Comparative analysis of Gene Finding tools when applied to *Trichoderma* genomes

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Applied Computing of Computer Science
University of Saskatchewan
Saskatoon

By

Connor Burbridge, Dave Schneider, Tony Kusalik

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

placeholder

Acknowledgements

I would like to acknowledge both Dr. Kusalik and Dave Schneider for their excellent support and mentorship during both COVID and my MSc. project. I would also like to thank Brendan Ashby and Dr. Leon Kochian for providing both data and additional financial support.

Contents

List of Tables

List of Figures

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
CDS	Coding sequence
Mb	Megabases
Kb	Kilobase
WGS	Whole Genome Shotgun (sequencing)
NGS	Next generation sequencing
PCR	Polymerase chain reaction
GMO	Genetically modified organism
CPU	Central processing unit
GC	Guanine cytosine
HMM	Hidden Markov model
UTR	Untranslated region
GFF	General feature format
BUSCO	Benchmarking Universal Single-Copy Orthologs
RSMI	Root, Soil and Microbial Interactions
MITE	Minature inverted-repeat transposable element
TIR	Terminal inverted repeat
TDR	Terminal direct repeat
maybe	Should I include tools like BLAST and resources like NCBI?

2 Background

2.1 Genomics

Genomics is a wide area of study focusing on the genomes of organisms from all varieties of life. A genome is a sequence of characters that contains the fundamental set of ‘rules’ used to create what we know as life. One can think of a genome as a set of instructions that our cells use in order to complete the tasks that make us function. A genome is comprised of tightly bundled sequences of DNA, which are stored in the nucleus of cells. These bundles of DNA contain sections known as genes, which can be thought of as the tools described by the set of instructions. These tools carry out a vast number of processes ranging from no known function at all to genes that are key in protecting against diseases cancer. Genomes can vary widely in size, ranging from small bacterial genomes of roughly 4 Mb up to approximately 149000 Mb. Piecing together genomes provides numerous opportunities to understand other ‘omics’ within cells, such as proteomics, metabolomics, transcriptomics and epigenomics.

2.2 Sequencing

Sequencing data is a pivotal form of data used in nearly all applications of Bioinformatics. To understand the processes used by organisms for survival, we must have an initial set of data points to work with. These sequences, referred to as reads after sequencing, are the foundation for solving problems ranging from taxonomical classification to the understanding or complex biological functions like signaling pathways. Reads may come in a variety of forms and formats depending on the desired application.

2.3 Whole Genome Shotgun Sequencing

Whole Genome Shotgun sequencing (WGS), is a method to produce a large number of genomic sequences from a sample of interest for the purpose of genome asembly. This is form of sequencing is quite common as it is has a wide variety of applications in research [2]. WGS involves slicing up genomic DNA into smaller segments. These small segments are then processed further resulting in a set of physical molecules that can be supplied to a compatible sequencing platform of which there are a variety. Modern sequencing platforms are comprised of next generation sequencing (NGS) and 3rd generation sequencing approaches.

2.4 Next Generation Sequencing - Illumina

Illumina™ [7] sequencing is one of the most popular NGS platforms currently available. Illumina sequencing produces a very large number of high quality short reads, typically between 75 and 250 base pairs in length. Sequencing libraries can be prepared to produce reads solely from one end of a sequence fragment (single-end) or both ends (paired-end). Advantages of paired end sequences are the additional context provided by the paired sequence on the opposite end of the fragment. This context is leveraged by read processing tools to identify features such as repetitive regions and genomic rearrangments, which can be significant in downstream analyses. Illumina sequence librairies are generated by first fragmenting the DNA samples, amplifying them via PCR, ligating adapters that allow the sequence to bind to the sequencing plate, and finally identifying each fragment's sequence of nucleotides using fluorescently-labeled nucleotides that bind to the fragments [18].

