

### 3. Assembly and annotation

- Genome assembly
- Metrics and assembly QA/QC
- Re-sequencing and variant calling
- Annotation

# *De novo* genome assembly

- Reconstructing long, continuous sequence (up to chromosomes) from millions of fragments (reads)



- Computationally difficult due to repetitive (identical or highly similar) DNA
- Levels of assembly: Reads > contigs > scaffolds (> chromosomes)

# Short read assembly

- Paired ends

(PE)

Forward read

Reverse read



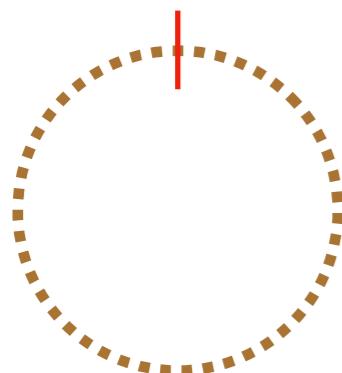
200–500 bp unknown sequence (insert)

Example file names

Hpue\_raw300\_F.fastq

Hpue\_raw300\_R.fastq

- Mate pairs



Forward read

Reverse read

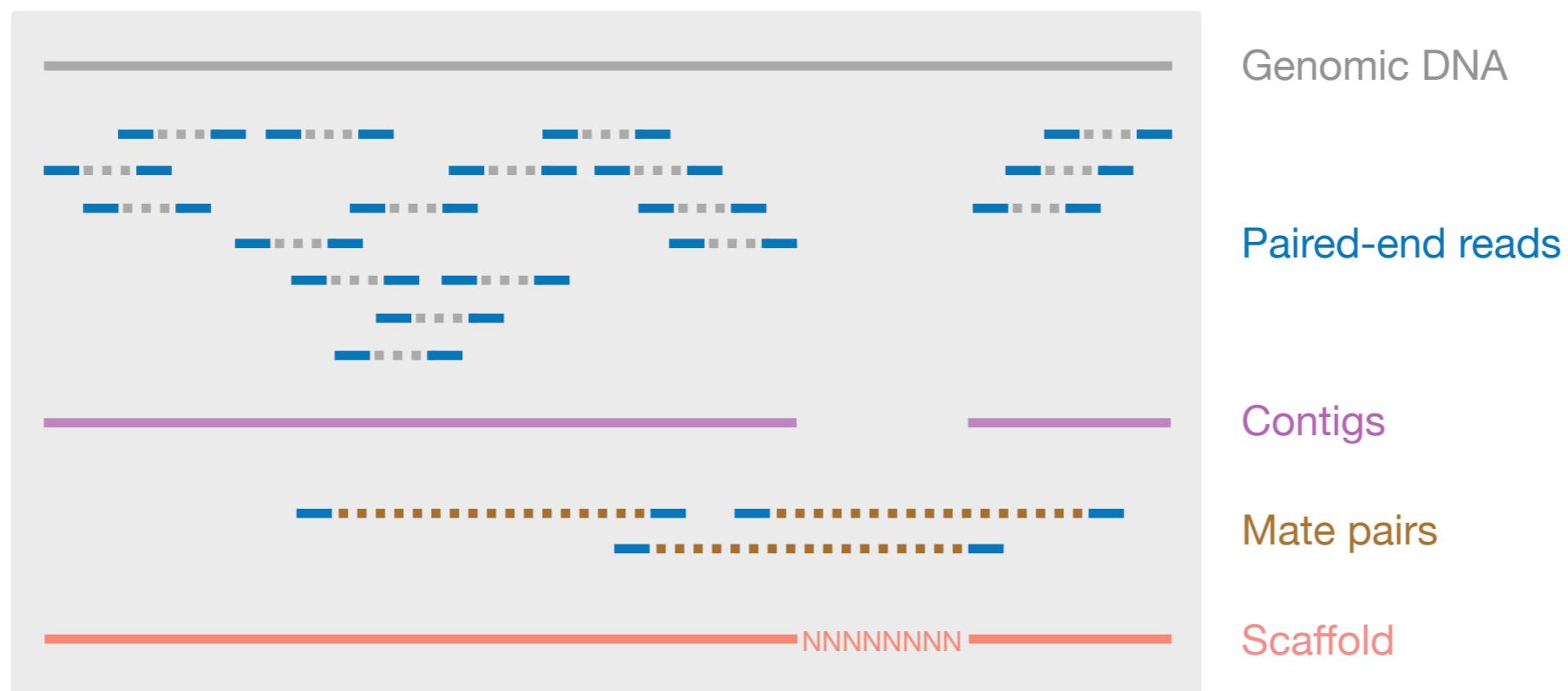
2–20 kb unknown sequence (long insert)

Example sequencing strategy

Library type	Insert sizes
5 × pair	250–580 bp
2 × mate	2.5, 4.3 kb

# Short read assembly

- Overlapping reads > contigs
- Ordered, oriented contigs with gaps of known position, length > scaffolds  
(gaps spanned by mate pairs)

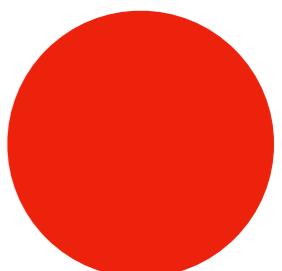


# Hybrid strategy

- Combine Illumina with long read sequencing
- Correct long reads with high coverage Illumina data
- Fill gaps in Illumina-based scaffolds with long reads (> up to chromosomes)

# Polishing and mapping

- Polishing?
- High-C, optical mapping?
- Merge / purge haplotigs



# Assembly example

## The FASTA format

```
cd workshop_zmt/project/3_assembly          # change directory  
zcat < Hpue_assembly_01_LG12.fas.gz | head -n 100    # look at compressed file (zcat),  
                                                       # display first 100 lines
```

```
>UXGA01000012.1 Hypoplectrus puella genome assembly, contig: LG12, whole genome shotgun  
sequence  
CCAAAAGTAGTTACAGTCCATGCGGCAAGTCCAGaactctctgcaccagactttTTTGAAGAAATGAGAGAGTTATGCG  
CCGGTCTAAACAAACTGAGGCTGCACTTTGAGCAGATTAAACAGCTCAGAAGGGACATGAGACCAAGATCAAGGCAAG  
ACTGGACCTCTGAAGAACCTTTGAACAACCTGTTCTCATCTCCTCAGCTGATAGCCAGTGGACCTGTCTTCATCT  
...  
TGGGGATTAAGAGACGAAACACGTTAAACTCACATGAAACACTAAACTGATAACATAGNNNNNNNNNNNNNNNNNNNN  
NNNTCGGTAGAAAAGTCTGTGCCTGTTGTCAGATGATTGAACTCATGAGCTGTCTACTGTCTGGTGTAT  
GTTAATCAAGTGTACTTCCGATAATGGAAACCTCTGATAACAGCTTTGTTGCCGCTGCTGATAAGCCCTGCAAAGC
```

# Assemblers

- Illumina: ALLPATHS, ABYSS, Discovar, Platanus, SOAPdenovo
- PacBio: Falcon, Canu (also Nanopore)
- 10x Genomics: Supernova
- Small genomes: SPAdes, Velvet, HGAP

# General advice

- Compare **alternate assemblies** (consider use case)
- Even finished assemblies should be **considered drafts**
- Expect lower quality in difficult regions (high repeat content, heterozygosity)

# Assessing assembly quality

- **Assembly metrics:** size distribution of contigs / scaffolds
- Gaps per Gb, unplaced contigs
- Base accuracy (Q score)
- Presence of false duplications
- **Gene completeness**
- Percentage of assembly assigned to chromosomes

# Assembly metrics

- Total size (compare to expected genome size)
- Number of contigs / scaffolds
- Largest scaffold
- N50: contig / scaffold size where 50% of assembly is found on contigs / scaffolds of equal or larger size (measure for **sequence continuity**)

Scaffolds: 530, 760, 1050, 610, 450, 800, 220, and 1200 kb

Reorder: 1200, 1050, 800, 760, 610, 530, 450, 220 kb

Sum / 2:  $5620 / 2 = 2810$

Add up until reached:  $1200 + 1050 + 800 > 2810$

N50 = 800 kb

Example calculation

# Assembly metrics

Using QUAST (command line tool or <http://cab.cc.spbu.ru/quast/>)

## Report Example

### E. coli single-cell assemblies

Aligned to "e.coli\_reference" | 4 639 675 bp | 50.79% G+C | 884 operons

All statistics are based on contigs of size  $\geq$  500 bp, unless otherwise noted (e.g., "# contigs ( $\geq$  0 bp)" and "Total length ( $\geq$  0 bp)" include all contigs).

