2020 Vertebrate Genomes Project (VGP) Workshop

January 15, 2020
Handlery Hotel San Diego,
950 Hotel Cir N, San Diego, CA 92108, USA
Garden Ballroom, 10:00 AM - 5:30 PM
@2020 Plant and Animal Genomes conference
https://www.intlpag.org/2020/

Contact: contact@genomeark.org
Twitter: @genomeark

Workshop Organizers:
Giulio Formenti, Ph.D., Rockefeller University, NY: gformenti@rockefeller.edu
Sadye Paez, Ph.D., MSPT, MPH, Rockefeller University, NY: spaez@rockefeller.edu
Constantina Theofanopoulou, Ph.D. Rockefeller University, NY: ktheofanop@rockefeller.edu
Erich D. Jarvis, Ph.D., G10K Chair, Rockefeller University, NY: ejarvis@rockefeller.edu

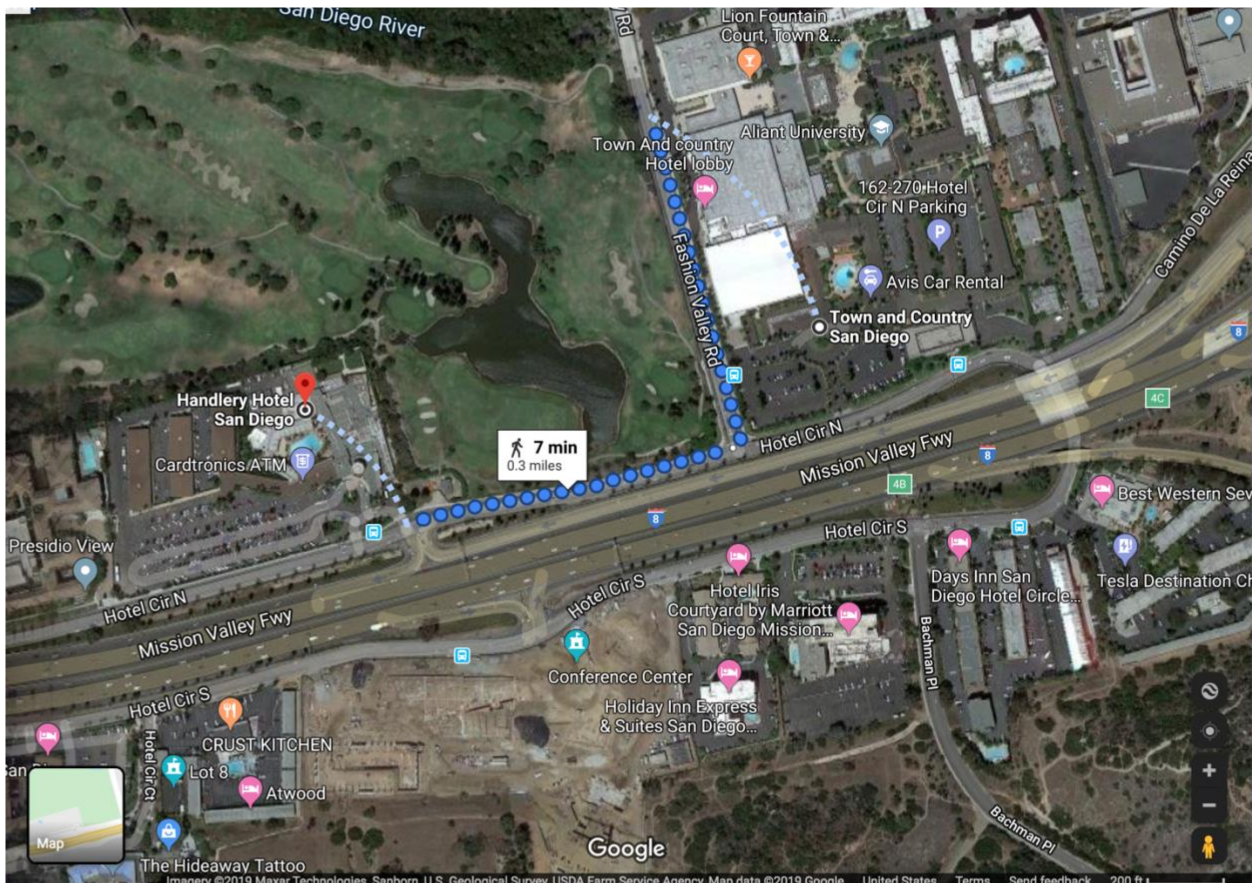
VGP Goals: The goal of the Vertebrate Genomes Project (VGP) is to generate high-quality, error-free, gapless, chromosome-level, haplotype phased, and annotated reference genome assemblies of at least one individual each representing all ~70,000 extant vertebrate species and to use those genomes to address fundamental questions in biology, disease, and conservation. We define a high-quality reference genome assembly as one that meets a minimum metric of: N50 contig $\geq$ 1 Mb; N50 scaffold $\geq$ 10 Mb; 90% of the genome assembled into chromosomes validated by at least 2 methods; an average QV40 or higher base call quality (called a 3.4.2.QV40 phased genome assembly). We are conducting the VGP in taxonomic phases from orders (Phase 1), families (Phase 2), and genera (Phase 3), to eventually all species (Phase 4). For Phase 1, the primary species selection criterion are species representing shared divergence times before and soon after the last mass extinction event 66 MYA, which results in over 260 lineages. Phase 1 is being supported in part by crowdfunding among scientist as well as institutional collaborations with other large genome efforts among participating scientists (Rockefeller University; HHMI; Wellcome Sanger 25G, UK Darwin Tree of Life; Max Planck Institute; Earth Biogenome Project to name a few). The G10K negotiated significantly discounted costs. We have an open-door policy for scientist to join the VGP, benefit from the discounts, and conduct science on the shared Genome Ark Library (https://vgp.github.io/) of high-quality phased reference genome assemblies. The successful outcome of Phase 1 will be leveraged to raise the necessary funds to sequence the genomes of all ~1K vertebrate families, then all ~10K genera, and finally all ~70K species.

