Index-Free De Novo Assembly and Deconvolution of Mixed Mitochondrial Genomes

Bennet J. McComish*†‡,[1,2], Simon F. K. Hills‡,[1,3], Patrick J. Biggs[1,4,5], and David Penny[1,2]

[1]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.

[2]Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand.

[3]Institute of Natural Resources, Massey University, Palmerston North, New Zealand.

[4]Massey Genome Service, Massey University, Palmerston North, New Zealand.

[5]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand.

†Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, PO Box 11-222, Palmerston North, New Zealand.

‡These authors contributed equally to this work.

*Corresponding author: Email: b.mccomish@massey.ac.nz

Accepted: 19 May 2010

Abstract

Second-generation sequencing technology has allowed a very large increase in sequencing throughput. In order to make use of this high throughput, we have developed a pipeline for sequencing and de novo assembly of multiple mitochondrial genomes without the costs of indexing. Simulation studies on a mixture of diverse animal mitochondrial genomes showed that mitochondrial genomes could be reassembled from a high coverage of short (35 nt) reads, such as those generated by a second-generation Illumina Genome Analyzer. We then assessed this experimentally with long-range polymerase chain reaction products from mitochondria of a human, a rat, a bird, a frog, an insect, and a mollusc. Comparison with reference genomes was used for deconvolution of the assembled contigs rather than for mapping of sequence reads. As proof of concept, we report the complete mollusc mitochondrial genome of an olive shell (*Amalda northlandica*). It has a very unusual putative control region, which contains a structure that would probably only be detectable by next-generation sequencing. The general approach has considerable potential, especially when combined with indexed sequencing of different groups of genomes.

Key words: multiplex sequencing, informatic deconvolution, control region, noncomplementary, molluscs.

Introduction

DNA sequence information is fundamental to our understanding of genome structure, function, and evolution. A major advance in sequencing methodology was introduced by the Sanger group in the 1970s, with the development of the chain-termination DNA sequencing reaction (Sanger et al. 1977). Sequencing has subsequently undergone increasing degrees of industrialization, with the introduction of fluorescent radiolabeled terminators and capillary electrophoresis, allowing the sequencing of entire genomes. In the last few years, however, so-called second-generation sequencing technologies have been developed using strategies such as pyrosequencing (Margulies et al. 2005) and sequencing by synthesis (Bentley 2006); strategies that are radically different from the Sanger dideoxy methodology.

Four commercial second-generation DNA sequencing platforms are now available: Roche's (454) Genome Sequencer FLX System, Illumina's Genome Analyzer (GA), Applied Biosystems' SOLiD System, and Helicos' HeliScope Single Molecule Sequencer. These all use a massively parallel approach, producing hundreds of thousands to tens of millions of sequence reads at a time; however, they are much shorter than Sanger dideoxy reads. Instead of creating a clone library (which could have ethics and/or genetic modification issues), the sample DNA is fragmented and the fragments are ligated to adapters, eliminating library construction and cloning host biases. At the time these experiments were carried out, a single run on the 454 system produced 400,000 reads of around 250 nt, a GA run produced over 40 million 36 nt reads, and a SOLiD run promised 86–114 million 35 nt reads. However, these output figures are all increasing rapidly as the technologies from each company are developed further. For example, a single GA run can currently produce 12–15 GB of sequence data (i.e., more than 10 million 75-bp paired-end reads per lane).

For robust phylogenetic reconstruction, it is highly advantageous to demonstrate concordance between independent data sets. In molecular data sets, this is often achieved by comparing results from nuclear data and mitochondrial and/or chloroplast data (e.g., Pratt et al. 2009). These data sets have often not been concordant due to the limited amount of sequence data being more indicative of aberrant histories of the gene involved rather than the evolutionary history of the genome (Nichols 2001). With the advent of second-generation sequencing, it has become increasingly possible to generate large quantities of data. Large multigene data sets are significantly less likely to be dominated by aberrant individual gene histories. It is therefore desirable to sequence both nuclear and organelle genomes. Due to issues such as nuclear copies of mitochondrial genomes, it is necessary to segregate organelle genomes from the nuclear sequence. However, the size of these genomes is such that much of the sequence will be wasted in many times more coverage than is needed.

If even a single lane of a GA flow cell is used to sequence something as small as a typical animal mitochondrial genome, there is a high degree of redundancy. For the 16.5-kb human mitochondrial genome, for example, raw coverage could be over 90,000×, and each read would be present in 300 copies. Current de novo sequence assembly algorithms perform well with much lower coverage. For example, Hernandez et al. (2008) successfully assembled a *Staphylococcus aureus* genome from 35-bp reads with a raw coverage of 48×.

A solution to this problem is sequencing a mixture of many organelle genomes; however, this leads to the difficulty of separating the individual genomes from the resulting short sequence reads. Clearly, a method is required to informatically allocate de novo contigs to a given genome, maybe via a pooling or an indexing strategy. There are many examples of pooling and indexing strategies in the literature, although none of them do exactly the same as the strategy we are proposing. Prior to next-generation sequencing, there were a variety of methodologies to look at pooling and/or indexing (see, e.g., Cai et al. 2001; Ng et al. 2006; Fullwood et al. 2009); however, these kinds of approach rely on finding segments in a genome for subsequent mapping and analyses but not for sequencing whole genomes. Illumina have developed and marketed their own indexing technology that allows up to 12 samples to be mixed in 1 lane of a GA flow cell. Using current protocols, each sample must be prepared individually, resulting in a linear cost increase for the number of samples under investigation. There is some cost reduction with the mixing of samples for running on the machine, but overall, this is still an expensive procedure. At the other end of the indexing continuum are new “hyperindexing” methods, such as DNA Sudoku (Erlich et al. 2009) and BARCRAWL (Frank 2009). However, again economies of scale mean that these approaches are useful for large numbers (thousands) of short sequences sometimes using multiple lanes and/or pooling, and so the sequencing of organellar genomes would not be appropriate with this approach either.

Our aim here is to test the hypothesis that for distantly related species (i.e., for highly divergent sequences), assembly should be straightforward and unambiguous. Where there is a high degree of similarity between two sequences, however, it becomes more difficult to assemble short reads unambiguously as there will be longer overlaps between reads from the different genomes. For these more similar genomes, we expect that indexing would be more appropriate, but we need to develop a method that could combine both approaches, index-free multiplexing and indexing. Ultimately, we would like to get the cost of a mitochondrial genome to under $100 but that is beyond the scope of our present work.

We first used combined simulated reads from a set of several animal mitochondrial genomes to explore the ability of sequence assembly algorithms to separate and assemble sequences from a mixture of reads from different sources. Once optimized, the same methods were successfully applied to reads from a single lane of a GA flow cell containing a mixture of 6 different mitochondrial genomes.

Mitochondrial sequences from 4 species were successfully assembled, thus establishing that it is possible to disambiguate and assemble a complete organellar genome from a mixture of sequence reads from more distantly related species. The complete mitochondrial genome of the neogastropod mollusc *Amalda northlandica* is reported in more detail, and we identify a novel putative regulatory element, most likely a reduced control region. This structural feature can, under certain assembly conditions, interfere with complete assembly of the genome, and this control feature is unlikely to be detected by classical sequencing techniques.

This approach is complementary to the indexing strategies mentioned above. Indexed sequencing will allow our approach to be used for several mixtures in a single run, with each mixture assigned a single index. This will enable us to sequence a large number of samples with a fraction of the sample preparation that would be required if we were to assign an index to each sample. The combination of index-free multiplexing and indexing should reduce costs considerably. In the application reported here, we use a disparate mixture of mitochondrial genomes (from humans to molluscs), but other combinations can certainly be used.

Methods

Simulations

Simulations were carried out using known animal mitochondrial genome sequences, which were downloaded and stored in a MySQL database. Custom Perl scripts (available from http://awcmee.massey.ac.nz/downloads.htm) were used to simulate 35-bp reads at random positions in the sequence and to introduce errors in these reads based on observed error profiles from previous GA sequencing experiments. Reads were then extracted from the database to simulate mixtures of different genomes in predefined ratios and written to files in FASTA format. A total of 4 million reads were extracted for each simulation, a conservative approximation to the number of usable reads produced on a single lane of a GA flow cell at the time of these experiments.

The simulated reads were assembled using Velvet version 0.7.26 (Zerbino and Birney 2008) and Edena version 2.1.1 (Hernandez et al. 2008), with a range of values for the hash length *k* (Velvet) or the minimum overlap between reads (Edena). The assembled contigs were aligned to the original genomes using the assembly tool of the Geneious package (v4.5.3; Drummond et al. 2008). Because the reference sequences were those used to generate the reads, stringent parameters were used for the alignment (minimum overlap 50 and overlap identity 98%). The contigs were also aligned to related reference sequences using less stringent parameters (minimum overlap 40 and overlap identity 60%) to test how closely related the reference needed to be to separate the contigs unambiguously.

