High-quality *De Novo* Genome Assembly and Intra-Individual Mitochondrial Instability in the Critically Endangered Kākāpō

Jonas Korlach[1], Jason T. Howard[2], Bruce Robertson[3], Sarah B. Kingan[1], Jill Muehling[1], Primo Baybayan[1], Richard Hall[1], Erich D. Jarvis[2,4]

[1]Pacific Biosciences, Menlo Park, CA 94025, USA; [2]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY 10065, USA; [3]University of Otago, Dunedin, New Zealand, [4]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

Abstract

The kākāpō (*Strigops habroptila*) is a large, flightless parrot endemic to New Zealand. It is highly endangered with only ~150 individuals remaining, and intensive conservation efforts are underway to save this iconic species from extinction. These include genetic studies to understand critical genes relevant to fertility, adaptation and disease resistance, and genetic diversity across the remaining population for future breeding program decisions.

To aid with these efforts, we have generated a high-quality *de novo* genome assembly using PacBio long-read sequencing. Using the new diploid-aware FALCON-Unzip assembler, the resulting genome of 1.06 Gb has a contig N50 of 5.6 Mb (largest contig 29.3 Mb), >350-times more contiguous compared to a recent short-read assembly of a closely related parrot (kea) species. We highlight the benefits of the higher contiguity and greater completeness of the kākāpō genome assembly through examples of fully resolved genes important in wildlife conservation (contrasted with fragmented and incomplete gene resolution in short-read assemblies), in some cases even providing sequence for regions orthologous to gaps of missing sequence in the chicken reference genome. We also highlight the complete resolution of the kākāpō mitochondrial genome, fully containing the mitochondrial control region which is missing from the previous dedicated kākāpō mitochondrial genome NCBI entry. For this region, we observed a marked heterogeneity in the number of tandem repeats in different mtDNA molecules from a single bird tissue, highlighting the enhanced molecular resolution uniquely afforded by long-read, single-molecule PacBio sequencing.

Motivation

Highly endangered:

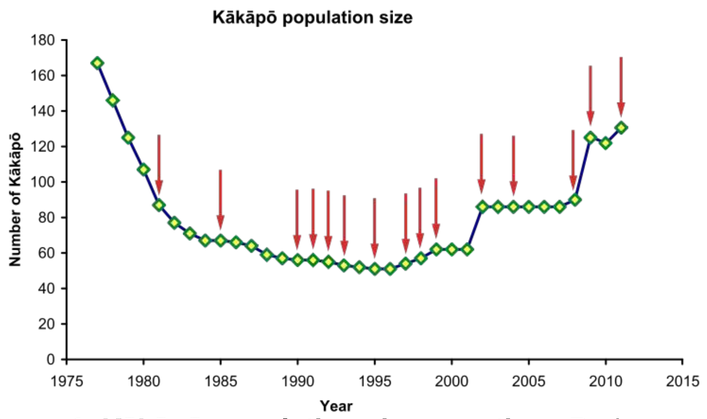
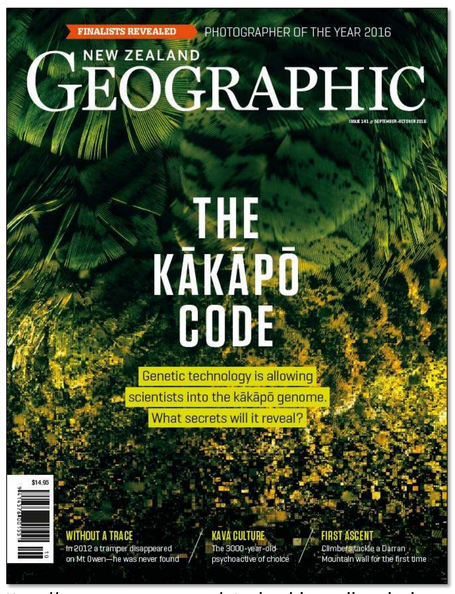


Figure 1. Kākāpō population size over time. Red arrows indicate breeding years. From: https://en.wikipedia.org/wiki/Kakapo#/media/File:Kakapo-population-size.svg

More background information:



https://www.nzgeo.com/stories/decoding-kakapo/

Sequencing & *De Novo* Assembly

Sequencing:

- 61 PacBio RS II SMRT Cells, 55.6 Gb total, P6-C4 chemistry, 6-hour movies

De Novo Assembly:

- Falcon-Unzip assembler
- Primary assembly size: 1.06 Gb
- 1668 contigs
- **Contig N50: 5.63 Mb**
- 140 Mb in alternative haplotig alleles

For comparison, this assembly represents a ~350-fold improvement in contiguity compared to the closest relative species (Kea, 0.016 Mb contig N50, ASM69687v1).

Significance Example

Aggrecan (*ACAN*) gene:

- Candidate gene for chondrodystrophy (defects during development)
- Conventional re-sequencing is difficult because of complex gene structure, including variable repeat regions

ACAN is fully resolved in the kākāpō assembly:

- Present on contig 32 (7.25 Mb)
- Predicted protein length[1]: 2333 a.a (18 exons)
- Repeat region: 26 repeat units (60 bp each)

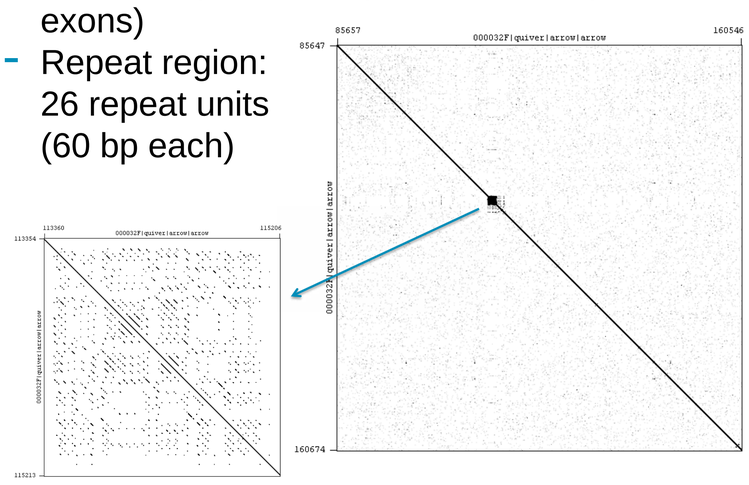


Figure 2. *ACAN* self-dot plot. The inset highlights the variable repeat region.