Mission statement for the Jan 2020 reference VGP workshop: From 2019-2020, we have been conducting analyses of some of the Phase 1 ordinal genomes, preparing publications, and finding means to complete the remaining ~80 species for Phase 1. The mission of the 2020 VGP workshop is to further advance these missions, and looking forward to Phase 2, by focusing on remaining challenges in Arc of the pipeline that goes from: 1) sample procurement and funding; 2) sequencing and assembly; 3) annotation and analyses; 4) publications of methods and biological findings; and 5) preparing for Phase 2.

Preparation: More historical background can be found in a 3-page G10K-VGP workshop document and watching the G10K-VGP 2018 year-end presentation: https://hhmi.zoom.us/recording/share/ebHgvjuUBfurGETQhhF5Idu0S9B3312WXzc-Apaak8-wIumekTziMw Access password: G10K. With these documents and presentation, please respect standard rules of scientific ethical conduct for credit and the G10K-VGP Embargo Data Use Policy: https://genome10k.soe.ucsc.edu/data-use-policies/

Messages from the PAG organizers: Please be advised that your workshop at PAG XXVIII is scheduled at the Handlery Hotel which is a 2-min shuttle bus ride or 8-min walk. We will have a dedicated shuttle bus running between the 2 hotels ALL day. Coffee breaks will also take place at the Handlery however; lunch will only take place at the Town & Country Hotel. Here is the link to the web version of the mobile app for conference abstracts: https://plan.core-apps.com/pag_2020

Below is a google map image of the location of Handlery Hotel where VGP meeting will occur, relative to the Town and Country conference center for the PAG meeting, in San Diego. Give yourself a minimum of 10-min of walk time, including finding rooms.



