**CCGP Genome Release Papers with the Journal of Heredity**

**CCGP Genome Resources** are brief reports of 1000-1500 words and up to four display items (including Table 1) that describe CCGP reference genome assemblies. They can be longer if you wish, but our target is short, succinct, rapidly-published release papers. It is important for the CCGP project that as many of our new reference genomes as possible, and hopefully all of them, are announced in this format.

* CCGP will cover the cost of publication, including open access fees.
* CCGP will contribute the majority of the technical details, including the methods, results, and data availability sections. Please note that these components will come from multiple team members and, as the primary author(s), it is the PIs responsibility to revise the provided sections so that the text is cohesive.
* The PI is responsible for the introduction, discussion, and other components specific to their species, including some of the results.

**Submission Instructions**

All Genome Resources papers should be submitted through the ScholarOne Manuscripts portal: [https://academic.oup.com/jhered/pages/Msprep\_Submission.](https://academic.oup.com/jhered/pages/Msprep_Submission) To ensure that the CCGP is billed properly, there must be 1) at least one author who is a member of the American Genetic Association (AGA), and 2) authors must note that the manuscript is part of the CCGP in their cover letter. Having one author who is an AGA member saves us a lot of money, and is a requirement for all papers.

**Template**

With approval from JOH, we have tailored the JOH Genome Release guidelines for CCGP releases. Please use the following template in preparation of your article. Highlighted areas describe who is primarily responsible for each section, although all authors should read, contribute as they can, and approve the entire manuscript.

# Title Page *(PI contributes)*

***Title:*** This should include the common and scientific name of the organism(s) and a succinct description of the resource. It should be no longer than 75 characters. Example suggested title: “The reference genome of the southwestern pond turtle, *Actinemys pallida*”.

***Author names:*** Provide complete first and last author name(s) as you wish them to be in the final publication. **Note:** Please see our advice on authorship for Genome Resource publications at the end of this template.

***Author affiliation(s):*** Each affiliation should be preceded by a superscript number that corresponds to the author list.

***Corresponding author email:*** provide their primary email address.

# *Running title:* A shortened version of the full title. Example suggested running title: “Southwestern pond turtle genome.”

# Abstract *(PI contributes)*

A single paragraph of about 150 words, without references to the text or literature cited. Do not use acronyms and complex abbreviations. Please emphasize the value of this species for California conservation and fundamental importance to ecosystem functionality, as appropriate.

# Keywords *(PI contributes)*

3-6 keywords or short phrases **not included in the title**. Please include ‘California Conservation Genomics Project’ and ‘CCGP’ unless they are in your title.

# Main Text

# *Introduction* *(PI contributes)* (Target: 300 words): Provide the rationale for generating the Genome Resource and its broader applications to a research community. Given that this is a CCGP project, please consider why this genome is important for California, including its role in California conservation or ecosystem-health studies. Note: Please cite the main JOH article describing the CCGP, (we will provide the reference when it is written and accepted, hopefully before the end of 2021) and any other relevant published CCGP Genome Resources papers for other species. If you are unsure if relevant CCGP Genome Releases have been finalized, please contact Erin Toffelmier or Courtney Miller.

# *Methods* (*PI and* *CCGP contribute)* Target: 200-300 words: Please include each of the following components. This will generally follow a standard CCGP format. If the genome release includes additional data generated independently of CCGP, please work with us to incorporate that information appropriately. For example, some projects are using previously collected HiC data to help with assembly. Contact Erin Toffelmier or Brad Shaffer if you need assistance.

***Biological Materials*: *(PI describes, CCGP can supplement if needed)*:** Describe the sources of the biological materials (including museum accessions, voucher IDs, etc.), including georeferenced coordinates of the specimen. If you are concerned about providing precise locality information for endangered species, please provide locality information at the level that you feel comfortable and note that more precise information is being withheld due to conservation concerns. Identify a contact person if more information is needed (probably the corresponding author).

***Nucleic acid library preparation (CCGP will provide)*:** Describe details of DNA extraction and preparation of nucleic acid libraries.

