and the standard deviation linearly correlated with the size of the repeat element (Spearman's  $\rho$  correlation = 0.79,  $p < 8.1 \times 10^{-16}$ ). Major deviations of  $\pm 0.2$  in length of repetitive elements and duplications are unlikely explained

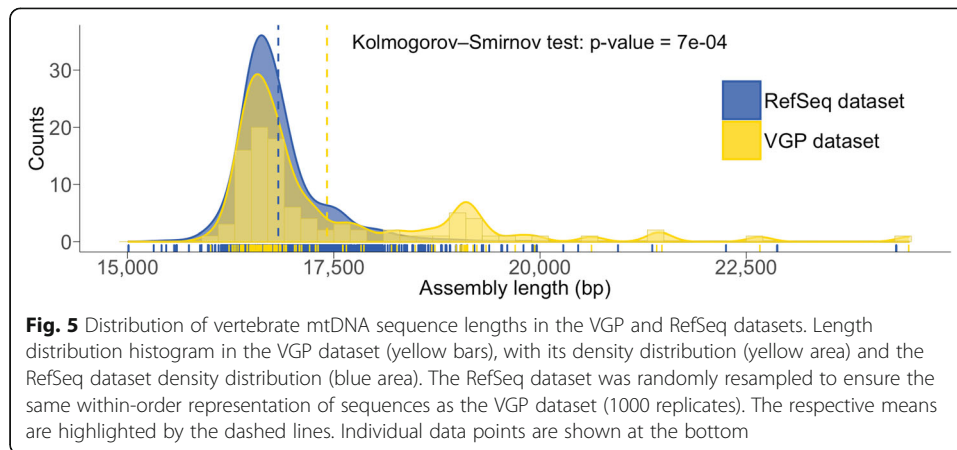
by sequencing errors. Rather, they can be explained by heteroplasmy. For example, the 84-bp tandem repeat of the kakapo was present 11 times in almost 80% of the reads that spanned the repeat region; but the remaining reads had from 7 to 10, 12, or 13 repeats (Fig. 4). Interestingly, in a few species, repetitive elements and duplications were missing in a few reads spanning the region, whereas the other reads supported the reference, indicative of heteroplasmy (Fig. 3b, circles). The presence of repeats or duplications was not associated with tissue type (repeats: Fisher's exact test,  $p = 0.11$ ,  $N = 125$ ; duplications: Fisher's exact test,  $p = 0.16$ ,  $N = 125$ ), suggesting that these are not transient, tissue-specific events.

For the human mitoVGP assembly, we had trio data (child and parents), including recently developed PacBio HiFi (High-Fidelity) technology, which generates  $\sim 10$ – $20$ -kbp-long reads having lower error profiles close to Illumina short reads [55]. We mapped the child and parental HiFi reads child CLR mitoVGP assembly. The HiFi read alignments showed uniform coverage across the child CLR mitoVGP assembly, and the identification of SNPs and indels at single molecule resolution that showed the expected inheritance from the mother (Fig. S9).

#### Length distribution of vertebrate mitogenomes

Having ascertained that several of our assemblies are likely more complete than the respective existing references, we compared the mtDNA sequence length distribution of mitoVGP assemblies with the overall distribution of RefSeq representative mitogenomes of vertebrates labeled as complete. A final RefSeq dataset containing 3567 sequences, excluding our VGP submissions (Additional file 11: Table S9), was randomly resampled to ensure the same within-order representation of sequences as the VGP dataset (1000 replicates). The two datasets consistently diverged for lengths  $> 18$  kbp, with longer sequences represented in the mitoVGP dataset (Fig. 5). This divergence was most influenced by duplications and repeats in non-mammalian species and suggests that the underrepresentation of repeats and duplications observed in our sample





is likely to affect many of the existing reference assemblies deemed complete. It also highlights the multimodal distribution of mitogenome length in vertebrates, with mitoVGP assemblies revealing at least a secondary peak due to the presence of repeats and duplications. This secondary peak is mostly contributed by a duplication involving *MT-CYB*, *MT-TT*, *MT-TP*, *MT-ND6*, and *MT-TE* that we have frequently observed in birds ( $N = 9$ ), and which is usually missed in short-read assemblies. Of note, genes were usually fully duplicated, the most frequent exception being *MT-CYB*, which was usually only partially duplicated and of varying length (120–609 bp; Additional file 2: Table S1, column AL). This may be due to a positional factor, where *MT-CYB* is at the beginning of the duplicated region.

## Discussion

We have demonstrated that mitoVGP, a new mitogenome assembler combining long and short reads, can successfully assemble high-quality mitogenomes in a variety of datasets. The number, quality and variety of complete mitogenome assemblies and datasets presented here allows a more accurate comparison of sequence data and assembly strategies than ever before. Our results suggest that when targeting the mitogenome, a careful design should be used to decide the sequencing technology, with further attention paid to the evaluation of assembly results. Tissue types with abundant mtDNA and libraries that avoid too stringent size selection should be preferred to ensure the presence of mtDNA reads in WGS experiments with long reads. Possibly due to their shallow size selection, current Nanopore library preparation protocols favor the availability of mtDNA reads over current PacBio CLR library protocols (see “Methods”). Alternatively, if stringent size selection is avoided, given their length and base accuracy, HiFi PacBio reads are an excellent candidate for future mitogenome studies, providing incontrovertible single molecule assessment.

