**Exploring the phylogenetic evolution and geographic transmission patterns of the rice blast *Magnaporthe oryzae* in Africa**

Chapter 1: Introduction

**Introduction & background**

Rice blast disease is caused by a fungal pathogen, *Magnaporthe oryzae.* Its outbreaks keep reoccurring in areas where rice is grown resulting in major losses to rice production globally and in turn threatens global food security. The disease has been found to be difficult to control but knowing the structure of the current population and its diversity will help in managing the disease (Dean et al., 2005; Zhong et al., 2018). The amount of rice destroyed by rice blast disease annually can feed up to 60 million people (Dean et al., 2005). The disease affects rice plants at all stages of growth. It infects all parts of the plant except the roots, the leaves and panicles being the most seriously affected (Chuwa, Mabagala, & Reuben, 2015). The spread of rice blast disease indicates the evolution of new pathotypes and enhancement of population diversity is unknown (Onaga, Wydra, Koopmann, Séré, & Von Tiedemann, 2015).

Previous global phylogenetic studies based on whole genome data only included a few strains from Africa. South East Asia was found to be the centre of origin of *M. oryzae*. Therefore, the pathogen originated from Asia and then spread to the rest of the world. Insufficient clues were provided regarding to the point of entry of the pathogen to Africa and the possible routes of migration. Inferring these routes and identifying the genetic diversity within strains will likely aid in making future predictions regarding epidemics, spread and management of the disease and in the proposition of new control strategies (Gladieux et al., 2018; Saleh et al., 2012; Zhong et al., 2018).

Some studies in Africa used molecular analysis techniques. They used non sequence based markers and some conserved genes. They showed very small differences between strains from major rice growing areas (Chuwa et al., 2015; Onaga et al., 2015). According to (Mutiga et al., 2017), they used genotyping by sequencing method and found evidence of differentiation and possible adaptation along geographic lines, though at a lower representation to allow confident phylogenetic positioning.

This study will focus on whole genome data obtained through an Illumina sequencing platform. 50 isolates of *M. oryzae* from various African countries will be primarily characterized based on the occurrence of SNPs throughout the genome. The study will also look at the evolution of particular genes such as *Pi9, Pita2, Pil2* known to be important in virulence and pathogenicity according to previous studies. Signatures of adaptation to different environmental conditions in these strains will be studied. Having a better understanding of the pathogen’s phylogeograhy will help in informing future measures taken towards combating its spread and in managing the disease.

However, there have been no studies on the phylogenetic evolution and geographic transmission patterns of rice blast *M. oryzae* in Africa, from the point of first entry and spread within the region. This will likely inform the efforts put in towards managing the disease. Consequently, that should have far reaching beneficial effects on consumer livelihoods globally.

**Problem statement**

Millions of people rely on rice as a source of livelihood in the world. However, global rice production is constrained or threatened by rice blast disease. With the expected increase in cultivation of rice, the disease will inevitably continue to spread and cause havoc in all the rice growing fields. The consequences are significant reductions in rice production as well as a strain on economies. If the disease causing pathogen is not contained, the global food security is threatened as rice will continue to die in the fields.

**Justification**

In order to contain the challenges posed by rice blast disease, it is important to investigate the phylogenetic transmission of the pathogen within Africa. This study will focus on whole genome data of 50 isolates from different parts of Africa. Results of the study will seek to fill in the knowledge gap regarding the genetic diversity of the pathogen within Africa as well as feed into the efforts being put in towards management and combating the spread of the disease.

**Objectives**

Main objective

1. To determine the possible point of entry of the *M. oryzae* pathogen to Africa and find any possible transmission patterns within Africa.

Specific objectives

1. To investigate insights into the evolution of genes that are known to be important for virulence and pathogenicity.
2. To examine the signatures of adaptation within the African strains.

**Hypothesis**

1. There are several possible transmission patterns of the pathogen within Africa.
2. There is a genetic diversity of the pathogen within the African strains.
3. The genes responsible for virulence and pathogenicity in the pathogen have not evolved.
4. There are several signatures of adaptation amongst the African strains.

**Scope/ limitations of the study**

This is a phylogeographic study that will explore the possible transmission patterns of *Magnaporthe oryzae*, a fungal pathogen, within Africa. Rice blast disease, caused by this pathogen, is a menace in all rice growing fields. It will feature 50 isolates of the pathogen. The project will be completed in 8 months, by June 2020.

