**Pangenome Analysis of *Zymoseptoria Tritici De Novo* Genome Assemblies**

**Abstract**

*Zymoseptoria tritici* is a necrotrophic pathogen that is a leading threat to wheat crops in Europe, and around the world. Host resistances and the use of fungicides to counter the pathogen have been outpaced by its rapid adaptive evolutionary processes. Due to the *Z. tritici* genome being split into core and accessory genes, it can utilize non-vital segments of the genome as a means of testing rapid, large-scale chromosomal rearrangements, which result in intra-population variation in chromosomal structure and gene volume. By considering a geographically diverse set of genome assemblies, we can prepare comparative analyses to explore the global set of genes, or pangenome, and define the core and accessory genomes of the pathogen. With extensive genomic information, we will create a workflow to explore the prevalence of accessory and singleton genes in *Z. tritici* and enable future identification of effector genes contributing to virulence and fungicide resistance.

**Introduction**

Wheat (*Triticum aestivum*) is one of the most widely grown crops in the world. Due to the importance of wheat production, threats to crop productivity can result in widespread food shortages and economic strain 1. A major threat to wheat exists in *Zymoseptoria tritici* (syn. *Mycosphaerella Graminicola*), which is a globally spread necrotrophic pathogen that causes Septoria Leaf Blotch (STB) in wheat. The pathogen has a two-part infection cycle that begins with penetration and stealth pathogenesis, then transitions into necrotrophic colonization that is characterized by host cell death 2. STB can lead to large-scale crop losses that are estimated to cost €280-1200 million per year in the EU alone, and traditional methods to counter it are becoming increasingly unreliable due to rapid evolutionary changes in *Z. tritici* 3. Without host resistance to the pathogen, the use of fungicides is the most common solution. However, recent increases in fungicide use have resulted in quicker pathogen adaptation, culminating in parallel evolution in fungicide resistance genes between several isolated groups of *Z. tritici* around the world 4. Specifically, numerous mutations in the CYP51 gene have impacted resistance to azole fungicides, which attempt to disrupt sterol biosynthesis and target the fungal cell membrane 5. Prolonging the effectiveness of fungicides is dependent on accurate knowledge of the mutations contributing to resistant pathogen phenotypes 6. As pathogen evolution continues to create similar issues in separate regions, the approach of looking at the entirety of the global gene pool has become more prevalent in recent years.

With dramatically reduced costs associated with whole genome sequencing, there is an unprecedented amount of data to use in genomic studies. With this volume of data, there has been a dramatic shift from single-strain genomics to studies of the pangenome, or full gene pool of a global population 7. The pangenome of a microbial species consists of a shifting set of genes that can vary between and within populations. This chromosomal volatility can impact genes that lead to functional variation and differences in expressed phenotypes 8. Genes in a pangenome are categorized into core and accessory subsets based on their prevalence in the global population; only genes found in all individuals are considered part of the core genome. Genes are split into ortholog groups, or sets of homologous genes, which are defined by protein sequence similarity. Organisms with large pangenomes are difficult for typical studies involving gene mapping, as comparisons to a single reference genome can exclude important genetic variations present on a larger scale 9. Ortholog identification enables the tracking of genes in genomes prone to large-scale chromosomal rearrangements, or functional redundancies via gene duplications, which are potential barriers to other forms of locus identification 10. Looking only at each genome loci to identify the volume of unique genes found also largely ignores gene function as well as both non-coding sequences and transposable elements.

*Z. tritici* has 13 core chromosomes and 8 accessory chromosomes. The reference genome isolate IPO323 had its full genome sequenced in 2011 via Sanger sequencing. Since this initial publication, many genome-wide studies have been published, including a 19-long-read genome assembly that examined the pangenome of this species. This initial study showed that approximately 60% of all ortholog groups were a part of the core, 30% were accessory, and 10% were singletons, or only found in a single genome 11. This level of genomic variability demonstrates the importance of looking at the entire pangenome, not just a single isolate.