2.5 3rd Generation Sequencing - Nanopore

Nanopore™ [44] sequencing data is relatively recent approach to sequencing projects. While Illumina reads are short, Nanopore reads are much larger, ranging from 10Kb to 300Kb depending on the approach used. Long reads are beneficial due to their ability to bridge the gaps between difficult to assemble regions when performing sequence assembly. An example of a difficult to assemble region would be a region with a high repeat content, where a large number of small repeats may be collapsed during the assembly process, resulting in an assembly that does not represent the true nature of the sequence being studied [27]. While Nanopore was previously known for having lower quality base calls when compared to Illumina, that is no longer the case at this time. Nanopore sequencing works by passing long seqments of genetic sequence through a membrane bound protein and measuring changes in electrical current, which is characteristic of the nucleotide at a given position.

2.6 Secondary Metabolites

While cellular products essential to an organism's viability are are of great interest, there are a vast number of products that are not essential, but provide great benefit to the organism[13]. These other cellular products are known as secondary metabolites.

2.7 Non-ribosomal Peptide Synthetases

pee pee poopoo yujie

connor

burbridhge

lalalal

dhsjkhdkj ashdjksah kj

Column 1	Column 2	Column 3
Row 1, Col 1	Row 1, Col 2	Row 1, Col 3
Row 2, Col 1	Row 2, Col 2	Row 2, Col 3

Table 2.1: A table with 3 columns and 2 rows.

2.8 *Trichoderma*

Trichoderma are a genus of ascomycete fungi found in a wide variety of soils, and are well-known for their use in biomanufacturing of cellulases and hemicellulases and their roles as non-toxic, avirulent opportunistic plant symbionts[46][23]. It is believed that *Trichoderma* initally evolved as a parasite or saprotroph of plants, due to their abundant production of cellulases and plant wall degrading enzymes[23]. In addition to saprotrophy, *Trichoderma* species are also mycoparasitic, feeding on other Ascomycete and closely related fungal species[15]. *Trichoderma* comprises a large number of species with several large clades.

In addition to these beneficial properties some strains *Trichoderma* is a well studied . Many strains of *Trichoderma* have been shown to provide resistance to pathogenic bacteria and other fungi in soils through the use of polyketides, non-ribosomal peptide synthetases and other secondary metabolites[46] (SMs). Despite growing knowledge and study of these SMs, the mechanisms behind their function are still unclear[5], especially in the case of new *Trichoderma* species and strains.

Crop resistance to environmental stressors is a necessity for crop health and overall crop yields. Current popular methods for crop protection involve the use of pesticides and genetically modified organisms, which can be expensive and potentially politically dividing in the case of GMOs[42]. In addition, crops suffer when soils are not sufficient for crop growth and health. Soil insufficiencies can result in drought stress as well as nutrient stress, leading to poor overall yields.

2.9 Novel *Trichoderma* Genomes

Recently, two strains of *Trichoderma* have been identified in the prairie regions of Alberta and Saskatchewan. These two strains, named Tsth20 and DC1, have been found to have beneficial properties when used as an inoculant for plants in the soils mentioned before. In addition to these beneficial properties, the two strains mentioned previously provide even further protection for plants in dry, salty soils and one strain also has potential for use as a bioremediation tool in soils contaminated with hydrocarbon content. Bioremediation and resistance to drought tolerance has also been investigated in other strains of *Trichoderma* as well[38]. However, little is known about the mechanisms at work in these strains, so DC1 and Tsth20 were sequenced by the Global Institute for Food Security (no publication yet) in an initial attempt to better understand the details of these genomes. While this research does not directly identify genomic elements related the beneficial properties of these genomes, it may serve as a foundation for future research of *Trichoderma*.

