Global biodiversity genomics and birds: a general overview

The human population has grown dramatically over the past centuries and decades. Globally, food and space are becoming increasingly scarce resources. Without going into the depth of this multi-faceted problem, this article highlights an aspect of why fundamental biodiversity research has the potential to contribute to easing this problem.

by Robert H. S. Kraus,
Department of Migration,
Max Planck Institute
of Animal Behavior,
78315 Radolfzell, Germany.
rkraus@ab.mpg.de

Due to an increasing world population more food needs to be produced in less space. A simple consequence is a necessary increase in efficiency. For yield to increase, agriculture has tried standardisation via monoculture but suffers the risks of a low diversity, high density strategy.

For example, clonal production of bananas was highly efficient for a variety called Gros Michel until the 1960s. Then, a fungal disease (Panama disease) wiped it out nearly completely.

Today, the Cavendish banana is the dominant clone. But history repeats itself and today's banana production is immensely threatened by a newly emerging fungal disease called Black Sigatoka.

Breeding for diversity

The generally accepted opinion among plant and animal breeders is that breeding for diversity, incorporating knowledge and resources from seed banks, fancy breeds and wild ancestors, is going to be a major factor.

A high diversity in global breeding stocks is thought to increase not only the availability of locally adapted breeds with increased yield under local conditions but also higher resilience against change and thereby increased sustainability.

Accompanying the human genome project Kruglyak and Nickerson (2001) wrote a seminal roundup on genetic variation: 'Variation is the spice of life'. In it, they highlight the importance of understanding every single SNP in the human genome and what could be achieved by describing genetic variation down to the level of the nucleotide.

This variation stems from the process of mutation and is maintained or acted against by selection.

The result is organismal evolution leading to the diversity of form and function. The term biodiversity is a modern fusion of biology and diversity and comes from the nature conservation community.

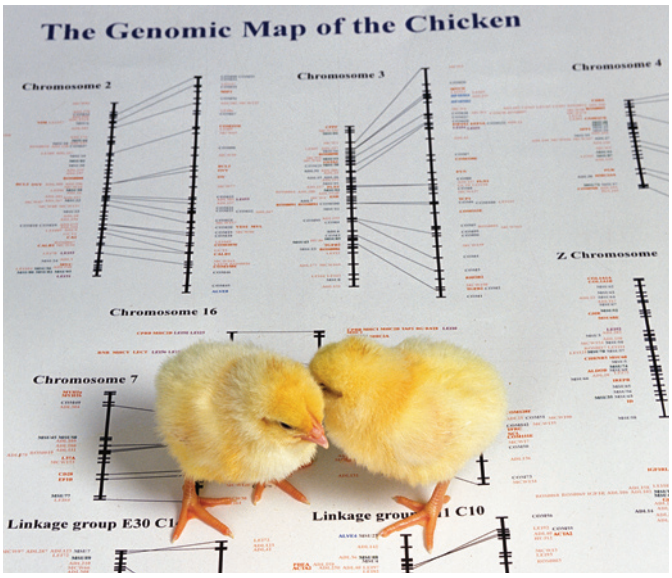
An early promoter of the term and its ideas and context to a broader public was E. O. Wilson with his book 'The diversity of life'. Defining biodiversity and what the concept entails is a broad field and there are whole books attempting to do justice to it – but here I refer to it in a rather technical way and in the context of genetic variation, particularly on the genomic level.

Biodiversity genomics

The throughput of genome sequencing has surpassed Moore's law of electronics and peaked in mostly Illumina short read sequencing becoming the driver of genome sequencing. There has been a profound impact on avian genomics as birds are among the most charismatic and important groups of animals.

One of the first whole genomes to be sequenced among vertebrates was the chicken, by traditional, expensive Sanger-based methods. But the sequence revolution with genomes of turkey and duck sounded the bell for multi-species genome projects.

A team of scientists formed the avian phylogenomics consortium because they argued that many small projects would not create a consistent framework for comparative genomics quickly and coherently enough. Not only did they stress the importance of a phylogenetic framework, they also assembled a consortium of scientists



to jointly work on 'the big bang of avian genomics'. This culminated in the publication of more than 50 articles based on 48 whole genomes of birds with good phylogenetic spread (roughly one species per bird order) at the end of 2014.

