STUDIES ON THE GENOME BIOLOGY AND EVOLUTION OF *ACANTHAMOEBA*

by

Morgan J. Colp

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2024

Dalhousie University is located in Mi'kma'ki,
the ancestral and unceded territory of the
Mi'kmaq. We are all Treaty people.

For Navi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Acanthamoeba castellanii has the potential to be a model organism for studying genome evolution and lateral gene transfer in eukaryotes. However, its complex and dynamic genome has posed a challenge to realizing the full potential of this system. To lay a foundation for developing *A. castellanii* as an experimental model, and to gain insight into eukaryote genome biology as a whole, I undertook a multi-pronged investigation. Generating high-quality reference genome sequences of *A. castellanii* strains Neff and C3 revealed inter-strain variation in gene content that conferred different functions to each strain. These chromosome-scale assemblies also allowed genome structure and organization to be predicted, demonstrating an inferred karyotype of 35 chromosomes for both strains, ranging from 100 Kbp to 2.5 Mbp in size. Artificial transformation experiments were performed to investigate how the *A. castellanii* genome responds to foreign DNA. Using nanopore sequencing and molecular biology experiments, a potential mechanism for transgene maintenance was elucidated, where incoming foreign DNA is tandemly duplicated and telomeres are added to the ends. This nascent linear molecule is maintained as a minichromosome bearing the transgenes, while also allowing for chromosomal integration. A large amount of long- and short-read sequence data were generated from genome assembly and analysis of transformation. These sequence read sets were used to better understand the ploidy in *A. castellanii*. Aneuploidy was suggested by the data, with significant variation in ploidy signal across chromosomes within an isolate, as well as across isolates. The predicted proteomes generated for the new *A. castellanii* Neff and C3 genomes were used to conduct a phylogenomic analysis for lateral gene transfer in these two strains. This analysis revealed a trend toward lateral gene transfers that may increase metabolic flexibility, and demonstrated how *A. castellanii* may be taking advantage of the genetic capabilities from organisms in its community to thrive in the environment. Collectively, these findings provide a deeper understanding of *Acanthamoeba* genome biology, and suggest genome biology across eukaryotes may be more dynamic than previously thought.

LIST OF ABBREVIATIONS USED

ARS	Autonomously replicating sequence
BUSCO	Benchmarking Universal Single-Copy Orthologs
dsDNA	Double-stranded DNA
EGFP	Enhanced green fluorescent protein
EGT	Endosymbiotic gene transfer
FISH	Fluorescent in situ hybridization
Gbp	Gigabase pairs
gDNA	Genomic DNA
GO	Gene ontology
Kbp	Kilobase pairs
LGT	Lateral gene transfer
Mbp	Megabase pairs
MBP	Mannose-binding protein
MW	Molecular weight
NCLDV	Nucleocytoplasmic large DNA virus
PAS	Page's amoeba saline
PCR	Polymerase chain reaction
PFGE	Pulsed-field gel electrophoresis
qPCR	Quantitative PCR
rDNA	Ribosomal DNA
SDS	Sodium dodecyl sulfate
SMRT	Single-Molecule Sequencing in Real Time
SNP	Single nucleotide polymorphism
SSC	Saline-sodium citrate
T-DNA	Transfer DNA

ACKNOWLEDGEMENTS

Over the course of my work on this thesis, there have been challenging or even demoralizing moments where it felt like valuable results were out of reach, but it is exceptionally gratifying to see all the products of my research brought together in this document. The growth I have experienced as a scientist and a person these last seven years is irreplaceable.

First, I express deep gratitude to my supervisor, John Archibald, for providing countless opportunities to grow and learn as a scientist and thinker, whether through research projects, facilitating collaborations, making it possible to attend several international conferences, or simply through our discussions over the years. Thank you to my supervisory committee, Andrew Roger, Claudio Slamovits, and John Rohde, for valuable advice and for giving me a gentle push in the right direction when it was needed to keep me on track. To Jon Jerlström-Hultqvist, nobody was as influential as you in the way I approach actually 'doing science' and thinking about research questions on a day-to-day basis in the lab. Thank you to Gina Filloramo for years of being an unwavering mentor, friend, and support during your time in the lab. Thank you to Marlena Dlutek for keeping the lab running, for technical help in the lab, and for making sure I stay on top of all of the logistical tasks and paperwork that I have a tendency to forget about. Thank you to Dr. Michael Gray, who has always shown me respect, warmth, and encouragement when we have crossed paths.