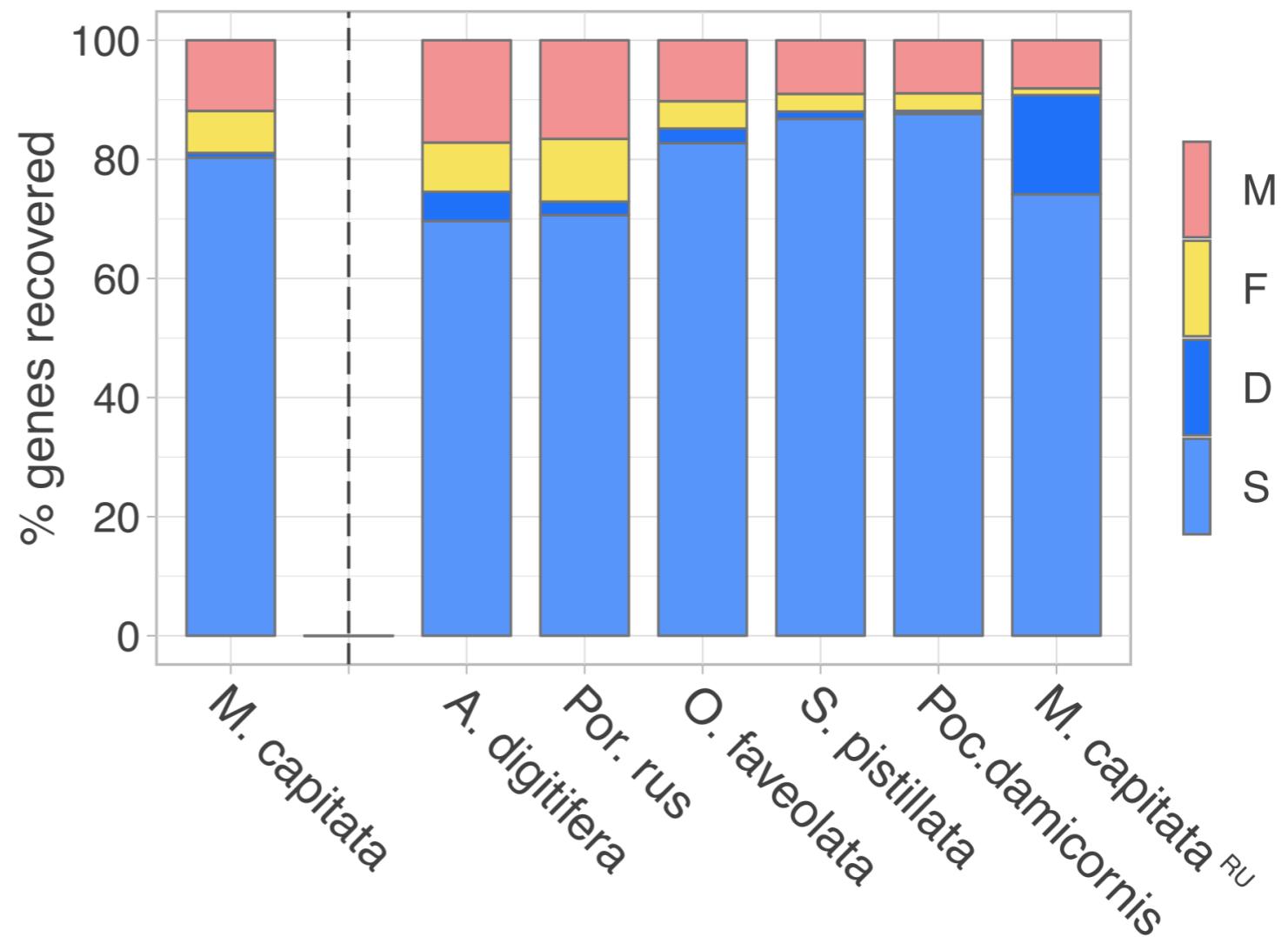
	Worst	Median	Best	Show heatmap
<b>Statistics without reference</b>				
# contigs	344	250	277	519
Largest contig	132 865	224 018	269 177	121 367
Total length	4 540 286	4 791 744	4 877 521	4 526 656
N50	33 616	96 947	106 927	20 445
<b>Misassemblies</b>				
# misassemblies	2	9	2	2
Misassembled contigs length	23 485	66 335	26 551	22 359
<b>Mismatches</b>				
# mismatches per 100 kbp	2.26	3.65	5.06	1.77
# indels per 100 kbp	0.7	0.2	0.7	0.92
# N's per 100 kbp	0	0	4.86	0
<b>Genome statistics</b>				
Genome fraction (%)	91.727	94.943	95.759	91.43
Duplication ratio	1.001	1.001	1.004	1.002
# genes	3767 + 160 part	4026 + 80 part	4046 + 102 part	3630 + 288 part
# operons	723 + 87 part	802 + 40 part	809 + 48 part	650 + 158 part
NGA50	32 051	96 947	110 539	19 791
<b>Predicted genes</b>				
# predicted genes (unique)	4258	4394	4417	4331
# predicted genes ( $\geq$ 0 bp)	4258	4394	4490	4331
# predicted genes ( $\geq$ 300 bp)	3643	3736	3784	3666
# predicted genes ( $\geq$ 1500 bp)	524	559	559	515
# predicted genes ( $\geq$ 3000 bp)	44	49	48	39

[Extended report](#)

# Gene completeness

Using BUSCO

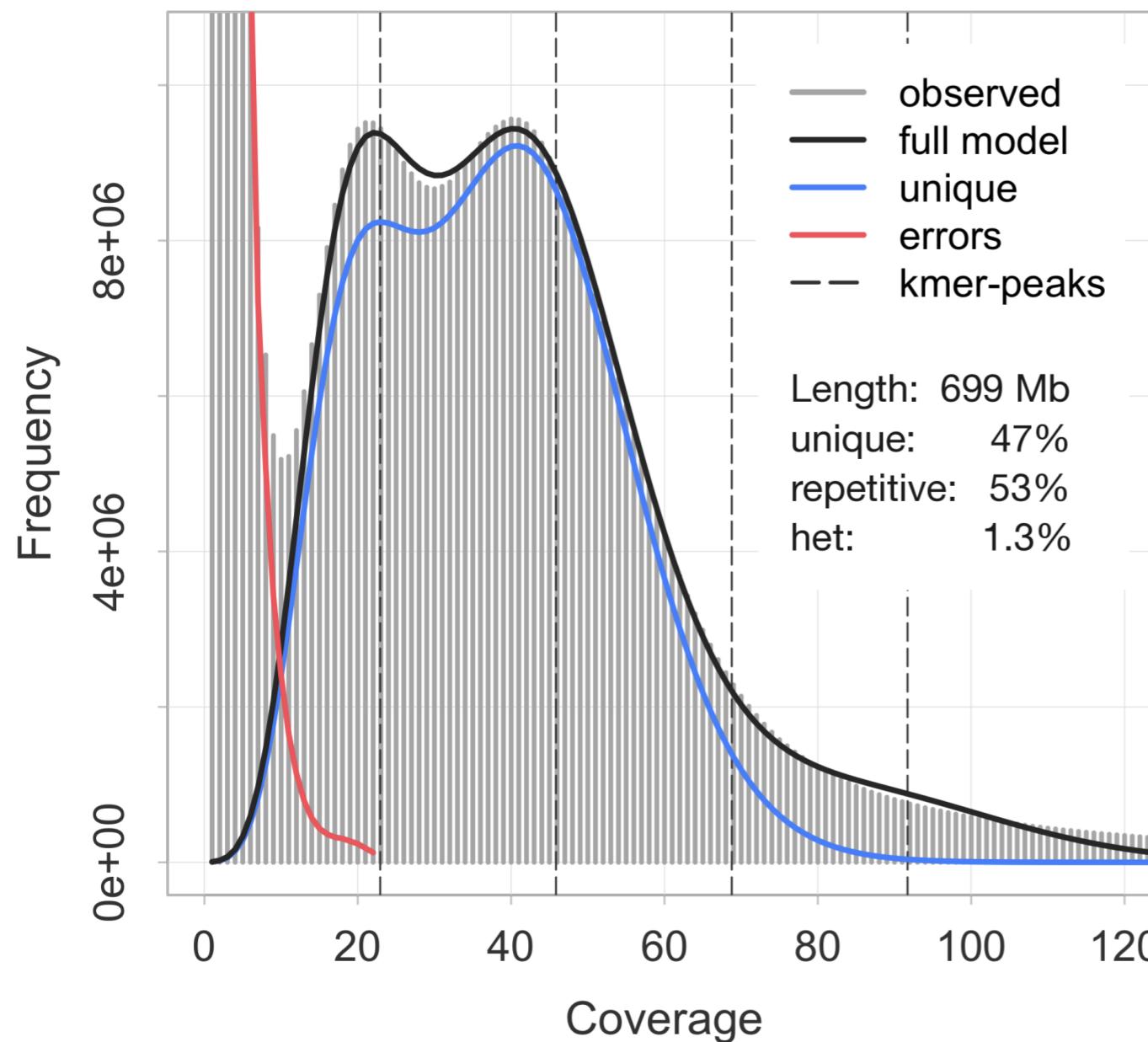
Quantifies assembly completeness based on presence of universal, highly conserved, single-copy genes (e.g. housekeeping genes)



Helmkampf et al. 2019 (Genome Biology and Evolution)

# Genome assessment using $k$ -mer profiles

## Using GenomeScope



Example:

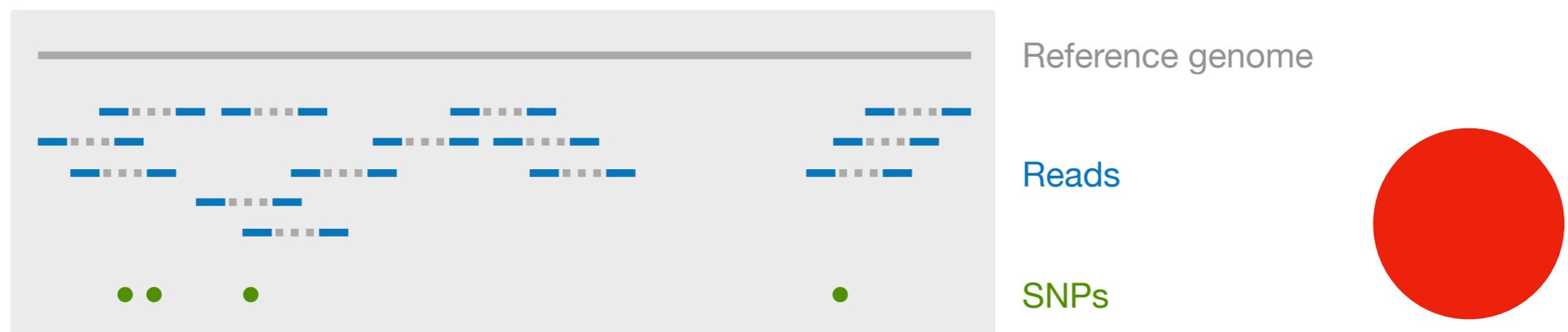
6-mers of 10 bp-sequence

AGTCAAGTCT  
AGTCAA  
GTCAAG  
TCAAGT  
CAAGTC  
AAGTCT

Helmkampf et al. 2019 (Genome Biology and Evolution)