ACAN protein is 11% larger in the kākāpō compared to chicken (Galgal5):

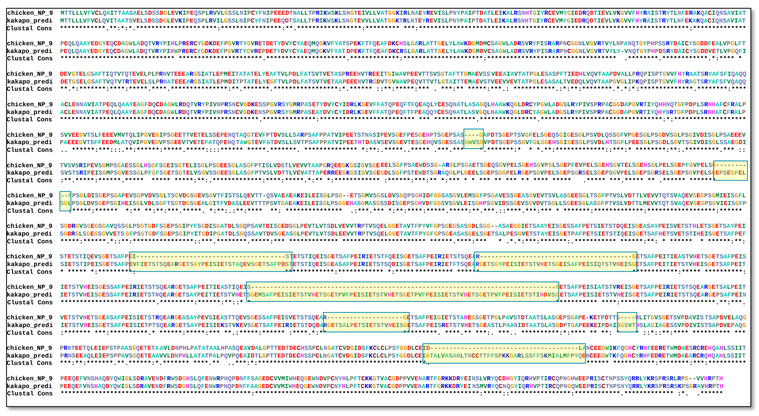
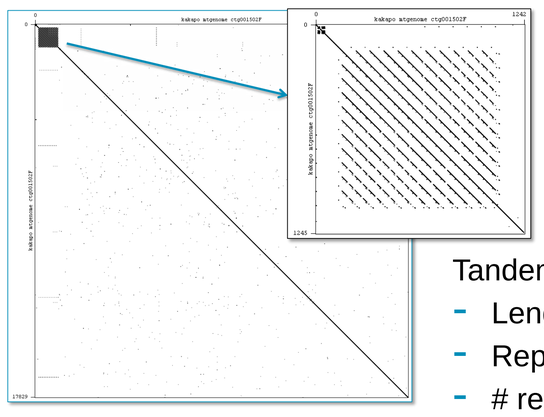


Figure 3. ACAN kākāpō *vs.* chicken comparison. Regions unique to the kākāpō are highlighted.

For comparison, *ACAN* is not fully resolved in the Illumina-based Kea assembly

Mitochondrial Genome Instability

Mitochondrial (mt) genome is fully resolved in the kākāpō assembly:



Tandem repeat:

- Length: ~0.9 kb
- Repeat size: 80 bp
- # repeat units: 11

Figure 4. Self dot plot of the kākāpō mt genome. The inset highlights the tandem repeat region.

For comparison, the dedicated NCBI kākāpō mt genome (AY309456.1) is missing the tandem repeat region. Also, the mt genome is completely missing in the Illumina-based Kea assembly.

Intra-individual tandem repeat instability:

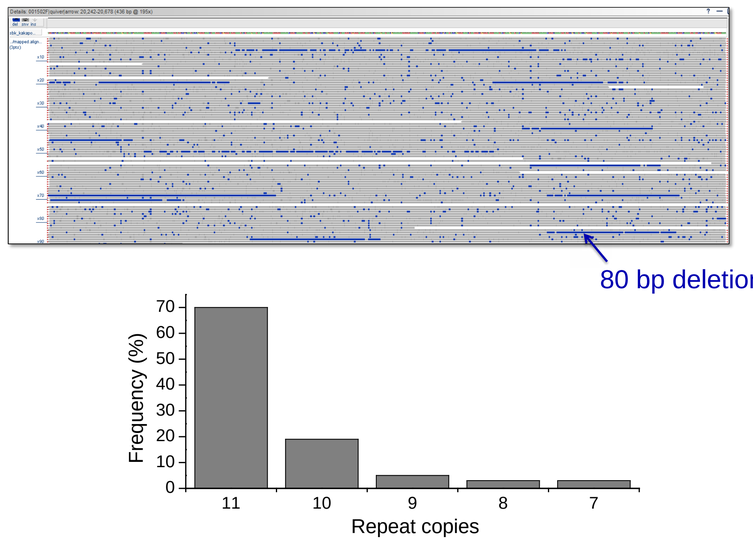


Figure 5. Tandem repeat instability. Individual PacBio reads highlight heterogeneity in the number of repeat units between different mitochondria, quantified in the histogram.

Conclusions

- One of the most contiguous and complete vertebrate genomes to date
- Full resolution of many gene previously fragmented in short-read assemblies, including genes relevant to conservation and breeding efforts
- Single-molecule, long-read nature of PacBio sequencing reveals intra-individual mitochondrial genome instability
- Will become the reference genome for this species, and the representative for the parrot order as part of the Vertebrate Genomes Project (VGP)
- Foundation for further chromosome-scale assembly with addition of scaffolding methods (Bionano, 10X, Hi-C)

References

[1]http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::genscan

Acknowledgements

The authors would like to thank everyone who helped generate data for this study.

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2018 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Advanced Analytical Technologies. All other trademarks are the sole property of their respective owners.