SNP discovery by resequencing the *Agapornis roseicollis* (Peach-faced lovebird) genome using Genome Analysis Toolkit

H. van der Zwan*, C. Visser2, M. Schoonen3 & R. van der Sluis4

1North-West University, Centre of Human Metabolomics, Potchefstroom, South Africa  
*Presenting and corresponding author (henriettevdz@gmail.com)  
2University of Pretoria, Department of Animal and Wildlife Sciences, Pretoria, South Africa

Summary

The Genome Analysis Toolkit (GATK) presents a method requiring relatively little bioinformatics skills for Single Nucleotide Polymorphism (SNP) discovery in non-model organisms. The recently assembled de novo reference genome of the peach-faced lovebird (an African parrot) was used to identify SNP markers for application in genetic screening tests for lovebird breeders. The parents of the reference individual were resequenced at a sequencing depth of 30x coverage on an Illumina HiSeq platform. The data generated was used to discover SNPs dispersed throughout the genome using the Genome Analysis Toolkit pipeline. Variants were discovered for the mother and father as well as from the combined genotypes of the two parents. For the mother 1 916 289 SNPs were discovered compared to 1 428 869 for the father indicating she was slightly more heterozygous than him. The number of SNPs discovered compared favourably with previously reported results in other avian genomes. These SNPs can be applied to develop genetic screening (such as parentage verification) tests for domesticated lovebirds, or in conservation strategies, for example species identification.

Keywords: GATK, bioinformatics, avian genomics, parrot breeding, parrot conservation

Introduction

*De novo* genome assembly of non-model organisms are becoming more common (Ellegren, 2014) and its application in domestic and wild parrots is not an exception (van der Zwan et al., 2017). The challenge, however, remains to discover Single Nucleotide Polymorphisms (SNPs) from these genomes for various applications in breeding (Vignal et al., 2002) and conservation (Hess et al., 2015). Seeb et al. (2011) classify two groups of SNP discovery methods namely Sanger sequencing and Next-generation sequencing (NGS). Popular Sanger sequencing methods include cloning and shotgun sequencing; expressed sequenced tags (ESTs) and Comparative Anchor Tagged Sequences (CATS) (Garvin et al., 2010; Seeb et al., 2011). Next-generation sequencing (NGS) methods include Restriction site Associated DNA sequencing (RADseq); various density SNP chips and whole genome resequencing (Seeb et al., 2011; Fleming et al., 2015).

Constraints to consider when choosing a SNP discovery method are the number of SNPs that will be needed for the study, cost involved as well as the level of bioinformatics experience required (Seeb et al., 2011). Sanger sequencing yields fewer SNPs per run compared to NGS data (e.g. Olsen et al., 2011, vs. Bowers et al., 2016) making NGS a better option when many SNPs are required. With Sanger sequencing no prior knowledge of the targeted area is known, resulting in the possibility of no SNPs being targeted. Excessive costs may be incurred in the form of primers amplifying areas without SNPs (Seeb et al., 2011) whereas NGS might be more expensive but covers the whole genome and thereby guarantees SNP discovery. NGS yield vast amounts of data and advanced bioinformatics skills may be
required (Seeb et al., 2011) to analyse it. However, The Genome Analysis Toolkit (GATK) is a structured command line interface programming framework designed to identify variants from NGS data (McKenna et al., 2010) and was developed with tools that are easily executed, making it accessible to individuals with moderate bioinformatics skills.

The parrot genus *Agapornis* (or lovebirds) consists of nine parrot species and are widely kept as pets (Forshaw, 1989; van den Abeele, 2016) but are also found in their natural habitat distributed over Africa and Madagascar (Forshaw, 1989). The de novo genome of *A. roseicollis* was sequenced, assembled and annotated previously (van der Zwan et al., 2017). In this paper we report the bioinformatics pipeline followed to discover SNPs in the resequenced parent genomes by applying the guidelines as set out by GATK. These SNPs can be used to develop genetic screening tests, e.g. individual or parentage verification panel or species identification tests. Since 55% of the 356 recognized parrot species (Forshaw, 1989; Pain et al., 2006) are affected by illegal trade and parrot breeding is globally a popular aviculture practice (van den Abeele, 2016), the application can have far-reaching outcomes.