Accessory genes in pathogens tend to have a disproportionately high impact on adaptiveness and virulence 12. It has also been found that genes on accessory chromosomes undergo evolutionary changes more rapidly than those of the core genome, leading to the concept of the two-speed genome 13. Accessory chromosomes provide pathogens the opportunity to test novel gene mutations, as high levels of gene insertions, deletions, and duplications in regions of non-essential genes can potentially benefit the pathogen in certain hosts 14. This capacity for rapid evolution gives plant pathogens an advantage over their hosts, as faster adaptive evolution can counter newly evolved host resistance and fungal pathogenicity factors 15.

The gene pool of *Z. tritici* is not only highly varied through an inter-population lens, but nearly the entire pangenome has been recorded within a single field of wheat 16. As such, even localized studies of *Z. tritici* would be lacking potentially critical information without taking a wider view of the pangenome into account. Utilizing analyses of multiple sequenced genomes enables the ability to identify chromosomal polymorphism in the pangenome.

We will be utilizing *de novo* genome assemblies from different populations and analyzing the gene frequency in the *Z. tritici* pangenome for each strain to establish the core and accessory genomes. Virulence is a quantitative trait consisting of different components that require strict regulation to enable proper host colonization 17. Despite the availability of a finished reference genome and many re-sequenced genomes, only a limited number of effector genes, which are genes that are critical for virulence, have been observed 18-20. ‌Most of these functionally characterized genes have been regulatory, rather than directly contributing to virulence 21. And while some *Z. tritici* genes that contribute to host cell death have been found, their purpose in pathogen infection and their respective molecular targets in wheat are still largely unknown 22.

We will use a selection of 20 *de novo* genome assemblies from 6 geographically diverse strains and annotate these genomes with a fungal-specific gene annotation pipeline. Subsequently, these gene annotations will be used to assess the pangenome content of these isolates. The pipelines refined during this project will be used to support a much larger analyses with over 400 genomes. Using this data, we can identify traits leading to virulence and fungicide resistance throughout the global population and find their respective frequencies in future work.

**Results**

**Gene Model Predictions**

We created gene predictions and evidence-drive annotations for our set of 20 samples, utilizing the pathogen *Verticillium longisporum* as a reference species. Our total number of pre-filtered genes was 220848, and after removing genes under 150 bp, we were left with 219762 total gene models. The average length of the post-filtered gene models was 1526 bp.

In our set, the number of gene models found ranged from 10398 (isolate ISY\_Ar\_18b) to 11558 (isolate a12\_3B\_14), with a global mean of 10988 and median of 11099. While there were only a few samples from each geographic region, the samples from Israel had significantly fewer annotated genes than other samples, while the samples from Australia had higher average and median gene volumes.

**Pangenome Analysis**

Our pangenome analysis found a total of 15809 ortholog groups in our samples. Of these, 51% (8070 orthogroups) belonged to the core genome, as they were present in all 20 of our samples. Additionally, 32.2% (5090 accessory orthogroups) were present in multiple but not all our samples. The remaining 16.8% (2649 singleton orthogroups) were only found in a single isolate. The average orthogroup cluster in our set was 401 bp, with the average minimum and maximum ranges per cluster between 360 bp and 436 bp. However, there was a direct correlation between the number of isolates a cluster was present in and the length of cluster. Core orthogroups had a mean length of 514 bp, and on average ranged from 469 bp to 552 bp, while clusters found in 10 or fewer of our samples were less than half as large and had significantly higher variance in their size. Singleton clusters only had an average length of 222 bp.

Graphical user interface, application, table, Excel

Description automatically generated

We examined pairwise identities between each isolate to find the mean percent identity of CGN matches between genomes. We found that an average of 97.6% identity overlap was found between isolates, after correcting for self-paired overlaps. When comparing only isolates from the same original sample location, the percent identity overlap rose to 98.2. We also compared the percent identity overlap between core and accessory orthogroups. There was a 98.1% core orthogroup identity overlap, while in the accessory clusters this number fell to 93.7%.

