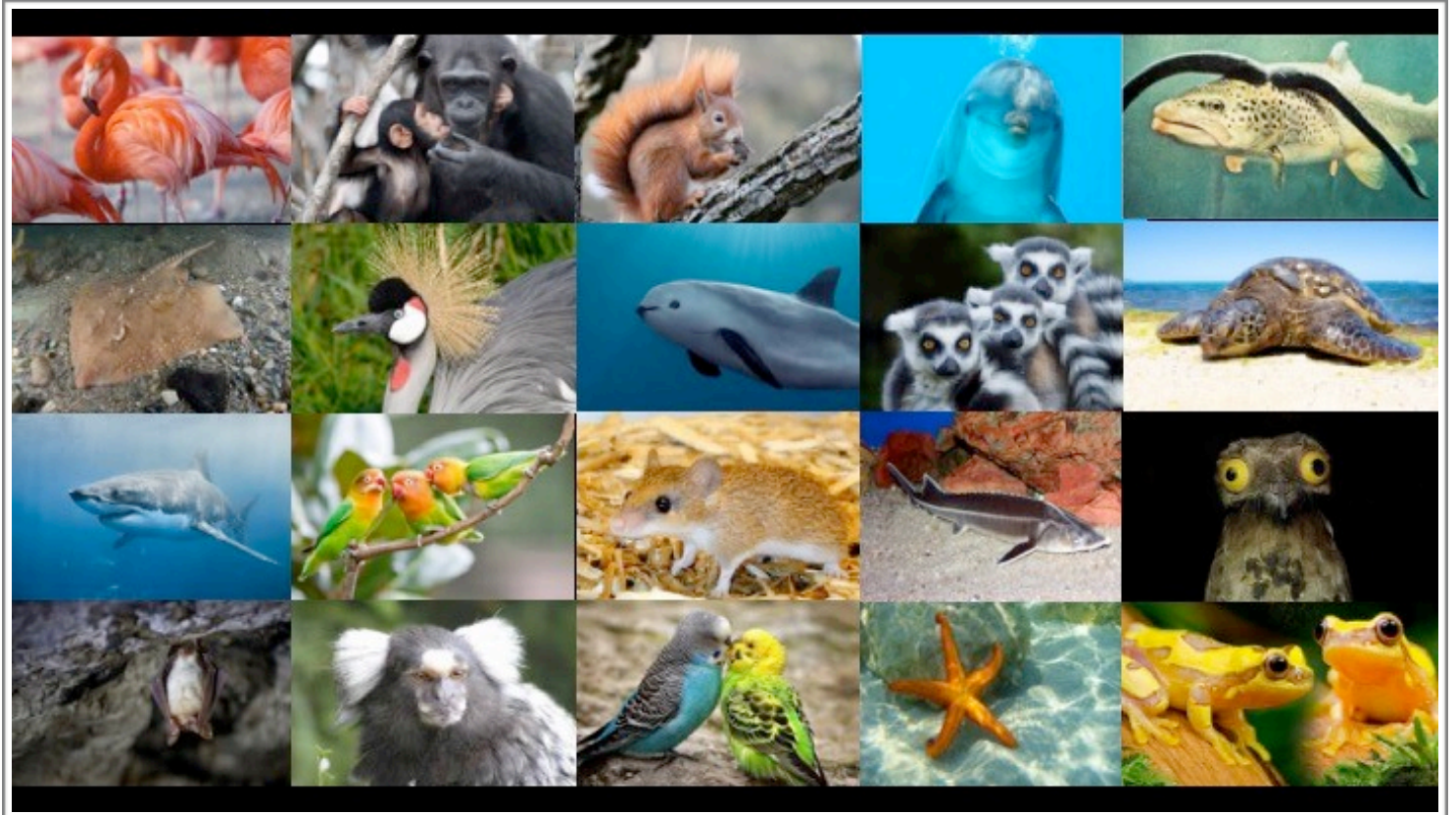


The G10K-VGP/EBP 2019 Meeting Agenda



***Advancing the missions of the Vertebrate Genomes Project,
Earth Biogenome Project, and other genome projects***

Tuesday - Friday, 27-30 August 2019

The Rockefeller University, Manhattan, NY, USA

Table of Contents

Conference Information

| | |
|-------------------------------------|---|
| • Conference organizers..... | 7 |
| • Goals..... | 7 |
| • Location..... | 8 |
| • Zoom Link, Official Sessions..... | 9 |

Overall Schedule

Tuesday, 27 August 2019

| | |
|--|----|
| • Invite Only Press Conference..... | 10 |
| • Registration and Lunch..... | 10 |
| • Welcome and Opening Remarks..... | 10 |
| • Plenary Lecture: Can biodiversity genomics save the world? The reality of conservation amidst the sixth mass extinction..... | 10 |
| • Opening General Session..... | 10 |
| • Status of VGP Phase 1 and Coordination | |
| • Diverse species present diverse challenges: Towards assembling gapless vertebrate genomes | |
| • VPG Tools and Formats | |
| • Annotating the VGP: The story so far | |
| • Reference quality genomes of six bats illuminate the genomics determinants and evolution of unique adaptation in bats | |
| • Sequencing the Green Branch of the Tree of Life | |
| • Darwin Tree of Life | |
| • Welcome Reception (sponsored by Arima Genomics)..... | 11 |

Wednesday, 28 August 2019

- Breakfast (sponsored by DNANexus).....12
- Concurrent Sessions.....12
 - VGP Assembly
 - Comparative Genomics
 - B10K
 - Insects
- Lunch (sponsored by Bionano Genomics).....12
- General Session Tech Talk: Bionano optical mapping for accurate genome assembly, comparative genomics, and haplotype segregation.....13
- Official Poster Session: Genome Assemblies Section and Big Projects Section.....13
- Concurrent Sessions.....13
 - VGP Annotation, Alignment and Data Coordination
 - VGP Sample Prep
 - Mammals
- Summaries, General Session.....13
- G10K Council Meeting (by invitation only).....14

Thursday, 29 August 2019

- Breakfast (sponsored by Dovetail Genomics).....15
- Concurrent Sessions.....15
 - VGP Assembly
 - Conservation Genomics
 - Bat 1K
 - GIGA
- Lunch (sponsored by Pacific Biosciences).....15
- General Session Tech Talk: New Capabilities of the PacBio Sequel II Sequencing System.....16
- Official Poster Session: Methods Section and Biological Discoveries Section.....16

(Thursday continued)

- Concurrent Sessions.....16
 - VGP Funding and Publications
 - Microbial Eukaryotes
- Summaries, General Session.....16
- Networking Reception (sponsored Oxford Nanopore Technologies).....16

Friday, 30 August 2019

- Breakfast.....17
- EBP Status and Future.....17
- National/Regional Projects.....17
 - Darwin Tree of Life
 - BRIDGE Colombia
 - Oz Mammals
 - EBP China/10XP
 - California Conservation Genomics Project
 - Genomics in the Anthropocene - Unlocking the Value of Smithsonian Science and Museum Collections
- Other Ongoing and Planned Projects/Discussion.....18
 - Catalonia Biodiversity Project
 - The Genome Alliance in Australia
 - 1000 Chilean Genomes
 - Swedish EBP Initiative
 - Taiwan BioGenomes Project
 - Indian Initiative on Earth Biogenome Sequencing
 - CanSeq150
- Lunch (sponsored by Illumina).....18
- General Session Tech Talk: Impact of Next-Generation Sequencing on Biology.....18
- EBP Working Group Meeting (by invitation only).....18
- ELSI Workshop.....18
- Summaries, General Session.....18

(Friday continued)

- Closing General Session: New Science Enabled by the VGP, EBP, other topics, and plans going forward.....19

Plenary Lecture (details).....20

Concurrent Sessions Abstracts and Programs

Wednesday morning:

- VGP Assembly.....22
- Comparative Genomics.....24
- B10K.....25
- Insects.....25

Wednesday afternoon:

- VGP Annotation, Alignment and Data Coordination.....27
- VGP Sample Prep.....29
- Mammals.....30

Thursday morning:

- VGP Assembly.....31
- Conservation Genomics.....31
- Bat 1K.....32
- GIGA.....33

Thursday afternoon:

- VGP Funding and Publications.....34
- Microbial Eukaryotes.....34

General Sessions, Tech Talks (details)

- Bionano optical mapping for accurate genome assembly, comparative genomics, and haplotype segregation.....35
- New Capabilities of the PacBio Sequel II Sequencing System.....36
- Impact of Next-Generation Sequencing on Biology (Illumina).....37

| | |
|--|----|
| One Tree, One Planet (art installation sponsored by Naziha Mestaoui, Douglas Soltis, Pamela Soltis, Robert Guralnick, Matt Gitzendanner, James Rosindell and Yan Wong)..... | 38 |
|--|----|

| | |
|----------------------------------|----|
| Posters At-A-Glance | 42 |
|----------------------------------|----|

Poster Abstracts

Wednesday:

- Genome Assemblies.....47
- Big Projects.....51

Thursday:

- Methods.....53
- Biological Discoveries.....58

| | |
|---------------------------|----|
| Social Media | 63 |
|---------------------------|----|

| | |
|-----------------------|----|
| Sponsors | 64 |
|-----------------------|----|

Conference Information

Conference Organizers

The Rockefeller University:

- Erich D. Jarvis, Ph.D.
- Sadye Paez, Ph.D., PT, MPH
- Lauren Shalmiyev, MPH

University of California, Davis

- Nicolette Caperello
- Harris Lewin, Ph.D.
- Stephen Richards, Ph.D.

Goals

The Genome 10K (G10K) Community of Scientists is a [consortium](#) of leading scientists representing research centers, zoos, museums, and academic institutions from around the world and is dedicated to coordinating international efforts for large-scale sequencing and analyses projects since 2009. The G10K is currently focusing on the [VGP](#) (Twitter @genomeark), whose goal is to generate near error-free and complete reference genome assemblies of approximately all 70,000 extant vertebrate species to address fundamental questions in biology, disease and conservation. The number of species has been revised upwards from 66,000 due to updates in species identification and classifications. These genome assemblies are being made publicly available via our [Genome Ark GitHub](#) and annotated in the NCBI and Ensembl databases. At last year's meeting, [we released 15 reference genome representing 14 species](#) in 13 vertebrate orders with each genome meeting our VGP 3.4.2.QV40 phased metric. At this year's meeting, we announce 100 new high-quality genomes, representing an additional 77 orders towards completing the ~260 orders of the VGP Phase 1.

The EBP is a consortium of independent partner institutions that have the common goal of sequencing and annotating the genomes of all 1.5 million named eukaryotic species (animals, plants, fungi, protozoa and other microbes). Since its official launch in November 2018 in London, there are [25 member institutions](#) and more than [20 large-scale projects](#) affiliated with EBP, including the VGP. The EBP is expected to create a new foundation for biology to drive solutions for preserving biodiversity and sustaining human societies.

Overarching G10K-VGP/EBP meeting aims:

1. Advance the missions of the VGP, EBP, and other large-scale genome projects;
2. Announce progress towards completing Phase 1 of the VGP;
3. Announce progress of the EBP and related projects;
4. Disseminate best current practices from the VGP and other projects for genome assemblies, curation, annotation, and biological analyses, while strategizing for further improvements in productivity, efficiency and quality; and
5. Create opportunities for collaboration, integration, cross-fertilization of ideas, and execution of plans for completing large-scale genome projects.

Location

[The Rockefeller University Campus](#)

Zoom Link: Official Sessions

Zoom link for all official sessions: <https://rocku.zoom.us/j/9048498644>

- Tuesday
 - 1:00-6:00pm: Welcome, Plenary lecture, Opening Session
- Wednesday
 - 12:30-1:30pm: Tech Talk by Bionano; and
 - 5:00-6:00pm: Summaries from B10K, Mammals, Insects, Comparative Genomics, and Assembly
- Thursday
 - 12:30-1:30pm: Tech Talk by Pacific Biosciences; and
 - 4:15-6:00: Summaries from Assembly, Alignment and Annotation, Conservation Genomics, and Sample Prep; Discussion on methodologies and coordination across national, regional, and taxon-specific projects
- Friday
 - 9:00-12:00pm: Morning session;
 - 12:30-1:30pm: Tech Talk by Illumina; and
 - 2:00-6:00pm: Afternoon session.

Meeting ID: 904 849 8644

One tap mobile

+16468769923,,9048498644# US (New York)

+16699006833,,9048498644# US (San Jose)

Dial by your location

+1 646 876 9923 US (New York)

+1 669 900 6833 US (San Jose)

Find your local number: <https://zoom.us/j/abOOQsm2Y6>

Join by SIP 9048498644@zoomcrc.com

Join by H.323

162.255.37.11 (US West)

162.255.36.11 (US East)

221.122.88.195 (China)

115.114.131.7 (India)

213.19.144.110 (EMEA)

103.122.166.55 (Australia)

209.9.211.110 (Hong Kong)

64.211.144.160 (Brazil)

69.174.57.160 (Canada)

207.226.132.110 (Japan)

Tuesday, 27 August 2019

Press Briefing

10:00-12:00 **Invite Only Press Conference**

G10K-VGP Program Director: Sadye Paez, Rockefeller University, NY, USA

(*meeting attendees may view a livestream in Carson Auditorium)

12:00-1:00 **Registration and Lunch**

Greenberg Atrium, Floor B, Collaborative Research Center

Opening General Session

Carson Family Auditorium, Floor B, Collaborative Research Center

1:00-1:05 **Welcome**

Erich D. Jarvis, Rockefeller University, NY, USA

Harris Lewin, University of California, Davis, CA, USA

1:05-1:15 **Opening Remarks**

Richard Lifton, President, Rockefeller University, NY, USA

1:15-2:15 **Plenary Lecture: Can biodiversity genomics save the world? The reality of conservation amidst the sixth mass extinction**

Rebecca Johnson, Director, Australian Museum Research Institute

2:00-6:00 **Opening General Session**

2:15-2:45 **Status of VGP Phase 1 and coordination**

Erich D. Jarvis, Rockefeller University, NY, USA

- 2:45-3:15 **Diverse species present diverse challenges: Towards assembling gapless vertebrate genomes**
Adam Phillippy, NHGRI
- 3:15-3:45 **VGP Tools and Format**
Richard Durbin, University of Cambridge and Gene Myers, MPI Dresden
- 3:45-4:00 Break: tea, coffee, and snacks provided with registration
Greenberg Atrium, Floor B, Collaborative Research Center
- 4:00-4:30 **Annotating the VGP: The story so far**
Fergal Martin, EBI, UK and Françoise Thibaud-Nissen, NCBI, D.C.
- 4:30-5:00 **Reference quality genomes of six bats illuminate the genomic determinants and evolution of unique adaptations in bats**
Michael Hiller, MPI Dresden
- 5:00-5:30 **Sequencing the Green Branch of the Tree of Life**
Pam Soltis, University of Florida
- 5:30-6:00 **Darwin Tree of Life**
Mark Blaxter, Wellcome Sanger Institute
- 6:00-8:00 Welcome Reception (sponsored by Arima Genomics)
Live Music: Carlitos Padron y Rumberos del Callejón (*Facebook @rumberosdelcallejon and Instagram @rumberos_del_callejon*)
Anna-Maria and Stephen Kellen BioLink

Wednesday, 28 August 2019

8:00-9:00 Registration and Breakfast (sponsored by DNANexus)
Greenberg Atrium, Floor B, Collaborative Research Center

9:00-12:00 **Concurrent Sessions**

VGP Assembly

Chair: Adam Phillippy, NHGRI

Location: Collaborative Research Center 506

Comparative Genomics

Chair(s): Joana Damas, UC Davis and Morgan Wirthlin, Carnegie Mellon University

Location: Collaborative Research Center 406

B10K

Chair: Josefin Stiller, University of Copenhagen

Location: Collaborative Research Center 306

Insects

Chair: Kevin Hackett, USDA

Location: Collaborative Research Center 206

10:30-11:00 Break: tea, coffee, and snacks provided with registration
Greenberg Atrium, Floor B, Collaborative Research Center

12:00-2:00 Lunch (sponsored by Bionano Genomics)
Greenberg Atrium, Floor B, Collaborative Research Center

- 12:30-1:30 **General Session Tech Talk:** Bionano optical mapping for accurate genome assembly, comparative genomics, and haplotype segregation
 Alex Hastie, Director of Customer Solutions, Bionano Genomics
 Carson Family Auditorium, Floor B, Collaborative Research Center
- 1:30-2:00 **Official Poster Session:** Genome Assemblies Section and Big Projects Section
 Greenberg Atrium, Floor B, Collaborative Research Center
- 2:00-5:00 **Concurrent Sessions**
VGP Annotation, Alignment and Data Coordination
Chair(s): Benedict Paten, UCSC, Françoise Thibaud-Nissen, NCBI, and Fergal Martin, EBI
Location: Collaborative Research Center 506
- VGP Sample Prep**
Chair(s): Jacquelyn Mountcastle, Rockefeller University and Sylke Winkler, MPI Dresden
Location: Collaborative Research Center 406
- Mammals**
Chair(s): Kathy Belov, University of Sydney; Rebecca Johnson, Australian Museum Research Institute; Elinor Karlsson and Kerstin Lindblad-Toh, Broad Institute, MIT
Location: Collaborative Research Center 306
- 4:00-4:30 Break: tea, coffee, and snacks provided with registration
 Greenberg Atrium, Floor B, Collaborative Research Center
- 5:00-6:00 **Summaries, General Session:** B10K, Mammals, Insects, Comparative Genomics, and Assembly
 Carson Family Auditorium, Floor B, Collaborative Research Center

7:00-9:00 **G10K Council Meeting** (by invitation only)

Thursday, 29 August 2019

8:00-9:00 Registration and Breakfast (sponsored by Dovetail Genomics)
Greenberg Atrium, Floor B, Collaborative Research Center

9:00-12:00 **Concurrent Sessions**

VGP Assembly

Chair: Adam Phillippy, NHGRI

Location: Collaborative Research Center 506

Conservation Genomics

Chair(s): Warren Johnson, Smithsonian; Oliver Ryder and Cynthia Steiner, San Diego Zoo Institute for Conservation Research

Location: Collaborative Research Center 406

Bat 1K

Chair(s): Liliana Davalos, Stony Brook University and Michael Hiller, MPI Dresden

Location: Collaborative Research Center 206

GIGA

Chair(s): Jose Lopez, Nova Southeastern University Ocean Center; Monica Medina, Penn State University; Adelaide Rhodes, Broad Institute MIT; Agosthino Antunes, University of Porto.

