Genome 10K Annual Conference

The Rockefeller University, NY

September 12-14, 2018

AGENDA

Table of Contents

Conference information

Conference organizers

Erich D. Jarvis, Ph.D., G10K Chair, The Rockefeller University: ejarvis@rockefeller.edu
Sadye Paez, Ph.D., MSPT, MPH, The Rockefeller University: spaez@rockefeller.edu
Lauren Shalmiyev, MPH, The Rockefeller University: lshanker01@rockefeller.edu
Beth Shapiro, Ph.D., UCSC: bashapir@ucsc.edu

Breakout session organizers

Richard Durbin, Ph.D., Wellcome Trust Sanger Institute and Cambridge: rd109@cam.ac.uk
Olivier Fedrigo, Ph.D., The Rockefeller University: ofedrigo@rockefeller.edu
Paul Flicek, Ph.D., Ensembl EBI: flicek@ebi.ac.uk
Warren Johnson, Ph.D., Smithsonian: johnsonwe@si.edu
Lisa Komoroske, Ph.D., UMass Amherst: lkomoroske@umass.edu
Harris Lewin, Ph.D., UC Davis: hlewin@gmail.com
Jacquelyn Mountcastle, MPH, The Rockefeller University: jmountcast@rockefeller.edu
Gene Myers, Ph.D., Max Planck Institute of Molecular and Cell Biology and Genetics: gene@mpi-cbg.de
Benedict Paton, Ph.D., UCSC: bpaten@ucsc.edu
Andreas Pfenning, Ph.D., Carnegie Melon University: apfenning@cmu.edu
Adam Phillippy, Ph.D., NHGRI, NIH: adam.phillippy@nih.gov
Oliver Ryder, Ph.D., San Diego Zoo: ORyder@sandiegozoo.org
Josefin Stiller, Ph.D., University of Copenhagen: josefinstiller@gmail.com
Emma Teeling, Ph.D., University College Dublin: Emma.Teeling@ucd.ie
Françoise Thibaud-Nissen, Ph.D., NCBI, NIH: thibaudf@ncbi.nlm.nih.gov
Sonja Vernes, D. Phil, Max Planck Institute for Psycholinguistics: Sonja.Vernes@mpi.nl
Tandy Warnow, Ph.D., University of Illinois at Urbana-Champaign: warnow@illinois.edu

Goals

The primary goals of the 2018 G10K annual conference are to advance the mission of its Vertebrate Genomes Project (VGP), to foster the development of methods for error-free haplotype phased genomes, and to further integrate with other related large-scale genome efforts. Specific goals include announcing the first data release of Phase 1 VGP genomes representing all vertebrate orders, planning the science to be conducted and published with Phase 1 genomes, holding a hackathon workshop for error-free complete genome assemblies, developing rapid and accurate alignment annotation methods, and conducting workshops for integrating with the Bird 10K (B10K), Bat 1K, Earth Biogenome, and other large-scale vertebrate genome projects.

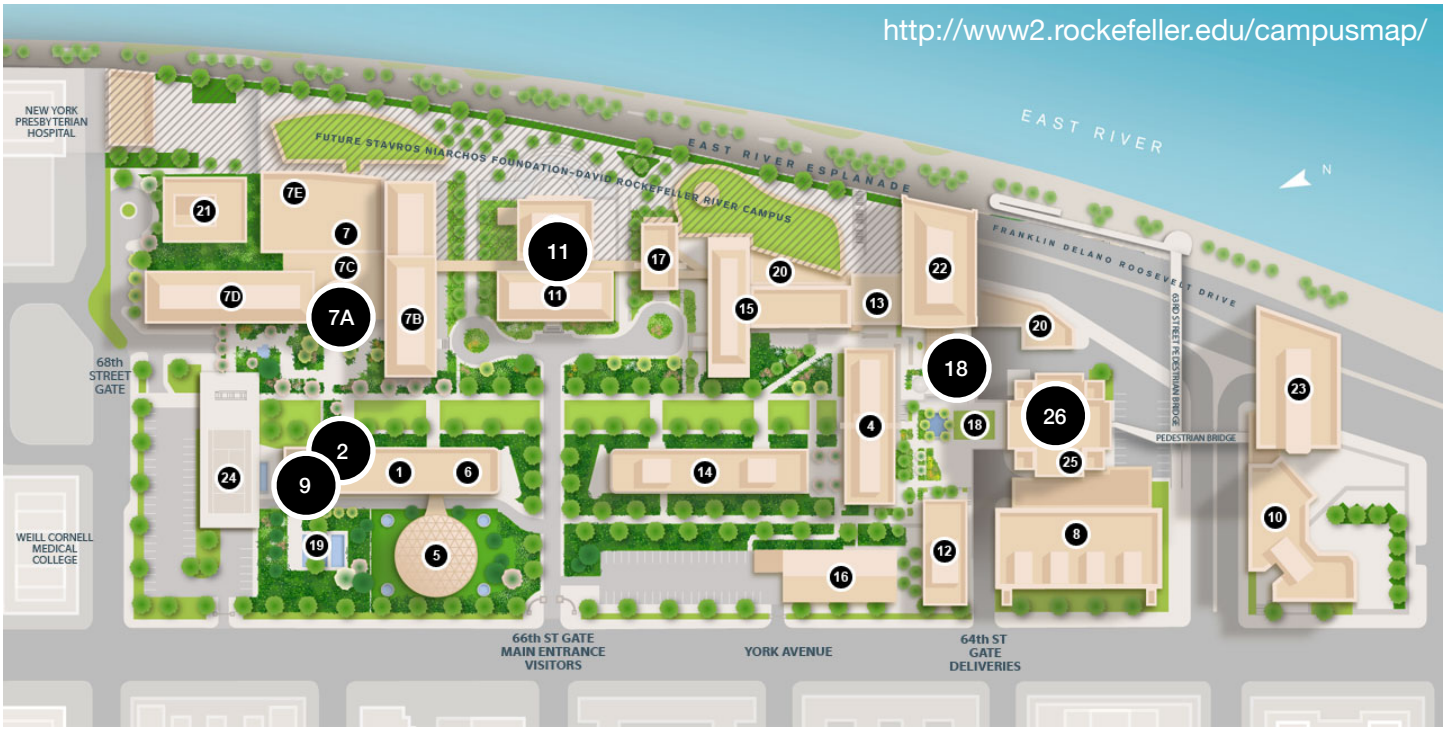
Preparation

In order for attendees to more easily follow the workshop presentation and discussions, they can watch the G10K-VGP 2017 year-end presentation: https://tinyurl.com/y8pc42f9 (Username: G10K; Password: Genomes) and read the VGP plan in brief here. With these documents and presentations, please respect standard rules of scientific ethical conduct for credit and the G10K-VGP Embargo Data Use Policy.

Location

Rockefeller University
1230 York Avenue (66th street entrance)
New York, NY, 10065
(212) 327-8437 (Lauren Shalmiyev's office number)

Most events will take place in the Collaborative Research Center (CRC) building on the North-East side of campus.



- 7A Collaborative Research Center (CRC)
- 11 Founders Hall
- 18 Peggy Rockefeller Plaza
- 26 Weiss Research Building
- 2 Abby Lounge & Dining Room
- 9 Faculty & Student Club

Day 1: Wednesday, September 12th, 2018

Press conference on G10K-VGP & first data release announcement

G10K-VGP Program Director: Sadye Paez, The Rockefeller University

Location: Great Hall, in Founders

10:00 AM - 11:00 AM **Press conference** (by invitation only)

11:00 AM - 12:00 PM **Press one-on-one interviews** (round tables for topics and persons)

12:00 PM - 12:50 PM Registration and Lunch

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

General session 1: Group reports

G10K Chair: Erich Jarvis, Rockefeller University

Location: Carson Family Auditorium, Floor B, Collaborative Research Center (CRC)

Zoom link: https://hhmi.zoom.us/j/460641369

1:00 PM - 1:10 PM **Welcome**
Richard Lifton, President of The Rockefeller University

1:10 PM - 1:30 PM **Purpose of G10K 2018 conference and status of VGP**
Erich Jarvis, The Rockefeller University

1:30 PM - 2:00 PM **Report from VGP sample prep group: Achieving ultra-High Molecular Weight DNA**
Jacquelyn Mountcastle & Olivier Fedrigo, The Rockefeller University