Based on significantly shortened protein sequence matches, we found 14645 fragments that were suspected protein fragments, either from a single fragment or multi-protein fusions. 11923 (81.4%) of these potential fragments were found within an orthogroup cluster, and the remaining 1591 (18.6%) were found between multiple clusters.

Excluding the 11394 unique ortholog groups found in our reference genome (isolate AUS\_1A4), we found, expectedly, that adding additional samples produced diminishing returns on unique orthogroups. Most of the isolates with higher unique ortholog volume were primary samples in certain geographic regions or specific sample locations, or even simply samples that had higher levels of singletons. While it appears to be a normal case of diminishing returns on new clusters, the introduction of samples from new geographic regions would likely result in additional spikes. Due to these spikes, it is difficult to predict the number of unique orthogroups that could be found through adding more isolates, but we can assume that our current number of 15809 clusters, while incomplete, is largely representative of the *Z. tritici* pangenome in these regions.

**Discussion**

With the assembly and analysis of 20 geographically diverse isolates, we created a high-quality *Z. tritici* pangenome and pangenome analysis pipeline to utilize on our entire set of 400 samples. We explored the vast levels of gene plasticity within the pathogen and found prominent levels of variability throughout the entire genome. The genomic variation due to large chromosomal shifts is demonstrated by the extensive size of the accessory genome, consisting of roughly half of the entire gene pool, as well as the prevalence of singleton genes.

With our gene annotations, we had expected to find that our samples from the Middle East (Israel and Iran) had a greater number of gene models than those from other regions. As the *Z. tritici* pathogen originated in the Fertile Crescent, the samples there would have had more opportunity to undergo beneficial gene duplications to adapt to local wheat host resistances. However, these samples tended to have the fewest number of gene models on average, especially the samples taken in Israel, whereas Australian isolates had significantly more. This may be due to differences in local host resistance evolutionary speeds that result in highly volatile *Z. tritici* isolates to better adapt to hosts in certain regions than others. Infecting less adaptive hosts doesn’t require the same level of chromosomal plasticity in pathogens, so they can retain virulence while utilizing less of the accessory genome, resulting in fewer overall genes in those regions.

We found in our subset of 20 isolates a lower percentage of ortholog groups considered to be a part of the core genome than in similar previous studies that also examined the *Z. tritici* pangenome. While we found a core genome consisting of 51% of our orthogroups, previous examinations of an initial 5 genomes and a follow-up study of 19 isolates found core genome sizes of 60% and 58.1%, respectively 8,11. We also found a higher percentage of singleton orthogroups (16.8%) than the previous study of 19 genomes (10%). While these differences in sizes are significant, this is more likely due to variation in the methods of identifying homologous proteins than differences in volatility between the isolate genomes themselves. Lowering the thresholds for both required length and identity percent would reduce the size of our pangenome and remove numerous singleton and accessory genes, likely resulting in comparative findings in gene presence-absence variations.

We found that the more isolate genomes an ortholog group was present in, the larger it tended to be. This is expected, as core genes tend to be more stable and have more complex functions that are vital to the pathogen’s survival. Accessory and singleton genes are often the result of whole or partial duplications, and undergo significant levels of insertions and deletions, causing a greater variance in gene length. Larger changes to the genome are more likely to result in phylogenic traits detrimental to the pathogen’s continued existence, so these larger duplicated regions would tend to be phased out of the pangenome.

We found a 97.6% identity overlap between all our isolates, after correcting for self-pairing, which rose to 98.2% when doing pairwise comparisons between isolates of the same sample location. While we expected the percent identity of isolates from the same locations to be higher, the difference when looking at the entire subset of 20 samples was almost negligible. This is likely due in part to the significant functional variation present even within specific populations, as many of these mutations are neutral and can remain in the gene pool despite being non-vital. More significant was the difference in percent identity overlap within core orthogroups and within accessory orthogroups. Core orthogroup clusters had a 98.1% identity overlap, while accessory clusters had only a 93.7% identity overlap. This emphasizes the extreme volatility of the accessory genome, categorized by rapid duplications, insertions, and deletions. Even genes in the same orthogroup have significant levels of mismatches and gaps when compared to each other; this genomic testing ground is what enables rapid adaptations to host resistance and fungicides.