To all members of the Archibald lab and the ICG community, I couldn't have asked for a more welcoming and supportive environment to have done this research. Thank you to Cédric Blais for being a thoughtful and diligent student, and now a friend and collaborator, and for pushing my intellectual boundaries in both cases. A special thank you to Shannon Sibbald, Kelsey Williamson, and Shelby Williams, who have more or less been on this journey with me from the start. This would have been an entirely different experience without you, and I don't know how I would have persevered if you weren't all persevering alongside me.

Thank you to everyone who has supported me from outside of the academic sphere. Thank you to my weightlifting coach Augie Westhaver and my teammates, especially Ava for her relentless positivity, encouragement, and support in and out of the gym. Having a space outside of my graduate studies to be an athlete, pursue goals, compete, learn, and be surrounded by a supportive community has been incredibly important for getting to this point.

Thank you to Josh Lowery and Jaden Berckmoes for putting up with all of the times I've disappeared for months at a time when I've been swamped, and welcoming me back like I'd never left.

Finally, thank you to my family for unconditional love, support, and belief in me, in this endeavour and through my whole life, including Sharon, Bruce, Will, Liam, Ethan, Kristian, Jiawen, and especially my parents. I certainly could not have achieved this without you.

CHAPTER 1 INTRODUCTION

This thesis had its genesis in the pursuit of an ambitious goal, one that ultimately eluded me: to develop *Acanthamoeba castellanii* as a model system for studying lateral gene transfer (LGT) in the lab. There was, and to a lesser extent still is, a large discrepancy between the number of convincing cases of LGT inferred from genome sequences, and our understanding of the mechanisms and frequency of such gene transfer events into eukaryotic organisms. For some authors, this knowledge gap has been a deal-breaker in accepting the existence or importance of this phenomenon in eukaryotes at all[1–4].

To support the notion of LGT as an important force in shaping eukaryote evolution, and more importantly to characterize this process, how it happens, and how often, a group of researchers in the Archibald lab at Dalhousie University began to develop a selection of microbial eukaryotes into experimental model systems. I took responsibility for one of the chosen protists, *Acanthamoeba castellanii* strain Neff. The goal of my project as originally conceived was to shore up genomic resources for *Acanthamoeba*, become comfortable with the genetic tools that were available, test various culturing conditions, and then perform experiments where exogenous DNA was presented to *Acanthamoeba* in different ways in order to observe and characterize cases of LGT using nanopore sequencing supplemented by molecular biology experiments.

The planned experiments for observing LGT in the lab were inspired by researchers seeking to answer similar questions about endosymbiotic gene transfer (EGT) from the chloroplast to the nucleus in tobacco plants[5]. In these experiments, chloroplast genomes were transformed with a kanamycin resistance gene that was controlled by a

cauliflower mosaic virus 35S promoter known to drive strong constitutive expression in the tobacco nucleus, but negligible expression in the chloroplast. After several rounds of plant regeneration, leaf cuttings from the resulting plants could be screened with kanamycin to detect EGT. The authors of this study were able to perform subsequent experiments and analyses based on plant crosses and factors such as the number of cells exposed to biolistic transformation to estimate the frequency of EGT.

To adapt this concept for studying lateral gene transfer, my plan was to use plasmids with a neomycin resistance marker driven from a native, constitutive *Acanthamoeba* promoter. Instead of the step where kanamycin resistance was transformed into the chloroplast in the tobacco EGT experiments, the goal was to present an antibiotic-resistant plasmid to the amoebae externally in a variety of forms. Some examples include feeding the amoeba with *E. coli* that were maintaining the plasmid at high copy number, introducing the plasmid freely into amoeba cultures as an analog to environmental DNA, and co-culturing *Acanthamoeba* with *Agrobacterium* that has been engineered to contain our neomycin resistance expression cassette within its T-DNA. After a period of exposure to the foreign DNA, the amoebae would be subjected to selection with G418 to screen for transfers. Recognizing the frequency of transfer would be low, I was prepared for the need to drastically scale up such experiments to improve the chance of getting results.