The statistics package R (version 2.8.1; R Development Core Team 2009) was used to examine the distribution of coverages for each set of contigs. If the coverage distribution showed discrete peaks corresponding to the 5 different genomes, the contigs were grouped according to their coverage. Each group was then assembled into supercontigs using Geneious. No reference was used for the supercontig assembly—separating the contigs into groups corresponding to the different mitochondria should eliminate the ambiguous overlaps that broke up the initial assembly (except in the case of repeats), so that each group of contigs will assemble into a small number of supercontigs.

Another approach used to separate contigs from different genomes was to align the contigs to a set of reference sequences using the Exonerate sequence alignment package (v2.2.0; Slater and Birney 2005). Exonerate was set to report the five best alignments for each contig and to output a table showing, for each alignment, the names of the contig and the reference, the beginning and end of the aligned region in each, and the score and percent identity. As with the Geneious alignments, this was performed using the source genomes and using genomes with differing degrees of relatedness. The resulting table was used to group the contigs according to which reference produced the highest scoring alignment, and Geneious was used to assemble each group into supercontigs.

Sequencing

Long-range polymerase chain reaction (PCR) products were generated from a diverse set of templates in order to create a mixture of templates to sequence using an Illumina GA. The organisms used were a human, a rat (bush rat, *Rattus fuscipes*), a bird (tawny frogmouth, *Podargus strigoides*), a frog (Hamilton's frog, *Leiopelma hamiltoni*), an insect (ground weta, *Hemiandrus pallitarsis*), and a mollusc (Northland olive, *A. northlandica*). PCR products of between ~1 and 8 kb were generated using primers specific to, and thermal cycling conditions optimized for, each DNA template (available from the authors). PCR products were processed by SAP/EXO digestion to remove unincorporated oligonucleotides and then quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc.). Aliquots were taken in order to have an approximately even relative molarity for all DNA fragments in the final mix. All samples were then pooled and processed for sequencing in one lane using the genomic DNA sample preparation kit from Illumina (part #1003806).

A 50-bp single read run was performed on an Illumina GA GA2 (Illumina, Inc.) according to the manufacturer's instructions. Unfortunately, there was an instrument problem at cycle 33, which meant that only 32 nt were usable. Due to the availability of the raw material for sequencing, the run was continued to completion. After sequencing, the resultant images were analyzed with the proprietary Illumina pipeline (version 1.0) using default parameters. This resulted in ~238 Mb of sequence, with 63% of the clusters passing the initial filtering step.

Additional assessment of an anomalous section of the *A. northlandica* mitochondrial genome was performed by traditional Sanger sequencing of a 300-bp PCR product spanning a region between *nad*5 and *cox*3. This PCR product was generated from a total genomic DNA sample using specifically designed primers (Anor_nad5_f1618: 5′-ATGTCACAAGCAAACCAAAAGATCC-3′ and Anor_cox3_r100: 5′-TTACTGTAATATACCCATATCCGTG-3′) and using *Taq* DNA polymerase (Roche Applied Science) under the manufacturer's recommended conditions. The PCR product was processed by SAP/EXO digestion and sequenced on an ABI3730 automated sequencer (Applied Biosystems) in both the forward and reverse directions using the specifically designed PCR primers. The resulting sequences and electropherograms were visualized using Geneious.

De Novo Assembly

Due to high error rates observed for bases 1–5 and 33 onwards in the control lane of the Illumina flowcell, the reads were trimmed before assembly, removing the first 5 bases and the last 18 to leave 27-bp reads consisting of bases 6–32 of the original reads.

Perl scripts were used to run Velvet with a range of values for the hash length *k* and the coverage cutoff and to extract the number of nodes, maximum contig length, and N50 (median length–weighted contig length—half of all bases assembled are in contigs of this size or longer) values reported by Velvet.

Table 1

Mitochondrial Sequences Referred to in this Study

Accession Number	Species	Common name	Reference
J01415[a],[b]	*Homo sapiens*	Human	Anderson et al. (1981)
NC_001807[b]	*H. sapiens*	Human	Ingman et al. (2000)
AJ428514[a]	*Rattus norvegicus*	Norway rat	Nilsson et al. (2003)
NC_001665[b]	R. *norvegicus*	Norway rat	unpublished
EU273708[b]	*Rattus praetor*	Spiny rat	Robins et al. (2008)
NC_008551[a]	*Ardea novaehollandiae*	White-faced heron	Gibb et al. (2007)
DQ780883[b]	*Pelecanus conspicillatus*	Australian pelican	Gibb et al. (2007)
NC_008540[b]	*Apus apus*	Common swift	unpublished
AB043889[a]	*Rana nigromaculata*	Dark-spotted frog	Sumida et al. (2001)
NC_006688[b]	*Alytes obstetricians*	Common midwife toad	San Mauro et al. (2004)
AY660929[a]	*Gryllotalpa orientalis*	Oriental mole cricket	Kim et al. (2005)
EU938374[b]	*Troglophilus neglectus*	Cave cricket	Fenn et al. (2008)
NC_007894[a]	*Sepioteuthis lessoniana*	Reef squid	Akasaki et al. (2006)
AB029616[b]	*Loligo bleekeri*	Bleeker's squid	Tomita et al. (1998), Sasuga et al. (1999)
DQ238598[a],[b]	*Ilyanassa obsoleta*	Eastern mudsnail	Simison et al. (2006)
NC_008098[b]	*Lophiotoma cerithiformis*	Turrid snail	Bandyopadhyay et al. (2006)
NC_008797	*Conus textile*	Cloth-of-gold cone	Bandyopadhyay et al. (2008)
NC_010090	*Thais clavigera*	Rock shell	unpublished
NC_011193	*Rapana venosa*	Veined rapa whelk	unpublished
NC_013239	*Terebra dimidiata*	Dimidiate auger shell	Cunha et al. (2009)
NC_013241	*Cancellaria cancellata*	Cancelate nutmeg	Cunha et al. (2009)
NC_013242	*Fusiturris similis*		Cunha et al. (2009)
NC_013243	*Conus borgesi*		Cunha et al. (2009)
NC_013245	*Cymbium olla*	Pata-del-burro	Cunha et al. (2009)
NC_013248	*Nassarius reticulates*	Reticulate nassa	Cunha et al. (2009)
NC_013250	*Bolinus brandaris*	Purple dye murex	Cunha et al. (2009)
GU196685	*Amalda northlandica*	Northland olive	This study

[a] Genomes from which simulated reads were extracted.

[b] Genomes used as references.

Because of the large numbers of contigs produced, a Perl script (available from http://awcmee.massey.ac.nz/downloads.htm) was used to automate the procedure of aligning contigs against the reference sequences and separating them to produce a FASTA file of contigs aligning to each of the references, along with a file containing those contigs that fail to align to any of the references. The same script also converted the de Bruijn graph of contigs for each assembly produced by Velvet to DOT format, so that the graph could be visualized using GraphViz (Gansner and North 2000).

Identification of coding regions of the sequenced portions of the mitochondrial genomes was achieved through comparison to published complete mitochondrial genome sequences available through GenBank.

Results

Simulations

Thirty-five base-pair reads were extracted for a human mitochondrial genome (GenBank accession number J01415; see table 1 for a list of the mitochondrial sequences used in this study), the white-faced heron, the dark-spotted frog, the oriental mole cricket, and the eastern mudsnail. These organisms were chosen as they represent a mixture similar to that used in our experimental run. To test the effect of having the genomes present at different concentrations, the reads were extracted in a ratio of 10:15:20:25:30, in several permutations. These simulated reads were then assembled using Velvet version 0.7.26 (Zerbino and Birney 2008) and Edena version 2.1.1 (Hernandez et al. 2008). The largest possible overlap (the largest hash length in Velvet) gave the highest N50 in all cases.

The two sets of contigs produced by Velvet and Edena were aligned to each of the 5 genomes in turn. Each contig mapped perfectly to one of the five genomes, indicating that there were no misassemblies and that all sequencing errors were eliminated by the high coverage.

Coverage distributions for both sets of contigs for a single permutation are shown in figure 1A. All permutations that were tested gave similar results, with a single peak corresponding to each of the five genomes. This meant that, for simulated reads, the coverage values could easily be used to separate the contigs into five groups, one for each genome. In practice, however, it has been reported that for GA reads, coverage is not uniform but is correlated to GC content, perhaps due to AT-rich fragments being more prone to denaturation than GC-rich fragments (Dohm et al.

