Draft *de novo* genome sequence of *Agapornis roseicollis* for application in avian breeding

Henriëtte van der Zwan*[,1], Francois van der Westhuizen[1], Carina Visser[2], Rencia van der Sluis[1]

[1] Centre for Human Metabolomics, North-West University, Potchefstroom, North-West, South Africa and [2] Department of Animal and Wildlife Sciences, University of Pretoria, Pretoria, Gauteng, South Africa

*To whom correspondence should be addressed.

Abstract

In aviculture, lovebirds are considered one of the most popular birds to keep. This African parakeet is known for its range of plumage colours and ease to tame. Plumage variation is the most important price-determining trait of these birds, and also the main selection criterion for breeders. Currently, no genetic screening tests for traits of economic importance or to confirm pedigree data are available for any of the nine lovebird species. As a starting point to develop these tests, the *de novo* genome of *Agapornis roseicollis* (rosy-faced lovebird) was sequenced, assembled and annotated. Sequencing was done on the Illumina HiSeq 2000 platform and the assembly was performed using SOAPdenovo v2.04. The genome was found to be 1.1 Gb in size and 16 044 genes were identified and annotated. This compared well with other previously sequenced avian genomes like the chicken, zebra finch and budgerigar. In order to assess genome completeness, the number of Benchmarking Universal Single-Copy Orthologs (BUSCOs) were identified in the genome. This was compared to other previously assembled avian genomes and the results indicated that the genome will be useful in development of genetic screening tests to aid lovebird breeders in selecting breeding pairs.

Keywords: avian genomics, whole genome sequencing, bioinformatics.

Contact: henriettevdz@gmail.com

Introduction

Birds from the genus *Agapornis* (commonly called lovebirds) form part of the sub-family Psittacinae, or parrots. There are 356 parrot species distributed over Africa, Asia, Australia and South America and are recognized by their blunt bill with a downward curving upper mandible [1]. There are nine lovebird species, eight of which are native to Africa and one to Madagascar [2]. *Agapornis roseicollis* (rosy-faced lovebird) is indigenous to the South-Western parts of Africa including Angola, Namibia and South-Africa [2, 3]. Wild flocks are still found, however lovebirds are best known as pets due to the fact that they are easily tamed and breed effortlessly in captivity [3].

There are 20 naturally occurring *A. roseicollis* colour variations, all of which are inherited as Mendelian traits, either autosomal or sex-linked recessive or dominant. Some of these variants are shown in Figure 1. The plumage colour dictates the price of the bird so breeders experiment by crossing birds of different plumage colours attempting to breed birds with unique plumage colourations. Some of the species can also be interbred to produce viable hybrid offspring, often with spectacular colour variations (Personal communication, Mr D van den Abeele).

Parrots have a unique mechanism of coloration and unlike most bird species do not utilize carotenoids to generate red, yellow, orange, green, blue or violet plumage colours [4, 5, 6, 7, 8]. Parrot pigments differ both chromotographically and spectrally from all other feather carotenoids [6, 8] and is called psittacofulvins. Psittacinae is the only aves family that express psittacofulvins in their plumage cells. Although the physiological or anatomical sources of these molecules are still unknown, it is strongly believed that diet does not play a role in parrot coloration.

Due to the strong economic influence, the main trait breeders are selecting for is colour. As most of the colours are inherited in a Mendelian recessive manner, chicks born from parents with colour variants, but that display wild type coloration, may be heterozygous for the rare colour. Since there is no screening test available, bird sellers cannot guarantee that the chick is a heterozygote of that colour, nor can they confirm the parentage of the chick. Developing a SNP-based test to genotype individuals could also result in a reduction of illegal parrot trade. It is estimated that 300 000 illegal parrots are sold in the United States every year. This can be lowered if genotyping of birds can be done on a routine basis [9].