After several months of experimentation using the *E. coli*-based foreign DNA as food I was left with no clear, actionable results to build from. At that point it was decided that the resources prepared for this project should be applied in parallel to lower-risk experiments aimed at improving the general understanding of *Acanthamoeba* genome

biology. The first such opportunity stemmed from my earlier steps taken to develop *Acanthamoeba* as an experimental model. To ensure any transfers that did occur could be properly characterized, I had re-sequenced and assembled the *Acanthamoeba castellanii* strain Neff genome using long- and short-read methods. The original reference genome sequence was highly fragmented and would not be suitable for detecting and characterizing LGT in the lab, especially with respect to a re-sequencing approach. Investing more time in this direction seemed likely to bear fruit in terms of genome assembly quality, a more comprehensive predicted proteome, and the opportunity to perform phylogenetic analysis to detect genes that may have been acquired by LGT. The other promising opportunity stemmed from transformation experiments I had done to ensure the construct intended for the LGT experiments behaved as expected in vivo. Characterizing the interaction between the *Acanthamoeba* genome and this foreign DNA was an opportunity to skip over the cell biological barriers to 'real' LGT and elucidate some of the genome biology that may be involved in the process.

1.1.1 Beyond 'typical' eukaryote genetics and genome biology

To guide the continuing study of eukaryote genome biology and evolution, it is helpful to first recognize some of the generalizations we currently make about eukaryotes and their genetics, especially in contrast with prokaryotes, but also in a vacuum.

In large part, sparse characterization of eukaryotic genome biology across the breadth of eukaryote diversity is responsible for our incomplete understanding of eukaryote LGT mechanisms. Beyond the most essential information processing pathways, features of eukaryote genome biology are known from phylogenetically

restricted subsets of the eukaryote tree of life. An overwhelming majority of eukaryotic diversity falls outside of the widely known clades containing animals, plants, and fungi, emphasizing the evolutionary and ecological importance of the remaining taxa, collectively referred to as 'protists'[6]. As 'omic' technologies rapidly advance, researchers have been characterizing an increasing number and diversity of protists. However, protist research has predominantly focused on systematics, comparative genomics, metabolism, ecology, and cell biology. In other words, researchers are interested in how protists are related, how they make a living, in which ways they resemble and differ from their nearest relatives on both a phenotypic and genotypic level, and how they impact their community in nature. One largely overlooked thread of investigation has been which genome maintenance pathways are shared by all eukaryotes and how different lineages have diverged from this core, either through innovation or differential loss. These pathways are likely to be highly relevant to the acquisition and use of foreign DNA by eukaryotes and understanding them is essential to understanding the process as a whole.

1.1.2 Amoebozoan genome biology

For many molecular biologists, the first thing to come to mind at mention of amoebozoan genomes is the famously extreme genomes in *Amoeba proteus*, thought to contain over 500 chromosomes[7], and *Polychaos dubium* (formerly *Amoeba dubia*), thought to have a total DNA content per cell of roughly 670 Gbp[8]. It is worth noting that this estimate of 670 Gbp is commonly described as the genome size of *P. dubium*, but without knowledge of the ploidy of this organism, it cannot be determined whether it corresponds to the length of the *haploid* genome, as is typically the case when discussing

genome size. A high level of polyploidy could explain this admittedly staggering amount of DNA per cell without a similarly unprecedented haploid genome size.

Based on our current knowledge, it is likely that genomes in the supergroup Amoebozoa are rich in additional variations on what we would think of as 'typical'. While not all examples are as striking as the first ones mentioned here, they all still serve to place bounds on the realm of possibilities, or perhaps more accurately to expand the boundaries of what was previously thought to be possible. The first sequenced amoebozoan, *Dictyostelium discoideum*, brings a few of these to the table; it maintains its ribosomal DNA (rDNA) as a high-copy extrachromosomal circle with a chromosomal master copy, rather than as expansive tandem arrays of the rDNA operon in subtelomeric regions as is found in common model animals, fungi, and plants[9]. It has also replaced canonical telomeric repeats with a short repeat of its own to serve the same purpose, potentially due to strong AT bias in its genome[9].