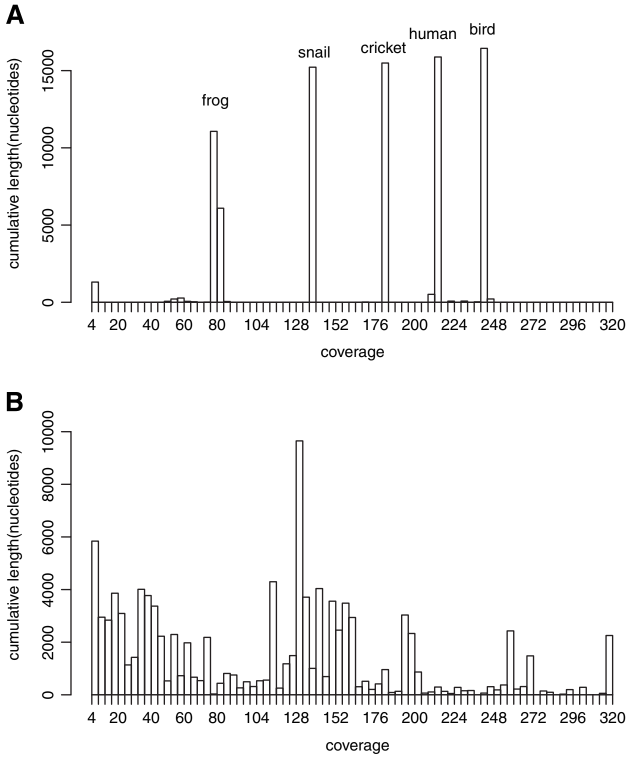


Fig. 1.—Coverage distributions. (*A*) Coverage, weighted by contig length, for contigs assembled from simulated reads by Velvet with $k = 31$. The sequences used in this simulation were human (25%), bird (30%), frog (10%), cricket (20%), and snail (15%). It is clear from these distributions that the contigs from each genome have tightly clustered coverage values, with the coverage for each genome directly proportional to the percentage of reads from that genome. (*B*) Coverage, weighted by contig length, for contigs assembled from biological data by Velvet with $k = 25$. Coverage for each genome is clearly not sufficiently uniform to be useful as a means of separating contigs. Coverages are given as *k*-mer coverage (see Zerbino and Birney 2008).

2008; Hillier et al. 2008). This highlights the need for caution when using simulations to test new methods.

To separate the contigs produced from simulated reads without using coverage information, Exonerate (Slater and Birney 2005) was used to align each contig against a set of reference sequences related to the mitochondrial genomes used to generate the reads. For the human, rat, bird, snail, and squid mitochondria listed above, the references used were another human mitochondrial genome (accession number NC_001807), another Norway rat (accession number NC_001665), the Australian pelican, the turrid snail, and Bleeker's squid, respectively. Because the relatedness between the reference and the original sequence was different for each genome (the same species for the human and rat and different orders for the bird) and the degree of sequence conservation varies across the genome, only the relative values of the alignment scores for each contig could be used to separate the mixture of contigs into its component genomes. For each contig of our Edena assembly, the best alignment identified corresponded to the correct reference, except for two short contigs from the control region of the bird, which failed to align to any of the reference sequences.

Once separated, each group of contigs was assembled to give one or more supercontigs for each genome. For the Velvet assembly of the permutation described above, the cricket and snail mitochondrial genomes each gave two contigs, which overlapped to form a single supercontig covering the whole of each genome. The human mitochondrial genome gave seven contigs, which formed a single supercontig covering the whole genome, although one of the overlaps was very short (seven bases). The bird and frog mitochondrial genomes contain tandem repeat regions, which could not be assembled from short reads. This would be the case regardless of whether they were sequenced separately or as part of a mixture (see Chaisson et al. 2004 and Kingsford et al. 2010 for analysis of the limitations of short reads for repeat resolution). However, the remainder of each genome was successfully assembled into two supercontigs. We obtained similar results for the other permutations we examined and for the Edena assemblies—there were small differences in the numbers of contigs produced, but these did not affect the assembly into supercontigs.

To test whether more closely related mitochondrial genomes could be separated in the same way, the exercise was repeated using the same human, bird, and snail mitochondrial genomes, together with a Norway rat (accession number AJ428514) and reef squid. The relatively closely related human and rat mitochondrial genomes were each broken up into a larger number of contigs (12 each), but these could still be separated by their different coverage levels, and each set then assembled into a single supercontig. Two short contigs (length 52 and 54 bp) had coverage equal to the sum of the expected coverages for the human and rat genomes and aligned with 100% sequence identity to both the human and rat references. These represent regions of the 16S ribosomal RNA gene that are conserved between the two species and were included in both sets of contigs.

The squid mitochondrial genome contained a duplicated region, which gave a 505-bp contig with twice the expected coverage. The double coverage made it possible to identify the contig as a repeat and to include it twice when assembling the contigs into a supercontig.

Our simulations thus confirmed that it is possible to assemble short reads from this type of mixture of mitochondrial genomes and to separate the assembled contigs into the individual components of the mixture.

Biological Data

For some of the organisms chosen, closely related reference genomes were available. We also chose some more difficult examples, for which the closest available reference was

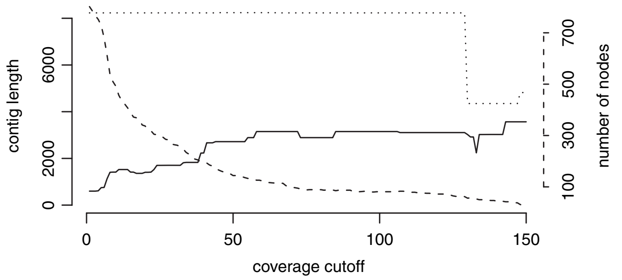


Fig. 2.—Assembly statistics for biological data. Median length–weighted contig length (N50, solid line), maximum contig length (dotted line), and number of nodes (dashed line) plotted against coverage cutoff for Velvet assemblies with hash length $k = 25$. Contig lengths are in k-mers (length in base pairs can be obtained by adding $k - 1$). Increasing the coverage cutoff eliminates low-coverage nodes, removing some branching in the graph and allowing some of the higher coverage nodes to merge. The distinct steps in the N50 plot may reflect different coverages for the different DNA fragments sequenced. The longest contig is stable, with a length of 8,231 for all coverage cutoffs up to 129, except that for coverage cutoffs between 45 and 64, 10 nt are added to one end of the contig to give a length of 8,241.

much more distant, for example, in a different taxonomic order in the case of the bird.

Trimmed 27-bp GA reads were assembled using Velvet. As expected, the best results were obtained with the longest possible hash length (25, giving 536 contigs with an N50, or median length–weighted contig length, of 598). The coverage cutoff parameter of Velvet was used to eliminate short low-coverage nodes (which are likely to be errors), giving considerably higher N50 values. It is likely that the six samples were present at different concentrations, so we expected that different values of the coverage cutoff would be optimal for each genome. The number of nodes, maximum contig length, and N50 values reported by Velvet with coverage cutoff values up to 150 are shown in figure 2. Assemblies with coverage cutoff set to 12, 26, 35, 45, and 58 were examined.

Probably because of the differences in GC content within a genome, coverage was not sufficiently uniform to separate the contigs belonging to the different genomes (see fig. 1*B*). Consequently, they were separated by aligning them to a set of reference genomes. The references used were the mitochondrial genomes of a human (accession number J01415), the spiny rat, the common swift, the common midwife toad, the cave cricket, and the eastern mudsnail. The degree of relatedness between the target and reference sequences was thus different in each case: for human, target and reference were two members of the same species; for the rat, different species of the same genus; and for the bird, frog, cricket, and mollusc, target and reference were in different families or even higher order taxa.

