In order to confirm the validity of our sequences and to test the de novo assembly against a reference assembly, we also assembled the junco transcriptome using GMAP [1] with standard parameters against the closest available draft genome assembly: the zebra finch the first passerine genome sequencing project [2], and a species that diverged from juncos approximately 25 million years ago [3]. Briefly, each read was independently aligned to the genome at each location with a match meeting the program's 24-mer criteria to ensure that large intron interruptions did not artificially reduce our sequence alignment rates.

Of the 1,180,500 reads from the junco transcriptome used for this assembly, 1,031,427 (87%) had a significant alignment to the zebra finch genome identified by the software, with an average of 92.9% identity. The high proportion of matches strongly suggests that the sequencing reads from the junco transcriptome are high quality. The high identity supports the relatively close relationship between junco and zebra finch.

In addition, we identified the overlap of locations on the zebra finch genome that had been aligned to a junco read in order to identify regions of the genome that were transcribed in un-interrupted fashion (i.e., exons). Junco sequences aligned in 116,201 regions, representing 26,651,558 unique basepairs, in the zebra finch genome (average length of 331 basepairs, 10.5-fold average depth of coverage). This length is within the empirically identified range of the length of exons in several taxa (150 to 370 bp; [4]), as expected if our sequence reads were from transcribed RNA rather than genomic DNA. The Ensembl release 63 of gene models for the zebra finch [5] contains a total predicted transcribed length of 24,471,883 base pairs. The fact that there are more matches than the predicted length of genes indicates that we have extensive transcriptome coverage.

Many of the reads that mapped to the zebra finch genome aligned to multiple locations (585,709; 56.78% of mapped reads); however, only 8,866 (1.51% of multiply mapped reads) matched to locations on multiple chromosomes, far less than expected by chance ($\chi^2$, d.f. = 1, $p < 0.001$). This finding strongly suggests that the junco 454 reads were split by introns located in the genomic DNA, which is to be expected in sequencing reads that contain only exons and not genomic DNA. Due to the limitation imposed by introns, the lengths of the assembled regions from the reference assembly are substantially shorter, and split into many more groups, than our de novo assembly. For this reason, we chose to use the de novo assembly for all further analyses.

References