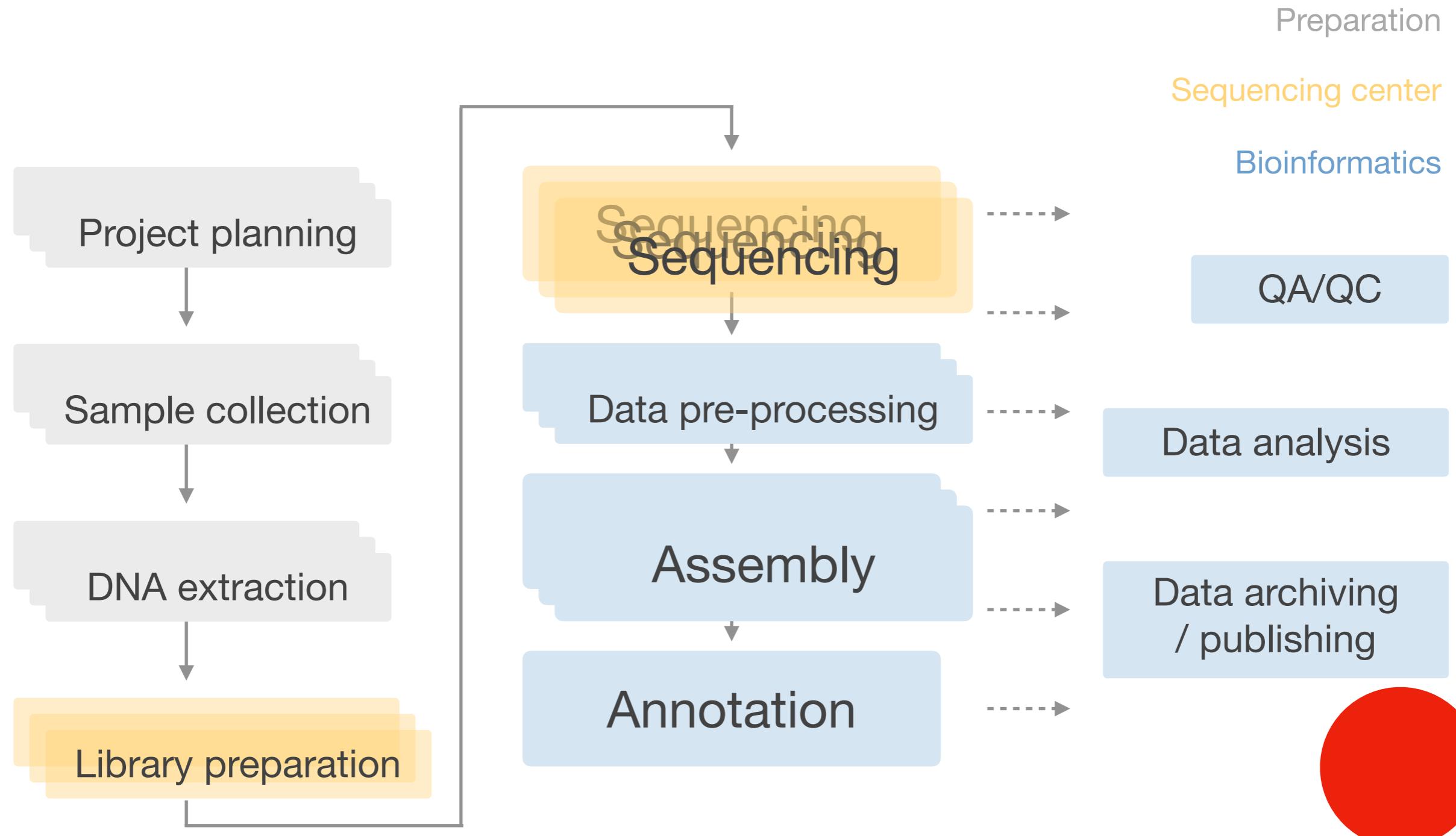
# Re-sequencing

- Prerequisite: high-quality reference genome
- Approach
  - low-coverage sequencing of multiple samples
  - aligning reads to reference (mapping)
  - identify differences > SNPs (variant calling / genotyping)



# Genome re-sequencing workflow

Legend

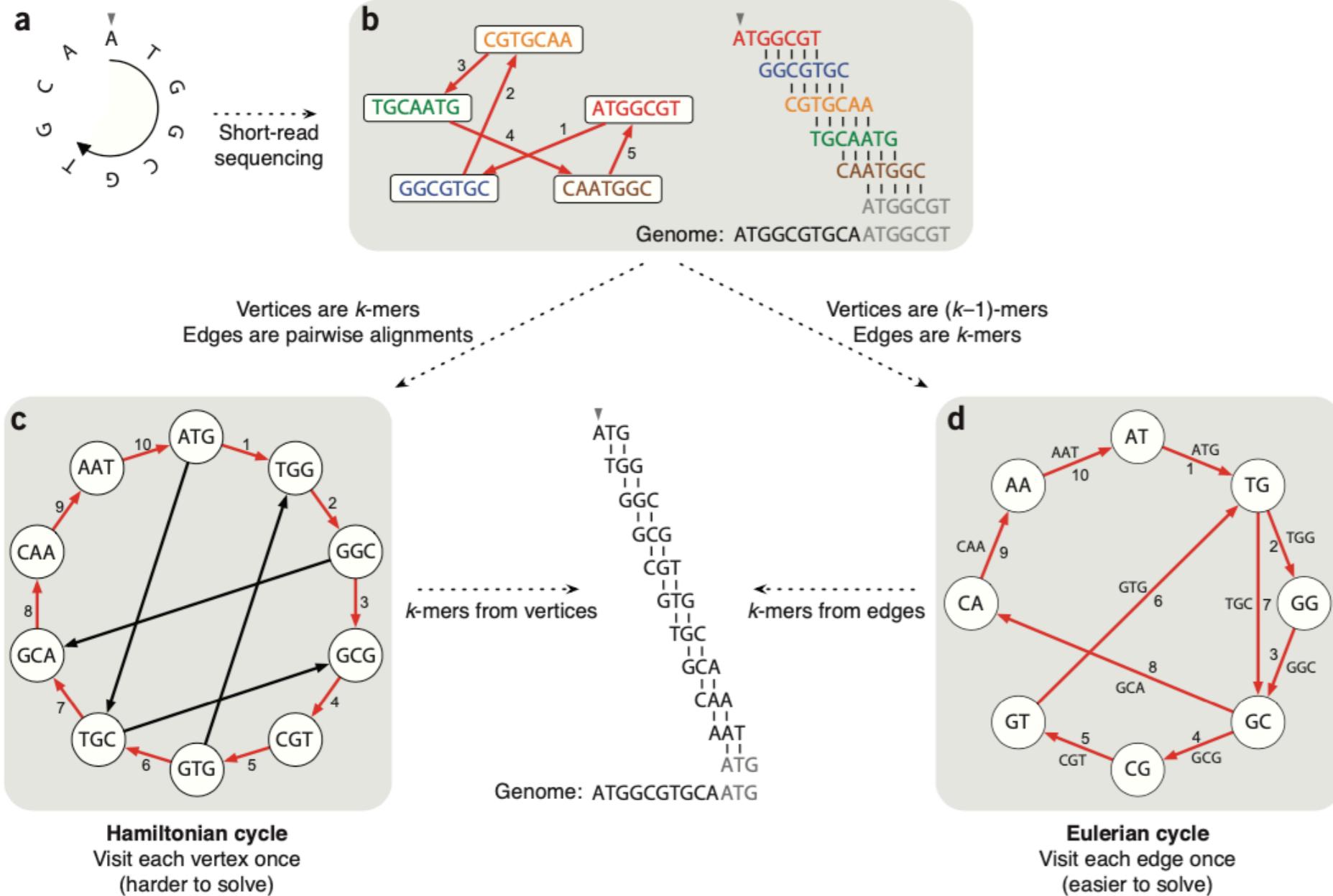


# Re-sequencing

## The VCF format

```
##fileformat=VCFv4.1
...
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sampleA sampleB sampleC sampleD sampleE
LG01 1258 . C T . PASS . GT 0|0 0|0 0|0 0|0 0|0
LG01 1277 . G T . PASS . GT 0|0 0|0 0|0 0|0 0|0
LG01 1292 . G A . PASS . GT 0|0 0|0 0|0 0|0 0|0
LG01 1365 . C A . PASS . GT 0|0 0|0 0|0 0|0 0|0
LG01 1373 . C T . PASS . GT 0|1 0|1 1|1 0|0 1|1
LG01 1398 . G C . PASS . GT 0|1 0|1 1|1 0|0 1|1
LG01 1403 . G C . PASS . GT 0|0 0|0 0|0 0|0 0|0
LG01 1494 . C T . PASS . GT 1|0 0|0 0|0 1|1 0|1
LG01 1495 . G A . PASS . GT 0|1 1|0 1|1 0|0 1|0
...
```

