

15 Sept 2025

# ***Legionella* Tools**

**Advanced Molecular Detection  
Southeast Region Bioinformatics**

This resource was made possible through funding provided under the Epidemiology and Laboratory Capacity for Prevention and Control of Emerging Infectious Diseases (ELC) Cooperative Agreement (CK24-0002), Project D: Advanced Molecular Detection to the Florida Department of Health. The conclusions, findings, and opinions expressed by authors do not necessarily reflect the official position of the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention.

# Overview

**Legsta**

**El Gato**

**Legiovue**

**Other Tool: Ksnp**

# Overview

## Legsta

### Purpose

- A tool is developed for the sequence-based typing (SBT) of *Legionella pneumophila* which enables standardized identification of allelic profiles and sequence types directly from the assembled genome contigs

### Usage

- Can be used by public health laboratories, and researchers to perform molecular typing for outbreak investigation, surveillance and comparative genomic studies

### Dependencies

- Installation of Legsta:
  - perl
  - isPcr
  - any2fasta

# Overview

## El Gato

### Purpose

- A tool is developed for genotyping *Legionella pneumophila* isolates using sequence bloom trees (SBTs), providing identification of sequence types from Illumina paired end short reads data

### Usage

- Can be used by public health laboratories, and researchers to assign sequence types for outbreak detection, genomic epidemiology and routine surveillance

### Dependencies

- Installation of elgato:
  - Python3
  - Snakemake
  - sourmash

# Overview

## Legiovue

### Purpose

- A tool is developed for comprehensive genomic analysis of *Legionella pneumophila* integrating quality control, assembly, subtyping, MLST, serogroup prediction and contamination screening

### Usage

- Can be used by public health laboratories, and researchers to generate standardized genomic surveillance reports, outbreak analyses, and comparative studies of *Legionella pneumophila* isolates

### Dependencies

- Nextflow
- Apptainer/Singularity
- Kraken2
- SPAdes
- Fastp
- ChewBBACA
- El\_gato

# Overall Workflow – Option 1

*Legionella* Sequences (FASTQs)

Phoenix/Sanibel  
(QC + Assembly)

Legsta

El Gato

Ksnp 3/4

Report Generation

# Resource Links

## Github Links

### BPHL-Molecular:

<https://github.com/BPHL-Molecular>

### Phoenix:

<https://github.com/CDCgov/phoenix>

### Sanibel:

<https://github.com/BPHL-Molecular/Sanibel>

### Report Generation:

<https://github.com/BPHL-Molecular/ReportGen>

## Previous Trainings

### Phoenix:

[https://github.com/StaPH-B/southeast-region/blob/master/trainings/office%20hours/20250609\\_Session50\\_Pipeline%20Training%20Pt%2012\\_PHoeNlx.pdf](https://github.com/StaPH-B/southeast-region/blob/master/trainings/office%20hours/20250609_Session50_Pipeline%20Training%20Pt%2012_PHoeNlx.pdf)

### Sanibel:

[https://github.com/StaPH-B/southeast-region/blob/master/trainings/office%20hours/20241028\\_Session39\\_Pipeline%20Training%20Part%202\\_Sanibel.pdf](https://github.com/StaPH-B/southeast-region/blob/master/trainings/office%20hours/20241028_Session39_Pipeline%20Training%20Part%202_Sanibel.pdf)

### Report Generation:

[https://github.com/StaPH-B/southeast-region/blob/master/trainings/office%20hours/20250818\\_Session53\\_Pipeline%20Training%20Pt%2016\\_ReportGen.pdf](https://github.com/StaPH-B/southeast-region/blob/master/trainings/office%20hours/20250818_Session53_Pipeline%20Training%20Pt%2016_ReportGen.pdf)



# Ksnp3/4 Info

Git Clone

<https://github.com/kissake/kSNP4>

```
#!/bin/bash -evx

# For debugging, use below as first line.
#!/bin/bash -vx

# Examples are default in Example/Example<#>/...

# Since this script will be in the source, default to making sure that the
# binary dependencies (Jellyfish, mummer, consensus, etc.) are available:

echo "Checking dependencies:"
[ -x jellyfish ]      && echo "... Jellyfish found in current working directory (cwd)"      || echo "Jellyfish missing"
[ -x mummer ]         && echo "... mummer found in cwd"                                || echo "mummer missing"
[ -x consensus ]      && echo "... consensus found in cwd"                            || echo "consensus missing"
[ -x FastTreeMP ]     && echo "... FastTreeMP found in cwd"                            || echo "FastTreeMP missing"
[ -x parsimonator ]   && echo "... parsimonator found in cwd"                        || echo "parsimonator missing"
[ -x kSNP4 ]          && echo "... kSNP4 found in cwd"                                || echo "kSNP4 missing"

MISSINGEXAMPLES=0
echo "Checking for required example files:"
[ -d Examples/ ] || MISSINGEXAMPLES=1
[ -d Examples/Example2/ ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/annotated_genomes ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/in_list2 ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/in_list ] || MISSINGEXAMPLES=1
[ -d Examples/Example2/Genomes ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/Genomes/Vc63-93_MOM5.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/Genomes/Vc01-ElTorN16961.fa ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/Genomes/VCM56.fa ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/Genomes/Vc523-80.fa ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/Genomes/VCLM3984-4.fa ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/Genomes/ERR579925.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example2/CommandLines.txt ] || MISSINGEXAMPLES=1
[ -d Examples/Example1/ ] || MISSINGEXAMPLES=1
[ -d Examples/Example1/Genomes ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_PE-0_0155.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_NJ-60.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_BeAr436087.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_FL93-939.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_ref_g121218484.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_Georgia97.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_PE6.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_Florida91-4697.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_PE-3_0815.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/Genomes/EEE_NorthAmerican_antigenic_variety.fasta ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/ExampleRuns/in_list ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/ExampleRuns/annotated_genomes ] || MISSINGEXAMPLES=1
[ -f Examples/Example1/CommandLines.txt ] || MISSINGEXAMPLES=1

if [ ${MISSINGEXAMPLES} -gt 0 ]
then
    echo "Missing critical files under Examples/; are you sure you have unpacked the Examples?"
    exit 1
else
    echo "... Found examples."
fi

# Typically the in_list file must be modified to run the examples.

# The output directory and log output locations can be specified. There may
# be hidden path dependencies such that the output directory must be a sub-
# directory of the current.

# Current working directory:
CURRENT= pwd

# Directory for caching
CACHE="${HOME}/kSNP/"
# CACHE="" # No caching.
```