Workshop Agenda

Introduction (will start promptly at 10:00AM)

10:00-10:20 **Where we have been and where we are going**
Erich Jarvis, G10K Chair, Rockefeller University, NY, USA

Workshop Session 1: Sample procurement, preparation, and sequencing (10:20-11:10)

Chair: Erich Jarvis, Rockefeller University, NY, USA

10:20–10:40 **Benchmarking ultra-high molecular weight DNA and tissue preservation protocols for the Vertebrate Genomes Project**
Jennifer Balacco, Vertebrate Genomes Lab, Rockefeller University, NY, USA

10:40–11:00 **Accuracy vs. length: comparing PacBio HiFi and Oxford Nanopore ultra-long sequencing for human genome assembly**
Adam Phillippy, NIH, Bethesda, MD, USA

11:00–11:10 **Discussion**

11:10-11:30 **Coffee/Tea Break**

Workshop session 2: Assembly, Annotation, and Evaluation (11:30-14:20)

Chair: Giulio Formenti, Rockefeller University, USA

11:30–11:50 **Scalable and cross-platform compatible genome assembly pipeline**
Arkarachai Fungtammasan, DNAnexus, San Francisco, CA

11:50-12:10 **Reference-free evaluation of assembly quality and improved trio binning**
Arang Rhie, NIH, Bethesda, MD, USA

12:10-12:30 **The little shop of assembly curation horrors**
Kerstin Howe, Wellcome Sanger Institute, UK

12:30-13:30 **Lunch, served at the workshop venue**

13:30-13:50 **Quantifying improvements and needed improvements of VGP assemblies**
Chul Lee, Seoul National University, Seoul, South Korea

13:50-14:10 **TOGA: Tool to infer orthologs from genome alignments**
Michael Hiller, Max Planck Institute, Dresden, Germany

14:10–14:20 **Discussion**

Workshop session 3: Biological discoveries
(14:20-15:10)
Chair: Sadye Paez, Rockefeller University, USA

14:20-14:40 **Genomes and population size: Historical demography of the vaquita**
Phillip Morin, Southwest Fisheries Science Center, NOAA, La Jolla, CA

14:40-15:00 **Genomic conservation of the critically endangered great Indian bustard**
Dushyant Singh Baghel, Nucleome Informatics, India

15:00-15:10 **Discussion**

15:10-15:30 **Coffee/Tea Break**

Workshop session 4: Future directions
(15:30-17:30)
Chair: Erich Jarvis, Rockefeller University, NY, USA

15:30-15:50 **Initial genome assemblies from the Darwin Tree of Life Project**
Shane McCarthy, Wellcome Sanger Institute & Cambridge University, UK

15:50-16:10 **The human pangenome project**
Benedict Paten, UCSC, Santa Cruz, CA, USA

16:10-16:30 **VGP-AI (Artificial Intelligence): a proposal**
Giulio Formenti, Rockefeller University

16:30-16:50 **Expanding the VGP network**
Michaela West, UMR GDEC 1095 INRA, France
Claudio Ciofi, University of Florence, Italy

16:50-17:30 **Discussion: Completing VGP Phase 1 and preparing for Phase 2**
Entire group

17:30 **Close of G10K-VGP 2019 workshop**

Issues that need resolution to consider for discussion at the workshop:

- Adding metric for haplotype phasing and other metrics
- Need an approach for complete phasing of haplotypes, at all steps
- Missing genomic sequence from VGP assemblies
- Sex chromosomes
- Scaling up production of genome assemblies, from sample collection to annotation
- Improvements to multilayered and reference free alignment approaches
- Initial studies with VGP Phase 1 data release genomes
- Planning VGP Phase 2

ABSTRACTS

Benchmarking ultra-High molecular weight DNA and tissue preservation protocols for the Vertebrate Genomes Project

Jennifer Balacco, Vertebrate Genomes Laboratory, Rockefeller University, NY, USA

Long read technologies require high quality, ultra-High Molecular Weight (uHMW) DNA (>5 µg and >150Kb). The Vertebrate Genomes Project (VGP) is an international and multidisciplinary project of the Genome10K (G10K) consortium, which has selected long-range genomic sequencing technologies that require large amounts of uHMW DNA. The current gold standard tissue preservation method for uHMW DNA is flash freezing with liquid nitrogen and storing at -80°C, but this is not always feasible for field collection. We present a comparative study of preservation methods for different environmental conditions and various sample types (muscle, soft tissue, and blood) across vertebrate groups (Mammals, Birds, Fish and Amphibians). This presentation includes current preliminary findings from this study, as well as a look towards future and developing methods in uHMW DNA extractions. The evidence gathered through this benchmarking investigation allows us to present much needed guidelines for field biologists.

