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

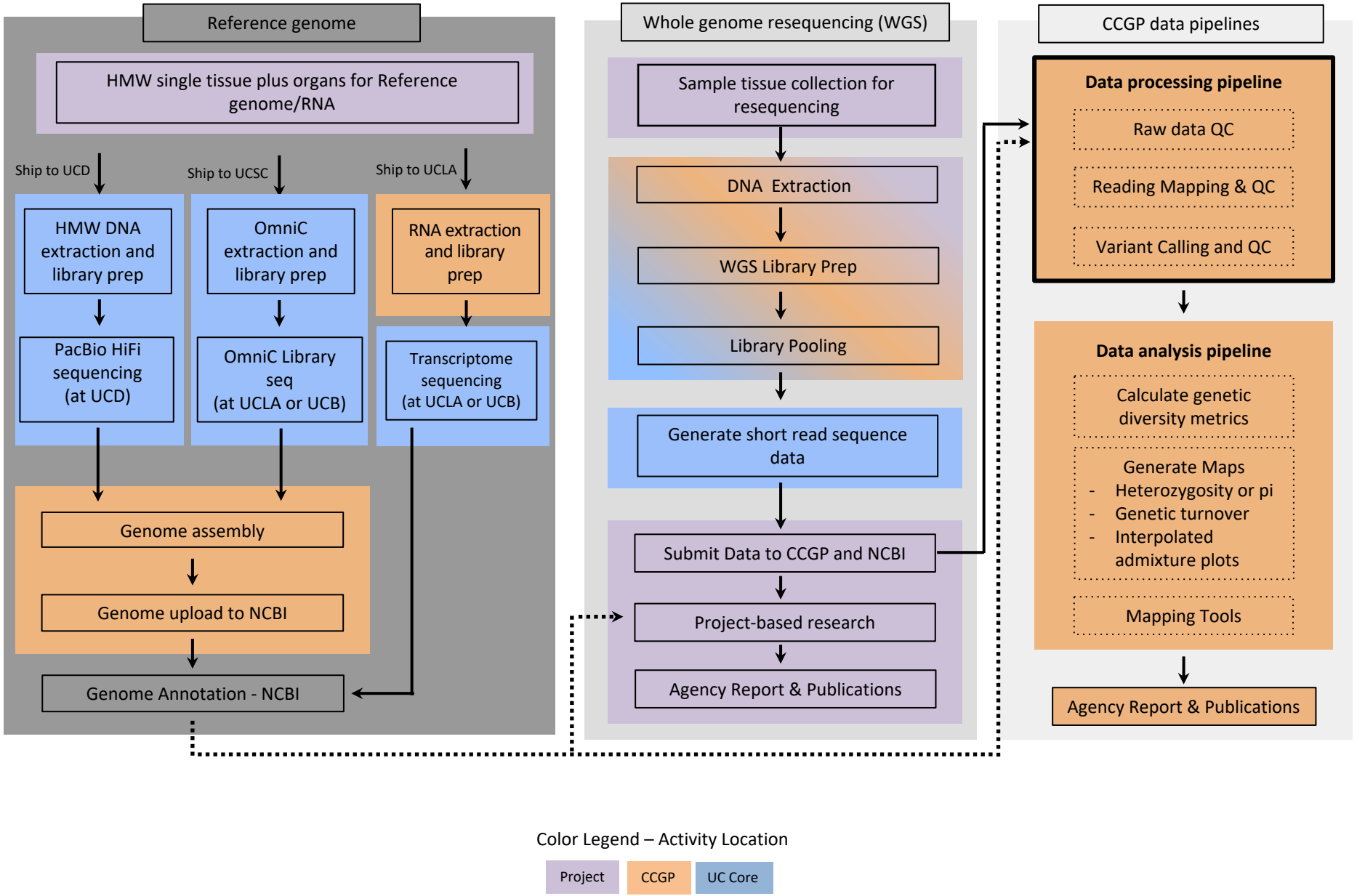
- Data for management/policy:
  - State and federal agencies
  - Private and public land managers

Today's goals:

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

Questions and Discussion

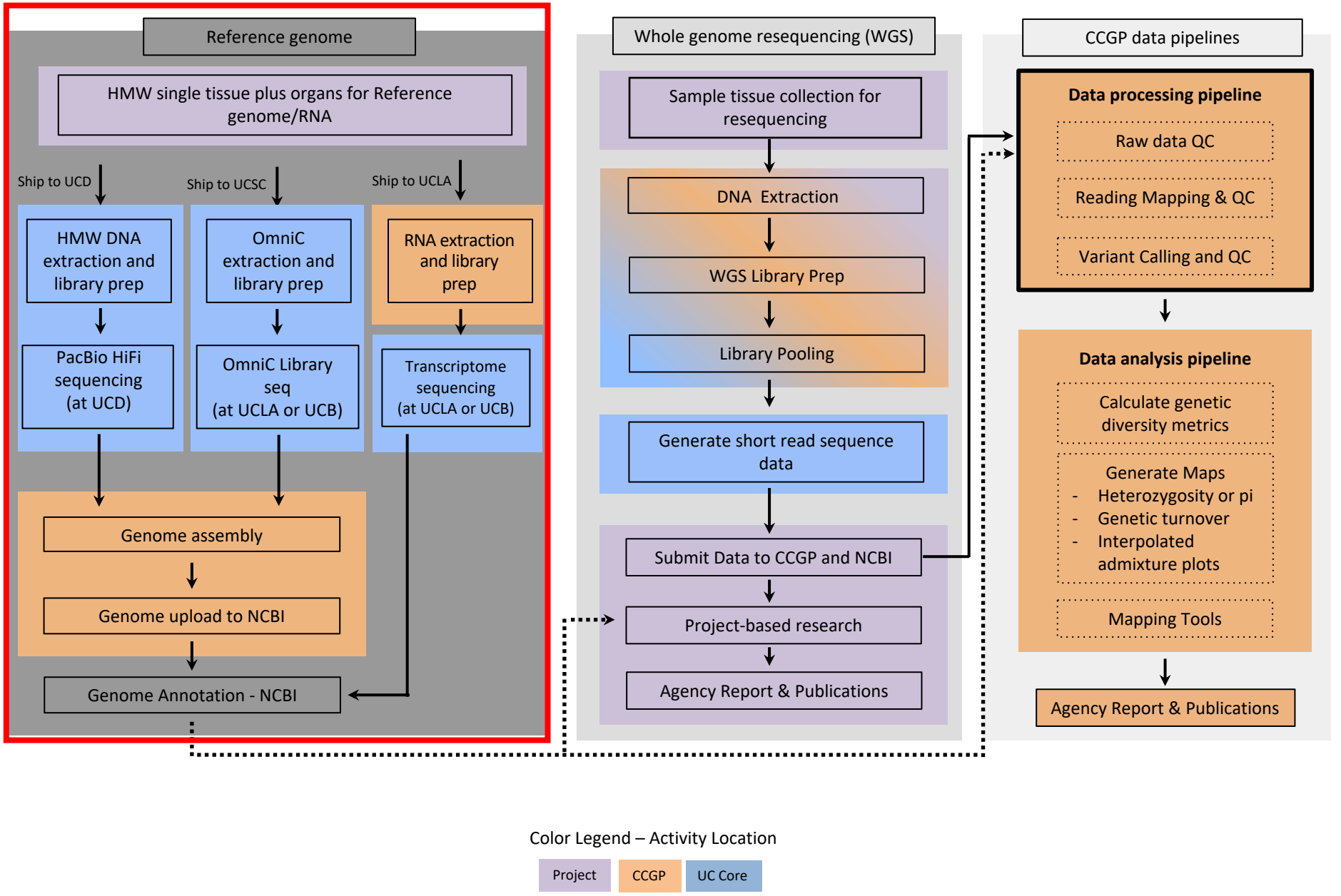
CCGP workflow



Color Legend – Activity Location

Project | CCGP | UC Core

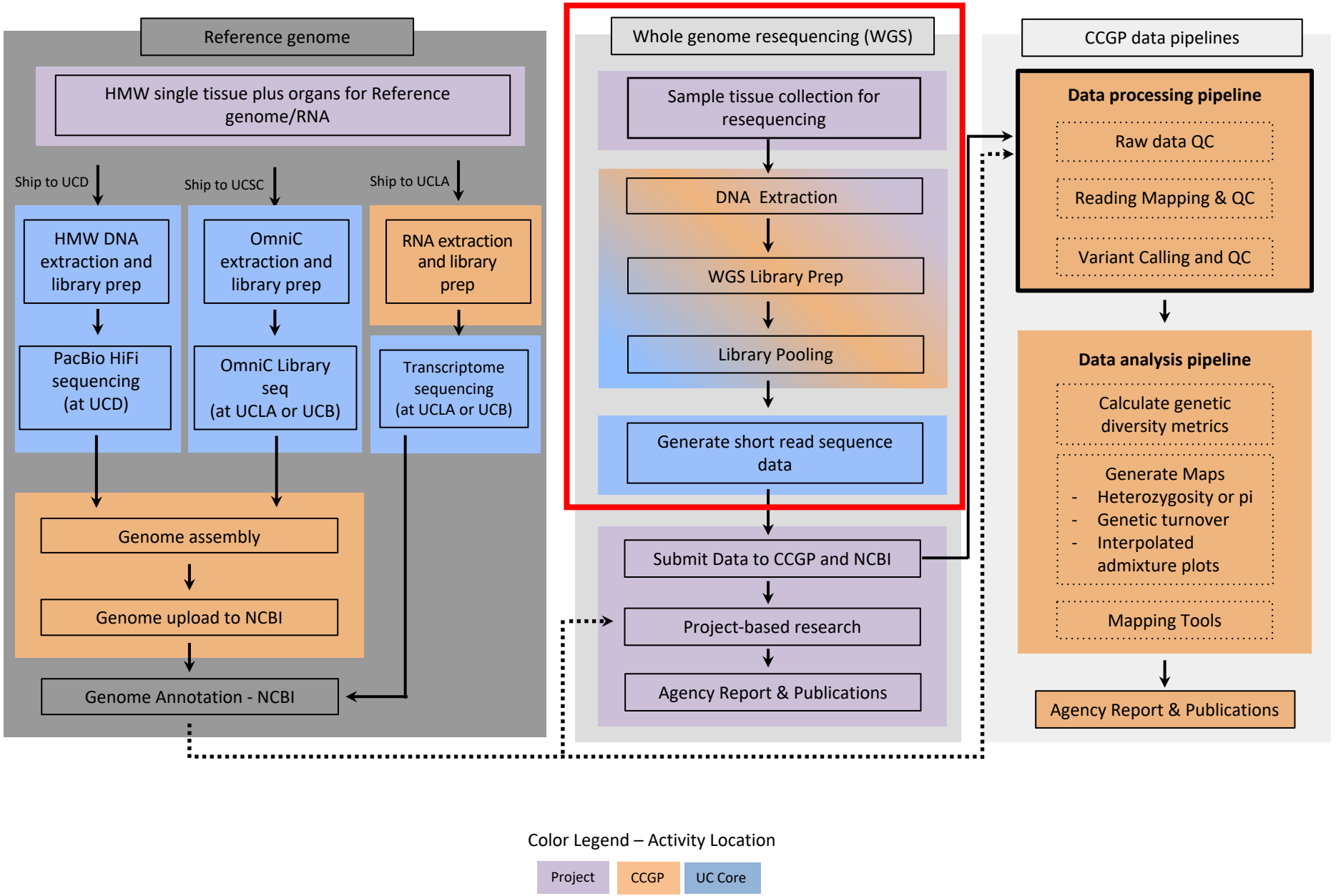
CCGP workflow



Color Legend – Activity Location

Project | CCGP | UC Core

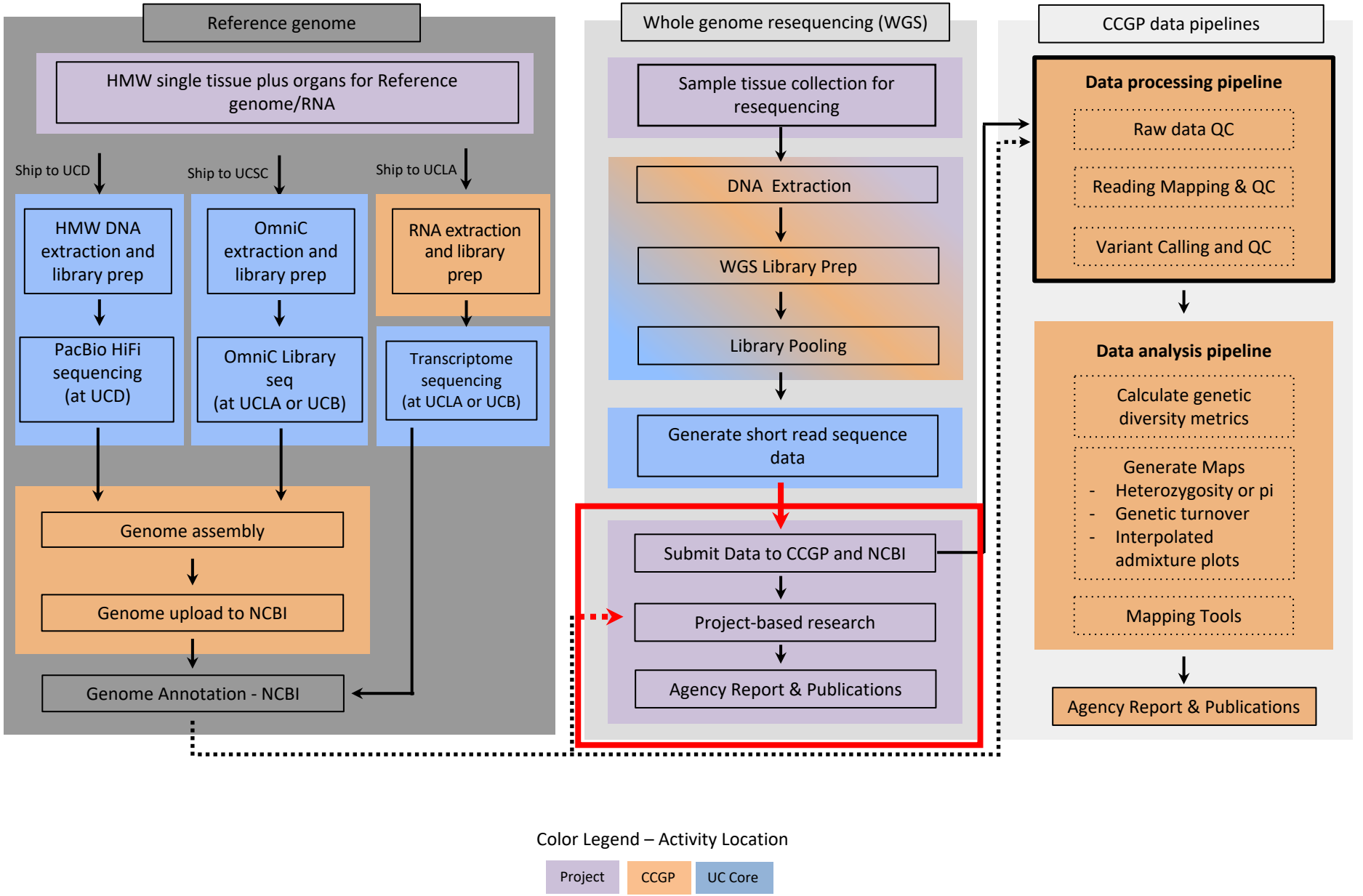
CCGP workflow



Color Legend – Activity Location

Project | CCGP | UC Core