**For More Information:**

<https://github.com/kissake/kSNP4/blob/%40/kSNP3.1.2%20User%20Guide%20.pdf>

# Legsta Workflow

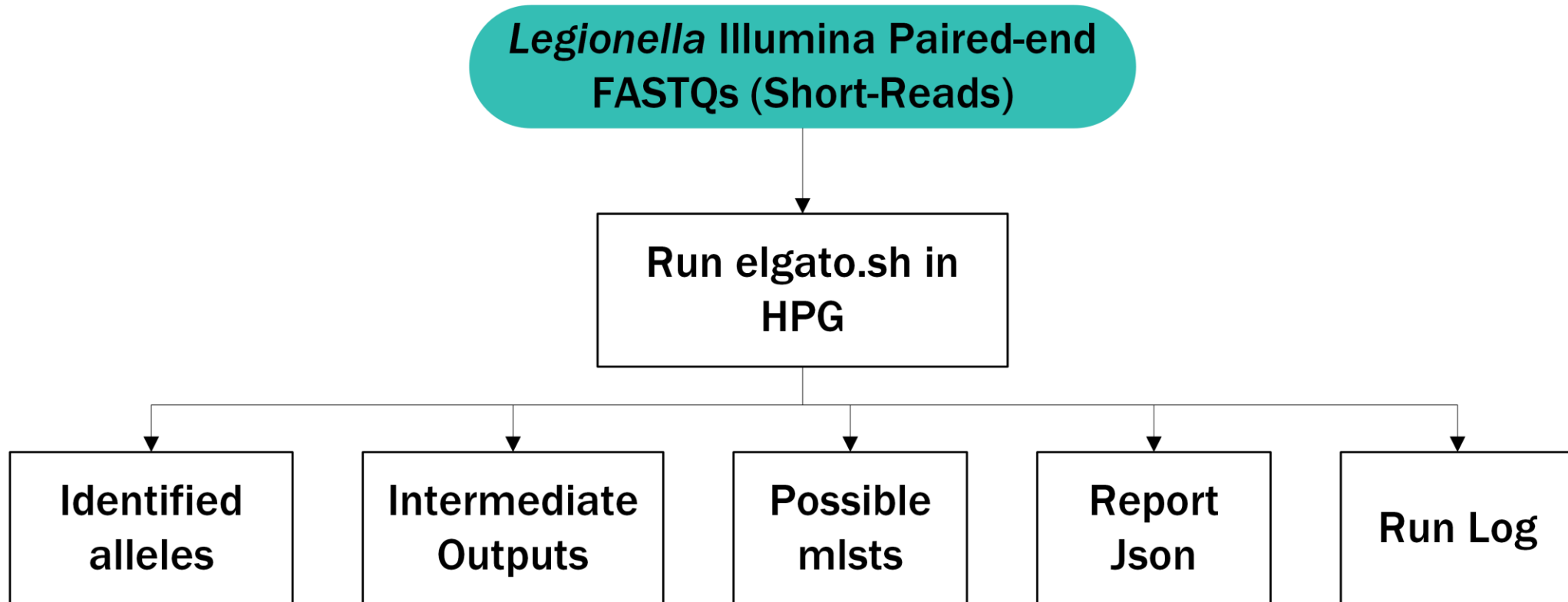
*Legionella* Assembled Contigs

*Legionella*  
*pneumophila* SBT  
Scheme Database

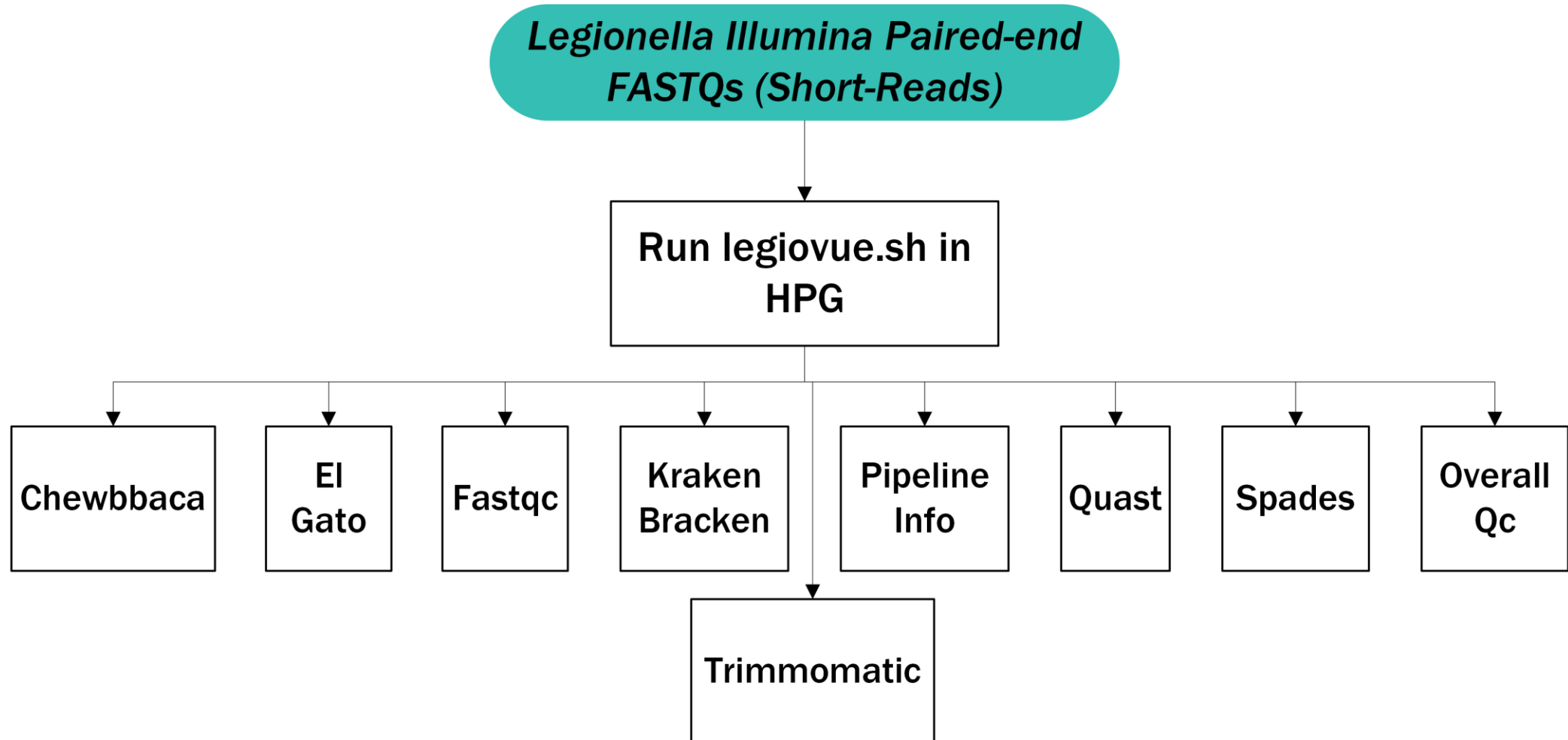
Run runlegsta.sh in  
HPG

Output-  
legsta\_result.tsv

# EL Gato Workflow



# Legiovue Workflow – Option 2



# Applications

1. Use assembled contig files from the lab, use **Legsta** to analyze the samples
2. Retrieve short read samples from NCBI database (DB), use **El Gato** to analyze the samples
3. Retrieve short read samples from NCBI DB, use **Legiovue** to analyze the samples