Chapter 2: Literature review

While there has been much research on the transmission patterns of rice blast disease globally, few researchers have taken into consideration how the disease causing pathogen *M. oryzae* entered Africa and spread to other countries. We need to perform population genomic analysis to find out the structure of the current African *M. oryzae* population as well as how the diversity is generated, to help in managing the disease. An African genetic subdivision of *M. oryzae* and its history of invasion.

Rice blast disease was first reported in Asia over three centuries ago. It has since spread to 85 countries and counting. Its adaptation to different environmental conditions is very high (Odjo, Ahohuendo, Onasanya, Akator, & Séré, 2011).

It was a global phylogenetic study. They used whole genome data of 50 different isolates of the disease causing pathogen from various times and places in the world. The amount of African strains used in this study were very few. They performed SNP calling. They found 6 lineages of the disease causing pathogen. They also found that the pathogen did separate a millennium ago. They used (Gladieux et al., 2018).

They looked at 100 isolates of *M. oryzae* from all over the world, from different time points. They assembled the short Illumina reads of whole genome data, then aligned to a reference genome, performed SNP calling, constructed phylogenetic trees, performed population genomic analysis and the estimated the time of divergence. Only a few African strains were involved in the study. They found 3 clades. They estimated the time of divergence to be about 700-1000 years ago. The estimated divergence times were based on the time the strain was collected and the SNPs (Zhong et al., 2018).

The pathogen reproduces asexually however sexual reproduction was found in just a few areas of South East Asia. The strains were genotyped using microsatellite markers which was a molecular biology technique. This was demonstrated using two opposite mating types, with one being female-fertile. They produced viable offspring. The sexual reproduction capability is then lost outside this known centre of origin (Saleh et al., 2012).

It was a global study looking at the genetic diversity of 1372 strains from all over the world. A few African isolates were included in the study. The different strains were genotyped using microsatellite markers which was a molecular biology technique. They looked at the genetic diversity within and without the strains. They found 3 clades with Asia being the origin where the pathogen was found to have been reproducing sexually, from where it spread to the rest of the world. Getting to understand the link between the evolutionary history and how diverse M. oryzae is today will be fundamental in foreseeing likely changes (Saleh, Milazzo, Adreit, Fournier, & Tharreau, 2014).

They looked at 88 isolates of the disease causing pathogen from East Africa, West Africa and the Philippines (Asia). They used a molecular technique whereby they used amplified fragment length polymorphism (AFLP) markers. They found that there was no population structure and a significant flow of genes was evident. The also analysed the pathogenicity of the isolates which showed that aggression was variable amongst the EA isolates, indicating possible racial diversity. They found that there is a possibility that the EA population either consists of only one genetic population or the flow of genes is quite significant (Onaga et al., 2015).

They examined how diverse virulence is and how genetically related 122 isolates of the pathogen are from African countries. They looked at SNPs from data generated by genotyping by sequencing. They found 7 clades of the pathogen differing in virulence (Mutiga et al., 2017).

Chapter 3: Methodology

* Detailed description of the methods to be used, and how data will be analysed

***M. oryzae* isolates**

This study will focus on 50 isolates of *M. oryzae* from different locations in Africa and a few from Asia. They were sequenced under Illumina HiSeq at BGI. BGI took care of adapter trimming and removal of low-quality reads.

During the sequencing, the genomes of the isolates were cut into millions of small pieces, called reads. The genome of each isolate is recovered by mapping the reads against a known reference genome (<ftp://ftp.ensemblgenomes.org/pub/fungi/release-44/fasta/magnaporthe_oryzae/dna/Magnaporthe_oryzae.MG8.dna.toplevel.fa.gz>). This project focuses on the individual genome differences from the species-wide consensus represented with the reference genome, and these differences are called variants. Variants are detected by investigating the differences between all the mapped reads and the reference at one position.

**Single Nucleotide Polymorphisms (SNPs) discovery and variant calling**

Quality control will be performed on the reads using fastqc (v0.11.7).

Sequenced raw reads will be indexed and then aligned to a reference genome using bowtie2 (v2.3.4.1) which outputs SAM files which are text-based files (Langmead & Salzberg, 2012).