The seminal papers about phylogeny and comparative genomics nicely illustrated how such collaborative initiatives can open new trajectories of comparative genomic research.

The strength of comparative genomics with a solid phylogenetic backbone were compelling and the avian phylogenomics project constituted itself as the bird 10,000 (B10K) community of research with a mission to sequence all bird species.

B10K and its phases

The B10 project is a community of more than 100 scientists. Its goal is the generation of representative draft genome sequences for all bird species. It is divided into four phases according to the taxonomic categories order, family, genus, and species.

Phase 1, the order level phase, was the big bang in avian genomics in 2014. In Phase 2, the family phase, ~240 families of birds are covered with at least one genome per family, and Phase 3 and Phase 4 will require the sequencing of ~2250 genera and a remaining ~8000 species, respectively.

Currently, B10K has finalised Phase 2 and has now covered nearly all families of the avian tree of life (218 of 236, 92%) with 363 draft genomes.

Taxonomic coverage has been a priority from the start. Of the 363 genomes in the data set, 268 genomes have not previously been published, representing 155 avian families for the first time.

The raw data for assembly typically stem from standard Illumina pipelines making use of paired-end and in many (but not all) cases mate-pair libraries. Gene completeness based on BUSCO scores is high with an average of 95%.

A few genomes are also based on third generation sequencing or benefitted from using long-range scaffolding strategies such as Dovetail or 10X, for the majority, assembly statistics are similar to those in B10K Phase 1.

A challenge in this data set was not only to produce whole genome alignments, for which new, reference-free methods were developed on the basis of the Cactus aligner, but also to establish orthology.

Continued on page 27

Continued from page 25

Due to gene duplications during further divergences between bird families one-to-many and many-to-many orthology relationships needed to be resolved with a strategy that goes beyond the regular reciprocal best hit method (sequence similarity).

Taking into account positional information from whole genome alignments (synteny) a total of 16,179 high-quality orthologs across all species could be identified, that is an increase of 8% over the reciprocal best hit method.

Genome data

The data of the B10K family phase along with details on whole genome alignment methods, orthology detection, bird family-by-family pan-genome evolutionary rate and conservation statistics will be published soon. Upon publication, all genome data will become available publicly, comprising 18.4 trillion base pairs (bp) of raw and 285 billion bp of assembled data.

Analogously to the big bang in avian genomics in 2014, a wave of publications describing analyses of these data by B10K members was set for 2020.

Evolutionary innovations in the biological class of birds are analysed with comprehensive, expert-curated databases of morphological, physiological, ecological and behavioural traits for linking phenotypes to genes.

Population genomic models are explored to reconstruct demographic history of species and clades to understand their adaption to previous environmental changes to better predict future threats to ecosystem diversity.

Duck immunogenetics

Except from increasing statistical power by looking at more markers, reference genomes also enable the study of functional variation, a field that has been restricted to only well-developed model systems until most recently. Functional variation responsible for dealing with, for example, environmental toxins have been proposed in conservation frameworks. Studies of functional regions have led to insights into the genomic basis of susceptibility to disease.

New ecology and evolution models arose during recent years due to advances in sequencing technology. One example is waterfowl. Ducks and other waterfowl have long become a model system to study diseases, especially those that can infect humans (zoonotic diseases), such as avian influenza. Therefore, understanding its immune system is of great importance.

My group has started to investigate immunogenetics of waterfowl species on genomic and transcriptomic scales.

In an experimental approach, we examined the molecular, physiological and behavioural responses of wild type mallards during an immune challenge in captivity. We triggered the immune response with non-infectious immunostimulants (LPS, poly I:C, and heat-killed S. aureus) and monitored it with state-of-the-art bio-loggers and high throughput sequencing technologies.

By tracking body temperature, heart-rate and whole-body acceleration (as a measure of activity) as well as white blood cell counts, we confirmed a phenotypic immune reaction. The core of the experiment was a time series analysis of the whole transcriptome cellular response in whole blood. We took blood samples during the acute phase response at 0h post stimulation (ps), 3h, 6h and 24h ps.

