SMART2: Multi-Library Statistical Mitogenome Assembly with Repeats

Fahad Alqahtani[1,2] and Ion I. Măndoiu[1]

[1] Computer Science & Engineering Department, University of Connecticut, Storrs, CT, USA, {fahad.alqahtani,ion.mandoiu}@uconn.edu
[2] National Center for Artificial Intelligence and Big Data Technology, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

Abstract. SMART2 is an enhanced version of the SMART pipeline for mitogenome assembly from low-coverage whole-genome sequencing (WGS) data. Novel features include automatic selection of the optimal number of read pairs used for assembly and the ability to assemble multiple sequencing libraries when available. SMART2 succeeded in generating mitochondrial sequences for 26 metazoan species with WGS data but no previously published mitogenomes in NCBI databases. The SMART2 pipeline is publicly available via a user-friendly Galaxy interface at `https://neo.engr.uconn.edu/?tool_id=SMART2`.

Keywords: Mitogenome assembly, multi-library assembly, low-coverage sequencing

1 Introduction

Mitochondria are cellular organelles present with very rare exceptions in all eukaryotic cells. In most animals, the mitochondria have their own genome, a double-stranded circular DNA molecule typically ranging in size between 15-20Kb that encodes 37 genes (2 ribosomal RNA genes, 13 protein coding genes, and 22 transfer RNA genes). The mitochondrial genome is inherited maternally, and has much higher copy number than the nuclear genome [24]. The small size, high copy number, and the presence of both coding and regulatory regions that mutate at different rates make the mitochondrial genome an ideal genetic marker. Indeed, mitochondrial sequences have been used in applications ranging from maternal ancestry inference and tracing human migrations [6] to forensic analysis [19]. The mitochondrial DNA has also become the workhorse of biodiversity studies since many non-model species do not have yet a sequenced nuclear genome [12,16].

To date, most such biodiversity studies have been based on sequencing a single gene fragment, such as the Cytochrome C oxidase I (COI) gene, which has been adopted as the preferred "barcode of life" [14,21]. Recently there have been a renewed appreciation for the improved accuracy of taxonomic and phylogenetic analyses performed based on complete mitogenome sequences assembled from low coverage whole genome shotgun (WGS) reads generated using next generation sequencing (NGS) technologies. Indeed, full length mitogenome sequences capture evolutionary events such as genome

rearrangements that are missed in single gene analyses [18]. Furthermore, the exponential decrease in NGS costs has led to an explosion in the number of WGS datasets generated from non-model organisms. For mammals alone, there are currently over two hundred species with paired-end WGS data available in the NCBI SRA database but for which no complete mitogenome is available. Recent studies have also demonstrated that WGS data of sufficient depth for reconstructing mitogenomes can be generated from preserved museum specimens [23], making the approach applicable to rare or even extinct species.

Leveraging the available WGS datasets to expand the number of complete mitogenomes requires bioinformatics pipelines that can assemble and annotate high-quality mitogenomes quickly and with minimal human intervention. Unfortunately, standard genome assemblers often fail to generate high quality mitochondrial genome sequences due to the large difference in copy number between the mitochondrial and nuclear genomes [13]. This has led to the development of specialized tools for reconstructing mitochondrial genomes from WGS data, mainly falling within three categories. *Reference-based* methods such as MToolBox [8] require the mtDNA sequence of the species of interest or a closely related species, which are often not available for the less-studied species of interest in biodiversity studies. *Seed-and-extend* tools such as MITObim [13] and NOVOPlasty [11] use a greedy approach to extend available seed sequences such as the COI but can have difficulty handling repetitive regions present in some mitochondrial genomes [16]. Finally, *de novo* methods such as Norgal [2] and plasmidSPAdes [7] use coverage-based filtering to remove nuclear WGS reads before performing assembly using the de Bruijn graph of remaining reads.

In [5] we introduced a hybrid method called *Statistical Mitogenome Assembly with RepeaTs* (SMART), which uses a seed sequence to estimate the mean and standard deviation of mtDNA k-mer counts, then positively selects reads with k-mer counts falling within three standard deviations of the estimated mean before performing *de novo* assembly. Experiments in [3] show that for low-depth WGS datasets the positive selection approach implemented by SMART yields higher enrichment for mtDNA reads than the negative selection of Norgal. Furthermore, SMART was shown to produce complete circular mitogenomes with a higher success rate than both seed-and-extend tools MITObim and NOVOPlasty and *de novo* assemblers Norgal and plasmidSPAdes.

In this paper we present an extension of the SMART pipeline, referred to as SMART2, that can take advantage of multiple sequencing libraries when available and automatically selects the optimal number of read pairs used for assembly. We also present experimental results comparing read filtering and assembly accuracy of SMART2 with that of existing state-of-the-art tools, along with the results of a pilot "orphan mitogenomes" project in which SMART2 was used to generate 15 complete and 11 partial mitogenomes for 26 mammals and amphibians without previously published mitogenomes. All novel mitogenomes have been submitted to GenBank as Third Party Annotation (TPA) sequences [9].