# Application Cont.

```
cd /blue/bphl-<state>/<user>/repos/bphl-molecular/  
git clone https://github.com/tseemann/legsta  
git clone https://github.com/appliedbinf/el_gato  
git clone https://github.com/phac-nml/legiovue  
mkdir analysis/  
cd analysis/  
cp /blue/bphl-<state>/<user>/repos/bphl-molecular/  
legsta  
cp /blue/bphl-<state>/<user>/repos/bphl-molecular/  
el_gato  
cp /blue/bphl-<state>/<user>/repos/bphl-molecular/  
legiovue
```

# 1. Application Cont.

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legsta0915/data/  
Name  
..  
GCA1_genomic.fna  
GCA2_genomic.fna  
GCA3_genomic.fna  
GCA4_genomic.fna  
GCA5_genomic.fna  
GCA6_genomic.fna  
GCA7_genomic.fna  
GCA8_genomic.fna  
GCA9_genomic.fna  
GCA10_genomic.fna  
GCA11_genomic.fna  
GCA12_genomic.fna  
GCA13_genomic.fna  
GCA14_genomic.fna  
GCA15_genomic.fna  
GCA16_genomic.fna  
GCA17_genomic.fna  
GCA18_genomic.fna  
GCA19_genomic.fna  
GCA20_genomic.fna
```

# 1. Application Cont.

```
GNU nano 5.6.1 runlegsta.sh
#!/bin/bash
#SBATCH --account=bphl-umbrella
#SBATCH --qos=bphl-umbrella
#SBATCH --job-name=Legsta
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=10
#SBATCH --mem=256gb
#SBATCH --time=48:00:00
#SBATCH --output=Legsta.%j.out
#SBATCH --error=Legsta.%j.err
#SBATCH --mail-user=nikhil.yengala@flhealth.gov
#SBATCH --mail-type=FAIL,END

# Load conda
module load conda || true
conda activate legsta

# Paths
DB="${CONDA_PREFIX}/db"
IN="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legsta0915/data/*.fna"
OUT="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legsta0915/legsta_results.tsv"

# Check database
test -f "${DB}/ispcr.tab" || { echo "DB missing ispcr.tab at ${DB}"; exit 2; }

# Run legsta
legsta --db "${DB}" "${IN}" > "${OUT}"

echo "[OK] results -> ${OUT}"
```



# 1. Application Cont.

**sbatch runlegsta.sh**

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legsta0915/
Name
└─ ..
└─ bin
└─ data
└─ db
└─ test
└─ Legsta.11650814.err
└─ Legsta.11650814.out
x └─ legsta_results.tsv
└─ LICENSE
└─ README.md
└─ runlegsta.sh
```

# 1. Application Cont.

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legsta0915/
Name
├── ..
├── bin
├── data
├── db
├── test
├── Legsta.11650814.err
├── Legsta.11650814.out
├── x legsta_results.tsv
├── LICENSE
├── README.md
└── runlegsta.sh
```

FILE	SBT	flaA	pilE	asd	mip	mompS	proA	neuA
/blue/bphl -		6	10	19	28	-	4	3
/blue/bphl	81	2	10	3	28	9	4	9
/blue/bphl	28	3	10	1	3	14	9	1
/blue/bphl	1362	2	10	3	28	9	4	207
/blue/bphl	28	3	10	1	3	14	9	1
/blue/bphl	68	3	13	1	28	14	9	3
/blue/bphl	1362	2	10	3	28	9	4	207
/blue/bphl	1362	2	10	3	28	9	4	207
/blue/bphl -		3	13	1	28	-	9	3
/blue/bphl	93	3	10	1	28	14	9	13
/blue/bphl	93	3	10	1	28	14	9	13
/blue/bphl	9	3	10	1	3	14	9	11
/blue/bphl -		3	4	1	1	-	9	1
/blue/bphl	121	2	10	3	10	9	4	6
/blue/bphl -		3	10	1	1	-	9	11
/blue/bphl	37	3	4	1	1	14	9	11
/blue/bphl -		4	10	11	15	-	1	6
/blue/bphl	37	3	4	1	1	14	9	11
/blue/bphl	1361	3	13	1	1	14	9	207
/blue/bphl	1361	3	13	1	1	14	9	207

## 2. Application Cont.

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/input/  
Name  
..  
logs  
SRR35011513_R1.fastq.gz  
SRR35011513_R2.fastq.gz  
SRR35011514_R1.fastq.gz  
SRR35011514_R2.fastq.gz  
SRR35024812_R1.fastq.gz  
SRR35024812_R2.fastq.gz  
SRR35024814_R1.fastq.gz  
SRR35024814_R2.fastq.gz  
SRR35024816_R1.fastq.gz  
SRR35024816_R2.fastq.gz  
SRR35038347_R1.fastq.gz  
SRR35038347_R2.fastq.gz  
SRR35038348_R1.fastq.gz  
SRR35038348_R2.fastq.gz  
SRR35056275_R1.fastq.gz  
SRR35056275_R2.fastq.gz  
SRR35056276_R1.fastq.gz  
SRR35056276_R2.fastq.gz
```

## 2. Application Cont.

```
GNU nano 5.6.1 elgato.sh
#!/bin/bash
#SBATCH --account=bphl-umbrella
#SBATCH --qos=bphl-umbrella
#SBATCH --job-name=elgato_loop
#SBATCH --ntasks=1
#SBATCH --nodes=1
#SBATCH --cpus-per-task=16
#SBATCH --mem=32gb
#SBATCH --time=06:00:00
#SBATCH --output=elgato_loop.%j.out
#SBATCH --error=elgato_loop.%j.err
#SBATCH --mail-user=nikhil.yengala@flhealth.gov
#SBATCH --mail-type=FAIL,END

INPUT_DIR=/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/input
OUT_DIR=/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun
THREADS=${SLURM_CPUS_PER_TASK:-8}

module load conda || true
conda activate elgato

mkdir -p "$OUT_DIR" logs
ELGATO_REPO=/apps/el_gato # optional for PDFs

shopt -s nullglob
for R1 in "$INPUT_DIR"/*_R1.fastq.gz "$INPUT_DIR"/*_R1*.fastq.gz; do
    [ -e "$R1" ] || continue
    base=$(basename "$R1")
    sample=${base%*_R1.fastq.gz}
    sample=${sample%*_R1*.fastq.gz}
    R2="$INPUT_DIR/${sample}_R2.fastq.gz"
    if [[ ! -f "$R2" ]]; then
        echo "[WARN] Missing R2 for $sample - skipping."
        continue
    fi

    OUT="$OUT_DIR/$sample"
    TMP="$OUT.tmp.$(date +%s).$$"

    rm -rf "$TMP" "$OUT"
    while [[ -e "$OUT" ]]; do sleep 0.5; done

    echo "[INFO] el_gato(tmp): $sample → $TMP"
    el_gato.py \
        --read1 "$R1" \
        --read2 "$R2" \
        --out "$TMP" \
        --threads "$THREADS" \
        --header

    mv "$TMP" "$OUT"

    if [[ -f "$OUT/report.json" && -f "$ELGATO_REPO/elgato_report.py" ]]; then
        python "$ELGATO_REPO/elgato_report.py" \
            --input_jsons "$OUT/report.json" \
            --out_report "$OUT/report.pdf" || true
    fi
done

echo "[DONE] All samples processed → $OUT_DIR"
```