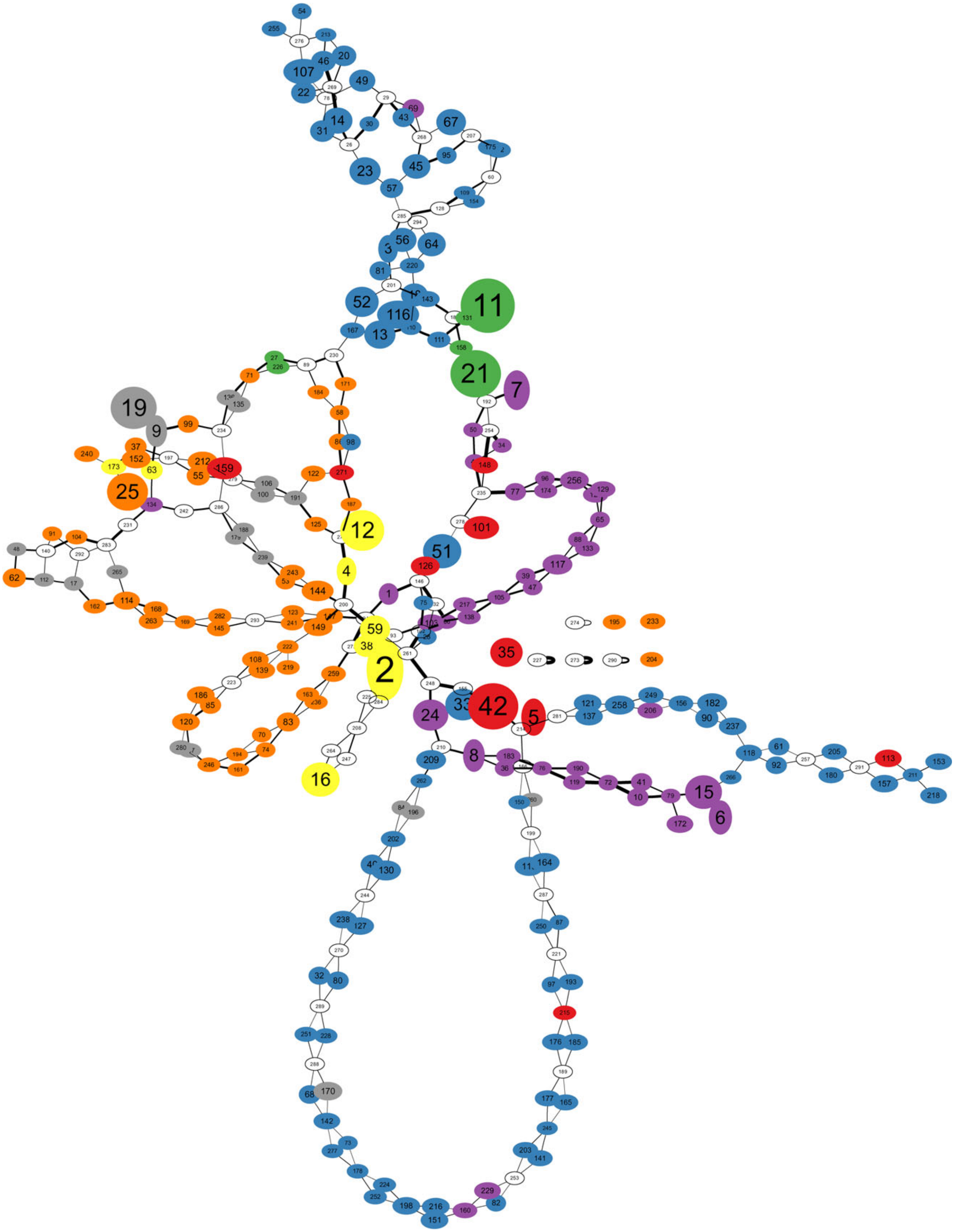
Of a total of 964 contigs for the 5 assemblies examined, 762 were correctly grouped into species in this first step. However, because the single best alignment for each contig was used regardless of the relative scores of alignments to the other references, 64 contigs were initially assigned to incorrect species. Where sequences are highly conserved (or highly divergent), contigs may align with similarly high (or low) scores to several references, thus it was expected that not all contigs would be assigned correctly by this method. A further 113 contigs failed to produce any alignments with scores above Exonerate's default threshold. However, all contigs belonging to the human and rat sequences were assigned correctly, presumably as a consequence of having closely related reference sequences for these organisms.

A second round of separation was carried out using the assembly graphs produced by Velvet. An example of an assembly graph, with each node colored according to the reference to which it aligned, is shown in figure 3. We used the graph to identify contigs that appeared to have been assigned to the wrong genome and to ascertain the origin of those contigs that failed to align using Exonerate. These were checked against the GenBank (Benson et al. 2009) nucleotide database using the web-based BlastN algorithm (Altschul et al. 1997). BlastN found closer alignments for most of these contigs than those to our reference genomes, as we expected, because GenBank contains many shorter sequences in addition to the relatively small number of whole mitochondrial genomes known. Such comparisons are therefore very useful in aiding assembly.

Any contigs that were connected in the assembly graph to contigs that aligned to different references were checked against GenBank. Node 206 of the assembly in figure 3, for example, aligned to the spiny rat, whereas the 2 neighboring contigs aligned to the common swift, but when checked against GenBank, the best alignment found for node 206 was to *Gallirallus okinawae* (Okinawa rail) mitochondrial DNA, so it was reassigned to the pool of bird contigs. Another example is node 75, which was found to match fragments of mitochondrial 16S sequence from the frogs *Leiopelma archeyi* and *Leiopelma hochstetteri* in the GenBank database with 100% identity, despite aligning more closely to our bird reference than to our frog reference. No useable alignments were found for 17 unmatched nodes, all of which grouped in the graphs with contigs that aligned to the insect reference, and these were assigned to the insect pool on the basis of their position in the graph. This general problem will certainly decrease as more complete genomes become available, but it still requires care at present.

Once separated, the contigs aligning to each reference were imported into Geneious (Drummond et al. 2008). Each set of contigs was assembled into supercontigs, and these supercontigs, along with any contigs not included in the supercontigs, were aligned against the reference.

The human sequence was a single long-range PCR product from a human Melanesian sample, the remainder of this mitochondrial genome having been sequenced in a previous


54
255
276
213
46
20
107
269
49
22
78
29
69
43
67
14
30
31
26
268
207
175
95
45
23
60
57
109
154
128
285
294
3
56
64
220
81
201
143
11
52
116
131
13
167
111
158
27
226
89
230
21
7
71
171
184
19
135
192
58
99
9
234
50
254
98
34
37
240
152
197
212
159
148
173
63
55
122
271
279
106
96
256
25
100
191
174
77
129
235
134
242
286
187
12
65
125
278
101
231
188
91
104
283
88
133
48
140
292
239
51
117
4
243
126
62
112
17
265
144
146
39
47
1
114
105
162
168
200
75
217
138
263
282
123
169
293
241
149
103
274
195
233
145
59
93
222
38
261
35
227
273
290
204
108
219
2
139
259
223
248
186
163
85
236
24
33
42
5
121
258
249
156
182
137
206
90
120
284
208
281
237
83
210
70
61
280
264
194
74
209
8
183
76
190
118
92
257
205
291
113
153
246
161
247
36
119
72
41
266
180
211
157
218
16
262
10
79
15
150
84
196
6
172
199
202
130
164
244
287
238
127
250
87
270
221
32
80
97
193
289
215
251
228
176
185
288
189
68
170
177
165
142
245
73
277
203
141
178
253
224
229
252
198
216
82
151
160

experiment. The best results were obtained with a coverage cutoff of 12, giving 3 overlapping contigs with a total length of 10,485 nt spanning from *cox*1 to 12S rRNA as expected. Higher coverage cutoff values still gave the same three overlapping contigs, except that the longest contig (and hence the overall length) was slightly shorter. This human Q2 haplotype will be reported separately and has the GenBank accession number GQ214521.

The best assembly for the mollusc sequence was obtained with the higher coverage cutoff values. Coverage cutoffs of 45 and 58 produced seven contigs that overlapped to form a single supercontig 15,361 bp in length, whose ends overlap by 7 bp. Although this overlap is short, it is within the *trn*H gene and is part of a short (18 bp) overlap between two long-range PCR products. All other overlaps between contigs were 19 bp or longer. The supercontig appears therefore to constitute the entire mitochondrial genome of *A. northlandica*. Lower coverage cutoffs gave six contigs, covering the whole genome except for a gap of 11 nt in the noncoding region between *trn*F and *cox*3. We discuss this genome in more detail below.

All coverage cutoff values gave similar results for the frog, with five contigs forming three supercontigs of 616, 1,385, and 6,361 bp at coverage cutoffs of 45 and 58. This represents almost all of the frog template loaded (long-range PCR was only able to generate one fragment representing approximately half of the frog mitochondrial genome). At the lower coverage cutoff values, six contigs were produced, but they still formed the same three supercontigs, although the longest was slightly shorter, at 6,350 bp.

The rat assembly was also largely unaffected by the coverage cutoff. However, there were two regions where polymorphisms were observed. These can be seen as crisscross patterns in the graph in figure 3—where the two sequences have diverged, they form a pair of parallel contigs both of which overlap with contigs on either side where the sequences are identical. These regions are in the 12S and 16S rRNA genes and in *cox*1. The two sequences observed in each of these regions were highly similar, and open reading frames were preserved. These might indicate the presence of nuclear DNA sequences of mitochondrial origin (numts; see Lopez et al. 1994; Richly and Leister 2004).

The contigs where the sequence was unambiguous were used in conjunction with further sequencing experiments to determine the complete mitochondrial genome sequence of *R. fuscipes*, extending the work of Robins et al. (2008). This sequence has the GenBank accession number GU570664 and will be published separately, along with the mitochondrial genome sequences of several other *Rattus* species.

The bird sequences show a more complicated pattern again, as can be seen in figure 3. It appears that, as well as containing the intended tawny frogmouth DNA, the sequencing reaction was contaminated with DNA from a common moorhen, a sandhill crane, and a red-fronted coot. Unfortunately, no reference sequences are available at present that can be used to distinguish these birds across the whole mitochondrial genome, and the problem appears to have arisen through tissue contamination (see later).

