California Conservation Genomics Project (CCGP)

A collaborative effort to conserve California flora and fauna using conservation & landscape genomics of threatened, commercially exploited and ecologically important species.

https://sites.lifesci.ucla.edu/eeb-CCGP/

Or google “CCGP UCLA”
Elements of the CCGP

- ~230 species of plants and animals
- 150 individuals per species
- High quality reference genome for all spp
- Whole genome resequencing (WGS)
- Best available GIS/imagery

Data for management/policy:
- State and federal agencies
- Private and public land managers
Today’s goals:

• Data flow & goals
• WGS: options & issues
• Reference genome progress updates

Questions and Discussion
CCGP workflow

Reference genome

- HMW single tissue plus organs for Reference genome/RNA

Data processing pipeline

- Raw data QC
- Reading Mapping & QC
- Variant Calling and QC

Data analysis pipeline

- Calculate genetic diversity metrics
- Generate Maps
  - Heterozygosity or pi
  - Genetic turnover
  - Interpolated admixture plots
- Mapping Tools
- Agency Report & Publications

CCGP data pipelines

- Agency Report & Publications

Project-based research

- Agency Report & Publications

HMW DNA extraction and library prep

- PacBio HiFi sequencing (at UCD)

OmniC extraction and library prep

- OmniC Library seq (at UCLA or UCB)

RNA extraction and library prep

- Transcriptome sequencing (at UCLA or UCB)

Whole genome resequencing (WGS)

- Sample tissue collection for resequencing

- DNA Extraction

- WGS Library Prep

- Library Pooling

- Generate short read sequence data

- Submit Data to CCGP and NCBI

- Project-based research

- Agency Report & Publications

- Genome assembly

- Genome upload to NCBI

- Genome Annotation - NCBI

- Ship to UCD

- Ship to UCSC

- Ship to UCLA

HMW DNA extraction and library prep

- Pacific HiFi sequencing (at UCD)

RNA extraction and library prep

- Transcriptome sequencing (at UCLA or UCB)

Whole genome resequencing (WGS)

- Sample tissue collection for resequencing

- DNA Extraction

- WGS Library Prep

- Library Pooling

- Generate short read sequence data

- Submit Data to CCGP and NCBI

- Project-based research

- Agency Report & Publications

- Genome assembly

- Genome upload to NCBI

- Genome Annotation - NCBI

- Ship to UCD

- Ship to UCSC

- Ship to UCLA

Color Legend – Activity Location

- Project
- CCGP
- UC Core
WGS General Guidelines

• Number of Samples per Species: ~150
• DNA Sequencing Facilities: Sequencing must be done at a UC facility (Davis, Berkeley, and UCLA)
• Sequencing Coverage: we’re aiming for 10x. Remember, organelles (bigger for plants), duplicate reads can eat up 10-20% of reads.
• Plan carefully with a good informatician!
All data need to pipe over to CCGP, and must be made publically available
Timeline & Public Data Release

• All resequencing should be completed by December 2021
• All resequencing data should be shared with CCGP as it is produced and no later than December 2021
• Data submission to NCBI Short Read Archive (SRA) should be completed by March 2022
• Data submission to the NCBI SRA is the responsibility of individual PI
• We are in the process of setting up an umbrella BioProject, please check back for updates or contact Erin (etoff@ucla.edu) for the BioProject ID
Submitting Your Data to CCGP

• Data generated by the CCGP Mini-Core will automatically be sent to CCGP

• For projects that do not utilize the Mini-Core, *either*:

1) Include CCGP in the submission process
   - Submission processes vary; this may not be feasible
   - Include Erin (etoff@ucla.edu) in your submission
   - Please submit metadata (see instruction document)

2) Submit fastq files and metadata after sequencing has been completed
   - We are working to setup a FTP server for easy data transfer
   - If you already have sequence data, please wait for an announcement
   - Submission will consist of form submission with project information and metadata followed by raw sequence upload to our servers via FTP
WGS Library Prep and Sequencing: Options

1. Do it yourself (read the instruction sheet, remember that to get the best price your libraries may need to be pooled with others)

2. Have your Core do it (easy, but expensive, and the cost is on you)

3. Use the CCGP Mini-Core (we’ll extract, make libraries, and pool within and across projects)
What the Mini-Core Needs

• High quality tissue or extracted gDNA
• We can only do CCGP samples
• We can only bill back from the same fund that you were allocated (so save some of it!)
## Mini-Core Pricing

<table>
<thead>
<tr>
<th>Mini-Core task</th>
<th>per sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA extraction from tissue + library preparation</td>
<td>$30</td>
</tr>
<tr>
<td>Library preparation from user supplied gDNA</td>
<td>$20</td>
</tr>
<tr>
<td>Additional QC, if needed</td>
<td>$4</td>
</tr>
<tr>
<td>Additional extraction, if needed</td>
<td>$6</td>
</tr>
<tr>
<td>EDTA cleanup (SPRI based), if needed</td>
<td>$3</td>
</tr>
</tbody>
</table>
Other Important Consideration