## 2. Application Cont.

SBATCH elgato.sh

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/
Name
└─ ..
└─ Accessory_scripts
└─ conda-recipe
└─ docs
└─ el_gato
└─ input
└─ logs
└─ resultsrun
└─ elgato.sh
└─ elgato_loop.10642747.err
└─ elgato_loop.10642747.out
└─ elgato_report.py
└─ LICENSE
└─ nextflow.config
└─ README.md
└─ run_el_gato.nf
└─ setup.py
```

## 2. Application Cont.

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/
Name
├── ..
├── Accessory_scripts
├── conda-recipe
├── docs
├── el_gato
├── input
├── logs
├── resultsrun
├── elgato.sh
├── elgato_loop.10642747.err
├── elgato_loop.10642747.out
├── elgato_report.py
├── LICENSE
├── nextflow.config
├── README.md
├── run_el_gato.nf
└── setup.py
```

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/
Name
├── ..
├── SRR35011513
├── SRR35011514
├── SRR35024812
├── SRR35024814
├── SRR35024816
├── SRR35038347
├── SRR35038348
├── SRR35056275
└── SRR35056276
```

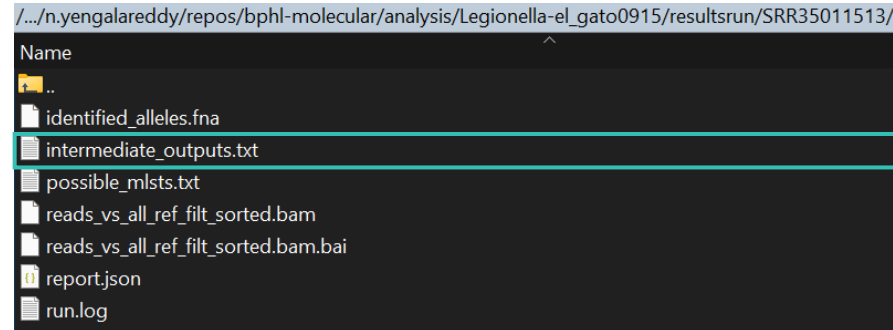
```
/.../n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513/
Name
├── ..
├── identified_alleles.fna
├── intermediate_outputs.txt
├── possible_mlsts.txt
├── reads_vs_all_ref_filt_sorted.bam
├── reads_vs_all_ref_filt_sorted.bam.bai
├── report.json
└── run.log
```

## 2. Application Cont.

```
./.../n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513/  
Name  
└─ ..  
├─ identified_alleles.fna  
├─ intermediate_outputs.txt  
├─ possible_mlsts.txt  
├─ reads_vs_all_ref_filt_sorted.bam  
├─ reads_vs_all_ref_filt_sorted.bam.bai  
├─ report.json  
└─ run.log
```

```
>flaA  
TCAAACCAACCATCAATTCAGTATCGGCAGCACAAAAGCTTCTTCTTGGTGGTATTGCCACGGCAACAGGAACAGAAGTAGCAGGTGCAGCAGCGGCAGATATTACTATCGCAATTGGAGGAGGAGCAGCAACAGTATTAACTCTTCTGCCAATTTACCGGGCACTAAACGGACA  
>pilE  
ATGCTCAGTCATGTACATTTTATGAAGAATAGCCGTATGAACAATCCGCTTTACCCGGTTGAAGTCTGATCAGCATGGTCATTATGGGAATTCGGTTTCAATTGTCCTATCCATCCTATTTACAATATATACAAAAATCCCGTCGTGCTGATGCTCATGCCATTGACACAAGATCAATCATTTTAGAACGCTGTTATTCACAGAAT  
TTTTCTATGCGGCAGCGTGTGGCGCCCTACCAGCATTTCTCTCAACAACGCGGAATGGGTATTATACATCAATATTTCAAACCTGACAGCCACAACGTATACCTTAACCGCAACCCCT  
>asd  
CGCTGTTGGTATTAGTCGTATTAAATGTGGCGACTTACCAGTCTGTTCCGGGACTGGCAAAAAGCCATCAGCGAGCTGGTGTCTCAGGTAGGCGATCTTTAAATGGGAGACCAGTAAATGTTCAAGTTTATCTCAGCAAAATGCTTTTAAATGCGCTTCTCATATTGATCAGTTTGAAGACAATGGTTATACCCGAGAAGAGATGAAGAT  
GGTCTGGGAACCCGCAAGATTATGAAGATGACAGCATATGGTTAACCTACAGCCGTGAGGTTCTGTTATTTATGGGCTTCTGAAGCGGTCATCTGGAGTTAAAAAAGCCTTTGACGGCTGACGATGCTCGCGCGCTTTGGCAAGGCACCGGCTTACTGTAGTGGATAATCTTTCTAAAGCAAGTTATCCACAGCAATTAA  
GAATGCACTCGGCATGATGATGTTTTGTAGGGCGCATGAAGACAGG  
>mip  
TGCCGATTTGGGGAAGAAATTTAAAAATCAAGGCATAGATGTTAATCCGGAAGCAATGGCTAAAGGCATGCAAGACGCTATGAGTGGCGCTCAATGGCTTTAAACGGAACAGCAATGAAAGACGTTCTTAAACAGTTTACGAAAGATTGATGGCTAAGCGTACTGCTGAATTCATAAAGAAAGCGGATGAAATAAAGTAAAGGGGAAGC  
CTTTTAACTGAAAAACAAAAACAGCCAGGCGTTGTTGATTGCCAAGTGGTTGCAATACAAAGTAATCAATCTGGAAATGGTGTAAACCCGAAAAATCGGATACAGTCATGTCGAATATACCTGGTCTGATTGATGGTACCGTTTTTGACAGTACCGAAAAAAGCTGGTAAGCCAGCAACGTT  
>mompS_1_native_locus_39_reads  
ATGACAGTGATCACTGGGCTGATTTTGTAACTGGCACAACACAAACAAGTGGGATGCTGTTAATGCTGAATTAGGTCAATTCTGATGATTTTCAAGCGCTAACAGAAAAATGCGTTTCCACGGCGGTGTTCAATACGCTCGCATGAAGCTGATGTGAACCGTATTTCATAAATCTTGCTTTAACGGGTCAACTCTAAGTCAATGGCT  
TTGGTCTCGCAGTGGTTTAGACATGAACATATGATTTGGCAATGGCTTTGGTGTATTGCTAAAGGCGCTGCTGCTATTCTGGTTGGTACCAGCGATTCTACGATGGAATCAACTTCATTACTGGTTCTAAAAATGC  
>mompS_2_non-native_locus  
ATGACAGTGATCACTGGGCTGATTTTGTAACTGGCACAACACAAACAAGTGGGATGCTGTTAATGCTGAATTAGGTCAATTCTGATGATTTTCAAGCGCTAACAGAAAAATGCGTTTCCACGGCGGTGTTCAATACGCTCGCATGAAGCTGATGTGAACCGTATTTCATAAATCTTGCTTTAACGGGTCAACTCCAAGTCAACGGCT  
TTGGTCTCGCAGTGGTTTAGACATGAACATATGATTTGGCAATGGCTTTGGTATTATGCTAAAGGAGCTGCTGCTATTCTGGTTGGTACCAGCGATTCTATGATGGAATAGGTTTCTGTTACTGGTTCTAAAAATGC  
>prpA  
TATGAAACCGCTACTGGACGGCAAAACAAATGACCTTTGGTATGGTGACACCATGATGATCTTTGGTTCTCTAGGTGTAGTGCGCATGGAATCAAGTCAAGGCTTCAACGAACAACATTCTGGACTTGAATATTTCCGTGAGTCTGGTGGCATGAATGAGTCGTTTTAGATATGGCTGCCCAAGCTGCTGAATATTATTCGGTTGGA  
AAAAACAGCTGGCAATCGGCCCGGAAATATGAAAGAAGACAGTGGTTATGACGCATTGCGTTACATGGATAAACAAGCGTGAATGGTATGCTCTATGATGATGATGATGATTTATGTTGGTGGTTAGATGTCATTACTCCAGTGGAGTGATACCACTATTATCTACATATTAGCCAATCAACCTAAT  
>neuA_neuH  
GATACCTTCGATGTGATTATCATCTGTTATGACCTTCTGTTAAGTTTCAGCAAAATGGAGGTTTTTTTGAACAGTGTATTGCTTTTACAACCAACTCTTCATTTCAGGAAGCCAGAAACCATAGACATGCTGTTGAAATACATCAAGCAACGGGAAAAAGTGTGTTTCAGTTAGTCCCATCTCTTTAAAGCCTTCTTGGTGTAGAAGCAATT  
GATAGCCAAGGCAATTGGTTAAGCCAGAGCTTTTTCAGGATCTGGAATTTATTGCAACAGAGAATCCAATTTATAAATTAATGGGTCAATTTATATTGCAACAGCTAAACAATATTCGAAATTAAGAGTTTTATAGC
```

## 2. Application Cont.



Output:

Average insertion size: 401.838 Average read length: 146.713

WARNING: The average read length is 146. El\_gato may struggle to resolve multiple alleles such as for momp5 with shorter read lengths. Assessing coverage of MLST loci by provided sequencing reads.

Command:

