2014

# High-coverage sequencing and annotated assemblies of the budgerigar genome

Ganeshkumar Ganapathy

Wesley C. Warren

et al

(GIGA)n SCIENCE

# High-coverage sequencing and annotated assemblies of the budgerigar genome

Ganapathy *et al.*

BioMed Central

Ganapathy *et al. GigaScience* 2014, **3**:11
http://www.gigasciencejournal.com/content/3/1/11

(GIGA)n SCIENCE

DATA NOTE Open Access

# High-coverage sequencing and annotated assemblies of the budgerigar genome

Ganeshkumar Ganapathy[1†], Jason T Howard[1†], James M Ward[2], Jianwen Li[3], Bo Li[3], Yingrui Li[3], Yingqi Xiong[3], Yong Zhang[3], Shiguo Zhou[4], David C Schwartz[4], Michael Schatz[5], Robert Aboukhalil[5], Olivier Fedrigo[6], Lisa Bukovnik[6,13], Ty Wang[2], Greg Wray[7], Isabelle Rasolonjatovo[8], Roger Winer[9], James R Knight[9], Sergey Koren[10,12], Wesley C Warren[11], Guojie Zhang[3*], Adam M Phillippy[10,12*] and Erich D Jarvis[1*]

## Abstract

**Background:** Parrots belong to a group of behaviorally advanced vertebrates and have an advanced ability of vocal learning relative to other vocal-learning birds. They can imitate human speech, synchronize their body movements to a rhythmic beat, and understand complex concepts of referential meaning to sounds. However, little is known about the genetics of these traits. Elucidating the genetic bases would require whole genome sequencing and a robust assembly of a parrot genome.

**Findings:** We present a genomic resource for the budgerigar, an Australian Parakeet (*Melopsittacus undulatus*) – the most widely studied parrot species in neuroscience and behavior. We present genomic sequence data that includes over 300× raw read coverage from multiple sequencing technologies and chromosome optical maps from a single male animal. The reads and optical maps were used to create three hybrid assemblies representing some of the largest genomic scaffolds to date for a bird; two of which were annotated based on similarities to reference sets of non-redundant human, zebra finch and chicken proteins, and budgerigar transcriptome sequence assemblies. The sequence reads for this project were in part generated and used for both the Assemblathon 2 competition and the first *de novo* assembly of a giga-scale vertebrate genome utilizing PacBio single-molecule sequencing.

**Conclusions:** Across several quality metrics, these budgerigar assemblies are comparable to or better than the chicken and zebra finch genome assemblies built from traditional Sanger sequencing reads, and are sufficient to analyze regions that are difficult to sequence and assemble, including those not yet assembled in prior bird genomes, and promoter regions of genes differentially regulated in vocal learning brain regions. This work provides valuable data and material for genome technology development and for investigating the genomics of complex behavioral traits.

**Keywords:** *Melopsittacus undulatus*, Budgerigar, Parakeet, Next-generation sequencing, Hybrid assemblies, Optical maps, Vocal learning

* Correspondence: zhanggj@genomics.cn; aphillippy@gmail.com; jarvis@neuro.duke.edu
†Equal contributors
[3]China National Genebank, BGI-Shenzhen, Shenzhen 518083, China
[10]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20740, USA
[1]Department of Neurobiology, Duke University Medical Center, Durham, NC 27710, USA
Full list of author information is available at the end of the article

## Data description

### Raw genome DNA sequence reads

DNA samples were obtained from a blood sample taken from a single male *Melopsittacus undulatus*, who we aptly named Mr. B. For Illumina sequencing, reads were generated at Duke University (16×), Illumina UK (54×), and BGI (219×) using Illumina's TruSeq [1] version2 or version3 chemistries (Table 1 and GigaDB [2]). The version3 chemistry reads through GC-rich regions, which are often found in promoters, more evenly than does version2 [3]. The insert sizes for the BGI libraries ranged from 220 bp to 40 Kbp, and the insert sizes for the Duke libraries ranged from 400–600 bp, in order to assist assemblies. Fragment sizes for the mate pair libraries, based on genome mapping, and the per base sequence quality distribution for the libraries are shown in GigaDB [2]. The Duke University Illumina libraries were sequenced at two different cluster densities: 8× coverage reads at the normal 420 k clusters/mm density and 8× coverage at a lower 350 k clusters/mm. The lower cluster density was used to increase the number of GC-rich regions sequenced. For PacBio sequencing, 6.76 Gbp (~5.5× coverage) of PacBio RS reads [4] were generated at Pacific Biosciences from two insert size libraries (7.5 K bp at 1.93× and 13 Kbp at 3.56×; PacBio reads error-corrected with Illumina can be downloaded from the supplementary webpage associated with [5]). With all reads combined, the total coverage exceeds 300× (assuming a haploid genome size of 1.23 Gbp) (Table 1), perhaps making Mr. B one of the most sequenced individual vertebrate animals as of to date. The read length distributions of these different types of reads are shown in Figure 1.

