|  |
| --- |
| ***Insert Laboratory Specific Name Here*** |
| **De novo Genome Assembly Training Form** |

|  |  |
| --- | --- |
| **Employee Name** | **Training Start Date** |
|  |  |

**Section I – *TELL* - Base Knowledge (Video and Reading Requirements)** *[select videos and documents relevant to your lab processes; add other videos and documents as appropriate; see examples of tool-specific line items in blue below]*

|  |  |  |
| --- | --- | --- |
| **Document Name** | **Trainee Initials** | **Date Watched** |
| [***De novo* genome assembly: what every biologist should know**](https://www.nature.com/articles/nmeth.1935) |  |  |
| [**Quast : Genome assembly evaluation tool**](https://github.com/ablab/quast/blob/master/README.md) |  |  |
| [***Running SPAdes***](https://github.com/ablab/spades/blob/spades_3.14.0/README.md#sec3) |  |  |
| [***ABySS : Assembling a paired-end library***](https://github.com/bcgsc/abyss/blob/master/README.md#assembling-a-paired-end-library) |  |  |
| [***PacBio assembly with SMRT portal***](http://sepsis-omics.github.io/tutorials/modules/pacbio/) |  |  |
| [***Assembly using Velvet***](https://angus.readthedocs.io/en/2016/week3/LN_assembly.html) |  |  |
| [***Shovill***](https://github.com/tseemann/shovill/blob/master/README.md) |  |  |
| [***Unicycler Assembly***](https://galaxyproject.github.io/training-material/topics/assembly/tutorials/unicycler-assembly/tutorial.html) |  |  |
| [***Geneious : De Novo Assembly***](https://www.geneious.com/tutorials/de-novo-assembly/) |  |  |
| [***Canu***](https://canu.readthedocs.io/en/latest/quick-start.html) |  |  |

**Section II – *SHOW* - Observation: Trainee observes the trainer perform all steps in the De novo Genome Assembly SOP**

|  |  |  |
| --- | --- | --- |
| **Discussion Points** | **Trainer Initials** | **Date** |
| What metrics are important to consider when evaluating assembly quality (using a tool like QUAST)? |  |  |
| What is the purpose of performing a de novo assembly within the context of the overall workflow?  |  |  |
| What is the significance of coverage depth? |  |  |
| What does N50 describe? |  |  |
| What is the difference between a FASTQ and a FASTA file? |  |  |
| What is the difference between de novo assembly and reference-based assembly? |  |  |
| *What file generated by SPAdes would contain important run information such as commands performed and errors?* |  |  |
| *Add additional questions the trainer should ask the trainee to determine level of understanding specific to your protocol.* |  |  |

**Section III – *DO* - Performance under Supervision: Trainee performs all steps in the sequencing SOP under direct trainer supervision**

Controls and/or sample(s) will be provided to the trainee. The trainee will:

*(Find example steps below, insert steps specific to your lab)*

1. *Acquire/select sample reads for de novo genome assembly.*
2. *Perform de novo genome assembly with tool of your choosing (e.g., SPAdes, ABySS, HGAP/Celera, Velvet)*
3. *Properly save results in appropriate directories*

Successful performance criteria: All samples are processed into assembly files within required parameters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Performance Assessment** | **Yes** | **No** | **Trainer Initials** | **Date** |
| *Acquired and selected appropriate sample file(s) for de novo genome assembly* | o | o |  |  |
| *Perform de novo genome assembly with tool of your choosing (e.g. SPAdes, ABySS, HGAP/Celera, Velvet)* | o | o |
| *Properly save results in appropriate directories* | o | o |
| **Comments:** |

**Section IV – *APPLY -*Independent Performance: Trainee individually executes all steps in the sequencing SOP**

Previously run sample(s) will be provided to the trainee. The trainee will:

*(Find example steps below, insert steps specific to your lab)*

1. *Acquire/select sample reads for de novo genome assembly.*
2. *Perform de novo genome assembly with tool of your choosing (e.g., SPAdes, ABySS, HGAP/Celera, Velvet)*
3. *Properly save results in appropriate directories*

Successful performance criteria: All samples are processed into assembly files within required parameters.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Performance Assessment** | **Yes** | **No** | **Trainer Initials** | **Date** |
| *Acquired and selected appropriate sample file(s) for de novo genome assembly* | o | o |  |  |
| *Perform de novo genome assembly with tool of your choosing (e.g. SPAdes, ABySS, HGAP/Celera, Velvet)* | o | o |
| *Properly save results in appropriate directories* | o | o |
| **Comments:** |

**Section V – Employee Attestation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attestations** | **Yes** | **No** | **Trainee Initials** |
| I read and understand the procedures listed in the required reading. | **o** | **o** |  |
| I had an opportunity to discuss my questions with the trainer. | **o** | **o** |  |
| I am satisfied with the explanations provided to me; all my questions were answered. | **o** | **o** |  |

**Section VI – Review and Signatures**

|  |  |  |
| --- | --- | --- |
| **Trainee Name** | **Signature** | **Date** |
|  |  |  |
| **Trainer Name** | **Signature** | **Date** |
|  |  |  |
| **CLIA Technical or General Supervisor (as applicable)** | **Signature** | **Date** |
|  |  |  |

**Appendix A – Trainer Discussion Topic Answer Sheet**

**What metrics are important to consider when evaluating assembly quality (using a tool like QUAST)?**

N50 (defined as the minimum contig length needed to cover 50% of the genome) is often considered when considering assembly quality (higher is better), however there are a number of other metrics that should be reviewed based on what you want to do with the assembly. Others include L50, N90, NG50, etc. Many of these metrics may be interdependent. NG50 (created by authors of the Assemblathon competition) for example, is the same as N50 except that it is 50% of the known or estimated genome size that must be of the NG50 length or longer.

**What is the purpose of performing a de novo assembly within the context of the overall workflow?**

*\*The answer to this question will depend on your lab’s workflow, and should account for downstream analyses that will utilize the assembly and overall focus of the pipeline or workflow\**

**What is the significance of coverage depth?**

Depth of coverage is a measure indicating the number of times a nucleotide is read during a sequencing run. Having a greater depth of coverage can increase the confidence in the final results.

**What does N50 describe?**

The minimum contig length needed to cover 50% of the genome.

**What is the difference between a FASTQ and a FASTA file?**

Both formats are used for storing sequence data however, FASTQ files were created to allow for encoding of Phred quality scores (which estimate the probability of having correctly identified a given nucleotide) and are mostly used to store short-read sequence data. FASTA files are mostly used to store reference quality sequence and by contrast, it was not designed with a standardized way to encode these types of quality scores.

***\*Insert answers to additional discussion topics here (including tool-specific questions)\****

***e.g. What file generated by SPAdes would contain important run information such as commands performed and errors?***

*SPAdes log file (spades.log)*