Aligning the contigs to the common swift reference genome showed a single sequence stretching from the middle of the 12S rRNA gene to *trn*M, with a small gap in 12S rRNA. From *nad*2 to *atp*6, there were two parallel sequences, and from the end of *cox*3 to the middle of *cyt*B, there were three. Comparing contigs to the GenBank nucleotide database using BlastN showed that the sequence covering 12S rRNA to *trn*M matched tawny frogmouth sequence fragments: 1 partial 12S rRNA sequence and 1 sequence covering *trn*L, *nad*1, *trn*I, and *trn*Q. Of the two parallel sequences from *nad*2 to *atp*6, one gave an exact match to existing partial *cox*1 and *atp*8 sequences for the Southern American common moorhen *Gallinula chloropus galeata* and the other gave an exact match to an existing partial *cox*1 sequence for the red-fronted coot *Fulica rufifrons*. At the *cyt*B locus, where there were three parallel sequences, one was found to match tawny frogmouth, the second matched common moorhen, and the third matched the sandhill crane *Grus canadensis*. One of the 3 sequences also matched an existing tawny frogmouth fragment covering part of *nad*1 and *trn*H, *trn*S, and *trn*L and another matched an existing sandhill crane fragment covering part of *cox*3, *nad*3, and *trn*G.

Many of the bird contigs had relatively low coverage values (because the presence of contaminants meant that the overall sequence length was much longer than expected), so that when assembly was carried out with a higher coverage cutoff, they were eliminated or two parallel contigs were merged to form a single contig.

DNA was extracted from a sandhill crane sample in our laboratory alongside the tawny frogmouth sample. However, neither common moorhen nor red-fronted coot have ever been studied in this laboratory (nor are the species present in this country), so it is likely that either the tawny frogmouth or the sandhill crane tissue sample was contaminated with DNA from these two species before our laboratory

⟵
Fig. 3.—Assembly graph. The assembly graph for sequences assembled by Velvet with $k = 25$ and cov_cutoff = 26. Nodes are colored according to the reference sequence to which the corresponding contigs align: green for human, purple for rat, blue for bird, red for frog, yellow for mollusc, and orange for insect. Gray nodes failed to align to any of the references, and white nodes are shorter than $2k - 1$ (Velvet does not output contigs for these nodes). The area of each node is proportional to the length of the sequence it represents, and the width of an edge between nodes is proportional to the number of reads that connect those nodes. The human, mollusc, and frog sequences are assembled into relatively small clusters of long contigs, whereas the insect, bird, and rat show more complex chains of shorter contigs. The reasons for these patterns are discussed in the text.

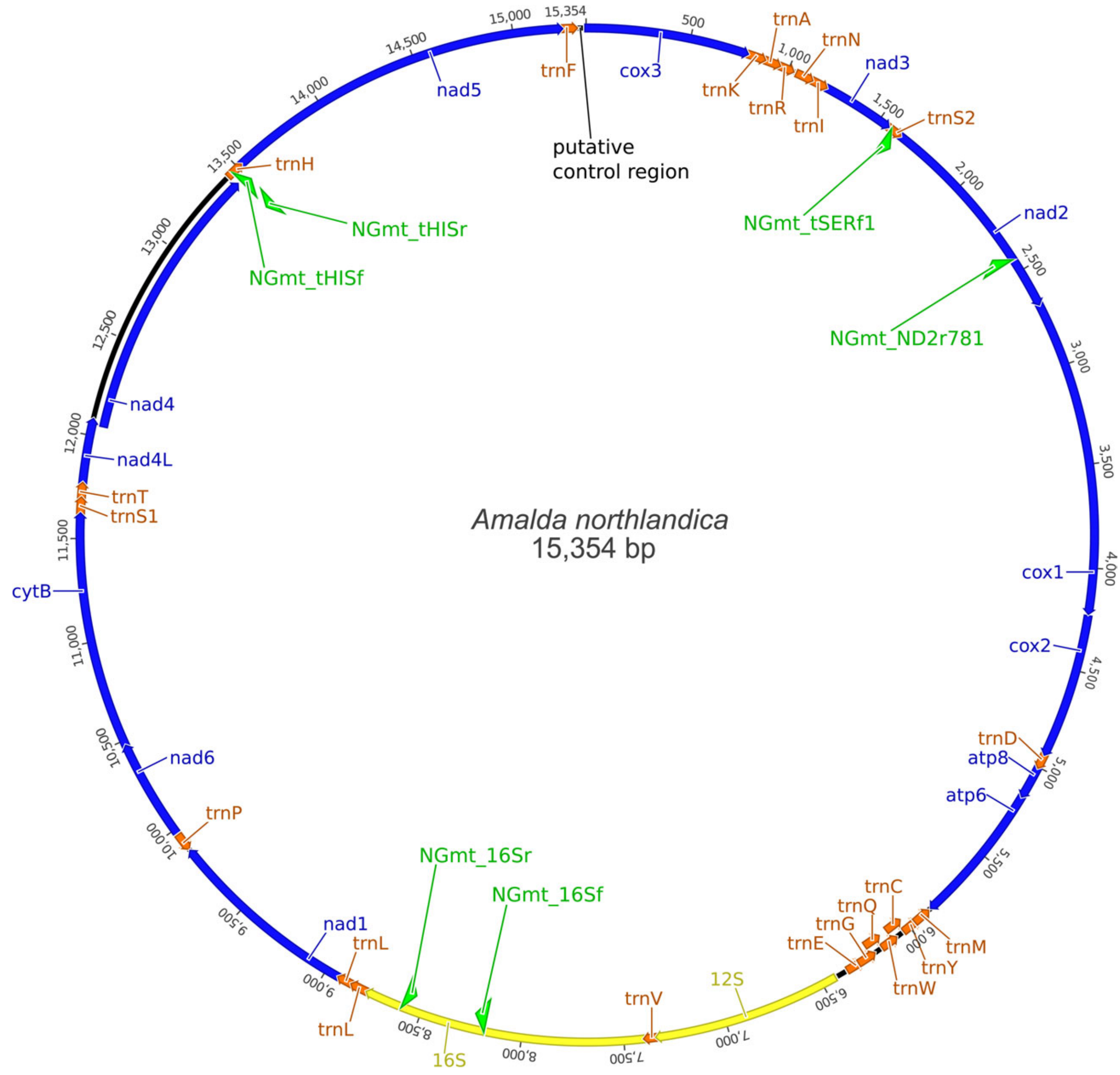


FIG. 4.—The *A. northlandica* complete mitochondrial genome. Arrowheads depict the direction of transcription. Genes with offset annotations (*trn*C, *trn*Q, and *nad*4) overlap with genes preceding them. Binding sites for the primers used to generate the long-range PCR products are indicated in green.

received them. Using the same scalpel for dissecting different birds is a possible explanation. This highlights the need for good laboratory practice—the high dynamic range of these DNA sequencing techniques means that minute traces of DNA will be amplified and sequenced.

As with the bird, the insect sequences show a rather convoluted assembly, with regions where two or three sequences align in parallel to the same region of the reference. The insect sequences, however, appear to be nuclear DNA sequences of mitochondrial origin as we were unable to identify open reading frames corresponding to the genes to which the sequences align. A possible solution to this problem would be the isolation of whole mitochondria, followed by DNA extraction from these mitochondria. This would exclude nuclear DNA, thereby eliminating the contribution of any nuclear copies of mitochondrial genes to the resulting sequence reads.

Amalda northlandica Mitochondrial Genome

The mitochondrial genome of *A. northlandica* is 15,354 bp in length and contains 13 protein-coding genes, 2 ribosomal RNA genes, and 22 tRNA genes (figure 4) and has

