De Novo PacBio Long-read Assembled Avian Genomes Correct and Add to Genes Important in Neuroscience and Conservation Research

Jonas Korlach[1], Gregory Gedman[2], Lindsey Cantin[2], Jason Howard[2], Primo Baybayan[1], Sarah Kingan[1], Richard Hall[1], Jenny Gu[1], Chen-Shan Chin[1], Bruce Robertson[3], Andrew Digby[3], Oliver A. Ryder[4], Erich D. Jarvis[2]
[1]PacBio, Menlo Park, CA, [2]The Rockefeller University, New York, NY & Howard Hughes Medical Institute, [3]University of Otago, Dunedin, New Zealand, [4]San Diego Zoo Institute for Conservation Research, Escondido, CA

Abstract

To test the impact of the quality of long- versus short-read genome assemblies on biological research, we applied PacBio long-read sequencing in conjunction with the new, diploid-aware FALCON-Unzip assembler to a number of bird species. These included: the zebra finch, for which a consortium-generated, Sanger-based intermediate-read length reference exists; Anna's hummingbird, for which a short-read reference exist, generated by the Avian Phylogenomics Consortium phase I; and two critically endangered bird species (kākāpō and ʻalalā) of high importance for conservations efforts, whose genomes had not previously been sequenced and assembled.

All PacBio *de novo* genome assemblies had contiguities in the megabase range (contig N50s ranging between 5.4 and 7.7 Mb), representing a 150-fold improvement over the zebra finch genome reference, and a 200-fold improvement over the hummingbird reference. Allele-resolved contigs of this size range translated into the resolution of thousands of gaps present in the previous finch reference and hummingbird assemblies, correction of erroneous sequence flanking those gaps, correction of misassemblies in the previous assemblies and resolution of complex repeat structures, as well as resolution of allelic differences between the two chromosome haplotypes that caused assembly errors in the haploid references. RNA-Seq coverage was higher on the PacBio-based long-read assemblies, demonstrating more complete gene assembly. For the first time, we were able to assemble the complete genome structure of many critical genes in neuroscience and conservation research. These findings demonstrate the impact of higher-quality, phased and gap-less assemblies *vs.* fragmented, incomplete scaffold-based assemblies in genomic research.