## Conclusions

The Genbank nucleotide database, one of the most complete DNA sequence archives, contains thousands of animal mtDNA sequences. A search in the animal subset (June 2020, keywords “mitochondrion AND complete”), yielded over 100,000 mitochondrial sequences, of which 75,565 were vertebrates. When the sequence is reported as *complete* in the metadata, we suggest it should imply that the full mtDNA sequence,

circular in the case of vertebrate mitogenomes, has been assembled with no gaps. However, when compared with our long-read mitogenome assemblies, a proportion (at least 15%) of short-read vertebrate assemblies publicly available in Genbank/RefSeq repositories and labeled as complete, are missing repeats and gene duplications. Unfortunately, these assemblies usually lack the publicly available supporting raw sequence data, making it difficult, or at times impossible, to evaluate their quality. Despite this limitation, we have shown that the discrepancy between mitoVGP and previous submitted Genbank/RefSeq assemblies is due to the use of long reads that span repetitive elements and duplicated genes. The presence of repeats and duplications in over half of the species herein assembled indicates that their occurrence is a principle of mitochondrial structure rather than an exception. Given the relatively high frequency of these elements, even in the “simple” case of vertebrate mitogenomes, the completeness of many currently available reference sequences can be further improved. Therefore, caution should be exercised before claiming the complete assembly of a mitogenome from short reads alone.

## Methods

### VGP data generation

PacBio and 10x Genomics datasets for all VGP species were generated following protocols detailed in our companion paper on the nuclear assembly pipeline [35] and in Mountcastle et al. (in preparation). The final dataset ( $N = 125$ ) includes one invertebrate (the common starfish, *Asterias rubens*) and two individuals for the zebra finch (*Taeniopygia guttata*, one male and one female). A summary of the approaches employed for the samples analyzed in this work is provided in Additional file 2: Table S1. Briefly, total genomic DNA (gDNA) was obtained using a variety of state-of-the-art approaches for High Molecular Weight (HMW) DNA extraction available mostly at three different sequencing facilities of the contributing to the VGP (<https://vertebrategenomesproject.org/>): The Rockefeller University Vertebrate Genome Laboratory in New York, USA; the Wellcome Trust Sanger Institute in Hinxton, UK; and the Max Planck Institute in Dresden, Germany. This includes the Bionano plug protocol for soft tissue (Cat. No. 80002) and nucleated blood (Cat. No. 80004), MagAttract HMW DNA Kit for blood and tissue (Cat. No. 67563), and Phenol-Chloroform extraction. Library preparation followed standard protocols as suggested by the datasheets. In several cases, the DNA was fragmented using the Megaruptor at various fragment sizes between 15 and 75 kbp. In other cases, the DNA was fragmented by needle shearing. Importantly, libraries were usually size-selected to enrich for HMW fragments, and the range of size selection varied widely between 7 and 40 kbp. Both PacBio CLR and 10x Genomics linked reads were generated for all species in the VGP dataset, except the common starfish (*Asteria rubens*) and the chimp (*Pan troglodytes*) for which 10x was replaced with standard Illumina library preparation and publicly available data (SRX243527), respectively. For the human trio, we generated ~ 10 kbp CCS libraries for all samples.

### Nanopore data generation

For the Nanopore datasets, total gDNA was obtained from tissue using Genomic-tip 100/G (Qiagen) for the spotty (*Notolabrus celidotus*), thorny skate (*Amblyraja radiata*),



and hourglass treefrog (*Dendropsophus ebraccatus*), following the manufacturer's protocol. For the sand lizard blood, total gDNA was extracted using Nanobind CBB Big DNA kit (Circulomics) as described by the manufacturer. All extracted gDNA underwent size selection using the Short Read Eliminator kit (Circulomics) to deplete fragments < 10 kb. The resulting material was then prepared for sequencing using the Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies Ltd) and sequenced using R9.4.1 flowcells (FLO-PRO002) on the PromethION device (Oxford Nanopore Technologies Ltd). Flowcell washes and library re-loads were performed when required. Interestingly, at matched coverage, most Nanopore datasets contained a larger amount of mtDNA reads compared to PacBio (average fold change 3.9, Additional file 9: Table S7). The sole exception was the thorny skate, where the two datasets are comparable (1.6 fold more reads in the PacBio dataset). The PacBio dataset for the thorny skate was one of the very first VGP datasets produced, and at that time, size selection was not as stringent.

### Mitogenome assembly pipeline

The mitoVGP pipeline was designed to simultaneously take advantage of the availability of long reads (especially PacBio) and short reads (usually 10× linked reads) data from the same individual. This condition is met for all genomes sequenced under the VGP nuclear genome pipeline, as well as for Nanopore datasets. The pipeline is fully automated, and it is composed of a series of single-node, fully parallelized, Bash scripts designed to run in a Linux environment. The amount of resources required is minimal and will only affect speed. Command line code to reproduce our results for each assembly using mitoVGP is provided in Additional file 2: Table S1.