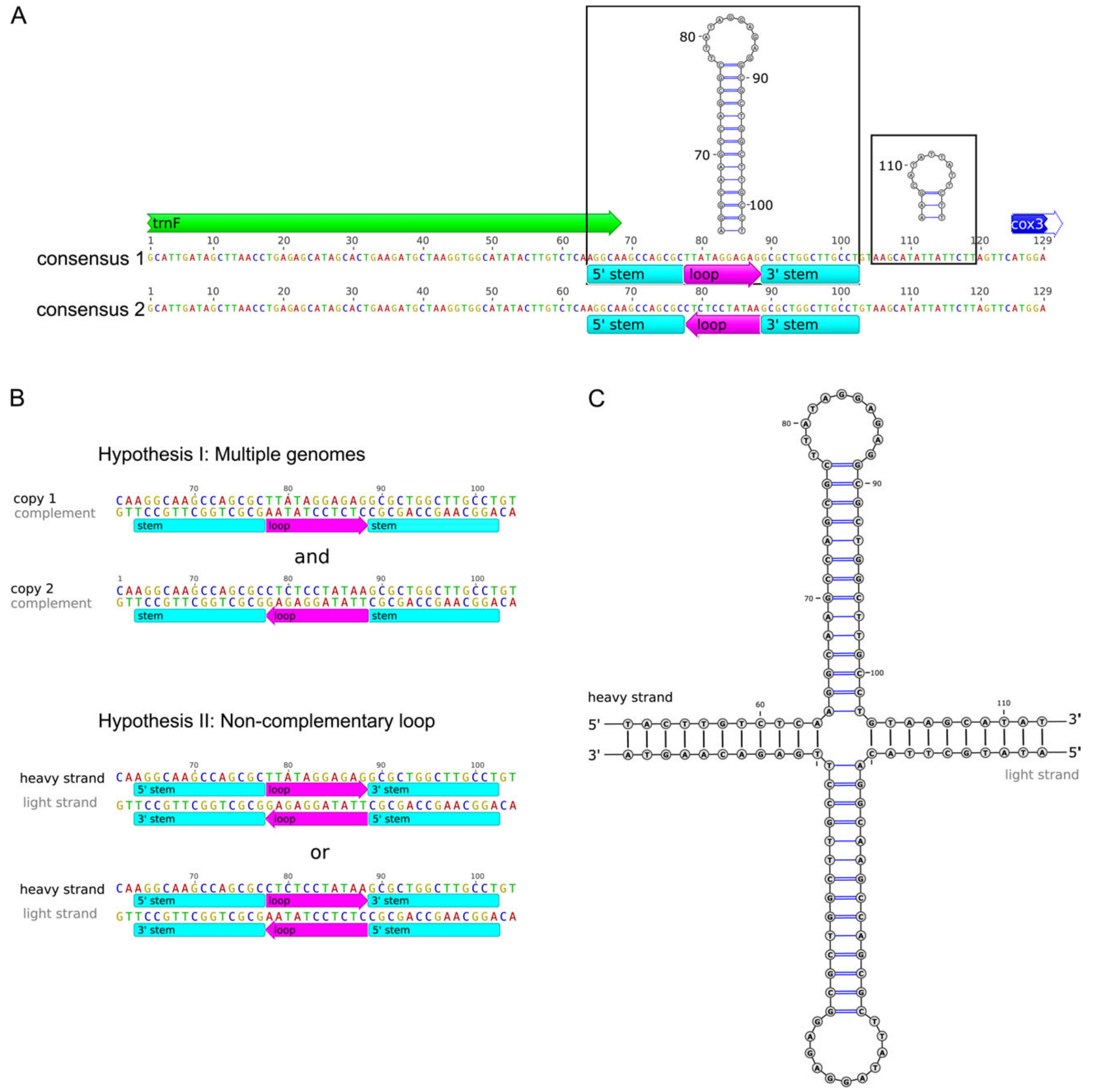
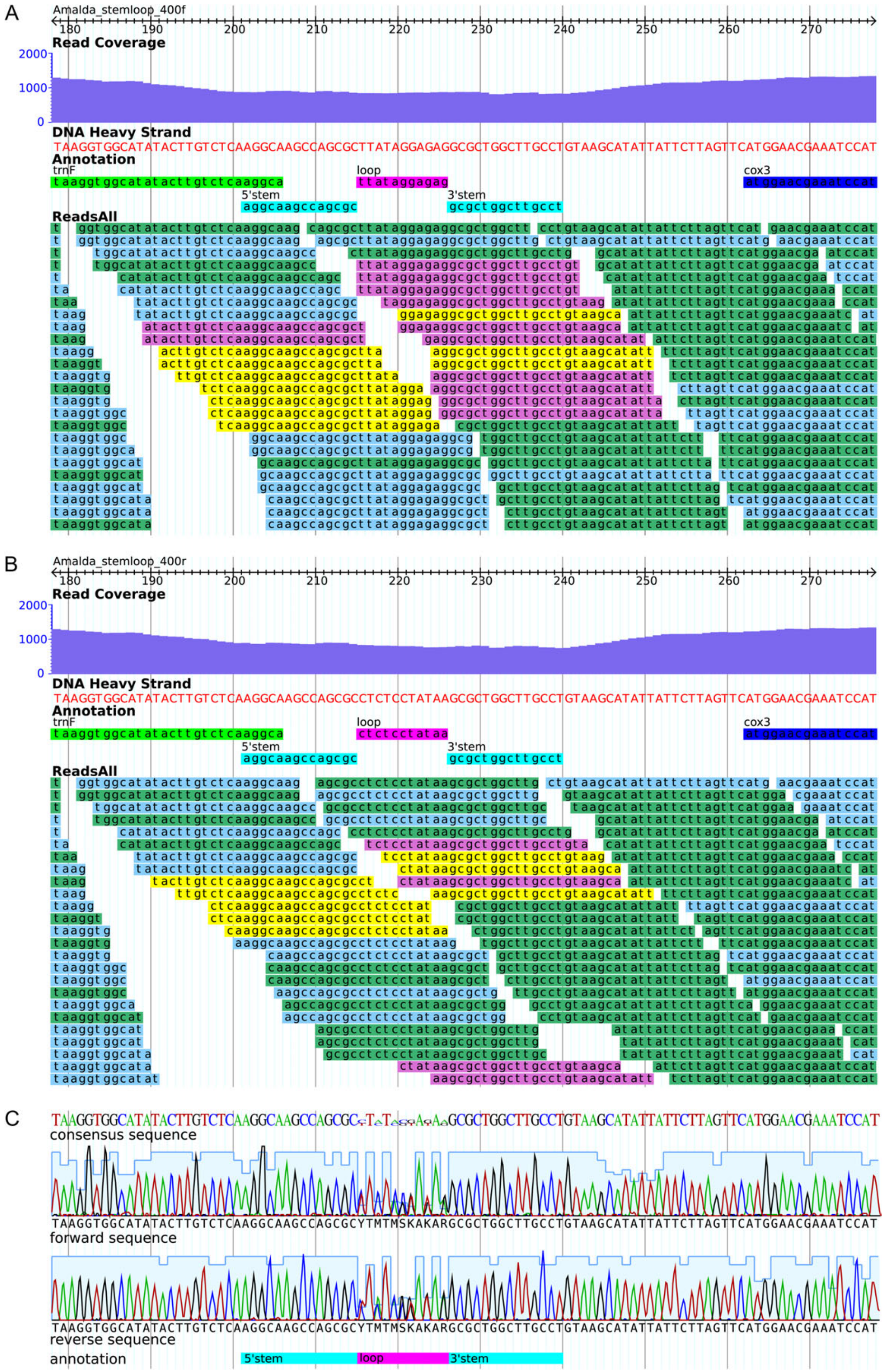

A
80
90
70
100
110
trnF
cox3
1 10 20 30 40 50 60 70 80 90 100 110 120 129
consensus 1 GCATTGATAGCTTAACCTGAGAGCATAGCACTGAAGATGCTAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGA
5' stem
loop
3' stem
consensus 2 GCATTGATAGCTTAACCTGAGAGCATAGCACTGAAGATGCTAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGA
5' stem
loop
3' stem
B
Hypothesis I: Multiple genomes
copy 1
complement
70 80 90 100
CAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGT
GTTCCGTTCGGTCGCGAATATCCTCTCCGCGACCGAACGGACA
stem
loop
stem
and
copy 2
complement
1 70 80 90 100
CAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGT
GTTCCGTTCGGTCGCGGAGAGGATATTCGCGACCGAACGGACA
stem
loop
stem
Hypothesis II: Non-complementary loop
heavy strand
70 80 90 100
CAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGT
5' stem
loop
3' stem
light strand
GTTCCGTTCGGTCGCGGAGAGGATATTCGCGACCGAACGGACA
3' stem
loop
5' stem
or
heavy strand
70 80 90 100
CAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGT
5' stem
loop
3' stem
light strand
GTTCCGTTCGGTCGCGAATATCCTCTCCGCGACCGAACGGACA
3' stem
loop
5' stem
C
80
90
70
100
heavy strand
60
110
5'
3'
3'
5'
light strand

Fig. 5.—The *trn*F–*cox*3 intergenic region of the *A. northlandica* mitochondrial genome. (*A*) The position and inferred structure of stem–loop elements in this region; the positions of the *trn*F gene and the initial bases of the *cox*3 gene are also indicated. The smaller predicted stem–loop reduces the overall stability of both structures. (*B*) Two hypotheses could explain the sequence data: hypothesis I: there is a mixture of 2 mitochondrial genome copies that differ in the orientation of the loop sequence; or hypothesis II: there is a single genome that contains a noncomplementary region, which could exist in either of 2 possible orientations. (*C*) Hypothesis II suggests the formation of a double-stem structure in double-stranded DNA.

the GenBank accession number GU196685. All protein-coding genes begin with the standard ATG start codon with the exception of *nad*6, which starts with an ATA codon. All the protein-coding genes terminate with standard TAA or TAG codon. The gene composition and order is consistent with neogastropod complete mitochondrial sequences currently available in GenBank (*Ilyanassa obsoleta*, *Lophiotoma cerithiformis*, *Conus textile*, *Thais clavigera*, *Rapana venosa*, *Terebra dimidiata*, *Cancellaria cancellata*, *Fusiturris similis*, *Conus borgesi*, *Cymbium olla*, *Nassarius reticulatus*, and *Bolinus brandaris*). In addition, a novel structural element (outlined below) was identified during assembly of the *A. northlandica* mitochondrial genome sequence. This structure may represent a reduced mitochondrial control region (which has not yet been identified in neogastropod molluscs).