Location: Collaborative Research Center 106

10:30-11:00 Break: tea, coffee, and snacks provided with registration
Greenberg Atrium, Floor B, Collaborative Research Center

12:00-2:00 Lunch (sponsored by Pacific Biosciences)
Greenberg Atrium, Floor B, Collaborative Research Center

- 12:30-1:30 **General Session Tech Talk:** New Capabilities of the PacBio Sequel II Sequencing System
Jonas Korlach, CSO, Pacific Biosciences
Carson Family Auditorium, Floor B, Collaborative Research Center
- 1:30-2:00 **Official Poster Session:** Methods Section and Biological Discoveries Section
Greenberg Atrium, Floor B, Collaborative Research Center
- 2:00-3:30 **Concurrent Sessions**
VGP Funding and Publications
Chair: Erich D. Jarvis, Rockefeller University
Location: Carson Family Auditorium, Floor B, Collaborative Research Center
- Microbial Eukaryotes**
Chair: Neil Hall, Earlham Institute
Location: Collaborative Research Center 306
- 3:45-4:00 Break: tea, coffee, and snacks provided with registration
Anna-Maria and Stephen Kellen BioLink
- 4:00-5:00 **Summaries, General Session:** Assembly, Alignment and Annotation, Conservation Genomics, and Sample Prep
- 5:00-6:00 **Discussion on methodologies and coordination across national, regional, and taxon-specific projects**
- 6:00-8:00 Networking Reception (sponsored by Oxford Nanopore Technologies)
Art and Science Installation: One Tree, One Planet
Live Music: The Torosjan Trio
Anna-Maria and Stephen Kellen BioLink

Friday, 30 August 2019

8:00-9:00 Registration and Breakfast (sponsored by Pacific Biosciences)
Greenberg Atrium, Floor B, Collaborative Research Center

Morning Session

Carson Family Auditorium, Floor B, Collaborative Research Center

9:00-9:30 **EBP Status and Future**
Harris Lewin, UC Davis, CA, USA

National/Regional Projects

9:30-9:50 **Darwin Tree of Life UK-EBP**
Mark Blaxter, Wellcome Sanger Institute

9:50-10:10 **BRIDGE Colombia**
Federica DiPalma, Earlham Institute

10:10-10:30 **Oz Mammals**
Rebecca Johnson, Australian Museum Research Institute

10:15-10:45 Break: tea, coffee, and snacks provided with registration
Greenberg Atrium, Floor B, Collaborative Research Center

10:30-10:50 **EBP China/10KP**
Xin Liu, BGI Shenzhen

10:50-11:10 **California Conservation Genomics Projects**
H. Bradley Shaffer, UCLA

- 11:10-11:30 **Genomics in the Anthropocene - Unlocking the Value of Smithsonian Science and Museum Collections**
W. John Kress, Smithsonian Institute
- 11:30-12:00 **Other Planned and Ongoing Projects**
- Catalonia Biodiversity Project: Montserrat Corominas, Societat Catalana de Biologia and University of Barcelona
 - The Genome Alliance in Australasia: David Burt, University of Queensland
 - 1000 Chilean Genomes: Miguel Allende, University of Chile
 - Swedish EBP Initiative: Kerstin Lindblad-Toh, Broad Institute MIT
 - Taiwan BioGenomes Project: Shu-Miaw Chaw, Academia Sinica
 - Indian Initiative on Earth Biogenome Sequencing: Saloni Mathur, National Institute of Plant Genome Research, New Delhi, India
 - CanSeq150: Steven Jones, Genome Sciences Centre, BC Cancer
- 12:00-2:00 Lunch (sponsored by Illumina)
Greenberg Atrium, Floor B, Collaborative Research Center
- 12:30-1:30 **General Session Tech Talk:** Impact of Next-Generation Sequencing on Biology
Gary P. Schroth, Distinguished Scientist, Illumina
Carson Family Auditorium, Floor B, Collaborative Research Center
- 12:00-2:00 **EBP Working Group Meeting** (by invitation only)
- 2:00-3:30 **ELSI Workshop**
Melissa Goldstein, Milken Institute School of Public Health
Anna-Maria and Stephen Kellen BioLink
- 3:30-4:00 **Summaries, General Session:** Bat 1K, GIGA, and Microbial Eukaryotes

Anna-Maria and Stephen Kellen BioLink

4:00-4:15 Break: tea, coffee, and snacks provided with registration

4:15-6: **Closing General Session**

New Science Enabled by the VGP, EBP, other topics, and plans going forward

Anna-Maria and Stephen Kellen BioLink

Plenary Lecture

Can biodiversity genomics save the world? The reality of conservation amidst the 6th mass extinction

Rebecca Johnson

Twitter @DrRebeccaJ | Instagram @drrebeccaj

E rebecca.johnson@austmus.gov.au |

<http://australianmuseum.net.au/staff/rebecca-johnson> |

www.linkedin.com/in/dr-rebecca-johnson



Abstract

In May 2019 the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) group released their global assessment report outlining the unprecedented decline of the Earth's natural life support systems. The IPBES report estimated that 1 million species are threatened with extinction directly attributable to anthropogenic activity such as land clearing, climate change, and environmental crime. In one view, this report is merely the latest to represent such sobering expert views regarding the future of global biodiversity – and by extension human and ecosystem wellbeing. However, the IPBES report goes beyond other published works, by including sobering metrics supporting the assertion of a 6th mass extinction event, to directly show that biodiversity loss is a result of human impacts and it is anticipated that global threat levels will only increase in coming decades.

Genomic tools have precipitated one of the great revolutions in biological sciences, and its connection to biodiversity conservation has long been obvious for scientists from the genetics and genomics communities. It is worth asking, however, as genomics has become more accessible, cheaper, and faster, whether these rapid advances have actually mitigated the planetary biodiversity crisis. In other words, is the threat of the 6th extinction too large for genomics to make a worthwhile contribution? I aim to demonstrate actual and potential benefits of biodiversity genomics through specific examples including the work of the koala genome consortium, a conservation genomics project that has revealed important information about koala adaptations and population genomics while also having impact in government policy in managing koalas.

There are still good reasons to be optimistic. The global scientific community has coalesced around G10K (sequencing vertebrate genomes) and the Earth Biogenome Projects, which offer a form of future-proofing through genomic vouchers and datasets. The increasing sophistication and power of genomic tools may compel scientists to ask the best questions for their use. Downstream of basic research, it is also clear we need to communicate better, beyond our community, and make the case at fora outside of academia, where we can best showcase the translation of conservation genomics research in best practises, recommendations, or real leadership for science in society.

Biography

Professor Rebecca Johnson is Director of the Australian Museum Research Institute, a wildlife forensic scientist, conservation geneticist and chief investigator of the Koala Genome Consortium. Rebecca is a member of the Australian Academy of Forensic Sciences; NSW Branch President of the Australian & New Zealand Forensic Science Society; and director of Membership & Outreach for the Society for Wildlife Forensic Science.

With an honours degree from the University of Sydney and PhD from La Trobe University Melbourne in the field of molecular evolutionary genetics and has worked as a molecular geneticist, in Australia and the USA before joining the Museum in 2003. She established the Museum as one of the global leaders in the field of wildlife forensics and conservation genomics through the ISO17025 accreditation of the Australian Centre for Wildlife Genomics facilities (one of only six such laboratories globally).

In April 2015, she became director of the Australian Museum Research Institute (the first female science director in the Museum's 192 year history). She is one of 32 individuals certified as a wildlife forensic scientist globally, and is one of only two experts appointed by the Federal Environment Minister as an examiner in wildlife forensics under the Commonwealth legislation. She is an adjunct Professor at the University of Sydney and a conjoint Professor at the University of New South Wales.

In July 2017 Rebecca was named one of the 30 inaugural "SuperStars of STEM" by Science and Technology Australia. She was awarded the 2016 University of Sydney, Faculty of Science Alumni Award for Professional Achievement and in September 2016 was also announced as one of The Australian Financial Review and Westpac "100 Women of Influence" in the Innovation category. Rebecca has also received a Chief Executive Women (CEW) scholarship to attend the INSEAD business school for executive leadership and in 2018 was recognised as a "Vogue 2018 Game Changer" in the Tech & Innovation category; named one of CEO Magazine's '10 leading business women in Australia' for 2018; awarded the 2018 Eureka Prize AMRI medal recognising research excellence; and named one of Harper's Bazaar's "Woman of the Year" for 2018.

Rebecca is passionate about conservation, reducing the illegal wildlife trade and the importance of STEM education in contributing to positive environmental outcomes.

Concurrent Session Abstracts and Programs

***Wednesday, 28 August 2019: Morning Program
9:00-12:00 pm***

VGP Assembly: Retrospective

Chair: Adam Phillippy

Location: CRC 506

This 1st assembly session will be a retrospective on progress and remaining challenges faced by the VGP assembly group, recently partnered with the EBP, towards achieving high-quality de-novo genome assemblies that meet the VGP metric and beyond for complete and error free genomes. All conference participants, studying all types of organisms, are welcomed.

9:00 / **Introduction:** Adam Phillippy, NHGRI

Day 1 format: retrospective talks and discussion

9:10 / **Curation status and lessons learned:** Kerstin Howe, Sanger

9:30 / **Heterozygosity and haplotig purging:** Shane McCarthy, Sanger

9:50 / **Polishing and trios:** Arang Rhie, NHGRI

10:10 / **Relative contributions of technology:** Harris Lewin, UC Davis

10:20 / **Summary discussion**

10:30 / Break

11:00 / **Mitochondrial genome assembly:** Giulio Formenti, Rockefeller

11:20 / **Nanopore / SHASTA:** Benedict Paten, UCSC

11:40 / **Crowdsourcing and training:** Marcela Uliano da Silva, Berlin Center for Genomics

11:50 / **Summary discussion**

Running list of challenges

How can we scale?

- Goal of 6 finished and validated assemblies per week

How can the DNAnexus pipeline be improved?

- PreQC, validation?

- Defined inputs/outputs for each stage

How should we contig diploid genomes?

- PacBio or Nanopore?

- Better haplotype separation prior to contigging?

- Fix contig phase issues prior to scaffolding?

- Assemble phased unitigs rather than pseudo-haplotypes?

How should we scaffold diploid genomes?

- Reorder or drop particular technologies?

- Stick with iterative scaffolding? (i.e. each data type added in succession)

- Switch to integrated scaffolding? (i.e. all data types considered at once)

How can we automatically validate the genomes?

- Develop an agreed feature set for validation

- Automatically map 10X, BioNano, Hi-C data to final assemblies

- Telomere and centromere annotation

- Hi-C maps automatically generated and browseable

How can we cultivate new methods for long-read, diploid, phased assembly?

- Provide formatted reads, contigs, scaffold linkage information

- How would we evaluate potential alternative assemblies?

How should we format the final assemblies?

- Two haplotypes, plus a maximally continuous pseudo haplotype?

What are our plans for emerging technologies?

- PromethION sequencer (100 Gb+ throughput per flow cell)

Plans for better cooperation

- Interoperability of tools

- Better agility

- More in-person meetings?

Revisit minimum VGP assembly metrics

- Definition of “chromosome-scale”?

- What kind of minimum support is necessary

Comparative genomics

Chairs: Morgan Wirthlin, Joana Damas

Location: CRC 406

The comparative genomics breakout session will present some of the exciting biological findings from ongoing multispecies vertebrate genome projects thus far, as well as plans for proposed comparative studies to be conducted with high-quality genomes currently being generated by multiple large-scale, international genome sequencing efforts. Talks by leading experts and emerging pioneers will emphasize the broader lessons learned for meeting the challenges of performing comparative genomics in the massively high-throughput era. The format of this session will be short presentations (15 minutes), followed by questions (5 minutes) and a guided open discussion at the end. The session participants include G10K-VGP members conducting multispecies genome analysis, such as the learning and language group, those conducting phylogeny, genome evolution, and trait analyses studies. This session is open to all conference attendees.

Program:

- 9:00 **Welcome and introduction from the chairs**
- 9:10 **Scaling up to comparative genomics with 240 mammals**
Elinor Karlsson, Broad Institute MIT
- 9:30 **Investigating convergent molecular specializations across vocal learning taxa requires high quality assemblies and annotations**
Greg Gedman, Rockefeller University
- 9:50 **Vocal learning in the era of massively high-throughput genomics: a case study on new approaches for relating genomes to complex phenotypes**
Morgan Wirthlin, Carnegie Mellon University
- 10:10 **Coffee break**
- 10:30 **An enduring frontier: early results and future directions of comparative genomics in squamate reptiles**
Daren Card, Harvard University
- 10:50 **Evaluating genome assemblies using comparative genomics tools**
Joana Damas, University of California Davis
- 11:10 **Chromosomes – much more than DNA**
Jennifer Graves, La Trobe University
- 11:20 **Aiming for chromosomal DNA sequences**
Richard Durbin, Wellcome Sanger Institute
- 11:30 **Discussion: Imagining the Future of Comparative Genomics**
- What are the biggest challenges currently faced in comparative genomics and what strategies are needed to solve them?
 - What types of genetic differences can we look for to associate with traits in comparative studies?
 - What can we do with the VGP genomes (goals, projects)?
 - What opportunities exist for forming new collaborative comparative genomics working groups?

B10K

Chair: Josefin Stiller

Location: CRC 306

<https://global.gotomeeting.com/join/473983685>

The B10K breakout session will focus progress made with and plans for B10K bird genomes, including the draft genomes of the 360+ family level species, and the B10K reference genomes of VGP Phase 1.

9-10.30 Talks

- **VGP assembly of Yellow-throated sandgrouse:** Chul Lee, Seoul National University (Lightning talk, 5 min).
- **New models of protein sequence evolution:** Edward Braun, University of New Mexico (15 min).
- **Large-scale multi-locus species tree estimation using divide-and-conquer:** Tandy Warnow, University of Illinois -Urbana (15 min).
- **Towards scalable phylogenetic network inference:** Luay Nakhleh, Rice University (15 min).
- **Genome-wide species tree estimation: challenges and recent advances:** Siavash Mirarab, UCSC (15 min).