```
samtools coverage -r flxA:351-532 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | cut -f 1,4,5,6,7,8,9; samtools coverage -r asd:351-532 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r mip:350-751 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r mmp:367-718 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r proA:388-724 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r neuA:390-703 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r neuH:350-703 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r neuA:207-350-700 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r neuA:211:350-703 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r neuA:212:350-700 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9; samtools coverage -r neuA:215:350-703 /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/reads_vs_all_ref_filt_sorted.bam | tail -1 | cut -f 1,4,5,6,7,8,9
```

Output:

#name	numreads	covbases	coverage	meandepth	meanbases	meanmap
flxA	532	352	100	237.566	38.8	53.7
pilE	739	333	100	222.135	31	52.7
asd	1062	473	100	246.442	38.8	56.1
mip	919	402	100	233.552	38.8	58
momp5	1584	352	100	446.386	38.7	55.8
proA	866	400	100	213.931	38.9	51.6
neuA	786	354	100	210.438	38.6	56
neuH	0	0	0	0	0	0
neuA_207	0	0	0	0	0	0
neuA_211	0	0	0	0	0	0
neuA_212	0	0	0	0	0	0
neuA_215	0	0	0	0	0	0

Minimum depth of flxA locus is 186.  
Minimum depth of pilE locus is 145.  
Minimum depth of asd locus is 156.  
Minimum depth of mip locus is 126.  
Minimum depth of momp5 locus is 310.  
Minimum depth of proA locus is 128.  
Minimum depth of neuA locus is 125.

Best match of each identified sequence determined using BLASTN

Command:

```
blastn -query /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513.tmp.175587962.3701908/identified_alleles.fna -db /blue/bphl-florida/n.yengalareddy/conda/envs/elgato/bin/db/all_loci.fasta -outfmt '% s% id s% len' -max_target_seqs 50000 | sort -k4,3 -k12,12g | sort -m -rg -w -k3,1
```

Output:

sseqid	slen	pidcent	length	mismatch	gapopen	qstart	qend	sstart	send	eval	bitscore	qlen	slen
flxA	182	100.000	333	0	0	1	333	1	333	1.07e-04	137	182	182
pilE	100.000	333	0	0	1	333	1	333	2.34e-178	616	333	333	
asd	473	100.000	473	0	0	1	473	1	473	0.0	874	473	473
mip	100.000	402	0	0	1	402	1	402	0.0	743	402	402	
momp5_native_locus_39_reads	100.000	352	0	0	1	352	1	352	0.0	651	352	352	
momp5_non-native_locus	100.000	350	7	0	3	352	3	352	4.16e-176	608	352	352	
proA	100.000	405	0	0	1	405	1	405	0.0	749	405	405	
neuA	100.000	354	0	0	1	354	1	354	0.0	654	354	354	



## 2. Application Cont.

```
./.../n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/resultsrun/SRR35011513/  
Name  
..  
identified_alleles.fna  
intermediate_outputs.txt  
possible_mlsts.txt  
reads_vs_all_ref_filt_sorted.bam  
reads_vs_all_ref_filt_sorted.bam.bai  
report.json  
run.log
```

Sample	ST	flaA	pilE	asd	mip	mompS	proA	neuA_neuAH	mompS_reads_support	mompS_reads_against
SRR35011513	1534	3	10	1	1	1	9	11	39	0
SRR35011513	Novel ST*	3	10	1	1	NAT	9	11	0	4

# 3. Application Cont.

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/input/  
Name  
..  
SRR35202208_R1.fastq.gz  
SRR35202208_R2.fastq.gz  
SRR35202209_R1.fastq.gz  
SRR35202209_R2.fastq.gz  
SRR35202210_R1.fastq.gz  
SRR35202210_R2.fastq.gz  
SRR35202211_R1.fastq.gz  
SRR35202211_R2.fastq.gz  
SRR35202212_R1.fastq.gz  
SRR35202212_R2.fastq.gz  
SRR35202213_R1.fastq.gz  
SRR35202213_R2.fastq.gz  
SRR35202214_R1.fastq.gz  
SRR35202214_R2.fastq.gz  
SRR35202215_R1.fastq.gz  
SRR35202215_R2.fastq.gz  
SRR35202216_R1.fastq.gz  
SRR35202216_R2.fastq.gz  
SRR35202217_R1.fastq.gz  
SRR35202217_R2.fastq.gz  
SRR35202218_R1.fastq.gz  
SRR35202218_R2.fastq.gz  
SRR35202219_R1.fastq.gz  
SRR35202219_R2.fastq.gz
```

