CHAPTER I

Introduction

1.1. Background

Viruses rank among the most numerous biological agents found across a wide range of ecosystems, often outnumbering their hosts by a substantial margin. The diversity is characterised by a large genetic diversity, with many viruses being discovered annually due to advancements in sequencing technology and the ability to assemble viral genomes directly from metagenomes without cultivation (Lobb et al., 2023). Among the viral groups, giant viruses, which refer to the viruses from the phylum Nucleocytoviricota, challenge the understanding of viruses due to their large size and genetic complexity. As the name suggests, giant viruses are characterised by their large size, comparable to that of bacteria and archaea. They also possess huge genomes, comprising a wide range of genes that they have acquired through continuous genetic transfer involving diverse viral and cellular lineages (Rodrigues et al., 2016).

The Nucleocytoviricota phylum comprises a monophyletic and heterogeneous group of viruses with double-stranded DNA, associated with infection across a wide range of hosts, including metazoans, protists, and mammals (Queiroz et al., 2023). The first of its kind, *Mimivirus*, was discovered in 2003, and subsequently, various giant viruses have been identified (Athira & Antony, 2023). These newly discovered giant viruses have been classified into several putative new viral groups. The research by Pitot et al. (2024) concludes that the most recent taxonomy of giant viruses identifies two distinct classes of giant viruses. Megaviricetes is the first group, encompassing the orders *Imitervirales*, *Pandoravirales*, *Pimascovirales*, and *Algavirales*. In contrast, the Pokkeviricetes group comprises the orders *Asfuvirales* and *Chitovirales*.

The expansion and findings of new taxa may be attributed to the giant virus genes. It may have expanded through repeated gene duplications and mutations throughout evolutionary time. Most giant virus groups share key biological features, particularly in their replication cycle, which occurs within the cytoplasm of host cells during infection (Talbert et al., 2023). Additionally, they encode proteins essential for virus morphogenesis. Among them, genes linked to various microbial metabolic pathways, such as photosynthesis and the tricarboxylic acid (TCA) cycle, were commonly identified, which implies that giant viruses potentially affect the metabolic capacities of their hosts. However, these genes still require further exploration, as their exact functional potential remains unknown (Belhaouari et al., 2022).

Due to their unique characteristics and expanding diversity, many questions have been raised about their origin and evolution (Aherfi et al., 2018). The formation of taxa within this viral group became more frequent and rapid, making it difficult to determine the evolutionary relationships between the giant virus groups. This confusion highlights the need to establish clear criteria and standardised methods for identifying specific giant virus taxa. Traditional genomic studies often rely on a single reference genome to represent an entire species. However, genetic differences can still be observed even at the strain level, underscoring the need for more comprehensive approaches to understand the diversity of giant viruses (Matthews et al., 2024).

One of the recently discovered methods is the creation of pangenome models, which represent the complete set of genomic elements in a given species or clade. The pangenome contrasts with the reference-based genomic approach, which focuses more on the sequence of a specific genome (Rodrigues et al., 2022). Pangenomes are particularly important in giant viruses, as their genome plasticity and diversity make this method a crucial tool for comparative analysis among families, extending to the species level (Sun & Ku, 2021). Genes from a particular order or family are the primary focus of standard pangenomic analysis, which also identifies a core (genes found across all strains), a shell (genes found across multiple strains), and a cloud (genes found in a few strains) pangenome. (Eizenga et al., 2020). This approach allows for a better understanding of shared genes across groups and helps distinguish each of the giant virus groups.

Therefore, this research aims to investigate several aspects of the evolution, diversity, and distinctive characteristics of these gigantic viruses by constructing gene-sharing networks and performing pangenome analysis among giant virus taxa, based on the presence and absence of genes within different groups. Herein, the functional pangenome will be constructed based on the reannotation of metagenome-assembled genomes (MAGs) from the giant virus metagenome, collected from various marine and soil locations, with comparisons to existing complete genomes and MAGs from publicly available databases. Furthermore, it will feature phylogenetic reconstructions of the genes that comprise the core genome of these taxa, along with a detailed analysis of the unique genes identified in the sample, presented as a heatmap. Any other unique genes found in the sample will be annotated and analysed to determine their novel existence in giant viruses. This research will contribute to a deeper understanding of giant virus diversity and their genomic diversity, which could thus be extended to reveal their ecological roles and evolutionary significance.

1.2. Objective

The primary objective of this research is to enhance the insights into the diversity, evolutionary patterns, and functional features of giant viruses (Nucleocytoviricota) by examining giant virus metagenome-assembled genomes (GVMAGs) through pangenome analysis, presence-absence mapping, gene-sharing networks, and phylogenetic relationships, thereby clarifying their classification and evolutionary trajectories.

1.3. Hypothesis

Based on the research background and objectives, distinct gene-sharing patterns are hypothesised to reveal the relationship and taxonomic diversity among specific giant virus taxa. The functional pangenome analysis is also hypothesised to identify a set of core genes and highlight the unique genes that may contribute to their evolutionary and novel functions.