2.10 Genome Assembly

Sequence assembly has been a long-standing problem in the field of bioinformatics[28]. Determining the correct order and combination of smaller subsequences into an accurate complete sequence assembly is computationally difficult in terms of compute resources such as memory, CPU cycles and storage required for input sequences[28]. In addition to these difficulties, there can be other issues encountered during asssembly due to the nature of the data or genomes themselves, such as low quality base calls for long read data, which is not necessarily the case today, or the inherent content of genomes themselves using repetitive regions as an example. Insufficient data may result in short, fragmented assemblies, depending on the size of the genomes, while sequence data that is not long enough can fail to fully capture repetitive regions in an assembly. A wide range of assembly tools have been developed with their own unique approaches to the genome assembly problem, so it is important to use an appropriate assembler for the task at hand, and also important to evaluate the assembly thoroughly.

Genome assembly tools generally approach the assembly problem using a graph-based approach. The most common graph-based approach is the de Bruijn graph assembly [12]. A graph in this context, is set of nodes (k-mers from sequences) connected by edges (overlaps between k-mers). Traversing through this graph results in longer subsequences that ultimately result in a set of sequences referred to as an assembly. In the early years of long read sequence data, sequencing platforms encountered difficulties producing consistently high scores for base calls when sequencing. To combat this, some assembly workflows may also include a polishing or correction step once the initial assembly is completed in which high quality short read sequences are used as supplemental information to correct low quality regions in the assembly. These low quality base calls are typically not present in modern long read sequencing approaches as the methodology and quality of calls have improved drastically. While the polishing step is arguably unnecessary in modern assemblies, the polishing programs remain available should researchers be interested in applying additional reads for polishing.

One approach to aid in the previously mentioned issue of assembly correctness is to use a combination of long and short reads in what is known as a hybrid assembly. Combining both highly accurate short reads with deep coverage along with less accurate but much longer reads can produce high quality genome assemblies that capture long repetitive regions. Hybrid assembly approaches have been shown to produce high quality assemblies in a wide variety of organisms as the combine long read data with short data to produce assemblies that properly represent long repetetive regions with additionaly high quality Illumina sequences for correction. Once assembled, the sequences must also be evaluated with measures such as N50, L50, coverage, average contig length and total assembled length to ensure that the genomes are well assembled, at least based on these metrics[28]. Following appropriate assembly protocols is essential to the further success of a project as downstream processing such as annotation depends on a high-quality assembly.

2.11 Identification of Anomolous Genomic regions

One important aspect of interest when assembling any form of sequence is GC content or percent GC of the assembled sequence. Large regions of anomolous GC content may be of interest to researchers as they may contain repetitive regions and unique features responsible for traits specific to the organism in question.

2.12 Repeat Identification and Masking

Repeat identification within assembled genomes is a problem that needs to be considered during the genome annotation process. Regions with long repeats can have a significant impact on genome assembly as well as gene finding due to the limitation of short reads used in some assemblies[43]. Short reads may be unable to bridge or cover entire repeat regions within a genome, so it is important to consider the use of long reads from technologies such as Nanopore or PacBio™[35] to provide a complete picture of these regions when pursuing a new genome assembly project. It is also possible for repetitive regions to contain genes as well, making for an interesting investigation in regards to *Trichoderma*, as fungal genomes have been shown to contain many repeat regions with a high concentration of A and T nucleotides[45]. Once these repetitive regions have been identified, the genome could be masked as a method to mark these regions for downstream processing if desired, as these regions may be poorly assembled and may result in found genes that do not truly exist in those regions. However, this may not be as common today, as repetetive regions have been shown to contain genes as well[40]. This may affect the gene finding process described later and may be an interesting topic to look into considering the large number of available gene finding programs.

2.13 Centromere Identification

A centromere is a region of a chromosome that is crucial for the proper cell division. These regions are the main anchor for microtubules, which are fibers that attach to centromeres to separate chromosomes during both mitosis and meiosis. Centromeres are critical to the survival of an organism, with malfunctions in the process of cell division usually resulting in potential disease and fatal outcomes[33]. Centromeric sequences can be comprised of several different genetic components, with repetititve regions being the most prevalent in the forms of satellite DNA and transposable elements. In addition to centromeric regions, there are flanking pericentric regions with their own properties, including potential candidates for small-interfering RNAs[33]. Identification and consideration of centromeric regions may prove useful when comparing the outputs of gene finding tools, as the underlying properties and structure of the genetic sequence differ in comparison to typical coding regions of DNA.