Similar to other methods available to assemble organelle genomes [39, 40, 56], mitoVGP general workflow starts by selecting putative mitochondrial reads from a long-read WGS dataset based on their similarity with an existing reference of the same or other species, even distantly related ones. In mitoVGP 2.2, this is achieved using pbmm2 v1.0.0 (<https://github.com/PacificBiosciences/pbmm2>), the official PacBio implementation of Minimap2 (<https://github.com/lh3/minimap2>) [57] for PacBio long reads, and using directly Minimap2 v2.17 for Nanopore reads. MitoVGP currently supports a variety of different PacBio chemistries. For RSII chemistries, the aligner blasr v5.3.3 was employed via the -m option (<https://github.com/PacificBiosciences/blasr>). Reads were individually aligned to a mtDNA reference sequence using default parameters and allowing for unique alignments. The reference can be of the same species, or that of a closely-to-distantly related species, since even with default parameters pbmm2/Minimap2/blasr search similarity cut-offs are relatively loose, to account for the high error rate of noisy long reads [57]. We have experimentally determined in one VGP dataset (Anna's hummingbird) that, when mapping with pbmm2 and *C. elegans* mitogenome indel/substitution rates [58], an edit distance between the reference and the sample as large as 20% will decrease the number of available reads by 41% (total Gbp decrease 25%), and therefore having no substantial impact in long read availability for assembly in most cases. For comparison, mouse (*Mus musculus*, NC\_005089.1) and zebrafish (*Danio rerio*, NC\_002333.2) edit distance is 33.1%. The use of even more distant reference sequences should be possible using lower stringency in the mapper.

Moreover, such references do not have to be complete, since even short matches on a fragmented reference will allow fishing out long reads, potentially spanning the gaps in the reference.

The next step of the pipeline involves the de novo genome assembly of the long reads extracted from the WGS dataset using the long read assembler Canu v1.8 (<https://github.com/marbl/canu>) [37]. After the assembly, since the Canu output may contain more than one contig, we used BLAST [59] to identify and filter out contigs originating from lower quality mtDNA reads or failed overlaps as well as from the inclusion of nDNA reads. The sequence of the putative mitocontig was then polished using Arrow (<https://github.com/PacificBiosciences/GenomicConsensus>) in the case of PacBio datasets, or one round of Racon (<https://github.com/isovic/racon>) and one of Medaka (<https://github.com/nanoporetech/medaka>) in the case of Nanopore datasets. The sequence was further refined with a round of polishing using short-read data, where the WGS short reads were mapped with Bowtie2 v2.3.4.1 [60] to extract putative mtDNA reads, variant calling on the alignments performed with Freebayes v1.0.2 [61], and consensus generated with Bcftools v1.9 using parameters optimal for a haploid genome and to left-align and normalize indels.

Because Canu was developed to assemble large linear nuclear chromosomes rather than short circular genomes, it leaves large overlaps at the sequence ends [37]. In a few cases, particularly when PacBio reads were missing adapter sequences resulting in sub-reads having multiple copies of the same DNA sequence, these overlaps can become exceptionally long, including the mitochondrial sequence read through multiple times. A custom script was developed to remove the overlaps (<https://github.com/gf777/ mitoVGP/blob/master/scripts/trimmer>). Contrary to other methods that are based on the sequence of the overlaps alone (e.g., Circlator [38] <https://sanger-pathogens.github.io/circlator/>), the script is based on the deconvolution of the repetitive elements using MUMmer matches (<https://mummer4.github.io/>) [62], short-read mapping to the sequence using Bowtie2 [60] and coverage-based definition of the reliable ends. A final round of short-read polishing was performed using the same approach as previously described. Although we used 10x Genomics linked short-reads, mitoVGP accepts any short-read dataset in FASTQ format. Finally, tRNAs were identified using tRNA scan-SE [63], and the sequence was automatically oriented to start with the conventional tRNA Phenyl-Alanine (*MT-TF*).

#### **mitoVGP parameters**

For 34 species, default mitoVGP parameters were adjusted as detailed in Additional file 2: Table S1. Relevant parameters include a query coverage cutoff to filter spurious BLAST [59] matches during the identification of the putative mitochondrial contig (-p option, applied in 13 cases), the maximum read length cutoff applied to long reads extracted with the aligner in the first step (-f option, 19 cases), Canu defaults (which can be adjusted directly in mitoVGP using the -o option, 1 case), and the MUMmer cutoff for extending matches (-z option, 4 cases). The query coverage cutoff is important to identify and remove reads that only share a very loose similarity with the reference, such as repeats or small common motifs. Very short spurious matches can be filtered with a small value (e.g., -p 5), but if the reference is considered complete

and closely related, query coverage can safely be increased (e.g., -p 70) as it is generally robust to the low accuracy of long reads. Filtering out reads that are significantly longer than the expected mitogenome assembly size (e.g., 25 kbp) using the maximum read length cutoff is a complementary strategy that in some cases can considerably improve the quality of the assembly, since these reads will include NUMTs with flanking nuclear sequence. Without filtering these reads will complicate the assembly graph, often resulting in assembly failure. Canu defaults can be changed for a variety of reasons, such as different error rates due to chemistry or quality of the dataset. When only a few mtDNA reads are available, the option *stopOnLowCoverage* needs to be tweaked (e.g., -o “stopOnLowCoverage = 9”). The MUMmer cutoff should normally be adjusted only in a few cases, for instance in the presence of very large repeats. Nanopore-based assemblies were generated using the same pipeline. For the spotty and the hourglass treefrog mitoVGP was run with default parameters. For the sand lizard, the -f option was adjusted to 25,000 and -p was adjusted to 5. For the thorny skate, the -f option was adjusted to 18,000.

### Statistical analyses

Fisher’s exact tests and  $\chi^2$  tests were performed with the respective primitive R functions using default parameters. Fisher’s exact test was used when the counts in each category were not sufficient for a reliable  $\chi^2$  test. In  $\chi^2$  tests, simulated *p* values were generated using 10,000 replicates.

The set of reliable mtDNA reads used for statistical analyses was defined by the reads that covered the reference by at least 70% of their length using BLAST [59] matches using the reference as query and the reads as reference (formula: reference length/read length  $\times$  query cover). For three species with phylogenetically distant reference sequences, i.e., large-scale four-eyed fish (*Anableps anableps*), warty frogfish (*Antennarius maculatus*), and blunt-snouted clingfish (*Gouania willdenowi*), the mitoVGP assembly was used to extract the reads to avoid underestimating read counts.