In an attempt to address the lack of genetic screening tests for lovebird breeders worldwide, the first step was to perform a *de novo* genome sequencing and assembly of *A. roseicollis*. Three parrot genomes namely the Puerto Rican parrot (*Amazona vittata*) [10] scarlet macaw (*Ara macao*) [11] and the budgerigar (*Melopsittacus undulates*) [12] have been sequenced previously. In 2014, Zhang *et al*. published the genomes of 48 bird species including that of the budgerigar [13].

Here we describe the first *de novo* attempt to sequence, assemble and annotate the genome of *A. roseicollis.*

Materials and methods

Sample selection

One adult male (one year in age), that was born and bred in captivity in Belgium, was selected as the *Agapornis roseicollis* reference genome. The criteria set for sample selection was the availability of a minimum of five generations pedigree data as well as information on all plumage colours of its ancestors. After ethical approval was obtained from the North-West University AnimCare committee (Ethics number NWU00348-15-S5), 400 µL blood collected by a veterinarian in an EDTA tube was taken from the bird. Blood samples were shipped on ice to Germany for sequencing.

Whole genome sequencing

Genomic DNA was isolated from the blood sample using the Machery Nagel Blood Mini kit according to manufacturer’s protocol with one change in that 10 µl blood diluted with 190 µl PBS were used as starting material. Sequencing was performed by Eurofins Genomics in Germany. The sequencing was done at an average total depth of 100X coverage.

Three shotgun libraries were constructed consisting of fragment sizes of 300, 550 and 750 kb respectively. Creation of the Shotgun library was done by using commercially available kits (NEBNext® Ultra™ DNA Library Prep Kit for Illumina, article number E7370) according to the manufacturer's instructions. 1 µg of DNA was fragmented using a Covaris E210 Instrument (Covaris Inc., Woburn, MA) according to manufacturer's instructions. End-repair, A-tailing and ligation of indexed Illumina Adapter, size selection and amplification was performed accordingly. The resulting fragments were cleaned up and pooled. Additionally, one 3kb and one 8kb mate-pair-like library (LJD library, Eurofins Medigenomix, Ebersberg, Germany, proprietary protocol) were prepared based on the mate-pair library protocol from Illumina, modified by using adaptor-guided ligation of genomic fragments which achieves higher accuracies. The resulting fragments were cleaned up and pooled.

The libraries were analysed on the Illumina HiSeq 2000 platform with chemistry version 4 and sequencing was performed using manufacturer's instructions. Paired-end sequencing using 125bp read length was performed on a HiSeq machine using v4 chemistry (HiSeq Control Software 2.2.38). For processing of raw data RTA version 1.18.61 and CASAVA 1.8.4 was used to generate FASTQ-files.

Genome assembly

The lovebird genome assembly and annotation was performed at BGI in collaboration with the Bird 10K (B10K) genome project [14]. The lovebird genome was assembled using *de novo* software SOAPdenovo v2.04 [15]. Filtering of low quality reads were done on the basis of k-mer frequency error correction by removing reads with more than 10% ambiguous bases. Reads were filtered and removed if more than 40% of bases were of low quality and duplicate reads were removed.

De Bruijn graphs were constructed by SOAPdenovo [15] and then tips were clipped, bubbles merged and low coverage links removed. Contigs were collected using kmer lengths of 69 and all usable libraries reads were mapped to contig sequences to construct scaffolds. Pair-end reads were used to fill gaps between scaffolds as one end of the read will align with a contig and the other end with the gap. The Gapcloser module of SOAPdenovo [15] was used to fill gaps within scaffolds. The genome was uploaded onto NCBI with accession number NDXB01000000.

Genome annotation

Transposable Elements were identified by executing RepeatProteinMask in RepeatMasker 4.0.5 [16] by comparing the assembly against the Repbase TE library (Repbase-17.06). This identifies transposable elements by aligning the genome sequence to a self-taken curated transposable element protein database. The default parameters of RepeatModeler [16] was used to build a *de novo* lovebird repeat library and this resulted in consensus sequences and classification information for each repeat gene family.