***DNA Sequencing and Genome Assembly (CCGP will provide)*:** List the sequencing platform(s) used, with detailed assembly and scaffolding methods. To encourage reproducibility, programs (including the version of the software) used for data processing should be clearly cited and be listed in a table **(Table 1**, see example below**)** and presented in a logical format, which is usually the order in which analyses were performed. This will follow a standard CCGP format.

***Specific parameters or variables used in each of the programs should be clearly stated (CCGP will provide)*:** If custom codes and scripts are used for data processing, authors should indicate where they can be accessed. CCGP will deposit custom codes and scripts in repositories (probably GitHub) and provide links.

# Results *(CCGP will provide)*

Target: 250-300 words. Include standard metrics of assembly completeness and coverage, including estimated genome size, N50 (and/or k-mer) statistics for contigs and scaffolds, longest contigs, number of gaps, and BUSCO scores. This will follow a standard CCGP format, and will be summarized in **Table 2** (see example, below).

# Discussion *(PI contributes)*

Target: 300 words. Include a succinct discussion of the value of the resource, its potential value for conservation and biodiversity studies generally, and what we have learned compared to related genomes. For example, how does this genome compare to early estimates based on C-values, or other indirect estimates? How does it compare to closely related species, members of the same family, and other relevant comparisons? As we publish more CCGP species, there may be comparative information in our published Genome Release papers from the same taxonomic group/clade; if so, please highlight them as appropriate.

# Funding *(PI contributes)*

If the sole funding is CCGP (which is paying for all reference genome components except sample collection), then simply insert the following sentence:

"This work was supported by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224].”

If in addition you wish to acknowledge other funding, then provide the full funding agency name(s), with complete and accurate grant numbers in brackets. Separate different funding agencies by a semi-colon.

# Acknowledgements *(PI Contributes)*

This depends on decisions about authorship (see section at the end of this template). At a minimum, we request that the following be included in recognition of the work done at the UC Davis and UC Berkeley core labs and to the UCSC group. Additional information in this section is optional, and may include individuals who are not listed coauthors, and details of collection permits or ethics protocols relevant to the sample collection.

“PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Core and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data.”

**If this species is one of our Illumina supported species (this information will be included with the notification that the genome is nearing completion), please additionally include:**

Partial support was provided by Illumina for Omni-C sequencing.

# Data Availability *(CCGP contributes)*

[*JOH Data Availability Policy*](http://academic.oup.com/jhered/pages/Policies)List all GenBank accession numbers and Dryad DOIs (where applicable), following the example below. Accession numbers for data are required upon submission and data must be made available immediately upon acceptance; manuscripts lacking this information will be returned without review. JOH will pay the cost to deposit relevant data in Dryad for all published articles. If your manuscript is accepted, you will receive a link to deposit your data in Dryad (if appropriate), and your specific Dryad DOI will be included in your published article. Where datasets already have a DOI, please cite the data in the text and place the citation in the References section.

We do not anticipate that CCGP projects will require Dryad data sets, but please contact us if you feel that this would be relevant to your reference genome.

The CCGP NCBI BioProject ID is PRJNA720569. NCBI accession numbers and links are often the final details to be generated during the submission process, so we recommend writing the manuscript with placeholders. Our reference team will be in contact with PIs during the submission and writing process to provide updates as they become available.

*Example Data Availability text:*  
Data generated for this study are available under NCBI BioProject PRJNA720569. Raw sequencing data for sample YYYY (NCBI BioSample YYYY) are deposited in the NCBI Short Read Archive (SRA) under SSSS1-SSSS2. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: [www.github.com/ccgproject/ccgp\_assembly](http://www.github.com/ccgproject/ccgp_assembly)

# References *(PI and CCGP contributes)*

These can be formatted in any readable style at submission, although the authors are responsible for their accuracy. CCGP will provide a standard list of references for lab protocols and software used for genome assemblies. Please keep this brief—generally around 10 references should do it.

# Tables and Figures *(CCGP and PI contribute, see below)*

All display items should be cited in the order that they appear. This includes any Supplemental Figures or Tables (we do not expect there to be Supplemental materials for CCGP submissions, but they are allowed). Define all symbols and abbreviations used in legends and captions. Legends and Captions should stand alone. The first sentence of each should describe the contents of the display item, and be interpretable without reference to the Main Text. Footnotes may be added. Table 1 will include information on DNA sequencing and genome assembly protocols (see example below). We provide a list of recommended Tables and Figures below.