2:00 PM - 2:30 PM **Report from VGP assembly working group 1: Progress towards complete and error-free genome assemblies**
Adam Phillippy, NHGRI, NIH

2:30 PM - 3:00 PM **Report from VGP assembly working group 2: Analyses of remaining errors that need to be fixed in reference genomes**
Harris Lewin, UC Davis & Kerstin Howe, Wellcome Trust Sanger Institute

3:00 PM - 3:15 PM Coffee break

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

3:15 PM - 3:45 PM **Report from VGP alignment group: Lessons learned from B10K and 200 mammal projects**
Benedict Paten, UCSC

3:45PM - 4:15 PM **Report from VGP annotation group: Scaling up, RNASeq vs IsoSeq, haplotypes and more**
Françoise Thibaud-Nissen, NCBI & Paul Flicek, Ensembl EBI

4:15PM - 4:45PM **Report from a VGP comparative genomics group: Vocal learning and lessons learned for comparative genomics**

Erich Jarvis, The Rockefeller University & Sonja Vernes, Max Planck Institute for Psycholinguistics

4:45PM - 5:15PM **Open discussion on reports**: **Strengths and weaknesses**

5:30 PM - 7:30 PM Welcome reception and VGL tour

Hors d'oeuvres and drinks at the Peggy Rockefeller Plaza (South-East corner of campus). If inclement weather, event will take place in Weiss Research Building, East Room.

Tour of Vertebrate Genome Lab (VGL) on 7th Floor Weiss building adjacent to Plaza

Day 2: Thursday, September 13th, 2018

8:00 AM - 9:00 AM Breakfast

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

** Set up posters in same location **

9:00 AM - 9:15 AM **Discussion on goals of concurrent sessions**

G10K Chair: Erich Jarvis, The Rockefeller University

Location: Carson Family Auditorium, Floor B, Collaborative Research Center (CRC)

Breakout session 1: Assembly, B10K, and Bat1K

9:15 AM -12:00 PM **Breakout session 1A: Assembly hackathon towards complete and error-free assemblies, v1 assembly retrospective**

Chairs: Adam Phillippy, Gene Myers, Richard Durbin, & Olivier Fedrigo

Location: Room 206, Collaborative Research Center (CRC)

9:15 AM -12:00 PM **Breakout session 1B: B10K genomics project integration**

Chair: Josefin Stiller

Location: Room 102, Collaborative Research Center (CRC)

9:15 AM -12:00 PM **Breakout session 1C: Bat1K genomics project integration**

Chairs: Emma Teeling & Sonja Vernes

Location: Room 106, Collaborative Research Center (CRC)

10:45 AM - 11:00 AM Coffee break

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

12:00 PM - 1:00 PM Lunch and Poster session

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

Breakout session 2: Alignment & annotation, EBP, and sample prep

1:30 PM - 4:30 PM **Breakout session 2A: Alignment and annotation hackathon**

Chairs: Benedict Paten, Françoise Thibaud-Nissen, & Paul Flicek

Location: Room 506, Collaborative Research Center (CRC)

3:45 PM - 4:45 PM **Breakout session 2B: Earth Biogenome Project integration**

Chair: Harris Lewin

Location: Room 306, Collaborative Research Center (CRC)

1:30 PM - 4:30 PM **Breakout session 2C: Sample preparation workshop**

Chairs: Jacquelyn Mountcastle & Olivier Fedrigo

Location: Room 406, Collaborative Research Center (CRC)

3:30 PM - 3:45 PM Coffee break

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

4:45 PM - 5:45 PM **Summaries of the 6 breakout groups**

G10K Chair: Erich Jarvis

Location: Carson Family Auditorium, Floor B, Collaborative Research Center (CRC)

6:30 PM - 8:30 PM Dinner

Abby Dining room

Day 3: Friday, September 14th, 2018

8:00 AM - 9:00 AM Breakfast

Abby Lounge

General session 2: Stakeholders

Chair: Beth Shapiro

Location: Carson Family Auditorium, Floor B, Collaborative Research Center (CRC)

Zoom link: https://hhmi.zoom.us/j/460641369

9:00 AM - 10:00 AM **Report on funding thus far obtained, needed, and planned**
Erich Jarvis, The Rockefeller University

10:00 AM - 10:30 AM **Scientific Journal's interest in Genomics**
Orli Bahcall (Nature). Overview of current publishing initiatives for genomics
Steve Mao (Science). Science magazines' vision in genomics
Laurie Goodman (GigaScience). Future vision of GigaScience in genomics.

10:30 AM - 10:45 AM Coffee break

1st floor lobby, Collaborative Research Center (CRC)

10:45 AM - 12:00 PM **Assigning VGP genome projects and publications timeline**
Erich Jarvis, The Rockefeller University & Beth Shapiro, UCSC

12:00 PM - 1:00 PM Lunch and Poster session

Faculty & Student Club, beneath Abby Lounge & Dining Room

Breakout session 3: Assembly, comparative Genomics, and conservation

1:00 PM - 3:30 PM **Breakout session 3A: Assembly Hackathon towards complete and error-free assemblies, v2 assembly plans**
Chairs: Gene Myers, Adam Phillippy, Richard Durbin, & Olivier Fedrigo
Location: Room 406, Collaborative Research Center (CRC)

1:00 PM - 3:30 PM **Breakout session 3B: Comparative genomics with VGP genomes**
Chairs: Sonja Vernes, Andreas Pfenning, & Tandy Warnow
Location: Room 306, Collaborative Research Center (CRC)

1:00 PM - 3:30 PM **Breakout session 3C: Conservation efforts to be conducted with VGP genomes**

Chairs: Warren Johnson, Oliver Ryder, & Lisa Komoroske

Location: Room 106, Collaborative Research Center (CRC)

2:00 PM - 2:15 PM Coffee break

1st floor lobby, Collaborative Research Center (CRC)

General session 3: Conclusions and decisions

G10K Chair: Erich Jarvis, The Rockefeller University

Zoom link: https://hhmi.zoom.us/j/460641369

3:30 PM - 4:30 PM **10-minute summaries of the 3 breakout groups and meeting summary**

Location: Carson Family Auditorium, Floor B, Collaborative Research Center (CRC)

4:30 PM - 5:30 PM **G10K Council meeting (by invitation only)**

Location: Room 506, Collaborative Research Center (CRC)

Breakout session abstracts and programs

Breakout session 1A: Assembly hackathon towards complete and error-free assemblies — v1 assembly retrospective

Chairs: Adam Phillippy; Gene Myers, & Richard Durbin

Location: Room 206, Collaborative Research Center (CRC)

Time: September 13th, 2018 9:15 AM -12:00 PM

Abstract:

The focus of the assembly group breakout session 1A will be to summarize issues identified in the current best genome version 1 (v1) assemblies, and brainstorm on solutions to improving them for the Phase 1 VGP 260+ ordinal genomes in progress. Some of the remaining major issues identified in the current assemblies to date are due to current algorithms not handling heterozygosity and repeats properly, not handling large differences between species, and potentially requiring longer, more accurate reads to get through long repeats. The format of this session will be brief 10 minute presentations by discussion leaders, with encouraged questions and open discussions. The session participants include the G10K-VGP assembly group that meets weekly online, and is open to all G10K conference attendees.

Program:

9:15 / **Introduction**

Adam Phillippy, NHGRI, NIH

- Retrospective analysis of v1 assemblies
- Goal is to organize and understand challenges in assembling v1 genomes

9:30 / **Contigging challenges**

Shane McCarthy, Wellcome Trust Sanger Institute

- High repeat, high heterozygosity genomes
- Phasing issues, retained haplotigs, switch errors, and more
- Repeat masking pros and cons

9:45 / **Scaffolding challenges**

Arang Rhie, NHGRI, NIH

- High repeat, high heterozygosity genomes
- Retained haplotigs and scaffold interleaving
- Chimeric joins and missed breaks
- Error propagation from step to step

10:00 / **Curation challenges**

Kerstin Howe, Wellcome Trust Sanger Institute

- Common issues found with v1 assemblies
- Scaling curation efforts to many genomes

10:15 / **Operational challenges**

Olivier Fedrigo, The Rockefeller University

- Metadata and sample tracking
- Contamination and data QC
- AWS and DNAnexus
- Raw data and assembly submission

10:30 / **Hi-C update**

Zev Kronenberg, Phase Genomics

- FALCON-Phase and other happenings

10:45 / **Coffee break**

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

11:00 / **Wrap-up**

- Organize and prioritize outstanding assembly issues
- What worked and what didn't work in v1?
- Plans for a v1 assembly paper

Breakout session 1B: Bird10K genomics project integration

Chair: Josefin Stiller

Location: Collaborative Research Center building conference Room 102

Zoom link: https://hhmi.zoom.us/j/460641369

Time: September 13th, 2018 9:15 AM -12:00 PM

Abstract:

The focus of the Bird 10,000 genomes (B10K) project breakout session will be to summarize lessons learned from new analyses of the 48-bird Avian Phylogenomics Project representing all bird orders, the currently unpublished 360+ draft avian genomes representing all families, planned studies with the family level genomes, and integration with the reference VGP. The format of this session will be presentations, open discussions, and proposed solutions. The session participants include B10K members that meets bi-monthly online, and is open to all G10K conference attendees.