2.14 Gene Finding Methods

Gene finding (or gene annotation) has been a long standing computational problem in bioinformatics, which concerns itself with identifying potential genes within assemblies based on patterns or pre-existing experimental evidence considered by the gene finding program. This process is critical for unraveling and understanding the complex processes occurring in all forms of life. In a general sense, gene finding programs operate by searching for patters or indicators showing that a gene of feature may be present. The most basic indicators being start and stop codons, with splice sites in between should the sequence match the applied model. The results produced by gene finding tools can vary considerably for a number of reasons, including quality of the assembly, the intrinsic model used by the gene finder, filtering criteria, and even the nature of the organism and assembly itself. Given the broad applications, choice of gene finding tools, and the variability of assemblies being considered, it is important that we gain a deeper understanding of these tools prior to putting them to use.

There are two common methods for gene finding, those methods being *ab initio* methods, where programs search for patterns and gene structures, and similarity or evidence-based searches, which use prior information such as RNAseq data, expressed sequence tags and expressd protein sequences to identify genes within a new genome[16]. Complicating the process more is the introduction of introns and alternative splicing in eukaryotes, making it possible for one gene to have several possible transcripts at the same locus. An example of an *ab initio* method would be GeneMark-ES[25], while an evidence based tool would be Braker2[10]. *Ab initio* gene finders typically predict genes using a Hidden Markov Model (HMM)[16]. These predictions are based on 'signals' or features associated with a gene, such as the usual start, stop, exon and intron portions of a gene as well as upstream promoter sequences and more. In this case, these signals would be considered states in the terminology associated with HMMs. Gene finders wish to predict these states based on observations, or sequences presented to the model. HMMs in gene finding tools are trained beforehand and then applied to a sequence. This means that a gene finding program may not be trained in the context of any assembly provided to it, and thus may miss genes that are unique to the assembly in question. On the other hand, while still relying on HMMs for a 'base' set of predictions, evidence-based gene finding tools leverage new evidence that may be outside the scope of the pre-existing model[**?**]. As an example, an evidence-based model would be useful in a situation where you are interested in annotating a new assembly for a non-model organism. The addition of experimental data provides context specific to your assembly of interest while still retaining the predictions from existing HMM models.

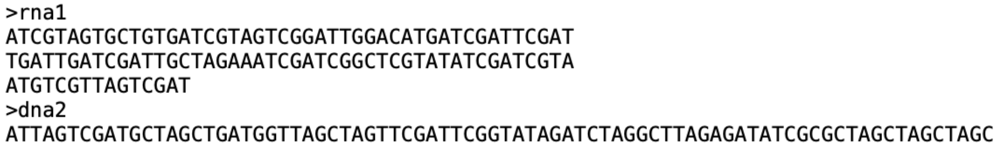
There are also other aspects of gene finding tools that are important to consider. These include features such as whether or not the gene finders find non-coding RNAs, annotation of 5' and 3' UTR regions, and in the case of ab-initio methods, the assumptions made by the underlying models used for gene finding. These features and others can influence a user's decision on which gene finding tool to consider and will complicate comparative analysis of multiple gene finding tools. (citation needed somewhere in here)