### Fosmid Library

To validate the assemblies in the Assemblathon 2 competition, a fosmid library was created from sheared genomic DNA (35–40 Kbp) of Mr. B [6]. Ten pools of clones were generated and sequenced using Illumina as described in [7]. Each pool of reads was individually assembled using Velvet [8]. The fosmid assemblies have been deposited at GigaDB [2]).

### Transcriptome Reads

454 FLX transcriptome reads were generated from brain RNA isolated from two males, neither of whom was Mr. B. An initial set of sequencing runs of both males was conducted at Washington University at St. Louis, producing 89.2 Mb of transcriptome sequence as reported in [9] (NCBI accession numbers SRR029329–30) and were assembled using Newbler [10] into 19,198 contigs. An additional 21× coverage (run label GK0K2XF01) was generated at Duke University from one of the males.

### Assemblies

We present three hybrid assemblies: 1) Budgerigar 454-illumina hybrid v6.3 using the CABOG assembler; 2) Budgerigar PBcR hybrid using the CABOG assembler; and 3) Budgerigar illumina-454 hybrid using the SOAPdenovo*2* assembler. The first two assemblies were annotated, after which, optical-map assisted mega-scaffolds were constructed based on them. As of yet, the SOAPdenovo*2* assemblies have not been annotated or aligned to optical maps. The quality statistics of these assemblies are in listed in Table 2, and brief descriptions of their construction and relative quality are provided in Additional file 1.

### Validating sequence assemblies with optical maps

Optical Mapping is a single molecule system for the construction of ordered restriction maps of whole genomes [11], and it has been used to guide and validate sequence assemblies [12]. An optical map for the budgerirgar genome was created, using a method described in Additional file 1. The optical map contigs ranged in size from 2 Mbp to 74 Mbp and spanned over 900 Mbp with a resolution of 13.94 Kbp (i.e., one non-redundant SwaI every 13.94 Kbp). The contigs were then aligned to *in silico* restriction maps generated from Budgerigar_v6.3 and PBcR assembly scaffolds in order to validate the scaffolds. An approximate 859.21 Mb of the optical maps aligned to the Budgerigar_v6.3 assembly, in 146 scaffolds with 3 or more SwaI restriction fragments (excluding ends and fragments less than 0.4 Kbp). Of these 146 scaffolds, 43 appeared chimeric (i.e., aligned to two or more optical map contigs). For the PBcR assembly, 796.63 Mbp optical map contigs aligned, in 673 scaffolds. Of the 673 scaffolds, only 51 were chimeric. Thus, while the Budgerigar_v6.3 assembly has a higher N50 scaffold metric and hence longer scaffolds compared to the PBcR assembly, 30% the

**Table 1 Summary of genomic reads**

| | Library sizes | Total reads | Total BP (Mb) | Coverage (assuming 1.23 Gbp genome size) |
|---|---|---|---|---|
| **454** | Shotgun, 3 kb, 8 kb, 20 kb mate pair | 41,898,557 | 19,736 | 15.4× |
| **Illumina** | 220, 230, 500, 400–600, 800, 2 kb, 5 kb, 10 kb, 20 kb, 40 kb paired end | 561,074,047 | 356,597 | 289× |
| **Pacific Biosciences** | 7.5Kb, 13 kb | 4,176,242 | 6,763 | 5.5× |
| **Combined** | | 607,148,846 | 383,096 | 309.9× |