Program:

9:15 / **Introduction**
Josefin Stiller, University of Copenhagen

9:20 / **Multiple sequence alignment and multi-locus species tree estimation**
Tandy Warnow, University of Illinois at Urbana-Champaign

9:40 / **Overview and current status of analyses of the B10K**
Josefin Stiller, University of Copenhagen

10:00 / **Sampling strategy for the B10K and the way forward**
Robb Brumfield, Lousiana State University

10:20 / **Whole genome Cactus alignments and annotations**
Benedict Paten and Joel Armstrong, UCSC

10:40 / **Network methods for phylogenomics**
Luay Nakhleh, Rice University

11:00 / **Effects of trimming phylogenomic data**
Brant Faicloth, Louisiana State University

11:20 / **Data problems in comparative genomics**
Ruta de Fonseca, University of Copenhagen

11:40 / **Adaptive genomics of avian traits**
Agostinho Antunes, University of Porto

Breakout session 1C: Bat1K genomics project integration

Chairs: Emma Teeling & Sonja Vernes

Location: Collaborative Research Center building conference Room 106

Time: September 13th, 2018 9:15 AM -12:00 PM

Abstract:

The focus of the Bat 1,000 genomes (Bat1K) project breakout session will be to summarize the current progress of generating bat reference genomes, planned studies with the bat genomes for Phase 1 VGP, planned studies for a family level bat genome Phase 2 VGP, and integration with the reference VGP. The format of this session will be presentations and open discussions. The session participants include Bat1K members that meets semi-monthly online, and is open to all G10K conference attendees.

Program:

9:15 / **Introduction from the chairs**

9:30 / **Goals of the Bat1K project**

10:00 / **Current progress of pilot phase of Bat1K**

10:30 / **Collaborations and data sharing in Bat1K**

10:45 / **Coffee break**

Greenberg atrium, Floor B, Collaborative Research Center (CRC)

11:00 / **Directed Discussion**

Breakout session 2A: Alignment and annotation hackathon

Chairs: Benedict Paten, Françoise Thibaud-Nissen, & Paul Flicek

Location: Collaborative Research Center building conference Room 506

Time: September 13th, 2018 1:30 PM - 4:30 PM

Abstract:

The alignment and annotation group breakout session will summarize the progress and status of the current draft bird B10K 300 families and mammal 200 family genome project alignments. Annotation groups from NCBI, Ensembl and UCSC will discuss progress over the past six months, plans for 2019 and the value of broadly sampled transcriptomic data as ordinal genomes start to become available at greater frequency. Important issues for overall integration such as ideal data flow, data presentation and distribution and required metadata to efficiently complete the alignments and annotation will also be covered. The format of this session will be brief presentations and open discussions. The session participants include the G10K-VGP

alignment and annotation groups lead by UCSC, NCBI, and Ensembl leadership, and is open to all G10K conference attendees.

Program:

1:30 / **Alignment topics**

- Updates on 200 mammal and 300 bird genome alignments - (20 mins, Joel Armstrong, Benedict Paten, Mark Diekhans)
- Discussion and plans for G10K species / pan vertebrate alignments - (25 mins, all)

2:15 / **Annotation topics**

- Ensembl annotation progress and plans - (10 mins, Fergal Martin)
- Plans for CAT annotations - (10 mins, Mark Diekhans)
- Plans for NCBI annotations (10 mins, Françoise Thibaud-Nissen)
- Discussion and plans for G10K species annotations (45 mins, all, lead by discussion leaders identified below):
    - Strengths, weaknesses and limitations of our approaches (Fergal Martin)
    - Genomic features to annotate: genes, repeat masking, regulation (Mark Diekhans)
    - The need for broadly sampled transcriptomic data (Fergal Martin)

3:30 / **Break**

Greenberg atrium, Floor B, Collaborative Research Center

3:45 / **Integration topics**

- Discussion (30 mins, all, lead by discussion leaders identified below):
    - Data presentation, flow and distribution (Paul Flicek)
    - Metadata requirements for efficient annotation and alignment (Françoise Thibaud-Nissen).
- Summary / Next steps / Action Items (15 mins)

Discussion leaders:

- Françoise Thibaud-Nissen, NCBI
- Paul Flicek, Ensembl EBI
- Fergal Martin, Ensembl EBI
- Leanne Haggerty, Ensembl EBI
- Benedict Paten, UCSC
- Joel Armstrong, UCSC
- Mark Diekhans, UCSC

Breakout session 2B: Earth Biogenome Project integration

Chair: Harris Lewin

Location: Room 306, Collaborative Research Center (CRC)

Time: September 13th, 2018 3:45 PM - 4:45 PM

Abstract:

The focus of the Earth Biogenome Project (EBP) breakout session will be to summarize the current status of the project to sequence representatives of all eukaryotic species, lessons learned from the VGP, and develop plans for integration with the reference VGP and other related vertebrate genomes project. The format of this session will be a brief presentation and open discussions. The session participants include EBP members that meet monthly online, and is open to all G10K conference attendees.

Program:

3:45 / **Introduction and discussion**

- Overview of EBP, Harris Lewin, UC Davis
- Plans for EBP November 2018 workshop and launch event
- Scaling production of reference genomes for eukaryotes: Opportunities and Challenges

Breakout session 2C: Sample prep group

Chairs: Jacquelyn Mountcastle & Olivier Fedrigo

Location: Room 406, Collaborative Research Center (CRC)

Zoom link: https://hhmi.zoom.us/j/610542882

Time: September 13th, 2018 1:30 PM - 4:30 PM

Abstracts:

The focus of the sample preparation group breakout session will be to summarize issues identified in the current Phase 1 VGP genomes in progress, lessons learned to achieve producing high molecular weight DNA for genome sequencing and at scale for the VGP and beyond. Discussions will include challenges with field collection, tissue types, preservation methods, bioarchive, HMW isolation protocols, blanket permits for national and international exchanges of tissue, and required metadata. The format of this session will be brief presentations and open discussions. The session participants include the G10K-VGP sample prep group that meets bi-monthly, and is open to all G10K conference attendees.

Program:

1:30 / **Introduction**

Olivier Fedrigo, The Rockefeller University

1:35 / **Tissue preservation tests for HMW gDNA**

Sylke Winkler, Max Planck Institute of Molecular and Cell Biology and Genetics

1:50 / **Discussion**

- Tissue preservation project
- Publication
- Future directions

2:05 / **Collecting samples for G10K/VGP in international locations outside of sequencing hubs**

Bob Murphy, University of Toronto

2:20 / **Successes and outstanding challenges for sample collection in the marine environment**

Lisa Komoroske, UMass Amherst; and Phil Morin, NOAA

2:35 / **De novo genome assemblies utilizing in vitro cultured cells: What to hope for and to expect**

Oliver Ryder, San Diego Zoo

2:50 / **Discussion**

- Sample collection
- Metadata
- Permits

3:15 / **Fostering efficient integration of quality whole-genomes with long-term institutional archival of samples**

Warren Johnson, Smithsonian

3:30 / **Bionano Genomics high molecular weight gDNA isolation**

Steffen Oeser, Bionano Genomics

3:45 / **Nanobind High MW DNA extraction, library preparation for long-read sequencing, and optical mapping**

Kelvin Liu, Circulomics Inc.

