LECTURE 2

Piecing Together the Puzzle: Bacterial Assembly and Annotation

Daniel Gyamfi Amoako, **PhD**
University of Guelph, Canada (damoako@uoguelph.ca)
University of KwaZulu-Natal, South Africa (amoakod@ukzn.ac.za)
Amoako Analytics Hub *

* "Empowering One-Health in Africa through Bioinformatics: Free Machine Learning and AI Analytics Data Solutions”

Introduction to bacterial assembly and annotation

- Assembly and annotation are essential steps in analyzing genomic data
- **Assembly r**efers to the process of piecing together short reads of DNA sequence into longer, contiguous sequences (called contigs) that represent the genome of the organism.
- **Annotation** refers to the process of identifying and labeling specific genomic features (such as genes, regulatory elements, and repeat sequences) within the assembled genome.
- The goal of assembly and annotation is to create a complete and accurate representation of the bacterial genome, which can be used to understand the organism's biology and potential interactions with other organisms or environments.
- Annotation can be performed using a variety of tools and databases, including gene prediction algorithms, functional annotation databases, and comparative genomics tools.

Understanding de-novo and reference-based assembly methods

- **De-novo assembly** is a method of piecing together short reads into longer sequences without a reference genome.
- **Reference-based assembly** is a method of aligning short reads to a known reference genome to identify variations and generate a new assembly.
- **De-novo assembly** is more suitable for **non-model organisms** and can identify novel sequences, while **reference-based assembly** is more suitable for **closely related organisms** and can identify variations from a known reference genome.
- De-novo assembly can be **computationally intensive** and require large amounts of computing power, while reference-based assembly can be less computationally intensive and require less computing power.
- Both de-novo and reference-based assembly have their **strengths and weaknesses**, and the choice of method depends on the **research question**, the **available resources**, and the **characteristics of the organism** being studied.

Overview of some assembly tools

- **SPAdes:** A popular **de novo assembler** that can handle a wide range of sequencing data, including Illumina, PacBio, and Oxford Nanopore.
- **ABySS:** Another popular de novo assembler that is particularly effective for large and complex genomes.
- **IDBA-UD:** A de novo assembler that uses iterative de Bruijn graph construction to handle uneven sequencing depth and errors.
- Other popular de novo assemblers include Velvet, SOAPdenovo, and Canu.
- **Reference-based assemblers**, such as BWA-MEM and Bowtie, use a reference genome to align and assemble sequencing reads.
- **Hybrid assemblers,** such as SPAdes hybrid and DBG2OLC, combine de novo and reference-based approaches to improve assembly accuracy and completeness.
- Choosing the right assembly tool depends on factors such as the type and quality of sequencing data, the size and complexity of the genome, and the research question being addressed.

Quality checks for bacterial assemblies -1

- Quality checks for bacterial assemblies involve evaluating a variety of metrics, including contiguity, completeness, accuracy, and consistency.
- **Contiguity** refers to how well the assembled contigs/scaffolds represent the original genome, and can be assessed using metrics like N50, L50, and genome fraction.
- **Completeness** refers to the percentage of the original genome that is present in the assembly, and can be evaluated using tools like BUSCO and CheckM.
- **Accuracy** refers to the correctness of the assembly, and can be assessed by comparing the assembled genome to a reference genome or to other sequencing data.
- **Consistency** refers to the concordance between the assembly and other data, such as transcriptomics or proteomics data.

Quality checks for bacterial assemblies -2

- Some popular tools for assessing the quality of bacterial assemblies include QUAST, REAPR, and MUMmer.

- Quality checks should be performed at multiple stages of the assembly process, including pre-assembly, during assembly, and post-assembly. This allows for early detection and correction of errors, and can help ensure that the final assembly is as accurate and reliable as possible.

- Quality checks can also help guide decisions about which assembly parameters to use and how to optimize the assembly process. By identifying areas of weakness or improvement, quality checks can help researchers and bioinformaticians improve the quality and accuracy of their bacterial genome assemblies.

Contig ordering and scaffolding

- Contig ordering and scaffolding are crucial steps in the assembly process, as they help to create a more complete and accurate representation of the genome.
- The process of contig ordering involves arranging contigs in the correct order and orientation based on their relationships to one another. This can be done using various methods such as mate-pair sequencing or optical mapping.
- Scaffolding involves filling in gaps between contigs to create larger sequences called scaffolds. This can be done using paired-end sequencing, long-read sequencing, or other methods.
- Various software tools are available for contig ordering and scaffolding, including SSPACE, BESST, and LRScaf.
- The quality of the resulting scaffolds can be assessed using various metrics, such as N50 and genome completeness, to ensure that they are accurate and representative of the underlying genome.

Genome size estimation

- Genome size estimation is an important step in the bacterial assembly process that allows researchers to determine the size and complexity of a bacterial genome.
- Accurate genome size estimation can help in selecting appropriate sequencing strategies, evaluating the quality of genome assemblies, and predicting gene content and functional potential.
- Several methods are available for genome size estimation, including k-mer-based approaches, read coverage analysis, and mapping-based approaches.
- K-mer-based approaches involve counting the frequency of unique k-mers in a sequencing dataset and using this information to estimate genome size.
- Read coverage analysis and mapping-based approaches involve mapping reads back to an assembly and calculating coverage depth to estimate genome size.