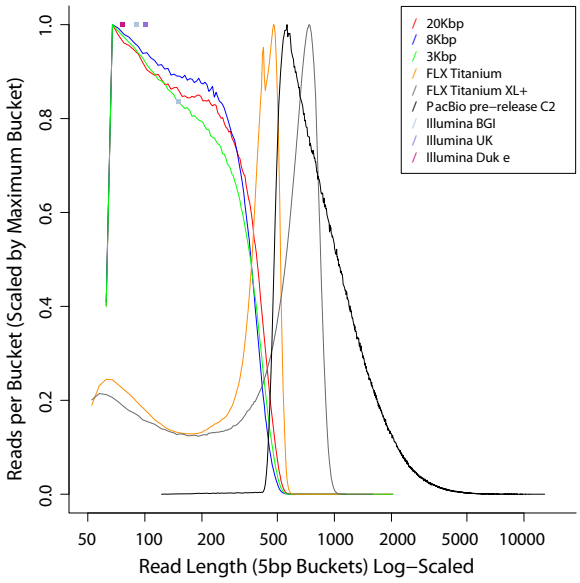


**Figure 1 The distribution of read lengths in 454, Illumina and PacBio budgerigar sequences.** The reads are binned into 5 bp buckets based on their lengths, and the fraction of reads (normalized by the size of the largest bucket) falling into each bucket is shown. Thus, curves shifted towards the right indicate longer read lengths. The reads labeled "20 Kbp", "8 Kbp" and "3 Kbp", "FLX Titanium" and "FLX Titanium XL+" are 454 reads. The reads labeled "PacBio pre-release C2" are uncorrected PacBio reads. The Illumina read lengths appear as colored square boxes, since these read lengths are uniform. The "Illumina Duke" reads are of length 76, The "Illumina UK" reads are of length 101, and the "Illlumina BGI" reads are of lengths 90 or 150. The longest reads come from PacBio sequencing, followed by 454 FLX + (i.e., FLX Titanium XL+) sequencing.

v6.3 scaffolds are chimeric, whereas only 7.6% of the PBcR assembly are chimeric.

### Optical map assisted assemblies

We took both Budgerigar_v6.3 and PBcR assemblies and filtered out alignments that did not extend to the end of either the genomic sequence scaffold or the optical map. The remaining high-quality alignments were then used to identify optical map alignments that bridged scaffolds, such that a single optical map aligned to the ends of at least two sequence scaffolds. We then iteratively extended the megascaffolds beyond pairs of sequence scaffolds, using three heuristics: (1) we limited the overhangs (i.e., the portion of the scaffold sequence that does not align to the optical map) to 2 Mbp total; (2) we bridged two scaffolds together only if the size of the gap separating them is less than 2 Mbp of Ns; and (3) if a sequence scaffold aligned to more than one optical map, we placed it into the largest one it aligns with. The above procedure slightly reduced the number of scaffolds from 25,212 to 25,163 in the Budgerigar_v6.3 assembly, and from 54,668 to 54,138 in the PBcR assembly. This relatively small change in number is expected as our procedure tended to join only sequence scaffolds that were already fairly large into even larger megascaffolds, since it is only possible to confidently align an optical map to a fairly large sequence scaffold bearing numerous SwaI restriction sites. However, this analysis substantially improved the scaffold N50 sizes from 10.6 Mbp to 13.8 Mbp in the Budgerigar_v6.3, and 1.7 Mbp to 7.3 Mbp in the PBcR assemblies, respectively (Table 2). Without limiting the length of the overhangs and gap sizes to 2 Mbp, the increase in N50 scaffold sizes in the Budgerigar_v6.3 is 17.1 Mbp (which we think could be an artifact). We speculate that some of the large gaps in the optical map correspond to centromeres or highly repetitive DNA that are difficult to assemble.

### Annotations

The Budgerigar_v6.3 and PBcR assemblies were annotated at BGI for protein coding genes by first generating a reference set of human, chicken and zebra finch proteins, and then aligning the reference set to the assemblies, and propagating annotations to 30% coverage of the reference at TBlastN, $E = 1e^{-5}$. For the Budgerigar_v6.3 assembly, the reference set comprised of human proteins from Ensembl 60 and a set of zebra finch and chicken proteins re-annotated based on these human proteins, using a custom BGI pipeline reported on separately (Jarvis *et al.* in preparation; Zhang *et al.*, in preparation). For the PBcR assembly, the reference set comprised of the Ensembl 60 human, chicken and zebra finch proteins. The propagation of these reference sets to the budgerigar assemblies is described in more detail in Additional file 1. Further, in the PBcR assembly, UTRs were annotated for 6,203 genes using the GK0K2XF01 transcriptome runs with a pipeline similar to the one described in [13]. The assembly annotations were then propagated to the corresponding sets of megascaffolds. No *de novo* gene annotations were performed.

The annotated Budgerigar assemblies had fewer genes (15,470 and 16,204 genes in the Budgerigar_v6.3 and PBcR assemblies respectively) than the published Zebra Finch (18,618 genes) and Chicken genome assemblies (17,108 genes in the 2011 Galgal4 assembly [14]). We believe the lower number of annotated genes in budgerigar assemblies is due to the differences in annotation methods rather than assembly completeness, for two reasons: (1) These annotations were produced based on similarities to zebra finch, chicken and human proteins, and hence they cannot contain more genes than the source genome annotations; and (2) The independent GenScan annotation of the Budgerigar_v6.3 assembly at the UCSC Genome Browser contains more genes than in zebra finch and chicken, 24,095 in total.