With our 20 isolates, we found 15809 unique ortholog groups, with diminishing returns on each new isolate added. However, due to higher variation levels between isolates from different geographic regions or even sample locations within the same region, additional samples from new sources could result in new spikes in *de novo* identified orthogroups. As this pathogen has more time to evolve and adapt, the global gene pool will only continue to increase, meaning utilizing genome assemblies of diverse, modern isolates will likely improve the coverage of the pangenome.

**Methods**

**Funannotate Pipeline**

We utilized a combination of *de novo* gene assemblies and data from past research from the Croll Lab, NCBI, and FungiDB, which can be found in the Availability of Data and Materials section. Our samples were taken from Israel, Iran, Switzerland, USA (Oregon), mainland Australia, and Tasmania. After initial reformatting of our assemblies, we began our funannotate (v. 1.8.1) (<https://github.com/nextgenusfs/funannotate>) pipeline. Funannotate is a software package designed specifically for the prediction and annotation of fungal genomes.

We ran an initial cleaning of our genome assemblies through funannotate with default parameters. This cleaning utilizes minimap2 (v. 2.1.7) to sort by size from smallest to largest, remove all contigs under 500 bp, then search through the remaining contigs up to those of the median length of contigs in the assembly to find and remove duplicates 23. The default parameters require both 95% identity and 95% coverage of overlap to be considered a duplicate contig.

Graphical user interface, application

Description automatically generated

The funannotate sort command was used with default parameters to re-order our remaining cleaned contigs from longest to shortest and relabel our input scaffolds to properly fit formatting requirements for both NCBI submission and AUGUSTUS (v. 3.3.3) inputs 24.

Graphical user interface, application, Word

Description automatically generated

Next, funannotate mask was run with default parameters to soft-mask repetitive regions via tantan (v. 23) 25. These soft-masked fasta outputs are then passed to the prediction step.

Graphical user interface, text, application

Description automatically generated

Gene predictions were generated using the funannotate predict command using *Verticillium longisporum* as a reference genome which was pre-trained from AUGUSTUS.

Graphical user interface, text, application

Description automatically generated

First, the proteins are mapped to the genome using DIAMOND (v. 0.9.30) and Exonerate (v. 2.4.0) 26,27. Funannotate then parses the soft-masked genome fasta files and pulls the data into a BED file. Default settings dictate a double-sided approach to handling the filtering of repetitive regions. Overlap filtering uses the soft-masked BED file and ignores modelled genes with 90% or greater repeated base pairs. Blast filtering compares the sequences of amino acids to a database of known transposons.

The soft-masked regions from the BED file are then directly passed to the *ab initio* predictors AUGUSTUS and GeneMark-ES (v. 4.61) to internally handle the repetitive regions 28. By default, GeneMark will ignore repeated regions shorter than 2 kb when making predictions. Gene predictions are then done by GeneMark-ES via iterative unsupervised training. BUSCO (v. 3.0.2) is called to find conserved gene models for training *ab initio* parameters and formatted into a gff3 file for EVidenceModeler (v. 1.1.1), which creates weighted consensus genome models to then be validated by a BUSCO hidden Markov model (HMM) search 29,30. Next, AUGUSTUS, SNAP (v. 2013-11-29), and GlimmerHMM (v. 3.0.4c) gene predictions are run using the *ab initio* parameters created by BUSCO 31,32. The AUGUSTUS predictions are separated into standard and high-quality groups. Then, all gene predictions from each of these tools are converted to gff3 format and passed through EVidenceModeler to create consensus gene models from all inputs, with the high-quality AUGUSTUS predictions weighted twice as heavily as each of the others.

Protein fasta files are created from these gene models, which then undergo quality control to remove models under 50 aa, models with span gaps, and those of transposable elements. TRNAscan-SE (v. 2.0.6) is used to detect and classify tRNA genes 33. Finally, tbl2asn (v. 25.8) (<https://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/>) is used to convert the output tbl annotation files to proper GenBank submission format. Funannotate also parses through NCBI errors and notifies of any invalid gene models.