The linear model was implemented in R using the primitive R function *lm*:

Number of long mtDNA reads  $\sim$  Taxonomic group + Tissue type + Size selection + DNA extraction + Fragmentation + Library prep + Total raw data (Gbp).

The fraction of the variance explained by each predictor was computed as Sum Sq/Total Sum Sq  $\times$  100. For size selection, when multiple libraries with different cutoffs were generated, the minimum cutoff was considered in the statistical analyses. All observations having a factor represented 3 times or less were excluded ( $N = 94$ ). Post hoc tests were performed with the *glht* function in the multcomp R library. We carefully checked model assumptions and further checked statistical significance of all terms with a permutation approach using the *lmp* function in the lmPerm R library. Results were always consistent, so for simplicity we report only the results of the parametric test.

### Measure of QV

QV of the mitogenome assemblies was assessed using a recently developed k-mer method called Merqury (<https://github.com/marbl/merqury>) [64]. The tool was run with default parameters and k-mer size 31, but given the approximate 50–100 $\times$

coverage of VGP datasets, k-mers with frequency < 100 were removed, to limit QV overestimation due to the inclusion of nDNA k-mers. Fifteen datasets were excluded from the analysis since mitogenome short-read coverage < 250× (Additional file 2: Table S1, column AE) overlapped substantially with the nDNA k-mers distribution and was similarly affected by the cutoff, preventing a reliable QV estimate.

### **Benchmarking with existing tools**

We run NOVOPlasty [40] with default parameters, using the same reference sequence employed for mitoVGP as bait, and trimming 10× datasets using proc10xG with the -a option (<https://github.com/ucdavis-bioinformatics/proc10xG>). Following NOVOPlasty authors' suggestion and in order for the results to be comparable to mitoVGP assemblies, the entire short-read set was employed. Despite the fact that both mitoVGP and NOVOPlasty assemblies run on machines with 32 cores and 383 GiB of memory, in the case of NOVOPlasty the large size of VGP datasets often led to memory issues as all the reads have to be simultaneously loaded into a hash table. Assemblies that initially failed with NOVOPlasty were rerun on two fat nodes, each with 64 cores and 1.534 and 3.096 TB of memory respectively, narrowing down to four the number of failed assemblies due to reproducible internal errors (European Toad, Eurasian black-cap, cow, Canada lynx). Among the successful assemblies, one assembly (Anna's hummingbird) was made of multiple contigs < 3 kbp and < 90% identity with the reference, one assembly (stoat) was made of multiple contigs with no identity with the reference, and one assembly (common pipistrelle) was made of a single 253 bp contig. These results were likely due to the inability of NOVOPlasty to identify an appropriate seed, and these assemblies were excluded from downstream analyses.

### **Annotation**

In all datasets, genes were annotated using MITOS v2.0.6 (<https://gitlab.com/Bernt/MITOS>) [65], with default parameters. For the VGP and Genbank/RefSeq comparison, since MITOS may sometimes generate two contiguous annotations out of a single gene, the annotations were manually curated to identify real missing genes and gene duplications. Duplications were considered complete if they overlapped in length at least 95% with the original gene (Additional file 2: Table S1, column AL). Simple and tandem repeats were annotated using WindowMasker v1.0.0, with default parameters. K-mers were counted combining both the mitoVGP assembly and its Genbank/RefSeq or NOVOPlasty counterpart, and the difference in repeat representation between the two datasets was measured as the difference in the number of bases annotated as repetitive for each species.

### **Coverage tracks**

Reads were remapped on the original assemblies and filtered for identity to the reference > 70% and length above ~ the repeat length + 2000 bp to ensure that they could anchor on both sides of the repetitive elements, avoiding contamination by nuclear repeats. For circos plots, reads were mapped to a 2-copy concatemer of the reference sequence two allow accurate read mapping at the edges.

### Phylogenetic tree

For the consensus tree, Timetree (<http://www.timetree.org/>) [66] was initially used to generate a topological backbone of all species with a mitoVGP assembly. The tree topology was edited using TreeGraph2 [67] to reflect the current hypotheses on the phylogeny of the major vertebrate clades [48–51].

### Repeat sequence similarity

To estimate repeat sequence similarity, repeat boundaries were identified using BLAST self-alignment without repeat masking and with word size = 16. 7-mers that could accommodate short repeat units were then generated for all repeat sequences using the *k*-mer-counter meryl (<https://github.com/marbl/meryl>), and the fraction of shared *k*-mers was calculated as the total number of *k*-mers common to both datasets divided by the total number of *k*-mers of both datasets.

### Measure of heteroplasmy

The coordinates of repeats and duplications were identified by manual inspection of BLAST [59] off-diagonal matches with no repeat masking. The deviation from the reference was assessed by mapping the full WGS read set to the reference with pbmm2 v1.0.0. Confident mtDNA reads were identified in the same way used to define the read set for the linear model, the reads fully overlapping the repeat were filtered using bedtools [68] and hard-clipped with jvarkit [69].

### RefSeq sequence analysis

In order to collect RefSeq mtDNA sequences from Genbank, on March 30, 2020, we conducted a custom search using the following keywords “*mitochondrion AND srcdb\_refseq[PROP] AND complete [TITLE] NOT complete cds [TITLE] NOT isolate NOT voucher\_03302020*”. To assign taxonomy and generate random samples belonging to the same orders of the VGP datasets, we used the R package *taxonomizr*. Within-order resampling was conducted with 1000 replicates.