Species Sequenced

Zebra finch

Anna's hummingbird

Kākāpō

ʻAlalā

Phased, Diploid *De Novo* Genome Assemblies

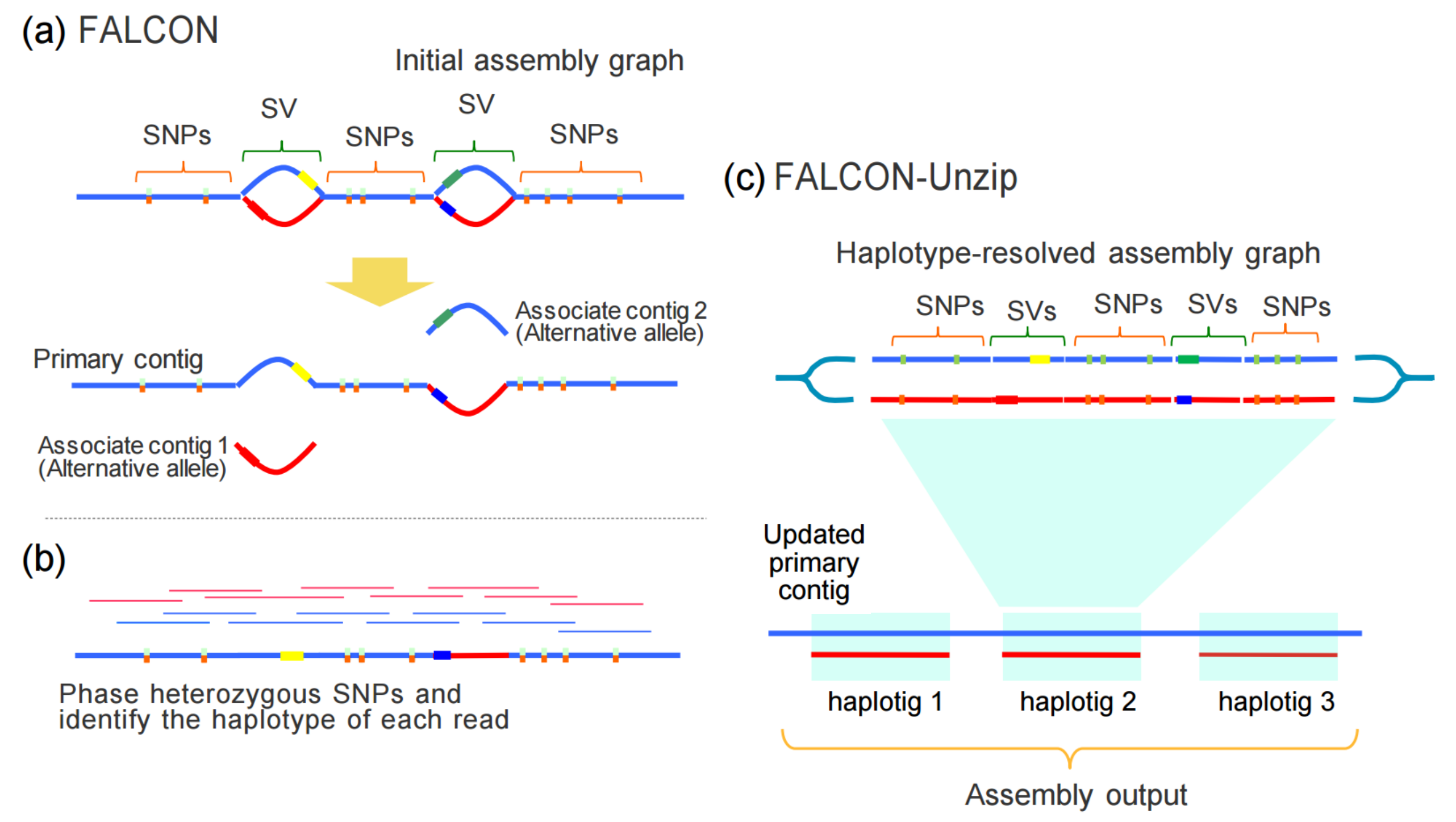
(a) FALCON
Initial assembly graph
SNPs
SV
SNPs
SV
SNPs
Primary contig
Associate contig 2 (Alternative allele)
Associate contig 1 (Alternative allele)
(b)
Phase heterozygous SNPs and identify the haplotype of each read
(c) FALCON-Unzip
Haplotype-resolved assembly graph
SNPs
SVs
SNPs
SVs
SNPs
Updated primary contig
haplotig 1
haplotig 2
haplotig 3
Assembly output

Figure 1. FALCON-Unzip assembler. From Chin et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13(12), 1050-1054.

Assembly Results

Zebra finch:

	Reference genome GCF_000151805 version 3.2.4	FALCON-Unzip Primary assembly	Improvement	FALCON-Unzip Alternate haplotype assembly
Genome size	1.23 Gb	1.14 Gb		0.84 Gb
Contig N50	0.04 Mb	5.81 Mb	**150-fold**	2.74 Mb
# of contigs	124,806	1,159	**108-fold**	2,188

Anna's hummingbird:

	Illumina assembly GCF_000699085	FALCON-Unzip Primary assembly	Improvement	FALCON-Unzip Alternate haplotype assembly
Genome size	1.11 Gb	1.01 Gb		1.01 Gb
Contig N50	0.03 Mb	5.37 Mb	**201-fold**	1.07 Mb
# of contigs	124,820	1,076	**116-fold**	4,895

Kākāpō & ʻAlalā:

	Kākāpō		ʻAlalā	
	FALCON-Unzip Primary assembly	FALCON-Unzip Alternate haplotype assembly	FALCON Primary assembly	FALCON Alternate haplotype assembly
Genome size	1.06 Gb	0.14 Gb	1.09 Gb	0.07 Gb
Contig N50	5.6 Mb	0.13 Mb	11.0 Mb	0.05 Mb
# of contigs	1,668	1,482	1,026	1,435