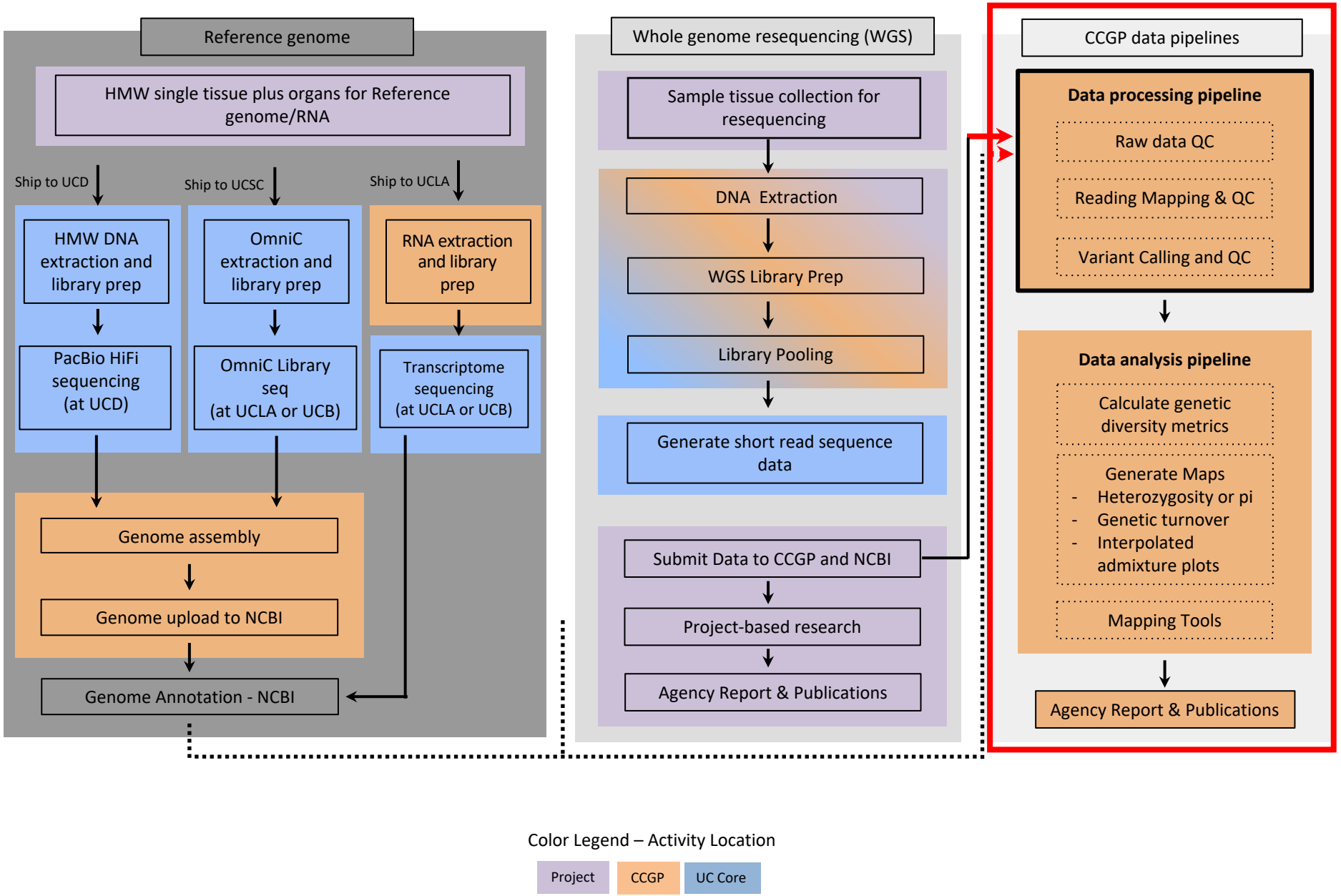
CCGP workflow



Color Legend – Activity Location

Project | CCGP | UC Core

CCGP workflow



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

Mini-Core task	*per sample*
DNA extraction from tissue + library preparation	$30
Library preparation from user supplied gDNA	$20
Additional QC, if needed	$4
Additional extraction, if needed	$6
EDTA cleanup (SPRI based), if needed	$3