Gene annotation was done in two parts – homology based gene prediction and gene function annotation. During the homology based gene prediction stage, Ensembl gene sets (release 60) [17] of the chicken (*Gallus gallus*), zebra finch (*Taeniopygia guttata*) and human (*Homo sapiens*) genomes were used to annotate the protein-coding genes. These species were selected based on the fact that their genomes were sequenced with a very high coverage and the annotation of genes will more likely be correct. If future gene studies are to be conducted on this genome it is advisable to use closer related species, e.g. the budgerigar, genomes to re-annotate the genome. This gene set's protein sequences were used as reference templates for homology-based gene predictions. The gene function annotation phase relied on the annotation of the motifs and domains of the reference genes by using InterPro. Databases in the public domain were used for this part of the annotation. Gene products were presented by Gene Ontology and retrieved from InterPro results. Lastly the reference genes were mapped to Swiss-prot database to find the best match for each gene [18].

The annotation pipeline was executed by firstly performing a rough alignment where protein sequences of the reference gene set was aligned using TBLASTN. The cut-off value was 1e^{-5}. genBlastA was used to find the corresponding gene loci with the blast hits. All loci with homologous block lengths shorter than 30% of the query protein were excluded. A precise alignment followed where sequences of candidate gene loci were extracted and using GeneWise v2.2.0, the precise alignment was concluded. MUSCLE v3.8.31 was then run on the predicted protein and reference protein, where-after predicted proteins were filtered out if shorter than 30 aa or had a percent identity less than 25% as well as pseudogenes. Lastly a non-redundant gene set was built to ensure that no gene overlaps were annotated.

Comparing avian genomes

The lovebird genome was compared to other avian genomes to assess its size, number of genes, scaffold N50 size, contig N50 size and gene completeness. The chicken (Gallus_gallus-5.0) (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*) genomes were selected as they are considered as the avian model organisms. The budgerigar (*Melopsittacus undulates)*, kea (*Nestor notabilis*) and Peregrine falcon (*Peregrine falcon*), Puerto Rican parrot (*Amazona vittata*) and Scarlet macaw (*Ara macao*) were included as they are close relatives of the lovebird.

Assessing gene completeness

Bradnam e*t al.* (2013) [21] reports the metrics used to assess the quality of a genome. In addition to Scaffold N50 length, Contig N50 length and the number of scaffold sequences that are gene sized, the number of core genes mapped in the annotation is also important. Making use of the open-source software that is implemented in Python, BUSCO v2 (Benchmarking Universal Single-Copy Orthologs) [22] the lovebird genome as well as seven other previously assembled bird genomes were assessed based on gene completeness using the above mentioned method. All genomes were compared to a set of 303 conserved eukaryote genes namely OrthoDB v9.1. OrthoDB v9.1 is a set of orthologs that is used by BUSCO v2 to assess gene completeness [23]. OrthoDB covers 5756 species including bacteria, eukaryotes, fungi, plants, archaea and viruses.

Results and Discussion

The lovebird genome was estimated to consist of 1 159 816 267 base pairs. Sequencing coverage of each sequencing library can be calculated by dividing the number of bases generated by the size of the genome. The sequencing coverage per library is shown in Table 1. The contig N50 and scaffold N50 lengths were 5 455 and 108 514, respectively. The G/C content of the genome was 43%.

Zhang *et al.* [13] reported the assembled genomes of 48 bird species. From these species, five species were selected for this study and the genomes compared. Two additional parrot species, the Puerto Rican parrot [10] and Scarlet macaw [11], was also included as shown in Table 2.