# Supplementary Material *(PI contributes, if relevant)*

This material should not be essential to comprehension of the article, but should be relevant to the article content. Supplementary Material cannot be altered after the paper has been accepted, and will not be edited. It should be referred to in the Main Text in the following manner: ‘See Supplementary Figure 1’ or ‘See Supplementary Table 1’ or ‘See Supplementary Material’.

# Example Tables *(CCGP contributes; PI contributes, if relevant)*

Table 1: Assembly Pipeline and Software Used. Software citations are listed in the text.

|  |  |  |
| --- | --- | --- |
| Assembly | Software | Version |
| Filtering PacBio HiFi adapters | HiFiAdapterFilt  <https://github.com/sheinasim/HiFiAdapterFilt> | Commit 64d1c7b |
| Kmer counting | Meryl | 1 |
| Estimation of genome size and heterozygosity | GenomeScope | 2 |
| *De novo* assembly (contigging) | HiFiasm | 1.8 |
| Long read, genome-genome alignment | Minimap2 | 2.16 |
| Remove low-coverage, duplicated contigs | Purge\_dups | 1.0.1 |

|  |  |  |
| --- | --- | --- |
| Scaffolding | | |
| OmniC mapping for SALSA | Arima Genomics mapping pipeline  <https://github.com/ArimaGenomics/mapping_pipeline> | Commit 2e74ea4 |
| OmniC Scaffolding | SALSA | 2 |
| Gap closing | YAGCloser  <https://github.com/merlyescalona/yagcloser> | Commit  XXXX |

|  |  |  |
| --- | --- | --- |
| OmniC Contact map generation | | |
| Short-read alignment | Bwa | 0.7.17-r1188 |
| SAM/BAM processing | Samtools | 1.11 |
| SAM/BAM filtering | pairtools | 0.3.0 |
| Pairs indexing | pairix | 0.3.7 |
| Matrix generation | Cooler | 0.8.10 |
| Matrix balancing | HiCExplorer | 3.6 |
| Contact map visualization | HiGlass | 2.1.11 |

|  |  |  |
| --- | --- | --- |
| Benchmarking | | |
| Basic assembly stats | QUAST | 5.0.2 |
| Assembly completeness | BUSCO | 5.0.0 |
| Merqury | 1 |

Table 2: Sequencing and assembly statistics, and accession numbers. *[Accession numbers and links are often the final details to be generated]*

|  |  |  |
| --- | --- | --- |
| Bio Projects  & Vouchers | CCGP NCBI Bio-project | PRJNA720569  http://www.ncbi.nlm.nih.gov/bioproject/XXXXXX |
| *Genus species* NCBI Bio-project | PRJNA720569 https://www.ncbi.nlm.nih.gov/bioproject/XXXXXX |
| NCBI Bio-sample | SAMNXXXXXXXX  https://www.ncbi.nlm.nih.gov/biosample/SAMNXXXXXXXX |
| Specimen identification number | <http://arctos.database.museum/guid/MVZ:Herp:111298>  https://www.gbif.org/occurrence/1145409359 |
| GenomeArk data link | https://vgp.github.io/genomeark/Genera\_species/ |
| Genome Sequence | PacBio HiFi long read runs (male/female) | 1 PACBIO\_SMRT (Sequel II) run: 1.7M spots, 20.5G bases, 11Gb downloads |
| OmniC Illumina sequencing | 1 Illumina NovaSeq 6000 run: 44.7M spots, 13.3G bases, 3.9Gb download |
| PacBio HiFi NCBI SRA Accession | SRXXXXXXXX  https://www.ncbi.nlm.nih.gov/sra/SRXXXXXXXX |
| OmniC Illumina NCBI SRA Accession | SRXXXXXXXX  https://www.ncbi.nlm.nih.gov/sra/SRXXXXXXXX |
| Genome Assembly  Primary (Alternate) | Assembly identifier | aEmyMar1 |
| HiFi Read coverage | 30X |
| Number of contigs | 184 (3,363) |
| Contig N50 (bp) | 115,383,402 (2,659,599) |
| Longest Contigs | XXXX |
| Number of scaffolds | 33 (108) |
| Scaffold N50 (bp) | 50,802,423 bp |
| Size of final assembly (bp) | 2,355,098,844 (2,259,568,388) |
| Gaps per Gbp | X (XX) |
| NCBI Genome Assembly Accession | GCA\_000XXXXXX.1  https://www.ncbi.nlm.nih.gov/assembly/GCA\_000XXXXXX.1 |
|  | Assembly Quality identifier | 7.C.Q44 |
|  | Base pair QV (Merqury) | P: Q 66.3, A: Q 66.35 |
| Assembly  Quality | Indel QV (Frame shift analysis) | P: Q 44.27, A: Q45.04 |
|  | k-mer completeness | P: 94.56% A: 90.79% |
|  | BUSCO completeness  Primary (C:S:D:F:M)  Alternate (C:S:D:F:M) | 98.40% : 97.00% : 1.40% : 0.90% : 0.70%  94.50% : 92.70% : 1.80% : 1.00% : 4.50% |
|  | Phased block NG50 | XX,XXX,XXX bp |
|  | Assigned Chromosomes % | XX% |
|  | Sex Chromosomes (if appropriate) | (Fragmented / 1 piece) |
|  | Organelles | (1 complete allele / fragmented for mitochondria/plastids) |
|  |  |  |