2.15 RefSeq

First, we will briefly discuss the RefSeq annotation process. RefSeq annotation is only applied to data that is submitted to NCBI. The RefSeq Eukaryotic Genome Annotation Pipeline[29] is a genome annotation process developed and maintained by NCBI. The pipeline is not directly publicly available to public users, and requires submission of data to NCBI. Once data is submitted to NCBI, the RefSeq annotation pipeline may be applied upon request only if the genome is the highest quality assembly for the species in question or if the genome is of significant interest to the scientific community, limiting reach of the annotation process to many users. The pipeline supplies existing RNAseq, CDS and protein sequences to NCBI's in-house gene prediction tool Gnomon, which produces trained models for gene prediction. While the tools used for alignment and processing of supporting sequence information are listed, the inner workings of Gnomon are not well documented, at least from the public perspective, and I was unable to find Gnomon in any compilable or executable form during my search. Recreation of the RefSeq pipeline would prove extremely challenging if not impossible without supporting information. Run times for this pipeline are difficult to determine due to the hidden nature of the pipeline, unknown compute resources and varying quantities of data used. The RefSeq annotation process produces comprehensive outputs, including CDS sequences, translated CDS, RNA from genomic sequences, proteins, feature counts and tables, and finally GFF and GTF formatted annotation files for these features.

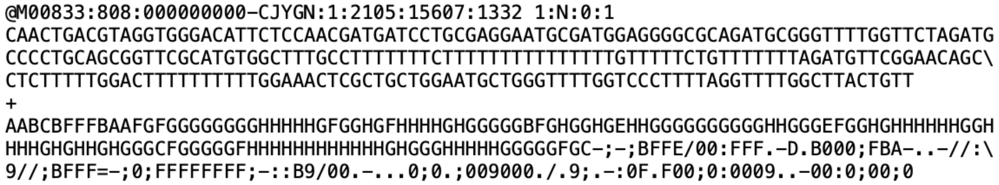
2.16 InterProScan

The outputs from gene finding tools are a set of potential genes that fit the model used by each tool. While they are considered genes, the use of the word gene is used in a very loose sense, in that these genes may or may not be functional or match any existing gene sequences from previous research. Typically, to confirm the 'correctness' of predicted genes, the outputs from a given tool are used in a sequence similarity search against a reference set of genes or a large datasbase comprised of multiple organisms. This approach is straightforward, but can introduce bias from database choice and also allows for vague or loose matches, depending on the parameters used and the interpretation of the results. Another approach is to use InterProScan, which is a tool used for functional annotation of proteins using evidence from a variety of databases [?]. The presence of some form of functional domain or annotated structure in a predicted gene sequence is reasonable evidence for the existence of a predicted gene. This approach also avoids the problems associated with similarity-based approaches.



```
>rna1
ATCGTAGTGCTGTGATCGTAGTCGGATTGGACATGATCGATTCGAT
TGATTGATCGATTGCTAGAAATCGATCGGCTCGTATATCGATCGTA
ATGTCGTTAGTCGAT
>dna2
ATTAGTCGATGCTAGCTGATGGTTAGCTAGTTCGATTCGGTATAGATCTAGGCTTAGAGATATCGCGCTAGCTAGCTAGC
```

Figure 2.1: Example of two FASTA sequence entries. One example with sequence characters split across multiple lines, and one showing all sequence characters on the same line.



```
@M00833:808:000000000-CJYGN:1:2105:15607:1332 1:N:0:1
CAACTGACGTAGGTGGGACATTCTCCAACGATGATCCTGCGAGGAATGCGATGGAGGGGCGCAGATGCGGGTTTTGGTTCTAGATG
CCCCTGCAGCGGTTCGCATGTGGCTTTGCCTTTTTTTCTTTTTTTTTTTTTTTGTTTTTCTGTTTTTTTAGATGTTCGGAACAGC\
CTCTTTTTGGACTTTTTTTTTTGGAAACTCGCTGCTGGAATGCTGGGTTTTGGTCCCTTTTAGGTTTTGGCTTACTGTT
+
AABCBFFFBAAFGFGGGGGGGGHHHHHGFGGHGFHHHHGHGGGGGBFGHGGHGEHHGGGGGGGGGGHHGGGEFGGHGHHHHHHGGH
HHHGHGHHGHGGGCFGGGGGFHHHHHHHHHHHHHGHGGGHHHHHGGGGGFGC-;-;BFFE/00:FFF.-D.B000;FBA-..-//:\
9//;BFFF=-;0;FFFFFFFF;-::B9/00.-...0;0.;009000./.9;.-:0F.F00;0:0009..-00:0;00;0
```

Figure 2.2: Example of the four lines in a FASTQ entry.