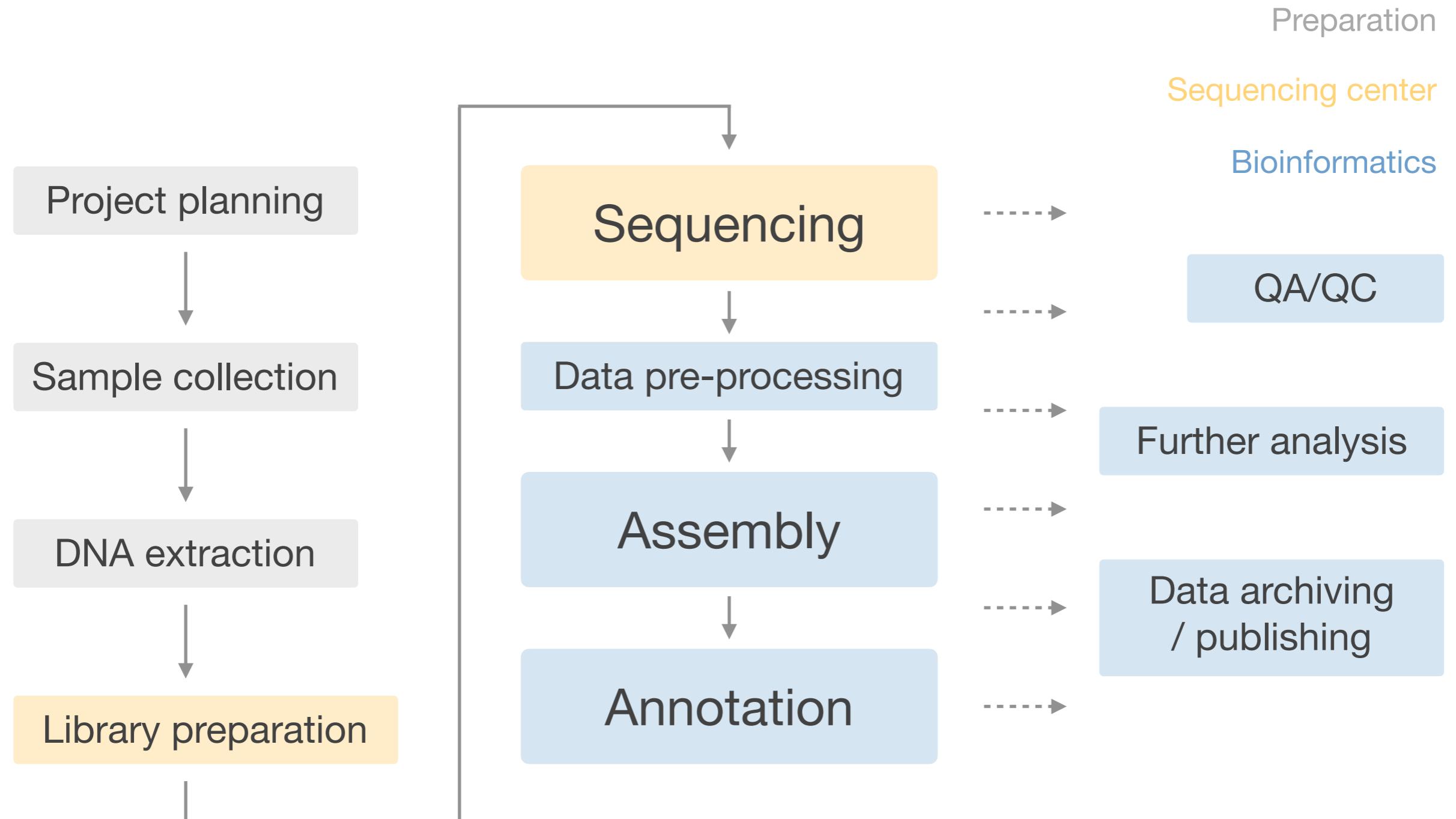
# De Bruijn graphs



Compeau and Tessler 2011 (Nature Biotechnology)

# *De novo* genome sequencing workflow

Legend



# Genome annotation

- Identifying the location, structure, and putative function of genetic elements
  - repetitive and mobile DNA
  - protein-coding genes
  - regulatory elements

```
AGTCATTATTCTGCATCAACTTAATCACACAAACCTATTGGCTCGTTCAGCAGAGAAAGCAGATCAGTGAGCGTTC  
AAGGACCAATCATATTGAAAATCACTGAGCGTTACTGATCTGCTTCTCTGCTGAACGCAGCCATTCTGCGCTAA  
ATCCTTCTATATATTATTCTGTAATTCTGTTACTGAGGATTAATAAATTATAAATGTATTACAGTCGAGGAAGCATAAAACCACAAAAATGGTA  
AGTTTCTAAGTATGTAGAGGATTAATAAATTATAAATGTATTACAGTCGAGGAAGCATAAAACCACAAAAATGGTA  
AAAATTCTATGCTTTCAATATATGCTAGATATGATCTTGTATATCATTATTATAATTGTTAACTGTAATCGT  
GATTTGCTATTCTCTTTCTAAATTCTGCTCTATCTACTATGAGAAATATAGATTCTATTTCATTTCTCAAATCT  
AGTTAATCTAAATATCAATGTATGTAATGAGTTTATTACCTGATCATATAGAAAAAGGCTCAGGAGAGCATTCAAC  
CTCCGCCAAGGCAAAGCTAAATTGGTCATCATTGCTAGCAATACGCCACCGCTAAGGTGAGACTGAGAAGTAGCACT  
TTAATGAATATAATCTTGATAAAATGTATACATATATACATATACATATTGTAAAAATTATATTCTATTGTAATT  
GTAATATTGATATAATCATCAATTGCAATTGGAAAACAATAGGAGTACTATGGAGTAGTCTCCAGCTCTGCTTG  
GAAGTCGGAGATTGAATACTATGCAATGTTAGCGAAGACTGGTGTGCATCATTACACGGGAATAACATCGAACTGGG  
TATAAAATCTATGCCAACTGGTGTACAGCGTAATGTACAGTTTAACTCCAATAAAATTCAAAACGTTTGATT  
TATTAATCTATATTAGTTATTTAGCAAGAGTAACATCAAAGATTCTCCTTAGATTTCACCGACTTGAAATA  
CTGAAGTTTCATTGTGTACATTAAACAAATCCCATTTCACGAATGCTAAGAATTGTACAATAAAATATAG  
AAAATCTTTAACATTATCCAAAAACTTTATACATATGTATATGTACATATACACCGCGCGCGTGT  
GTGCATCGTTCTCCAAATGATTGGTTGTTACAATAACGACAAATTGACCTTGCACCGACATGCAGTTACTGCA  
ATCTGCCACAGTTGTACATAGTCGCCAAATGCCGACAACGTGATTGAATATGATTGTTCCACATTAGTACAA  
ACACCCCAGGCTGGACATCGTTCCGATCATTGCTGCCACATCGAACACATTCTATTATAGATTAGTGCAGGAA  
ATAACAGCTAGATTGCTTTCAAAAACGGTATCCAATGTTCCCATGATTGTTATGGTAACTGCAGCTTCC  
TCGCCTCAACAGCAATGGACTGCAAGAAATATACAAATTATTAGAACATGCAATTGAGATTAGACATGATTAGCTA  
TTTTTATCTTCATACAAGTAATAGATATCAAACCTTAAAGCTACTTATGTTCTGCAATTCTCATTCA
```

# Repeat annotation

- Most genomes contain large amounts of repetitive and mobile DNA
- Repeat classes:

Repetitive Elements Identified in the *Montipora capitata* Genome Assembly

	Number	Total Length (Mb)	Fraction of Assembly (%)
Tandem repeats	160,985	16.5	2.7
Interspersed repeats	1,028,006	257.4	41.9
DNA elements	31,667	11.6	1.9
LTR elements	14,753	10.0	1.6
Non-LTR elements	110,728	42.6	6.9
Unclassified	870,858	193.2	31.4

- Repeats may have regulatory or structural functions, but can also bias downstream analyses

Helmkampf et al. 2019 (Genome Biology and Evolution)

# Repeat annotation

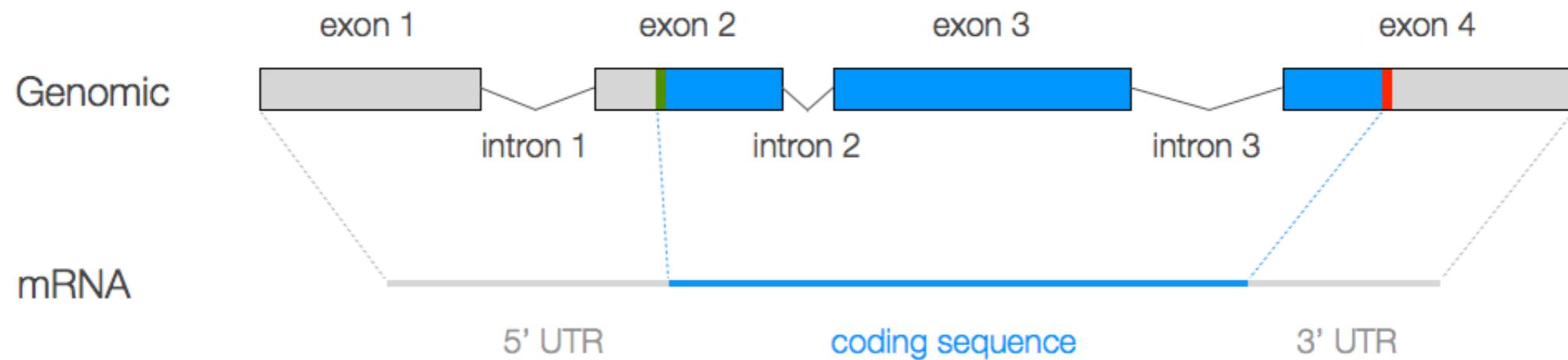
- **RepeatModeler** combines multiple repeat search engines and curated databases (e.g. Repbase) to identify repeats
- **RepeatMasker** marks repeats in assembly (replacement with Ns or soft-masking)

```
>Sc0000000
TAATGGCAGTCCAATCTGTACCCCTGATTATAATTCATTCAAATAATTCTTTTTATCCAATGATCAACCAAATGCATTGCTGCATTAAAAACTC
TTGTCAGCATTAAGTCCTCAGATATGGATCAACAAACACAGTAAAACCATTAAATCACTAAACTTAAAATAATAATTATTATTGAAGAATCTAAAtca
tcatttcttcacaggaacacacaagccaaacaaattgacctgttccaaacagattggcttcatalogtcagttggttctagcaattgcattcattggta
tcacagaggtaatgggttcaaattccattggagctgcctgaattttggtgtctTTTAAAGTGAUTCAATTCTATCTTCTCCCTGAATGCAGCG
TGTTAAGTATGCTTTTGAAACACAAGTTGAGAATCCTGTACTGCAAACAACTAAATCCTTAGTTGCAGTGTATacacacacacacacacatacTGGAA
AATTACGGGTAAGAAATGTCAATCTCTCCTCATTCCAGTTCTCTCCAAAATAATGTTGAAGTTGAGGTAATGGTGGACCACGTGAATAGTGAATTCCTCCT
GTTTCATCCTGTTATGTTACAATCACATGTACATTGATTAATGGAAATATAATCCCCCTTAGAACCGATTGCAGACAGTCATTCTTAACTTG
...