**Pangloss Pipeline**

Pangloss (v. 0.9.1) is a three-part pipeline for pangenome analyses of microbial eukaryotes 34. The first component is an optional gene and gene location prediction pipeline that overlaps with our funannotate predictions and was not utilized. The second part utilizes the Perl software PanOCT (v. 3.2) to construct a syntenic pangenome through ortholog clustering, along with an optional refinement using reciprocal homology between clusters of syntenic orthologs to account for micro-synteny loss due to assembly artifacts 35. Pangloss begins by performing an all-vs-all BLASTp (v. 2.9.0) search for the entire pangenome protein sequence dataset 36. Then PanOCT is called to construct the pangenome, with divisions between the core and accessory genomes. The four required inputs for running PanOCT without gene predictions are a full protein database, created through the combination of protein fasta files output from funannotate, a full set of attributes, which are a combination of modified gff3 files created by funannotate, an all-vs-all BLASTp results file which consists of funannotate’s scaffold fasta files, and a list of genomic isolate tags. The creation of these combined files from the sets of funannotate outputs is done in the initial step of the Pangloss pipeline. Pangloss was run without gene predictions, with all other parameters as default, and the sample AUS\_1A4 was used as the reference isolate.

Graphical user interface, text, application

Description automatically generated

The all-vs-all tabular BLASTp file is filtered, and results with an amino acid percent identity under 35%, a BLAST E-value cutoff greater than 10-5, or a minimum percent match length of query and subject less than 1% are excluded. PanOCT calculates the BLAST score ratio (BSR) using the lengths of each protein, taken from the input protein fasta file. The BSR is a classification of all hypothetical proteins, which is used for calculating the proteome similarity between multiple genomes.

PanOCT uses a combination of conserved gene neighborhood information (conservation of gene order and orientation), pairwise sequence identity, and frame-shift detection (the detection of amino acid insertions and deletions not in multiples of 3, causing shifts in all following translated proteins) to identify paralogous clusters in accessory genomes created by duplications and subsequent alterations of core gene models.

Before PanOCT is called, we added an additional custom python script to our pipeline to remove gene models under 150 bp from our combined protein database, to properly match our attributes file.

Graphical user interface, text, application

Description automatically generated

Then, we ran PanOCT with the following command.

Graphical user interface, text, application

Description automatically generated

PanOCT has several outputs, including the match table and match table ID files, which are both grouped by individual ortholog clusters. The match table file simply shows which clusters are present in each of the input genomes, including the reference. The match table ID file shows the percent identity for each protein to the reference genome, along with identifiers for each protein. The PanOCT ID file contains reference protein annotation and percent identities to orthologs in each target genome. The final default output is the frame-shift report, which consists of proteins suspected of being split due to frame-shifts. It shows the percentage of retained and ignored fragments that are considered part of the same frame-shifted gene for each protein fragment.

After PanOCT is complete, the Pangloss pipeline continues to attempt to fill in syntenic cluster gaps due to genomic events or assembly artifacts and refines the match table file. It then extracts the gene model clusters from the combined protein and scaffold databases and writes out protein and nucleotide sequence families to the match table Finally, the core and accessory proteins and nucleotides are output to 4 separate fasta files.

Utilizing a binarized form of the match table file (1 if ortholog group is present in isolate, 0 if not present), we calculated gene presence-absence variation within our set of 20 samples. We divided our orthogroups into core, accessory, and singleton categories; orthogroups that are present in all samples, those present in some samples but not all, and those present in only a single sample, respectively. We then took data from the cluster weight PanOCT output to determine average lengths of ortholog group clusters and compare them between the core and accessory genomes to try to discover any significant differences between the groups.

Next, we used the pairwise identity matrix to examine pairwise identities between each isolate. This matrix shows the mean percent identity of Conserved Gene Neighborhood matches between each genome. We corrected for self-paired overlaps to find and compare percent identity overlaps between all isolates and between isolates of the same region. Then we found the identity percentages of isolates within ortholog groups in the core and accessory genomes, respectively.