2 Methods

The SMART2 pipeline is deployed using a customized instance of the Galaxy framework [1] and is publicly available via a user-friendly Galaxy interface at `https://neo.engr.uconn.edu/?tool_id=SMART2` (see Fig. 1). The pipeline was designed for processing paired-end reads in fastq format from one or two WGS libraries. In addition to fastq files, the user specifies the sample name and a seed sequence in fasta format. By default

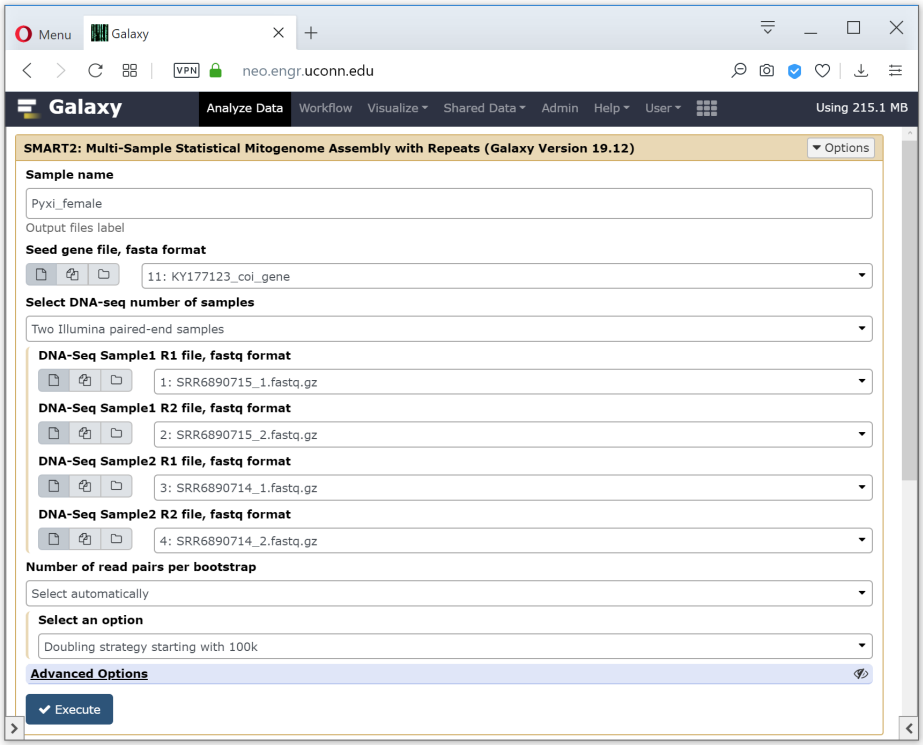


Fig. 1. Galaxy interface of SMART2.

the number of reads is selected automatically as described below, but the user can override the default and manually specify it. Advanced options also allow the user to change the default choices for the number of bootstrap samples (default is 1), k-mer size (default is 31), number of threads (default is 16), and the genetic code used for MITOS annotation (default is the vertebrate mitochondrial code).

The main steps of the SMART2 pipeline follow those of SMART with adaptations for multi-library inputs:

1. Automatic adapter detection and trimming, performed independently for each library.
2. Random resampling of a number of trimmed read pairs, either specified by the user or automatically determined using the doubling strategy described below.
3. Selection of mitochondrial reads based on coverage estimates of seed sequence k-mers – aggregated across libraries using one of the methods described below (2-dimensional Gaussian mixture modeling using MCLUST, Union, or Intersection).
4. Joint preliminary assembly of reads passing the coverage filter in the two libraries, performed using SPAdes.
5. Filtering of preliminary contigs by BLAST searches against a local mitochondrial database.
6. Secondary read filtering by alignment to preliminary contigs that have significant BLAST matches, performed independently for each library.