```

# Gene prediction

In eukaryotes, gene prediction is difficult due to coding sequence (exons) being interrupted by introns



# Gene prediction

- Complex pipelines (e.g. AUGUSTUS) combine multiple approaches
  - *ab initio* prediction (pattern-based)
  - extrinsic evidence can include
    - RNA-seq (transcripts)
    - homologous protein sequences
  - Sensitivity and specificity are increased by multiple rounds of training
  - Repeats should be masked
- Computational gene prediction always requires careful quality control

# Gene prediction

# The GFF format: gene structure, coding sequence

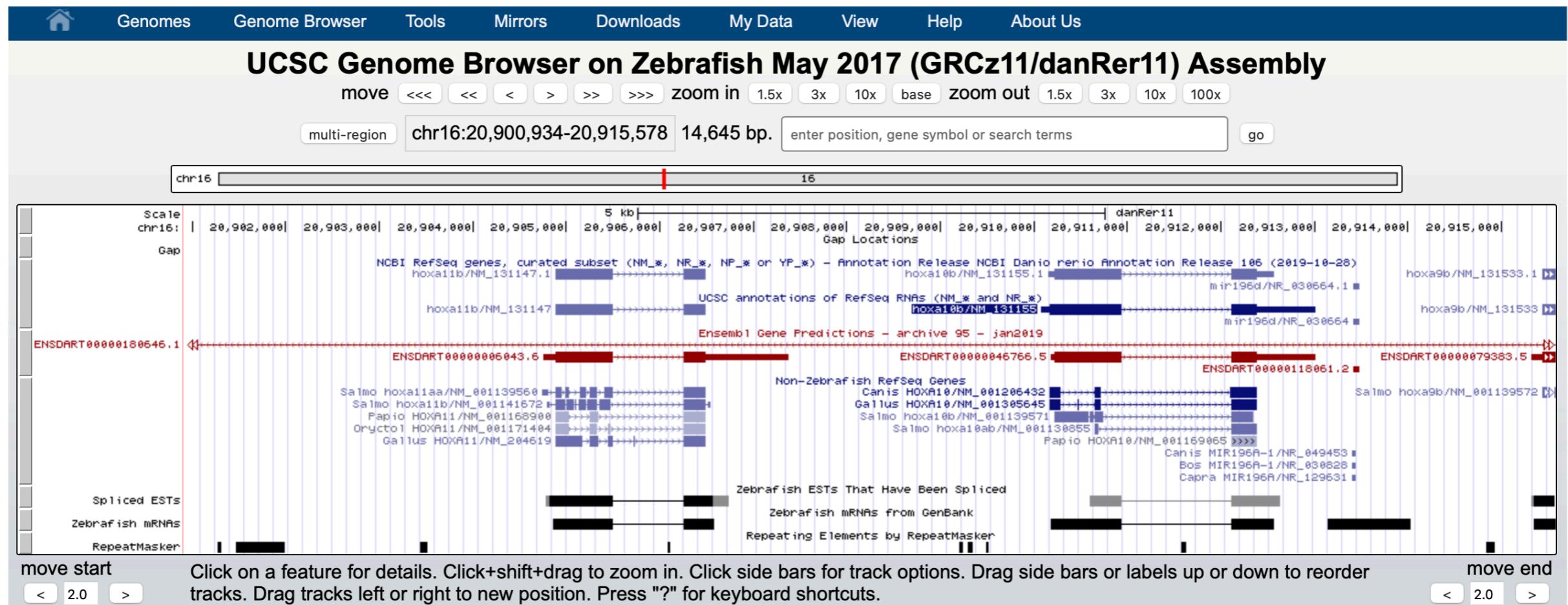
```
# start gene g29
Sc0000000 AUGUSTUS gene 656020 658053 0.9 - . g29
Sc0000000 AUGUSTUS transcript 656020 658053 0.9 - . g29.t1
Sc0000000 AUGUSTUS stop_codon 656020 656022 . - 0 transcript_id "g29.t1"; gene_id "g29";
Sc0000000 AUGUSTUS intron 656165 657273 1 - . transcript_id "g29.t1"; gene_id "g29";
Sc0000000 AUGUSTUS intron 657509 657956 1 - . transcript_id "g29.t1"; gene_id "g29";
Sc0000000 AUGUSTUS CDS 656020 656164 1 - 1 transcript_id "g29.t1"; gene_id "g29";
Sc0000000 AUGUSTUS CDS 657274 657508 1 - 2 transcript_id "g29.t1"; gene_id "g29";
Sc0000000 AUGUSTUS CDS 657957 658053 0.9 - 0 transcript_id "g29.t1"; gene_id "g29";
Sc0000000 AUGUSTUS start_codon 658051 658053 . - 0 transcript_id "g29.t1"; gene_id "g29";
# coding sequence = [atgaccgcgttcgaaagcgacgattctcaattggatgaagatagtgatcatggaaattgttcagcaaaacggttggccc
# caaaaagaaaaattcaggtcggtgcgcaggaatcaaaagaatcatatgcctgggtcccaagggtgatgtaactccagcgagcagtgacgatgaa
# ctggagcggaaacgtctttatgaactctgttaccttaaccagtcgagggtttccaaccctataccgtgagctccggagaggacgactttga
# tgaccttagtgtggagacatgaaaaagcgtcgagctaaaaacgaaagaattccgacaacgcgaagatacccacgatacacctgattgacagaca
# atgagggagagctgggttaacaccttaagccgccttacgtcaacgaaggtgcccacacttcggatgatgccaatctgaactggaaaaggaa
# aaatga]
# protein sequence = [MTRFESDDSQLDEDSDHGNCSAKRLAPRKISGRLRRKSKESYRLGPKVDPASSDDELERETSFMNSGTFKPVEGF
# SNPPIPVSSEDDFDDLSVGDMKKRSSLTKEFRQREDTHDTPDLTDNEGELGFNTFPASVNEGAHTSDDAESELEKEK]
```

# Gene prediction

- Gene structure (UTRs, exons, introns)
- Coding sequence
- Transcription status and splice variants
- Putative identity and function  
(homology-based, e.g. by BLAST versus [UniProt](#) database)
- Computational + manually curated gene models > [Official Gene Set](#)