### Graphical representations

For pairwise comparisons, we used R packages *ggplot2*, *gtable*, and *ggrepel*. To represent gene duplications in the VGP and Genbank/RefSeq comparison, we used the R package *shape*. For Circos plots, we used the R package *circlize*. The phylogenetic tree was annotated with iTOL (<https://itol.embl.de/>) [54]. For density plots, we used the R package *ggplot2*. For the correlation, we used R packages *ggplot2* and *ggpubr*. The repeat heatmap was generated using the R package *corrplot*. Figures of read alignments were generated using IGV [70].

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02336-9>.

**Additional file 1: Fig. S1.** Outline of the mitoVGP assembly pipeline. Fig. S2. Assembly success by the availability of long mtDNA reads. Fig. S3. PacBio CLR mitochondrial read counts in different tissues. Fig. S4. mitoVGP assembly results and comparisons. Fig. S5. Correlation between differences in repeat content and assembly length between the mitoVGP versus the Genbank/Refseq assemblies. Fig. S6. Paired comparisons of GC content between the mitoVGP assemblies and their Genbank/RefSeq counterparts. Fig. S7. Heatmap of *k*-mer-based sequence similarity



of repetitive elements. Fig. S8. Read length deviation from the reference in kbp. Fig. S9. Long accurate HiFi reads from the VGP human trio mapped to the child mitogenome assembly.

**Additional file 2: Table S1.** Metadata.

**Additional file 3: Table S2.** Success rate of the mitoVGP assembly pipeline by taxonomic group.

**Additional file 4: Table S3.** Analysis of variance of factors affecting the availability of long mtDNA reads.

**Additional file 5: Table S4.** Success rate of the mitoVGP assembly pipeline by tissue type.

**Additional file 6: Table S5.** Contingency table tissue type by taxonomic group in the VGP dataset.

**Additional file 7: Table S6.** Success rate of the mitoVGP assembly pipeline by size selection cutoff.

**Additional file 8: Supplementary Note 1.** Relationship between mtDNA sequencing, coverage and genome size.

**Additional file 9: Table S7.** Comparison of the mitogenome assembly from the same sample using PacBio and Nanopore read sets.

**Additional file 10: Table S8.** Analysis of repetitive sequences.

**Additional file 11: Table S9.** Genbank mtDNA sequences in this study.

**Additional file 12.** Review history.

### Acknowledgements

We thank the contributors of the VGP on the first 125 species for letting us use data for generating mitochondrial genome assemblies; in particular, they are Alexander N. G. Kirschel, Andrew Digby, Andrew Veale, Anne Bronikowski, Bob Murphy, Bruce Robertson, Clare Baker, Camila Mazzoni, Christopher Balakrishnan, Chul Lee, Daniel Mead, Emma Teeling, Erez Lieberman Aiden, Erica Todd, Evan Eichler, Gavin J.P. Naylor, Guojie Zhang, Jeremiah Smith, Jochen Wolf, Justin Touchon, Kira Delmore, Kjetill Jakobsen, Lisa Komoroske, Mark Wilkinson, Martin Genner, Martin Pšenička, Matthew Fuxjager, Mike Stratton, Miriam Liedvogel, Neil Gemmell, Piotr Minias, Peter O. Dunn, Peter Sudmant, Phil Morin, Sadequr Rahman, Qasim Ayub, Robert Kraus, Sonja Vernes, Steve Smith, Tanya Lama, Taylor Edwards, Tim Smith, Tom Gilbert, Tomas Marques-Bonet, Tony Einfeldt, Byrappa Venkatesh, Warren Johnson, Wes Warren, and Yury Bukhman. We are grateful to the Ngā Papatipu Rūnanga o Murihiku and the Ngāi Tahu for their support in generating the kakapo datasets. We also thank Aureliano Bombarely for his support in testing Organelle\_PBA on VGP datasets.

### Review history

The review history is available as Additional file 12.

### Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Authors' contributions

G. F. built the mitoVGP pipeline with support from A. R., D. H., M. R. C., S. K., S. M., A. F., A. M. P., and the mitoVGP assembly working group. G. F. performed the bioinformatic and statistical analyses with support from F. O. A., R. A., E. B., K. H., A. T., J. W., M. R. C., P. H., A. A., J. K., J. T., and the mitoVGP assembly working group. PacBio and 10x data were generated by O. F., B. H., J. M., I. B., J. B., S. W., K. O., J. S., E. B., J. D., M. S., and C. C. Nanopore datasets were provided by V.C., S. M., and D. F. W. K. and J. T. helped with metadata collection and submission of the assemblies to the public archives. G. F. and E. D. J. conceived the study and drafted the manuscript. All authors contributed to the final manuscript and approved it.

### Funding

A. R., S. K., and A. M. P. were supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. A. R. was also supported by the Korea Health Technology R&D Project through KHIDI, funded by the Ministry of Health & Welfare, Republic of Korea (HI17C2098). F. O. A. was supported by Al-Gannas Qatari Society and The Cultural Village Foundation-Katara, Doha, State of Qatar and Monash University Malaysia. G. F. and E. D. J. were supported by Rockefeller University start-up funds and the Howard Hughes Medical Institute. A.A. and M.R.C. received support from the Fondazione Cariplo project no. 2018–2045 and the Italian Ministry of Education, University and Research (MIUR) for Progetti PRIN2017 20174BTC4R and Dipartimenti di Eccellenza Program (2018–2022). R. D. and S. M. received funding from Wellcome grant WT207492.