Across time points and control and treatment groups we analysed differential gene expression of 120 samples against the published duck reference genome. Because of its availability and good annotation, we were able to not only count and list differentially expressed genes but also do GO and pathways analysis (such as KEGG).

This facilitates moving away from a gene-centric view of gene regulation to a pathway-centric view, where no gene acts in isolation and the knowledge of pathway regulation can be studied in depth.

We could show that the immunostimulants tested in this study were not only able to provoke an acute phase response of physiology and behaviour but were accompanied by the appropriate cellular and genetic responses, too. In a follow-up study we used these immunostimulants on wild caught mallards and released them to the wild to follow their natural disease behaviour with bio-loggers.

Jax et al. found that indeed sick mallards move less than healthy ones. This will be instrumental in detecting the arrival of, for example, avian flu waves by monitoring wild mallards with bio-loggers.

In another study, Jax et al. (2018a) used the experience from the pathway-centric view to explore evolutionary patterns in immune genes across pathways. While previous studies typically examined evolutionary patterns gene by gene, or gene family by gene family, and focused mostly on MHC, we used hybrid bait capture on a global sample of mallards (Sweden, Spain, Canada, Greenland) and five other duck species, and bioinformatically mined previously published goose genomes to obtain a set of 120 innate immune gene sequences across waterfowl species.

These genes were chosen to cover every single gene across some prominent immune pathways. Our aim was to examine patterns of genetic variation and evolution in immune genes involved in detection, signalling and response to pathogens to test different evolutionary patterns depending on network topology of the respective pathways.

Patterns of natural selection were calculated as traditional summary statistics on whole genes (dN/dS) but also for codon-specific signatures of selection. Most immune genes were under purifying selection but several genes indeed showed signatures of positive selection at specific codons.

This suggests that certain parts or motifs of the immune genes are evolutionarily constrained, possibly due to their role in signalling networks, while other parts are more likely to adapt, possibly to avoid genetically drifted or novel pathogens.

Further, detector molecules were overall more conserved than other immune genes, but specific positions in domains involved in pathogen recognition can be under strong positive selection. This suggests that the evolution of innate immune receptors is driven by host parasite co-evolution – very similar to MHC.

In these projects, we learned that pulling DNA sequences of non-model species from databases can be cumbersome and frustrating. In scientific papers genes are typically named by their gene symbol and this is often ambiguous. In the B10K framework 363 species will be analysed for the totality of their immune genes (the 'immunome') by Mueller et al.

Problems of ambiguity and identification of orthology or paralogy must be tackled by an overarching 'one-stop resource' combining well curated databases and internal B10K annotations. The identification of immune genes is based on gene ontology and similar databases.

With this information, the major public sequence repositories Ensembl and UniProt were mined and a comprehensive database of the avian immunome was constructed with subsequent input of all annotated 363 B10K family phase genomes. This database 'Immunome DB: a one-stop resource for avian immunogenomics' is built in an open-source relational SQL environment.

Conclusions

One could argue that the new era of high throughput sequencing leads to an end of the model organism. In some sense this is true, because many researchers already have reference quality genome assemblies for their particular bird species.

However, while more species have advanced in terms of genomic resources, the chicken genome and other traditional models have been improved even further, with considerable functional annotation by experimental evidence. More reference genomes lead to more model species for answering more specific questions, but in the end model species have become what they are because of, for example ease of access, short generation time, easy breeding and handling, and nowadays also relatively easy access to genetic engineering methods.

Co-operation of emerging models of biodiversity genomic projects such as B10K with traditional models is one very promising outcome of the sequencing revolution.

Comparative genomics can answer questions that are impossible to answer with just a single genome. The evolutionary history of every nucleotide in a genome can only be understood when looking at it in a multi-species framework. Eventually, this understanding is going to be instrumental to understand genotype-phenotype interactions and accurately predict marker-assisted breeding programmes or genetic engineering targets. ■

References are available from the author on request