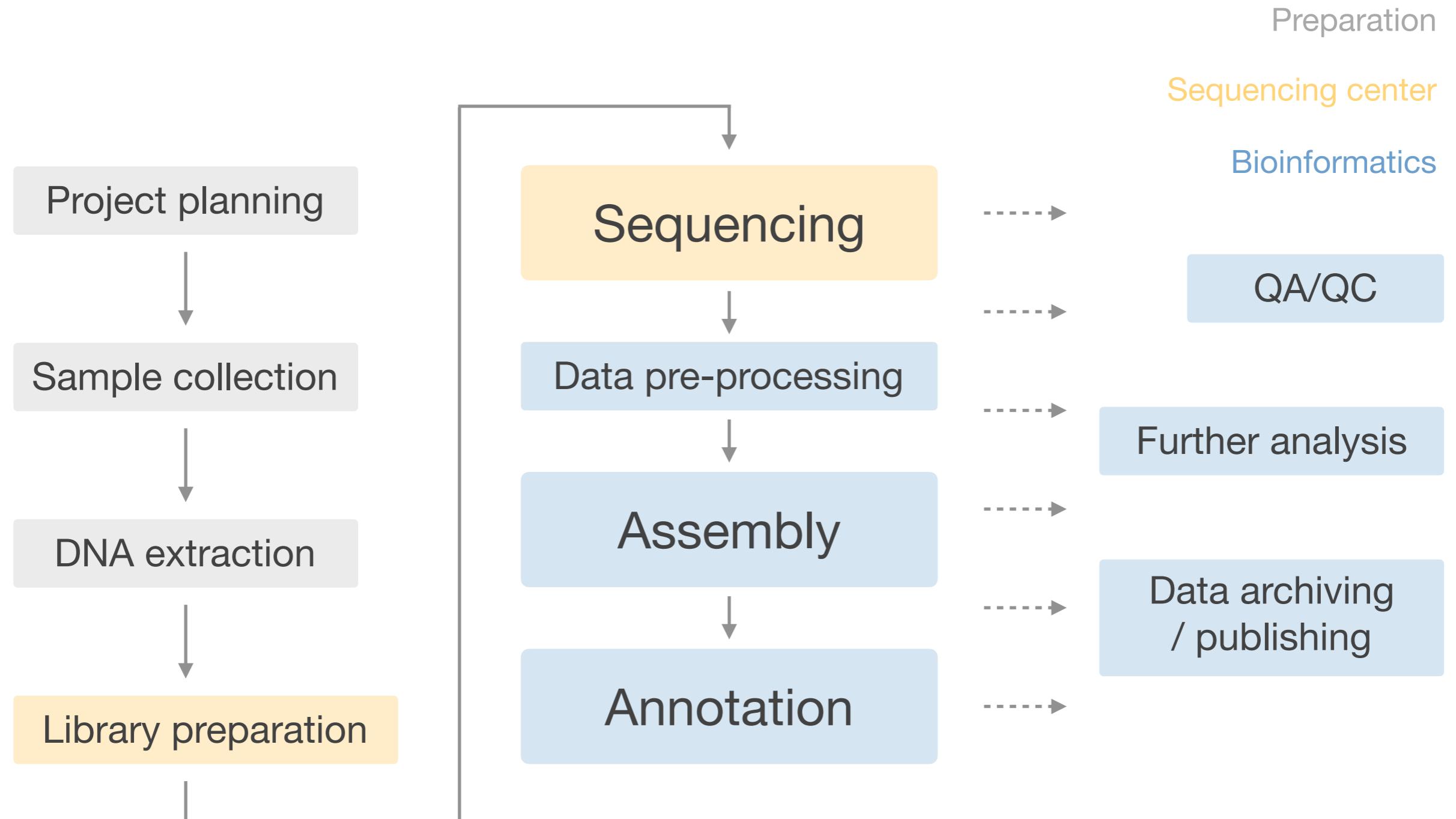
# Visualization

## GBrowse / JBrowse genome browser



# *De novo* genome sequencing workflow

Legend

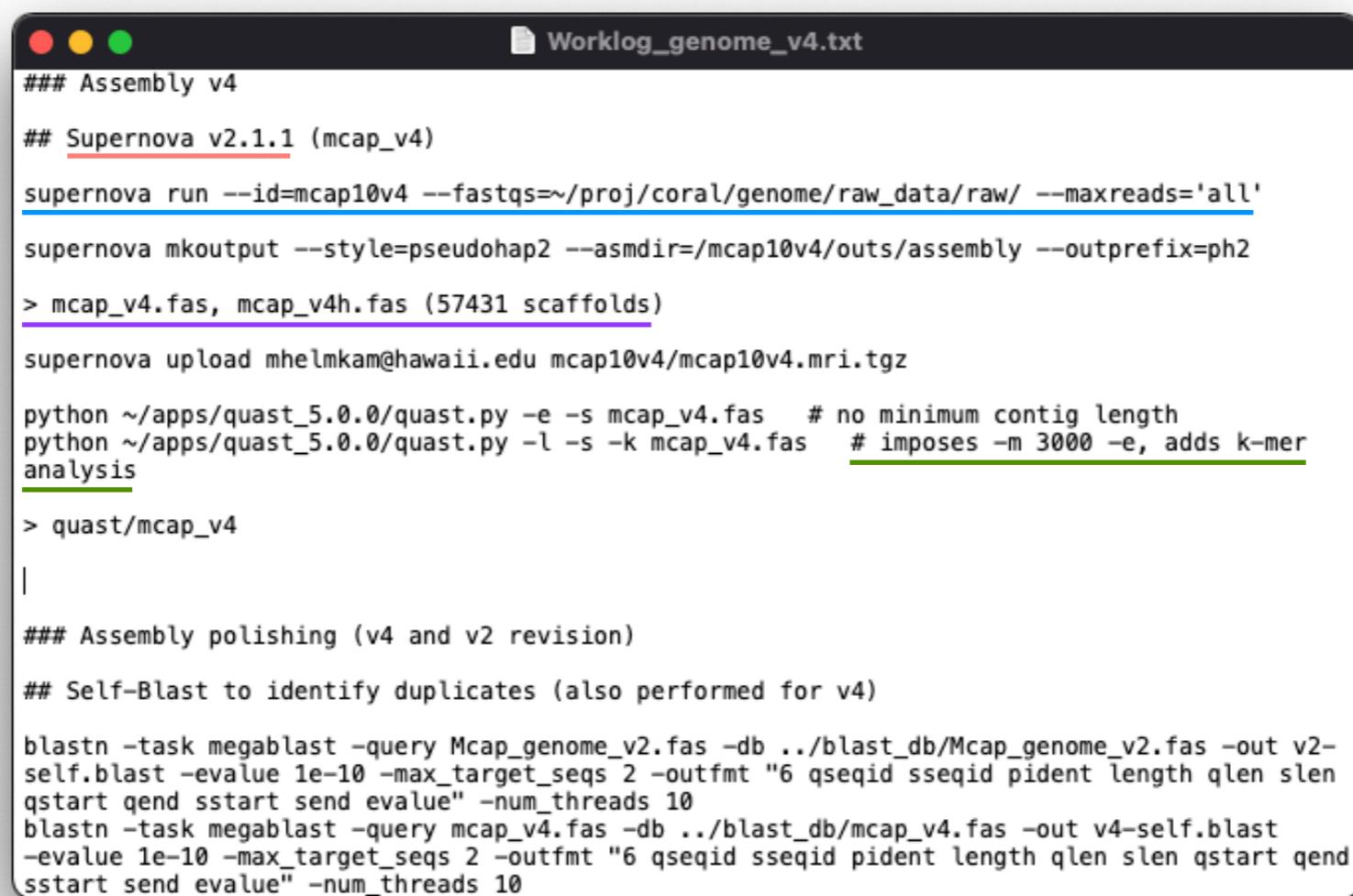


## 4. Data management

- Documentation
- Data organization
- Backing up / Version control
- Publishing / Sharing data

# Documentation

- Document your workflow in plain text format (Atom /TextEdit / Notepad)
- Include **code**, **software versions**, **summary results**, and **comments**



```
Worklog_genome_v4.txt
### Assembly v4
## Supernova v2.1.1 (mcap_v4)
supernova run --id=mcap10v4 --fastqs=~/proj/coral/genome/raw_data/raw/ --maxreads='all'
supernova mkoutput --style=pseudohap2 --asmdir=/mcap10v4/outs/assembly --outprefix=ph2
> mcap_v4.fas, mcap_v4h.fas (57431 scaffolds)
supernova upload mhelmkam@hawaii.edu mcap10v4/mcap10v4.mri.tgz
python ~/apps/quast_5.0.0/quast.py -e -s mcap_v4.fas # no minimum contig length
python ~/apps/quast_5.0.0/quast.py -l -s -k mcap_v4.fas # imposes -m 3000 -e, adds k-mer
analysis
> quast/mcap_v4
|
### Assembly polishing (v4 and v2 revision)
## Self-Blast to identify duplicates (also performed for v4)
blastn -task megablast -query Mcap_genome_v2.fas -db ../blast_db/Mcap_genome_v2.fas -out v2-
self.blast -eval 1e-10 -max_target_seqs 2 -outfmt "6 qseqid sseqid pident length qlen slen
qstart qend sstart send evalue" -num_threads 10
blastn -task megablast -query mcap_v4.fas -db ../blast_db/mcap_v4.fas -out v4-self.blast
-eval 1e-10 -max_target_seqs 2 -outfmt "6 qseqid sseqid pident length qlen slen qstart qend
sstart send evalue" -num_threads 10
```

# Data organization

## Formats

- Standard file formats (e.g. plain text, pdf, png)
- Standard data formats (e.g. FASTA, FASTQ, VCF, GFF)

# Data organization

## Naming scheme

- Uniform, persistent
- Comprehensive, includes important information
- No spaces and special characters (use underscores instead)
- Consistent file extensions (e.g. .fastq, .vcf, .gff)
- May include date (YYYYMMDD) or version number

# Data organization

## Naming scheme – Examples

What is it?

Species acronym

Hpue\_raw300\_F.fastq.gz

File / data format

Mcap\_genemodels\_1.1\_aa.fas

hyp155\_a\_0.33\_mac4\_5kb.raxml.log

Dataset name

Version number

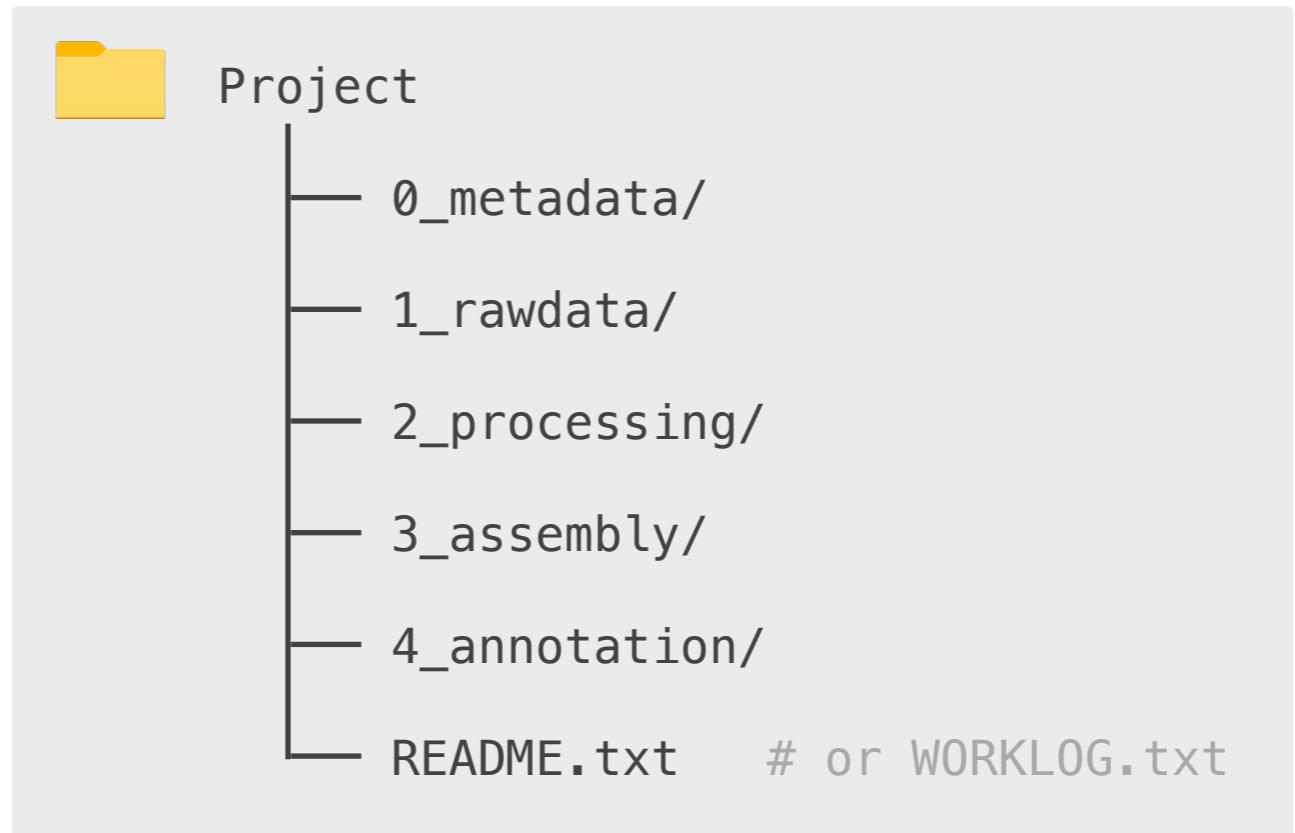
Important parameters

# Data organization

## Folder structure

- Be consistent
- Separate ongoing and completed work
- Consider links instead of duplicating files
- Include documentation (readme or worklog)