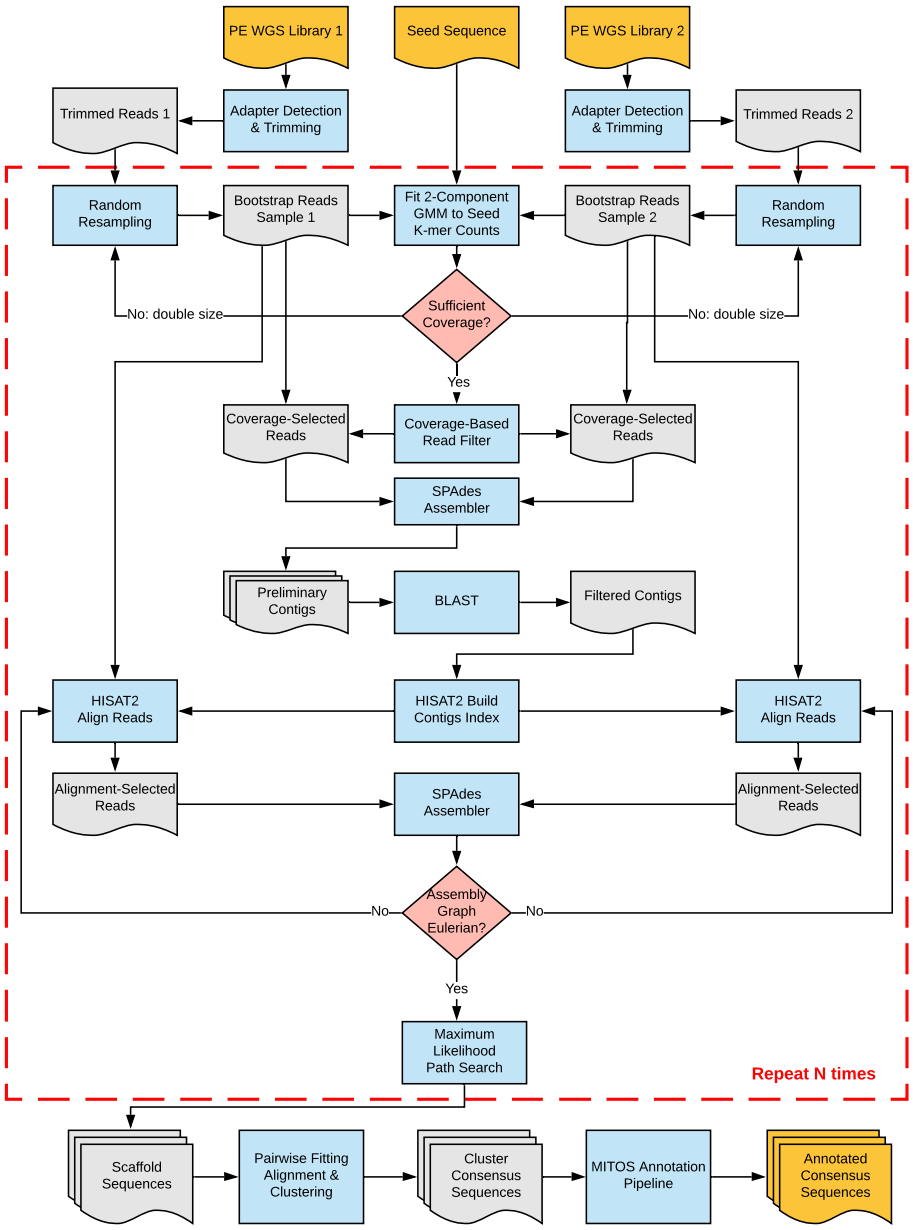
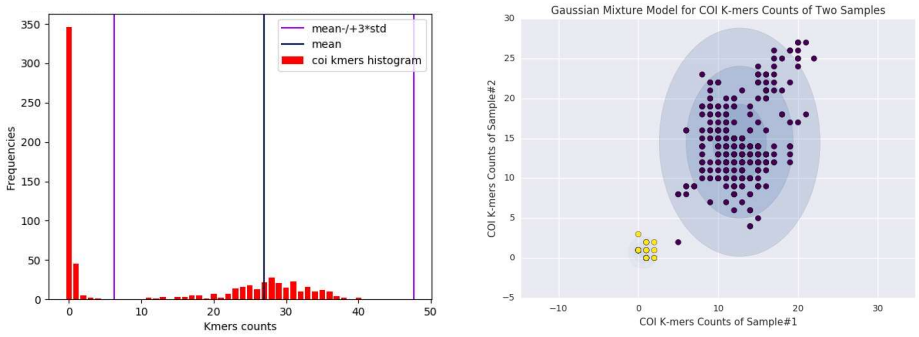


Fig. 2. SMART2 workflow.



(a) (b)

Fig. 3. Mitochondrial k-mer coverage distribution estimated by MCLUST using seed k-mer counts generated from (a) 800k read pairs sampled from library SRR630623 of the *Anopheles stephensi* dataset, and (b) 400k read pairs sampled from each of the two libraries of the *Anopheles stephensi* dataset.

7. Joint secondary assembly of selected reads, performed using SPAdes.
8. Iterative scaffolding and gap filling based on maximum likelihood.
9. Prediction and annotation of mitochondrial genes using MITOS.

As for SMART, steps 2-8 of SMART2 can be repeated a user-specified number of times to compute the bootstrap support for the assembled sequences. A detailed flowchart of the SMART2 pipeline is shown in Fig. 2.

2.1 Coverage-based k-mer classification

For a single library SMART2 uses the same method as SMART for classifying k-mers as mitochondrial or nuclear in origin. Specifically, SMART2 uses MCLUST [22] to fit a two-component Gaussian mixture model to the one-dimensional distribution of counts of seed sequence k-mers. The upper component of the fitted model is taken as a proxy for the corresponding mtDNA k-mer count distribution, and all k-mers that have a count within 3 standard deviations of the estimated upper component mean are classified as mitochondrial (see Fig. 3(a)).

For two libraries the natural extension of this approach would be to fit a two-component Gaussian mixture model to the *two-dimensional* distribution of counts of seed sequence k-mers (see Fig. 3(b)). Unfortunately experimental results in Section 3 show that this approach (referred to as "MCLUST") has relatively poor read filtering performance. Consequently, we implemented in SMART2 two alternative approaches for k-mer classification. Both rely on first independently classifying each k-mer as mitochondrial or nuclear based on fitting two-component Gaussian mixture models to the one-dimensional distributions of counts of seed sequence k-mers of each library. The "Union" method ultimately classifies a k-mer as mitochondrial if it is classified as such based on either one of the libraries, while the "Intersection" method does so if the k-mer is classified as mitochondrial according to *both* libraries.

Table 1. Multi-library WGS datasets with published mtDNA sequences.