4:00 / **Discussion**

- HMW gDNA
- VGP sample archives
- Current sample collection and HMW isolation guidelines

Breakout session 3A: Assembly hackathon towards complete and error-free assemblies— v2 assembly plans

Chairs: Gene Myers, Adam Phillippy, Richard Durbin, & Olivier Fedrigo

Location: Collaborative Research Center building conference Room 406

Time: September 14th, 2018 1:00 PM – 3:30 PM

Abstract:

The focus of the assembly group breakout session 2 will be to continue where the group left off on session 1 the previous day, and to develop plans for scaling to 6 genomes per week in 2018-2019 and eventually 125 genomes per week from 2020 onwards, and for version 2 (v2) assemblies of the Phase 1 VGP that are as complete and error-free as possible. The format of this session will be brief presentations, open discussions, and proposed solutions. The session participants include the G10K-VGP assembly group that meet weekly online, and is open to all G10K conference attendees.

Program:

1:00 / **Introduction and day 1 summary**

Adam Phillippy, NHGRI

- General strategy discussion followed by challenge-specific subgroups
- Arrive at a consensus path forward for v2 assemblies

1:15 / **Discuss overall strategy for v2 assemblies**

Gene Myers, Max Planck Institute of Molecular and Cell Biology and Genetics

- What needs to change from v1 to v2?
- Are new methods required?
- Are new data interfaces required?
- Do we continue to work as individual labs or combine efforts?
- Do we need to recruit more help?

2:00 / **Break**

Room 102, Collaborative Research Center (CRC)

2:15 / **Discuss specific challenges and solutions, breakout groups**

- Rank challenges identified on day 1

3:15 / **Wrap-up**

- Set responsibilities and goals for v2 assemblies

Breakout session 3B: Comparative genomics with VGP genomes

Chair: Sonja Vernes, Tandy Warnow, Andreas Pfenning

Location: Room 306, Collaborative Research Center (CRC)

Time: September 14th, 2018 1:00 PM – 3:30 PM

Abstract:

The focus of the comparative genomics breakout session will be to summarize lessons learned from ongoing analyses of large-scale draft vertebrate genome projects (10s to 100s of genomes), and plans for proposed studies to be conducted with the Phase 1 VGP 260+ ordinal genomes. The format of this session will be presentations and open discussions. The session participants include G10K-VGP members of the G10K-VGP vocal learning and language group that meets weekly online, and those conducting phylogeny, genome evolution, and trait analyses studies, and is open to all G10K conference attendees.

Program:

1:00 / **Introduction from the chairs**

1:05 / **Genome-scale phylogenetic tree inference**
Tandy Warnow, University of Illinois at Urbana-Champaign,

1:20 / **Phylogenetic networks**
Luay Nakhleh, Rice University

1:35 / **Chromosome evolution in vertebrates**
Harris Lewin, UC Davis

1:50 / **Tracing the Evolution of Tissue and Cell Type-Specific cis-Regulatory Elements across the Vertebrate Phylogeny**
Irene Kaplow, Carnegie Mellon University

2:00 / **Break**

2:15 / **How many genomes are needed for comparative genomic trait analyses**
James Cahill, The Rockefeller University

2:30 / **Rules of convergent amino acid and nucleotide evolution**
Chul Lee, Seoul National University

2:45 / **Directed Discussion**

- What can we/should we/do we aim to do with the Phase 1 VGP data release (goals, projects)?
- What types of genetic differences can we look for to associate with traits in comparative studies?
- What are the biggest challenges we face with comparative genomics and what do we need to solve them?
- How can we benefit from/take advantage of both long-read and short-read genomes to greatest benefit?
- Integrating expression/functional data (RNA-Seq, Iso-seq, Epigenomics, etc)

Breakout session 3C: Conservation efforts to be conducted with VGP genomes

Chair: Warren Johnson, Oliver Ryder, & Lisa Komoroske

Location: Collaborative Research Center building conference Room 106

Time: September 14th, 2018 1:00 PM – 3:30 PM

Abstract:

The focus of the conservation breakout session will be to present and discuss plans for proposed studies to be conducted with the Phase 1 VGP 260+ ordinal genomes, determine a standardized approach for conservation genomics within the VGP, discussion of how Phase 1 genomes might provide insights on the 5th mass extinction 66 MYA, updates on the 6th mass extinction paper from the G10K-VGP, and use of complete and error-free genomes for population management and conservation efforts. The format of this session will be presentations and open discussions. The session participants include members of the G10K-VGP conservation group, including those supporting the critically endangered Kakapo parrot, Vaquita dolphin, Leatherback Turtle, Golden Mantella fish, and Mexican salamander, among others, and is open to all G10K conference attendees.

Program:

1:00 / **Introduction from the chairs**

1:10 / **Brief presentation of ideas by session participants and speakers**

- Beth Shapiro, UCSC
- Oliver Ryder, San Diego Zoo
- Andrew Crawford, Universidad de los Andes
- Lisa Komoroske, UMass Amherst
- Federica Di-Palma, Earlham Institute
- Klaus-Peter Koepfli, Smithsonian
- Tomas Marques, University Barcelona

Poster abstracts

Posters schedule

Thursday session	Friday session
Poster A1	Poster C1
Poster D1	Poster A5
Poster C5	Poster C2
Poster B1	Poster A7
Poster C6	Poster A2
Poster A6	Poster A3
Poster B2	Poster A4
Poster C3	
Poster C4	

Posters will be available for viewing on Thursday and Friday. The above schedule reflects the formal presentation schedule for each poster.

A. Genome assembly

Poster A1: Evaluating genome and transcriptome variation across the Antarctic Notothenioid fish radiation to explore causes and consequences of adaptive speciation.

Iliana Bista[1,2], Shane McCarthy[2], Melody S. Clark[3], Thomas Desvignes[4], John Postlethwait[4], C.-H. Christina Cheng[5], H. William Detrich III[6], Walter Salzburger[7], Zemin Ning[1], William Chow[1], Jonathan Wood[1], Kerstin Howe[1], Eric Miska[8], Richard Durbin[1,2] and the Sanger Institute VGP team[1].

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

[2] Cambridge University, Department of Genetics, Downing Site, Cambridge, UK

[3] British Antarctic Survey, Natural Environment Research Council, Madingley Road, Cambridge, UK

[4] Institute of Neuroscience, University of Oregon, Eugene, USA

[5] Department of Animal Biology, University of Illinois at Urbana – Champaign, Urbana, USA

[6] Department of Marine and Environmental Sciences, Northeastern University, Nahant, USA

[7] Zoological Institute, University of Basel, Basel, Switzerland

[8] Gurdon Institute, University of Cambridge, Cambridge, UK

Correspondence: Iliana Bista, ib8@sanger.ac.uk and Richard Durbin, rd@sanger.ac.uk

Introduction: The Notothenioid radiation of Antarctic fish is one of the most dramatic examples of marine adaptive radiations, and is dominating the Southern Ocean in fish species richness and biomass. At the Wellcome Sanger Institute and in collaboration with the Vertebrate Genomes Project (VGP) and

Genome 10K, we are sequencing a total 25 species representing 5 Notothenioid families. We are producing draft reference assemblies using PacBio (50X) and 10X Chromium for 5 species *Cottoperca gobio, Trematomus bernachii, Harpagifer antarcticus, Gymnodraco acuticeps,* and *Pseudochaenichtys georgianus*. Additionally, we are generating 10X Genomics assemblies for an additional 12 species, and HiSeq X Illumina data for 8 species across the radiation.

Objectives: Assembling using long read sequencing will allow study of highly repetitive gene families (e.g. antifreeze glycoproteins, AFGPs) and detection of structural variation, whilst providing valuable insights into the genome evolution of this important marine radiation. Furthermore, we are generating transcriptome data for key tissues, with a specific interest in transposon control in the germline through the Piwi/piRNA small regulatory RNA pathway. Comparison of expression levels of piRNAs across 16 Notothenioid species and in relation to specific transposons and their abundance will enable investigation of mechanisms controlling genome size expansion in this group.

Results: So far, we have generated genomic and transcriptomic data for 22 and 12 species respectively (out of 25). Long read assemblies have been generated using a Falcon-unzip and Scaff10X pipeline, and 10X Chromium with Supernova 2.0. Assembly size ranges between 651Mb – 1Gb, increasing for the more derived species. For species *C. gobio* (Bovichthidae) only, we have generated a hybrid assembly using PacBio, 10X, BioNano maps and Hi-C to achieve VGP quality standards (Contig N50 5Mb, Scaffold N50 14.74 Mb).