Accuracy vs. length: comparing PacBio HiFi and Oxford Nanopore ultra-long sequencing for human genome assembly

Adam Phiillippy, NIH Bethesda, MD, USA

With a focus on resolving haplotypes and large segmental duplications, we have benchmarked both PacBio HiFi and Oxford Nanopore ultra-long sequencing of the human genome. PacBio HiFi reads exceed 99.9% accuracy per read, with average read lengths of 10–15 kbp, while Nanopore ultra-long reads can reach in excess of 100 kbp but with less than 95% accuracy. I will report on our experience with these two datatypes and their relative strengths and weaknesses for vertebrate genome assembly.

Scalable and cross-platform compatible genome assembly pipeline

Arkarachai Fungtammasan, DNAnexus, San Francisco, CA, USA

Scalability and reproducibility are the heart of a reliable production pipeline. For the past years, DNAnexus serves as one of the main production platforms for the VGP genome assembly. However, the challenge is the effort into making pipeline runnable on a different platform and the reproducibility among them. In this presentation, we demonstrate the prototype of the Docker/WDL pipeline that can be run on the DNAnexus cloud platform, local cluster, native cloud, and local cloud. This approach would allow us to reduce the maintenance effort and increase the reproducibility across platforms.

Reference-free evaluation of assembly quality and improved trio binning

Arang Rhie, NIH, Bethesda, MD, USA

Emerging long-read sequencing technologies enables the generation of more continuous assemblies than in the past, often exceeding a previously built reference. Reference panels for benchmarking phased variants have been built by the Genome In a Bottle (GIAB) consortium. However, this is available only for a few human genomes and is mostly limited to variants called in the current human genome reference. As recent assembly methods aim to generate complete diploid assemblies, an unbiased, reference-free approach to assess assembly and phasing quality is needed. Here, we present merqury, a set of tools for evaluating assemblies in a reference-free environment. Merqury utilizes meryl, a k-mer counter for collecting k-mers in the assembly and in more accurate sequence reads (often Illumina whole genome sequencing reads). By comparing the k-mer sets, we calculate base level accuracy (QV), which eliminates the conventional read mapping and variant calling steps for QV estimation. Overall k-mer copy number analysis (spectra-cn) is performed, as in KAT (Mapleson et al, 2017). When identifiable parental or ancestral k-mers are available, we can re-use them as haplotype-specific k-mers (hap-mers), as introduced in our previous work on trio binning (Koren et al, 2018). Using these hap-mers, we can plot haplotype concordances for each sequence (contig or scaffold) in blobs (bubbles). In addition, we provide a novel way to compute phased block statistics and switch error rate, along with N* plots for phased blocks and hap-mer coordinates for further investigation. Similar to the spectra-cn analysis, mercury enables us to visually detect over-/under- represented hap-mers in the assembly. Using merqury, we benchmarked state-of-the-art genome assemblies generated with recent long-read and long-range technologies from two human trios (NA12878 and HG002) and a female zebra finch trio from the Vertebrate Genomes Project. We conclude that our method is fast and robust for assessing assembly and phasing quality for diploid and

(pseudo) haploid genome assemblies, and propose this analysis to be conducted on any assembly outcome. All codes used in this analysis is publicly available at https://github.com/marbl/meryl.