**Table 2 Summary of assemblies**

| | Budgerigar_v6.3 | PBcR | Megascaffolds from Budgerigar_v6.3 + Optical Map | Megascaffolds from PBcR + Optical Map | Illumina + 454 SOAPdenovo2 | Zebra Finch [15] | Chicken *v4* [13]* | Chicken *v3* [16] | Peregrine Falcon [17] | Puerto Rican Parrot [21] | Macaw 1.1 [20] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Assembler | Celera CABOG [25] | PBcR assembler [5] | | | SOAPdenovo2 [26] | PCAP [27] | NA | PCAP [27] | SOAPdenovo [28,29] | Ray [30] | CLC Genomics Workbench |
| Sequence method | 454 FLX, FLX+, Illumina | PacBio corrected with Illumina, 454 FLx, FLx+ | 454 FLX, FLX+, Illumina, Optical Maps. | PacBio corrected with Illumina, 454 FLx, FLx+, Optical Maps. | Illumina, 454 FLx+ | Sanger | Sanger, 454 | Sanger v2.1 | Illumina | Illumina | Illumina, 454 FLx+ |
| Coverage | 14× | 17× | | | 137.59 Illumina, 6.85 FLx+ | 6× | 19.1× | 7.1×s | 107× | 26.9× | 26× |
| Genome size | 1.2Gbp | 1.2Gbp | 1.2Gbp | 1.2Gbp | 1.2Gbp | 1.2Gbp | 1.2Gbp | 1.05Gbp | 1.2Gbp | 1.58Gbp | 1.2 Gbp |
| Total bases in scaffolds | 1,117,358,947 | 1,219,132,003 | 1,118,758,630 | 1,241,439,339 | 1,169,860,945 | 1,224,525,252 | 1,046,932,099 | 1,047,124,295 | 1,174,046,505 | 1,164,566,833 | 997,000 |
| Number of scaffolds | 25,212 | 54,668 | 25,163 | 54,138 | 151,393 | 37,698 | 15,932 | 23,776 | 21,224 | 148,255 | 140,453 |
| Avg. scaffold size | 44,319 | 22,300 | 44,460 | 22,931 | 7,727 | 32,482 | 65,713 | 44,041 | 55,317 | 7,855 | Not available |
| N50 scaffold size | 10,614,387 | 1,705,751 | 13,823,040 | 7,280,340 | 13,497,021 | 10,409,499 | 90,216,835 | 11,125,310 | 3,891,469 | 19,470 | 15,968 |
| Largest scaffold size | 39,887,647 | 11,564,683 | 61,483,320 | 33,208,800 | 66,566,439 | 56,620,707 | 195,276,750 | 51,053,708 | 18,327,016 | 206,462 | 177,843 |
| Total gaps in scaffolds | 51,150 | 26,444 | 51,295# | 27,118 | 60810 | 124,736 | NA | NA | 77,368 | Not available | Not available |
| Number of Contigs | 70,863 | 77,556 | NA | NA | 212,203 | 126,053 | 27,027 | 85,191 | 98,540 | 259,423 | 214,754* |
| Avg. contig size | 15,334 | 15,344 | NA | NA | 4664 | 9,714 | 38,736 | 12,291 | 11,914 | 4,304 | Not available |
| N50 contig size | 55,633 | 102,885 | NA | NA | 51,034 | 38,549 | 279,750 | 45,280 | 28,599 | 6,983 | 6,366 |
| Largest contig size | 465,633 | 849,044 | NA | NA | 500,974 | 424,635 | NA | 624,663 | 247,807 | 75,003 | 87,225 |

*The Chicken v4 assembly consists of chromosomes and not scaffolds with explains the very high scaffold length statistics.
#The increased number of gaps in megascaffolds reflects the fact that each megascaffold may be merger of many original scaffolds with gaps in between them.