Genome Completeness Examples

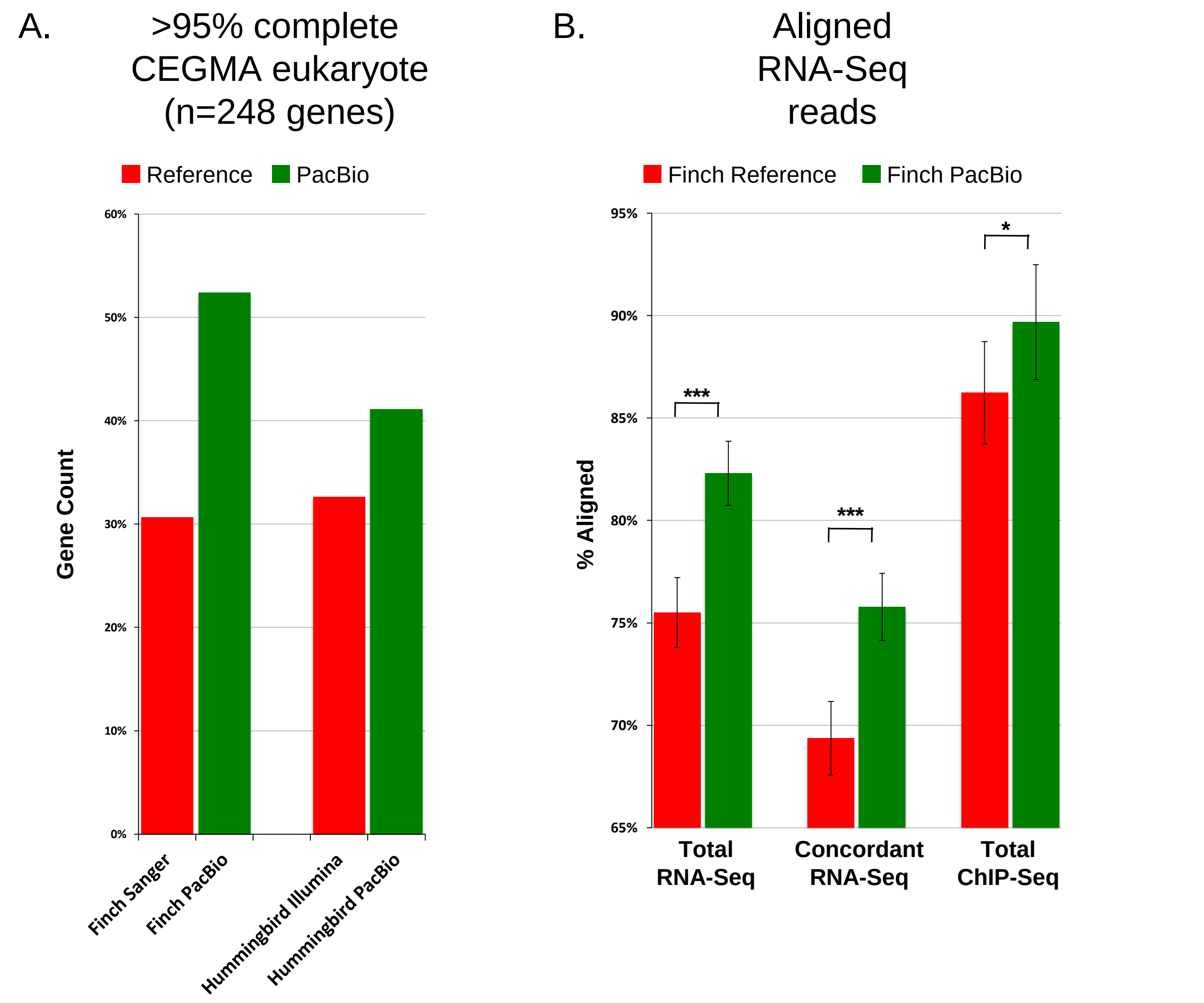

Figure 2. Gene completeness within assemblies. (A) Comparison to a 248 highly conserved core CEGMA eukaryote gene set using human genes (Parra *et al.* 2009), between the Sanger-based zebra finch and Illumina-based Anna's hummingbird references and their respective PacBio-based assemblies. We used a more stringent cut-off (>95%) for completeness than usually done (>90%). Gene count is the percentage of genes in each of the assemblies that meet this criterion. **(B)** Transcriptome and regulome representation within assemblies. Percentage of RNA-Seq and H3K27Ac ChIP-Seq reads from the zebra finch RA song nucleus mapped back to the zebra finch Sanger-based and PacBio-based genome assemblies. * p <0.05; ** p <0.002; *** p <0.0001; paired t-test within animals between assemblies; n = 5 RNA-Seq and n = 3 ChIP-Seq independent replicates from different animals.

Impact Examples

Zebra finch, examples:

	Reference genome (GCF_000151805, version 3.2.4)	FALCON-Unzip PacBio assembly