Conclusions: Overall this study will provide a deep genomic characterization of this important fish group, and an important platform to investigate the mechanisms of fish genome evolution.

Poster A2: Use of Nanopore sequence to validate and refine a mammalian genome

David Mohr and Alan Scott
Johns Hopkins University, MD, USA
Correspondence: afscott@jhmi.edu

We previously used linked reads and optical mapping to build a de novo genome from an endangered species. Overall, sequence contiguity was excellent with an N50 of 30 Mb and 98% of the predicted genome occurring in 170 scaffolds. However, we observed some regions of disagreement between DNA lengths estimated from the optical maps and the actual sequence. These occurred, as expected, at regions of repetitive sequence where they manifested as read pile-ups. Because nanopore sequencing is agnostic to repetitive motifs we made long DNA libraries using both standard and rapid library kits, ran these on the Oxford Nanopore R9.x flow cells and basecalled with Albacore v2.0.1. Reads were aligned to our existing scaffolds using minimap2, with 91-94% identity. We obtained reads up to 324 kb, most of which aligned extremely well with our published genome, thereby validating the optical map and linked read strategy used for assembly. Several regions were identified where the assembly will improve with the use of long reads. We conclude that nanopore sequencing is a valuable adjunct to more conventional NGS approaches and as throughput increases and accuracy improves will become an important tool in genome assembly.

Poster A3: Complete assembly of parental haplotypes with trio binning

Phillippy, Adam: Genome Informatics Section, NHGRI, NIH
Koren, Sergey: Genome Informatics Section, NHGRI, NIH
Rhie, Arang: Genome Informatics Section, NHGRI, NIH
Walenz, Brian: Genome Informatics Section, NHGRI, NIH
Dilthey, Alexander T. : Institute of Medical Microbiology, Heinrich-Heine-University Düsseldorf, Germany
Bickhart, Derek M. : ARS USDA, USA
Kingan, Sarah B. : Pacific Biosciences, Menlo Park, California, USA
Hiendleder, Stefan: The University of Adelaide, Roseworthy SA, Australia
Williams, John L. : The University of Adelaide, Roseworthy SA, Australia
Smith, Timothy P.L. : ARS USDA, USA
Correspondence: adam.phillippy@nih.gov

Introduction and Research Objectives: Reference genome projects have historically selected inbred individuals to minimize heterozygosity and simplify assembly. We challenge this dogma and present a new approach designed specifically for heterozygous genomes.

Methods: Prior approaches for assembling heterozygous diploid genomes only phase small variants or partially reconstruct the haplotypes. Our “trio binning” method uses short reads from two parental genomes to partition long reads from an offspring into haplotype-specific sets. Each haplotype is then assembled independently. The output of this process is a complete genome for each parental haplotype, containing all classes of haplotype variation assembled from the long reads, including single nucleotide, structural, and copy number variants.

Results: To demonstrate the effectiveness of trio binning on a heterozygous genome, we sequenced an F1 cross between cattle subspecies *Bos taurus taurus* and *Bos taurus indicus*, and assembled both parental haplotypes with NG50 haplotig sizes >20 Mbp each, surpassing the quality of current cattle reference genomes. In addition, both haplotypes approach 99.999% accuracy at the base level using PacBio data alone. Further application of this method to a benchmark human trio also achieved high accuracy and recovered complex structural variants missed by alternative approaches. Trio binning of both the human and cattle haplotypes successfully reconstructed highly heterozygous loci. For example, in human, both parental haplotypes of the Major Histocompatibility Complex (MHC) were accurately assembled and showed perfect human leukocyte antigen (HLA) gene typing accuracy. For cattle, many heterozygous regions between the newly assembled Angus and Brahman haplotypes intersected with previously identified quantitative trait loci (QTL), making it a suggestive candidate for adaptation among the cattle breeds.

Conclusion: Given the quality of the assemblies with this approach, we propose trio binning as a new best practice for diploid genome assembly that will enable platinum-quality reference genomes and new studies of haplotype variation and inheritance.

Poster A4: The VGP assembly working group: building and sharing the 1st batch of automated genome assemblies

Rhie, Arang: Genome Informatics Group, NHGRI, NIH
Phillippy, Adam: Genome Informatics Group, NHGRI, NIH
The VGP Assembly Working Group
Correspondence: arang.rhie@nih.gov

Introduction: The Genome10K Vertebrate Genomes Project (VGP) consortium aims to create a digital open-access genome library of at least one high-quality, near-gapless, phased and annotated chromosomal-level assembly of all extant vertebrate species.
Research Objectives / Rationale: The initial phase of this project focused on finishing 200 species from each vertebrate order, to a quality standard of >1 Mb N50 contig size, >10 Mb N50 scaffold size, average base quality >QV40, and 90% of the sequence assigned to chromosomes.
Methods: With the maturation of long-read sequencing and long-range scaffolding technologies, it is now possible to construct reference-grade assemblies, de novo, at reasonable cost. The VGP has begun collecting and sequencing ordinal samples using 4 such technologies: PacBio long reads, 10X Genomics linked reads, Bionano optical maps, and Arima Genomics Hi-C libraries. To generate assemblies in a uniform way, the VGP assembly working group designed an automated pipeline and applied the proposed pipeline to an initial set of 15 species including a mammal, bird, amphibian, reptile, skate and fish. The proposed pipeline begins with contig generation using PacBio, followed by haplotig purging. Then, scaffolding is performed in an iterative approach using 10X Genomics, followed by Bionano, and finally Hi-C. Additional rounds of polishing is performed using PacBio and 10X reads.
Results: So far, the genomes of 16 species have finished this process, with the majority meeting the VGP quality standard. After final curation, the genomes will be submitted to the public assembly databases. In the interim, the assemblies, intermediate files, and all the raw data are shared via an Amazon Web Services S3 bucket "s3://genomeark" (http://genomeark.s3.amazonaws.com). The assembly pipeline is publicly available on GitHub (https://github.com/VGP/vgp-assembly) and is being converted to a push-button cloud application by DNAnexus.
Conclusion: An improved, comprehensive strategy is under continued development, and these are aimed at better separation of haplotypes and scaffolding using an integrated approach.

Poster A5: Using the gEVAL genome browser to evaluate and improve draft assemblies.

William Chow, Joanna Collins, Sarah Pelan, James Torrance, Jonathan Wood, Ying Yan and Kerstin Howe

Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK

Correspondence: kj2@sanger.ac.uk

Introduction: Whilst there are workflows and ever evolving pipelines available for assembling new genomes, the challenge remains that results work favorably for one species may not work on another. This may be attributed to varying genome complexity and structure (e.g. heterozygosity and repeat content) and the performance of the applications to resolve haplotypes completely under these different conditions. Because of this, the need to assess these assemblies beyond standard length metrics is important before deciding whether they are good enough to be used as a reference genome.

Objectives: Our strategy to evaluate, but also actively improve assemblies, involves combining publicly available tools such as those used for contamination screening, separation of haplotigs from primary components and assessment of core gene presence, complemented by manual curation using our bespoke web-based genome assembly evaluation browser, gEVAL (https://geval.org.uk). gEVAL features analyses from a wide range of aligned data types such as optical maps, linked-reads, Hi-C, transcript sequences as well as whole genome alignments to other assemblies. It allows the user to quickly and easily access regions that lack concordance with multiple datatypes through issue lists and colour coding in the browser.

Results: The gEVAL browser has previously been used successfully in supporting the Genome Reference Consortium's release of the human, mouse, and zebrafish reference genomes as well as the Swine Genome Sequencing Consortium's pig reference. As part of the Genome 10k Vertebrate Genomes Project (VGP) assembly strategy, gEVAL has been used in the manual curation and evaluation of all draft assemblies.

Conclusions: This curation effort not only brings forward a higher quality assembly, but yield insights that are relayed back to our technical collaborators to further improve on data generation/quality and methods, whilst also allowing the VGP to create an improved workflow strategy.