### Comparisons to other avian assemblies

Our budgerigar genome assemblies were compared with the zebra finch, chicken, and falcon genomes [15-17]. The other assemblies from the Assemblathon 2 competition are available from GigaDB [18]. The zebra finch and chicken had similar contig and scaffold N50 values (38.5 kb and 10.4 Mb for zebra finch, and 279.8 kb and 90.2 Mb for chicken, respectively). In addition, since the Peregrine Falcon is the closest relative to parrots [19], we also compared the budgerigar genome assemblies to this bird. However, it was not possible to do an in depth comparison of these genomes to the recently sequenced Scarlet Macaw and Puerto Rican Parrot genomes [20,21], because both bird genomes had N50 scaffold sizes under 20,000 and N50 contig sizes under 7,000. A summary of assemblies, including the Scarlet Macaw and Puerto Rican Parrot, are shown in Table 2. Apart from the standard genome assembly quality statistics, we assessed the quality of the budgerigar assemblies along two other dimensions: (1) the coverage of highly conserved avian exons, and (2) the number of gaps 10 Kbp upstream and downstream of each gene (gene territories), and conversely, the number gene territories assembled without gaps. Of 3,288 highly conserved exons (>86% coverage across >87% of their length) we identified between chicken and zebra finch, 3,165 (96.25%) and 3,134 (95.31%) were covered with >86% identity across >87% of their length in the Budgerigar_v6.3 and PBcR assemblies respectively, pointing to good coverage of coding regions in these assemblies. The budgerigar assemblies had fewer gaps within the coding sequences and gene territories than all other avian genomes examined, except the newer unpublished Galgal4 chicken assembly that is similar to the budgerigar in that it is a hybrid that includes both short and long sequences (Sanger and 454 FLX+) (Figure 2). This suggests that our budgerigar assemblies have very well assembled genes and promoter regions.

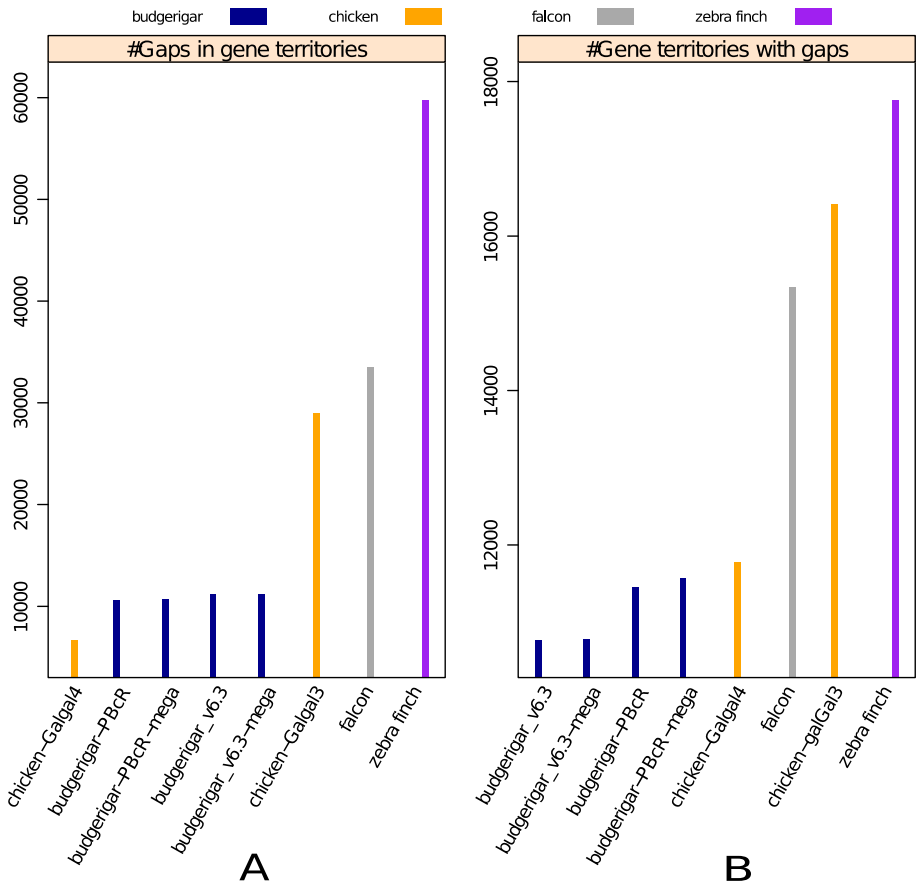


**Figure 2 Number of nucleotide gaps assess relative assembly incompleteness. A)** Shows the total number of gaps in genes and the surrounding 10,000 base pair regions upstream and downstream (collectively called gene territories). **B)** Shows the number of such gene territories with gaps. In both the panels, different species assemblies are colored differently, with the budgerigar assemblies shown in dark blue. The budgerigar assemblies with the "-mega" suffix are optical map enhanced versions of the Budgerigar_v6.3 and PBcR assemblies. The budgerigar assemblies have the highest numbers of gapless gene territories (right panel) and the fewest number of gaps of all assemblies except the recent chicken v4 assembly, which used a similar technology (left panel).