### Availability of data and materials

The mitogenome assemblies are available under the VGP BioProject PRJNA489243 as well as in the GenomeArk (<https://vgp.github.io/genomeark/>). Individual NCBI/ENA accession numbers are listed in Additional file 2: Table S1 column AA. The nuclear genome assemblies of these species are currently under the G10K embargo policy <https://genome10k.soe.ucsc.edu/data-use-policies/>. MitoVGP is available under BSD 3-Clause License through the VGP Github portal (<https://github.com/VGP/vgp-assembly>) along with a ready-to-use conda environment, fully commented code, and complete instructions and examples on how to run it. MitoVGP and the code to reproduce the analyses and the plots is available at <https://github.com/gf777/mitoVGP> [71].

### Declarations

#### Ethics approval and consent to participate

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

V. C., S. M., and D. F. are employees of Oxford Nanopore Technologies Limited. J. K. is Chief Scientific Officer of Pacific Biosciences.

**Author details**

<sup>1</sup>The Vertebrate Genome Lab, Rockefeller University, New York, NY, USA. <sup>2</sup>Laboratory of Neurogenetics of Language, Rockefeller University, New York, NY, USA. <sup>3</sup>The Howards Hughes Medical Institute, Chevy Chase, MD, USA. <sup>4</sup>Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. <sup>5</sup>Department of Biology and Biotechnology "L. Spallanzani", University of Pavia, Pavia, Italy. <sup>6</sup>Monash University Malaysia Genomics Facility, School of Science, Bandar Sunway, Selangor Darul Ehsan, Malaysia. <sup>7</sup>Tropical Medicine and Biology Multidisciplinary Platform, Monash University Malaysia, Bandar Sunway, Selangor Darul Ehsan, Malaysia. <sup>8</sup>Qatar Falcon Genome Project, Doha, State of Qatar. <sup>9</sup>Department of Environmental Science and Policy, University of Milan, Milan, Italy. <sup>10</sup>Department of Biology, New Mexico State University, Las Cruces, NM, USA. <sup>11</sup>Wellcome Sanger Institute, Cambridge, UK. <sup>12</sup>Department of Genetics, University of Cambridge, Cambridge, UK. <sup>13</sup>Max Planck Institute of Molecular Cell Biology & Genetics, Dresden, Germany. <sup>14</sup>Hoonygen, Seoul, Korea. <sup>15</sup>Pacific Biosciences, Menlo Park, CA, USA. <sup>16</sup>DNAnexus Inc., Mountain View, CA, USA. <sup>17</sup>Oxford Nanopore Technologies Ltd, Oxford Science Park, Oxford, UK. <sup>18</sup>Department of Biosciences, University of Milan, Milan, Italy. <sup>19</sup>Department of Biology, University of Florida, Gainesville, FL, USA.

Received: 17 August 2020 Accepted: 31 March 2021

Published online: 29 April 2021

**References**

- Karnkowska A, Vacek V, Zubáčová Z, Treitl SC, Petřelková R, Eme L, et al. A eukaryote without a mitochondrial organelle. *Curr Biol*. 2016;26(10):1274–84. <https://doi.org/10.1016/j.cub.2016.03.053>.
- Kolesnikov AA, Gerasimov ES. Diversity of mitochondrial genome organization. *Biochemistry*. 2012;77:1424–35.
- D'Erchia AM, Atlante A, Gadaleta G, Pavesi G, Chiara M, De Virgilio C, et al. Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion*. 2015;20:13–21. <https://doi.org/10.1016/j.mito.2014.10.005>.
- Cole LW. The evolution of per-cell organelle number. *Front Cell Dev Biol*. 2016;4:85.
- Mindell DP, Sorenson MD, Dimcheff DE. Multiple independent origins of mitochondrial gene order in birds. *Proc Natl Acad Sci U S A*. 1998;95(18):10693–7. <https://doi.org/10.1073/pnas.95.18.10693>.
- Satoh TP, Miya M, Mabuchi K, Nishida M. Structure and variation of the mitochondrial genome of fishes. *BMC Genomics*. 2016;17(1):719. <https://doi.org/10.1186/s12864-016-3054-y>.
- Gibb GC, Kardailsky O, Kimball RT, Braun EL, Penny D. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. *Mol Biol Evol*. 2007;24(1):269–80. <https://doi.org/10.1093/molbev/msl158>.
- Rob DeSalle HH. Evolutionary biology and mitochondrial genomics: 50 000 mitochondrial DNA genomes and counting; Wiley; 2017.
- Hebert PDN, Cywinka A, Ball SL, de Waard JR. Biological identifications through DNA barcodes. *Proc Biol Sci*. 2003;270(1512):313–21. <https://doi.org/10.1098/rspb.2002.2218>.
- Gibb GC, England R, Hartig G, McLenachan PAT, Taylor Smith BL, McCormish BJ, et al. New Zealand passerines help clarify the diversification of major songbird lineages during the Oligocene. *Genome Biol Evol*. 2015;7(11):2983–95. <https://doi.org/10.1093/gbe/ewv196>.
- Mindell DP, Fuchs J, Johnson JA. Phylogeny, taxonomy, and geographic diversity of diurnal raptors: Falconiformes, Accipitriformes, and Cathartiformes. In: Sarasola JH, Grande JM, Negro JJ, editors. *Birds of prey: biology and conservation in the XXI century*. Cham: Springer International Publishing; 2018. p. 3–32. [https://doi.org/10.1007/978-3-319-73745-4\\_1](https://doi.org/10.1007/978-3-319-73745-4_1).
- Miya M, Satoh TP, Nishida M. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. *Biol J Linn Soc*. 2005;289–306. <https://doi.org/10.1111/j.1095-8312.2005.00483.x>.
- Gibb GC, Condamine FL, Kuch M, Enk J, Moraes-Barros N, Superina M, et al. Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living Xenarthrans. *Mol Biol Evol*. 2016;33(3):621–42. <https://doi.org/10.1093/molbev/msv250>.
- Delsuc F, Kuch M, Gibb GC, Hughes J, Szpak P, Southon J, et al. Resolving the phylogenetic position of Darwin's extinct ground sloth (*Mylodon darwini*) using mitogenomic and nuclear exon data. *Proc Biol Sci*. 2018;285. <https://doi.org/10.1098/rspb.2018.0214>.
- Delsuc F, Kuch M, Gibb GC, Karpinski E, Hackenberger D, Szpak P, et al. Ancient mitogenomes reveal the evolutionary history and biogeography of sloths. *Curr Biol*. 2019;29:2031–42.e6.
- Harrison RG. Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends Ecol Evol*. 1989;4(1):6–11. [https://doi.org/10.1016/0169-5347\(89\)90006-2](https://doi.org/10.1016/0169-5347(89)90006-2).
- Galtier N, Nabholz B, Glémin S, Hurst GDD. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol*. 2009;18(22):4541–50. <https://doi.org/10.1111/j.1365-294X.2009.04380.x>.
- Allio R, Donega S, Galtier N, Nabholz B. Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol Biol Evol*. 2017;34(11):2762–72. <https://doi.org/10.1093/molbev/msx197>.
- Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. Identification of birds through DNA barcodes. *PLoS Biol*. 2004;2(10):e312. <https://doi.org/10.1371/journal.pbio.0020312>.
- Ward RD, Hanner R, Hebert PDN. The campaign to DNA barcode all fishes, FISH-BOL. *J Fish Biol*. 2009;74(2):329–56. <https://doi.org/10.1111/j.1095-8649.2008.02080.x>.