Poster A6: The Platinum genome of *Choloepus didactylus* (Pilosa, Xenarthra): a platform to study evolution of basal placental mammals

Marcela Uliano-Silva1,2, Sylke Winkler3, Eugene Myers3, Camila Mazzoni1,2

1Leibniz Institute for Zoo and Wildlife Research, Department of Evolutionary Genetics, Berlin, Germany

2 Berlin Center for Genomics in Biodiversity Research, Berlin, Germany

3 Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

Correspondence: marcela.uliano@izw-berlin.de

Introduction: Sloths, anteaters and armadillos compose a basal mammalian Order (Xenarthra) that evolved in South America ~59 Mya. Xenarthrans share several characteristics not present in other placental mammals, such as supplementary intervertebral articulations and intra-abdominal testes. The extant sloths, three-toed *Bradypus* and two-toed *Choloepus,* have both developed two of the slowest behaviors and metabolisms among mammals. Both sloth lineages present a series of anatomical adaptations to life suspended on the tree canopies, in contrast to their mostly terrestrial and giant extinct relatives. Such adaptations were acquired in parallel and represent a remarkable case of convergent evolution since *Bradypus* and *Choloepus* are only distantly-related, sharing a last common ancestor ~30 Mya. Despite their evolutionary peculiarities, Xenarthra genomics has been so far neglected.

Objectives: We aim to assemble the first chromosomal level (Platinum) genome for *Choloepus didactylus* to be a platform to study Xenarthra evolution and adaptations. Together with the Vertebrate Genomes Project (VGP) Assembly Group of Scientists, we are developing and applying a hierarchical assembly pipeline using multiple sequencing technologies.

Results: Long Pacbio reads (53x coverage, N50readlengh= 22Kb) were produced and assembled using Falcon and Falcon-Unzip resulting in a 3.7 Gb assembled genome in 3051 primary contigs with N50= 10Mb. Next, Purge haplotigs was applied and 1167 small contigs (200Kb to 1Mb in size) were flagged as haplotigs resulting in a Pacbio-purged assembly of 3.6Gb in length in 1884 primary contigs with a N50= 11Mb. The long size of the initial Pacbio reads seem to be an important factor for the generation of this highly-contiguous initial assembly, since analysis of the Chromium 10X reads suggests that *Choloepus didactylus* heterozygosity is high (0.9%). In the next steps we will apply Chromium 10X, Bionano maps and HiC reads for scaffolding and chromosome assignment.

Conclusions: The high-quality Platinum genome of *Choloepus didactylus* will well-support future molecular studies and phylogenies.

Poster A7: SMRT long-read sequencing and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (*Hirundo rustica rustica*)

Formenti, Giulio*, Department of Environmental Science and Policy, University of Milan (Milan, Italy).
Chiara, Matteo*, Department of Biosciences, University of Milan (Milan, Italy).
Poveda, Lucy, Functional Genomics Center of Zurich, University of Zurich, (Zurich, Switzerland).
Francoijs Kees-Jan, Bionano Genomics (San Diego, CA, USA).
Bonisoli-Alquati, Andrea, Department of Biological Sciences, California State Polytechnic University (Pomona, CA, USA).
Canova, Luca, Department of Biochemistry, University of Pavia (Pavia, Italy).
Gianfranceschi, Luca, Department of Biosciences, University of Milan (Milan, Italy).
Horner, David Stephen, Department of Biosciences, University of Milan (Milan, Italy).
Saino, Nicola, Department of Environmental Science and Policy, University of Milan (Milan, Italy).
* These authors contributed equally to the work
Correspondence: giulio.formenti@unimi.it

Introduction: The barn swallow (*Hirundo rustica*) is a migratory bird that has been the focus of a large number of ecological, behavioural and genetic studies.
Research Objectives/Rationale: To facilitate further population genetics and genomic studies, here we present a high-quality genome for the European subspecies (*Hirundo rustica rustica*).
Methods: We have assembled a highly contiguous genome sequence using Single Molecule Real-Time (SMRT) DNA sequencing and Bionano optical maps. We compared and integrated optical maps derived both from the Nick, Label, Repair and Stain and from Direct Label and Stain technologies.
Results: For our SMRT-only assembly, the direct labelling system more than doubled the assembly N50 with respect to the nickase system. The dual enzyme hybrid scaffold led to a further marginal increase in scaffold N50 and an overall increase of confidence in scaffolds. After removal of haplotigs, the final assembly is approximately 1.21 Gbp in size, with an N50 scaffold value of over 25.95 Mbp, representing an improvement in N50 of over 650-fold with respect to a previously reported assembly based on paired-end short read data.
Conclusions: This high-quality genome assembly represents a valuable resource for further studies of population genetics of the barn swallow and for studies concerning the evolution of avian genomes. It also represents one of the first genomes assembled combining SMRT sequencing with the new Bionano Direct Label and Stain technology for scaffolding, highlighting the potential of this methodology to contribute to substantial increases in the contiguity of genome assemblies.

B. Annotation

Poster B1: Fish clade annotation in Ensembl

Leanne Haggerty, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Osagie Izuogu, Denye Ogeh, Fergal J. Martin, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Correspondence: leanne@ebi.ac.uk

Introduction: The actinopterygians, or ray-finned fish, represent more than half of the planet's vertebrate biodiversity. However, until recently, fish genome sequence data was scarce: before 2013, the INSDC databases housed assemblies for just 31 fish species.Global sequencing and assembly efforts, including the VGP and Transcriptomes of the 1000 Fishes (Fish-T1K) projects, have created massive amounts of high-quality data. By April 2015, Fish-T1K reported that 7,000 genome-quality fish tissue samples had been collected, representing 51 of 71 orders. Today, INSDC have 213 fish assemblies available including 87 submitted this year.

Objectives: The surge in available sequence data for the fish creates an unprecedented opportunity to produce well-supported gene sets. In response to this we have recently annotated 41 fish species. By generating the annotations in parallel we produce gene sets in a consistent and efficient manner.

Results: Given the large evolutionary distances across the fish, we found that RNA-seq data were key to producing high quality gene annotations. In particular RNA-seq data represents the only method to reliably find novel genes and exons. Fish that had a broad sampling of data from tissues with complex transcriptomes such as brain, gonads and lung generally had more complete annotations. For species with RNA-seq data, in addition to our main annotation tracks, we have generated sample-specific RNA-seq gene tracks, providing a window into the transcriptome for these species. Depending on the species, the samples may be different tissues, development stages and/or different environmental conditions.

Conclusions: The collection of new fish annotations and other fish genome resources such as updated comparative genomics and variation data will be released in Ensembl version 94 (expected September 2018).

Poster B2: Optimizing *de novo* transcriptome assembly approaches for large mammalian gene families

Laurel R. Yohe[1,2], Kalina T. J. Davies[3], Stephen J. Rossiter[3], & Liliana M. Dávalos[2]

[1]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA

[2]Department of Geology & Geophysics, Yale University, New Haven, CT, USA

[3]School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom

[4]Consortium for Inter-Disciplinary Environmental Research, Stony Brook University, Stony Brook, NY, USA

Correspondence: lmdavalos@gmail.com

Introduction: RNA-seq and bait capture provide valuable biological information for a wide range of analyses, but meaningful interpretation of results requires well-assembled, high quality data sets. Many assembly approaches are optimized for single-copy genes, and few studies have assessed how these methods perform for large gene families with highly similar gene sequences, especially for *de novo* transcriptome assemblies for organisms without a reference genome.

Objectives: We evaluated how different assembly approaches performed in recovering mammalian olfactory receptors, the largest mammalian protein-coding gene family of the common vampire bat *Desmodus rotundus*.

Methods: Baits derived from the transcriptomes were used to capture and sequence the relevant gene. The vampire bat genome allowed us to characterize the expected set of olfactory receptors for the species. We compared two *de novo* assembly methods to genome-guided transcriptome assemblies and olfactory receptor amplicons sequenced using Sanger sequencing.

Results: From the genome, we discovered more than 300 intact olfactory receptors, double the current estimate. Bait-capture sequencing and genome-guided assemblies performed the best of the assembly approaches, recovering the most olfactory receptors that mapped to the genome. We found that the number of olfactory receptors is highly variable across methods. Although the transcriptome does not recover the complete receptor repertoire, it does assemble reliable sequences with low levels of chimeric contigs, and the resulting baits were the most effective method of sequencing olfactory receptors short of sequencing a high-quality genome. Much of the variation across assemblies is due to low coverage of baits or transcripts. The poor performance of the Sanger amplicons is likely due to primer biases, and the receptors recovered from the transcriptome represent a more diverse sampling of the gene family.