10.30- 11.45 Break

10.45-11.00: **Demonstration of the avian immunome database:** Ralf C. Mueller MPI Ornithology (Lightning talk, 5 min).

11.00-12.00 **Discussion**

Insects

Chairs: Kevin Hackett, Anna Childers, Monica Poelchau, Susan Brown

Location: CRC 206

Insect Zoom Link

<https://zoom.us/meeting/register/6036afd9af5a0a9a7510d14dfea9e911>

One tap mobile

+19294362866,,376132589# US (New York)

+16699006833,,376132589# US (San Jose)

Dial by your location

+1 929 436 2866 US (New York)

+1 669 900 6833 US (San Jose)

Meeting ID: 376 132 589

Find your local number: <https://zoom.us/j/ad3lbayM1j>

The 'Insects' breakout session will be an interactive meeting on arthropod-specific genome project issues. In this session, we aim to: start a conversation on how the arthropod research community at large can leverage momentum from the Earth BioGenome Project; discuss funding models for arthropod genome sequencing; and identify arthropod-specific solutions to technical problems that arise during genome projects. Presenters will direct and invite

discussions on each of these topics. We encourage anyone interested in sequencing insect or arthropod genomes and contributing to the EBP project to attend.

BACKGROUND SESSION. Chair: Monica Poelchau (USDA-ARS)

-9:00 am: **Welcome** (Kevin Hackett, USDA-ARS)

-9:05 am: **Motivation for sequencing insects** (Stephen Richards, UC Davis)

- Rationale for sequencing insects.

-9:20 am: **Introduction of participants** (Monica Poelchau, USDA-ARS)

- All session attendees will briefly answer a set of questions, provided by the chairs, about the genome projects that they are involved in.

-9:35 am: **Funding discussion** (Stephen Richards, UC Davis)

- Funding models for arthropod genome sequencing.

TECHNICAL SESSION. Chair: Susan Brown (Kansas State University)

-10:00 am: **Infrastructure – vouchersing** (Jon Coddington, Smithsonian Institution)

- Concepts of vouchersing and discussion of vouchersing best practices.

-10:15 am: **Infrastructure – databases** (Monica Poelchau, USDA-ARS; Christopher Childers, USDA-ARS)

- Introduce the concept of genome databases, and explain why they are important for data management and the Earth BioGenome Project.

-10:30 am: Coffee Break

-10:45 am: **Extraction, sequencing and assembly issues for insects** (Scott Geib, USDA-ARS; Brian Scheffler, USDA-ARS; Anna Childers, USDA-ARS)

- Present and discuss several topics relevant to insect DNA/RNA extraction, sequencing, and assembly, including but not limited to: low input methods; inbreeding strategies; sequencing and assembly strategies appropriate for the diversity of arthropods (ex: small vs. large body size, small vs. large genome size, haploid vs. diploid, short vs. long generation time). Unique solutions for unique challenges.
- Discussion of assembly standards appropriate for arthropods.

-11:30 am: **Coordination of sequencing projects.** (Anna Childers, USDA-ARS; Monica Poelchau, USDA-ARS)

- Broach the topic of how to coordinate distributed genome sequencing projects across Arthropoda.

-11:45 am: **General discussion/wrap-up.**

- Discussion of any outstanding issues for arthropod genome sequencing.

Wednesday, 28 August 2019: Afternoon program 2:00-5:00pm

VGP Annotation, Alignment and Data Coordination Breakout

Chairs: Fergal Martin, Françoise Thibaud-Nissen, Benedict Paten

Location: CRC 506

In order to facilitate downstream science it is crucial that we have scaleable and consistent results in terms of both gene annotation and genomic alignments. Furthermore, annotation and the distribution of annotation, whether done de novo or via whole-genome alignments, require careful data coordination, as it sits at a mid point in the data chain, between raw data and data analyses. In this breakout session we will discuss a variety of topics related to the above including: strategies on how to maximise the distance in cross-species transcriptome annotation; how to assign gene function across the vertebrate clade; assigning confidence levels to genes and transcripts; how to deal with hard to annotate edge cases; and how genomic alignments can be utilised for annotation of closely related species/sequences (e.g. breeds, strains, haplotypes etc.). We will conclude with a discussion on how to coordinate across the VGP data chain including: challenges emerging from coordination of Phase 1; centralised versus decentralised data coordination; and FAANG as an example for implementing data standards.

Gene annotation (2-3:30pm)

2:00-2:15: **Invited talk: A universal vertebrate gene nomenclature: a case study of the oxytocin/vasotocin ligand and receptor family.** Constantina Theofanopoulou, University of Barcelona and Rockefeller University (presenting online)

Open discussions:

- Assigning function in a scalable manner
- Understanding/maximising the distance we can use transcriptomic data
- Transcript confidence levels and building a vertebrate transcript library
- Examples of difficult annotations, how to do them and what we can learn

Break (3:30-3:45pm)

Genomic alignments and annotation of closely related things (3:45-4:30pm)

3:45-4:00: **Invited talk: False gene losses and gains corrected by VGP assemblies.** Juwan Kim (presenting), Chul Lee, and Byung June Ko, Seoul National University.

4:00-4:15: **Invited talk: Precise annotation of Evolutionary Breakpoint Regions and Homologous Synteny Blocks with multi-species genome alignment using "Syntenic Orthology Algorithm" to reconstruct genome rearrangements in Mammals.** Steve O'Brien Dobzhansky Center

Open discussions:

- Discussion about whole genome alignment-based annotation methods

- Creating new reference standard annotations in the absence of manual curation
- The cost/benefit ratio and use cases for annotating haplotypes/strains/breeds/sexes

Data coordination (4:30-5:00pm)

Open discussions:

- Emerging challenges from the current state of coordination in the VGP
- Centralised versus decentralised coordination
- Examples of successful data coordination efforts

VGP Sample Prep

Chairs: Jacquelyn Mountcastle and Sylke Winkler

Location: CRC 460

The VGP Sample Prep breakout session will focus on lessons learned and plans going forward for generating high quality molecular weight and pure DNA for generating reference-grade genome assemblies, permits for obtaining and transporting tissues among participants of the project, and ethics. All conference attendees are welcome.

| Time | Topic | Discussion Leaders |
|---|--|--|
| <u>VGP Sample Prep Group</u> | | |
| 2:00-2:40 | Tissue preservation manuscript update: preliminary data | Jacquelyn Mountcastle Rockefeller University |
| | Sample prep requirements for Hi-C | Anthony Schmitt Vice President, Arima Genomics, Inc |
| | Circulomics: New isolation protocols for a variety of sample types | Kelvin Liu CEO and Founder of Circulomics, Inc |
| | Sample quantity: How low can we go? | Jen Balacco Rockefeller University |
| 2:40–3:10 | <i>Group Discussion 30 min</i> | |
| <u>“Beyond Vertebrates” – Darwin Project/EBP</u> | | |
| 3:10–3:30 | DNA quality and purity over diverse species, and experiences from the SciLife Lab | Olga Pettersson Uppsala University |
| 3:30-3:50 | <i>Group Discussion 20 min</i> | |
| <u>Sample Collection Challenges, Permits, and Ethics</u> | | |
| 3:50-4:00 | Best practices for communication and ethics for sampling in the field | Andrew Crawford Universidad de los Andes |
| 4:00-4:15 | <i>Coffee Break</i> | Bob Murphy University of Toronto |
| 4:15– 4:30 | Permits for import and export in US and Europe: CITES/USFWS, Nagoya, etc. | Sylke Winkler MPI-CBG/ Camila Mazzoni Berlin Center for Genomics in Biodiversity Research |
| 4:30 - 5:00 | <i>Group Discussion 30 min</i> | |

Mammals

Chairs: Kathy Belov, Rebecca Johnson, Elinor Karlsson, Kerstin Lindblad-Toh

Location: CRC 306

The focus of the mammal session will be to summarize the current status of mammal genomics globally, including lessons learned from single taxon through to multispecies and higher level (subclass to family level) mammalian genome sequencing projects. Specific initiatives covered in this session will include the 200 mammals and Australian mammals projects along with results and updates from individual genome consortia. The format of this session will be brief 20 minute presentations (15 minute + 5 minute for questions), with questions and open discussions. The purpose of this session is to share findings of current sequencing initiatives but also provides the opportunity for the G10K-VGP community to participate in a global horizon scan of mammal genomics (excluding the Bat1K initiative which will have its' own session).

Program:

2:00-2:10 – **Welcome and introduction** -Rebecca Johnson (Australian Museum)

2.10-2.30 – **Marsupial genomes for conservation and drug development** – Kathy Belov (University of Sydney)

2.30-2.50 – **Centromeres in wallabies and gibbons** – Rachel O'Neil (University of Connecticut)

2.50-3.10 - **Aquatic adaptation & fur trade devastation: a deep dive into the genomes of the sea otter and giant otter** – Anabel Beichman (UCLA)

3.10-3.30 - **The evolutionary journey to gigantism: Capybara genome reveals the complex evolution of extreme body size** - Santiago Herrera Alvarez (Universidad de los Andes)

3.30-3.50 QUICK COFFEE BREAK

3.50-4.10 – **Mammals of Colombia** – Federica di Palma (Earlham Institute)

4.10-4.30 - **Reconstruction of the mammalian ancestral karyotype** - Joana Damas (University of California Davis)

4.30-4.50 – **200 mammals project** – Kerstin Lindblad-Toh (Uppsala University and Broad Institute)

4.50-5.00 – **Horizon scan** – what other mammalian genomes are currently being sequenced?

Thursday, 29 August 2019: Morning Program

9:00-12:00pm

VGP Assembly: Prospective talks and future plans

Chair: Adam Phillippy

Location: CRC 506

This 2nd assembly session will consist of prospective presentations and discussion of the VGP assembly group and recently with EBP collaborators, towards solving the remaining challenges for achieving complete and error free genomes. All conference participants, studying all types of organisms, are welcomed.

9:00 / **Introduction:** Adam Phillippy, NHGRI

9:10 / **Formats and methods:** Gene Myers, MPI Dresden

9:30 / **Heterozygosity and phasing:** Richard Durbin, Wellcome Sanger Institute

9:50 / **Sequencing technology updates:** Sergey Koren, NHGRI

10:10 / **Gene loss in assemblies:** Chul Lee, Seoul National University

10:20 / **Summary discussion**

10:30 / Break

11:00 / **Phasing methods:** Shilpa Garg, Harvard Medical School

11:20 / **Segdup resolution / remote presentation:** Mark Chaisson, University of Southern California

11:40 / **Pan-genomics and variation graphs.** Erik Garrison, UCSC

11:50 / **Summary discussion**

Conservation Genomics

Chairs: Oliver Ryder, Cynthia Steiner, and Warren Johnson

Location: CRC 406

Devil is in the detail: genetic rescue of Tasmanian devils: Kathy Belov, University of Sydney

Genomes and population size: Historical demography of the vaquita: Phil Morin, NMFS/NOAA

Genomic diversity in ‘alalā, and plans for resolving the genetic architecture of hatching failure with ‘alalā and kakapo: Jolene Sutton, University of Hawaii, Hilo

Large-scale comparative genomics as a tool for conservation biology: Diane Genereux, Broad Institute

Conservation genomics of the critically endangered northern white rhinoceros: Cynthia Steiner, San Diego Zoo Institute for Conservation Research

Applying genomics to empower management of *ex situ* populations of endangered species: examples from the black-footed ferret and dama gazelle: Klaus-Peter Koepfli, Smithsonian Conservation Biology Institute

Building a community to accelerate wildlife conservation – The iConserve initiative: Karine A. Viaud-Martinez, Illumina Inc.

The Genetic Species Knowledge Index: mapping the landscape of vertebrate data: Dalia Amore Conde, University of Odense and Species360

Bat 1K Breakout session

Chairs: Liliana M. Dávalos, Michael Hiller

Location: CRC 206

The Bat1K project aims to sequence the genome of all living bats. Although currently in its pilot phase, the project has: produced reference-quality assemblies of six bat species, generated highly complete gene annotations by integrating a variety of evidences; used genome-scale data sets to estimate the phylogenetic position of bats within Laurasiatheria; and performed a number of genome-wide screens to uncover genomic determinants of exceptional traits. Despite sparse sampling across bats, the project has already identified key genomic regions associated with trait diversity within bats. In the Breakout session, we will present workflows and new methods for genome annotation and comparison that have worked well in our previous work and likely have general applicability. A special focus will be on issues and open challenges that we encountered in assembly, gene annotation, and genome comparison. We will close with an outlook of the sequencing plans to generate a genome assembly of at least one member of each bat family. The format of this session will be presentations and open discussions. The session is open to all conference attendees.

Program:

9:00 / **Introduction from the chairs**

9:15 / **TOGA: Tool to infer Orthologs from Genome Alignments** (Michael Hiller, MPI Dresden)

9:45 / **Issues and challenges in gene annotation and genome comparisons** (David Jebb, MPI Dresden)

10:45 / **Coffee break**

11:00 / **The genomics of why are there so many species of bats** (Liliana Dávalos, Stony Brook University)

11:30 / **Open discussion and outlook of Bat1K Phase 1** (Michael Hiller, Liliana Dávalos)

12:00 **Lunch break**

Update from the Global Invertebrate Genome Alliance (GIGA)

Chairs: Jose Lopez, Monica Medina, Adelaide Rhodes, Agosthino Antunes

Location: CRC 106

The Global Invertebrate Genome Alliance/Community of Scientists (<http://GIGA-cos.org>) is a collaborative network of diverse researchers dedicated to promoting genomics research of spineless (mostly aquatic) animals, the invertebrates. Our goals support and align with the Earth Biogenome Project's (EBP) primary mission – to full genome sequences of most extant eukaryotes. For this conference session, we have gathered experts who will discuss their research and the latest progress. Many technical hurdles abound such as difficult genome assemblies, database gaps and computational processing shortfalls for assembly and annotation in the small laboratories that comprise most of GIGA. Nonetheless, recent advances include the full genome sequencing of invertebrate taxa within Cnidaria (Anthozoa, Endocnidozoa, and Medusozoa), Mollusks (Gastropoda, Bivalvia), Placozoa, and Planaria. The format of this session will encompass in depth presentations, open discussions, and potential solutions. Time permitting, the last segment of the session will be an open forum for Q and A.

9:00 / **Welcome** - Joe Lopez, Nova Southeastern University Ocean Center

9:20 / **Microbial-Host Codevelopment in the Upside Down Jellyfish *Cassiopea xamachana*** - Monica Medina, Penn State University

9:40 / **Caribbean corals have rebounded from major climatic variations in the past two million years** - Carlos Prada, University of Rhode Island

10:00 / **Mollusk genomics** - Juliette Gorson, Hunter College

10:20 / Coffee break

10:40 / **Update from UK Darwin Tree of Life project, nematode genomes** - Mark Blaxter, Welcome Sanger Institute

11:00 / **Examples of adaptive genomics** - Agosthino Antunes, University of Porto

11:20 / **Cloud Computing Opportunities for EBP Training** - Adelaide Rhodes, Broad Institute MIT

11:40 / **Do emergent technologies (10X Genomics, PacBio and Hi-C) help molluscan genome reconstruction?: Assembling a reference-quality genome for *Solemya velum* (Bivalvia: Protobranchia)** - Vanessa Gonzalez, Smithsonian

Thursday, 29 August 2019: Afternoon program

2:00-3:30pm

VGP Funding and Publications

Chair: Erich D. Jarvis

Location: Carson Family Auditorium

This session will focus on updating the community on publications in progress on the 1st wave of papers from the VGP Phase 1, proposed plans for the 2nd wave of publications. Also to be discussed are proposed plans for raising the remaining funds to complete Phase 1 of the VGP, start Phase 2 of family level genomes, and integration with other related projects, including EBP, Darwin UK EBP, Bat1K, B10K, among others. We will also hear from a journal editor about the future of publishing in genomics.