DUSP1	gross misassembly 77 a.a. missing (exon 2, part of exon 1) gap-flanking sequence region errors	Complete for both alleles
FOXP2	9 gaps within and upstream of gene including promoter region gap-flanking sequence region errors	Complete for both alleles
EGR1	3 gaps surrounding gene including promoter region gap-flanking sequence region errors	Complete for both alleles
SLIT1	14 gaps within and surrounding gene 193 a.a. missing (exons 1, 27, and part of 35) gap-flanking sequence region errors	Complete for both alleles

DUSP1, zebra finch & Anna's hummingbird:

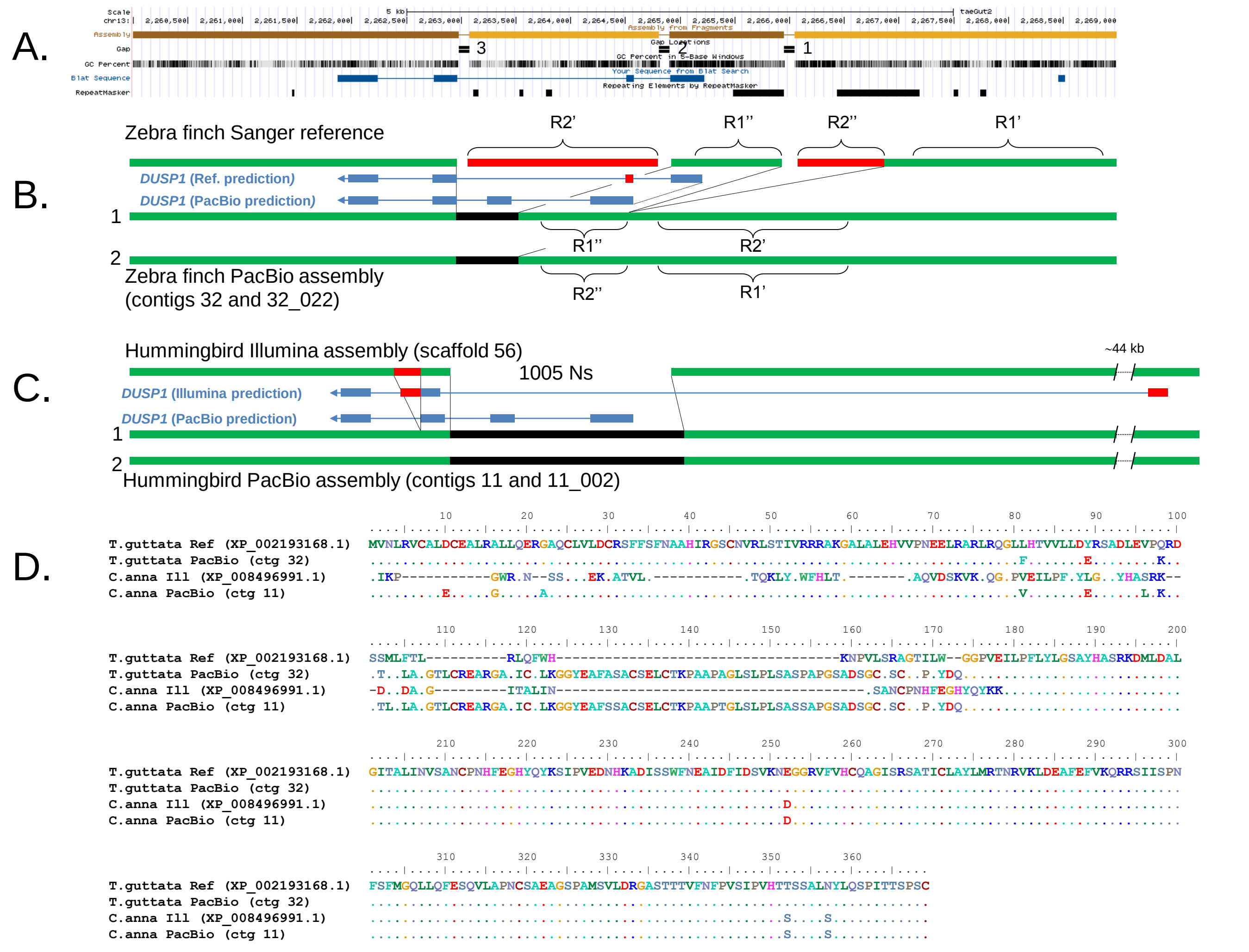
A.
Scale
chr13:
5 kb
taeGut2
Assembly
Gap
GC Percent
Blat Sequence
RepeatMasker
3
2
1
B.
Zebra finch Sanger reference
R2'
R1''
R2''
R1'
DUSP1 (Ref. prediction)
DUSP1 (PacBio prediction)
1
2
Zebra finch PacBio assembly (contigs 32 and 32_022)
C.
Hummingbird Illumina assembly (scaffold 56)
~44 kb
1005 Ns
DUSP1 (Illumina prediction)
DUSP1 (PacBio prediction)
Hummingbird PacBio assembly (contigs 11 and 11_002)
D.
T.guttata Ref (XP_002193168.1)
T.guttata PacBio (ctg 32)
C.anna Ill (XP_008496991.1)
C.anna PacBio (ctg 11)