Two other sequenced amoebozoans, *Acanthamoeba castellanii*[10] and *Entamoeba histolytica*[11], appear to exhibit polyploidy, estimated at ~25*n* and variation within the range of ~4*n* to 40*n*, respectively. *E. histolytica* also seems to show marked variation in length of homologous chromosomes across isolates of the species, although the mechanism of this variation has not been identified with certainty, and it seems to have also replaced its telomeres, potentially with tRNA tandem arrays. Notably, although genes encoding parts of the meiotic machinery have been detected in *A. castellanii* and *E. histolytica*[12], these are thought to be asexual like fellow polyploids *A. proteus* and *C. chaos*, while the haploid or diploid dictyostelids do appear to undergo sexual reproduction[12–15].

To date, efforts to produce reference genome sequences for amoebozoans have largely been driven by some other impetus than purely genomic exploration; pathogenicity, anaerobiosis, endosymbiosis, and sociality/multicellularity are among the features of particular interest in species that have been sequenced so far[9,16–18], each providing sufficient motivation to overcome the complexity of amoebozoan genomics. It is possible, though, that the intriguing aspects of each of these examples has also led to evolutionary pressures on their genomes that would differ from those experienced by a 'generic' free-living amoebozoan, thus potentially eliminating each one as a representative free-living amoebozoan genome.

The genus *Acanthamoeba* contains facultative human pathogens and ubiquitous inhabitant of natural and artificial environments[19]. *Acanthamoeba castellanii* strain Neff, a laboratory-cultured representative of this genus, is one of the few amoebozoans with a published genome sequence and analysis[20]. Despite bearing some of the features mentioned above (pathogenicity[21], anaerobiosis[22], and a range of endosymbionts[23–26]), *Acanthamoeba* spp. participate in those lifestyles facultatively, and are commonly found as a phagotrophic, free-living amoeba in soil and freshwater environments that are generally oxic[27,28]. Additionally, when grown in culture for laboratory study, conditions are oxic, although it is conceivable that localized microoxia may arise in cultures with high cell density in the absence of agitation. Arguably, this makes *Acanthamoeba* our best current representative of free-living amoebozoan genomes, recognizing that there is no guarantee it is not exceptional in one or more ways. With no better alternative, though, broadening and deepening our understanding of the *Acanthamoeba* genome could serve as a valuable foundation for generalizing genome biology within this supergroup, which

could be subsequently amended as we amass more information from a broader swath of amoebozoan diversity. Having this baseline understanding of amoebozoan genomes could facilitate future analyses of other amoebozoans by reducing the number of surprising and potentially confusing results that would need to be interpreted.

1.1.3 Variations on ploidy across eukaryote diversity

In a discussion of ploidy as it appears across the tree of eukaryotes, haploidy would be a natural starting point. Throughout this thesis, 'haploid' will be used as a synonym of 'monoploid' meaning bearing a single set of chromosomes, rather than its alternative definition, bearing the same number of chromosomes as a gamete. In the field of eukaryotic microbiology, where sexual cycles can look very different than in well-known multicellular organisms, appear to be entirely absent, or not be known at all, this usage tends to be the norm. While the dominant life stage of sexual eukaryotes is frequently not haploid, there is a haploid stage in most eukaryotic species; the typical sexual cycle involves alternation between haploid and diploid life stages.

Probably the most well-recognized deviation from this paradigm would be the extensive polyploidy observed among plant species. It has been estimated that more than 70% of flowering plant species have undergone an increase in ploidy level at some point in their evolutionary history[29], and polyploidization is thought to be one of the drivers of plant speciation[30]. There are two main categories of polyploids as seen in plants, autopolyploids that arise from reproduction between closely related individuals and allopolyploids that arise from hybridization between taxa that have diverged from one another[31]. A primary mechanism for the generation of both types is the fusion of gametes

that were not properly reduced in meiosis such that they now carry the same number of chromosome sets expected of somatic cells[32,33]. Generally, while there can be some instability in the ploidy level of newly arisen polyploids, established polyploids will go on to maintain their new ploidy level in much the same way as they had gone about maintaining their previous one.

While polyploidy is not found in animals at nearly the scale it is in plants, examples do exist[34]. The fishes appear to have a long history of polyploidy, with ancestral genome duplications dating back to the origin of ray-finned fishes, and even prior to the divergence of tetrapods from fishes[35]. The last common ancestor of the fishes was diploidized following this event, as were the ray-finned fishes; that is, much of the redundancy created by polyploidy was eliminated, and chromosomes returned to segregating as homologous pairs. Moving toward the present day, there have been polyploidization events observed at the order, family, genus, and species level across various fish lineages. Some notable examples include tetraploid goldfish and carp found in the otherwise diploid Cyprinidae[36], white sturgeons appearing to be ancient octoploids[37], and tetraploidy in the common ancestor of Salmonidae, which have since re-diploidized[34]. Other animal groups exhibiting polyploidy are reptiles, amphibians, and insects as well as many other invertebrates[34,38]. Like in plants, at least some of these instances have been attributed to the fusion of unreduced gametes[37]. There are also cases of polyploidization in somatic cells of otherwise diploid animals, including in mammalian liver cells[39].