2:00-2:20 / **Update and plans for 1st and 2nd waves of VGP publications and funding:** Erich D. Jarvis, Rockefeller University

2:20-2:40 / **Publishing Genomics in Nature family of journals:** Orli Bahcall, Senior Editor, Nature

2:40-3:00 / **A Science editors view of genomics heading into 2020:** Laura Zhan, Senior Editor, Science

3:00-3:30 / **Open discussion** on publications and funding

Microbial Eukaryotes

Chair: Neil Hall

Location: CRC 306

Microbial Eukaryotes Zoom Link <https://zoom.us/j/804992544> (meeting ID: 804 992 544)

One tap mobile
+16465588656,,804992544# US (New York)
+14086380968,,804992544# US (San Jose)

Dial by your location
+1 646 558 8656 US (New York)
+1 408 638 0968 US (San Jose)

Find your local number: <https://zoom.us/j/auQk9klfX>

This session will focus on the challenges faced and successes encountered on producing high quality assemblies from very small animals and single cell microbes.

2:00 pm- **De novo eukaryotic genome assembly through PCR library amplification of subnanogram DNA samples:** Christopher Laumer (remote link), European Bioinformatics Institute

2:25 **Single cell protist sequencing for the Darwin Tree of Life project:** Neil Hall, Earlham Institute,

2 50 **Lowering input requirements & microbial eukaryote PacBio sequencing examples:** Jonas Korlach, Pacific Biosciences

3:15 **Sequencing anything in any environment:** Daniel Fordham, Oxford Nanopore Technologies

3:40 **Sequencing very small animals and microbes:** Mara Lawniczac, Wellcome Sanger Institute

4: 05 **How not to sequence an insect:** Stephen Richards, Baylor College of Medicine

Tech Talks

Bionano optical mapping for accurate genome assembly, comparative genomics, and haplotype segregation

Alex Hastie, Ph.D., Director, Customer Solutions, Bionano Genomics



Wednesday, August 28, 2019, 12:30-1:30pm

Carson Family Auditorium, Collaborative Research Center

Abstract

Today, accurate and structurally intact genomes are finally being efficiently assembled for many diverse species thanks to the availability of new genomics technologies. Bionano optical mapping is a key component thanks to the extremely long read lengths which assure structural accuracy and contiguity even through complex repeat regions of the genome. Thanks to its high throughput and low cost, researchers are able to apply optical mapping to multiple individuals to make genome variation discoveries independent of sequencing for structural variations. Optical mapping has been used to compare different species and detect somatic variations. One remaining challenge for genome assembly is efficient separation of haplotypes. This can be accomplished with Bionano optical mapping to a high degree with an individual but can be even more efficiently separated by use of a trio approach. Bionano optical mapping will continue to be a valuable tool for plant and animal genome research including gold standard genome assembly, structural variation analysis, and haplotype separation.

Biography

Alex Hastie earned a PhD in Molecular and Cellular Biophysics and Biochemistry, Roswell Park Cancer Institute where he studied the role of various transcription factors and the structure of promoters in the control of gene expression. Following his PhD, he undertook postdoctoral training in biochemistry relating to cell stress at the Max Planck Institute for Biochemistry in Munich, Germany. His current role at Bionano Genomics is Director of Customer Solutions. His responsibilities include management of the customer support department. He also leads an applications lab which has the goal to demonstrate the Bionano technology, advance molecular biology and bioinformatics methods and to develop new applications for biological analysis.

New Capabilities of the PacBio Sequel II Sequencing System

Jonas Korlach, CSO, Pacific Biosciences



Thursday, August 29, 2019, 12:30-1:30pm

Carson Family Auditorium, Collaborative Research Center

Abstract

The Sequel II System has an 8-fold higher throughput compared to its predecessor, thereby allowing many applications to be carried out faster and more economically, as well as enabling new application areas. I will present on the latest advances and resulting applications, including a new sequencing chemistry which substantially increases the yield of long & accurate HiFi sequencing reads, and including new developments in the areas of further lowering the required DNA input amount ahead of sequencing, as well as de novo assembly algorithms and full-length RNA sequencing for genome annotation.

Biography

Jonas Korlach has been Chief Scientific Officer of Pacific Biosciences since July 2012, and has been with PacBio since 2004. He co-invented the SMRT technology with Stephen Turner, Ph.D., Pacific Biosciences Founder and Chief Technology Officer, when the two were graduate students at Cornell University. Dr. Korchach is the recipient of multiple grants, an inventor on 70 issued U.S. patents and 61 international patents, and an author of over 100 scientific studies on the principles and applications of SMRT technology, including publications in Nature, Science, and PNAS. In 2013, Dr. Korchach was honored by the Obama White House as an Immigrant Innovator “Champion of Change.” He received both his Ph.D. and his M.S. degrees in Biochemistry, Molecular and Cell Biology from Cornell, and received M.S. and B.A. degrees in Biological Sciences from Humboldt University in Berlin, Germany.

Impact of Next-Generation Sequencing on Biology

Gary P. Schroth, Distinguished Scientist and Vice President, Illumina, San Diego, CA



Friday, August 30, 2019, 12:30-1:30pm

Carson Family Auditorium, Collaborative Research Center

Abstract

It was just over 10 years ago that Illumina first launched the Genome Analyzer II, the world's first high-throughput sequencing platform. That system could produce one billion bps of sequence data per run, which helped ignite a new revolution in genomics. The past decade has seen further advances in data throughput, cost reductions, improvements in data quality and development of many new applications in next-generation sequencing. This talk will highlight the latest improvements in the science and technology of next-generation sequencing (NGS), and how these enable entire new areas of research in biology, agriculture, and medicine.

Biography

Dr. Schroth is currently a Vice President and Distinguished Scientist at Illumina where he directs the Core Applications Group in Product Development, based in San Diego. His department is responsible for core library prep kit and application development. Gary obtained his Ph.D. in biochemistry from the University of California at Davis and has been working in the field of NGS for over a decade as part of Illumina (and Solexa). In his research Gary uses NGS to study genomics, gene structure, expression and regulation and applies this to projects in the fields of cancer, microbiology and infectious disease. Over the course of his career Dr. Schroth has been an author on more than 95 peer reviewed research papers and holds 17 U.S. patents.

One Tree, One Planet

An art and science collaboration about the connection between humans and all life forms that share our planet

“One Tree, One Planet” Background

The “One Tree, One Planet” project is a collaboration of art and science by internationally acclaimed artist Naziha Mestaoui and Florida Museum scientists Douglas Soltis, Pamela Soltis, Robert Guralnick and Matt Gitzendanner, and OneZoom creators James Rosindell and Yan Wong. The team created the two-story interactive projection, titled “One Tree, One Planet,” which highlights the connection between humans and all life forms with which we share our planet. The project also celebrates Earth’s diversity of animals, plants and microbes represented by the Tree of Life, an immense network of relationships that links all species. Audience members may interact with the projection and see their faces and heartbeats next to those of other species. Music is paired with the projection by assigning notes to conserved DNA sequences shared across all species to create a symphony of life. The project originally launched at the University of Florida in Gainesville with a week-long celebration in November 2017 including public lectures, outdoor projections and a commemorative beer launch. The projection was displayed at a subsequent event at First Magnitude Brewing Co. in April 2018.

New “One Tree, One Planet” App

The Florida Museum of Natural History launched a “One Tree, One Planet” app on April 8, 2019, which allows users to interact with the Tree of Life on their cells phones and also offers challenges to emphasize the small, individual changes that each of us can do that cumulatively will have big positive outcomes for the future of our planet. The app is designed with three main areas. The first area is for use during the Tree of Life Projection. Users first capture an image of their face and heartbeat for use in the live projection and are also invited to connect with another species and learn a new perspective on biodiversity. In the second area of the app, participants are then invited to make a difference by taking challenges which help reduce our environmental impact. Some of the challenges include choosing sustainable seafood options and planting native wildflowers. Users will be able to “level up” as they complete challenges and see how their accomplishments compare to others using the app. The third part of the app includes free exploration of the fractal representation of the Tree of Life that maps the evolutionary history and relationships between more than 2 million species.

Appstore : <https://itunes.apple.com/app/id1405547430>

Google Playstore: <https://play.google.com/store/apps/details?id=org.onetreeoneplanet>

“One Tree, One Planet” Credits

“One Tree, One Planet” was produced at the University of Florida with support from the 1923 Fund, UF Research Opportunity Seed Fund Grant, UF Biodiversity Institute, UF Genetics Institute, UF Office of the Provost and the Florida Museum of Natural History.

Team Members

Naziha Mestaoui

Mestaoui is a Paris-based environmental artist and architect whose work creates immersive and sensory experiences by blending space, imagery and technology. Through her art, she invites us to use technologies to reconnect with nature, creating a dynamic that can inspire our future. Mestaoui's work has been exhibited around the globe including the Museum of Modern Art in New York, the Centre Georges Pompidou in Paris and the Museum of Photography in Tokyo.



Douglas Soltis

Soltis, the principal investigator of the “One Tree, One Planet” series, is a distinguished curator in the Florida Museum of Natural History Laboratory of Molecular Systematics and Evolutionary Genetics and a distinguished professor in the University of Florida department of biology. His research interests focus on plant evolution and phylogeny. He was elected to the National Academy of Sciences in 2017 and, along with his wife Pamela Soltis, was the joint awardee of the 2006 Asa Gray award which recognizes lifetime achievement in plant systematics. The Soltises were also awarded the Darwin-Wallace Medal in 2016 by the Linnean Society of London in recognition of major advances in evolutionary biology.



Pamela Soltis

A distinguished curator in the Florida Museum Laboratory of Molecular Systematics and Evolutionary Genetics, Soltis studies plant diversity with an emphasis on the origin and evolution of flowering plants, plant genome evolution and conservation genetics. She is also the director of the University of Florida Biodiversity Institute. She was elected to the National Academy of Sciences in 2016 and to the American Academy of Arts and Sciences in 2017, along with her husband Douglas Soltis. She currently serves on a National Academies of Sciences Engineering and Medicine committee to investigate the value and future of biological collections.

**Robert Guralnick**

An associate curator of biodiversity informatics at the Florida Museum of Natural History, Guralnick researches biodiversity with a focus on spatiotemporal changes in genetic and species diversity. He uses an integrative approach to global change biology and his work is also geared toward the mobilization and re-use of already collected biodiversity records like biocollections records.

Matt Gitzendanner

Gitzendanner is a scientist with the University of Florida department of biology and the Florida Museum of Natural History. He focuses on genomics, bioinformatics and population and conservation genetics and has considerable computational expertise in dealing with big data and building large relationship trees.

James Rosindell (OneZoom)

Rosindell is a research fellow/lecturer in biodiversity theory and science outreach at Imperial College London. Together with Yan Wong and others, he created the OneZoom Tree of Life explorer website www.onezoom.org which provides the interactive tree explorer engine for “One Tree, One Planet”. OneZoom is now run as an independent registered charity for which he and Wong both serve as board members. In his other research work, Rosindell is interested in models of biodiversity and their applications in conservation, ecology and evolution.

Yan Wong (OneZoom)

Wong is an evolutionary biologist with expertise in maths, genetics and computing. He was a lecturer in evolutionary biology and ecology at the University of Leeds, then worked in professional science outreach, and now carries out research at the Big Data Institute in Oxford on computational techniques for handling large genetic datasets. He co-authored “The Ancestor’s Tale” book (2016) with Richard Dawkins and has presented on numerous television and radio science shows. Wong has worked closely with James Rosindell on the OneZoom project since 2014, including the application of OneZoom as part of “One Tree, One Planet”.

News articles

- Alligator, 'One Tree, One Planet' event series promotes biodiversity - https://www.alligator.org/news/one-tree-one-planet-event-series-promotes-biodiversity/article_590ea52a-c01a-11e7-b98e-cb75551ee28c.html
- Alligator, First Magnitude Brewing Co. to host One Tree Beer event - https://www.alligator.org/the_avenue/first-magnitude-brewing-co-to-host-one-tree-beer-event/article_797c2bd4-c9cf-11e7-bea8-6b58ac102317.html
- WCJB - <https://www.wcjb.com/content/news/Florida-celebrates-Biodiversity-457603153.html>
- CBS4 - <https://mycbs4.com/news/local/tree-of-life-app-challenges-users-to-focus-on-sustainability-04-08-2019>

Websites

- Naziha's website - <http://nazihamestaoui.com/one-tree-one-planet-between-art-and-science/>
- Florida Museum OTOP website - <https://www.floridamuseum.ufl.edu/event/tol/>
- TreeTender website - <https://www.treetender.org/>
- Event Trailer - <https://www.facebook.com/FloridaMuseum/videos/treetender/10155734210393955/>
- UF BioLink - <https://www.ufbiolink.org/events/2017/11/16/one-tree-one-planet-world-premiere>
- TreeTender Trailer on Vimeo - <https://vimeo.com/254398812>
- Full TreeTender video on Vimeo - <https://vimeo.com/248181788>

Posters: At-A-Glance

By abstract title, authors, and affiliations

Genome Assemblies Section

1. The first skink genomes: Assembling and annotating genomes for *Lerista*, a nascent model system for the evolution of limblessness

Card, Daren C.1,2,; Hutchinson, Mark N.3; Donnellan, Stephen C.3; and Edwards, Scott V.1,2

1 Department of Organismic & Evolutionary Biology, Harvard University, Cambridge, MA, USA

2 Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA

3 South Australian Museum, Adelaide, South Australia, Australia

2. Conservation genomics applied to the Balearic shearwater

Cuevas-Caballé, Cristian1; Ferrer-Obiol, Joan1; Rozas, Julio1; González-Solís, Jacob2; and Riutort, Marta1

1 Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Catalonia, Spain.

2 Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Catalonia, Spain.

3. Preliminary genomic assemblages of two marine arthropods, *Chionoecetes opilio* (Crustacea: Decapoda) and *Nymphon striatum* (Chelicerata: Pantopoda)

Jeong, Jin-Hyeop; Lee, Damin; and Kim, Won

Seoul National University, Seoul 08826, Korea

4. The bilby genome project

Peel, Emma; Brandies, Parice; Hogg, Carolyn; and Belov, Katherine

School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia

5. Slothomics: the first chromosome-level genome of the slowest existing mammalian group

Uliano-Silva, Marcela1,2; Winkler, Sylke3; Myers, Eugene3; and Mazzoni, Camila1,2

1 Leibniz 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

6. Assembling the genome of the fungus fly, *Sciara coprophila*, to enable studies on its unique chromosome biology

Urban, John M.1; Foulk, Michael S.2,3; Bliss, Jacob E.2; Coleman, C. Michelle4; Lu, Nanyan4; Mazloom, Reza4; Brown, Susan J.4; Spradling, Allan C.1; and Gerbi, Susan A.2

1 Carnegie Institution for Science, Department of Embryology, 3520 San Martin Drive, Baltimore, Maryland 21218, USA

2 Brown University Division of Biology and Medicine, Department of Molecular Biology, Cell Biology, and Biochemistry, Providence, RI, USA

3 Mercyhurst University, Department of Biology, Erie, PA, USA

4 Kansas State University Division of Biology, KSU Bioinformatics Center, Manhattan, KS, USA

7. Near-complete genome assembly of three amphioxus species

Xu, Luohao¹ and Huang, Zhen²

¹ Dept. of Molecular Evolution and Development, University of Vienna, Austria

² Fujian Key Laboratory of Developmental and Neural Biology, Fujian Normal University, China

Big Projects Section

8. Growing Research Capability in Colombia: A Shared Vision on Protecting Biodiversity to Achieve Sustainability and Peace.

Azcarate, Juan¹; Valderrama, Natalia¹; Di Palma, Federica¹; and The GROW Colombia project consortium²

¹ The Earlham Institute, Norwich Research Park, Norwich, United Kingdom

² In the UK: The Earlham Institute, University of East Anglia, Aberystwyth University, Natural History Museum and Eden Project. In Colombia: Universidad de los Andes, Humboldt Institute and the International Center for Tropical Agriculture (CIAT). International: University of Sydney.