Species	Library ID	Read Length	% mtDNA	Seed ID	Seed Length	Reference ID	Reference Length
A. stephensi	SRR630623	2×101	0.041%	MK726121	704	KT899888	15,371
	SRR630669	2×101	0.041%				
A. funestus	SRR630620	2×101	0.03%	MK300232	709	MG742199	15,349
	SRR630619	2×101	0.032%				
D. mauritiana	SRR1560275	2×76	1.033%	HM630860	560	AF200830	14,964
	SRR1560276	2×76	1.120%				
P. major	SRR2961765	2×100	0.313%	GQ482300	694	NC_040875	16,777
	SRR2961767	2×100	0.313%				
P. humilis	SRR765709	2×101	0.215%	EU382177	620	KP001174	16,758
	SRR765710	2×101	0.294%				

2.2 Automatic selection of bootstrap sample size

The number of read pairs in a bootstrap sample has a significant effect on the quality of resulting assembly. Too small a number of reads may produce fragmented assemblies due to lack of coverage for some regions. Too large a number may be detrimental by increasing the complexity of the assembly graph and making it more difficult to remove tangles generated by sequencing errors. In the original version of SMART [5] the number of read pairs in a bootstrap sample is specified by the user, and this can lead to many trial-and-error runs to find the optimal coverage.

In SMART2 we implemented a simple doubling strategy for automatically selecting the number of read pairs used in each bootstrap sample. Based on SMART experiments with manually specified numbers of read pairs we noted that a mean read coverage of the mitochondrial genome between 20× and 40× generates complete mitogenomes with high success rate. Unfortunately, it is difficult to analytically estimate the number of read pairs that yields a mitochondrial coverage in this range since the percentage of mitochondrial reads in real WGS datasets can vary by orders of magnitude [3] and the exact sizes of the nuclear and mitochondrial genomes are often not known *a priori*. For a single WGS library, SMART2 starts with 100,000 read pairs and then iteratively doubles the number of pairs until reaching an estimated mean mitochondrial read coverage of 20× or more. For two WGS libraries, SMART2 uses a similar doubling strategy starting with 100,000 read pairs and stopping when the *sum* of the mean mitochondrial read coverages estimated from the two libraries is 20× or more.

3 Results and Discussion

3.1 Comparison of coverage-based filters and assembly accuracy on WGS datasets from species with published mitogenomes

For a detailed assessment, including evaluating the effectiveness of the SMART2 coverage-based filters and comparing assembly accuracy with previous methods we used five two-library datasets from species with published mitogenomes. The datsets are comprised of three insects (*Anopheles stephensi*, *Anopheles funestus*, and *Drosophila mauritiana*) and two birds (*Parus major* and *Pseudopodoces humilis*). Accession numbers and basic statistics for the five datasets are provided in Table 1.

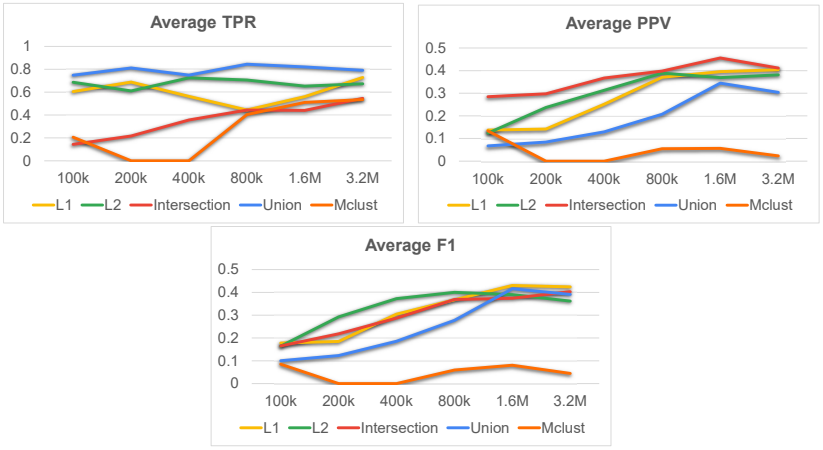


Fig. 4. Accuracy of single and multi-library coverage-based filters on 100k-3.2M read pairs randomly selected from libraries in Table 1.

Fig. 4 plots the *True Positive Rate* (TPR), *Positive Predictive Value* (PPV), and F1 score (harmonic mean of TPR and PPV) achieved by the MCLUST, Union, and Intersection filters of SMART2 as the total number of read pairs is varied between 100k and 3.2M. All values are averages over the five species in Table 1. For comparison we include the average TPR, PPV, and F1 score of single library filters (L1 and L2). The results underscore the poor performance of the 2-dimensional mixture model (MCLUST), and the different tradeoffs achieved between TPR and PPV by the Union and Intersection filters. Specifically, for a fixed number of reads, the Union filter typically achieves a higher TPR but lower PPV than single library filters, while the Intersection filter does the opposite. In these experiments, the Intersection filter yields an F1 score comparable with single library filters for the lower range of tested number of read pairs, but both the Union and Intersection filters converge towards the performance of single library filters as the number of read pairs exceeds one million.