Conclusions: We recommend implementing bait capture approaches for the most complete representation of large gene families.

C. Comparative genomics

Poster C1: Genomic insights from accelerated regions in vocal learning birds

James A. Cahill[1], Joel Armstrong[2], Alden Deran[2], Carolyn Khoury[1], Benedict Paten[2], David Haussler[2,3], Erich D. Jarvis[1,4]

[1] Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY 10065

[2] Jack Baskin School of Engineering, University California Santa Cruz, Santa Cruz, California 95064, USA

[3] UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA 95064;

[4] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

Correspondence: ejarvis@rockefeller.edu

Introduction: Vocal learning, the ability to mimic sounds from conspecifics and the environment, is a key component of spoken-language shared between humans and a limited number of non-human groups including: cetaceans, bats, pinnipeds, elephants, oscine songbirds, parrots and hummingbirds. Each of these groups likely developed their vocal learning independently but they share substantial phenotypic convergence at the behavioral, physiological and gene expression levels, suggesting the existence of shared underlying pathways for vocal learning.

Research Objectives/Rationale: We aim to test this hypothesis, by identifying regions under positive selection in vocal learning birds. We ask if there is convergent positive selection in the same genes across vocal learning lineages or in genes with shared functions.

Methods: We generated a whole genome alignment of 33 bird species including members of all three vocal learning bird groups and their close non-learning relatives using Progressive Cactus. Then we tested if there were accelerated regions in each vocal learning bird clade and compared them to one another and to known existing databases for vocal learning related traits.

Results: We find high densities of accelerated regions in avian vocal learners near genes with known speech functions and neurodevelopmental functions. Convergence between vocal learning birds accelerated regions occurs at a higher rate than is expected by chance and is clustered in the genome. Moreover, vocal learning birds' accelerated regions are overrepresented in genes associated with human accelerated regions and Autism Spectrum Disorders.

Conclusion: Vocal learning bird's accelerated regions reveal known and novel candidate genes for involvement in the development and maintenance of vocal learning. Our results suggest that comparative analysis of vertebrate genomes may yield insights into the evolution of human traits, such as language, and potentially aid in the identification genes relevant to human disease.

Poster C2: Convergent amino acid substitutions in avian vocal learning clades – not how many genes, but who

Lee, Chul[1], Kim, Joowan[1], Kim, Heebal[1], Jarvis, Erich D[2]

[1]Seoul National University, Republic of Korea.,

[2]Rockefeller University and HHMI, USA

Correspondence: heebal@snu.ac.kr and ejarvis@rockefeller.edu

Introduction: Vocal learning, the ability to imitate vocalizations, is a convergent trait rarely observed in independent lineages in animals (songbirds, parrots, hummingbirds, human, etc.). Avian and primate vocal learners share convergent brain structures and gene expression alterations in vocal communication brain regions. However, it is still unclear if coding variants correlates with convergent gene expression and vocal learning. The recent big bang of genome sequences of 48 avian species that span the phylogeny of modern birds brought about an unprecedented opportunity to investigate genomic variants specific to vocal learning clades.

Research Objectives/Rationale: Here, we analyzed avian genomes with three vocal learning clades to determine if the behavior and neural convergence is associated with molecular convergences related to gene product alterations.

Methods: We upgraded and performed comparative genomic approaches to detect vocal learner-specific convergent variants at the amino acid, codon, and nucleotide levels. Generating and comparing proper control sets, we investigated a preponderance and biological functions of genomic convergences of vocal learning birds. Evolutionary analyses were applied to find key candidate genes.

Results: We discovered that regardless of the species combination, the product of origin-branch lengths correlated with the number of convergent substitutions. This correlation did not differ for vocal learners, meaning that the number of convergent substitutions found did not differ from phylogenetic expectation. Further, there was no difference in correlations of convergent vs divergent amino acid variants compared to random control sets. Nevertheless, the genes with convergent substitutions in vocal learners were enriched for learning and for genes with specialized regulation in vocal learning brain regions, and this was the only species combination to have with these enrichments. A key candidate gene, *DRD5*, was supported by multiple lines of evidences, including human-specific variants compared to non-human primates.

Conclusions: Our findings reveal insights into macro-evolution of vocal learning and principles of convergent gene evolution.

Poster C3: Further resolution of hypotheses on convergent brain regions for learned song in songbirds and speech in humans

Gregory L Gedman[1,2], Andreas Pfenning[3], Morgan Wirthlin[3], Bettina Haase[1], Olivier Fedrigo[1], Erich D. Jarvis[1,2]

[1]The Rockefeller University, New York, NY,, USA.

[2]Howard Hughes Medical Institute, Chevy Chase, MD, USA.

[3]Dept of Computational Biology, Carnegie Mellon University, Pittsburgh, PA

Correspondence: ejarvis@rockefeller.edu

Introduction: Vocal learning is a rare, complex, convergent behavior seen in several independent lineages of birds and mammals, including humans, and is the basis for learned speech. To explain this convergent behavior, several competing hypotheses have been proposed for convergent song/speech brain regions in song learning birds and humans. For example, one hypothesis proposed by Doupe and colleagues is that songbird RA and HVC of the song production pathway are broadly analogous to human laryngeal motor cortex (LMC) and Broca's area respectively; another proposed by Jarvis and colleagues is that HVC and RA are analogous to cortical layers 3 and 5 of LMC, respectively, whereas the songbird LMAN vocal learning nucleus is analogous to layer 3 of Broca's area. Recent transcriptome analyses using microarrays from our group supported the hypothesis that RA is analogous to layer 5 of human LMC, but findings for HVC were inconclusive. In contrast, recent brain cooling experiments from the Long lab concluded that HVC was more similar to Broca's area than to LMC.

Objectives: These inclusive and alternative results for HVC may be attributed to not having expression profiles of the cell populations surrounding HVC, limited number of genes on microarrays, computational tools that prevented mapping of two or more avian brain regions to one region in humans with different cortical layers, or to the inability to cool specific cortical layers.

Methods: To test these alternatives, we conducted mRNA-seq experiments of all four major song nuclei (Area X, LMAN, HVC, and RA) and surrounding cell populations of the zebra finch (a songbird; *Taeniopygia guttata*), and compared them with human (cortical column) and macaque (cortical layer) microarray brain data from the Allen Institute for Brain Science. We developed new hypothesis-driven computational tools that allowed for cortical layer specific analyses within the same region.

Results: These new studies strongly confirmed that RA has a specialized molecular profile most similar to the human LMC and layers 5/6 neurons of primate primary motor cortex (PMC). HVC was found to have a specialized profile similar to human LMC, more so than to Broca's area, and specifically to layer 2 neurons of primate PMC. The arcopallium motor pathway cells adjacent to RA and the nidopallium cells adjacent to HVC (as well as LMAN) also shared significant molecular specializations with layers 5 and 2/3 of the PMC.

Conclusions: These new findings with many more genes and more brain regions support the hypothesis that RA and HVC are analogous to different cortical layers of human LMC, as well as support the broader nuclear-to-layer hypothesis of avian and mammalian brain cell type homologies.

Poster C4: Enhancing" vocal learning: Gene regulatory specializations in song circuits

Lindsey Cantin[1], Morgan Wirthlin[2], James Cahill[1], Gregory Gedman[1], Caitlin Gilbert[1], Thomas Carrol[1], Andreas Pfenning[2], Erich D. Jarvis[1,3]

[1]The Rockefeller University, New York, NY, USA

[2]Carnegie Mellon University, Pittsburgh, PA, USA

[3]Howard Hughes Medical Institute, Chevy Chase, MD USA

Correspondence: ejarvis@rockefeller.edu

Introduction: Vocal learning is a rare trait shared by a small number of distantly related species, including songbirds, parrots, hummingbirds, humans, cetaceans, bats, elephants and pinnipeds. As a result of convergent evolution, these species (at least as studied in birds and humans) share similar vocal learning behavior development and analogous vocal learning brain circuitry with specialized gene expression profiles. A previous study in our lab identified 55 genes with convergent gene expression specializations between the human laryngeal motor cortex (LMC) and the songbird robust nucleus of the arcopallium (RA) when normalized to their surrounding motor regions. When disrupted, some of these genes in humans are associated with communication disorders and understanding their regulation may provide key insights into the genetic causes of these disorders.