Other Important Consideration

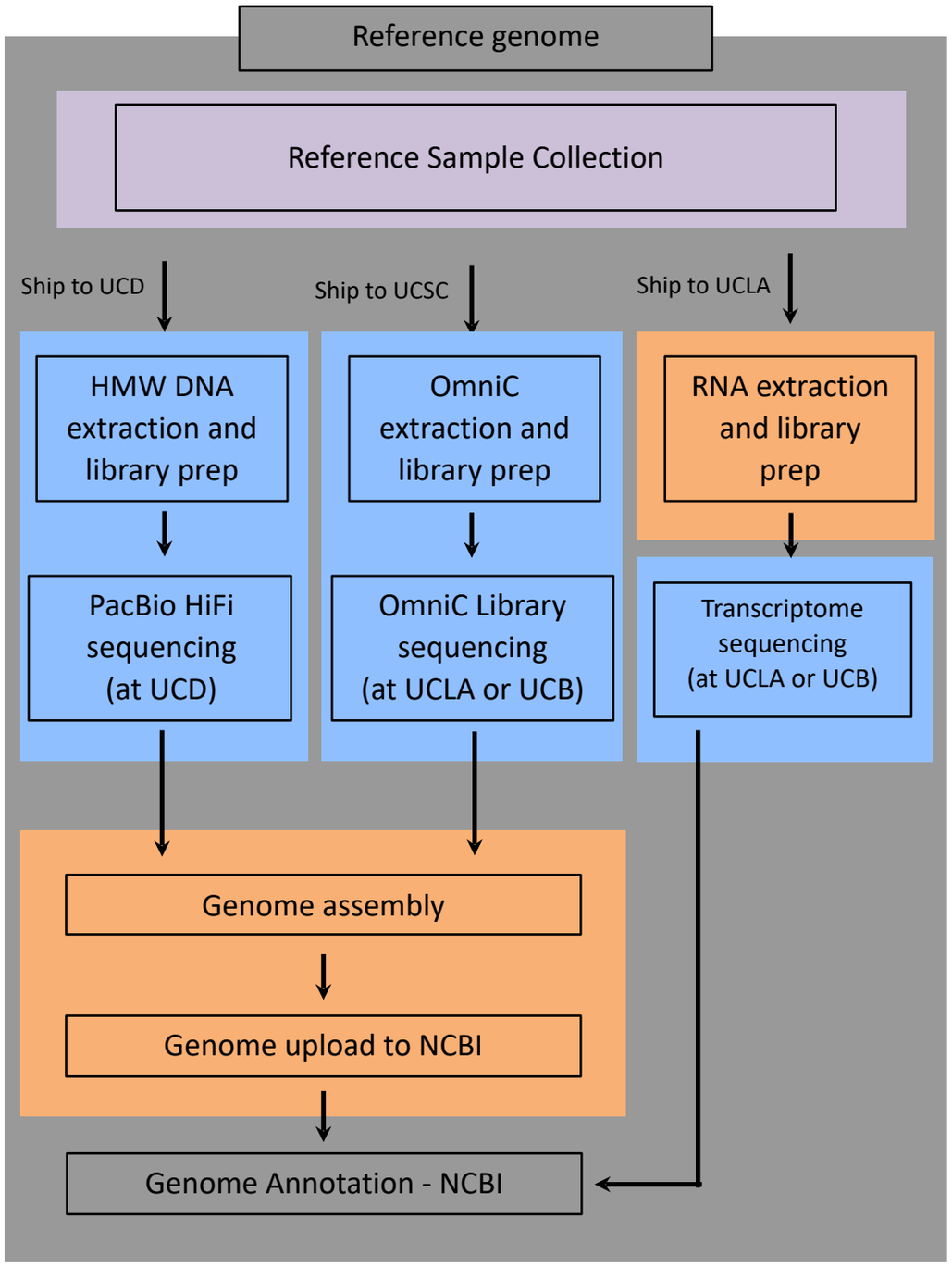
- Plants, slimy inverts are the hardest
- Only fresh material!
  - No herbarium sheet material
  - No formalin preserved
  - No “hard” tissue (feathers, bone, nail)
- Plants, at least 1 g fresh leaf
- Small inverts, at least 0.2 g fresh material
- Verts, at least 0.2 g fresh material
- Nucleated RBCs,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