21. Ivanova NV, Clare EL, Borisenko AV. DNA barcoding in mammals. *Methods Mol Biol.* 2012;858:153–82. [https://doi.org/10.1007/978-1-61779-591-6\\_8](https://doi.org/10.1007/978-1-61779-591-6_8).
22. Kornienko IV, Faleeva TG, Oreshkova NV, Grigoriev SE, Grigorieva LV, Putintseva YA, et al. Structural and functional organization of the mitochondrial DNA control region in the woolly mammoth (*Mammuthus primigenius*). *Mol Biol.* 2019;53(4):560–70. <https://doi.org/10.1134/S002689331904006X>.
23. Huang Z, Shen Y, Ma Y. Structure and variation of the Fringillidae (Aves: Passeriformes) mitochondrial DNA control region and their phylogenetic relationship. *Mitochondrial DNA A DNA Mapp Seq Anal.* 2017;28(6):867–71. <https://doi.org/10.1080/24701394.2016.1199023>.
24. Bronstein O, Kroh A, Haring E. Mind the gap! The mitochondrial control region and its power as a phylogenetic marker in echinoids. *BMC Evol Biol.* 2018;18:80.
25. Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, et al. Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proc Natl Acad Sci U S A.* 1989;86(16):6196–200. <https://doi.org/10.1073/pnas.86.16.6196>.
26. Picardi E, Pesole G. Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat Methods.* 2012;9(6):523–4. <https://doi.org/10.1038/nmeth.2029>.
27. Mackiewicz P, Urantówka AD, Krocak A, Mackiewicz D. Resolving phylogenetic relationships within Passeriformes based on mitochondrial genes and inferring the evolution of their mitogenomes in terms of duplications. *Genome Biol Evol.* 2019;11(10):2824–49. <https://doi.org/10.1093/gbe/evz209>.
28. Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet.* 2001;69(5):1113–26. <https://doi.org/10.1086/324024>.
29. Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.* 2010;6(2):e1000834. <https://doi.org/10.1371/journal.pgen.1000834>.
30. Maude H, Davidson M, Charitakis N, Diaz L, Bowers WHT, Gradovich E, et al. NUMT confounding biases mitochondrial heteroplasmy calls in favor of the reference allele. *Front Cell Dev Biol.* 2019;7:201. <https://doi.org/10.3389/fcell.2019.00201>.
31. Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Hum Mutat.* 2004;23(2):125–33. <https://doi.org/10.1002/humu.10304>.
32. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res.* 2014;42(20):12640–9. <https://doi.org/10.1093/nar/gku1038>.
33. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J.* 2020;18:9–19. <https://doi.org/10.1016/j.csbj.2019.11.002>.
34. Koepfli K-P, Paten B. Genome 10K Community of Scientists, O'Brien SJ. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci.* 2015;3(1):57–111. <https://doi.org/10.1146/annurev-animal-090414-014900>.
35. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature.* 2021. <https://doi.org/10.1038/s41586-021-03451-0>.
36. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods.* 2016;13(12):1050–4. <https://doi.org/10.1038/nmeth.4035>.
37. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27(5):722–36. <https://doi.org/10.1101/gr.215087.116>. PMID: 28298431. PMCID: PMC5411767.
38. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* 2015;16(1):294. <https://doi.org/10.1186/s13059-015-0849-0>.
39. Soorni A, Haak D, Zaitlin D, Bombarely A. Organelle\_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics.* 2017;18. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5219736/>. [cited 2020 Mar 28]
40. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 2017;45:e18.
41. Soares AER, Novak BJ, Haile J, Heupink TH, Fjeldså J, Gilbert MTP, et al. Complete mitochondrial genomes of living and extinct pigeons revise the timing of the columbiform radiation. *BMC Evol Biol.* 2016;16(1):230. <https://doi.org/10.1186/s12862-016-0800-3>.
42. Pratt RC, Gibb GC, Morgan-Richards M, Phillips MJ, Hendy MD, Penny D. Toward resolving deep neoaves phylogeny: data, signal enhancement, and priors. *Mol Biol Evol.* 2009;26(2):313–26. <https://doi.org/10.1093/molbev/msn248>.
43. Harrison GLA, McLenachan PA, Phillips MJ, Slack KE, Cooper A, Penny D. Four new avian mitochondrial genomes help get to basic evolutionary questions in the late cretaceous. *Mol Biol Evol.* 2004;21(6):974–83. <https://doi.org/10.1093/molbev/msh065>.
44. Qin P-S, Tao C-R, Yin S, Li H-M, Zeng D-L, Qin X-M. Complete mitochondrial genome of *Lacerta agilis* (Squamata, Lacertidae). *Mitochondrial DNA.* 2014;25(6):416–7. <https://doi.org/10.3109/19401736.2013.809436>.
45. Doyle JM, Katzner TE, Bloom PH, Ji Y, Wijayawardena BK, DeWoody JA. The genome sequence of a widespread apex predator, the golden eagle (*Aquila chrysaetos*). *Plos One.* 2014;9(4):e95599. <https://doi.org/10.1371/journal.pone.0095599>.
46. Yang C, Wang Q-X, Li X-J, Yuan H, Xiao H, Huang Y. The mitogenomes of *Gelochelidon nilotica* and *Sterna hirundo* (Charadriiformes, Sternidae) and their phylogenetic implications. *Mitochondrial DNA Part B.* 2017;2(2):601–3. <https://doi.org/10.1080/23802359.2017.1372709>.
47. Inoue JG, Miya M, Tsukamoto K, Nishida M. Mitogenomic evidence for the monophyly of elopomorph fishes (Teleostei) and the evolutionary origin of the leptocephalus larva. *Mol Phylogenet Evol.* 2004;32(1):274–86. <https://doi.org/10.1016/j.ympev.2003.11.009>.
48. Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, et al. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science.* 2011;334(6055):521–4. <https://doi.org/10.1126/science.1211028>.
49. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science.* 2014;346(6215):1320–31. <https://doi.org/10.1126/science.1253451>.