Gap filling in bacterial assembly

- Gaps in a bacterial genome assembly refer to regions where the sequence is missing or ambiguous.
- Gap filling is the process of resolving these regions to generate a more complete genome assembly.
- The most common approach for gap filling is to use additional sequencing data (e.g., PacBio, Nanopore) that can span across the gaps and provide longer reads.
- Other gap filling methods include PCR-based approaches or using information from related genomes to infer the missing sequence.
- After gap filling, it is important to perform quality checks to ensure that the assembly is accurate and complete, and to update the annotation accordingly.

Overview of some annotation tools

- **Prokka:** a rapid and simple annotation pipeline for prokaryotic genomes that uses various tools for gene prediction, functional annotation, and comparative genomics.
- **RAST:** a web-based annotation server for bacterial and archaeal genomes that uses various tools for gene calling, functional annotation, metabolic pathway analysis, and comparative genomics.
- **NCBI PGAP:** the NCBI prokaryotic genome annotation pipeline that provides a comprehensive annotation of prokaryotic genomes, including identification of genes, functional annotation, and comparative analysis.
- **BASys:** a web server that provides automated annotation of bacterial genomes, including identification of genes, functional annotation, and comparative analysis.
- **IMG-ER:** a web-based system for microbial genome annotation that includes various tools for gene prediction, functional annotation, comparative analysis, and data integration.
- **BV-BRC (formally called PATRIC):** a web-based annotation server for bacterial and virus genomes that uses various tools for gene calling, functional annotation, metabolic pathway analysis, and comparative genomics.

Benefits and limitations of automated annotation

- Benefits of automated annotation tools include speed, efficiency, and consistency in annotation, as well as the ability to annotate large datasets quickly.
- Automated annotation tools use existing databases and algorithms to predict genes, functions, and pathways, and can provide insights into potential biological processes and pathways.
- Automated annotation can be particularly useful for initial annotation of a genome, providing a good starting point for further manual curation and refinement.
- Limitations of automated annotation tools include potential errors or inaccuracies in gene prediction and functional annotation, as well as the lack of context-specific information.
- The use of automated annotation should always be complemented with manual curation, validation, and refinement to ensure accuracy and reliability of the results, especially when using the annotations for downstream analysis or interpretation.

Quality checks for bacterial annotations

- Quality checks are essential for ensuring the accuracy and completeness of bacterial annotations, as small errors or omissions can have a significant impact on downstream analyses and interpretations.
- Common quality checks for bacterial annotations include assessing completeness, accuracy, consistency, and comprehensiveness of the annotation, as well as checking for potential contamination or misannotation.
- Quality checks can be conducted using various tools and metrics, such as BUSCO, CheckM, and annotation consistency scores.
- Collaboration and manual curation by experts in the field can help improve the accuracy and quality of automated annotations.
- Regular updates and re-annotation of bacterial genomes can help ensure that the annotations reflect the most current knowledge and technology, and improve the overall quality of genomic databases.

Understanding functional annotation

- Functional annotation is the process of assigning biological functions to genes and other genomic elements.
- It involves the identification of conserved domains, motifs, and other features that can provide clues about the functions of genes and other genomic elements.
- Functional annotation can be done using a variety of tools, including homology-based methods, gene ontology (GO) analysis, and pathway analysis.
- It can provide important insights into the biological processes and pathways that are active in a given organism, as well as potential drug targets and disease mechanisms.
- However, functional annotation is not always straightforward, and can be subject to errors and limitations such as incomplete reference databases, annotation bias, and functional redundancy.

Annotating novel genes and functional domains

- Functional annotation can help identify novel genes and functional domains in bacterial genomes, providing insights into the biological mechanisms underlying bacterial physiology and pathogenesis.
- Gene prediction tools can be used to identify putative genes in bacterial genomes, but functional annotation is required to assign biological meaning to these predicted genes.
- Several databases and tools are available to aid in functional annotation of bacterial genomes, including Gene Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG), and InterProScan.
- The identification of novel genes and functional domains in bacterial genomes can lead to the discovery of new therapeutic targets, vaccine candidates, and diagnostic markers for bacterial infections.
- The annotation of novel genes and functional domains can be challenging, as it requires a combination of computational predictions and experimental validation to confirm the function of putative genes and domains.

Comparative genomics

- Comparative genomics involves the analysis of genomic features and variations among different organisms to infer evolutionary relationships and identify functional differences.
- Phylogenetic analysis is a key tool in comparative genomics and involves the construction of evolutionary trees to depict the relationships between different organisms based on shared features or genetic similarities.
- Comparative genomics and phylogenetic analysis can help identify genes or pathways that are unique to specific bacterial species or lineages, and provide insights into the evolution of these organisms and their adaptations to different environments.
- Tools for comparative genomics and phylogenetic analysis include alignment software such as MAFFT, MUSCLE, and ClustalW, and phylogenetic inference methods such as Maximum Likelihood and Bayesian Inference.
- Quality checks for comparative genomics and phylogenetic analysis involve assessing the accuracy and reliability of the data, selecting appropriate analysis tools and methods, and incorporating relevant metadata and information to ensure the validity of the results.