We then used the fragments fusion output file to identify probable protein fragments within our ortholog groups. By examining only high-quality matches between or within orthogroups, suspected protein fragmentations and fusions are predicted when the proteins have significant differences in length. We are also able to identify the number of potential fragments found within clusters, and those found between clusters.

Lastly, with our PanOCT outputs, we could see how the global ortholog group pool increased with each additional isolate added to our study. With this, we can attempt to predict the size of the global orthogroup and genome size of the entire *Z. tritici* species.

The next step to add to our pipeline would be to run a gene ontology analysis to include functional annotations to our ortholog data. We can find secreted proteins and effector candidates and compare to them to previous association mapping studies. Using known effector genes, we can identify their prevalence within our complete set of samples to extensively explore the parallel evolution between geographically isolated *Z. tritici* strains. The analysis of effector genes will also enable us to examine the divisions between effector gene presence in core, accessory, and singleton clusters.

**Software and Setup**

A substantial portion of our time was spent initializing and combining these two separate pipelines into a larger workflow. Funannotate’s extremely in-depth functions rely on a plethora of other dependencies. There are 11 Python packages, 27 Perl modules, 30 other external dependencies, 10 databases, and 7 environmental variables that all need to be properly set up to initialize and properly run the pipeline. After spending a lot of time unsuccessfully attempting to properly set up all these dependencies locally, we chose to use the funannotate that was setup on the University of Birmingham’s Birmingham Environment for Academic Research (BEAR) portal. However, many of the dependencies there were either outdated or improperly set up, so a handful of manual workarounds and local installations were required before funannotate could be run. The time required for funannotate’s clean and predict functions were significant barriers, as the cleaning took up to 36 hours for some samples, and the predictions nearly 60 hours for a single sample. We had initially attempted to run our pangenome analysis on all 400 of our samples, so even with the use of batch array jobs, it took a very significant amount of time for each of the longer steps in our workflow.

Due to the CPU, memory, and time requirements of the funannotate preparation and prediction pipeline, batch array jobs were needed to submit scripts for the funannotate clean and predict steps, and regular batch jobs were submitted for the sorting and masking stages.

Our two pipelines are both designed for eukaryotic genome analyses but are not easily compatible. Due to the differences between funannotate outputs and the inputs required for Pangloss, manual scripting was needed to properly re-format each of the three necessary files. Specifically, the gff3 outputs from funannotate were converted to attribute files using modified scripts initially created for preparing outputs from FunGAP (v1.0.1) for Pangloss 37,38. These tab-delimited files require the contig or scaffold ID, the protein identifier, the start position, the stop position, the annotation, and genome identifier.

Like our funannotate pipeline, we ran into many issues with our Pangloss pipeline. Instead of being able to run it straight through, we had to use Pangloss to create our all-vs-all BLASTp file, then manually run PanOCT after we fixed issues with file formatting that prevented a straightforward workflow. Due to time limitations, we had to manually create a handful of simple Python scripts to analyze summary statistics from several of our PanOCT output files, instead of using Pangloss scripting to do this for us. We also were unable to reach the gene ontology and functional annotation stages of the workflow, which would have enabled us to identify the presence, loci, and volatility of known effector genes affecting both virulence and fungicide resistance.

**Conclusion**

Our subset of 20 isolates provided an opportunity to test our pipeline that we will expand to our entire set of 400 isolates. As we go more in-depth in our pangenome analysis to search for effector genes, we can potentially learn more about the functions behind the genes causing virulence and fungicide resistance.

**Further Work**

For future work, our initial funannotate predictions will be completed with RNASeq evidence provided to improve the quality of our predictions. We will also utilize funannotate’s annotate command to create functional annotations to pipe into our Pangloss workflow. Along with our total set of 400 samples, we will also include known reference genome IPO323. We will use this is our reference isolate when running our Pangloss pipeline. Once we have our functional annotations and have run gene ontology on our ortholog groups, we will also create visualizations to better demonstrate patterns in genetic groupings, and the presence of known effector genes.