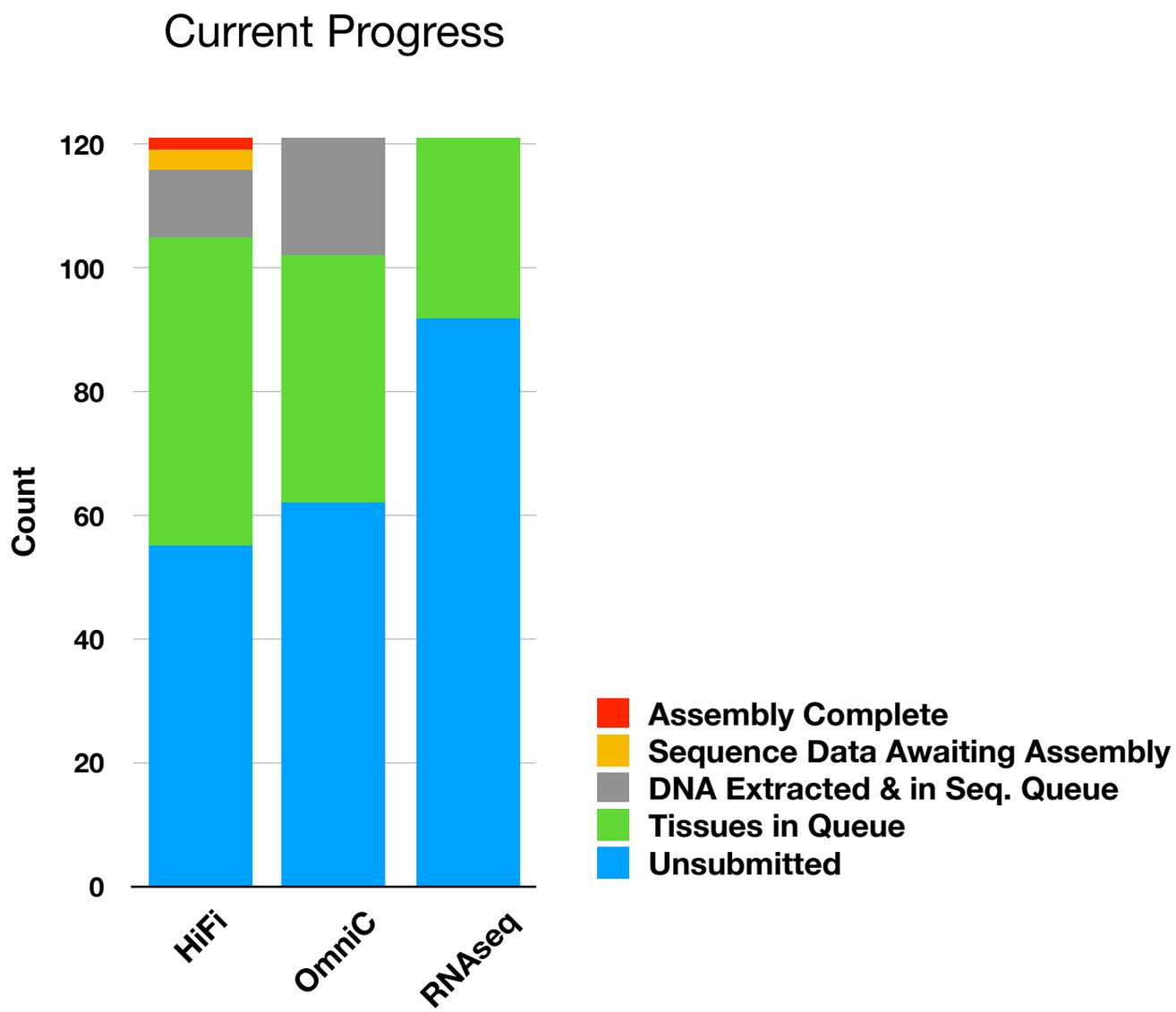
More Important Consideration

- 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



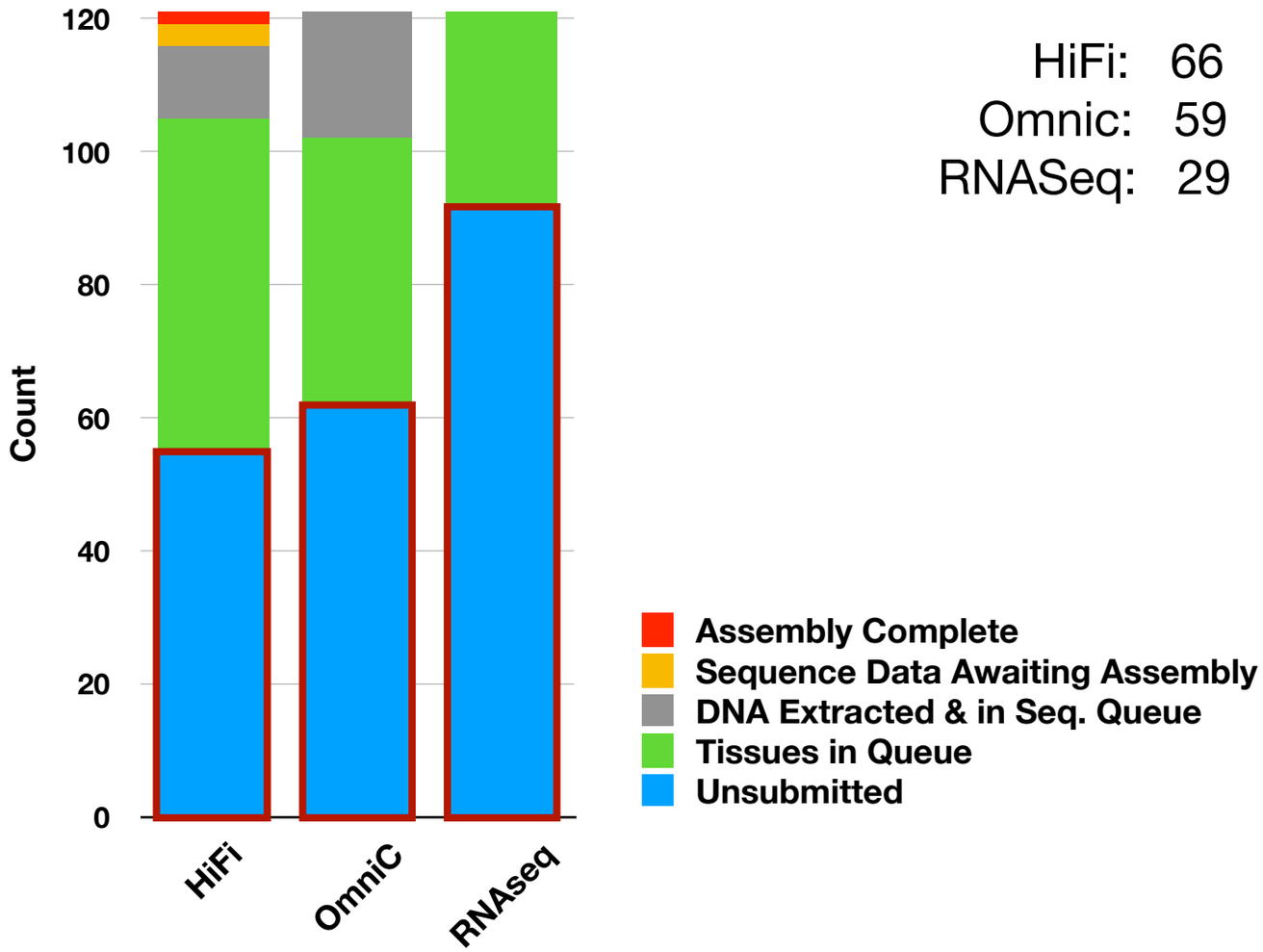