# 3. Application Cont.

```
GNU nano 5.6.1                                legiovue.sh
#!/bin/bash
#SBATCH --account=bphl-umbrella
#SBATCH --qos=bphl-umbrella
#SBATCH --job-name=legiovue
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --mem=264G
#SBATCH --time=24:00:00
#SBATCH --output=legiovue.%j.out
#SBATCH --error=legiovue.%j.err
#SBATCH --mail-user=nikhil.yengala@flhealth.gov
#SBATCH --mail-type=FAIL,END

# Modules / env
module load conda
source /apps/conda/25.3.1/etc/profile.d/conda.sh
conda activate Nextflow

# Apptainer only
module load apptainer

# Paths you control
FASTQ_DIR="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-el_gato0915/input"
KRACKEN_DB="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/kraken2_standard_8"
OUTDIR="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/results"
WORKDIR="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/work"

# Cache
export NXF_SINGULARITY_CACHEDIR="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915"
export APPTAINER_CACHEDIR="$NXF_SINGULARITY_CACHEDIR"
mkdir -p "$NXF_SINGULARITY_CACHEDIR" "$OUTDIR" "$WORKDIR"

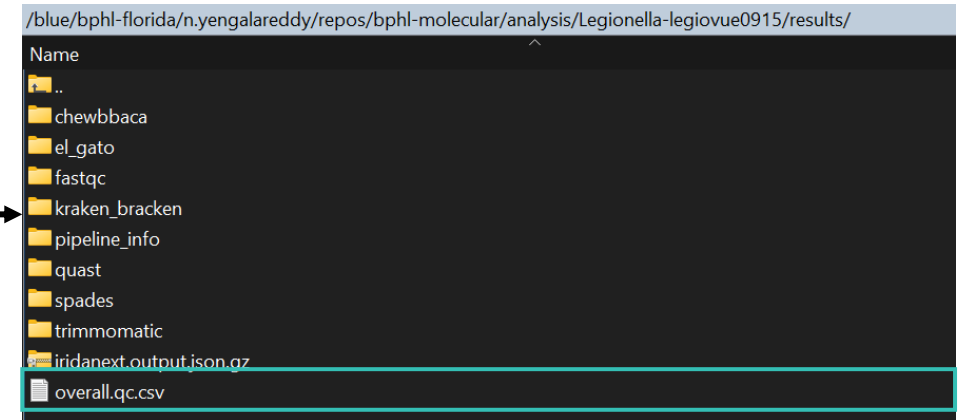
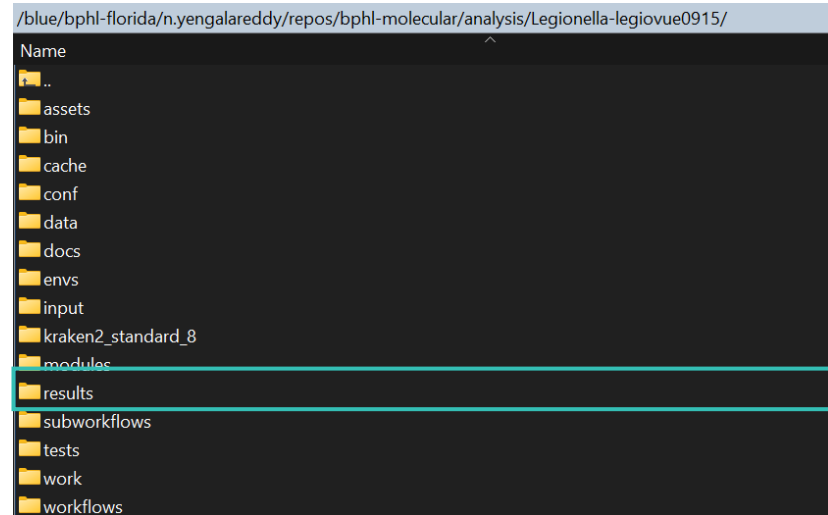
# Run LegioVue
nextflow run phac-nml/legiovue \
  -profile singularity \
  --fastq_dir "$FASTQ_DIR" \
  --kraken2_db "$KRACKEN_DB" \
  --outdir "$OUTDIR" \
  --work-dir "$WORKDIR" \
  --max_cpus "${SLURM_CPUS_PER_TASK:-16} \
  --max_memory 264.GB
```

# 3. Application Cont.

sbatch legiovue.sh

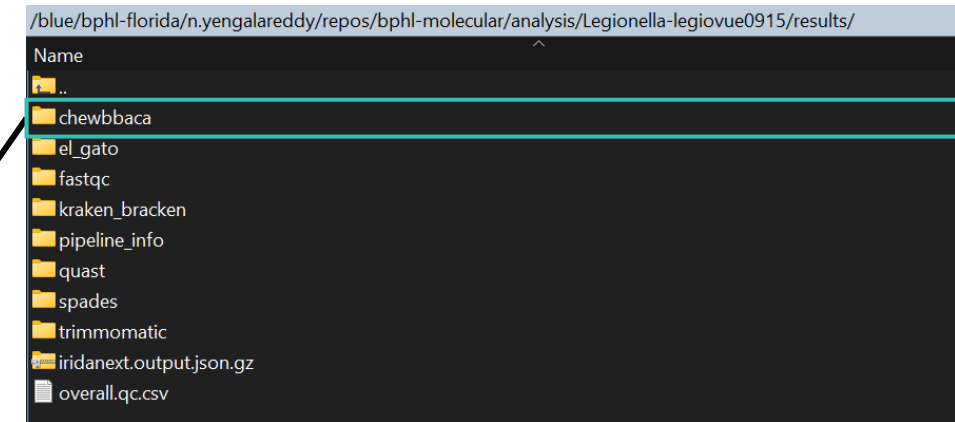
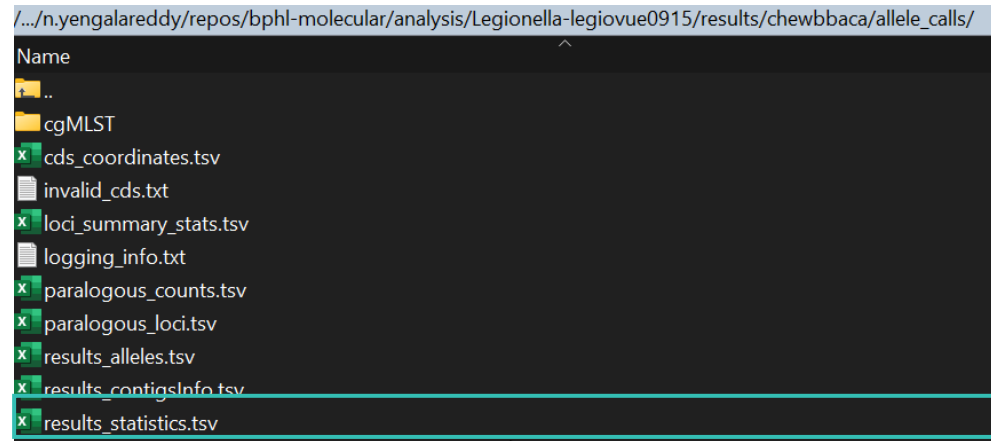
```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/  
Name  
└─ ..  
└─ assets  
└─ bin  
└─ cache  
└─ conf  
└─ data  
└─ docs  
└─ envs  
└─ input  
└─ kraken2_standard_8  
└─ modules  
└─ results  
└─ subworkflows  
└─ tests  
└─ work  
└─ workflows
```

