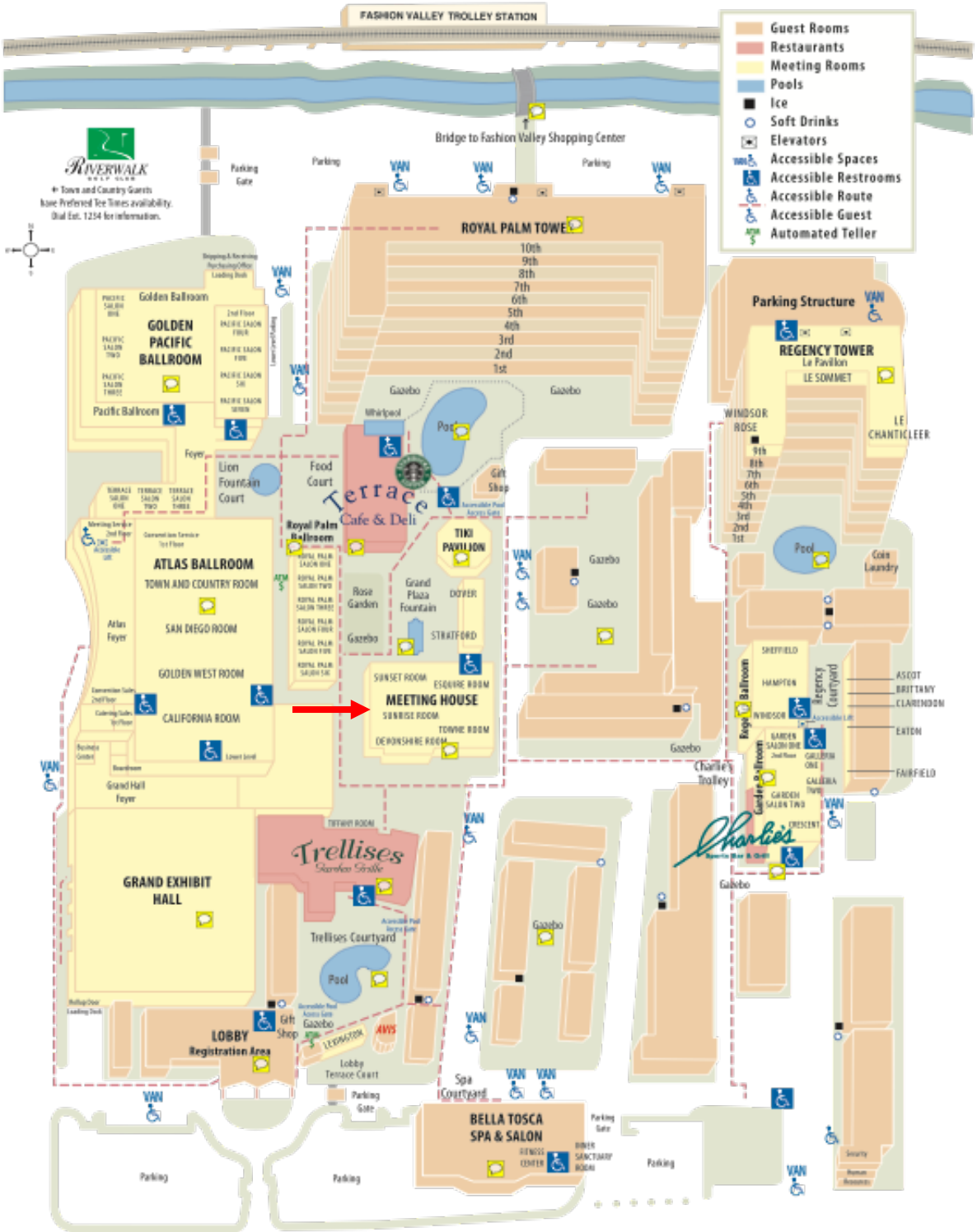
Genome 10K Workshop: Vertebrate Genomes Project

January 16, 2019
Town & Country San Diego, Sunrise Room, 9:30 AM - 4:30 PM
2019 Plant and Animal Genomes conference.

Contact: contact@genomeark.org
Twitter: @genomeark

Workshop Organizers:
Erich D. Jarvis, Ph.D., G10K Chair, Rockefeller University, NY: ejarvis@rockefeller.edu
Sadye Paez, Ph.D., MSPT, MPH, Rockefeller University, NY: spaez@rockefeller.edu