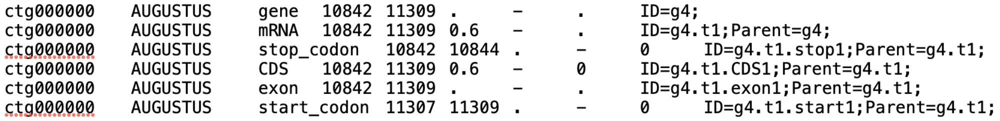
2.17 File Formats

2.17.1 FASTA

One of the most popular formats for sequences of DNA, RNA and amino acids is the FASTA format. The FASTA format consists of one or more entries containing two or more lines. The first line of an entry is the ID line, which must begin with a greater-than (>) character, followed by an ID and any other pertinent inormation for the following sequence. The greater-than character is the indicator that a new sequence has begun. The following line(s) contain the actual sequenced nucleotides or amino acids, which can be contained on one line or split across many lines. An example of multiple FASTA entries are shown in figure 2.1.

2.17.2 FASTQ

Another popular sequencing format is the FASTQ format. This format is very similar to the FASTA format but with the addition of two more lines per sequence entry and a change to the character indicating the beginning of a new sequence entry. An example of a FASTQ entry is shown in figure 2.2. In FASTQ formatted entries, the greater-than (>) character is swapped with the at (@) character. The IDs for the sequence also follow a specific format, which provide information about the sequencing run and flowcell that the read was sequenced on. This information can then be traced back to the sequencing experiment in the case that there were errors or anomalies in the output from the experiment. Following the ID is the string of base calls. The third line in a FASTQ entry is a plus (+) character, which indicates that the sequences character line has finished. Following the plus character is another sequence of characters, this time indicatintg the quality of basecall for the corresponding nucleotide base calls in the second line. The quality information included in FASTQ files are used to assess the quality of a sequencing run and extensively used in downstream processing steps, most notably in alignments.



```
ctg000000    AUGUSTUS    gene   10842  11309  .      -      .      ID=g4;
ctg000000    AUGUSTUS    mRNA   10842  11309  0.6    -      .      ID=g4.t1;Parent=g4;
ctg000000    AUGUSTUS    stop_codon   10842  10844  .      -      0      ID=g4.t1.stop1;Parent=g4.t1;
ctg000000    AUGUSTUS    CDS    10842  11309  0.6    -      0      ID=g4.t1.CDS1;Parent=g4.t1;
ctg000000    AUGUSTUS    exon   10842  11309  .      -      .      ID=g4.t1.exon1;Parent=g4.t1;
ctg000000    AUGUSTUS    start_codon  11307  11309  .      -      0      ID=g4.t1.start1;Parent=g4.t1;
```

Figure 2.3: An example of GFF entries for a single gene.

2.17.3 General Feature Format - GFF

General feature format (GFF) is a popular format for storing information about features relative to a position on an genetic sequence, and comprises a large portion of annotation results from this work. These features can be whatever the user desires, as long as the feature entry follows the required GFF guidelines. Relative to a reference sequence, each GFF entry contains the following tab-delimited columns: sequence ID, source, feature type, start position, end position, score, strand, phase, and a semi-colon delimited list of attributes. GFF files are widely supported accross bioinformatics tools, making them highly versatile while also remaining relatively simple in nature but also allowing for storage of more complicated items via the attributes column. One significatn useage of GFF files is in visualization of features against the reference sequence from which they were derived. Most genome viewers (or browsers) support GFF files as input, allowing intuitive visualization of many features when overlayed on a reference sequence. An example of a GFF entry can be seen in figure 2.3.