**Supplementary Data and Materials**

Example of a bash array job script used for our funannotate prediction step can be found at

<https://github.com/ConnerSick/Ztritici/blob/main/Bear_Array_Predict_Script.txt>.

Example of the script used to reformat funannotate gff3 files to proper Pangloss attribute file input can be found at <https://github.com/ConnerSick/Ztritici/blob/main/Gff3_to_Sh_Formatting_Script.txt>.

Reference genome transcript fasta and gff3 files from the Croll Lab can be found at <https://github.com/crolllab/datasets/tree/master/19-isolate_pangenome_BMC_Biology_2020> and from FungiDB at <https://fungidb.org/fungidb/app/record/dataset/NCBITAXON_336722> and <https://fungidb.org/fungidb/app/record/dataset/NCBITAXON_1276537>.

NCBI RNA-Seq data can be found at <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA559981>.

**References**

1. Orton, E. S., Deller, S., & Brown, J. K. M. Mycosphaerella graminicola: from genomics to disease control. Mol. Plant Pathol. 12, 413–424 (2011).
2. Dean, R., *et al*. The Top 10 fungal pathogens in molecular plant pathology. Molecular Plant Pathology, 13(4), pp.414–430. doi:10.1111/j.1364-3703.2011.00783.x (2012).
3. Figueroa, M., Hammond-Kosack, K. E., & Solomon, P. S. A review of wheat diseases—a field perspective. Mol Plant Pathol 19:1523–1536 (2018).
4. McDonald, M. C., *et al*. Rapid Parallel Evolution of Azole Fungicide Resistance in Australian Populations of the Wheat Pathogen Zymoseptoria tritici. Applied and Environmental Microbiology, 85(4). doi:10.1128/aem.01908-18 (2019).
5. Cools, H. J., & Fraaije, B. A. Update on mechanisms of azole resistance in Mycosphaerella graminicola and implications for future control. Pest Manag Sci 69:150–155 (2013).
6. Lucas, J. A., Hawkins, N. J., & Fraaije, B. A. The evolution of fungicide resistance. Adv Appl Microbiol 90:29–92 (2015).
7. ‌Vernikos, G. S. A Review of Pangenome Tools and Recent Studies. The Pangenome, pp.89–112. doi:10.1007/978-3-030-38281-0\_4 (2020).
8. Plissonneau, C., Hartmann, F. E., & Croll, D. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biology, 16(1). doi:10.1186/s12915-017-0457-4 (2018).
9. Marschall, T., *et al*. Computational pan-genomics: status, promises and challenges. Briefings in Bioinformatics, p.bbw089. doi:10.1093/bib/bbw089 (2016).
10. Badet, T. & Croll, D. The rise and fall of genes: origins and functions of plant pathogen pangenomes. Current Opinion in Plant Biology, [online] 56, pp.65–73. doi:10.1016/j.pbi.2020.04.009 (2020).
11. Badet, T., *et al*. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biology, 18(1). doi:10.1186/s12915-020-0744-3 (2020).
12. Wu, Y., Zaiden, N., & Cao, B. The Core- and Pan-Genomic Analyses of the Genus Comamonas: From Environmental Adaptation to Potential Virulence. Front. Microbiol. 9:3096. doi: 10.3389/fmicb.2018.03096 (2018).
13. Stukenbrock, E. H., *et al*. Whole-Genome and Chromosome Evolution Associated with Host Adaptation and Speciation of the Wheat Pathogen Mycosphaerella graminicola. PLoS Genetics, 6(12), p.e1001189. doi:10.1371/journal.pgen.1001189 (2010).
14. Croll, D., & McDonald, B. A. The Accessory Genome as a Cradle for Adaptive Evolution in Pathogens. PLoS Pathogens, 8(4), p.e1002608. doi:10.1371/journal.ppat.1002608 (2012).
15. Jones, J. D. G., & Dangl, J. L. The plant immune system. Nature. 444:323–9 (2006).
16. Zhan, J., Pettway, R. E., & McDonald, B. A. The global genetic structure of the wheat pathogen Mycosphaerella graminicola is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. Fungal Genet. Biol. 38, 286–297 (2003).