Assembly accuracy results generated by SMART2 and three other tools (Norgal [2], NOVOPlasty [11], and PlasmidSPAdes [7]) on the datasets described in Table 1 are given in Table 2. The number of read pairs used for assembly, indicated in last column for both single and two-library runs, was selected using the doubling strategy implemented described in Section 2. For each method, the assembled sequence length and percentage identity to the published reference are typeset in bold when the reconstructed sequence is a circular genome.

On all datasets Norgal failed to generate any contigs or generated nuclear rather than mitochondrial contigs, consistent with the poor performance reported for low-coverage WGS data in [3]. NOVOPlasty generated circular mitogenomes from two of the ten libraries, but failed on one library, and generated only incomplete mitogenomes from the remaining seven. PlasmidSPAdes generated circular mitogenomes from three of the ten libraries, while SMART2 succeed on five of the ten single-library runs and two of the five two-library runs.

Table 2. Assembled sequence length and percentage identity to the published reference for low-coverage WGS datasets from species with published mitogenomes. Numbers in bold indicate a complete circular mitogenome.

Species	Library	Norgal	NOVOPlasty	PlasmidSPAdes	SMART2	#Pairs
A. stephensi	SRR630623	nuclear	2,962 66.6%	14,974 99.5%	15,153 99.6%	800k
	SRR630669	nuclear	2,428 99.8%	15,324 99.5%	**15,412** **99.5**	800k
	Both	N/A	N/A	N/A	15,283 99.8%	2×400k
A. funestus	SRR630620	-	2,105 99.6%	12,819 41.6%	13,424 99.5%	800k
	SRR630619	-	2,402 99.4%	15,176 99.5%	13,369 99.4%	800k
	Both	N/A	N/A	N/A	10,502 99.5%	2×400k
D. mauritiana	SRR1560275	nuclear	14,922 99.9%	15,411 96.5%	15,462 96.7%	400k
	SRR1560276	nuclear	9,327 99.9%	15,245 97.9%	15,643 95.3%	400k
	Both	N/A	N/A	N/A	15,397 97%	2×200k
P. major	SRR2961765	-	**16,774** **99.8%**	**16,791** **99.7%**	**16,814** **99.6%**	1,6M
	SRR2961767	nuclear	**16,774** **99.8%**	**16,790** **99.7%**	**16,813** **99.6%**	1.6M
	Both	N/A	N/A	N/A	**16,814** **99.6%**	2×800k
P. humilis	SRR765709	nuclear	-	16,852 98.8%	**16,797** **99.1%**	1.6M
	SRR765710	-	8,139 99.5%	**16,774** **99.3%**	**16,797** **99.1%**	800k
	Both	N/A	N/A	N/A	**16,797** **99.1%**	2×400k

3.2 SMART2 assembles novel mitogenomes

In a pilot project to assemble "orphan mitogenomes" for species with publicly available WGS data but no published mitogenome sequence we ran SMART2 on WGS datasets from 18 mammals (*Abrocoma cinerea*, *Arvicola amphibius*, *Babyrousa babyrussa*, *Canis rufus*, *Coendou bicolor*, *Cratogeomys planiceps*, *Ctenodactylus gundi*, *Cuniculus paca*, *Grammomys surdaster*, *Heteromys oasicus*, *Hippotragus niger kirkii*, *Hippotragus niger niger*, *Pipistrellus pipistrellus*, *Pusa hispida saimensis*, *Rhacophorus chenfui*, *Sciurus carolinensis*, *Sorex palustris*, and *Urocitellus parryii*) and 8 amphibians (*Agalychnis moreletii*, *Brachycephalus ferruginus*, *Brachycephalus pombali*, *Cycloramphus boraceiensis*, *Hyla arborea*, *Hylodes phyllodes*, *Melanophryniscus xanthostomus*, and *Oophaga pumilio*). Basic information about the 26 datasets is given in Table 3. The number of read pairs was selected automatically using the doubling strategy for all datasets except

Table 3. WGS datasets from 26 metazoans without published mitogenomes. mtDNA content was estimated by aligning the reads against the SMART2 assembly only when the latter was a complete sequence. The number of read pairs was selected automatically by using the doubling strategy described in Section 2 except for the three species marked with a dagger for which it was manually increased after automatic selection failed to assemble a complete circular mitogenome. A "*" indicates datasets for which all available read pairs were used.