The size of the lovebird genome was comparable to all seven the other genomes except the Puerto Rican parrot. The larger size of the Puerto Rican parrot could be due to a poor assembly and lower genome coverage since it is larger than any of the avian genomes sequenced by Zhang *et al.* (2014) [13]. The lovebird genome was sequenced at 100x coverage and the scaffold N50 lengths were shorter than comparative genomes. This could be as a result of shorter read lengths resulting in shorter contigs and scaffolds. Zhang *et al.* (2014) also note that the use of 20 kb insert libraries will increase these metrics and it was not included in this study. Since scaffold N50 length does not give an indication on gene content, and in this study gene function is of greater importance than genome completeness, the lower N50 scaffold and contig values are not of great concern [21, 22].

During the genome annotation, 15 045 gene coding sequences were identified and 999 non-coding sequences were identified. This compares well with the other genomes, with the exclusion of the chicken genome, where between 18 618 (zebra finch) and 14 074 (kea) genes were annotated (Table 2). The chicken genome (v Gallus_gallus-5.0) was found to have 26 640 genes [19]. It should however be noted that this is the fifth draft of the chicken genome and that it has been sequenced using Sanger sequencing as well as various other Next Generation Sequencing platforms and therefore is more complete and accurate.

By comparing the genome to a eukaryotic dataset comprising of 303 genes from OrthoDB v9.1 [23] we found 258 or 85.2% complete BUSCOs. Simão *et al.* (2015) [22] defines a complete BUSCO as a gene with a length within two standard deviations of the BUSCO gene group mean length. 7.3% of the genes were fragmented which is defined as genes only partially recovered and 7.5% of genes were missing indicating that they were not identified at all, as displayed in Figure 2.

BUSCOs were also analysed in the other seven avian genomes that were compared and the results shown in Table 3.

The number of complete BUSCOs identified from the lovebird genome corresponds well with the number of complete BUSCOs identified in other genomes such as the budgerigar, chicken and zebra finch. The lower number of complete BUSCOs identified in the kea, Puerto Rican parrot and Scarlet macaw genomes could be due to lower genome coverage during sequencing.

Conclusion

In conclusion, this study and results provides the first genomic information in the form of a full genome sequence, assembly and comparison of the genus Agapornis. The results indicate that the lovebird genome contains an adequate number of core genes to be useful in further research such as the identification of SNPs for parentage verification, investigating the genetic basis linked to colour variation and development of tests to distinguish hybrids of different species.

Acknowledgement

We want to thank Mr Dirk van den Abeele for his help on selection of the samples. Without the help of the Avian Phylogenomics Project, in particular Prof Guojie Zhang, Prof Tom Gilbert and Prof Erich Jarvis as well as Dr Cai Li and Dr Yuan Deng from BGI, the assembly and annotation would not have been possible.

Funding sources

This work was supported by the Technology and Human Resources for Industry Programme (THRIP) which forms part of the Department of Trade and Industry of the South African government (project ID number TP13082831254) and Lumegen Laboratories.