17. Zhan, J., Torriani, S. F., & McDonald, B. A. Significant difference in pathogenicity between MAT1-1 and MAT1-2 isolates in the wheat pathogen Mycosphaerella graminicola. Fungal Genet. Biol. 44:339-346 (2007).
18. Goodwin, S. B., *et al*. Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7: e1002070 (2011).
19. Croll, D., Zala, M., & McDonald, B. A. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. PLoS Genet. 9: e1003567 (2013).
20. Marshall, R., *et al*. Analysis of two in planta expressed LysM effector homologs from the fungus Mycosphaerella graminicola reveals novel functional properties and varying contributions to virulence on wheat. Plant Physiol. 156:756–69 (2011).
21. ‌Palma-Guerrero, J., *et al*. Comparative Transcriptome Analyses in Zymoseptoria tritici Reveal Significant Differences in Gene Expression Among Strains During Plant Infection. Molecular Plant-Microbe Interactions®, 30(3), pp.231–244. doi:10.1094/mpmi-07-16-0146-r (2017).
22. ‌Karki, S. J., *et al*. A small secreted protein from Zymoseptoria tritici interacts with a wheat E3 ubiquitin ligase to promote disease. Journal of Experimental Botany, [online] 72(2), pp.733–746. doi:10.1093/jxb/eraa489 (2020).
23. Heng, L. Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics, Volume 34, Issue 18, Pages 3094–3100, <https://doi.org/10.1093/bioinformatics/bty191> (2018).
24. Stanke, M., & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Research, 33, W465-W467 (2005).
25. Frith, M. C. A new repeat-masking method enables specific detection of homologous sequences, Nucleic Acids Research 2011 39(4): Page e23, <https://doi.org/10.1093/nar/gkq1212> (2011).
26. Buchfink, B., Reuter, K., & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND, Nature Methods 18, 366–368. doi:10.1038/s41592-021-01101-x (2021).
27. Slater, G. S. C., & Birney, E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31. <https://doi.org/10.1186/1471-2105-6-31> (2005).
28. Lomsadze, A., Burns, P. D., & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research, July 2, doi: 10.1093/nar/gku557 (2014).
29. Manni, M., *et al*. BUSCO: Assessing genomic data quality and beyond. Current Protocols, 1, e323. doi: 10.1002/cpz1.323 (2021).
30. Haas, *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology, 9:R7. doi:10.1186/gb-2008-9-1-r7 (2008).
31. Korf, I. Gene finding in novel Genomes. BMC Bioinformatics, 5:59 (2004).
32. Salzberg, S. L., *et al*. Interpolated Markov models for eukaryotic gene finding. Genomics 59(1): 24-31 (1999).
33. Chan, P. P., *et al*. tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. Nucl. Acids Res. 49:9077–9096 (2021).
34. McCarthy, C. G. P., & Fitzpatrick, D. A. Pangloss: a tool for pan-genome analysis of microbial eukaryotes. Genes, 10(7):521. Doi: 10.3390/genes10070521 (2019).
35. Fouts, D. E., *et al*. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res. 2012 Dec;40(22): e172. doi: 10.1093/nar/gks757. Epub 2012 Aug 16. PMID: 22904089; PMCID: PMC3526259 (2012).
36. Camacho, C., *et al*. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421. doi: 10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857 (2009).
37. Gourlie, R., *et al.* Dissecting the Pyrenophora tritici-repentis (tan spot of wheat) pangenome. bioRxiv 2022.03.07.483352; doi: <https://doi.org/10.1101/2022.03.07.483352> (2022).
38. Min, B., Grigoriev, I. V., & Choi, I. FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. Bioinformatics, Volume 33, Issue 18, Pages 2936–2937, <https://doi.org/10.1093/bioinformatics/btx353> (2017).