9. USDA-ARS's Ag100Pest Initiative: The Genomics of Pest Control

Childers, Anna K.¹; Coates, Brad ²; Geib, Scott ³; Poelchau, Monica F. ⁴; Childers, Chris P. ⁴; Scheffler, Brian ⁵; and Hackett, Kevin⁶

¹ Bee Research Laboratory, USDA-ARS, Beltsville, MD, USA

² Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA, USA

³ Tropical Crop and Commodity Protection Research Unit, USDA-ARS, Hilo, HI, USA

⁴ Knowledge Services Division, USDA-ARS National Agricultural Library, Beltsville, MD, USA

⁵ Brian Scheffler, Genomics and Bioinformatics Research Unit, USDA-ARS, Stoneville, MS, USA

⁶ Kevin Hackett, Office of National Programs, USDA-ARS, Beltsville, MD, USA

10. Reference Quality Insect Genomes: Applications for Sustainable Protein and Biodiversity

Dossey, Aaron T¹; Chu, Clay¹; and Oppert, Brenda²

¹ All Things Bugs LLC and Invertebrate Studies Institute

² USDA Agricultural Research Service, Center for Grain and Animal Health Research

11. The i5k Workspace@NAL provides genome database services for orphaned arthropods

Poelchau, Monica¹; Chiang, Li-Mei²; Hsiao, Yi²; Hsu, Min-Chen³; Lin, Chun-Hung³; Wu, Chia-Tung⁴; and Childers, Christopher⁵

¹ Knowledge Services Division, USDA-ARS National Agricultural Library, Beltsville, MD, USA

² Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

³ Min-Chen Hsu, Graduate Institute of Communication Engineering, National Taiwan University

⁴ Graduate Institute of Computer Science and Information Engineering, National Taiwan University

⁵ Knowledge Services Division, USDA-ARS National Agricultural Library, Beltsville, MD, USA

12. An Earth BioGenome Project Progress Meter.

Richards, Stephen; Caperello, Nicolette D.; Lewin, Harris A.; and Representing members of the Earth BioGenome Project Working Group

UC Davis Genome Center, University of California at Davis, Davis CA.

Thursday:

Methods

13. **Genome-wide False Duplications Identified by VGP Assembly**

Byung June Ko¹; Chul Lee²; Juwan Kim²; Erich D. Jarvis³, and Heebal Kim^{1,2}

¹ Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University Seoul, Republic of Korea

² Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

³ The Rockefeller University and HHMI, USA

14. **A novel computational approach to haplotype-resolved assembly using sequence graphs**

Shilpa Garg^{1,5}; Yichen Wang²; John Aach¹; Heng Li³; Richard Durbin⁴; and George Church^{1,5}

¹ Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

² Boston University, Boston, Massachusetts, USA

³ Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

⁴ Department of Genetics, University of Cambridge, Cambridge, United Kingdom

⁵ Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts, USA

15. **Ensembl 2019**

Haggerty, Leanne; Allen, Jamie; Billis, Konstantinos; Girón, Carlos García; Hourlier, Thibaut; Izuogu, Osagie; Ogeh, Denye; Martin, Fergal J.; Howe, Kevin; and Flicek, Paul

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

16. **Visualization of Genome Scaffolding using Hi-C Paired-end Reads**

Harry, Ed and Ning, Zemin

Wellcome Sanger Institute, Hinxton, Cambridge, UK

17. **Single chromosome sequencing to improve genome studies**

Iannucci, Alessio; Ferguson-Smith, Malcolm; Pereira, Jorge Claudio; Rovatsos, Michail; Kichigin, Ilya G.; Makunin, Alex I.; Pokorná, Martina J.; Altmanová, Marie; Trifonov, Vladimir A.; Kratochvíl, Lukáš; Stanyon, Roscoe R.; Lind, Abigail L.; Pollard, Katherine S.; Bruneau, Benoit G.; and Ciofi, Claudio

University of Florence, Italy

18. **Vertebrate genome assembly and annotation projects as course-based undergraduate research experiences**

Jue, Nathaniel K.¹; Slown, Corin¹; Rocha, Luis A.²; Willis, Stuart C.²; Johnson, Shannon; and Vrijenhoek, Robert C.³

¹ California State University, Monterey Bay

² California Academy of Sciences

³ Monterey Bay Aquarium Research Institute

19. **False Gene Losses Corrected by VGP Assembly**

Kim, Juwan^{1¶}; Lee, Chul^{1¶}; Ko, Byung June²; Rhie, Arang³; VGP assembly group; Kim, Heebal^{1,2,6*}; and Jarvis, Erich D.^{4,5*}

¶ Both authors contributed equally to this work.

1 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

2 Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

3 Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA

4 Laboratory of Neurogenetics of Language, Rockefeller University, 1230 York Avenue, New York, NY 10065, USA

5 Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815, USA

6 C&K genomics, Seoul, Republic of Korea

20. The past, present, and future of Arima-HiC for genome assembly: An overview of Arima-HiC sample prep and scaffolding of VGP genomes

Schmitt, Anthony; De La Torre, Chris; Reid, Derek; Mac, Stephen; Zhou, Xiang; Tan, Catherine; Won, Melissa; and Selvaraj, Siddarth

Arima Genomics, Inc., San Diego, CA, USA

21. Building an assembly army for the VGP: challenges and first successes of building high-quality genomes by the Berlin student's team

Uliano-Silva, Marcela; Driller, Maximilian; Caswara, Calvinna; Vafadar, Majid; Rhie, Arang; Jarvis, Erich D.; and Mazzoni, Camila

Biological Discoveries Section

22. Analysis of Microsatellite Abundance in Mammalian Genomes; Revisiting Peto's Paradox

Jeong, Heesul and Kim, Heebal^{1,2,3*}

1 Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea, 151-742

2 Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Korea, 151-742

3 Institute for Biomedical Sciences, Shinshu University, Nagano, Japan

23. Rules of Amino Acid Convergences: Not How Many, but Who in Avian Vocal Learning Clades

Lee, Chul¹; Cho, Seoae¹; Kim, Kyuwon¹; Yoo, Dongahn¹; Han, Jae Yong¹; Lee, Hong Jo¹; Gedman, Gregory²; Pfenning, Andreas³; Kim Heebal^{*2}; and Jarvis, Erich D.^{*2}

1 Seoul National University

2 Rockefeller University and HHMI

3 Carnegie Mellon University

24. AgriVectors: Sequencing, omics resources and systems biology portal for plant pathosystems and arthropod vectors of plant diseases

Saha, Suryal¹; Hunter, Wayne²; Mueller, Lukas A.³; and The AgriVectors Consortium³

1 Boyce Thompson Institute, Ithaca, NY 14853

2 USDA ARS, U.S. Horticultural Research Laboratory, Ft. Pierce, FL 34945

3 Boyce Thompson Institute, Ithaca, NY 14853

25. Whole-genome alignments of publicly available high-quality bird genome assemblies highlight phylogenetic profiles of structural variants

Secomandi, Simona¹; Formenti, Giulio²; Rhie, Arang³; Chiara, Matteo¹; Poveda, Lucy⁴; Francoijs, Kees-Jan⁵; Bonisoli-Alquati, Andrea⁶; Canova, Luca⁷; Gianfranceschi, Luca¹; Horner, David Stephen¹; Jarvis, Erich D. ²; and Saino, Nicola⁸

¹ Department of Biosciences, University of Milan (Milan, Italy);

² The Rockefeller University (New York, NY, USA)

³ National Human Genome Research Institute, National Institutes of Health (Bethesda, Maryland, USA)

⁴ Functional Genomics Center of Zurich, University of Zurich, (Zurich, Switzerland)

⁵ Bionano Genomics (San Diego, CA, USA).

⁶ Department of Biological Sciences, California State Polytechnic University, Pomona (Pomona, CA, USA)

⁷ Department of Biochemistry, University of Pavia (Pavia, Italy)

⁸ Department of Environmental Science and Policy, University of Milan (Milan, Italy)

26. Origins and evolution of extreme longevity in the adaptive radiation of rockfish

Kolara, Sree Rohit Raj¹; Stubbs, Alexander¹; Chatla, Kamalakar¹; Jainese, Conner²; Seeto, Katelin²; Bachtrog, Doris¹; Love, Milton S²; and Sudmant, Peter H¹

¹ Integrative Biology, University of California, Berkeley

² Marine Science Institute, University of California, Santa Barbara

27. Genomic signatures of the landlocked adaptations in Taiwan gobies (*Rhinogobius* spp.)

Wang, Tzi-Yuan¹; Huang, Shih-Pin¹; Wu, Yu-Wei²; Liao, Te-Yu³; Chaw, Shu-Miaw¹; Wang, Feng-Yu⁴

¹ Biodiversity Research Center, Academia Sinica, Nankang, Taipei, Taiwan.

² Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan

³ Department of Oceanography, National Sun-Yet-Sen University, Kaoshiung, Taiwan

⁴ Taiwan Ocean Research Institute, National Applied Research Laboratories, Kaohsiung, Taiwan

28. Novel approach to detect interspecies-level directional selection based on divergence from ancestral sequence and polymorphism data

Yoo, DongAhn¹; Lee, Chul¹; and Kim, Heebal^{1,2,3,4*}

¹ Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

² Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

³ C&K Genomics, C-1008, H Businesspark, 26, Beobwon-ro 9-gil, Songpa-gu, Seoul, Republic of Korea

⁴ Department of Interdisciplinary Genome Sciences and Cell Metabolism, Institute for Biomedical Sciences, ICCER, Shinshu University, 8304 Minami-Minowa, Kami-Ina, Nagano 399-4598, Japan

Poster Abstracts

Wednesday:

Genome Assemblies Section

1. The first skink genomes: Assembling and annotating genomes for *Lerista*, a nascent model system for the evolution of limblessness

Card, Daren C.1,2;; Hutchinson, Mark N.3; Donnellan, Stephen C.3; and Edwards, Scott V.1,2

1 Department of Organismic & Evolutionary Biology, Harvard University, Cambridge, MA, USA

2 Museum of Comparative Zoology, Harvard University, Cambridge, MA, USA

3 South Australian Museum, Adelaide, South Australia, Australia

The genomics revolution has initiated a renaissance in diverse areas of biology, especially in studies of non-model taxa. However, squamate reptiles – a major vertebrate lineage – have been largely overlooked, despite exhibiting numerous interesting natural history characteristics. One example best known from snakes is the repeated evolution of serpentine body plans characterized by limb loss or reduction and elongation of the axial trunk. This major morphological transition has occurred over 20 times among squamates, with skink lizards alone containing at least half of the known origins of snake-like body forms. The skink genus *Lerista* offers a particularly tractable system for studying the evolution of serpentine body plans. *Lerista* species possess various limb morphologies, ranging from an ancestral pentadactyl limb to total limblessness, with several combinations of intermediate digit and limb loss. Moreover, the series of phenotypes within the genus *Lerista* appears to have convergently evolved over the past 20 million years. We aim to integrate morphological, developmental, and genomic data to decipher the molecular mechanisms underlying convergent morphological evolution within this clade. A critical resource for this research is the presence of high-quality reference genomes, but genomic resources are completely lacking for skink lizards. To fill this gap, we used a combination of 10x Chromium and Hi-C sequencing to produce high quality genome assemblies for two *Lerista* species: the pentadactyl-limbed *L. bougainvillii* and the limb-reduced *L. edwardsae*. For each genome, we describe repeat element composition and characterize the results of our gene annotations. Finally, we queried genes and regulatory regions known to play a role in limb development and provide a preliminary look at the genomic underpinnings of limb reduction in *L. edwardsae*. Overall, these resources illuminate the genomic landscapes of a major tetrapod lineage and are invaluable for our ongoing research examining convergent limb loss in *Lerista* skinks.

2. Conservation genomics applied to the Balearic shearwater

Cuevas-Caballé, Cristian1; Ferrer-Obiol, Joan1; Rozas, Julio1; González-Solís, Jacob2; and Riutort, Marta1

1 Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Catalonia, Spain.

2 Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals and Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Catalonia, Spain.

The Balearic shearwater (*Puffinus mauretanicus*) is the most threatened bird in Europe. Listed as critically endangered by the IUCN, the species population undergoes an annual decline of 7.4-14%. Decimated by longline bycatch, invasive mammals, plastic ingestion and light pollution, some studies predict that the species could become extinct by 2070 if the breeding population falls into an extinction vortex. Conservation genomics take advantage of

genome-scale data to assess populations viability and rapidly inform resource managers and policy makers. Here, we assembled a high-quality *P. mauretanicus* genome to scrutinize the genomic effects of the population decline, reconstruct its historical demography and gain insight into the molecular evolution of Procellariiformes. The hybrid assembly (Illumina + ONT) consisted of 4169 scaffolds, with a N50 of 2.1 Mb. The genome has 1.21 Gb, with a 9.95% of repeats, and a BUSCO completeness of 94.8%. We annotated a total of 22179 genes with a BUSCO completeness of 88.2%. We also assembled and annotated the mitogenome (21978 bp), finding similar duplications to the present in Audubon's shearwater (*Puffinus lherminieri*). We inferred the historical population size of the species with PSMC, which according to previous literature reveals that the human colonization of the Balearic Islands circa 5000 years ago halved the *P. mauretanicus* population. Finally, we did a comparative genomics study with other 8 species of Procellariiformes. This reference genome will be the keystone for future fine-scale studies of the species population genomics based on resequencing of 28 individuals. The analyses of these genomes will allow us to assess the current state of the different colonies of the species, its inbreeding and the possible introgression with its sister species, the Mediterranean shearwater (*Puffinus yelkouan*), for which we will also obtain genomes from individuals all around its distribution.

3. Preliminary genomic assemblages of two marine arthropods, *Chionoecetes opilio* (Crustacea: Decapoda) and *Nymphon striatum* (Chelicerata: Pantopoda)

Jeong, Jin-Hyeop; Lee, Damin; and Kim, Won
Seoul National University, Seoul 08826, Korea

Marine arthropods contain more than 60,000 species and demonstrate extensively diverse modes of life and bodyplans. The vast majority of marine arthropods belong to paraphyletic Crustacea and arachnid-excluded Chelicerata. Decapod crustaceans have significant economic values as edible marine invertebrates. Sea spiders are key taxa to understand the early arthropod evolution due to their archaic origins. Despite of their importance, only a few cases of De novo genome assemblies have been reported so far. We sequenced and assembled genomes of *Chionoecetes opilio* and *Nymphon striatum* for the first time. An individual of *C. opilio* was collected from the offshore of Yeongdeok-gun, South Korea at 2019.03.14. 40 individuals of *N. striatum* were collected from Sacheon-hang, South Korea at 2018.07.12 by SCUBA diving. Their DNA samples were extracted for the De novo genome sequencings. A pair of 350 bp insert-sized Illumina paired-end and a PacBio Sequel read libraries were sequenced for the each species. Genome surveys were conducted by K-mer analysis. Wtdbg2 (for *C. opilio*) and HGAP4 (for *N. striatum*) assembled contig-level draft genome sequences respectively. We sequenced 105.60 Gb paired-end reads and 201.36 Gb PacBio reads for *C. opilio* and 136.28 Gb paired-end reads and 84.83 Gb PacBio reads for *N. striatum*, respectively. Genome surveys estimated the genome sizes of *C. opilio* and *N. striatum* as 1.89 Gb and 567.83 Mb. Preliminarily assembled *C. opilio* genome was 1.98 Gb long and composed of 45,098 contigs with 112.2 Kb of N50. The preliminary assemblage of *N. striatum* was 727.55 Mb long and composed of 2,947 contigs with 357.6 Kb N50. Our preliminary draft genomes of *C. opilio* and *N. striatum* were highly heterozygous. Merging haplotigs improved *N. striatum* genome and reduced overestimated genome size. Currently, re-assembling of *C. opilio* genome using Falcon and genomic annotation analyses for both species are under progress.