Species	Run ID	Read Length	% mtDNA	#Pairs Used	Seed ID
Abrocoma cinerea	SRR8885043	2×151	1.490	2,000,000†	AF244388
Agalychnis moreletii	SRR8327212	2×182	NA	1,600,000	EF125031
Arvicola amphibius	ERR3316036	2×151	0.002	51,200,000	LT546162
Babyrousa babyrussa	ERR2984475	2×100	0.022	12,800,000	AY534302
Brachycephalus ferruginus	SRR5837605	2×251	NA	856,599*	HQ435708
Brachycephalus pombali	SRR5837604	2×251	NA	846,282*	HQ435714
Canis rufus	SRR8066613	2×101	0.565	400,000	U47043
Coendou bicolor	SRR8885018	2×151	3.372	100,000	U34852
Cratogeomys planiceps	SRS4613652	2×151	2.537	100,000	AY545541
Ctenodactylus gundi	SRR8885020	2×151	0.246	400,000	U67301
Cuniculus paca	SRS4613635	2×151	0.371	400,000	JF459150
Cycloramphus boraceiensis	SRR4019528	2×305	NA	1,776,547*	KU494395
Grammomys surdaster	SRS4524074	2×151	0.689	10,000,000†	KY753991
Heteromys oasicus	SRR8885041	2×151	0.965	200,000	ABCSA423-06
Hippotragus niger kirkii	SRS4184270	2×101	0.017	25,600,000	AF049388
Hippotragus niger niger	SRR8366604	2×101	0.012	51,200,000	AF049393
Hyla arborea	SRR2157967	2×101	NA	10,000,000†	JN312692
Hylodes phyllodes	SRR4019434	2×305	NA	1,055,455*	DQ502873
Melanophryniscus xanthostomus	SRR5837589	2×251	NA	977,403*	KX025607
Oophaga pumilio	SRR7627571	2×49	NA	3,200,000	KX574023
Pipistrellus pipistrellus	ERR3316150	2×151	0.007	25,600,000	HM380206
Pusa hispida saimensis	ERR2608991	2×170	0.098	1,600,000	JX109798
Rhacophorus chenfui	SRR5248583	2×300	NA	3,477,603*	KP996818
Sciurus carolinensis	ERR3312500	2×151	2.791	100,000	JF457099
Sorex palustris	SRR8451745	2×150	NA	6,400,000	MG421461
Urocitellus parryii	SRR8263911	2×151	0.609	200,000	KX646821

A. cinerea, *G. surdaster*, and *H. arborea*, for which we manually increased the number of read pairs after automatic selection failed to assemble complete circular genomes.

As shown in Table 4, out of the 26 datasets, SMART2 generated 15 complete circular mitogenomes and 11 partial mitogenomes, for a total of 403,541 bp. NOVOPlasty and PlasmidSPAdes generate only 5 and 1 complete circular mitogenomes, respectively. As seen in Fig. 5, when all three methods succeed, agreement between the assembled sequences is very high. However, NOVOPlasty and PlasmidSPAdes have a much higher failure rate than SMART2, generating a total of only 258,538 bp and 224,818 bp of mitogenomic sequences, respectively.

To further assess the accuracy of mitogenomes assembled by SMART2 we performed a joint phylogenetic analysis with published complete mitogenome sequences of up to two species in the same family, whenever the latter could be identified (see Table 4 for accession numbers). The joint phylogeny annotated using iTOL [17] is shown

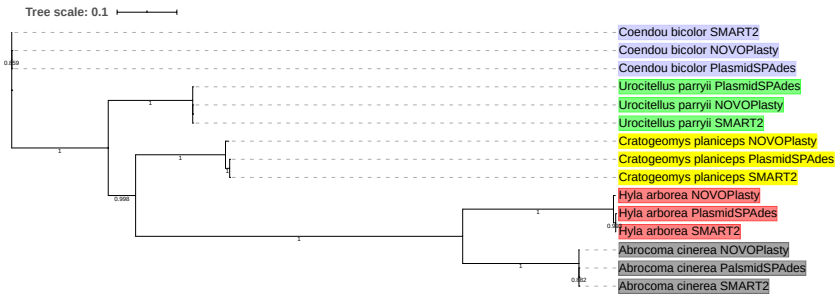


Fig. 5. Phylogenetic tree of mitogenomes assembled by SMART2, NOVOPlasty, and PlasmidSPAdes.

in Figure 6. The phylogeny was constructed using FastTree [20] with 10,000 bootstraps and the jModelTest [10] model of sequence evolution from a multiple alignment generated using MAFFT [15]. The phylogeny places the sequences of each family within independent clades, supporting the accuracy of SMART2 assemblies. Assembly accuracy is further supported by the completeness of MITOS annotations (see Table 4 for the number of annotated genes for each species). All mtDNA sequences assembled by SMART2 for the 26 species in the pilot project have been submitted to GenBank as Third Party Annotation (TPA) sequences (see Table 4 for TPA accession numbers).

4 Conclusions

In this paper we presented SMART2, an enhanced pipeline that can assemble high quality mitochondrial genomes from low coverage WGS datasets with minimal user intervention. SMART2 succeeded in generating mitochondrial sequences – including 15 complete circular mitogenomes – for 26 metazoan species with WGS data but no previously published mitogenomes in NCBI databases. An additional complete mitogenome assembled using the multi-library feature of SMART2 will be published separately [4]. The SMART2 pipeline is publicly available via a user-friendly Galaxy interface at `https://neo.engr.uconn.edu/?tool_id=SMART2`.