50. Malmstrøm M, Matschiner M, Tørresen OK, Star B, Snipen LG, Hansen TF, et al. Evolution of the immune system influences speciation rates in teleost fishes. *Nat Genet.* 2016;48(10):1204–10. <https://doi.org/10.1038/ng.3645>.
51. Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, et al. Phylogenetic classification of bony fishes. *BMC Evol Biol.* 2017;17(1):162. <https://doi.org/10.1186/s12862-017-0958-3>.
52. Wang X, Liu N, Zhang H, Yang X-J, Huang Y, Lei F. Extreme variation in patterns of tandem repeats in mitochondrial control region of yellow-browed tits (*Sylviparus modestus*, Paridae). *Sci Rep.* 2015;5(1):13227. <https://doi.org/10.1038/srep13227>.
53. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. *Genome Biol.* 2013;14(5):R51. <https://doi.org/10.1186/gb-2013-14-5-r51>.
54. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019;47(W1):W256–9. <https://doi.org/10.1093/nar/gkz239>.
55. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019;37(10):1155–62. <https://doi.org/10.1038/s41587-019-0217-9>.
56. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 2013;41(13):e129. <https://doi.org/10.1093/nar/gkt371>.
57. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34(18):3094–100. <https://doi.org/10.1093/bioinformatics/bty191>.
58. Konrad A, Thompson O, Waterston RH, Moerman DG, Keightley PD, Bergthorsson U, et al. Mitochondrial mutation rate, spectrum and heteroplasmy in *Caenorhabditis elegans* spontaneous mutation accumulation lines of differing population size. *Mol Biol Evol.* 2017;34(6):1319–34. <https://doi.org/10.1093/molbev/msx051>.
59. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357–9. <https://doi.org/10.1038/nmeth.1923>.
61. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]*. 2012;1207.3907v2.
62. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *Plos Comput Biol.* 2018;14:e1005944.
63. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol.* 2019:1–14. [https://doi.org/10.1007/978-1-4939-9173-0\\_1](https://doi.org/10.1007/978-1-4939-9173-0_1).
64. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. doi: <https://doi.org/10.1101/2020.03.15.992941>
65. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol Phylogenet Evol.* 2013;69(2):313–9. <https://doi.org/10.1016/j.ympev.2012.08.023>.
66. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, Timetrees, and divergence times. *Mol Biol Evol.* 2017;34(7):1812–9. <https://doi.org/10.1093/molbev/msx116>.
67. Stöver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics.* 2010;11(1):7. <https://doi.org/10.1186/1471-2105-11-7>.
68. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
69. Lindenbaum P. Jvarkit: java-based utilities for Bioinformatics. *FigShare.* 2015;10:m9.
70. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat Biotechnol.* 2011;29(1):24–6. <https://doi.org/10.1038/nbt.1754>.
71. Formenti G. gf777/ mitoVGP: Paper release 1. 2021. Available from: <https://zenodo.org/record/4636722>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