Conflict of interest: none declared

References

1. Forshaw J.M. Parrots of the World. Princeton University Press; 2010.
2. Dilger, W.C. The Comparative Ethology of the African Parrot Genus *Agapornis*. Ethology, 1960; 17(6), 649-685.
3. Van den Abeele, D. Lovebirds Compendium. About pets. 2016.
4. Dyck, J. Structure and spectral reflectance of green and blue feathers of Rose-faced Lovebirds (*Agapornis roseicollis*). Biol Skr Dan Vid Selsk. 1971a; 18, 1-67.
5. Dyck, J. Structure and colour-production of the blue barbs of *Agapornis roseicollis* and *Cotinga maynana*. Zeitschr Zellforsch. 1971b; 115, 17-29.
6. Hudon, J. and Brush, A.H. Identification of carotenoid pigments in birds. Methods Enzymol, 1992; 213, 312-321.
7. McGraw, K.J. and Nogare, M.C. Carotenoid pigments and the selectivity of psittacofulvin-based coloration systems in parrots. Comp Biochem Physiol B.; 2004 138, 229-233.
8. Hill, G.E. and McGraw, K.J. (Editors). Bird Coloration Mechanism and Measurements Volume I. Harvard University Press. 2006.
9. Pires, S. The illegal parrot trade: A literature review. Global crime; 2012, Vol 13:3
10. Oleksyk, T.K., Pombert, J-F., Siu, D., Mazo-Vargas, A., Ramos, B., Guiblet, W., Afanador, Y., Ruiz-Rodriguez, C.T., Nickerson, M.L., Logue, D.M., 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. GigaScience; 2012, 1, 14.
11. Seabury, C.M., Dowd, S.E., Seabury, P.M., Raudsepp, T., Brightsmith, D.J., Liboriussen, P., Halley, Y., Fisher, C.A., Owens, E., Viswanathan, G., Tizard, I.R. Multi-Platform Draft de novo Genome Assembly and Comparative Analysis for the Scarlet Macaw (*Ara macao*). PLOS ONE; 2012, 8:e62415.
12. Ganapathy, G., Howard, J.T., Ward, M.J., Li, J., Li, B., Li, Y., Xiong, Y., Zhang, Y., Zhou, S., Schwartz, D.C., Schatz, M., Aboukhalil, R., Fedrigo, O., Bukovnik, L., Wang, T., Wray, G., Rasolonjatova, I.,

Winer, R., Knight, J.R., Koren, S., Warren, W.C., Zhang, G., Phillippy, A.M. & Jarvis, E.D. High-coverage sequencing and annotated assemblies of the budgerigar genome. GigaScience; 2014, 3:11.

13. Zhang, G., Li, B., Gilbert, M.T., Jarvis, E.D., Wang, J., The Avian Genome Consortium. Comparative genomics reveals insights into avian genome evolution and adaptation. Science; 2014, 346, 6215.
14. Zhang, G. Bird sequencing project takes off. Nature*,* 2015, 522, 34.
15. 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, D.W., 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. GigaScience; 2012, 1, 18.
16. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 . http://www.repeatmasker.org. 2013.
17. Flicek, P., Ahmed, I., Amode, R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard B., Pritchard, E., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken, B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J.P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., Searle, S.M.J. Ensembl 2013. Nucleic Acids Research; 2012, 1, 8.
18. The UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acid Research*. 43.
19. Warren, W.C., Hillier L.W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic, C., Bouk, N., Pruitt, K.D., Thibaud-Nissen, F., Schneider, V., Mansour, T.A., Brown, C. T., Zimin, A., Hawken, R., Abrahamsen, M., Pyrkosz, A. B., Morisson, M., Fillon, V., Vignal, A., Chow, W., Howe, K., Fulton, J.E., Miller, M.M., Lovell, P., Mello, C.V., Wirthlin, M., Mason, A.S., Kuo, R., Burt, D.W., Dodgson, J.B., Cheng, H. H. A new chicken genome assembly provides insight into Avian genome structure. G3; 2017, 7, 109-117
20. Warren, W.C., Clayton, D.F., Ellegren, H., Arnold, A.P., Hillier, L.W., Künstner, A., Searle, S.,

White, S., Vilella, A.J., Fairley, S., Heger, A., Kong, L., Ponting, C.P., Jarvis, E.D., Mello, C.V., Minx, P., Lovell, P., Velho, T.A.F., Ferris, M., Balakrishnan, C.N., Sinha, S., Blatti, C., London, S.E., Li, Y., Lin, Y-C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M. The genome of a songbird. Nature; 2010, 464, 757–762.

21. Bradnam, K.R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W-C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E.D., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M.D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B.M., Wang, J., Worley, K.C., Yin, S., Yiu, S-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S. & Korf, I.F. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience; 2013, 2, 10

22. Simão, F.A. Waterhouse. R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics; 2015, 31, 19

23. Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simao, F.A., Ioannidis, P., Seppey, M., Loetscher, A., & Kriventseva, E. V. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Research; 2015, 45