References

1. Afgan, E., *et al.*: The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research **46**(W1), W537–W544 (2018)
2. Al-Nakeeb, K., Petersen, T.N., Sicheritz-Pontén, T.: Norgal: extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. BMC bioinformatics **18**(1), 510 (2017)
3. Alqahtani, F., Mandoiu, I.: Statistical mitogenome assembly with repeats. Journal of Computational Biology, online ahead of print (2020), `https://doi.org/10.1089/cmb.2019.0505`
4. Alqahtani, F., Duckett, D., Pirro, S., Măndoiu, I.I.: Complete mitochondrial genome of water vole, *Microtus richardsoni*. in preparation (2020)

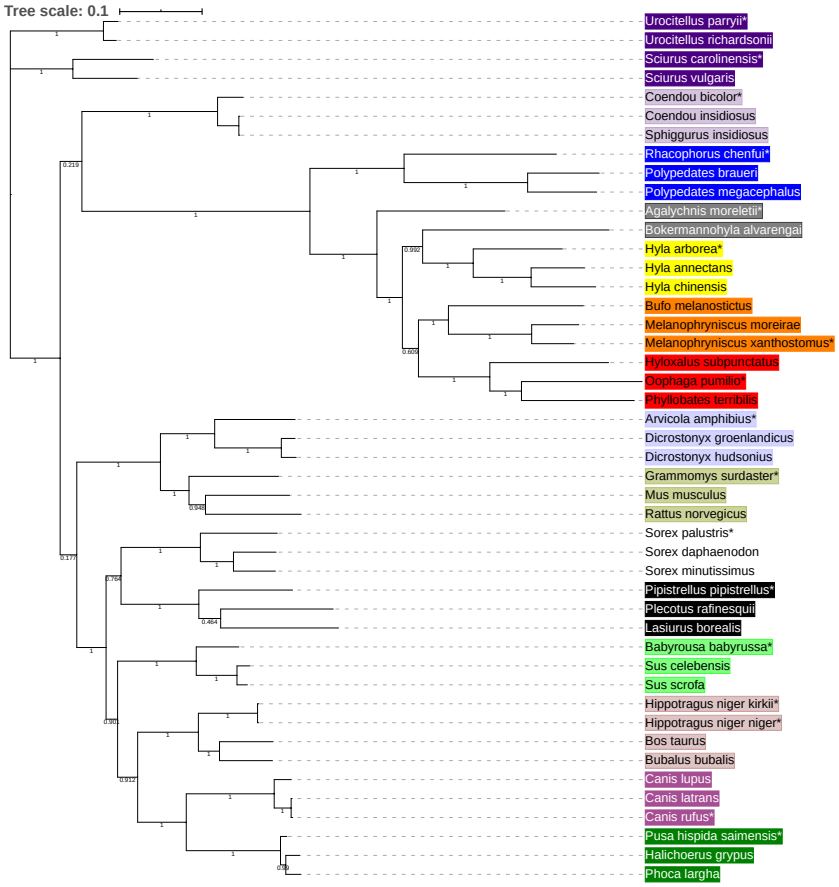


Fig. 6. Phylogenetic tree comparing SMART2 mitogenomes with published mitogenomes of related species.

5. Alqahtani, F., Măndoiu, I.I.: Statistical mitogenome assembly with repeats. In: 8th IEEE International Conference on Computational Advances in Bio and Medical Sciences (2018)
6. Alves-Silva, J., da Silva Santos, M., Guimarães, P.E., Ferreira, A.C., Bandelt, H.J., Pena, S.D., Prado, V.F.: The ancestry of brazilian mtdna lineages. The American Journal of Human Genetics **67**(2), 444–461 (2000)
7. Antipov, D., Hartwick, N., Shen, M., Raiko, M., Lapidus, A., Pevzner, P.A.: plasmidSPAdes: assembling plasmids from whole genome sequencing data. Bioinformatics **32**(22), 3380–3387 (2016)
8. Calabrese, C., Simone, D., Diroma, M.A., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., Attimonelli, M.: MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics **30**(21), 3115–3117 (2014)
9. Cochrane, G., Bates, K., Apweiler, R., Tateno, Y., Mashima, J., Kosuge, T., Mizrachi, I.K., Schafer, S., Fetchko, M.: Evidence standards in experimental and

inferential insdc third party annotation data. Omics: a journal of integrative biology **10**(2), 105–113 (2006)
10. Darriba, D., Taboada, G.L., Doallo, R., Posada, D.: jmodeltest 2: more models, new heuristics and parallel computing. Nature methods **9**(8), 772 (2012)
11. Dierckxsens, N., Mardulyn, P., Smits, G.: NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic acids research **45**(4), e18–e18 (2016)
12. Gupta, A., Bhardwaj, A., Sharma, P., Pal, Y., et al.: Mitochondrial DNA-a tool for phylogenetic and biodiversity search in equines. Journal of Biodiversity & Endangered Species **2015** (2015)
13. Hahn, C., Bachmann, L., Chevreux, B.: Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic acids research **41**(13), e129–e129 (2013)
14. Hebert, P.D., Ratnasingham, S., de Waard, J.R.: Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London. Series B: Biological Sciences **270**(suppl_1), S96–S99 (2003)
15. Katoh, K., Misawa, K., Kuma, K.i., Miyata, T.: Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic acids research **30**(14), 3059–3066 (2002)
16. Kurabayashi, A., Sumida, M.: Afrobatrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes. BMC genomics **14**(1), 633 (2013)
17. Letunic, I., Bork, P.: Interactive tree of life (itol) v4: recent updates and new developments. Nucleic acids research (2019)
18. Li, W.X., Zhang, D., Boyce, K., Xi, B.W., Zou, H., Wu, S.G., Li, M., Wang, G.T.: The complete mitochondrial dna of three monozoic tapeworms in the caryophyllidea: a mitogenomic perspective on the phylogeny of eucestodes. Parasites & Vectors **10**(1), 314 (2017)
19. Melton, T., Holland, C., Holland, M.: Forensic mitochondria dna analysis: Current practice and future potential. Forensic science review **24**(2), 101 (2012)
20. Price, M.N., Dehal, P.S., Arkin, A.P.: Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution **26**(7), 1641–1650 (2009)
21. Ratnasingham, S., Hebert, P.D.: BOLD: The barcode of life data system (http://www.barcodinglife.org). Molecular Ecology Notes **7**(3), 355–364 (2007)
22. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal **8**(1), 205–233 (2016)
23. Trevisan, B., Alcantara, D.M., Machado, D.J., Marques, F.P., Lahr, D.J.: Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. PeerJ **7**, e7543 (2019)
24. Veltri, K.L., Espiritu, M., Singh, G.: Distinct genomic copy number in mitochondria of different mammalian organs. Journal of cellular physiology **143**(1), 160–164 (1990)