4. The bilby genome project

Peel, Emma; Brandies, Parice; Hogg, Carolyn; and Belov, Katherine
School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia

The greater bilby (*Macrotis lagotis*) is a burrowing nocturnal marsupial endemic to Australia, and the last surviving member of the Thylacomidae family. The bilby was once distributed over arid and semi-arid Australia, however widespread population decline has resulted in the species being listed as vulnerable on the IUCN red list. Sequencing the genome of this unique marsupial will offer crucial insights into the genetic basis of arid zone adaptations, and other important processes such as immunity and reproduction. This project will provide a reference genome for

ongoing conservation projects to directly inform species management. The bilby reference genome was created using high molecular weight DNA from spleen tissue collected from a single female captive bilby housed at Perth Zoo, Australia. 10x chromium technology was used for library preparation without size selection, and 2 x 150bp paired-end reads were sequenced across one S1 lane on a NovaSeq 6000. Raw reads were assembled de novo using Supernova, producing a draft genome of 3.2GB with a scaffold N50 of 462kb and containing 86% complete mammalian BUSCOs. For the transcriptomes, high quality RNA was extracted from eleven tissues collected from the same individual. TruSeq total RNA libraries were prepared, and 2 x 150bp paired-end reads sequenced across one S1 lane on a NovaSeq6000. Raw reads were trimmed with Trimmomatic, assembled de novo using Trinity and annotated using the Trinotate pipeline. This resulted in eleven individual tissue transcriptomes, ranging in size from 400-900MB with complete vertebrate BUSCO scores from 84% to 89%. Analysis is ongoing, including genome scaffolding using HiC and automated annotation with Maker. Manual annotation of the genome and transcriptomes has already lead to the discovery of a suite of antimicrobial peptides and characterisation of important immune families such as the major histocompatibility complex.

5. Slothomics: the first chromosome-level genome of the slowest existing mammalian group

Uliano-Silva, Marcela^{1,2}; Winkler, Sylke³; Myers, Eugene³; and Mazzoni, Camila^{1,2}

¹ Leibniz 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

Modern sloths are members of Xenarthra, the only placental mammalian Superorder that evolved in the New World. They present extreme adaptations to a life on the trees, being the world's only inverted quadruped as a result of several fore-limbs adaptations, a lower-than-average-muscle content and extremely low basal metabolic rates. Little is known about the genomics of the group. We have assembled the first chromosome level genome for the two-toed sloth *Choloepus didactylus* following the Vertebrate Genomes Project Assembly Pipeline (v1.5). It uses four data types and different softwares hierarchically to assemble genomes to chromosomes, including a final step of curation: (Pacbio and FalconUnzip, Purge Haplotigs, Chromium 10X and Scaff10x, Bionano and Solve, Arima Hi-C and Salsa2). Before manual curation, *C. didactylus* was assembled in a total of 3.6 Gb into 668 scaffolds with a N50 = 111 Mb. Nucleotide synteny analysis (mummer) showed entire correspondence of *C. didactylus* scaffolds and complete human chromosomes, in accordance with previous cross-species chromosome painting, evidencing that the VGP (v.15) Assembly pipeline builds complete chromosomes even before manual curation. At this stage, 94% of the genome is assembled in 32 scaffolds, in agreement with most *C. didactylus* karyotypes (2n=65). A first evaluation of the repetitive content showed that 43% of *C. didactylus* genome is composed of repeats, with LINE1 representing 33% of the elements. A preliminary Maker annotation was produced and imputed to an orthology analysis (Orthofinder2) with 11 other species representing all the largest mammalian clades, and yielded a total of 2596 single copy orthologs and a maximum likelihood tree indicating Afrotheria to be more basal than Xenarthra within Eutherian mammals, a scenario that is still largely debated. Currently, *C. didactylus* genome is in the final stages of manual curation and will be the first Xenarthra representative in the Phase 1 of the VGP Ordinal Project.

6. Assembling the genome of the fungus fly, *Sciara coprophila*, to enable studies on its unique chromosome biology

Urban, John M.¹; Foulk, Michael S.^{2,3}; Bliss, Jacob E.²; Coleman, C. Michelle⁴; Lu, Nanyan⁴; Mazloom, Reza⁴; Brown, Susan J.⁴; Spradling, Allan C.¹; and Gerbi, Susan A.²

¹ Carnegie Institution for Science, Department of Embryology, 3520 San Martin Drive, Baltimore, Maryland 21218, USA

² Brown University Division of Biology and Medicine, Department of Molecular Biology, Cell Biology, and Biochemistry, Providence, RI, USA

³ Mercyhurst University, Department of Biology, Erie, PA, USA

⁴ Kansas State University Division of Biology, KSU Bioinformatics Center, Manhattan, KS, USA

The fungus fly, *Sciara coprophila*, is rich with opportunities to study chromosome biology. Specific genomic loci are amplified, entire chromosomes are eliminated, and a single nucleus can contain thousands of copies of each chromosome. *Sciara* has five distinct chromosomes (X, II, III, IV, and L). Nevertheless, chromosome elimination and non-disjunction events distort the expected numbers of chromosomes in some cells. Whereas the initial diploid germline genome has ten chromosomes (two of each) and oocytes have five, sperm have seven and a fertilized egg results in twelve. Additional elimination events re-balance the germline and restructure the somatic lineage by eliminating 1-2 paternal X and all L chromosomes. Studies into how these chromosomal events are regulated are handicapped by the lack of a genome sequence. We approached assembling the fungus fly genome with multiple technologies: Illumina (103x), PacBio (44x), and Oxford Nanopore (11x). We generated many assemblies using these datasets and evaluated them with a battery of reference-free metrics. We chose a subset with the best evaluations for scaffolding with optical maps from BioNano Genomics, and more recently Hi-C from Phase Genomics. We produced highly contiguous assemblies with multi-megabase contigs using long reads. The assemblies generated from the combination of PacBio and Oxford Nanopore datasets typically ranked higher than PacBio-only assemblies. Non-hybrid assemblies performed better than hybrid assemblies. Ultimately, the Canu assembler gave us the best assembly. BioNano maps increased the NG50 three-fold. Hi-C gave chromosome-scale scaffolds. RNA-seq datasets from a combination of embryos, larvae, pupae, and adult flies from both sexes were used to facilitate annotation of the final genome sequence. Finally, both PacBio and Oxford Nanopore data gave us the opportunity to explore DNA modifications in the *Sciara* genome. The genome sequence will be an invaluable resource for studying the interesting chromosome dynamics in *Sciara coprophila*.

7. Near-complete genome assembly of three amphioxus species

Xu, Luohao¹ and Huang, Zhen²

¹ Dept. of Molecular Evolution and Development, University of Vienna, Austria

² Fujian Key Laboratory of Developmental and Neural Biology, Fujian Normal University, China

Amphioxus is a basal lineage of chordate, and is an important model in developmental and evolutionary biology. Despite that, a good-quality genome is still lacking. In this study, we assemble the genomes of three amphioxus, florida lancelet (*Branchiostoma floridae*, Bf), Japanese lancelet (*B. japonicum*, Bj) and Chinese lancelet (*B. belcheri*, Bb), using Pacbio long-reads. The contig N50 reach 10.0M, 6.2M and 6.0M in Bb, Bf and Bj respectively. More than 98.6% of the contig sequences are anchored into chromosomes through Hi-C based scaffolding. The assemblies successfully capture telomeric and centromeric sequences for most chromosomes, and reveal a telocentromeric organization of amphioxus chromosomes. While inter-chromosomal rearrangements are rare, with only three such events observed, intra-chromosomal rearrangements are frequently seen among the three amphioxus species. Compared with vertebrates, amphioxus have a twice larger percentage of segmental duplications in their genomes. The duplicated genes are enriched for DNA binding the G-protein coupled receptor protein signaling pathway, similar to what has been found for the retained genes after whole genome duplication in vertebrates. Finally, we identified the ZW sex chromosome pair in each of the amphioxus genome, and found they are not homologous with each other and show different degrees of sex chromosome differentiation. The almost complete assembly of the amphioxus genomes provide an important opportunity for the study on chromosome evolution of chordate and the evolution of sex chromosome turnover.

Amphioxus; Cephalochordata; Genome assembly; Comparative genomics

Big Projects Section

8. Growing Research Capability in Colombia: A Shared Vision on Protecting Biodiversity to Achieve Sustainability and Peace.

Azcarate, Juan¹; Valderrama, Natalia¹; Di Palma, Federica¹; and The GROW Colombia project consortium²

¹ The Earlham Institute, Norwich Research Park, Norwich, United Kingdom

² In the UK: The Earlham Institute, University of East Anglia, Aberystwyth University, Natural History Museum and Eden Project. In Colombia: Universidad de los Andes, Humboldt Institute and the International Center for Tropical Agriculture (CIAT). International: University of Sydney.

Colombia is one of 17 countries considered “megadiverse” by the United Nations Environment Programme (UNEP). The national catalogue of biodiversity includes up to 55 thousand species of animals and plants, 3,652 of them endemic, representing around 10% of all known species on earth. Following the peace agreement in Colombia in November 2016, we now have an opportunity to study the country’s staggeringly rich native biodiversity. In a GROW Colombia Research Councils UK (RCUK) Global Challenges Research Fund (GCRF) international collaborative project, UK and Colombian institutions are working together in this multidisciplinary project. Our aims include strengthening Colombian research capacity in the biological sciences, computational biology and socio-economics to develop robust coordinated activities under a shared vision centred on biodiversity as a means to achieve sustainability and peace. The project includes natural diversity, agricultural diversity and socio-economics of biodiversity research programmes, and a broad set of activities focussed on enhancing research capability by improving researcher skills, as well their access to research information and resources. The GROW Colombia project relies on the complementary expertise and strengths of an alliance which is equipped to develop robust coordinated solutions around biodiversity, with implications for its preservation, global health and development. Here we describe the project, its three programmes, partners, progress, expected outcomes and concrete impacts.

9. USDA-ARS’s Ag100Pest Initiative: The Genomics of Pest Control

Childers, Anna K.¹; Coates, Brad²; Geib, Scott³; Poelchau, Monica F.⁴; Childers, Chris P.⁴; Scheffler, Brian⁵; and Hackett, Kevin⁶

¹ Bee Research Laboratory, USDA-ARS, Beltsville, MD, USA

² Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA, USA

³ Tropical Crop and Commodity Protection Research Unit, USDA-ARS, Hilo, HI, USA

⁴ Knowledge Services Division, USDA-ARS National Agricultural Library, Beltsville, MD, USA

⁵ Brian Scheffler, Genomics and Bioinformatics Research Unit, USDA-ARS, Stoneville, MS, USA

⁶ Kevin Hackett, Office of National Programs, USDA-ARS, Beltsville, MD, USA

USDA-ARS is pledging its support for the Earth BioGenome Project (EBP) and reaffirming its commitment to i5K, the 5000 arthropod genomes initiative, through the Ag100Pest initiative. The Ag100Pest goal is to produce annotated, reference quality genome assemblies for the top 100 US arthropod agricultural pests. The Ag100Pest prioritization team is considering approximately 400 species nominations and establishing selection criteria for arthropod pests of US field crops, livestock, bees, trees, and stored products as well as foreign pest species considered potential invasive threats to US agriculture. We have also formed teams focusing on extraction methods, sequencing, assembly and post-assembly support in the i5k Workspace@NAL, a genome database for data visualization and manual annotation of gene sets. We are striving to generate PacBio long-read data from a single individual, although this presents a significant challenge as many important species have small body sizes. We will present sequencing and assembly progress from the first year of the project. Beyond genomes, Ag100Pest teams are developing best practices that will benefit the entire arthropod genomics community. It is along those lines that we invested in the development

of an open-source functional annotation pipeline specifically tailored for arthropods. The Ag100Pest initiative is interested in collaborating with the broader arthropod research community in this effort. We hope our successes will encourage the community to initiate additional projects to help fill the EBP ark.

10. Reference Quality Insect Genomes: Applications for Sustainable Protein and Biodiversity

Dossey, Aaron T¹; Chu, Clay¹; and Oppert, Brenda²

¹ All Things Bugs LLC and Invertebrate Studies Institute

² USDA Agricultural Research Service, Center for Grain and Animal Health Research

Insects are a critical part of life on earth and a valuable resource for sustainable human existence. Recent sobering studies demonstrate what biologists know: mass extinction is real, and happening in real-time due to human destruction of natural habitats. With historic levels of biodiversity loss and an increasing human population, it is critical to reduce human consumption from earth and its ecosphere. Already 70% of agricultural land and 30% of the land on earth is used for livestock. Diversification of our food supply is critical for food security. The good news is insects hold promise as a sustainable yet unexplored solution. They utilize less energy, feed, land and water than other livestock and contribute less to climate change and pollution (Dossey et al., 2016). Our research on insects as sustainable food ingredients involves sequencing genomes of the most commercially produced insects on earth, crickets and mealworms, obtaining reference quality genomes of *Acheta domesticus* (house cricket), *Grylloides signalus* (banded cricket), and *Tenebrio molitor* (yellow mealworm). We also have genetically engineered crickets and mealworms. These data are being used to improve insects as food crops by increasing nutrient production and disease resistance, as well as applications for non-food bioproduction, such as vaccines, bioactive peptides and antimicrobials. The Invertebrate Studies Institute (ISI, a 501c3 non-profit) is sequencing genomes of insects to understand biodiversity at the molecular level, starting with phasmids. Phasmids are "stick" insects from the Order Phasmatodea, a small but highly diverse order, with over 3,000 described species. They have various mechanisms of reproduction and some have wings and fly, while many have a diverse array of defense mechanisms, including chemical sprays and camouflage while others are brightly colored and aposematic. Sequence data from phasmids will drive discovery-based research and contribute to the goals of the Earth Biome Project.

11. The i5k Workspace@NAL provides genome database services for orphaned arthropods

Poelchau, Monica¹; Chiang, Li-Mei²; Hsiao, Yi²; Hsu, Min-Chen³; Lin, Chun-Hung³; Wu, Chia-Tung⁴; and Childers, Christopher⁵

¹ Knowledge Services Division, USDA-ARS National Agricultural Library, Beltsville, MD, USA

² Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

³ Min-Chen Hsu, Graduate Institute of Communication Engineering, National Taiwan University

⁴ Graduate Institute of Computer Science and Information Engineering, National Taiwan University

⁵ Knowledge Services Division, USDA-ARS National Agricultural Library, Beltsville, MD, USA

Genome databases are critical infrastructure for transforming genome assemblies into effective tools for scientific discovery and knowledge. Databases provide a central access point for data, usually around a focal taxonomic group; tools for interacting with the data; services and expertise for data improvement and maintenance; and community data management support. The i5k Workspace@NAL (<https://i5k.nal.usda.gov>), a genome database for orphaned arthropod species, provides 1) a data access point to find and retrieve arthropod genomic and gene data, and 2) curation resources to manually improve existing genome annotations. i5k Workspace goals are to 1) improve the quality and depth of data and metadata; 2) facilitate and improve community annotation; 3) analyze and improve our platform architecture; 4) provide training and advice on manual annotation and data management; and 5) continue collaborations on joint software and standards development and implementation. Since 2013, the i5k Workspace has provided genome database services for orphaned arthropods via the Tripal and Apollo software, as well as the in-house developed Genomics Workspace software. As of July 2019, the i5k Workspace hosts 71 organisms and their genome assemblies. We have facilitated manual curation of over 15,000 gene models. We have

also released Official Gene Sets for 13 organisms, most recently using our updated GFF3Toolkit software (<https://github.com/NAL-i5K/GFF3toolkit/>). To facilitate data migration after genome assembly updates, we have developed workflows to update GFF3 files of the organisms that we host to new genome assemblies (https://github.com/NAL-i5K/coordinates_conversion and <https://github.com/NAL-i5K/remap-gff3>). Finally, we hold webinars every other month on topics of interest to the i5k Workspace community. The i5k Workspace's activities should provide enhanced infrastructure for arthropod genome data, resulting in improved scientific outcomes.

12. An Earth BioGenome Project Progress Meter.

Richards, Stephen; Caperello, Nicolette D.; Lewin, Harris A.; and Representing members of the Earth BioGenome Project Working Group

UC Davis Genome Center, University of California at Davis, Davis CA.

The Earth BioGenome Project (EBP) is an audacious project to sequence the genomes of 1.5 million described eukaryotic species. The EBP was formally launched in Nov 2018 at the Wellcome Trust London. Given the scale of the project, it is important to keep track of progress towards multiple goals, such as funding commitments, sequencing ready samples collected, and genomes sequenced. Although there will be multiple different priorities, the EBP roadmap is based on major phasing of the project is by phylogenetic wave. The EBP roadmap has 3 phases: phase 1: representatives of the approximately 9,500 taxonomic families, phase 2: representatives of ~180,000 taxonomic genera, and phase 3: all known eukaryotic species. Other goals in the roadmap include the establishment of pilot nodes and project agreements, and the agreement on standards during phase 1. Revising taxonomy at the end of phases 1, 2, and 3. Later goals include informing biodiversity preservation and conservation and enhancing ecosystem services starting during phase 2. We present the current state of progress towards the EBP goal at the start of the project. This will include status on funding commitments and genomes publicly available, as well as their qualities, and sample availability for both phase 1 and phase 3 goals.