Moving away from the relatively ‘orderly’ polyploidy in plants and animals, additional examples can be found across eukaryote diversity. Fungi were previously not

thought to be especially prone to polyploidy, but it is now known to have occurred on several occasions across fungal diversity[40]. Strains of some well-known species such as *Saccharomyces cerevisiae*, *Candida albicans*, and *Cryptococcus neoformans* from diverse environments have been found to be polyploid but unlike in animals and plants, many of the observed polyploidization events are thought to have happened asexually. Many fungi also demonstrate aneuploidy, where ploidy level is not the same across all chromosomes, and there are some species where this appears to be a normal part of their life cycle[41]. Some other examples of polyploidy scattered across eukaryotes can be found in brown algae[42,43], multicellular red algae[44], diatoms[45], and oomycetes[46], all of which undergo a sexual cycle which appears to have been a contributing factor to their changes in ploidy level.

Finally, among microbial eukaryotes, a variety of more dynamic life cycles can be found with respect to nuclear DNA content and how it changes. A common feature of these life cycles is significant variation in the DNA content per nucleus or per cell beyond the typical fluctuations that occur over the course of standard meiotic and mitotic processes[47]. In more precise terms, these organisms either synthesize additional full sets of chromosomes in a process known as endoreplication, or they undergo genome amplification where only select portions of the genome are replicated. There is somewhat more variety in the ways different organisms reduce their nuclear DNA content back to its original level. In the case of endoreplication and reduction, this phenomenon is known as cyclic polyploidy.

A couple of well-known examples have already been mentioned briefly above: the amoebozoans *A. proteus* and *E. histolytica*. In *A. proteus*, the DNA content is amplified

during interphase to roughly three times the original amount, then the excess is eliminated in a process called chromatin extrusion prior to the next round of mitosis[48]. In *E. histolytica*, DNA content can vary by about 10-fold over the course of the life cycle, which appears to be the result of very loose synchronization between DNA replication, nuclear division, and cell division. The genome can replicate multiple times before any nuclear division occurs, and nuclei can divide several times before cytokinesis occurs, which can give rise to either uninucleate or multinucleate polyploid cells[11]. It does not appear that any special mechanism of reduction is required in this case.

There are some other quite complex life cycles found within Rhizaria that involve massive DNA replication, and subsequent massive proliferation of daughter nuclei to generate large numbers of gametes[49]. This has recently been characterized in the foraminiferan *Allogromia laticollaris*, where most of the life cycle is spent in an endoreplicated state, with only brief moments of haploidy and diploidy[50]. This organism resets back to true haploidy by a process called *Zerfall* where the large nucleus containing the highly endoreplicated genome (over 10,000-fold) dissolves, releasing its contents into the cytoplasm to be repackaged into much smaller haploid nuclei. The parent cell is then partitioned into many gametes, each with one of these haploid nuclei.

1.1.4 *Acanthamoeba* and its diverse bacterial and viral associates

Free living amoebae, including *Acanthamoeba*, are famous for harbouring both viruses and diverse bacteria as intracellular symbionts[51]. In *Acanthamoeba*, the first confirmation of bacterial endosymbionts came from a 1975 study by Proca-Ciobanu and colleagues[52]. This discovery was made through observation of the intracellular bacteria in

transmission electron micrographs, but a taxonomic identification was not made. The endosymbionts in this *Acanthamoeba* strain were found to be cytoplasmic and not confined by a vacuole. A decade later, Hall and Voelz published the next observation of intracellular bacteria in *Acanthamoeba*[53], although they report that these endosymbionts had been present in this strain at the American Type Culture Collection dating back at least six years. Taxonomic identification was also not made in this study.