# 3. Application Cont.



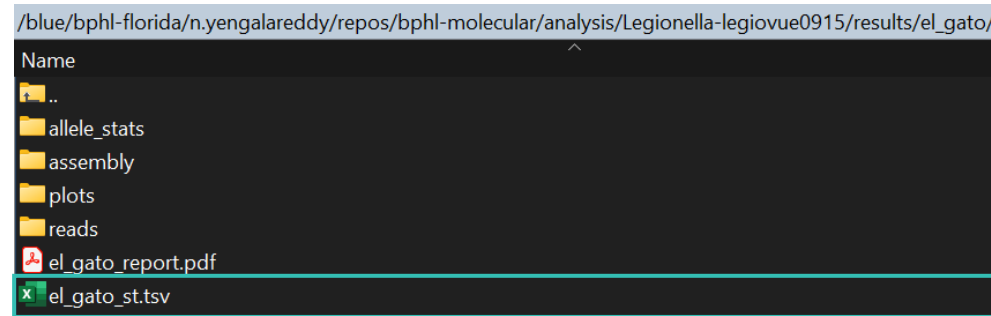
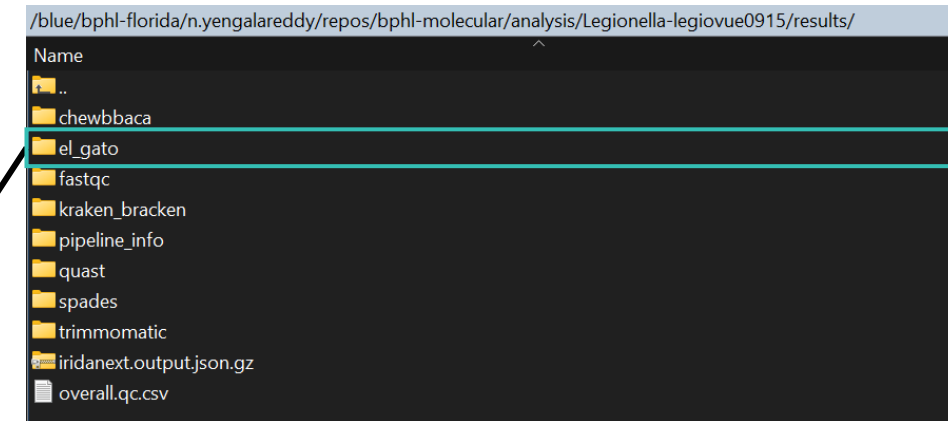
sample	lfn_abundance	num_paired_trimmed_reads	pct_paired_reads_passing_qc	n50	num_contigs	pct_gc	assembly_len	largest_contig	st	st_approach	chewbbaca_exc	chewbbaca_inf	chewbbaca_pct_exc	final_qc_score	qc_status	qc_message
SRR35056275	100	395181	82.66	152529	52	38.24	3407337	430849	731	reads	1481	1	97.37	5.52	PASS	
SRR35024812	100	122629	64.2	154930	56	38.23	3397052	381621	731	assembly	1118	360	73.5	5.54	WARN	low_read_count;low_exact_allele_calls
SRR35038348	88.73	1227389	47.11	140080	73	38.2	3717698	427957	2256	reads	1044	403	68.64	4.43	WARN	low_exact_allele_calls
SRR35011513	100	1402083	43.77	436412	26	38.39	3541164	1156536	1534	reads	1081	403	71.07	6	WARN	low_exact_allele_calls
SRR35056276	100	454800	82.67	192859	41	38.24	3396639	584080	731	reads	1482	0	97.44	5.81	PASS	
SRR35011514	99.99	1036447	47.13	237429	26	38.19	3431729	560894	1400	reads	1054	432	69.3	5	WARN	low_exact_allele_calls
SRR35038347	99.78	1279480	46.23	237717	26	38.22	3510744	560129	1400	reads	1484	1	97.57	5	PASS	
SRR35024814	100	134737	64.06	186785	46	38.23	3395970	430849	731	assembly	1480	1	97.3	5.76	WARN	low_read_count
SRR35024816	100	187292	69.54	187246	53	38.24	3395987	430849	731	assembly	1482	0	97.44	5.77	PASS	

# 3. Application Cont.



FILE	EXC	INF	PLOT3	PLOT5	LOTSC	NIPH	NIPHEM	ALM	ASM	PAMA	LNF	Invalid CDSs	Classified_CDSs	Total_CDSs
SRR35011513	1081	403	0	1	0	13	0	10	4	0	9	0	1531	3174
SRR35011514	1054	432	0	0	0	8	2	11	2	0	12	0	1520	3058
SRR35024812	1118	360	0	2	0	9	1	12	5	0	14	0	1517	3007
SRR35024814	1480	1	0	0	0	0	10	12	5	0	13	0	1518	3008
SRR35024816	1482	0	0	0	0	1	9	12	5	0	12	0	1519	3010
SRR35038347	1484	1	0	0	0	1	9	11	3	0	12	0	1520	3140
SRR35038348	1044	403	0	1	0	9	2	10	5	0	47	0	1486	3367
SRR35056275	1481	1	0	0	0	2	9	11	5	0	12	1	1520	3008
SRR35056276	1482	0	0	0	0	0	10	12	5	0	12	0	1519	3011

### 3. Application Cont.



Sample	ST	flaA	pilE	asd	mip	mompS	proA	neuA_neuAH	approach
SRR35011513	1534	3	10	1	1	1	9	11	reads
SRR35011514	1400	11	14	16	1	15	13	207	reads
SRR35024812	731	7	10	17	12	29	11	9	assembly
SRR35024814	731	7	10	17	12	29	11	9	assembly
SRR35024816	731	7	10	17	12	29	11	9	assembly
SRR35038347	1400	11	14	16	1	15	13	207	reads
SRR35038348	2256	16	21	33	37	41	1	222	reads
SRR35056275	731	7	10	17	12	29	11	9	reads
SRR35056276	731	7	10	17	12	29	11	9	reads

# 3. Application Cont.

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/results/

Name

- ..
- chewbbaca
- el\_gato
- fastqc
- kraken\_bracken
- pipeline info
- quast
- spades
- trimmomatic
- iridanext.output.json.gz
- overall.qc.csv

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/results/quast/

Name

- ..
- aligned\_stats
- basic\_stats
- genome\_stats
- report.html
- report.pdf
- scored\_quast\_report.csv
- transposed\_report.tsv