Thursday:

Methods

13. Genome-wide False Duplications Identified by VGP Assembly

ByungJune Ko¹; Chul Lee²; Juwan Kim²; Erich D. Jarvis³, and Heebal Kim^{1,2}

¹ Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University Seoul, Republic of Korea

² Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

³ The Rockefeller University and HHMI, USA

One of the key processes to construct reference genomes is identifying and minimizing the causes of genome assembly errors. Recently produced high-quality reference genomes of the VGP make it possible to observe the typical error in previous reference genomes and newly generated genomes. Here we found that paralogues in short-read based genome are a major source of artifacts causing false duplication from highly heterozygous loci. We hypothesized that the false duplicated sequences have the same origin as from VGP primary and alternate haplotype assemblies, generated by phasing. We aligned the previous Sanger-based reference genome of zebra finch (*Taeniopygia guttata*) with VGP-level high quality primary genome and alternate assembly from the same animal using the Cactus alignment tool. We found false duplicated sequences genome-wide only in the old Sanger-based reference genome, 2.7% of the total genome length. We constructed phylogenetic trees for clustering on putative duplicated locus, and obtained 6,579 topologically reliable trees from 6,724 homology blocks containing duplicated

sequences only in old assembly. The 6,579 trees were merged to a final consensus tree, and we found that the duplicated sequences only in old genome were distinctly clustered with either the VGP primary genome and alternate-one, respectively. The genetic distance of sequences in the loci where false duplications occurred were larger than homozygous loci. Almost all (99.8%) of the false duplications loci had a gap between it and the rest of the genome, and this was extremely higher than the randomly estimated probability on the genome (30.4%), indicating that genome assemblers have the possibility of identifying such false duplications. Preventing and correcting false duplications will become more important, especially for polyploid species, which we predict will have greater false duplication errors even when sequenced with long-reads.

14. **A novel computational approach to haplotype-resolved assembly using sequence graphs**

Shilpa Garg^{1,5}; Yichen Wang²; John Aach¹; Heng Li³; Richard Durbin⁴; and George Church^{1,5}

¹ Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

² Boston University, Boston, Massachusetts, USA

³ Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

⁴ Department of Genetics, University of Cambridge, Cambridge, United Kingdom

⁵ Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, Massachusetts, USA

Reconstructing high-quality phased assemblies for related and unrelated individuals has important applications in comparative and evolutionary genomics. Through major genomics sequencing efforts such as the Personal Genome Project and the Vertebrate Genome and the Earth BioGenome Project (VGP-EBP), a variety of sequencing datasets of diploid genomes are becoming available. Current assembly approaches collapse haplotype sequences and are not designed to incorporate long-, short-read and long-range data in a haplotype-aware manner. Thus, building a haplotype-aware assembler capable of producing accurate and chromosomal-scale diploid genomes of any species, while being cost-effective in terms of sequencing costs, is a pressing need of the genomics community. We present a novel sequence graph based approach to diploid assembly that combines the advantages of accurate Illumina data, long-read Pacific Biosciences (PacBio) data and long-range information (from trios or Hi-C). In this approach, we construct an assembly graph from accurate Illumina data while retaining different alleles from individuals. We thread long-reads through this graph and partition these reads to different haplotypes. In unrelated cases, we additionally consider Hi-C haplotypes threaded through the graph. We demonstrate the effectiveness of our approach on simulated pseudo-diploid yeast genomes with different heterozygosity rates, and real data from human chromosome (chr22). We show that we require as little as 30× Illumina data and 15× PacBio data from each individual in a trio to generate chromosomal-scale phased assemblies. For unrelated cases, we show that we require as little as 45× PacBio data and 30× Hi-C data for high-quality phased assemblies. Additionally, we show that we can detect and phase variants from generated phased assemblies.

15. **Ensembl 2019**

Haggerty, Leanne; Allen, Jamie; Billis, Konstantinos; Girón, Carlos García; Hourlier, Thibaut; Izuogu, Osagie; Ogeh, Denye; Martin, Fergal J.; Howe, Kevin; and and Flicek, Paul

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

In response to the global genome sequencing and assembly efforts, Ensembl have accelerated work on annotation, generating almost 100 new annotations in the past year. These annotations span the vertebrate tree, including updates for key species, annotations for new species, and even annotations for multiple strains/ breeds of socioeconomically important species. The Ensembl gene annotation system generates high-quality, well-supported gene sets in an automated and parallel manner, ensuring consistency and efficiency. The system is under constant development to stay relevant and keep up with demand. Ensembl gene annotations are created using four main sources of evidence; 1. Long-read IsoSeq data, 2. Short-read RNASeq data, 3. Projected annotations from a suitable reference, and 4. A select set of vertebrate proteins from UniProt. At each locus, low quality transcript models are

removed, and the data are collapsed and consolidated into a final gene model plus its associated non-redundant transcript set. Where possible, priority is given to models derived from full-length transcriptomic data over homology data. Coverage of the vertebrates in Ensembl has been dramatically improved, for example, we have completed 51 fish annotations, including several cichlids, three inbred strains of Japanese Medaka, and the Atlantic herring. We have also added 19 bird annotations, including three species of kiwi and the Chilean tinamou. We have continued to expand the farm animal annotations with the introduction of both the maternal and paternal haplotypes for the *Bos indicus* x *Bos taurus* hybrid cattle assemblies and 12 non-reference pig breeds. In addition to all of the aforementioned, we have annotated many other species including koala, polar bear, alpine marmot, red fox, and tuatara, to name but a few. The annotated assemblies are available in Ensembl. Where available, annotated genomes include RNA-seq data, which can be viewed on the Ensembl genome browser. All data is available via the FTP site, REST API or Perl API.

16. **Visualization of Genome Scaffolding using Hi-C Paired-end Reads**

Harry, Ed and Ning, Zemin

Wellcome Sanger Institute, Hinxton, Cambridge, UK

Chromosome conformation capture techniques provide methods which can be used to quantify long-range interactions between genomic loci. These locations or contact points are nearby in 3D space, but may be separated by many bases in the linear genome. Mapped against a draft genome assembly, short read pairs obtained from cross-linking Hi-C library preparations are a fantastic data source for genome scaffolding and also the mapping information may be used in sorting chromosome structures as well as detecting mis-assembly errors. Summarising a set of interactions that potentially range over a whole genome and at every length scale is challenging, and has typically been achieved by visualising the interactions in contact matrices, or contact maps. Various workflows and file formats have been proposed to produce these maps, all with relatively high time and computational costs; both in producing and interacting with viewing the maps. We present new methods for producing, viewing and real-time editing of paired sequence contact maps, based on texture compression and GPU hardware acceleration. Our method produces maps directly from SAM formatted alignments in real-time, i.e. by reading in from a unix pipe, meaning that maps can be produced by directly from alignments, or from BAM / CRAM files. A human-sized genome can be processed from a BAM file in a few hours and only take tens of megabytes of disk space. Our map viewer uses OpenGL as it's rendering engine. Textures are directly loaded in their block compressed form. Users can pan and zoom their view in real time with no noticeable refresh time or input lag. Our viewer also features real-time editing of contact maps, allowing users to interactively re-scaffold genome sequences.

17. **Single chromosome sequencing to improve genome studies**

Iannucci, Alessio; Ferguson-Smith, Malcolm; Pereira, Jorge Claudio; Rovatsos, Michail; Kichigin, Ilya G.; Makunin, Alex I.; Pokorná, Martina J.; Altmanová, Marie; Trifonov, Vladimir A.; Kratochvíl, Lukáš; Stanyon, Roscoe R.; Lind, Abigail L.; Pollard, Katherine S.; Bruneau, Benoit G.; and Ciofi, Claudio
University of Florence, Italy

The production of valid chromosome-specific assemblies remains a significant challenge, even in the age of next-generation sequencing (NGS). Parallel sequencing of single chromosomes is an elegant approach to determine chromosome content and assign genome scaffolds to chromosomes. Physical chromosome isolation may be obtained either via flow sorting or mechanical microdissection. Both techniques can provide enough chromosome-specific genetic material to be used in NGS experiments. Here, we describe two case studies where chromosome isolation was used to assign genome scaffolds of the Komodo dragon (*Varanus komodoensis*) and the European pond turtle (*Emys orbicularis*) to chromosomes. Chromosome isolation was performed for *V. komodoensis* by flow sorting, while for *E. orbicularis* we used mechanical microdissection. In both cases chromosome-specific genetic material was used to prepare NGS libraries for sequencing with an Illumina platform. Chromosome-specific reads were then mapped onto reference genomes in order to assign each scaffold to the chromosomes. Results provided information of

chromosome content and allowed investigation of chromosomes evolution. Moreover, for *Varanus komodoensis* chromosome assignment revealed the gene content of sex chromosomes. The production of a high resolution genome for these species is important for investigating reptile genome evolution and to provide a high-quality reference that can be used to support population genomic studies of wild populations. Beside being a valid technique to assign genome to chromosomes, sequencing of isolated chromosomes can be useful to assist de novo assembly of vertebrate genomes that are particularly difficult to assemble, such as polyploid or very large genome sequences.

18. **Vertebrate genome assembly and annotation projects as course-based undergraduate research experiences**

Jue, Nathaniel K.¹; Slown, Corin¹; Rocha, Luis A.²; Willis, Stuart C.²; Johnson, Shannon; and Vrijenhoek, Robert C.³

¹ California State University, Monterey Bay

² California Academy of Sciences

³ Monterey Bay Aquarium Research Institute

Course-based Undergraduate Research Experiences (CUREs) are high impact educational experiences that can significantly affect both science identity and academic development of students. Training in genomics, which includes areas not traditionally part of undergraduate education in biology, is needed to prepare contemporary students for professional careers. Participation in a genome project is an excellent opportunity for students to gain needed training and experience in genomics. We have developed a course model that includes participation in a vertebrate genome assembly and annotation CURE using de novo genome assembly projects. To use CUREs to generate research contributions to novel genome assembly and annotation projects. All undergraduate researchers used a suite of established de novo assembly tools for genome (e.g. SPAdes, Platanus, and Meraculous) and transcriptome (e.g. Trinity, TransAbyss, Oasis) data, genome assessment methods (e.g. BUSCO and REAPR), annotation pipelines (e.g. Maker, Blast2GO, Trinotate), and variant calling pipelines (e.g. GATK and samtools) to generate novel scientific contributions to these genome projects. Multiple iterations of this course produced contributions to a transcriptome used to annotate a published genome for white sharks, a draft genome for the hemichordate *Poeciliopsis monacha* (N50=659,732 bp; 75% complete BUSCOs), and a preliminary annotation of a near-chromosome-level assembly for the sex-changing fish *Pseudanthias squamipinnis* (N50=27.9 Mb; 90.5% complete BUSCOs). Despite successes, results were not yet publishable by end of the semester. Of the participating students (40% underrepresented minorities), 30% were accepted into PhD programs and 40% went onto graduate programs; the majority of the remaining are working professionally in the sciences. This inquiry-based learning model provided students with meaningful educational experience and contributed to the progress of research in de novo genome assembly and annotation. These projects suggest that appropriate engagement with undergraduate researchers may be a promising means of advancing collaborative genome projects.

19. **False Gene Losses Corrected by VGP Assembly**

Kim, Juwan^{1¶}; Lee, Chul^{1¶}; Ko, Byung June²; Rhie, Arang³; VGP assembly group; Kim, Hee-ba^{1,2,6*}; and Jarvis, Erich D.^{4,5*}

[¶] Both authors contributed equally to this work.

¹ Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

² Department of Agricultural Biotechnology and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

³ Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA

⁴ Laboratory of Neurogenetics of Language, Rockefeller University, 1230 York Avenue, New York, NY 10065, USA

⁵ Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815, USA

⁶ C&K genomics, Seoul, Republic of Korea

Gene loss can play an important role in the evolution of various traits, including loss or gain of traits. However, many current assemblies based only on sequencing platforms with short (<1,000 bp) read lengths have been found to contain various errors that can lead to false gene losses. Here, we compared two different versions of platypus genome assembly based on short-read only and the long-read, more complete, Vertebrate Genomes Project (VGP) platypus assembly (supported by Guojie Zhang) to detect false gene loss by using the program Comparative Annotation Toolkit (CAT). We found numerous cases of false gene loss in the existing short read length genome assembly where each of them could be classified into one of the following eight categories: 1) totally missing gene; 2) gene mapped with below 50% of coding sequences; 3) fragmented gene; 4) intra-scaffold split gene; 5) intron-exon junction disruption; 6) frameshift; 7) Ns in coding sequences; and 8) premature stop codon. These types of errors have been summarized by the combination of assembly platforms and submodules within CAT. In order to support annotation projection at with gene synten, we compared the order of genes within each assembly. This analyses confirmed the high contiguity of the new VGP assembly but also some genomic regions with a different order of genes at the chromosomal level between the short-read and long-read based VGP assemblies. The new VGP assembly corrected most of false gene loss cases and substantially reduced the error rate.

20. The past, present, and future of Arima-HiC for genome assembly: An overview of Arima-HiC sample prep and scaffolding of VGP genomes

Schmitt, Anthony; De La Torre, Chris; Reid, Derek; Mac, Stephen; Zhou, Xiang; Tan, Catherine; Won, Melissa; and Selvaraj, Siddarth

Arima Genomics, Inc., San Diego, CA, USA

A predominant approach to genome assembly involves scaffolding contigs into chromosomes, often leveraging chromosome-spanning linked-reads obtained from Hi-C. This unique value of Hi-C data has resulted in its broad utilization across assembly projects, including the Vertebrate Genomes Project (VGP). Our objective is to provide superior Hi-C products, services, and support to serve the genome assembly community. Here, as a VGP technology partner, we demonstrate the utility of high quality Arima-HiC data for chromosome-scale scaffolding of VGP genomes, and highlight ongoing technology developments to better serve the evolving needs of genome assembly projects. Hi-C data was generated using the Arima-HiC kit. Arima-HiC data was integrated into the VGP Phase I pipeline, combining PacBio sequencing, 10X Genomics linked-reads, Arima-HiC linked-reads, and Bionano optical maps. The automated Arima-HiC workflow was developed on the Agilent Bravo liquid handler. To best serve the genome assembly community with consistent high quality Hi-C data, we developed Arima-HiC technology. Our kits and services incorporate optimized protocols for various sample sources (plants, animals), types (tissue, blood, cells), preservation methods and sample quantities. Through our ongoing partnership with the VGP, we have currently delivered 56 high quality Arima-HiC datasets from tissue and blood samples derived from 18 fish, 22 birds, 11 mammals, 4 reptiles, and 1 amphibian. Of these, 39 genomes have incorporated Arima-HiC data to produce high quality draft or complete annotated assemblies, all with >10Mb scaffold NG50s (AVG=73.9Mb). Going forward, we focus on meeting the unique requirements of forthcoming assembly projects, such as developing automated and ultra-low input workflows, and new Hi-C chemistries for improved base polishing and haplotype phasing. Arima-HiC technology is a powerful tool for generating accurate chromosome-scale genome assemblies. The technology has been broadly validated, such as the extensive utilization of Arima-HiC technology to scaffold VGP genomes. We now seek to develop additional workflow and sample prep innovations to meet the growing needs of genome assembly projects.