Color Legend – Activity Location

Project | CCGP | UC Core

CCGP Reference Genome Update

Current Progress

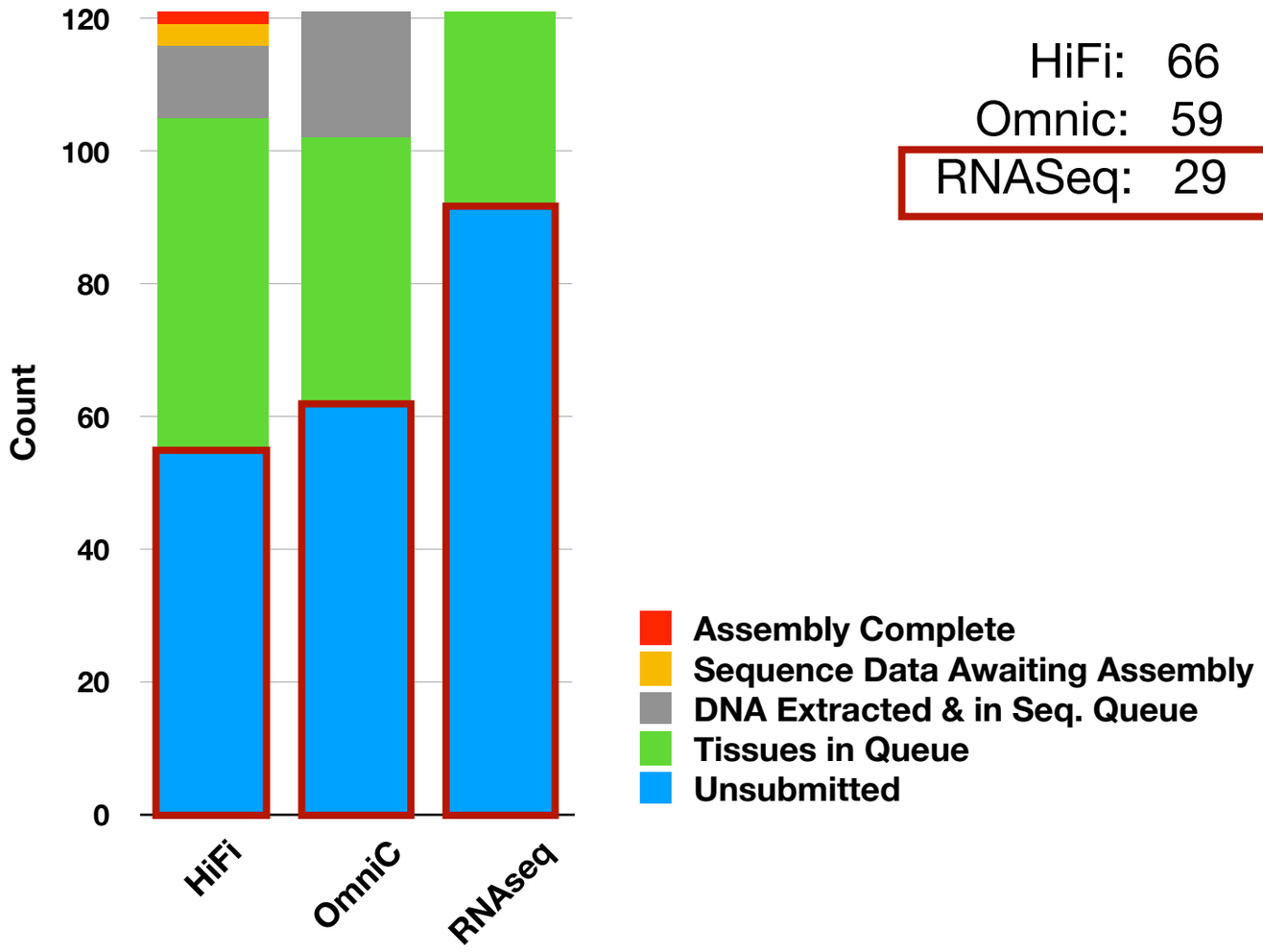


Tissue Submissions
(121 expected)