## Example



# Backing up / transferring data

- Keep raw data on dedicated hard drive
- Process and analyze working copies only
- Confirm data integrity using checksums after transfer

```
rsync -arvz <source_file> <destination_file> # improved copy (-a: archive mode, -r:  
# include subdirectories, -z: compress  
# during transfer)  
  
md5 <file> # calculate checksum (not working in Git  
# Bash)
```

# Version control

## Typical git usage

```
git init                                # create git repository based on current folder  
git status                               # show repository status  
git add *                                 # track all files (exceptions in .gitignore)  
git commit -m "title"                      # record changes  
  
git push                                  # push to remote GitHub repository (if set up)
```

## Example .gitignore file

```
1_rawdata/  
*.fastq.gz  
*.fastq
```

# Publishing / sharing genome data

FAIR principles – **findable, accessible, interoperable, reusable**

Key attributes:

- unique and persistent identifier
- clear and detailed metadata
- standardized data format and metadata vocabulary
- accessible data usage license (e.g. open access)

# Public sequence databases



# Public sequence databases

NCBI Genbank (<https://www.ncbi.nlm.nih.gov/genbank/>)

The screenshot shows the NCBI Nucleotide database homepage. At the top, there's a search bar with 'Assembly' selected. To the right of the search bar is a 'Search' button. Below the search bar, there's a 'COVID-19 Information' box containing links to CDC, NIH, SARS-CoV-2 data, HHS, and Español resources. The main content area is titled 'Nucleotide' and describes the database as a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery. On the left sidebar, under 'Using Nucleotide', there are links to Quick Start Guide, FAQ, Help, GenBank FTP, and RefSeq FTP. In the center sidebar, under 'Nucleotide Tools', there are links to Submit to GenBank, LinkOut, E-Utilities, BLAST, and Batch Entrez. On the right sidebar, under 'Other Resources', there are links to GenBank Home, RefSeq Home, Gene Home, SRA Home, and INSDC. At the bottom left, it says 'You are here: NCBI > DNA & RNA > Nucleotide Database'. At the bottom right, there's a 'Support Center' link.

Select database:  
Assembly

Search term:  
Scleractinia

Useful  
search tags  
[ACCN]  
[ORGN]  
[AUTH]  
[TITL]

# Public sequence databases

## NCBI Genbank | Assembly report

The screenshot shows the NCBI Genbank assembly report for the organism *Montipora capitata* (stony corals). The assembly is named **Mcap\_UHH\_1.1**. Key details include:

- Organism name:** [Montipora capitata \(stony corals\)](#)
- Isolate:** Colony #1
- BioSample:** [SAMN10221787](#)
- BioProject:** [PRJNA495325](#)
- Submitter:** University of Hawaii
- Date:** 2019/07/02
- Assembly level:** Scaffold
- Genome representation:** full
- RefSeq category:** representative genome
- GenBank assembly accession:** GCA\_006542545.1 (latest)
- RefSeq assembly accession:** n/a
- RefSeq assembly and GenBank assembly identical:** n/a
- WGS Project:** [RDEB01](#)
- Assembly method:** Supernova v. 2.1.1
- Expected final version:** yes
- Genome coverage:** 69.0x
- Sequencing technology:** Illumina HiSeq

IDs: 3590631 [UID] 11597468 [GenBank]

[History](#) ([Show revision history](#))

[Comment](#)

[Global statistics](#)

Total sequence length: 614,509,607

**Global statistics**

Total sequence length	614,509,607
Total ungapped length	571,886,739
Gaps between scaffolds	0
Number of scaffolds	27,865
Scaffold N50	185,537
Scaffold L50	747
Number of contigs	50,174
Contig N50	24,266
Contig L50	6,689
Total number of chromosomes and plasmids	0
Number of component sequences (WGS or clone)	27,865

Send to: [Download Assembly](#)

Access the data

- BLAST the assembly
- Run Primer-BLAST
- Full sequence report
- Statistics report
- FTP directory for GenBank assembly
- NCBI Datasets NEW

Assembly Information

- Assembly Help
- Assembly Basics
- NCBI Assembly Data Model

Related Information

- BioProject
- BioSample
- Genome
- PubMed
- Taxonomy

# Public sequence databases

## NCBI Genbank | BioSample

The screenshot shows a web browser window for the NCBI BioSample database. The URL in the address bar is [ncbi.nlm.nih.gov/biosample/SAMN10221787/](https://ncbi.nlm.nih.gov/biosample/SAMN10221787/). The page title is "Sperm sample of Montipora cap...". The main content area displays detailed information about a BioSample entry for a sperm sample of *Montipora capitata*. Key details include:

- Identifiers:** BioSample: SAMN10221787; Sample name: Sample #1; SRA: SRS3943439
- Organism:** *Montipora capitata* (cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Montipora)
- Package:** MIGS: eukaryote; version 5.0
- Attributes:**
  - isolate:** Colony #1
  - isolation source:** sperm
  - collection date:** 2016-07-06
  - broad-scale environmental context:** coral reef
  - local-scale environmental context:** tidal pool
  - environmental medium:** sea water
  - estimated size:** 700 Mb
  - geographic location:** USA: Hawaii, Waiopae tide pools
  - isolation and growth condition:** filtered sea water
  - latitude and longitude:** 19.4986 N 154.8183 W
  - number of replicons:** 2
  - ploidy:** diploid
  - propagation:** sexual
- Description:** Sperm sample from *Montipora capitata* gamete bundles released during spawning event (colony #1)  
Keywords: GSC:MiS;MIGS:5.0
- BioProject:** PRJNA495325 *Montipora capitata* isolate:Colony #1  
Retrieve [all samples](#) from this project

On the right side, there are sections for "Related information" (BioProject, SRA, Nucleotide, Assembly, Taxonomy) and "Recent activity" (listing previous searches like "Sperm sample of Montipora capitata" and "Montipora capitata isolate:Colony #1").

# Public sequence databases

## NCBI Genbank | BioProject

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (WGS master)	1
SRA Experiments	1
PUBLICATIONS	
PubMed	1
PMC	1
OTHER DATASETS	
BioSample	1
Assembly	1

Assembly details:

Assembly	Level	WGS	BioSample	Taxonomy
GCA_006542545.1	Scaffold	RDEB00000000	SAMN10221787	Montipora capitata

SRA Data Details

Parameter	Value
Data volume, Gbases	55
Data volume, Mbytes	27719

Display Settings: [Advanced](#) [Browse by Project attributes](#)

Montipora capitata isolate:Colony #1

Accession: PRJNA495325 ID: 495325

Montipora capitata isolate:Colony #1 Genome sequencing and assembly

The rice coral *Montipora capitata* is widely distributed throughout the Indo-Pacific, and comprises one of the most important reef-building species in the Hawaiian Islands. [More...](#)

Send to: [Related information](#) [Recent activity](#)

Assembly

BioSample

Full text in PMC

Genome

Nucleotide

PubMed

SRA

Taxonomy

WGS master

See Genome Information for [Montipora capitata](#)

NAVIGATE ACROSS

5 additional projects are related by organism.

Recent activity

Turn Off Clear

Montipora capitata isolate:Colony #1 BioProject

Sperm sample of Montipora capitata biosample

Mcap\_UHH\_1.1 - Genome - Assembly - NCBI Assembly

montipora capitata AND (latest[filter] AND all[filter] NOT anomalous[filter]) (1) Assembly

See more...

Accession: PRJNA495325

Data Type: Genome sequencing and assembly

Scope: Monoisolate

Organism: [Montipora capitata](#) [Taxonomy ID: 46704]  
Eukaryota; Metazoa; Cnidaria; Anthozoa; Hexacorallia; Scleractinia; Astrocoeniina; Acroporidae; Montipora; Montipora capitata

Publications: Helmkampf M et al., "Draft Genome of the Rice Coral *Montipora capitata* Obtained from Linked-Read Sequencing.", *Genome Biol Evol*, 2019 Jul 1;11(7):2045-2054