• Plants, slimy invert are the hardest
• Only fresh material!
  • No herbarium sheet material
  • No formalin preserved
  • No “hard” tissue (feathers, bone, nail)
• Plants, at least 1 g fresh leaf
• Small invert, at least 0.2 g fresh material
• Verts, at least 0.2 g fresh material
• Nucleated RBCs, 30μl; anucleated, 1200μl
• gDNA, 125ng – 500ng, quality check is on you
More Important Consideration

• Ship on Monday or Tuesday (UC Mail = 😞)
• Ship legally (declare dry ice, etc.)
• Use FedEx
• **Submit all samples for a project at once**
• Sample quality:
  • If >5 are poor, we’ll stop and contact you
  • If 5 or fewer poor, we’ll just keep going
• We will send to a Core (probably UCLA), let you know, and have billing sent to you
• We will also have data piped to us
• We will keep samples/extracts for 6 months
Erin to talk about Reference Genome progress...
CCGP Reference Genome Update

Reference genome

Reference Sample Collection

HMW DNA extraction and library prep
PacBio HiFi sequencing (at UCD)

OmniC extraction and library prep
OmniC Library sequencing (at UCLA or UCB)

RNA extraction and library prep
Transcriptome sequencing (at UCLA or UCB)

Genome assembly

Genome upload to NCBI

Genome Annotation - NCBI

Current Progress

- Assembly Complete
- Sequence Data Awaiting Assembly
- DNA Extracted & in Seq. Queue
- Tissues in Queue
- Unsubmitted

Color Legend – Activity Location

Project  CCGP  UC Core
Current Progress

- **HiFi**: 66
- **OmniC**: 59
- **RNASeq**: 29

Tissue Submissions
(121 expected)

- HiFi: 66
- Omnic: 59
- RNASEq: 29
CCGP Reference Genome Update

Current Progress

<table>
<thead>
<tr>
<th>Tissue Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi</td>
<td>66</td>
</tr>
<tr>
<td>Omnic</td>
<td>59</td>
</tr>
<tr>
<td>RNASeq</td>
<td>29</td>
</tr>
</tbody>
</table>

Tissue Submissions (121 expected)

- HiFi: 66
- Omnic: 59
- RNASeq: 29

RNASeq Tissues:
- Submit up to 5 different tissue types
- Different individuals & life stages

For Reference Tissue Submission Info:

https://sites.lifesci.ucla.edu/eeb-CCGP/specimens/

Contact Erin: etoff@ucla.edu
CCGP Reference Genome Update

Current Progress

Tissue Submissions (121 expected)

<table>
<thead>
<tr>
<th></th>
<th>HiFi</th>
<th>Omnic</th>
<th>RNASeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submit</td>
<td>66</td>
<td>59</td>
<td>29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>In process Library Prep &amp; Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiFi</td>
<td>16 (2 completed assemblies)</td>
</tr>
<tr>
<td>Omnic</td>
<td>19 (first deep seq run in January)</td>
</tr>
<tr>
<td>RNASeq</td>
<td>0 (extractions start with opening of mini core)</td>
</tr>
</tbody>
</table>

Count

Assembly Complete
Sequence Data Awaiting Assembly
DNA Extracted & in Seq. Queue
Tissues in Queue
Unsubmitted
Current Progress

Timeline

Tissues Submission: July 2021
Data Generation: November 2021
Assembly: December 2021
CCGP Reference Genome Update

PacBio HiFi Long Read Assembly

Western pond turtle
* Emyx marmorata

- Estimated genome size (related species): 2.6 Gbp
  - More like 2.3 Gbp?
- After 3 SMRT Cells: ~30X (2.6 Gbp)
  - Output 80.4 Gb

**General statistics**

<table>
<thead>
<tr>
<th>Primary</th>
<th>Alternate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of genome</td>
<td>2,355,098,844</td>
</tr>
<tr>
<td># Sequences</td>
<td>184</td>
</tr>
<tr>
<td>Contig N50</td>
<td>115,383,402</td>
</tr>
<tr>
<td>Contig L50</td>
<td>6</td>
</tr>
<tr>
<td>Longest contig</td>
<td>306,034,025</td>
</tr>
<tr>
<td># Gaps</td>
<td>0</td>
</tr>
</tbody>
</table>

**BUSCO Scores (n=954, metazoal)**

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>D</th>
<th>F</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>98.40%</td>
<td>97.00%</td>
<td>1.40%</td>
<td>0.90%</td>
<td>0.70%</td>
</tr>
<tr>
<td>A</td>
<td>94.50%</td>
<td>92.70%</td>
<td>1.80%</td>
<td>1.00%</td>
<td>4.50%</td>
</tr>
</tbody>
</table>

**Mercury (kmer) analysis**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>base-call QV (hap)</td>
<td>66.03</td>
<td>66.35</td>
</tr>
<tr>
<td>k-mer completeness (hap)</td>
<td>94.56</td>
<td>90.79</td>
</tr>
<tr>
<td>base-call QV (full)</td>
<td>66.18</td>
<td></td>
</tr>
<tr>
<td>k-mer completeness (full)</td>
<td>99.27</td>
<td></td>
</tr>
</tbody>
</table>

Image from Yathin S Krishnappa (CC BY-SA 3.0)