A
Amalda_stemloop_400f
180
190
200
210
220
230
240
250
260
270
Read Coverage
2000
1000
0
DNA Heavy Strand
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCTTATAGGAGAGGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
Annotation
trnF
taaggtggcatatacttgtctcaaggca
loop
ttataggagag
cox3
atggaacgaaatccat
5'stem
aggcaagccagcgc
3'stem
gcgctggcttgcct
ReadsAll
B
Amalda_stemloop_400r
180
190
200
210
220
230
240
250
260
270
Read Coverage
2000
1000
0
DNA Heavy Strand
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCCTCTCCTATAAGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
Annotation
trnF
taaggtggcatatacttgtctcaaggca
loop
ctctcctataa
cox3
atggaacgaaatccat
5'stem
aggcaagccagcgc
3'stem
gcgctggcttgcct
ReadsAll
C
consensus sequence
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCYTMTMSKAKARGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
forward sequence
TAAGGTGGCATATACTTGTCTCAAGGCAAGCCAGCGCYTMTMSKAKARGCGCTGGCTTGCCTGTAAGCATATTATTCTTAGTTCATGGAACGAAATCCAT
reverse sequence
annotation
5'stem
loop
3'stem

An Unusual Control Region?

A very unusual feature of the assembly was that under certain coverage cutoff regimes, a fragment of the mitochondrial sequence was omitted. This 11-bp fragment was found to be in an intergenic region between *trn*F and *cox*3, and it is surprising that such an apparently short region should disrupt assembly. In order to identify possible causes of incomplete assembly, the noncoding intergenic spacers were analyzed for secondary structure formation. This could also elucidate structural features, such as the origin of replication and control region, which have not yet been identified in neogastropod molluscs. The highly variable 3′ and 5′ domains of the rRNA genes mean that the precise boundaries of 12S rRNA and 16S rRNA are not yet known. Due to this uncertainty, the regions flanking the rRNA genes were not considered.

The longest intergenic spacer in *A. northlandica* is located between the genes *trn*F and *cox*3. It is 56 bp in length and contains 2 predicted secondary structural elements, a strong stem–loop element and a second small stem–loop element (fig. 5A). Of the remaining intergenic sequences in the *Amalda* mitochondrial genome, only 7 are longer than 10 bp. All of these 7 exhibit some secondary structure (as predicted by the program MFold; Zuker et al. 1999). Including sequence of *trn*F showed that the initial stem in the intergenic spacer overlaps with the 3′ end of the acceptor stem of the tRNA by 5 bp. This initial stem of 14 bases is by far the strongest secondary structure in the intergenic regions (−20.03 kcal/mol). The presence of the short second possible stem–loop reduces the stability of the combined structure to −19.85 kcal/mol.

The incomplete assembly observed for lower coverage cutoff regimes (see earlier) was identified to be the result of a loss of 11 nt representing the complete loop of the structure shown in figure 5A. This appeared to suggest that palindromic sequence of sufficient size may cause the loss of sequence during assembly under specific cutoff regimes. However, further analysis of the sequence coverage of this region revealed that identical (not complementary) but reversed sequence existed in both the forward and reverse directions of the loop region of this structure (fig. 6). Although we are able to confirm the sequence of the loop region, we are unable to show in which of 2 possible orientations this sequence exists naturally in the *Amalda* mitochondrial genome. Re-examination of this region with Sanger sequencing confirmed the presence of ambiguous base calls within the expected 11-bp section. The Sanger sequence also confirms that this anomalous region is not the result of an artifact introduced in the Illumina sequencing or short-read assembly (fig. 6C).

It is not yet possible to confirm whether this structure represents either the control region or an origin of replication. There are no clear homologies with known structures or known conserved sequence blocks associated with either structure. However, this region can be identified in 6 published neogastropods (*I. obsoleta*, *T. clavigera*, *R. venosa*, *F. similis*, *B. brandaris*, and *N. reticulatus*), where the size is nearly identical (56–58 bp). The predicted secondary structures are very similar (data not shown) with well-conserved sequences for the stem structure (see fig. 5), but the nucleotide sequences for the remainder of the region are quite divergent in these species. The mitochondrial genomes of *C. cancellata*, *C. olla*, *L. cerithiformis*, *C. textile*, and *C. borgesi* are all longer; have more complex predicted secondary structures; and, with the exception of *Cancellaria*, have no significant sequence homology to the previously mentioned neogastropods. The remaining published neogastropod (*T. dimidiata*) has a considerably larger intergenic region in this position that exhibits no clear homology with the other known neogastropod mitochondrial genome sequences.

In addition, the positions of other structure-bearing intergenic regions are not conserved across the known neogastropod mitochondrial genomes. For example, an intergenic region of 25 bp is observed in *Amalda* between *nad*1 and *trn*P, whereas most of the known neogastropod sequences have some intergenic sequence at this position only 5 have a region that is larger than 10 bp. Furthermore, there is no unambiguously homologous sequence in these variable intergenic regions. It remains uncertain whether homologous structures exist at different positions in the other mitochondrial genomes.

Discussion

These results show that, given an appropriate reference sequence for each genome under consideration, it is possible

←

Fig. 6.—Gbrowse visualizations of short reads from the *A. northlandica* mitochondrial control region showing reads present in either orientation and electropherograms confirming the sequence. Parts (*A*) and (*B*) show a representative sample of 27-bp sequence reads across each orientation. The loop sequence between the stems is shown in magenta in the “Annotation” track. Short reads are shown in the forward and reverse strands (blue and green, respectively). The reads that give directionality to the loop sequences (i.e., that cross the boundary of either the 5′ stem or 3′ stem into identifiable sequence) are shown in the forward (yellow) and reverse (pink) strands. Part (*C*) shows Sanger sequence confirmation of ambiguous nucleotide sites at the positions predicted by the short-read mapping in (*A*) and (*B*) above. Electropherograms show the base calls for the nucleotide sequence reads in both the forward and reverse directions. Sequence quality scores are indicated for each site as a histogram in parallel with the electropherograms. Scores range from 55 for high-quality base calls to 12 for the lowest quality call of the ambiguous nucleotide positions. The sequence shown includes only the 100 bases that align with the short-read assemblies shown in (*A*) and (*B*) and comes from a sequence fragment of length 300 bp.

to assemble short reads from a mixture of mitochondrial genomes and deconvolute the resulting contigs without the need to index the reads. The reference sequence for each genome must be considerably closer to that genome than to any of the others, but it is not necessary for the references to separate the sequences perfectly as the assembly graph can be used to identify spurious alignments, as well as to reallocate contigs that fail to align to any of the references.

In principle, the same approach could be applied to other mixtures of sequences, for example, chloroplast genomes. We have successfully assembled chloroplast genomes from short-read data (data not shown), although not yet from a mixture.

The main difficulties encountered in assembling the genomes in this study were not due to problems in separating the contigs but due to problems with sample preparation, namely the presence of numts and contamination. These same issues would have arisen if the six genomes had been sequenced separately. It is clear that it is important to have high-quality DNA samples for de novo assembly. Any contamination can lead to ambiguities which make it difficult to distinguish between the sample and the contamination. This issue is significantly compounded if the contamination is closely related to the target sequence, relative to the reference sequence used (e.g., 2 birds), with varying degrees of sequence incompleteness or incorrect contigs generated depending on the level of relatedness. However, contamination will normally only affect assembly of the most closely related sequence, leaving the other samples unaffected. In the absence of contamination and numts, we would expect fewer contigs to be produced, making the process of deconvolution considerably simpler.

In generating the complete mitochondrial genome sequence of the mollusc *A. northlandica*, we have characterized a novel structural element in a mitochondrial genome. The identification of apparently identical DNA sequence in both the heavy and light strands of this structure leads to two possible explanations (see fig. 5B):

1. that separate mitochondrial genome molecules exist in an individual, differing only in alternative orientations of the sequence of this loop or
2. that the sequence on both strands of the DNA molecule is identical in this loop and therefore noncomplementary in double-stranded DNA.

It is difficult to envisage a functional explanation for the first hypothesis. However, extrapolating from the second hypothesis, it could be suggested that this noncomplementary sequence enforces the formation of a functionally important structural element in double-stranded DNA (fig. 5C). One difficulty with this hypothesis is how such a noncomplementary region would be replicated. RNA mediation is a possible solution and could be involved in an initiation process. Furthermore, given that the identical loop sequences are in opposite directions on each DNA strand, this might impart a directionality to each strand (e.g., for replication). Similar stem structures have been proposed for bidirectional transcriptional promoters in vertebrate mitochondrial genomes (L'Abbé et al. 1991; Ray and Densmore 2002), but the suggestion of noncomplementary DNA in the double-stranded mitochondrial genome is, as far as we are aware, unprecedented. Such an arrangement could be a result of the contraction of the mitochondrial genome in neogastropod molluscs, and the structure we have identified may represent a highly reduced control region. It is extremely unlikely that traditional Sanger sequencing is capable of characterizing this novel sequence feature, although it might be detectable as a region of poor-quality sequence. Indeed, several reported neogastropod mitochondrial genomes share sequence and structural homology with the stem structure shown here for *Amalda*, but there is very little sequence homology seen for the loop. Furthermore, the sequence of the mitochondrial genome of *I. obsoleta* is reported with ambiguous bases in the region homologous to the *Amalda* loop, alluding to the presence of ambiguous sequence that we predict would be observed in Sanger sequence of this region. It is probable that the case reported here is not limited to *Amalda*. A detailed characterization of the structure and evolutionary significance of the genomic region that we have identified here will be reported elsewhere.