# Example Figures *(CCGP contributes; PI contributes)*

Figure 1. (*PI contributes)* An image of the study species, including copyright permissions as needed. This should be high quality, with a brief description including Latin binomial and common name in the legend. Multi-panel figures showing different life stages, or habitat shots, can also work well.

Figure 2. (*CCGP contributes)*. OmniC contact map to give an idea of the level of chromosome assembly. An example is shown below from the CCGP big berry manzanita (*Arctostaphylos glauca*) release paper. 2A: Kmer spectra (DRAFT). 2B: BlobToolKit Snailplot showing N50 metrics for *A. glauca* assembly ddArcGlau1 and BUSCO scores for the embryophyta set of orthologues. *[This will be regenerated once NCBI has given the final approval of a genome]* 2C: Contact Map of Primary/Alternate assembly *[This will be regenerated once NCBI has given the final approval of a genome]*

**Figure 2**

**Authorship of JOH Genome Resources Papers for CCGP**

Ultimately, authorship, including author order, is up to the lead author and their team. Our general feeling is that, for genome resource papers, the technical support provided by individuals who carried out extractions, library preps, and genome assembly is a much greater contribution to the final paper than normal, and that the individuals who worked on your individual genome may deserve authorship. We recognize that this is somewhat unusual, but it is also becoming standard for genome release papers. After all, the genome being released is the centerpiece of the paper, and these individuals made it happen.

The benchwork for your reference genome took place at two UC campus facilities (high molecular weight DNA extraction and HiFi library preparation at UC Davis, Omni-C extraction and library preparation at UC Santa Cruz). Methods may be contributed by these teams. All data assembly, bioinformatics, data and assembly QC, relevant writing, and final uploading was completed by Merly Escalona, (UC Santa Cruz). Troubleshooting and project coordination, sometimes including assisting with the final manuscript may include Erin Toffelmier or Brad Shaffer.

Which of these individuals should be considered as potential co-authors depends on your approach to authorship, the difficulty of the assembly (we can advise on this issue), and the degree to which any individual provided help with the paper itself, either in writing or revisions.

We have only two requests. First, try your best to be fair. Give credit where it is due, but don’t overdo it. Second, if you are considering including one or more people as coauthors, make sure to contact them first, and emphasize that they must read the manuscript, provide comments, and sign off on it. The names of the people who assisted with data generation will be provided to each project when the genome nears completion.

[Click here to access the list of individuals (and additional information such as their email, institution, role, etc.) that may have contributed to generating your reference genome.](https://docs.google.com/spreadsheets/d/1vrngY3aW-_tEt2j54Sn7xPMZUrHpV1Zq/edit?usp=sharing&ouid=102639935461879242978&rtpof=true&sd=true)

Please contact Brad Shaffer (brad.shaffer@ucla.edu) or Erin Toffelmier (etoff@ucla.edu) if we can provide any guidance or help on this.