De novo Assembly of the Chinese Bamboo-Partridge Genome and Comparison with Chicken

Gabriel Rognon, Diane Esquerre, Sylvain Foissac, Alain Vignal, Jean-Luc Coville, Bertrand Bed'Hom, Michèle Tixier-Boichard, Thomas Faraut

De novo assembly of the Chinese bamboo-partridge genome and comparison with chicken

Gabriel Rognon[1], Diane Esquerré[1], Sylvain Foissac[1], Alain Vignal[1], Jean-Luc Coville[2], Bertrand Bed'hom[2], Michèle Tixier-Boichard[2] and **Thomas Faraut**[1]

[1]GenPhySE, INRA Toulouse, [2]Gabi, INRA Jouy-en-Josas, France. Thomas.Faraut@toulouse.inra.fr

Context

History of chicken domestication at the genome level is not completely resolved, in particular the possible contribution of Gallus species different from Red junglefowl Gallus gallus. In the context of a larger project (Domestichick), which addresses the question of chicken domestication using molecular approaches, we undertook the sequencing and de novo assembly of the Chinese bamboo-partridge genome (Bambusicola thoracica), sister taxon of four Gallus species. The Bambusicola genome will is used, in this project, as an outgroup for the genetic diversity analysis performed on 4 wild gallus species. In order to pinpoint potential genome rearrangements between the Bambusicola genome and the Gallus gallus reference genome sequences, a *de novo* genome assembly was performed.

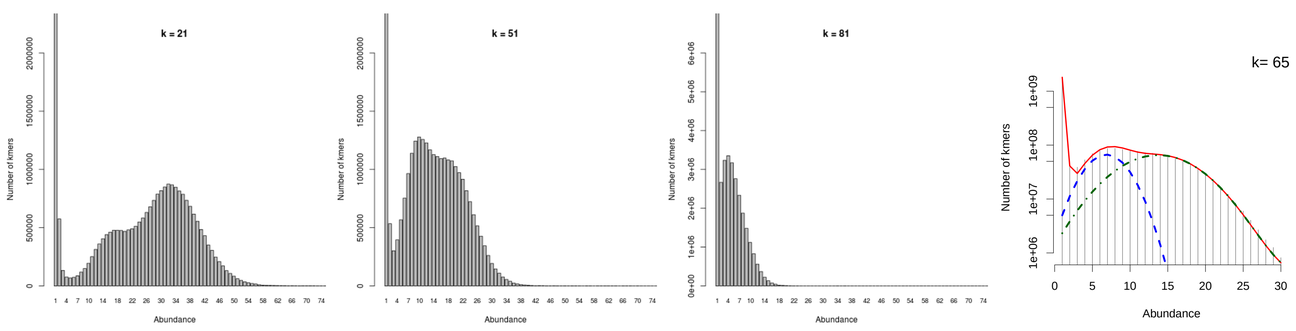
Project Overview

We used three different libraries (paired-end, 3 kb and 6 kb mate-pairs), sequenced each at 50X coverage. Taking advantage of the close phylogenetic relationship between Bambusicola and the chicken, confirmed by the ability to align more than 95% of the reads against the chicken genome with 95% identity, the reads were first partitioned according to chicken chromosomes based on sequence similarities after mapping using the ngm mapper [1]. The corresponding sequences were then *de novo* assembled within each chromosome. Following teh recomandation made by the assemblathon competitions we tested 5 different assembly softwares : masurca [2], minia [3], platanus [4], soap [5] and velvet [6].

Methods

k-mer size selection

The selection of the best k value for the k-mers is an important feature of any de Bruijn graph-based assembly approach. The following graphs shows the influence of the k value on the distribution of k-mers. The last figure on the right shows the decomposition of the k-mer dstribution into homozygote k-mers (blue line), homozygote k-mers (green line) and erroneus k-mers (left pic). The value of $k = 65$ was selected for every chromosome.



Assembly software comparison

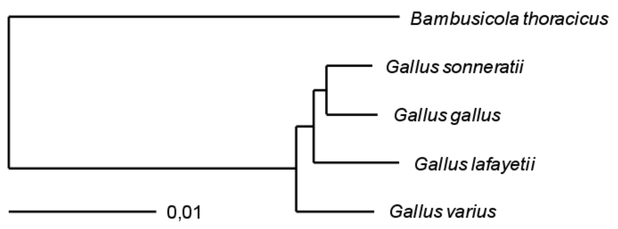
Assembly software were first compared using only pair-end sequences for a single chromosome. With our data set, the masurca software showed a remarkable improvment (10-fold increase of the contig N50 value). The resulting assemblies were analysed using the REAPR pipeline [7]. Results are shown for chromosome10 (gga10).

	masurca(pe)	masurca(se)	minia	platanus	soap	velvet
contigs (> 500bp)	1 779	10 831	9 330	8 959	11 189	10 542
total size (Mb)	20,06	19,66	17,02	17,47	15,67	14,44
largest contig (kb)	140,8	59,6	35,14	37,56	35,14	38,68
mean (kb)	8363	1518	766	437	460	413
N50 (kb)	32,03	2,46	2,21	2,62	1,66	1,58
L50	189	2022	2083	1807	2499	2227
misassemblies	11	5	6	5	10	3
(Re)aligned reads						
pe re-aligned (%)	99,13	96,88	96,45	98,83	99,17	99
pe re-aligned concordant (%)	96,74	82.63	73.47	78.35	64.13	59.48
mp aligned (%)	98.1	96.4	96.21	96.68	99.04	98.99
mp aligned concordant (%)	73.23	12.47	3.63	10.93	5.73	6.29
Alignment with the chicken genome						
non-aligned contigs	1250	6758	5690	5202	6041	5319
genome fraction (%)	21,2	35,8	36,1	35,2	34,8	33,2
longest alignment (kb)	115,13	30,38	31,8	27,84	30,37	34
REAPR diagnostics						
error free bases (%)	80,6	63,9	55,7	52	50	48,11

Perspectives

- First comparison with the chicken genome indicate very few chromosomal rearrangements
- The undergoing work aims at analysing the candiate *evloutionnary disrupted* scaffolds in the ligth of the associated assembly diagnostics to be able to distinguish chimeric scaffolds from true *evloutionnary disrupted* scaffolds.

Bambusicola thoracicus



Phylogeny of the genus Gallus and of Bambusicola.

Data

Each library was sequenced at 47X sequence coverage

Pair-end	476,967,832 seq
Mate-Pair 3kb	490,807,312 seq
Mate-Pair 6kb	453,480,038 seq

Final assembly

The resulting assembly totalizes 920 Mb of sequences assembled into scaffolds with a N50 scaffold length of 2.5 Mb. The resulting scaffolds were aligned to the chicken chromosomes. The extent of conservation between the newly assembled Bambusicola genome and the chicken genome are beginning to be analysed.

References

[1] Sedlazeck F, Rescheneder P and von Haeseler A (2013) NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 29:2790-2791.

[2] Zimin *et al*. (2013) The MaSuRCA genome assembler *Bioinformatics*

[3] Chikhi R and R G (2013) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for molecular biology* 8:22

[4] Kajitani *et al*. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research* 24:1384-95

[5] Luo *et al*. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18

[6] Zerbino D and Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* 18:821-9

[7] Hunt *et al*. (2013) REAPR: a universal tool for genome assembly evaluation. *Genome biology* 14:R47