The unusual arrangement of sequence in this structure was detectable in short-read sequencing as it led to an apparently structure-mediated loss of sequence during contig generation. The extent to which this prevails is unknown as such an arrangement has never been described. However, clearly, the development of new DNA sequencing technologies might allow the discovery of features that were intractable with earlier techniques.

The utility of complete mitochondrial genome sequences to the analysis of molluscan phylogenetic relationships is reinforced with the addition of the *A. northlandica* sequence. Neogastropoda represent a lineage that appears to have undergone a rapid diversification. Standard analysis of nucleotide sequence is often insufficient to resolve deep relationships in such cases (e.g., birds; Pratt et al. 2009). It is thought that structural organization of mitochondrial genomes ("rare genomic changes") could be used to resolve uncertainties in deep relationships in molluscs (Boore 2006). As the gene content and order of known neogastropod mitochondrial genomes is identical, positional data for genes will not be informative. However, positional information for intergenic spacer regions can provide important additional data. When the *Amalda* sequence is compared with the 12 known neogastropod sequences, a tantalizing picture of lineage-specific arrangements of structure-bearing intergenic spacers emerges. However, very little can be concluded from such a small sample of molluscs. Fortunately, as methods are developed to enable the deconvolution of mixed samples from second-generation sequencing runs, large numbers of mitochondrial genomes or other short genomic regions can now be quickly and cost-effectively generated. Through sufficient sampling of maximally informative taxa, inference of phylogenetic relationships of molluscan lineages will then be robust and free of the bias associated with insufficient taxon sampling and inadequate sequence coverage to achieve resolution.

The mixture strategy that we have developed can readily be combined with an indexing approach. For example, if we wish to sequence mitochondrial genomes from, say, 12 birds, 12 molluscs, 12 insects, and 12 human individuals, rather than using 48 index tags, we could use 12, each with a mixture consisting of 1 bird, 1 mollusc, 1 insect, and 1 human. A single set of 4 reference sequences could then be used to separate all 12 mixtures.

It should be noted that the approach developed here is very general in that it can be applied to a wide range of mixtures of DNA sequences. One that we have simulated is a mixture with a chloroplast and several mitochondria (data not shown), but in principle, any mixture could be used, provided that for each sample we have a reference sufficiently close to separate that sample from the other components of the mixture. However, whatever mixture is tried, we would strongly advocate that the simulation approach be used to test that the software can successfully separate the mixture before committing to the cost of an actual run.

Acknowledgments

The authors wish to thank Trish McLenachan, Chris Corser, Gillian Gibb, Judith Robins, Renae Pratt, and Jan Binnie for sample preparation and Lorraine Berry and Maurice Collins of the Allan Wilson Centre Genome Service for sequencing. We would also like to thank Bruce Marshall of the Museum of New Zealand Te Papa Tongarewa for formally identifying the *Amalda northlandica* specimen and lodging it in the national collection (voucher number NMNZ M.289187). This work was supported by the Allan Wilson Centre for Molecular Ecology and Evolution.

Literature Cited

Akasaki T, et al. 2006. Extensive mitochondrial gene arrangements in coleoid Cephalopoda and their phylogenetic implications. Mol Phylogenet Evol. 38:648–658.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Anderson S, et al. 1981. Sequence and organization of the human mitochondrial genome. Nature. 290:457–465.

Bandyopadhyay PK, Stevenson BJ, Cady MT, Olivera BM, Wolstenholme DR. 2006. Complete mitochondrial DNA sequence of a Conoidean gastropod, *Lophiotoma* (*Xenuroturris*) *cerithiformis*: gene order and gastropod phylogeny. Toxicon. 48:29–43.

Bandyopadhyay PK, et al. 2008. The mitochondrial genome of *Conus textile, cox*I- *cox*II intergenic sequences and Conoidean evolution. Mol Phylogenet Evol. 46:215–223.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26–D31.

Bentley DR. 2006. Whole-genome re-sequencing. Curr Opin Genet Dev. 16:545–552.

Boore JL. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol Evol. 21:439–446.

Cai W-W, Chen R, Gibbs RA, Bradley A. 2001. A clone-array pooled shotgun strategy for sequencing large genomes. Genome Res. 11:1619–1623.

Chaisson M, Pevzner P, Tang H. 2004. Fragment assembly with short reads. Bioinformatics. 20:2067–2074.

Cunha R, Grande C, Zardoya R. 2009. Neogastropod phylogenetic relationships based on entire mitochondrial genomes. BMC Evol Biol. 9:210.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36:e105.

Drummond AJ, et al. 2008. Geneious v4.0 [cited 2008 December 22]. Available from http://www.geneious.com/.

Erlich Y, et al. 2009. DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. 19:1243–1253.

Fenn JD, Song H, Cameron SL, Whiting MF. 2008. A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. Mol Phylogenet Evol. 49:59–68.

Frank DN. 2009. BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. BMC Bioinformatics. 10:362.

Fullwood MJ, Wei C-L, Liu ET, Ruan Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res. 19:521–532.

Gansner ER, North SC. 2000. An open graph visualization system and its applications to software engineering. Software Pract Exper. 30:1203–1233.

Gibb GC, Kardailsky O, Kimball RT, Braun EL, Penny D. 2007. Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. Mol Biol Evol. 24:269–280.

Hernandez D, Francois P, Farinelli L, Osterås M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res. 18:802–809.

Hillier LW, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods. 5:183–188.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature. 408:708–713.

Kim I, et al. 2005. The complete nucleotide sequence and gene organization of the mitochondrial genome of the oriental mole cricket, *Gryllotalpa orientalis* (Orthoptera: Gryllotalpidae). Gene. 353:155–168.

Kingsford C, Schatz MC, Pop M. 2010. Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics. 11:21.

L'Abbé D, Duhaime JF, Lang BF, Morais R. 1991. The transcription of DNA in chicken mitochondria initiates from one major bidirectional promoter. J Biol Chem. 266:10844–10850.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol. 39:174–190.

Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437:376–380.

Ng P, et al. 2006. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic Acids Res. 34:e84.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol Evol. 16:358–364.

Nilsson MA, Gullberg A, Spotorno AE, Arnason U, Janke A. 2003. Radiation of extant marsupials after the K/T boundary: evidence from complete mitochondrial genomes. J Mol Evol. 57:S3–S12.

Pratt RC, et al. 2009. Toward resolving deep Neoaves phylogeny: data, signal enhancement, and priors. Mol Biol Evol. 26:313–326.

R Development Core Team. 2009. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ray DA, Densmore L. 2002. The crocodilian mitochondrial control region: general structure, conserved sequences, and evolutionary implications. J Exp Zool. 294:334–345.

Richly E, Leister D. 2004. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 21:1081–1084.

Robins JH, et al. 2008. Dating of divergences within the *Rattus* genus phylogeny using whole mitochondrial genomes. Mol Phylogenet Evol. 49:460–466.

San Mauro D, García-París M, Zardoya R. 2004. Phylogenetic relationships of discoglossid frogs (Amphibia:Anura:Discoglossidae) based on complete mitochondrial genomes and nuclear genes. Gene. 343:357–366.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74:5463–5467.

Sasuga J, et al. 1999. Gene contents and organization of a mitochondrial DNA segment of the squid *Loligo bleekeri*. J Mol Evol. 48:692–702.

Simison WB, Lindberg DR, Boore JL. 2006. Rolling circle amplification of metazoan mitochondrial genomes. Mol Phylogenet Evol. 39:562–567.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31.

Sumida M, et al. 2001. Complete nucleotide sequence and gene rearrangement of the mitochondrial genome of the Japanese pond frog *Rana nigromaculata*. Genes Genet Syst. 76:311–325.

Tomita K, Ueda T, Watanabe K. 1998. 7-Methylguanosine at the anticodon wobble position of squid mitochondrial tRNA^{Ser}GCU: molecular basis for assignment of AGA/AGG codons as serine in invertebrate mitochondria. Biochim Biophys Acta. 1399:78–82.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zuker M, et al. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In Barciszewski J, Clark BFC, editors. RNA biochemistry and biotechnology. NATO ASI Series, Dordrecht (NL): Kluwer Academic Publishers. 11–43.

Associate editor: B. Venkatesh