Table 4: Mitochondrial sequences assembled for 26 metazoans without previously published mitogenomes. Sequence lengths typeset in bold indicate complete circular mitogenomes. Up to two complete mitogenomes from species in the same family were used for phylogenetic validation.

Species	Related Species	GenBank Accession	NOVO Plasty	Plasmid SPAdes	SMART2	TPA ID	Protein Coding	tRNA	rRNA
A. cinerea	NA	NA	17,047	16,863	**16,759**	BK010962	13	22	2
A. moreletii	*Bokermannohyla alvarengai*	NC_036493	12,236	18,107	15,781	BK010959	13	22	2
A. amphibius	*Dicrostonyx groenlandicus* *D. hudsonius*	NC_034313 NC_034307	291	Failed	**16,359**	BK010955	13	22	2
B. babyrussa	*Sus scrofa* *Sus celebensis*	MK251046 KJ789952	10,737	Nuclear	**16,645**	BK010954	13	22	2
B. ferruginus	NA	NA	6,846	16,135	9,806	BK010965	11	8	0
B. pombali	NA	NA	Failed	No match	9,847	BK010964	11	8	0
C. rufus	*Canis lupus* *Canis latrans*	KF857179 NC_008093	9,149	16,585	**16,474**	BK011186	13	22	2
C. bicolor	*Coendou insidiosus* *Sphiggurus insidiosus*	NC_021387 JX312693	16,630	16,771	**16,687**	BK010970	13	22	2
C. planiceps	NA	NA	16,438	16,682	**16,534**	BK010972	13	22	2
C. gundi	NA	NA	264	16,315	**16,101**	BK010971	13	22	2
C. paca	NA	NA	16,719	No match	**16,627**	BK010958	13	22	2
C. boraceiensis	NA	NA	16,933	No match	15,654	BK010967	13	22	2
G. surdaster	*Mus musculus* *Rattus norvegicus*	KY018919 KM577634	9,398	Nuclear	**16,308**	BK010969	13	22	2
H. oasicus	NA	NA	16,401	7,655	**16,401**	BK010960	13	22	2
H. niger kirkii	*Bos taurus* *Bubalus bubalis*	AF492351 MK234704	16,508	Timeout	**16,508**	BK011057	13	22	2
H. niger niger	*Bos taurus* *Bubalus bubalis*	AF492351 MK234704	Timeout	Failed	**16,506**	BK011056	13	22	2

Continued on next page

Table 4 Continued from previous page

Species	Related Species	GenBank Accession	NOVO Plasty	Plasmid SPAdes	SMART2	TPA ID	Protein Coding	tRNA	rRNA
H. arborea	*Hyla annectans* *Hyla chinensis*	NC_025309 NC_006403	16,757	15,958	15,751	BK010919	13	22	2
H. phyllodes	NA	NA	16,037	17,261	10,479	BK010968	11	13	0
M. xanthostomus	*Melanophryniscus moreirae* *Bufo melanostictus*	NC_037378 NC_005794	5,000	16,670	15,953	BK010963	13	22	2
O. pumilio	*Phyllobates terribilis* *Hyloxalus subpunctatus*	NC_037380 NC_037379	Failed	No match	15,856	BK010961	13	22	2
P. pipistrellus	*Lasiurus borealis* *Plecotus rafinesquii*	NC_016873 NC_016872	252	Failed	**16,458**	BK010957	13	22	2
P. hispida saimensis	*Phoca largha* *Halichoerus grypus*	FJ895151 NC_001602	258	16,667	**16,499**	BK011058	13	22	2
R. chenfui	*Polypedates braueri* *Polypedates megacephalus*	NC_042797 AY458598	14,999	Nuclear	14,441	BK010966	12	22+1dup	2
S. carolinensis	*Urocitellus richardsonii* *S. vulgaris*	NC_031209 NC_002369	7,025	16,610	**16,537**	BK010956	13	22	2
S. palustris	*Sorex daphaenodon* *Sorex minutissimus*	NC_044107 NC_042196	16,151	Nuclear	16,108	BK011027	13	22	2
U. parryii	*Urocitellus richardsonii* *Sciurus vulgaris*	NC_031209 NC_002369	16,462	16,539	**16,462**	BK011059	13	22	2