HiFi: 66
Omnic: 59
RNASeq: 29

CCGP Reference Genome Update

Current Progress



Tissue Submissions
(121 expected)

HiFi: 66
Omnic: 59
RNASeq: 29

RNASeq Tissues:

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

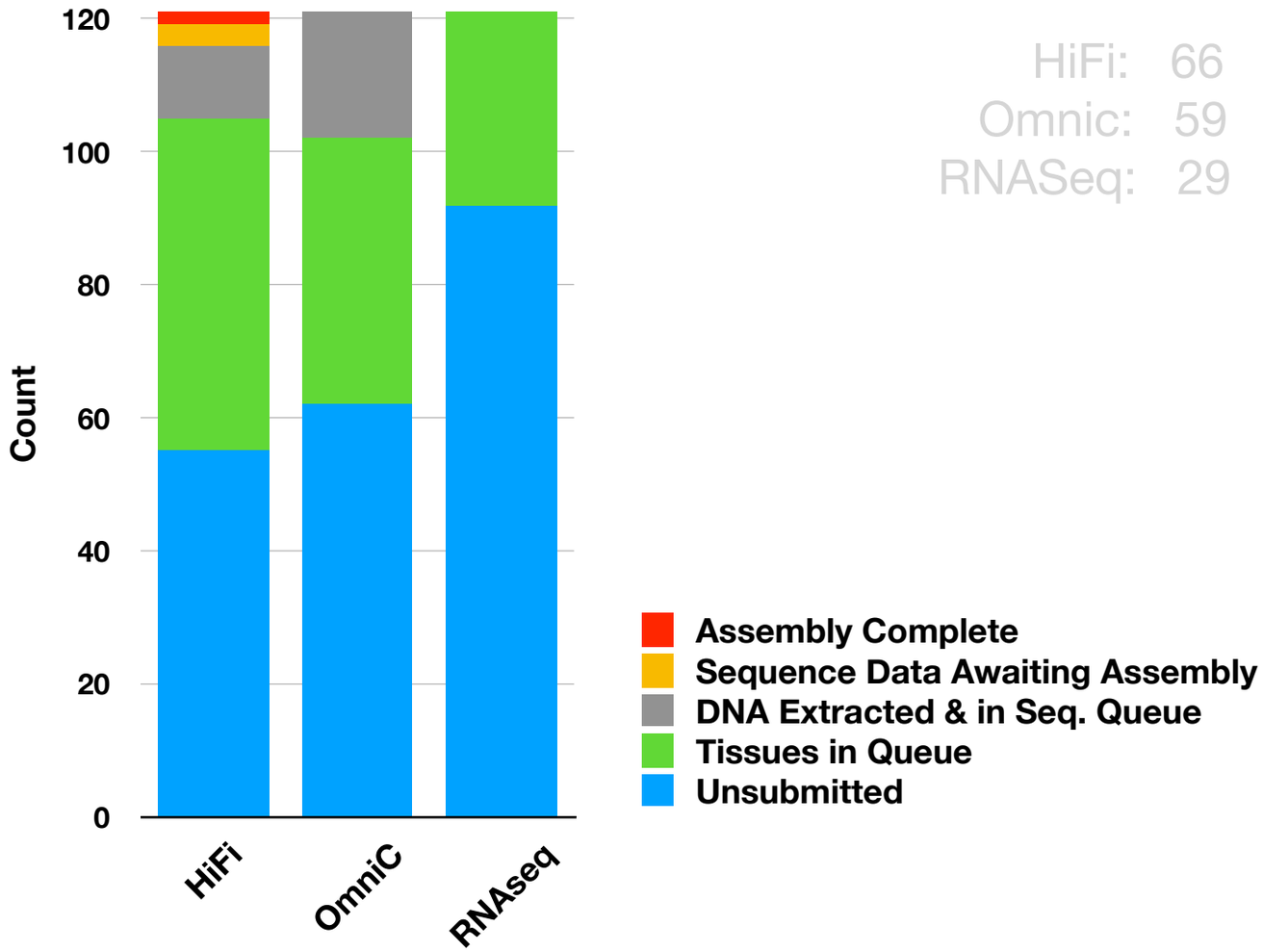
For Reference Tissue Submission Info:

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

Contact Erin: etoff@ucla.edu

CCGP Reference Genome Update

Current Progress



Tissue Submissions (121 expected)