Report

	SRH35011513 contigs	SRH35011514 contigs	SRH35024812 contigs	SRH35024814 contigs	SRH35024816 contigs	SRH35038347 contigs	SRH35038348 contigs	SRH35056275 contigs	SRH35056276 contigs
# contigs (>= 0 bp)	61	73	136	117	115	95	141	134	108
# contigs (>= 1000 bp)	21	24	49	42	42	26	67	44	36
# contigs (>= 5000 bp)	19	19	38	32	32	19	52	33	27
# contigs (>= 10000 bp)	15	18	34	30	30	19	41	32	26
# contigs (>= 25000 bp)	14	17	29	27	26	19	52	27	23
# contigs (>= 50000 bp)	12	14	21	19	19	15	18	22	17
Total length (>= 0 bp)	3541164	3431729	3397052	3395970	3395987	3510744	3717688	3407337	3396639
Total length (>= 1000 bp)	3532382	3424405	3379588	3381187	3378485	3501351	3703205	3389951	3383408
Total length (>= 5000 bp)	3529662	3414331	3354148	3361271	3355493	3483913	3667020	3370220	3365522
Total length (>= 10000 bp)	3505282	3409208	3327628	3347423	3342080	3483913	3598039	3364538	3359840
Total length (>= 25000 bp)	3484976	3384950	3252887	3300897	3282322	3483913	3421851	3388740	3315881
Total length (>= 50000 bp)	3388102	3237274	2952667	2992848	3024263	3399788	3093253	3102136	3084520
# contigs	26	26	56	46	53	26	73	52	41
Largest contig	1156536	560894	381621	430849	560129	427957	430849	584080	584080
Total length	3535727	3425598	3384697	3383608	3386699	3501351	3707774	3395094	3386567
Reference length	3504074	3504074	3504074	3504074	3504074	3504074	3504074	3504074	3504074
GC (%)	38.39	38.39	38.23	38.23	38.24	38.22	38.20	38.24	38.24
Reference GC (%)	38.39	38.39	38.39	38.39	38.39	38.39	38.39	38.39	38.39
NG50	436412	237429	154930	186785	187246	237717	140880	152529	192898
NG50	436412	237429	154930	178142	187246	237717	157797	149211	192898
NG0	79195	89425	45429	47711	80004	80004	32936	58640	60165
NG0	79195	80004	33822	45424	44554	80004	36340	45429	45429
aUK	572684.2	301212.3	183456.8	207411.9	187904.6	292827.6	204661.9	184732.3	266937.8
aUG	577857.4	294466.5	177206.5	200293.2	191275.5	292600.0	216559.4	176987.0	257966.9
L50	3	5	7	7	7	5	7	8	5
L50	3	5	7	7	7	5	6	8	5
L80	9	12	23	21	20	13	30	22	17
L80	9	13	26	23	22	13	25	24	19
# misassemblies	21	3	55	55	55	5	64	56	57
# misassembled contigs	0	2	23	21	21	0	20	23	20
Misassembled contigs length	2918647	358847	2711139	2857588	2889567	583958	2148456	2777312	3099133
# local misassemblies	6	2	34	36	34	3	39	34	35
# scaffold gap ext. mis.	0	0	0	0	0	0	0	0	0
# scaffold gap loc. mis.	0	0	0	0	0	0	0	0	0
# unaligned mis. contigs	0	9	1	1	1	8	9	1	1
# unaligned contigs	1 + 8 part	5 + 17 part	6 + 26 part	6 + 26 part	9 + 24 part	1 + 20 part	14 + 42 part	7 + 23 part	4 + 21 part
Unaligned length	243371	2747861	409887	401923	408460	3024996	1178478	402401	402051
Genome fraction (%)	93.896	19.338	84.789	85.028	85.027	19.291	72.095	85.042	85.074
Duplication rate	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.003	1.000
# N's per 100 kbp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
# mismatches per 100 kbp	418.20	2749.62	2182.30	2190.79	2189.70	2745.81	2924.09	2201.00	2194.29
# indels per 100 kbp	31.54	74.22	61.07	61.71	61.44	72.60	79.00	62.08	62.00
Largest alignment	1068771	220992	146099	146099	146099	220992	93874	146713	146109
Total aligned length	2393615	677129	2971812	2979928	2980561	2537188	2989609	2982150	2982150
NA50	215780	59686	60921	62261	67657	20889	62891	63827	63827
NA50	215780	-	54414	54414	59686	-	21888	62891	62891
NA90	18428	-	-	-	-	-	-	-	-
NA90	31154	-	-	-	-	-	-	-	-
aRNA	420910.8	21120.3	64731.7	67961.0	68267.6	20390.0	28164.3	68453.8	70566.2
aUGA	424713.0	20647.3	62505.5	65605.4	65801.9	20377.1	29801.5	66301.9	66109.9
L40	4	4	17	17	17	4	42	17	15
L40	4	-	19	18	18	-	38	17	17
L60	23	-	-	-	-	-	-	-	-
L60	21	-	-	-	-	-	-	-	-

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



# 3. Application Cont.

```
./bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/results/trimmomatic/
Name
..
SRR35011513.summary.txt
SRR35011513_paired_R1.fastq.gz
SRR35011513_paired_R2.fastq.gz
SRR35011513_unpaired_R1.fastq.gz
SRR35011513_unpaired_R2.fastq.gz
SRR35011514.summary.txt
SRR35011514_paired_R1.fastq.gz
SRR35011514_paired_R2.fastq.gz
SRR35011514_unpaired_R1.fastq.gz
SRR35011514_unpaired_R2.fastq.gz
SRR35024812.summary.txt
SRR35024812_paired_R1.fastq.gz
SRR35024812_paired_R2.fastq.gz
SRR35024812_unpaired_R1.fastq.gz
SRR35024812_unpaired_R2.fastq.gz
SRR35024814.summary.txt
SRR35024814_paired_R1.fastq.gz
SRR35024814_paired_R2.fastq.gz
SRR35024814_unpaired_R1.fastq.gz
SRR35024814_unpaired_R2.fastq.gz
SRR35024816.summary.txt
SRR35024816_paired_R1.fastq.gz
SRR35024816_paired_R2.fastq.gz
SRR35024816_unpaired_R1.fastq.gz
SRR35024816_unpaired_R2.fastq.gz
SRR35038347.summary.txt
SRR35038347_paired_R1.fastq.gz
SRR35038347_paired_R2.fastq.gz
SRR35038347_unpaired_R1.fastq.gz
```

```
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Legionella-legiovue0915/results/
Name
..
chewbbaca
el_gato
fastqc
kraken_bracken
pipeline_info
quast
spades
trimmomatic
iridanext.output.json.gz
overall.qc.csv
```

Input Read Pairs: 3203333  
Both Surviving Reads: 1402083  
Both Surviving Read Percent: 43.77  
Forward Only Surviving Reads: 1130211  
Forward Only Surviving Read Percent: 35.28  
Reverse Only Surviving Reads: 81548  
Reverse Only Surviving Read Percent: 2.55  
Dropped Reads: 589491  
Dropped Read Percent: 18.40

# Conclusion

- ✓ Fundamentals
- ✓ Installation and Setup in HPG
- ✓ Successfully Executed Job Queries
- ✓ Generated Output Files



# Advanced Molecular Detection

## Southeast Region Bioinformatics

# Questions?

[bphl-sebioinformatics@flhealth.gov](mailto:bphl-sebioinformatics@flhealth.gov)

**Molly Mitchell, PhD**

Bioinformatics Supervisor

[Molly.Mitchell@flhealth.gov](mailto:Molly.Mitchell@flhealth.gov)

**Nikhil Reddy, MS**

Bioinformatician

[Nikhil.Yengala@flhealth.gov](mailto:Nikhil.Yengala@flhealth.gov)

**Sam Bernhoft, MPH**

Bioinformatician

[Samantha.bernhoft@flhealth.gov](mailto:Samantha.bernhoft@flhealth.gov)