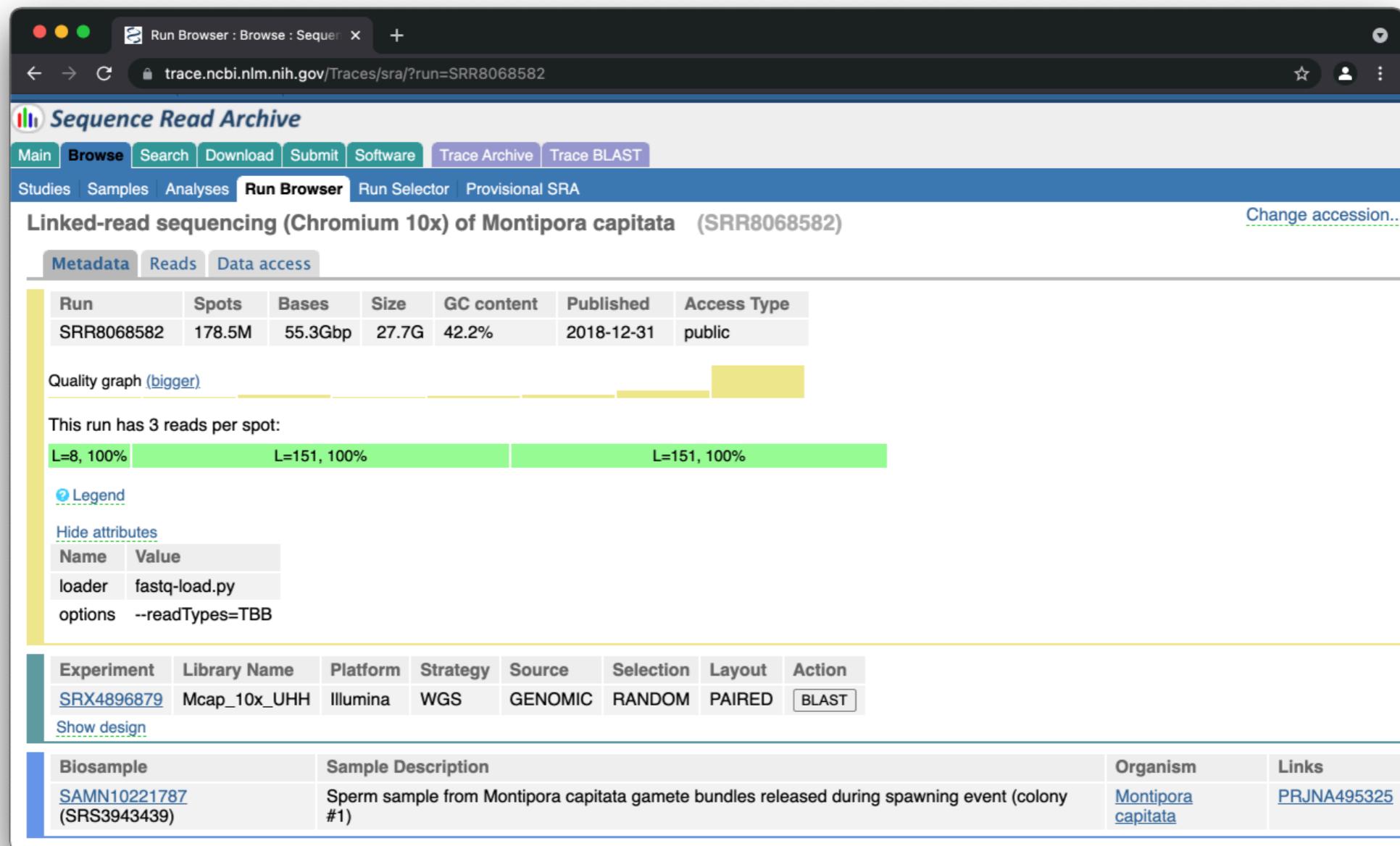
Grants:

- "Pacific High Island Evolutionary Biogeography: Impacts of Invasive Species, Anthropogenic Activity and Climate Change on Hawaiian Focal Species" (Grant ID EPS-0903833, United States National Science Foundation, Established Program to Stimulate Competitive Research)
- "CREST Center in Tropical Ecology and Evolution in Marine and Terrestrial Environments" (Grant ID 0833211, United States National Science Foundation, Center for Research Excellence in Science and Technology)
- "Understanding Biotic Response to Environmental Change in Tropical Ecosystems Through a Place-Based Context" (Grant ID 1345247, United States National Science Foundation, Center for Research Excellence in Science and Technology)

Submission: Registration date: 2-Jul-2019  
University of Hawaii

# Public sequence databases

## NCBI Genbank | Sequence Read Archive (SRA)



# Genome submission

- NCBI genome submission guide:

<https://www.ncbi.nlm.nih.gov/genbank/genomesubmit/>

- Submission portal login (<https://submit.ncbi.nlm.nih.gov/subs/genome/>)
  - Log in with NCBI, Google, or ORCID etc. account
  - Start new or edit existing submission

# Genome submission

The screenshot shows the 'Submission Portal' interface for the National Library of Medicine. The top navigation bar includes links for Home, My submissions (which is currently selected), Manage data, Groups, Templates, and My profile. The user 'mhelmkampf' is logged in. The main content area is titled 'Your submissions' and features a 'Start a new submission' section with options like GenBank, BioProject, Sequence Read Archive, BioSample, Genome, Supplementary Files, and API. Below this is a table of 8 submissions:

Submission	Title	App	Group	Status	Updated
SUB4625258	Montipora capitata genome assembly	WGS		✓ Genomes: Processed (Details)	Jul 02 2019
SUB4632316	Montipora capitata Genome sequencing and assembly, Oct 15 '18	Sequence Read Archive (SRA)		✓ SRA: Processed SRR8068582 Download metadata file with SRA accessions View and manage my SRA submission data	Oct 17 2018
SUB4611388	Montipora capitata genome sequencing and assembly	BioProject		✓ BioProject: Processed PRJNA495325 : Montipora capitata genome sequencing and assembly (TaxID: 46704) Locus Tag Prefixes: • D8914	Oct 09 2018
SUB4611414	Sperm sample of Montipora capitata	BioSample		✓ BioSample: Processed (Details)	Oct 09 2018
SUB4139132	Corvus hawaiiensis diploid de novo genome assembly	WGS		✓ Genomes: Processed (Details)	Aug 16 2018

New submission  
/ register:  
[BioProject](#)  
[BioSample](#)  
[SRA](#)  
[Genome \(WGS\)](#)

# Genome submission

- Provide project description, sample metadata, technical details (method, coverage etc.), contact information, publication if applicable
- Upload assembly as single **FASTA file** (minimum contig length 200 bp)
  - Chromosome-level?
  - What do 'Ns' represent?
- Upload reads to SRA in FASTQ format
- Batch upload for multiple genomes

# Genome submission

- Automated validation:
  - Contamination
  - Adapter sequences
  - Duplicate contigs
- After successful review, **accession numbers** will be issued
- **Public release** at requested date or publication date, whichever comes first
- Data may run through database annotation pipelines

# Dryad

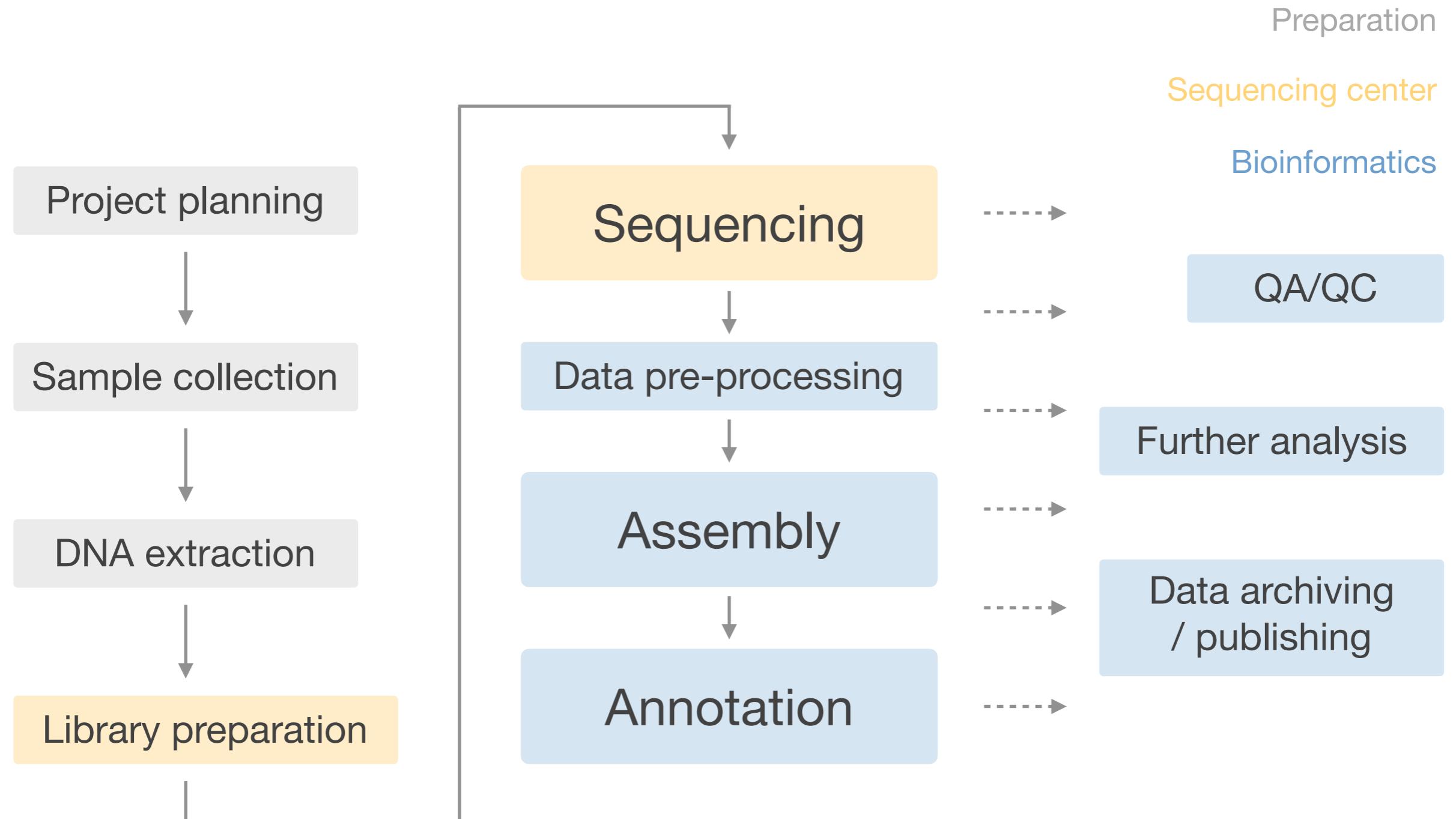
- <https://datadryad.org>
- Free open-access repository for data in any format  
(e.g. GFF, official gene set FASTA, genotype data in VCF format)
- May or may not be linked to publication
- Data receive permanent DOI

# Code and workflow documentation

GitHub – Example: [https://k-hench.github.io/hamlet\\_radiation/](https://k-hench.github.io/hamlet_radiation/)

# *De novo* genome sequencing workflow

Legend



# Thank you!

Please leave feedback for this course here (Google form):

<https://tinyurl.com/9hucxfvh>

Questions?

Email me at [martin.helmkampf@leibniz-zmt.de](mailto:martin.helmkampf@leibniz-zmt.de)