Using the online CoGe tool [22-24], we assessed the structural similarities between the various budgerigar assemblies and other avian assemblies [25-30], by computing the level of coding sequence synteny among assemblies. The highest numbers of genes in synteny were observed, as expected, between a budgerigar assembly and the optical map assisted version of the same assembly (Figure 3A). However, the number of genes in synteny between the Budgerigar_v6.3 and the PBcR assemblies was similar to the number of genes in synteny between budgerigar and falcon (Figure 3A, B). Further, the number of genes in synteny did not strictly reflect phylogenetic relationships, as the zebra finch and budgerigar, close relatives [19], had a lower level of synteny than budgerigar and chicken. In addition, a number of inversions were observed even in the syntenic dotplots between the original budgerigar assemblies and their optical map-assisted assemblies (88 inversions between Budgerigar_v6.3 and Budgerigar_v6.3_mega; 209 inversions between PBcR and PBcR_mega, plots shown in GigaDB [2]). This suggests that synteny based on CoGE syntenic maps is affected by the quality of the assemblies and the characteristics of the synteny algorithm. Thus, the number of genes in synteny computed using the available methods is only a rough measure of the actual structural similarity between the assemblies compared.

In summary, this study shows that the budgerigar genomic resource we have generated has provided [5,6] (and is still expected to provide more) valuable data and material for genome technology development and for further investigating complex behavioral traits at the genomics level.

All procedures on live animals were approved by the Institutional Animal Care and Use Committee of Duke University.

## Availability and requirements

The genomic sequence reads have been deposited in NCBI's sequence read archives (SRA) and the EBI's ENA archive, under the same project accession number ERP002324. The SOAPdenovo2 assembly has been submitted to GigaDB by the Assemblathon 2 team and is available at GigaDB [18]. Other supporting resources that have been deposited in GigaDB [2] are:

- Duke University brain transcriptome reads.
- Budgerigar_v6.3, PBcR assemblies (contigs and scaffolds) and optical map assisted megascaffolds based on these two assemblies (two contigs and four scaffolds in total).
- The per base sequence quality distribution of the paired end and mate paired libraries. The estimated fragment length distribution of the mate paired libraries. Peptide and coding sequences (CDS) for the Budgerigar_v6.3 and PBcR assemblies.
- Gene annotations and Repeat Masker annotations for the scaffolds.
- Optical map alignments of Budgerigar_v6.3 and PBcR assemblies in Microsoft Excel and XML formats and software (Gnomspace.rar) to view the XML alignments.
- The optical map dataset.

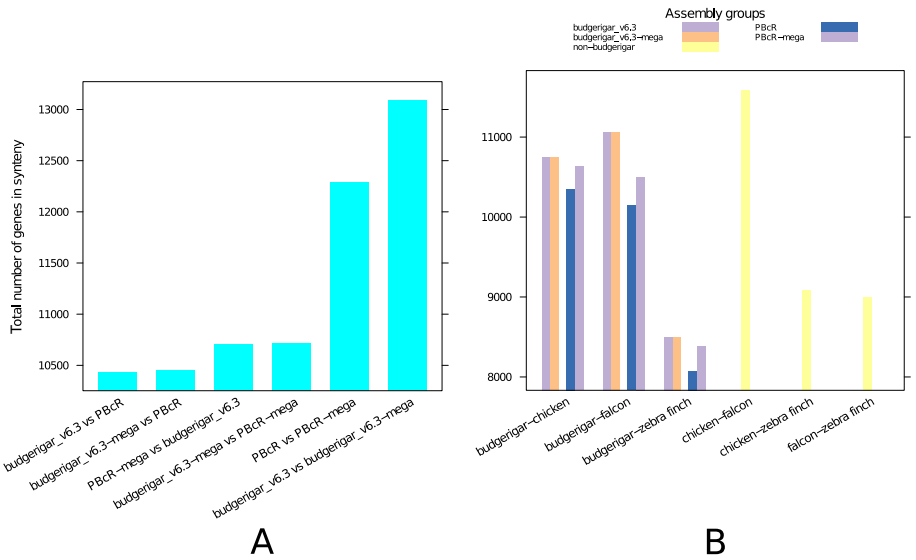


**Figure 3 The number of genes that are part of a syntenic block between different budgerigar assemblies (A) and between budgerigar and non-budgerigar assemblies (B).** The numbers were calculated from CoGE syntenic dotplots (not shown), as the total number of genes represented in syntenic blocks. The y-axis limits have been cut off close to the minimum value in the plot to show a more detailed spread of values.