2.18 BLAST and tblastn

The basic local alignment search tool (BLAST)**??** has been a popular tool in the bioinformatics space, used to align biological sequences and measure their similarity. The original blast program was developed for alignment of nucleotide sequences, but other blast tools have been developed as well. One of these tools is tblastn which was developed to align reverse-translated protein sequences to another nucleotide sequence. To align the protein sequences, the query nucleotide sequences are translated in all six reading frames and then aligned to the proteins. Tblastn is particularly useful for identifying functional or conserved proteins in a nucleotide sequence.

2.19 BUSCO

The Benchmarking Universal Single-Copy Orthologs (BUSCO)**??** tool was developed to evaulate assemblies and subsequent annotations from the perspective of gene orthology. As genomes diverge evolutionarily, it is expected that some genes will be conserved as they are required for basic function. The BUSCO (Benchmarking Universal Single-Copy Orthologs) tool and datasets were developed to assess completeness of an annotation in comparison to evolutionarily conserved genes.

References

[1] *Kolmogorov–Smirnov Test*, pages 283–287. Springer New York, New York, NY, 2008.

[2] Jill Adams. Complex Genomes: Shotgun Sequencing. *Nature Education*, 1(1):186, 2008.

[3] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, oct 1990.

[4] S Andrews. FastQC: a quality control tool for high throughput sequence data., 2010.

[5] Scott E Baker, Giancarlo Perrone, Nathan M Richardson, Antonia Gallo, and Christian P Kubicek. Phylogenomic analysis of polyketide synthase-encoding genes in Trichoderma. *Microbiology (Reading, England)*, 158(Pt 1):147–154, jan 2012.

[6] Derek W Barnett, Erik K Garrison, Aaron R Quinlan, Michael P Strömberg, and Gabor T Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics (Oxford, England)*, 27(12):1691–1692, jun 2011.

[7] Simon Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433–438, jun 2004.

[8] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014.

[9] Mark Borodovsky and Alex Lomsadze. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics*, Chapter 4(SUPPL.35):4.6.1–4.6.10, sep 2011.

[10] Tomáš Brůna, Katharina J Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics*, 3(1):lqaa108, mar 2021.

[11] Brad Chapman. BCBio Python Package.

[12] Phillip E C Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, nov 2011.

[13] Arryn Craney, Salman Ahmed, and Justin Nodwell. Towards a new science of secondary metabolism. *The Journal of Antibiotics*, 66(7):387–400, 2013.

[14] Rina Dechter. chapter 12 - Temporal Constraint Networks. In Rina Dechter, editor, *Constraint Processing*, The Morgan Kaufmann Series in Artificial Intelligence, pages 333–362. Morgan Kaufmann, San Francisco, 2003.

[15] Irina S Druzhinina, Komal Chenthamara, Jian Zhang, Lea Atanasova, Dongqing Yang, Youzhi Miao, Mohammad J Rahimi, Marica Grujic, Feng Cai, Shadi Pourmehdi, Kamariah Abu Salim, Carina Pretzer, Alexey G Kopchinskiy, Bernard Henrissat, Alan Kuo, Hope Hundley, Mei Wang, Andrea Aerts, Asaf Salamov, Anna Lipzen, Kurt LaButti, Kerrie Barry, Igor V Grigoriev, Qirong Shen, and Christian P Kubicek. Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus Trichoderma from its plant-associated hosts. *PLoS genetics*, 14(4):e1007322, apr 2018.

[16] Girum Fitihamlak Ejigu and Jaehee Jung. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology*, 9(9), sep 2020.

[17] Python Software Foundation. Python.

[18] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[19] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, 29(8):1072–1075, apr 2013.

[20] Jiang Hu, Junpeng Fan, Zongyi Sun, and Shanlin Liu. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, 36(7):2253–2255, apr 2020.