21. Building an assembly army for the VGP: challenges and first successes of building high-quality genomes by the Berlin student's team

Uliano-Silva, Marcela; Driller, Maximilian; Caswara, Calvinna; Vafadar, Majid; Rhie, Arang; Jarvis, Erich D.; and Mazzoni, Camila

The Vertebrate Genomes Project's goal of assembling 70,000 thousand reference genomes encompassing all vertebrates presents unprecedented logistical challenges, such as (i) the need for specialized manpower required to run and monitor genome assembly outputs and (ii) the need for pipeline optimization to decrease computational costs. We gathered volunteer bioinformatics masters students from the Free University in Berlin to assemble VGP genomes, and here we present their first results, and discuss the challenges they encountered during the work. Monitored by experienced researchers (Uliano-Silva, Rhie et al.), three students (Caswara, Driller, Vafadar) have assembled at least one genome each, using the VGP v1.5 Assembly Pipeline integrated into the DNAnexus platform. Throughout the process, they learned how to manipulate four data types (Pacbio long reads, 10X Genomics linked reads, Bionano next-generation maps and Arima Hi-C reads), and they learned how to run several programs integrated into the pipeline, while producing high-quality genomes for the scientific community. They were instructed over several hours of video conferences, dedicating 3-4 working hours weekly to the project, and they assembled both avian and mammalian genomes. The diverse nature of these genomes presented them with distinct computational challenges that were reported back to the experienced researchers for pipeline optimization. They have assembled three complete genomes so far (*Sterna hirundo*, *Merops nubicus* and *Catharus ustulatus*). These genomes meet the VGP quality standards (3.2.4.QV40) and are at the final stages of manual curation. The work of specialized volunteers allows (i) a learning opportunity for assembling high-quality genomes, (ii) scaling up the production of genomes and (iii) feedback for pipeline optimization. This learning through doing pipeline for training provides mutual benefits for both the project and for the students.

Biological Discoveries Section

22. Analysis of Microsatellite Abundance in Mammalian Genomes; Revisiting Peto's Paradox

Jeong, Heesul and Kim, Heebal^{1,2,3*}

¹ Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Korea, 151-742

² Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Korea, 151-742

³ Institute for Biomedical Sciences, Shinshu University, Nagano, Japan

Microsatellites (MSs) are rapidly evolving component in a genome with higher mutation rates (locus/generation) than other mutation types like point mutation (nucleotide/generation). Giving more diversity in a genome, MS polymorphism is often related with phenotypic variation, as MS instability is often reported in colorectal tumors and other various cancer types. Interestingly, mammals show the highest MS abundance among all eukaryotic classes of which genome sequences available, implying their significant role in mammalian genome evolution. Meanwhile, there is an ironic phenomenon for mammals which is known as Peto's paradox since mammalian species have constant rate of cancer irrelevant with their body size and lifespan. Body size and lifespan are risk factor for carcinogenesis since they are related with number of cell division in a lifetime, but uniform cancer rate in vast range of mammalian species strongly implies underlying global mechanism for cancer repression in the clade. In this context, the fact that previous study has reported negative and significant correlation between MS content in a genome and body size in 31 mammalian species may give another view for the paradox. Here, with use of recent version of repeat sequence database in the RepeatMasker program, we re-confirmed the tendency and found palindromic subset of repeat sequences that have significant correlation with the risk factors. We also observed scarcity of the repeat sequences in avian lineage including chicken, supporting its specific presence in mammalian genomes. We prioritized testing the repeats' relationship with functional units, checking presence of enriched genes near the repeats. Interestingly, from the list of enriched genes, we could find some cancer-related genes like IDH1. We expect that newly releasing high-quality genome and annotation data from VGP project will enable us to validate

the process relied on UCSC data, which is also suitable for further study to understand the candidate repeat sequences in context of Peto's paradox.

23. Rules of Amino Acid Convergences: Not How Many, but Who in Avian Vocal Learning Clades

Lee, Chul¹; Cho, Seoae¹; Kim, Kyuwon¹; Yoo, Dongahn¹; Han, Jae Yong¹; Lee, Hong Jo¹; Gedman, Gregory²; Pfenning, Andreas³; Kim Heeal^{*2}; and Jarvis, Erich D.^{*2}

¹ Seoul National University;

² Rockefeller University and HHMI;

³ Carnegie Mellon University.

Vocal learning, the ability to imitate vocalizations with convergent neural pathways in specific brain regions, is a convergent trait observed in a rare few animal lineages. Molecular mechanisms of the behavioral and neural convergence still remain open questions. Here, we analyzed avian genomes including vocal learning clades and found principles of amino acid convergences to explain the trait and random effects. For vocal learning birds and clade-wide control sets, we identified genetic variants at amino acid, codon and nucleotide levels, illuminated correlations between the variants and the *product of origina branch lengths*, and confirmed no preponderance of molecular convergences in avian vocal learners. Nevertheless, genes with convergent amino acid substitutions vocal learning birds were enriched for learning. A subset of these convergent genes and their convergent substitutions were under positive selection in avian vocal learners, and these same genes showed a higher proportion (75%) of differential expression in song nuclei compared to the closest control set (27%) or singleton orthologous gene set (35%). A key candidate gene was the D1B dopamine receptor (DRD5), which was supported by multiple lines of evidence for vocal learning, as well as genes human specific amino acid substitutions like FOXP2. Our findings reveal insights into macro-evolution of vocal learning.

24. AgriVectors: Sequencing, omics resources and systems biology portal for plant pathosystems and arthropod vectors of plant diseases

Saha, Surya¹; Hunter, Wayne²; Mueller, Lukas A.³; and The AgriVectors Consortium³

¹ Boyce Thompson Institute, Ithaca, NY 14853

² USDA ARS, U.S. Horticultural Research Laboratory, Ft. Pierce, FL 34945

³ Boyce Thompson Institute, Ithaca, NY 14853

Arthropod vectors of pathogens cause enormous economic losses and are a fundamental challenge for sustainable increases in food production. To more effectively fight plant diseases, data pertaining to a disease system needs to be consolidated, made searchable and amenable to data mining. The proposed AgriVectors platform is an open access and comprehensive resource for growers, researchers and industry working on plant pathogens and pathosystems spread by arthropod vectors. We are sequencing a range of hemipteran arthropod genomes for developing comparative genomics resources for important insect vectors. The portal connects established public repositories with pathosystem-specific data repositories. The AgriVectors system will provide tools to enable technologies such as RNAi, CRISPR, screening bioassays, etc. to leverage current and emerging knowledge across disciplines. The portal will be based on the Citrusgreening.org (<https://citrusgreening.org/>) community resource for the Huanglongbing pathosystem that was developed as a model for systems biology of tritrophic disease complexes. It includes a biochemical pathway database for each organism in this disease complex, and an expression atlas for the psyllid vector and citrus host. The hemipteran genomes will be sequenced with low input DNA and RNA protocols on the Pacbio Sequel2 instrument. Hi-C will be used for long range scaffolding. The genomes of Pear psylla, Lime psyllid, African citrus psyllid, Potato psyllid, Bindweed psyllid, Tarnished plant bug, Western big-eyed bug, Citrus mealybug, Longtailed mealybug, leafhopper and aphid species will provide a foundation for comparative analysis across hemipteran species. The AgriVectors portal will extend this model beyond gene-centric omics data to the broader Pathosystem-wide information, with integrated pest management, behavioral, plant health, soil health and climate

data to incorporate rapid phenotyping information from research trials, building a foundation for more effectively identifying solutions to combat plant diseases.

25. Whole-genome alignments of publicly available high-quality bird genome assemblies highlight phylogenetic profiles of structural variants

Secomandi, Simona¹; Formenti, Giulio²; Rhie, Arang³; Chiara, Matteo¹; Poveda, Lucy⁴; Francoijs, Kees-Jan⁵; Bonisoli-Alquati, Andrea⁶; Canova, Luca⁷; Gianfranceschi, Luca¹; Horner, David Stephen¹; Jarvis, Erich D. ²; and Saino, Nicola⁸

¹ Department of Biosciences, University of Milan (Milan, Italy);

² The Rockefeller University (New York, NY, USA)

³ National Human Genome Research Institute, National Institutes of Health (Bethesda, Maryland, USA)

⁴ Functional Genomics Center of Zurich, University of Zurich, (Zurich, Switzerland)

⁵ Bionano Genomics (San Diego, CA, USA).

⁶ Department of Biological Sciences, California State Polytechnic University, Pomona (Pomona, CA, USA)

⁷ Department of Biochemistry, University of Pavia (Pavia, Italy)

⁸ Department of Environmental Science and Policy, University of Milan (Milan, Italy)

The advent of new sequencing technologies has allowed the assembly of vertebrate genomes of unprecedented quality. These genomes have the potential to unveil phylogenomics profiles of structural rearrangements at the greatest level of resolution. Here we performed systematic comparisons within a collection of high-quality genomes from 17 closely and distantly related bird species, including the barn swallow (*H. rustica*) draft genome assembly produced in 2018 at the University of Milan and 10 genome assemblies made available within the first phase of the Vertebrate Genomes Project. All genomes were repeat-masked using WindowMasker and RepeatMasker. They were then aligned to the chicken (*Gallus gallus*) reference genome using Minimap2, generating the alignment coordinates. A custom script was used to process the alignments coordinate files. The regions of the chicken genome that consistently aligned to the other bird genomes included in the dataset were considered for the detection of structural variants. We detected and analyzed four major types of structural variations SVs (expansions, inversions, inter chromosomal translocations, translocations of uncertain type). In general, a high degree of concordance was observed between structural variants profiles and the phylogeny of the species included in the study. The approach can be successfully employed to detect structural rearrangements. While offering indirect evaluation of the quality of the genomes herein considered, this result further demonstrates the value of high quality genome assemblies in comparative genomic studies.

26. Origins and evolution of extreme longevity in the adaptive radiation of rockfish

Kolora, Sree Rohit Raj¹; Stubbs, Alexander¹; Chatla, Kamalakar¹; Jainese, Conner²; Seeto, Katelin²; Bachtrog, Doris¹; Love, Milton S²; and Sudmant, Peter H¹

¹ Integrative Biology, University of California, Berkeley

² Marine Science Institute, University of California, Santa Barbara

The maximum lifespan of vertebrates ranges from 5-weeks in the pygmy goby to 400 years in greenland sharks representing 4000-fold variation in longevity. However, the evolutionary mechanisms shaping this extensive variation remain unknown. Rockfish of the genus *Sebastes* exhibit extreme diversity in lifespan ranging from ten years to more than two centuries, providing an exceptional model system to study longevity. Furthermore, these species have undergone a recent extensive adaptive radiation into more than 100 different lineages across diverse habitats over the past 6 million years. Rockfish are one of the most rapidly speciating vertebrate clades indeed. To understand the genetic basis of variation in longevity, adaptation and speciation we sequenced the genomes of 61 different Pacific coast rockfish species ranging from Alaska to Chile including seven reference quality genome assemblies. We find that rockfish species exhibit an exceptional range in diversity (heterozygosity 0.002-0.009) spanning a range similar to

that observed across all chordates. Unlike other ray-finned fishes duplications appear to have played a minimal role in speciation. Repetitive elements however, particularly retro-elements and associated transposase gene families, show substantial clade specific expansions. Furthermore, we find lineage specific selection of acetylcholine signalling pathways in the genus *Sebastes* may have played a crucial role in their adaptive radiation and extreme longevity phenotypes. Together our data represents a rich resource to study the genetic basis of variation in longevity and the processes governing speciation.

27. **Genomic signatures of the landlocked adaptations in Taiwan gobies (*Rhinogobius* spp.)**

Wang, Tzi-Yuan¹; Huang, Shih-Pin¹; Wu, Yu-Wei²; Liao, Te-Yu³; Chaw, Shu-Miaw¹; Wang, Feng-Yu⁴

¹ Biodiversity Research Center, Academia Sinica, Nankang, Taipei, TAIWAN. Email: tziyuan@gmail.com

² Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan

³ Department of Oceanography, National Sun-Yet-Sen University, Kaoshiung, Taiwan

⁴ Taiwan Ocean Research Institute, National Applied Research Laboratories, Kaohsiung, Taiwan

The genus *Rhinogobius* (subfamily: Gobionellinae; Family: Oxudercidae) includes 66 recognized species, ten of which are indigenous to Taiwan. *Rhinogobius* is small, streamlined in shape and dwells mainly in freshwater of tropical and temperate regions in eastern Asia. Its species are often sexually dimorphic. The female generally attaches its eggs onto stones/substrates, where the male deposits adherent sperm-containing trails and then guards the clutch until the eggs hatch. Distinct from other genera of gobioid fishes that inhabit diverse environments such as coral reefs, seashores, brackish estuaries, and freshwater rivers, the *Rhinogobius* are strictly amphidromous or landlocked. The freshwater goby *R. rubromaculatus* is restricted to Taiwan and regarded as an excellent biological indicator of water quality and trophic conditions. In this study, we aimed to decipher the genomic signatures of this freshwater goby's landlocked adaptation. We anticipate that our gathered data will significantly fill gaps in the goby genome database. The estimated flow cytometry genome size (FCGS) of *R. rubromaculatus* is 1.23 ± 0.10 Gb/1 C. We produced an assembly of 946.8 Mb, the largest among the sequenced gobies, which primarily comprises $16.68 \times$ PacBio and $18.63 \times$ Nanopore long reads, with $N50 = 1.498$ Mb. The genome is about 77% of the FCGS estimate, but its completeness is estimated to be 89.6% using BUSCO. The quality of our goby genome could be regarded as by far the best in the family Oxudercidae. We are also sequencing the transcriptome of *R. rubromaculatus* and its two closely related species for future comparison.

28. **Novel approach to detect interspecies-level directional selection based on divergence from ancestral sequence and polymorphism data**

Yoo, DongAhn¹; Lee, Chul¹; and Kim, Heebal^{1,2,3,4*}

¹ Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea

² Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

³ C&K Genomics, C-1008, H Businesspark, 26, Beobwon-ro 9-gil, Songpa-gu, Seoul, Republic of Korea

⁴ Department of Interdisciplinary Genome Sciences and Cell Metabolism, Institute for Biomedical Sciences, ICCER, Shinshu University, 8304 Minami-Minowa, Kami-Ina, Nagano 399-4598, Japan

McDonald and Kreitman test (MKT) is a straightforward and simple non-parametric approach for detecting selection signature without relying on complex evolutionary model. This approach compares protein coding genes of two species by measuring the degree of functional single nucleotide divergences (or substitutions) between species and functional single nucleotide polymorphisms within species. One inherent limitation of MKT is its uncertainty in determining the direction of selection. Here, we developed a pipeline that can resolve this issue via ancestral sequence reconstruction. The pipeline was assessed by investigating recent human evolution using population genome data of Great Ape lineage and forward simulation. As a result of population data analysis, twice as many positively selected genes of human were detected in the standard MKT (722 and 740 when compared against

chimpanzee and bonobo) than the novel (395), with average functional substitutions of ~ 9 per gene. Tracing conservation of individual substitutions with the genomes from Great Ape lineage, we found that almost half of these functional substitutions are likely to emerge outside human lineage and result in false positive calls of up to ~ 370 genes. Similarly, non-functional polymorphisms presumably originated from pan lineage restrains standard MKT from detecting genes including *SALL3* related to GO term “neurogenesis”, and we also found that these variants could influence up to ~ 900 non-selected genes. In comparison with maximum likelihood-based dN/dS, higher number of genes were commonly found with the novel MKT compared to the standard MKT, while applying the novel method also strengthened the results of dN/dS by filtering out genes with excessive functional polymorphisms. Finally, using the forward simulation data, we traced the accumulated variants with positive fitness index to define positively selected genes and evaluated the performance of the novel MKT in detecting them, highlighting the robustness and strength of the novel pipeline over the standard MKT.

Social Media

Hashtags:

#genomes2019

#vgp2019

#ebp2019

Twitter:

- VGP: @genomeark
- G10K: @Genome10K
- EBP: @EBPgenome
- Sanger: @SangerVGP
- MPI-CBG Dresden: @mpicbg
- Plenary lecture, Rebecca Johnson: @DrRebeccaJ

Twitter: Technology partners:

- Arima: @ArimaGenomics
- Bionano: @bionanogenomics
- DNANexus: @dnanexus
- Dovetail: @DTGenomics
- Illumina: @illumina
- Nanopore: @nanopore
- Pacific Biosciences: @pacbio

Websites:

- VGP: <https://vertebrategenomesproject.org>
- EBP <https://www.earthbiogenome.org>
- GenomeArk: <https://vgp.github.io>
- G10K: <https://genome10k.soe.ucsc.edu>
- Bat1K: <https://bat1k.ucd.ie>
- B10K: <https://b10k.genomics.cn>
- Sanger Genomes 25 Project: <https://www.sanger.ac.uk/science/collaboration/25-genomes-25-years>
- UCSC VGP Genome Browser: <https://hgdownload.soe.ucsc.edu/hubs/VGP/>

Sponsors