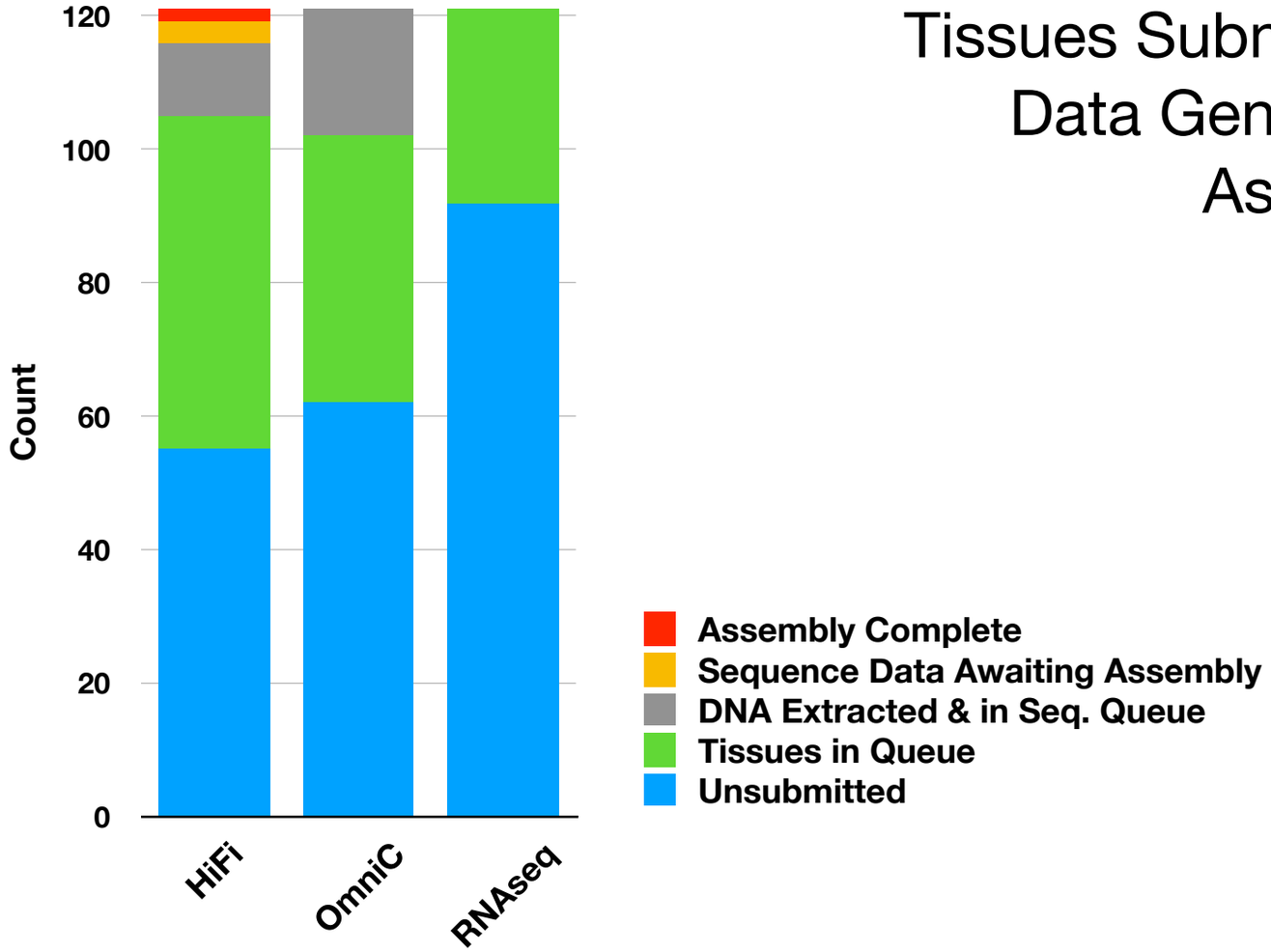
HiFi: 66
Omnic: 59
RNASeq: 29

In process Library Prep & Sequencing

HiFi: 16 (2 completed assemblies)
Omnic: 19 (first deep seq run in January)
RNASeq: 0 (extractions start with opening of mini core)

CCGP Reference Genome Update

Current Progress



Timeline

Tissues Submission: July 2021

Data Generation: November 2021

Assembly: December 2021

CCGP Reference Genome Update

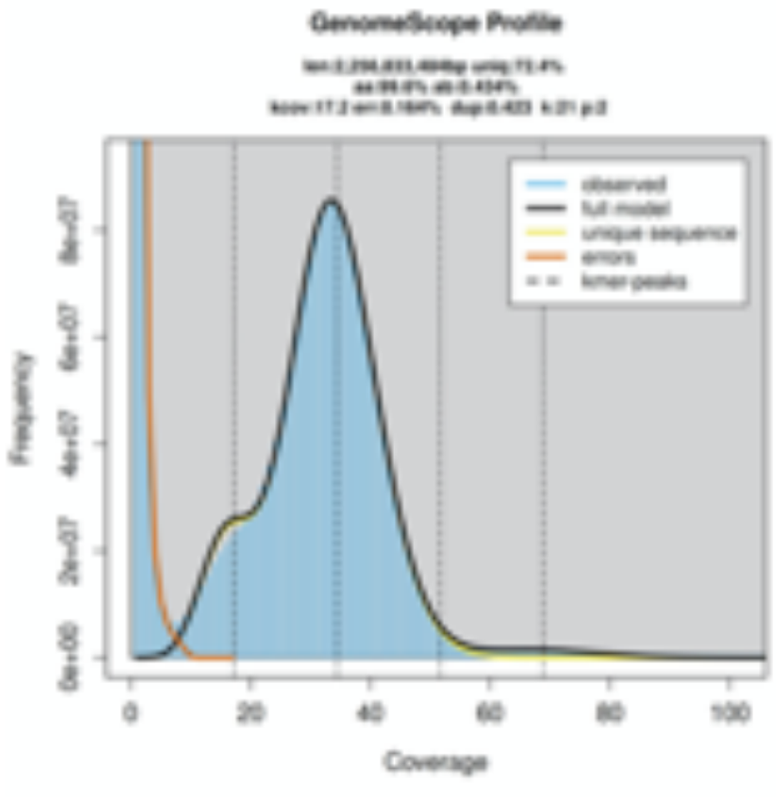
PacBio HiFi Long Read Assembly

Merly Escalona
UCSC

Western pond turtle

Emys marmorata

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



General statistics	Primary	Alternate
Length of genome	2,355,098,844	2,259,568,388
# Sequences	184	3,363
Contig N50	115,383,402	2,659,599
Contig L50	6	233
Longest contig	306,034,025	17,432,642
# Gaps	0	0

BUSCO Scores (n=954, metazoa)*					
	C	S	D	F	M
P	98.40%	97.00%	1.40%	0.90%	0.70%
A	94.50%	92.70%	1.80%	1.00%	4.50%

- (C)omplete and (S)ingle
- (C)omplete and (D)uplicated
- (F)ragmented
- (M)issing

Merqury (kmer) analysis	P	A
base-call QV (hap)	66.03	66.35
k-mer completeness (hap)	94.56	90.79
base-call QV (full)	66.18	
k-mer completeness (full)	99.27	