Visualizing and interpreting annotation results

- Visualization tools can help researchers to interpret and compare the results of genome annotations, allowing them to identify patterns and gain insights into the functional roles of genes and gene products.
- Common visualization tools for genome annotations include Circos, GView, and Artemis, among others.
- Visualization can help researchers to identify patterns such as synteny, conserved domains, and gene clusters, and compare annotations across different genomes or different regions of the same genome.
- By combining annotation results with additional data, such as transcriptomic or proteomic data, researchers can gain a more complete picture of gene function and regulation.
- Effective visualization and interpretation of annotation results is essential for understanding the biological significance of genomic data and informing downstream analyses and experimental design.

Evaluating annotation quality

- **Consistency and completeness:** The annotation should be consistent throughout the genome, with similar annotations for genes that share functions, and the annotation should be as complete as possible, with all relevant information captured.
- **Evidence-based annotation:** The annotation should be based on solid evidence, such as experimental data or similarity to well-annotated genes in related species, and not solely on computational predictions.
- **Avoidance of errors:** The annotation should be free from errors such as misidentifications or incorrect functional assignments, which can lead to erroneous downstream analyses.
- **Comparison with reference genomes:** The annotation should be compared with annotations of closely related reference genomes to ensure consistency and identify any potential errors or gaps in annotation.
- **Verification of selected genes:** A subset of the annotated genes should be verified experimentally to ensure that the predicted functions are accurate, and to identify any errors or omissions in the annotation.

Integration of metadata and relevant information

- Integration of metadata and other relevant information can help in accurate functional annotation by providing additional context for gene function.
- Metadata can include information such as sample source, growth conditions, and environmental factors that can impact gene function and regulation.
- Other relevant information such as protein-protein interactions, gene expression data, and pathway analysis results can provide further insight into the biological significance of annotated genes.
- Integration of metadata and relevant information can help in identifying common genetic features among bacterial strains, and provide insights into evolutionary relationships among different strains and their adaptation to different environments.
- Effective integration of metadata and relevant information requires proper data management, standardization, and access to relevant databases and software tools.

Challenges and limitations

- The accuracy and completeness of bacterial assemblies can be affected by low quality and coverage sequencing data, sequencing errors, and assembly biases.
- The choice between de novo and reference-based assembly methods can be influenced by the availability of high-quality reference genomes and the potential for misassembly in de novo methods.
- Annotation quality can be impacted by the choice of annotation tools and databases, incomplete and inaccurate functional annotations, and errors in gene prediction and functional annotation.
- Automated annotation can be prone to errors and should be complemented with manual curation and validation of results.
- Bacterial genomes can be highly diverse, with complex and dynamic evolution, which can pose challenges for accurate assembly and annotation, as well as for comparative and functional analysis.

Future directions and advances

- Improvements in long-read sequencing technologies and their ability to produce more accurate and contiguous assemblies.
- Integration of various omics data (e.g., transcriptomics, proteomics) to improve annotation accuracy and completeness.
- Development of machine learning and artificial intelligence-based methods for automated and more accurate annotation of bacterial genomes.
- Continued improvements in bioinformatics software and tools for assembly and annotation, as well as their integration into more user-friendly pipelines.
- Increased focus on population genomics and strain-level diversity to better understand the evolution and spread of bacterial pathogens, and to develop more effective treatments and control strategies.

Summary

- Quality of sequencing data is critical to achieving accurate bacterial assembly and annotation.
- Choosing appropriate assembly methods and tools is important, and it can depend on the type of data, the research question, and the desired outcomes.
- Evaluation of assembly quality is a critical step in the process and includes methods for checking the completeness, accuracy, and correctness of the assembly.
- Automated annotation tools can save time and improve efficiency, but manual curation is still necessary for accurate and comprehensive functional annotation.
- Integration of metadata, relevant information, and expertise can aid in the interpretation of the results and help in making informed decisions.

References

- Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824-834. doi:10.1101/gr.213959.116
- Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 2009;19(6):1117-1123. doi:10.1101/gr.089532.108
- Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420-1428. doi:10.1093/bioinformatics/bts174
- Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153
- Brettin T, Davis JJ, Disz T, et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep. 2015;5:8365. doi:10.1038/srep08365
- Tatusova T, DiCuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614-6624. doi:10.1093/nar/gkw569
- Aziz RK, Bartels D, Best AA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9:75. doi:10.1186/1471-2164-9-75
- Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(Database issue):D61-65. doi:10.1093/nar/gkl842
- Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996-1004. doi:10.1038/nbt.4229
- Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100-3108. doi:10.1093/nar/gkm160

Q & A

Open forum for discussion and questions from the audience.

THANK YOU