[21] Jiang Hu, Zhuo Wang, Zongyi Sun, Benxia Hu, Adeola Oluwakemi Ayoola, Fan Liang, Jingjing Li, José R Sandoval, David N Cooper, Kai Ye, Jue Ruan, Chuan-Le Xiao, Depeng Wang, Dong-Dong Wu, and Sheng Wang. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology*, 25(1):107, 2024.

[22] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

[23] Christian P Kubicek, Andrei S Steindorff, Komal Chenthamara, Gelsomina Manganiello, Bernard Henrissat, Jian Zhang, Feng Cai, Alexey G Kopchinskiy, Eva M Kubicek, Alan Kuo, Riccardo Baroncelli, Sabrina Sarrocco, Eliane Ferreira Noronha, Giovanni Vannacci, Qirong Shen, Igor V Grigoriev, and Irina S Druzhinina. Evolution and comparative genomics of the most common Trichoderma species. *BMC Genomics*, 20(1):485, 2019.

[24] Chengzhi Liang, Long Mao, Doreen Ware, and Lincoln Stein. Evidence-based gene predictions in plant genomes. *Genome research*, 19(10):1912–1923, oct 2009.

[25] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O Chernoff, and Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20):6494–6506, 2005.

[26] Mosè Manni, Matthew R Berkeley, Mathieu Seppey, and Evgeny M Zdobnov. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, 1(12):e323, 2021.

[27] Vivien Marx. Method of the year: long-read sequencing. *Nature Methods*, 20(1):6–11, 2023.

[28] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167, 2013.

[29] NCBI. NCBI Eukaryotic Annotation Pipeline, 2024.

[30] NCBI. NCBI SRA Toolkit, 2025.

[31] Geo Pertea and Mihaela Pertea. GFF Utilities: GffRead and GffCompare. *F1000Research*, 9, 2020.

[32] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015.

[33] Miroslav Plohl, Nevenka Meštrović, and Brankica Mravinac. Centromere identity from the DNA point of view. *Chromosoma*, 123(4):313–325, aug 2014.

[34] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[35] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, proteomics and bioinformatics*, 13(5):278–289, oct 2015.

[36] P Rice, I Longden, and A Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, 16(6):276–277, jun 2000.

[37] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

[38] Biancamaria Senizza, Fabrizio Araniti, Simon Lewin, Sonja Wende, Steffen Kolb, and Luigi Lucini. Trichoderma spp.-mediated mitigation of heat, drought, and their combination on the Arabidopsis thaliana holobiont: a metabolomics and metabarcoding approach. *Frontiers in Plant Science*, 14, 2023.

[39] Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.

[40] R Keith Slotkin. The case for not masking away repetitive DNA. *Mobile DNA*, 9(1):15, 2018.

[41] Mario Stanke, Ana Tzvetkova, and Burkhard Morgenstern. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, 7(1):S11, 2006.

[42] Ioannis S Arvanitoyannis Theodoros H. Varzakas and Haralambos Baltas. The Politics and Science Behind GMO Acceptance. *Critical Reviews in Food Science and Nutrition*, 47(4):335–361, 2007.

[43] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46, nov 2011.

[44] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, 2021.

[45] David J Winter, Austen R D Ganley, Carolyn A Young, Ivan Liachko, Christopher L Schardl, Pierre-Yves Dupont, Daniel Berry, Arvina Ram, Barry Scott, and Murray P Cox. Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae. *PLOS Genetics*, 14(10):1–29, 2018.

[46] Sheridan L Woo, Rosa Hermosa, Matteo Lorito, and Enrique Monte. Trichoderma: a multipurpose, plant-beneficial microorganism for eco-sustainable agriculture. *Nature Reviews Microbiology*, 21(5):312–326, 2023.

[47] Bradley W Wright, Mark P Molloy, and Paul R Jaschke. Overlapping genes in natural and engineered genomes. *Nature Reviews Genetics*, 23(3):154–168, 2022.