Figure 3. Comparison of *DUSP1* assemblies. (A) UCSC Genome browser view of the Sanger-based zebra finch *DUSP1* assembly, highlighting four contigs with three gaps, GC content, Blat alignment of the NCBI gene prediction (XP_002193168.1, blue), and repeat sequences. **(B)** Resolution of the region by the PacBio-based zebra finch assembly, filling the gaps (black) and correcting erroneous reference sequences in repeat regions (red) and gene predictions (blue). **(C)** Resolution and correction to the hummingbird Illumina-based assembly with the PacBio-based assembly (same color scheme as in *B*). **(D)** Multiple sequence alignment of the DUSP1 protein for the four assemblies in *B* and *C*, showing numerous corrections to the zebra finch Sanger-based and hummingbird Illumina-based protein predictions by both PacBio-based assemblies.

Conclusions

- PacBio *de novo* long-read assemblies generate high-quality phased, diploid genomes
- >100-fold improvement in contiguity over previous references
- Correction of misassemblies, gaps, erroneous sequences flanking gaps, and resolution of allelic differences
- Improved transcriptome & regulome representation

References & Acknowledgements

Parra G, et al. (2009) Assessing the gene space in draft genomes. *Nucleic Acids Research* 37, 289-297.

We thank Art Arnold for sending us the carcass with tissue used to create the original Sanger-based zebra finch assembly, and Claudio Mello and Peter Lovell again for their previous capture of the hummingbird used for the reference assembly. We thank the PacBio sequencing (Christine Lambert, Jill Muehling) and assembly (Matthew Seetin, Jane Landolin, Lawrence Hon) teams for help with sequencing and assembly, and Winston Timp and Rachael Workman for early access to their Iso-Seq reads of the Ruby-throated hummingbird. We thank Alejandro Burga for noticing indel errors in the original PacBio haploid assembly that lead us work with others to find a correction with phasing, and several members of the Jarvis lab (Mathew Biegler, Ha Na Choe, and Constantina Theofanopoulou) for help with manual analyses of gene models between the assemblies. This work was supported by HHMI funds to E.D.J.