The little shop of assembly curation horrors

Kerstin Howe, Wellcome Sanger Institute, UK

Based on the experience gathered during more than a decade of reference assembly curation as part of the Genome Reference Consortium, we have taken on the task to assess and correct errors in the assemblies produced by the Vertebrate Genomes Project (VGP), and now also the Darwin Tree of Life Project. During this work, we discovered numerous shortcomings of existing algorithms, which could be improved as a result of our feedback, and helped optimising the assembly pipeline by influencing inclusions, exclusions and order of processes. The most immediate impact of our work lies in the correction we apply to the genome assemblies we assess. Our assembly curation pipeline using gEVAL is currently the only assembly process that evaluates all generated data types simultaneously, rather than in succession, and therefore allows to make highly informed decisions on sequence manipulations. Our processes detect contamination and erroneous duplications left behind by insufficient separation of haplotypes, significantly reduce the impact of missed joins and misjoins generated by scaffolding algorithms and create chromosome level scaffolds with reliable structure, leading to identification of autosomes and sex chromosomes. With an average of 150 interventions (breaks, joins, removal of erroneous duplications) per Gb of sequence, we increase the scaffold N50 by an average of 74% per assembly and assign more than 97% of the sequence to chromosome-level scaffolds. With 88 VGP assemblies alone now having gone through our pipeline, we have a story to tell about what is wrong in that assembly you thought was the best you can possibly produce – and how to make it better.

Quantifying improvements and needed improvements of VGP assemblies

Chul Lee, Seoul National University, Seoul, South Korea

High-quality reference genome assemblies of species of necessary to address fundamental questions in biology and disease. However, previous reference genome assemblies constructed with short reads have critical limitations for doing so. Based on recent cutting-edge technologies in sequencing and assembly, the international Vertebrate Genomes Project (VGP) have made substantial improvements in the quality of genome assemblies. The VGP is applying these improvements with the ambitious goal to generate at least one high-quality reference assemblies of all living ~70,000 vertebrate species. Here, we quantify the advancements in the quality of reference genomes, by comparing new high contiguity VGP assemblies we generated with old genome assemblies of the extant vertebrates, including the zebra finch, Anna's humming bird, platypus, and Korean giant mudskipper. We found thousands of errors in the previous assemblies that were corrected in the new VGP assemblies. These errors include false gene gains caused by phasing errors and eight types of false gene losses associated with GC contents and repeat content. These findings provide novel strategies to measure unprecedented influences of the innovative high-quality reference genome assemblies.

TOGA: Tool to infer Orthologs from Genome Alignments

Michael Hiller, Max Planck Institute, Dresden, Germany

Annotating coding genes is typically one of the first steps in annotating newly-sequenced genomes. Using these annotations to detect functionally-relevant changes in genes, such as patterns of positive selection or gene loss, it is necessary to distinguish orthologous from paralogous genes. I will present TOGA (Tool to infer Orthologs from Genome Alignments), a new method that uses whole genome alignments to integrate detecting intact or lost genes, determining gene orthology, and annotating orthologous genes in a comparative manner. In contrast to current orthology-detection methods that rely on similarity between gene sequences, TOGA extracts rich information on how the genomic context around genes aligns between species and uses machine learning to accurately distinguish orthologs from paralogs. As an exemplary analysis, TOGA has been used in the Bat1K project to obtain a comprehensive set of 12,931 orthologous genes from 48 mammalian genomes, which enabled a comprehensive screen for positively selected genes in bats and provided insights into mammalian phylogeny. In the era of large genome sequencing efforts, TOGA provides an integrative framework combining the identification and annotation of orthologous as well as lost genes that scales to hundreds and likely thousands of genomes of related species.

Genomes and population size: Historical demography of the vaquita

Phillip Morin, Southwest Fisheries Science Center, NOAA, La Jolla, CA
The vaquita is the smallest and most critically endangered marine mammal, with fewer than 20 remaining in the wild. As part of the Vertebrate Genome Project, we generated a reference genome sequence of the vaquita, consisting of 21 chromosome-length autosome scaffolds and the X-chromosome scaffold, with a scaffold N50 of 115 Mb. Genome-wide heterozygosity is the lowest of any mammalian species (0.01%), but heterozygosity is evenly distributed across the chromosomes, consistent with long-term small population size at genetic equilibrium, rather than resulting from a recent population bottleneck or inbreeding. Historical demography of the vaquita also indicates long-term population stability at fewer than 5000 individuals for over 100,000 years. Together, these analyses indicate that the vaquita genome has had ample opportunity to purge highly deleterious alleles and maintain diversity necessary for population health and adaptive potential.