## Additional file

**Additional file 1: Supplementary materials.**

### Abbreviations

CABOG: Celera assembler with the best overlap graph; CoGE: Comparative genomics; PBcR: Pac bio corrected reads; XML: Extensible markup language.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

JH, GG, JW, JL, BL, OF, LB, TW, GW, IR, RW, JK, WW, GZ, and EDJ contributed to generating and analyzing the genomic reads. SK, JW, AP, MS, RA, WW, EDJ contributed to the genome assemblies. SZ, DCS, MS, RA worked on generating the optical maps and optical map assemblies. JH, JW, OF, LB, TW, GW, WW, AP, EDJ contributed to generating and analyzing the transcriptome reads. GG, JH, and EDJ wrote the paper. All authors read and approved the final manuscript.

### Authors' information

JH, EJ, GZ are members of the Bird 10 K project.

### Acknowledgements

We thank Graham Alexander (Duke Institute for Genome Sciences & Policy [IGSP]) for his work with the 454 sequencing, James Furbee (Roche) for his role in coordinating the sequencing of the 454 MP libraries and for assisting in the optimization of the 454 FLX + chemistry, Fangfei Ye and Nicholas Hoang (both from Duke IGSP) for their work with the Illumina sequencing, and Xiaoxia Qin, from Duke IGSP, for her advice on assembling the budgie genome. We are very appreciative of Tin Le (Gentris Corporation) for his efforts on coordinating the low-density Illumina sequencing and his role in optimizing the TruSeq3 approach. We also thank Brian Kelly, Edwin Hauw and Swati Ranade (Pacific Biosciences) for supervising and assisting with the PacBio sequencing. Optical mapping was supported in part by NHGRI R01HG000225 (DCS) and R01HG004348 (1K; subcontract to DCS). We thank Roche, Illumina and Pacific Biosciences corporations for providing sequencing and computational resources. Finally, we thank the G10K group and the Assemblathon2 group for including Budgerigar as one the model genomes in the Assemblathon2 competition.

### Author details

[1]Department of Neurobiology, Duke University Medical Center, Durham, NC 27710, USA. [2]National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health, Research Triangle Park, Raleigh, NC 27709, USA. [3]China National Genebank, BGI-Shenzhen, Shenzhen 518083, China. [4]Department of Chemistry, The Laboratory for Molecular and Computational Genomics, Laboratory of Genetics and Biotechnology Center, University of Wisconsin, Madison, WI 53706, USA. [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY 11724, USA. [6]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27710, USA. [7]Department of Biology, Center for Systems Biology, Duke University, Durham, NC 27710, USA. [8]Illumina Cambridge Ltd, Cambridge, UK. [9]454 Life Sciences, Branford, Connecticut 06405, USA. [10]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20740, USA. [11]The Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA. [12]National Biodefense Analysis and Countermeasures Center, Frederick, MD 21702, USA. [13]Advanced Liquid Logic Morrisville, Morrisville, NC 27560, USA.