**Materials and Methods**

**Sample selection and sequencing**

The sequencing, assembly and annotation of the de novo *A. roseicollis* genome is described in van der Zwan et al. (2017). After ethical approval (Ethics number NWU00348-15-S5) was obtained, blood from both the parents of the reference bird was collected. Genomic DNA isolation as well as sequencing library construction and sequencing is described in van der Zwan et al. (2017). Two shotgun libraries were constructed for each parent consisting of fragment sizes of 300 and 550 kb respectively and sequenced at a depth of 30x coverage for both birds.

**SNP discovery**

All command line arguments are presented in Supplementary File 1. The reads from the two parent birds were aligned against the reference genome of the offspring using the Burrow-Wheeler Aligner (BWA-MEM) algorithm (Li, 2013). During sequencing, the same fragment could be sequenced more than once causing duplicate reads. The duplicate reads shouldn’t be analysed as different reads and therefore the program Picard (v 2.10.9) (http://broadinstitute.github.io/picard) marked all duplicates by identifying the 5’ end of all read pairs and matched different pairs at the same orientation. The best read pairs were flagged and not removed so that GATK could interpret the flags as duplicates of the same fragment.

Variant discovery was performed using The Genome Analysis Toolkit (GATK) v 3.8-0 (McKenna et al., 2010) in three steps as shown in Table 1. The first step was to discover raw variants which include SNPs, Indels (insertions or deletions), mixed variants (combinations of SNPs and indels at a single position), MNP (multi-nucleotide polymorphisms) and symbolic variants. The second step involved combining the genotypes of the two birds to create one genotype file (the “Combined genotype file”).

<table>
<thead>
<tr>
<th>Step</th>
<th>GATK Module</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call variants</td>
<td>HaplotypeCaller</td>
<td>All raw variants per sample</td>
</tr>
<tr>
<td>Join genotypes</td>
<td>HaplotypeCaller</td>
<td>Combined genotype</td>
</tr>
<tr>
<td>Filter variants</td>
<td>SelectVariants and VariantsFiltration</td>
<td>Filtered variants</td>
</tr>
</tbody>
</table>

The third step of variant discovery was to filter the discovered raw variants to include
only SNPs. Since the lovebird is a non-model organism, hard filtering had to be applied to the raw variants dataset. Each raw variant was allocated parameter values and by applying filters to these values, variants not complying with specific quality metrics could be excluded from the call set. All indels were excluded and thereafter all SNPs adhering to the specific parameter values were included in the final SNP set. The parameters and values used during hard filtering as well as a definition of each parameter (de Summa et al., 2017) are given in Supplementary File 2. The raw filter parameters were applied to both the separate files of the mother and father as well as the combined genotype file.

Results & Discussion

The number of raw variants, SNPs and heterozygote SNPs that were discovered for the mother, father and the combined files as well as discovered variants and SNPs from previous avian studies, are shown in Table 2.

Table 2: Variants and SNPs discovered for the Mother, Father and combined genotype and comparisons with variants discovered in other avian genomes.

<table>
<thead>
<tr>
<th></th>
<th>Mother</th>
<th>Father</th>
<th>Combined</th>
<th>Ground tit</th>
<th>Scarlet macaw</th>
<th>Chicken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw variants</td>
<td>2 156 950</td>
<td>1 601 584</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>4 462 467</td>
</tr>
<tr>
<td>SNPs only</td>
<td>1 916 289</td>
<td>1 428 869</td>
<td>1 667 639</td>
<td>1 723 688</td>
<td>951 507</td>
<td>N/A</td>
</tr>
</tbody>
</table>

For the mother, 26% more raw variants and SNPs were discovered compared to what was discovered for the father, indicating that she was slightly more heterozygous than him. The combined file included 87% of the number of SNPs that was discovered for the mother. This implies that many of the SNPs discovered were shared between the mother and father, indicating that these SNPs will be found in the greater lovebird population since the mother and father were unrelated. In the study by Flemming et al. (2015) where Leghorn and Fayoumi chicken lines were resequenced and SNPs detected using GATK from the reference genome (Table 2), 3 223 583 of the raw variants that were discovered were novel SNPs, not previously discovered in the chicken genome. However, 16 individuals per line were analysed which could influence the number of variants discovered. The SNPs discovered in the current study compared well to the number of SNPs discovered from other avian genomes such as the ground tit (Pseudopodoces humilis) (Cai et al., 2013) and scarlet macaw (Ara macao) (Seabury et al., 2013) (Table 2). It should however be noted that these SNPs were discovered directly from the reference genome and not by resequencing.