Genomic conservation of the critically endangered great Indian bustard

Yadvendradev Jhala1, Yellapu Srinivas1, Sutirtha Dutta1, Vishnupriya Kollipakam1, Harish K2, Abhishek Tripathi2, Bipin Kumar2, Dushyant Singh Baghel2
1- Wildlife Institute of India, Dehradun, India; 2- Nucleome Informatics Pvt Ltd, Hyderabad, India
The Great Indian bustard (GIB) (*Ardeotis nigriceps*), Otidae family is endemic to India. Less than 150 individuals survive primarily in the western arid State of Rajasthan. The species is critically endangered and on Schedule I of the Wildlife Protection Act (1972). Primary threats to GIB are from power-line collisions, historical hunting, egg and chick predation by non-native predators, and habitat loss to agriculture. A conservation breeding program has been initiated as an insurance against extinction as well as to reintroduce/supplement wild populations after abating threats. A full genome sequencing of one GIB using PacBio Sequel at 63X and Illumina at 82X using 10X genomics library data and resequencing of two birds using Illumina Paired-End approach at 41x was done. The genome was assembled using WTDBG2, scaffolded with 10X datasets followed by polishing into 1269 scaffolds covering 99.16% of the genome having an N50 of 38Mb and a BUSCO score of 95% with the assembled genome size of 1.192 GB. Variant calling and tertiary analysis identified 3.3MB SNPs; 370,000 to 394,457 polymorphic SSR markers and 3000 SVs. Heterozygosity for the Great Indian Bustard at kmer-21 was 0.37% and for Houbara Bustard (Clamydotis macqeenii) was 0.56%. Using RNA Seq as transcript evidence, 16688 protein-coding genes were predicted and annotated. Similarly, the assembly of the mitochondrial genome and annotation resulted in the prediction of 14 protein-coding, 22 tRNAs and 2 rRNA genes. Otidae seems to evolved 120 MYA during which time they have gained 317 gene families and lost 773 in comparison to eight available avian genomes that had 14924 gene families, with 4675 single copy ortholog families. GIB and Houbara appear to have diverged form a common ancestor 30 MYA. The GIB has 27 unique gene families while Houbara has 3 unique gene families. PSMC revealed a population crash in GIB around 35000 YA which incidently coincides with the advent of humans into India. Genomic data will assist in determining founders and pedigree to maintain heterozygosity in the captive population for successful conservation breeding program.

Initial genome assemblies from the Darwin Tree of Life Project

Shane McCarthy, Wellcome Sanger Institute and Cambridge University, UK
The Darwin Tree of Life Project is a major UK initiative to help meet the visionary objectives of the Earth BioGenome Project. Nearly all phyla and over a third of described families have a representative species living in the UK. Furthermore, the UK is an intensively studied location, with a rich network of professional taxonomists, industrial monitoring agencies, and amateur experts as well as long term sites of ecological research. The Darwin Tree of Life Project aims to sequence all 60,000 described species present in the UK within the next ten years. We are establishing a network of collaborating centres to isolate, identify, and process the specimens, to set up new pipelines and workflows to process large numbers of species through DNA preparation, sequencing, assembly, gene finding and annotation, and to establish systems to disseminate and distribute the data while promoting large scale analysis and exploitation. We present initial assembly results and lessons learned so far from about 30 species, predominantly arthropods, sequenced and assembled since the project launch in November.