Received: 7 October 2013 Accepted: 3 June 2014
Published: 8 July 2014

### References

1. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53–59.
2. Ganapathy G, Howard JT, Koren S, Phillippy A, Zhou S, Schwartz D, Schatz M, Aboukhalil R, Ward JM, Li J, Li B, Fedrigo O, Bukovnik L, Wang T, Wray G, Rasolonjatovo I, Winer R, Knight JR, Warren W, Zhang G, Jarvis ED: **De novo high-coverage sequencing and annotated assemblies of the budgerigar genome.** *GigaSci Database* 2013. http://gigadb.org/dataset/100059.
3. **Illumina HiSeq.** 2000. [www.illumina.com/Documents//products/brochures/brochure_truseq_v3_advancements_for_hiseq_systems.pdf]
4. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, DeWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, *et al*: **Real-Time DNA Sequencing from Single Polymerase Molecules.** *Sci* 2009, **323:**133–138.
5. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM: **Hybrid error correction and de novo assembly of single-molecule sequencing reads.** *Nat Biotechnol* 2012, **30:**693–700.
6. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W-C, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, *et al*: **Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.** *GigaSci* 2013, **2:**10.
7. Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, Simmons LE, Gammill HS, Rubens CE, Santillan DA, Murray JC, Tabor HK, Bamshad MJ, Eichler EE, Shendure J: **Noninvasive Whole-Genome Sequencing of a Human Fetus.** *Sci Transl Med* 2012, **4:**137–176. 137ra76.
8. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
9. Künstner A, Wolf JBW, Backström N, Whitney O, Balakrishnan CN, Day L, Edwards SV, Janes DE, Schlinger BA, Wilson RK, Jarvis ED, Warren WC, Ellegren H: **Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species.** *Mol Ecol* 2010, **19:**266–276.
10. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376–380.
11. Lin J, Qi R, Aston C, Jing J, Anantharaman TS, Mishra B, White O, Daly MJ, Minton KW, Venter JC, Schwartz DC: **Whole-Genome Shotgun Optical Mapping of Deinococcus radiodurans.** *Science* 1999, **285:**1558–1562.
12. Zhou S, Bechner MC, Place M, Churas CP, Pape L, Leong SA, Runnheim R, Forrest DK, Goldstein S, Livny M, Schwartz DC: **Validation of rice genome sequence by optical mapping.** *BMC Genomics* 2007, **8:**278.
13. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl Automatic Gene Annotation System.** *Genome Res* 2004, **14:**942–950.
14. **Gallus gallus 4.0 Assembly.** [http://www.ncbi.nlm.nih.gov/assembly/GCA_000002315.2]
15. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TAF, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin Y-C, George J, Sweedler J, Southey B, Gunaratne P, Watson M, *et al*: **The genome of a songbird.** *Nature* 2010, **464:**757–762.
16. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, Dodgson JB, Map G, Fingerprint Assembly SA, Chinwalla AT, Cliften PF, Clifton SW, Delehaunty KD, Fronick C, Fulton RS, Graves TA, Kremitzki C, Layman D, Magrini V, McPherson JD, Miner TL, Minx P, Nash WE, Nhan MN, Nelson JO, Oddy LG, *et al*: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:**695–716.
17. Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H, Chen Y, Xia J, Luo Q, Xu P, Chen Y, Liao S, Cao C, Gao S, Wang Z, Yue Z, Li G, Yin Y, Fox NC, Wang J, Bruford MW: **Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle.** *Nat Genet* 2013, **45:**563–566.
18. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G,

Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, *et al*: **Assemblathon 2 assemblies.** *GigaSci Database* 2013. http://dx.doi.org/10.5524/100060.

19. Hackett SJ, Kimball RT, Reddy S, Bowie RCK, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han K-L, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T: **A Phylogenomic Study of Birds Reveals Their Evolutionary History.** *Sci* 2008, **320:**1763–1768.
20. Seabury CM, Dowd SE, Seabury PM, Raudsepp T, Brightsmith DJ, Liboriussen P, Halley Y, Fisher CA, Owens E, Viswanathan G, Tizard IR: **A Multi-Platform Draft de novo Genome Assembly and Comparative Analysis for the Scarlet Macaw (Ara macao).** *PLoS One* 2013, **8:**e62415.
21. Oleksyk TK, Pombert J-F, Siu D, Mazo-Vargas A, Ramos B, Guiblet W, Afanador Y, Ruiz-Rodriguez CT, Nickerson ML, Logue DM, Dean M, Figueroa L, Valentin R, Martinez-Cruzado J-C: **A locally funded Puerto Rican parrot (Amazona vittata) genome sequencing project increases avian data and advances young researcher education.** *GigaSci* 2012, **1:**14.
22. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M: **Finding and Comparing Syntenic Regions among Arabidopsis and the Outgroups Papaya, Poplar, and Grape: CoGe with Rosids.** *Plant Physiol* 2008, **148:**1772–1781.
23. Lyons E, Freeling M: **How to usefully compare homologous plant genes and chromosomes as DNA sequences.** *Plant J* 2008, **53:**661–673.
24. **CoGE.** [http://genomevolution.org/CoGe/]
25. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 2008, **24:**2818–2824.
26. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.** *GigaSci* 2012, **1:**18.
27. Huang X, Wang J, Aluru S, Yang S-P, Hillier L: **PCAP: A Whole-Genome Assembly Program.** *Genome Res* 2003, **13:**2164–2170.
28. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20:**265–272.
29. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC-C, Zhou Y, Cao J, Sun X, Fu Y, *et al*: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463:**311–317.
30. Boisvert S, Laviolette F, Corbeil J: **Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies.** *J Comput Biol* 2010, **17:**1519–1533.

doi:10.1186/2047-217X-3-11
**Cite this article as:** Ganapathy *et al.*: High-coverage sequencing and annotated assemblies of the budgerigar genome. *GigaScience* 2014 **3**:11.