Objectives: Non-coding regulatory regions control the expression of genes. Identifying and comparing these regions across vocal learning species will provide clues into the molecular basis for evolution of this complex trait.

Methods: Native ChIP-seq and comparative genomic experiments were preformed to identify non-coding regulatory regions. H3K27ac is generally associated with active enhancer DNA.

Results: We identified several hundred H3K27ac peaks with differential activity between the songbird RA and the surrounding motor region, including those associated with specialized genes identified in our RNA-seq experiments. We also aligned the genomes of avian vocal learning species and their closest vocal non-learning relatives to identify convergent accelerated non-coding regions that may be playing a role in the gene specializations of avian vocal learning lineages. Some of the vocal learning specific accelerated regions overlap with the differential histone acetylation peaks.

Conclusions: These experiments will help to identify the genetic mechanism in which vocal learning evolves and will benefit research for understanding vocal communication disorders.

Poster C5: Genome resequencing of the complete order Crocodylia to investigate patterns of evolution

Ray, David A[1], Osmanski, Austin[1], Brittain, Katherine[2], Jones, Elizabeth[2], Suh, Alexander[3], and Gongora, Jaime[2]

[1]Department of Biological Sciences, Texas Tech University, TX 79409, USA.

[2]Sydney School of Veterinary Sciences, Faculty of Science, University of Sydney, NSW 2006, Australia.

[3]Department of Ecology and Genetics, Uppsala University, 752 36 Uppsala, Sweden.

Correspondence: david.a.ray@gmail.com

Abstract

Introduction: The order Crocodylia consists of approximately 23 extant species in three families; Alligatoridae (alligators and caimans), Crocodylidae (crocodiles) and Gavialidae (gharial). These species inhabit a range of tropical and subtropical environments across the Americas, Africa, Asia, and Oceania. Previous research by the International Crocodilian Genomes Working Group (ICGWG) has sequenced the genomes of the American alligator, Australian saltwater crocodile, and Indian gharial—representatives of all three extant crocodilian families. That project revealed ancestral patterns of genome evolution among archosaurs and established these taxa as an emerging model given their robust immune systems, low cancer incidence, low heterozygosity, and their unusually slow rate of genomic evolution. Despite these unique adaptations and the conservation status of many of the component taxa, the genomic basis of these adaptations remains poorly understood.

Objectives: To address this problem, a collaborative sequencing effort from the ICGWG is generating genome drafts of all remaining species of the order to improve our knowledge of the evolutionary history and unique biology of crocodilians.

Methods/Results: De novo and reference-based assemblies have been generated using a combination of sequencing methods. To date, all but three have been assembled and analyses to study whether the low evolutionary change found in our early genome work is extended at the family and species level. We also aim to examine what variations may have occurred at a geographical level within crocodile and alligators, with an interest to understanding adaptation to varying environments.

Conclusions: Together, these genome assemblies represent a molecular toolkit required to increase our understanding of the unique evolutionary history of this enigmatic yet charismatic order.

Poster C6: How to make a rodent giant: Capybara genome reveals the complex evolution of gigantism

Santiago Herrera-Álvarez[1], Elinor Karlsson[2], Oliver A. Ryder[3], Kerstin Lindblad-Toh[2,4], & Andrew J. Crawford[1]

[1] Department of Biological Sciences, Universidad de los Andes, Bogotá 111711, Colombia

[2] Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA

[3] San Diego Zoo Institute for Conservation Research, San Diego Zoo Global, Escondido, CA, 92027, USA

[4] Department of Medical Biochemistry and Microbiology, Uppsala University, 752 36 Uppsala, Sweden.

Correspondence: s.herrera706@uniandes.edu.co

Introduction: Gigantism is the result of one lineage within a clade evolving extremely large body size relative to its small-bodied ancestors, a phenomenon observed numerous times in animals. Theory predicts that the evolution of giants should be constrained by two tradeoffs. First, because body size is negatively correlated with population size, purifying selection is expected to be less efficient in species of large body size, leading to a genome-wide elevation of the ratio of non-synonymous to synonymous substitution rates (d_N/d_S). Second, gigantism is achieved through higher number of cells and higher rates of cell proliferation, thus increasing the likelihood of cancer. However, the incidence of cancer in gigantic animals is lower than the theoretical expectation, a phenomenon referred to as Peto's Paradox.

Objectives/Methods: To explore the genetic basis of gigantism in rodents and uncover genomic signatures of gigantism-related tradeoffs, we sequenced the genome of the capybara, the world's largest living rodent, and developed a comparative analysis with 15 additional rodent genomes.

Results: We found that d_N/d_S is elevated genome-wide in the capybara, relative to other rodents, implying a higher mutation load. Conversely, a genome-wide scan for adaptive protein evolution in the capybara highlighted several genes involved in growth regulation by the insulin/insulin-like growth factor signaling (IIS) pathway. Capybara-specific gene-family expansions included a putative novel anticancer adaptation that involves T-cell-mediated tumor suppression, offering a potential resolution to Peto's Paradox.

Conclusion: Based on our findings, we hypothesize that gigantism in the capybara evolved by an increase in cell proliferation through the ISS pathway, and the establishment of the T cell-mediated tumor suppression pathway as an anticancer adaptation. Furthermore, we showed that the capybara harbors an increased mutation load, possibly an inevitable outcome of an increase in body size.

D. Phylogenomics

Poster D1: Unlocking the phylogenomic (super)tree of birds (and other organisms)

Braun, Edward L[1,*], Kimball, Rebecca T[1], Oliveros, Carl H[2], Wang, Ning[3], Barker, F Keith[4], Field, Daniel J[5], Ksepka, Daniel T[6], Chesser, R Terry[7], Moyle, Robert G[8], Brumfield, Robb T[2,9], Faircloth, Brant C[2,9], and Smith, Brian Tilston[10]

[1]Department of Biology, University of Florida, Gainesville, FL 32607 USA; [2]Department of Biological Sciences, Louisiana State University, 202 Life Sciences Building, Baton Rouge, LA 70803 USA;

[3]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA;

[4]Department of Ecology, Evolution and Behavior and Bell Museum of Natural History, University of Minnesota, 1479 Gortner Ave, Saint Paul, MN 55108 USA;

[5]Department of Biology & Biochemistry, Milner Centre for Evolution, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom;

[6]Bruce Museum, One Museum Drive, Greenwich, CT 06830 USA;

[7]USGS Patuxent Wildlife Research Center, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560 USA;

[8]Biodiversity Institute, University of Kansas, 1345 Jayhawk Blvd., Lawrence, KS 66045 USA;

[9]Museum of Natural Science, Louisiana State University, 119 Foster Hall, Baton Rouge, LA 70803 USA;

[10]Department of Ornithology, Division of Vertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024 USA

Correspondence: ebraun68@ufl.edu

The evolutionary history of life is written in the genomes of extant organisms, but efforts to unlock that history face two major challenges. First, there are gaps in the availability of tissues necessary to generate genomic data. Second, phylogenetic estimation using genome-scale data requires substantial computational resources. The first challenge can be solved using sequence capture methods, which can leverage existing natural history collections to use older (even >100-year old) museum specimens. Supertree methods, which integrate source trees to yield a synthetic large-scale phylogeny in a computationally-efficient manner, have the potential to solve the second challenge. Our goal was to build the computational infrastructure able to generate and update a supertree that integrates phylogenetic analyses of sequence capture and whole genome data. To accomplish this, we built an efficient supertree pipeline to integrate phylogenomic trees using a backbone based on existing megaphylogenies (taxon-rich phylogenies). Using birds as a test case we were able to construct a phylogenomic supertree that includes 688 species, representing more than 6% of named bird species. The pipeline was very efficient, rapidly generating an avian supertree that represented an accurate synthesis of the available phylogenomic data. A time-calibrated version of our phylogenomic supertree supported a model whereby all three major avian clades (Palaeognathae, Galloanseres, and Neoaves) underwent radiations close to the Cretaceous-Paleogene (K-Pg) boundary. Supertree methods appear to

provide an efficient way to synthesize phylogenomic information. We believe that our approach will provide a way to produce an estimate of the tree of life that can be continually updated.

2018 G10K Conference Sponsors