The human pangenome project

Benedict Paten, UCSC, Santa Cruz, CA, USA

Here we will present the goals of the recent NIH-funded human pangenome project (HPP), which is to generate as error-free, gapless, complete, and correctly haplotype-phased genome assemblies as possible from a set of 350 persons capturing the full extent of human diversity (>99% of allelic variants >1% frequency), and to provide these genomes as a resource to the international community to enable addressing fundamental unanswered questions in biology and disease. We will employ a multi-platform approach informed by the VGP and other efforts, using cutting-edge long read and linked read technologies to obtain the highest quality phased genomes. We aim to achieve over 99% haplotype-phasing for at least 700 haploid genomes from 350 diploid samples. We further aim to finish these genomes to be gapless from telomere-to-telomere (T2T) for each chromosome. Aim 4 will evaluate the genomes for accuracy and completeness and perform initial variant calling to assess the level of human diversity. Our effort will specifically target regions of the genome that have been excluded by other efforts, including segmental duplications, centromeres, and acrocentric DNA. To achieve these aims we have assembled a team consisting of experts from around the world in consent ethics, sample collection, sample extraction, and high-quality genome sequencing, assembly, finishing, evaluation, curation, and comparative genomics, including from the VGP. Initial sequencing using multiple technologies has resulted in some of the highest 3rd generation human genome assemblies yet brought together. We expect the result will be a pan-human genome reference, representing important human diversity not present in the current reference genome. The data we generate will enable a fundamental shift in human genetics, fostering new discoveries from the single-nucleotide to chromosomal levels, and revealing a more accurate and global view of the human population.

Vertebrate Genomes Project - Artificial Intelligence (VGP-AI): a proposal

Giulio Formenti, Rockefeller University, NY

The last decade saw upheavals in computer science with improvement of machine learning and the advent of deep learning. Today, such Artificial Intelligence (AI) applications have penetrated most scientific fields. However, its use in genomics has proceeded relatively slowly, particularly in relation to genome assembly. The increasing complexity of genome assembly processes in terms of datasets, parameters and pipelines, and of many of their subsidiary operations, now seems to be well-suited for machine learning approaches. The expanding availability of publicly accessible genome sequencing data in the framework of the Vertebrate Genomes Project (VGP), thanks to its species and dataset diversity, appears appropriate for the development of supervised and unsupervised AI algorithms that can help to generate better genome assemblies. Applications include, but are not limited to: extracting relevant genome features from raw data (e.g. genome size, heterozygosity, and possibly even chromosome number); data quality control (e.g. dataset matching); data classification (e.g. binning without trios); variant calling and polishing; automated curation; and extraction of genome features from the curated assemblies (e.g. alignment-free, kmer-free phylogenetic distances). It is even possible to envisage an assembly neural network using an iterative approach aimed at maximizing relevant assembly metrics (e.g. coverage depth and breath, QV, % of aligned/used bases, continuity, and BUSCO scores). The large number of successes achieved by AI systems in other research domains suggest that it will be beneficial for genome assembly, potentially leading to more accurate assembly through automated fine-tuning of existing parameters. It may even lead to radically new pipelines, allowing the identification of strategies and patterns that escape human attention. The development of these AI applications would require inputs from different backgrounds. Therefore, a focus group operating in the framework of the VGP to develop AI-based applications aimed at improving the genome assembly process could be extremely valuable.

Expanding the VGP network

Michaela West, UMR GDEC 1095 INRA, France; Claudio Ciofi, University of Florence, Italy;

Europe is the main western part of the Palearctic biogeographic realm and a temperate region with a quite diverse wildlife distributed across a variety of habitats from southern swamps to northern tundra and alpine biomes. Europe has more than 340 fresh-water fish species, about 200 of them endemic, 85 species of amphibians with 56 endemisms, particularly in southern Europe, 150 reptiles, about 530 native species of birds, 445 of them breeding in the continent and islands and 230 species of mammals, about 70 of them endemic to Europe. Approximately 37% of fishes, nearly a quarter of amphibians, one fifth of reptiles, 13% of birds and 15 % of mammals are threatened with extinction at the European level as a result of threats including habitat loss, fragmentation and degradation, pollution, poaching and deliberate persecution, overharvesting, invasive alien species and climate change. Here, we introduce a very preliminary plan aimed at providing reference genomes, according to the VGP pipeline, of European vertebrates with

particular reference to endemisms and strict endemisms. We report on low-coverage genome sequence of quite distinct southern European endemisms as an example of a holistic approach which would both complement current EU efforts and devise a new integrated initiative for taxa of particular interest.