Town and Country
HOTEL & CONVENTION CENTER



VGP Goals: The goal of the Vertebrate Genomes Project (VGP) is to generate high-quality, error-free, near gapless, chromosome-level, phased, and annotated reference genome assemblies of at least one individual each representing all 66,000 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 ≥ 1 Mb; N50 scaffold ≥ 10 Mb; 90% of the genome assembled into chromosomes validated by at least 2 methods; an average QV40 or higher base call quality; and haplotypes phased as much as possible; we call this a 3.4.2.QV40 phased genome assembly. We call this a 3.4.2.QV40 phased metric, where the first three numbers are the exponents of the N50 contig, N50 scaffold, and level of chromosomal 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 is species representing shared divergence times before and soon after the last mass extinction event 66 MYA, which results in over 260 lineages. Analyses of these ordinal genomes is expected to be published in studies from 2019-2020, including in special journal issues. These studies will include: A genome-scale ordinal-level vertebrate family tree; Development of a more universal vertebrate gene nomenclature based on gene and genome evolution; Reconstruction of the common ancestor genome of all vertebrates; Determination of genomic signatures of specialized traits for each vertebrate class and species, including humans; and Determination of the genetics of why some lineages are more resistant to specific diseases or extinction; among other studies. 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 and UK Darwin Tree of Life; and Max Planck Institute). 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 of these 3.4.2.QV40 phased 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 ~66K-71K species.

Mission statement for the Jan 2019 reference G10K-VGP workshop: The mission of the 2019 G10K workshop is to further advance Phase 1 of the VGP, focusing on three main themes: 1) Analyses of the 1st data release assemblies of 14+ species for improving genome assembly quality and accuracy; 2) Scaling up production of genomes from sample collection and preparation, to sequencing, assembly and annotation; and 3) Initial comparative genomic studies of VGP genomes.

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:

Workshop Agenda

Zoom call in https://hhmi.zoom.us/j/320453351

Introduction (will start promptly at 9:30AM)

9:35-9:50 **Overcoming remaining challenges for completing Phase 1 VGP**
G10K Chair, Erich Jarvis, Rockefeller University, NY, USA

Workshop Session 1: Analyses of initial VGP assemblies and making improvements (9:45AM-11:40AM)

Chair: Erich Jarvis, Rockefeller University, NY, USA

9:45–10:15 **Analyses of the VGP 1st data release assemblies**
Arang Rhie, Adam Phillippy, & VGP assembly group, NIH, Bethesda, MD, USA

10:15–10:35 **Assembling complete mitochondrial genomes in VGP pipeline**
Giulio Formenti, University of Milan, Italy

10:35-10:45 **Coffee/Tea Break**

10:45–11:05 **Defining chromosomes in VGP assemblies**
Kerstin Howe, Wellcome Sanger Institute, UK

11:05-11:20 **A more advanced Arima Genomics' automated Hi-C approach with increased genome coverage for phasing and assembly**
Siddarth Selvaraj, Arima Genomics, San Diego, CA, USA

11:20–11:40 **Discussion on initial Phase 1 genomes**

Workshop session 2: Scaling up production of high quality reference genomes (11:40AM-2:50PM)

Chair: Olivier Fedrigo, Rockefeller University, USA

11:40-12:10 **Scaling-up production of high accuracy long reads and haplotype phased assemblies**
Jonas Korlach, Pacific BioSciences, Menlow Park, CA, USA

12:10-12:30 **Using PacBio circular consensus sequencing (CCS) to generate highly accurate assemblies**
Gene Myers, Max Planck Institute, Dresden, Germany

12:30-1:30 **Lunch, as part of PAG meeting for those that registered for the full conference**

1:30-1:50 **ONT horizons for higher accuracy long reads**
Iain MacLaren, Oxford Nanopore, Edinburgh, UK

1:50-2:10 **Bionano advances for chromosome-level haplotype-resolved *de novo* genome map assemblies**
Alex Hastie, Bionano Genomics, San Diego, CA, USA

2:10-2:30 **Scaling-up production of high molecular weight DNA**
Kelvin Liu, Circulomics, Baltimore, MD

2:30-2:50 **Group discussion for scaling-up to complete Phase 1 VGP**

2:50-3:00 **Coffee/Tea Break**

Workshop session 3: Comparative genomic examples for Phase 1 VGP genomes (3:00PM-4:30PM)
Chair: Joel Amstrong, UCSC, Santa Cruz, CA, USA

3:00-3:30 **Lessons learned from the 360 B10K & 200 mammal family-level genomes for the cactus reference-free alignment algorithm**
Joel Amstrong, UCSC, Santa Cruz, CA, USA

3:30-3:50 **Lessons learned from analyses of single nucleotide and amino acid convergences**
Chul Lee, Seoul National University, Seoul, South Korea.

3:50-4:10 **Example improvements to understanding gene family evolution with higher quality assemblies**
Erich Jarvis, VGP comparative genomics, Rockefeller University, NY, USA

4:10-4:30 **Summary of workshop outcomes**

4:30 **Close of G10K-VGP 2019 workshop**

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

- Adding and defining a value to the metric for haplotype phasing
- Need an approach for complete phasing of haplotypes, at all steps
- Missing genomic sequence from VGP assemblies
- Sex chromosome
- Mitochondrial genomes
- Defining what is an assembled chromosome and chromosomal-level assembly
- Comprehensive all species VGP database
- Scaling up production of genome assemblies, from sample collection to annotation
- Improvements to multilayered and reference free alignment approaches
- Planning initial studies with VGP 1-2nd data release genomes (>65 species)

2019 G10K-VGP workshop company sponsors