Samtools (v1.8) will then be used to convert the SAM files to BAM files (a binary counterpart of the SAM file). It is easier for computer programs to work with binary files. Samtools will also be used to sort and index the resulting BAM files. After aligning the reads to the genome, the alignments resulting from that are usually in a random order with respect to their positions in the reference genome. Any meaningful downstream analysis to be carried out later would require a further manipulation of the BAM file. The BAM file is sorted using Samtools, such that the reads are ordered on the basis of their alignment coordinates on each chromosome. After that, the genome sorted BAM file is indexed, to allow quick extraction of alignments overlapping particular genomic regions. Some viewers such as IGV also require that the genomes are indexed (Li et al., 2009).

BCFTOOLS (v1.8) will be used to perform variant calling and generate VCF files. They will then be filtered to produce a list of positions where at least one of the VCF files has a polymorphism (SNP) (Li, 2011a, 2011b)

IGV (v2.4.5) and IGVTOOLS (v2.3.98) will be used for whole genome viewing (Robinson, Thorvaldsdóttir, Wenger, Zehir, & Mesirov, 2017; Thorvaldsdóttir, Robinson, & Mesirov, 2013)

Genome analysis toolkit (GATK v4.0.10.1) and Picard (v2.8.2) tool will then be used for variant discovery and genotyping (Zhong et al., 2018). Duplicates are bad for variant calling. Picard is sued to detect and mark the duplicates. The resulting BAM file contains only one fragment from each duplicate group. The duplicate fragments are given a flag and so are not used downstream. This is done using Picard’s MarkDuplicates program. Local realignment is then performed to eliminate ambiguity of alignment at indel sites, around adjacent SNPs and around homo-polymer runs flanked by adjacent SNPs. It is performed using RealignerTargetCreator and IndelRealigner programs both of which are found in GATK (v3.7.0). The mate information in BAM is then fixed using the FixMateInformation program in Picard (Heldenbrand et al., 2018) – for citing GATK only, yet to find a citation for Picard.

Using snakemake (v5.7.0) to run variant calling. It will apply the Picard and GATK tools. A config.json file is used to define the path to the data that will be manipulated by the defined tools (Köster & Rahmann, 2012).

**Phylogenetic analysis**

Construct a phylogenetic tree of the SNP data using BioNJ, adistance based phylogeny reconstruction algorithm (Gascuel, 1997). The tree will be a pseudo alignment made of SNP positions.

**Population genetics and genomics**

Principal component analysis (PCA) - it is a method of summarizing data by reducing the dimensionality of a data set consisting many variables correlated with each other, wither heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.

STRUCTURE analysis – structure (v3.2.2) is a program that uses multi-locus genotype data (SNPs, SSRs. AFLPs and RFLPs) to assign individuals to a population to a population. One of the outputs from STRUCTURE is the Q matrix, which gives a probability that an individual belongs to a subpopulation (Pritchard, Stephens, & Donnelly, 2000). The Q matrix is then incorporated as a fixed effect in the “Unified-Mixed Model” for association analysis (Yu et al., 2006)

Tip calibration dating - the molecular dating of phylogenetic trees is a growing discipline using sequence data to co-estimate the timing of evolutionary events and rates of molecular evolution. Using the Bayesian Evolutionary Analysis by Sampling Trees (BEAST) software. BEAST (v1.8.3) (Suchard et al., 2018).

Work plan

Work plan - in a Gantt chart format

* Showing the year/months that the research activities will be carried out

|  |  |
| --- | --- |
| **Activity** | **Timeline (Month and year)** |
| Developing project concept | 2nd Sept, 2019 to 30th Sept, 2019 |
| Proposal writing | 1st Oct, 2019 to 31st Oct, 2019 |
| Proposal presentation at Pwani University | 4th to 8th Nov, 2019 |
| Data analysis | 11th Nov, 2019 to 31st Mar, 2020 |
| MSc. Dissertation writing | 1st Apr, 2020 to 10th June, 2020 |
| MSc. Dissertation presentation at Pwani University | 22nd to 26th June, 2020 |
| Corrections/ revisions of MSc. Dissertation and writing of manuscript to be submitted to a journal | July to Sept 2020 |

Budget

References

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> ## for citing fastqc tool kit