The discovery of SNPs for application in animal breeding in non-model domestic organisms has been challenging (Olsen et al., 2011; Ellegren 2014) as even with a de novo reference genome there are no SNP databases available such as in the case of cattle (Ogden, 2011). However, GATK allows for SNP discovery from non-model organism genomes and small sample sizes (de Summa et al., 2017) with relatively limited bioinformatic skills. The use of the parents of the reference genome individual allows the user to validate that a discovered variant adhere to Mendelian laws. This is useful in the development of genetic screening tests such as parentage verification since each variant is verified to be an inherited variant. One challenge is that the reference genome only has one allele at each position, complicating verification.

SNP discovery from a small population pool could lead to ascertainment bias which is defined as the systematic deviation from the expected allele frequency distribution that occurs due to sampling of only a few individuals (Garvin et al., 2010; Seeb et al., 2011).
Therefore, the set of SNPs discovered in this study should only serve as a foundation from which to test more individuals.

This method is, to our knowledge, the first attempt to discover SNPs by resequencing the reference genome individual’s parents of any species. In this study the recently assembled de novo genome of the peach-faced lovebird (A. roseicollis) (van der Zwan et al., 2017) was used to discover SNPs by resequencing the parents of the reference individual’s genomes. It can be a valuable approach to follow, especially discovering SNPs for domesticated non-model species.

**Conclusion**

The application of the set of discovered SNPs in wild and domestic parrot conservation and breeding can have a major effect globally. In parrot breeding the development of a parentage verification panel could improve overall record keeping and breeding strategies. The SNPs could also be used in the development of a species identification tests which can improve parrot conservation attempts.

**References**


Cai, Q., Qian, X., Lang, Y. 2013. Genome sequence of ground tit *Pseudopodoces humilis* and its adaptation to high altitude. Genome Biol. 14: R29


Supplementary File 1

Command line arguments used.

**BWA: Align reads**
bwa mem -M -t 16 ref.fa read1.fq read2.fq > aln.sam

**Picard: Mark duplicates**
java -jar picard.jar MarkDuplicatesWithMateCigar \
I=input.bam \nO=mark_dups_w_mate_cig.bam \nM=mark_dups_w_mate_cig_metrics.txt

**GATK: Call variants**
java -jar GenomeAnalysisTK.jar \n-T HaplotypeCaller \n-R reference.fa \n-I preprocessed_reads.bam \n--genotyping_mode DISCOVERY \n-stand_emit_conf 10 \n-stand_call_conf 30 \n-o raw_variants.vcf

**GATK: Join genotypes**
java -jar GenomeAnalysisTK.jar \n-T GenotypeGVCFs \n-R abc.fasta \n-V sample1.g.vcf \n-V sample2.g.vcf \n-V sampleN.g.vcf \n-o output.vcf

**GATK: Filter variants**

*Extract the SNP from the call set*
java -jar GenomeAnalysisTK.jar \n-T SelectVariants \n-R reference.fa \n-V raw_variants.vcf \n-selectType SNP \n-o raw_snps.vcf

*Extract indels from the call set*
java -jar GenomeAnalysisTK.jar \n-T SelectVariants \n-R reference.fa \n-V raw_HC_variants.vcf \n-selectType INDEL \n-o raw_indels.vcf
Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.
### Supplementary File 2
### Parameters and values used during hard filtering of SNPs and indels.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value used</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality (QUAL)</td>
<td>N/A</td>
<td>A Phred-based probability of a false positive variant.</td>
</tr>
<tr>
<td>Quality by Depth (QD)</td>
<td>&gt;2.0</td>
<td>The quality score normalized by the coverage of that area available. In areas of high coverage variants will have a higher QUAL score which is then compensated for by the QD score.</td>
</tr>
<tr>
<td>Fisher Strand (FS)</td>
<td>&lt;10.0</td>
<td>A parameter used to correct for sequencing bias using the Fisher’s exact test. This parameter evaluates if one allele (reference or alternate) is favoured above another and the output is a Phred-scaled p-value.</td>
</tr>
<tr>
<td>Root Mean Square (MQ)</td>
<td>&gt;50.0</td>
<td>Indicates the quality of all reads at a locus.</td>
</tr>
<tr>
<td>Mapping Quality Rank Sum Test (MQRankSum)</td>
<td>&gt;-5.0</td>
<td>Indicate the difference in the mapping qualities of the reference and alternate alleles.</td>
</tr>
<tr>
<td>Read position rank sum test (ReadPosRankSum)</td>
<td>&lt;-8.0</td>
<td>Indicates if there is bias in the position of the reference and alternate alleles.</td>
</tr>
</tbody>
</table>