By the 1990s and 2000s, it was becoming clear that the various endosymbiotic bacteria in *Acanthamoeba* generally came from four bacterial lineages: Alphaproteobacteria, Betaproteobacteria, Bacteroidota, and Chlamydiota[54,55]. In each case, it appears that the *Acanthamoeba* symbionts are closely related to obligate endosymbionts of other organisms. On the side of *Acanthamoeba*, at the time of a review by Horn and Wagner in 2004, 25% of known clinical and environmental *Acanthamoeba* isolates contained obligate intracellular bacteria[54]. Horn and Wagner also observed that among the reported cases of bacterial endosymbionts in *Acanthamoeba*, the proteobacterial symbionts were always found directly in the cytoplasm while Bacteroidetes and most Chlamydiae were found in vacuoles.

In terms of specific bacterial genera or species identified as *Acanthamoeba* symbionts, well-known examples include *Candidatus* Caedibacter and *Candidatus* Paracaedibacter from Holosporaceae, the human pathogen *Legionella pneumophila* and related species, *Mycobacterium* spp., a couple of *Rickettsia*-like symbionts, and the chlamydial genera *Parachlamydia* and *Neochlamydia*[56]. *Mycobacterium avium* and *L. pneumophila*, known to be human pathogens, were found to have increased virulence and resistance to disinfection when amoeba-grown rather than grown in laboratory culture

broth, suggesting that amoebae may play a role in priming some bacterial pathogens of humans for infection[57]. Beyond their specific role in bacterial virulence, *Acanthamoeba* spp. and other free-living amoebae are increasingly being seen as a potential environmental reservoir for intracellular bacteria, where they provide a niche for bacterial replication, but may also protect the bacteria from external stresses[56,58].

The intracellular niche provided by amoebae may also play a role in shaping the evolution of these bacteria, especially through LGT. On more than one occasion, two unrelated bacterial species have been found simultaneously inhabiting an *Acanthamoeba* population[59,60]. A study[61] of *Rickettsiales* evolution in the context of *Acanthamoeba* symbiosis identified plasmids in amoeba endosymbionts for the first time, invoking the possibility of conjugation between bacteria within amoeba cells. This study also demonstrated evidence of ongoing LGT among amoeba symbionts within *Rickettsiales*, as well as evidence of past LGT between *Rickettsiales* and other amoeba symbionts. Many of the genes involved in these transfers were related to amoeba-symbiont interactions.

In its capacity as a 'professional host', *Acanthamoeba* has also played a supporting role in the study of viral diversity due to its association with several members of Megaviricetes, a viral class characterized by members with exceptionally large physical size as well as exceptionally large dsDNA genomes. Mimivirus, the first 'giant virus' to be discovered, was reported in 2003 to have been isolated from *Acanthamoeba polyphaga* found in a water tower sample[62]. Mimivirus was initially thought to be an intracellular bacterium due to its unprecedented 400 nm capsid size. A Mimivirus genome sequence published the following year revealed an equally impressive genome

size of 1.2 Mbp with over 1200 open reading frames[63].

Since this discovery, several additional closely related giant viruses have been discovered to the extent that the order Megavirales (now roughly equivalent to the class Megaviricetes) was proposed to contain them taxonomically[64]. Many of these novel viral strains were isolated from environmental *Acanthamoeba* isolates or through co-culturing methods using *Acanthamoeba*. Members of this class have continued to set records for capsid and genome size among viruses, with *Pandoravirus salinus* exhibiting a 2.5 Mbp genome[65], and *Pithovirus sibericum* having a capsid of around 1.5 μm in length[66]. Interestingly, *P. sibericum* and a close relative *Mollivirus sibericum* were found in permafrost and thought to be over 30,000 years old, but were possible to cultivate and study in *Acanthamoeba*[67]. Accompanying the discovery of these many viruses of *Acanthamoeba* have been discoveries of fragments of their genomes in the genomes of their hosts[68–70]. It has even been suggested that if the host range is broad enough, viruses could facilitate gene flow between eukaryote species. To add one more layer of complexity to this web of interactions, it is now known that viral factories of Mimiviridae, the regions of infected cells where virion production occurs, can in turn be parasitized by smaller viruses known as virophages, which also have left genomic footprints in their eukaryotic hosts[71,72]. Given the sophisticated network of interactions surrounding *Acanthamoeba* spp., it is not difficult to imagine how *Acanthamoeba*'s evolution could be shaped by these relationships through LGT or by other evolutionary pressures.

1.1.5 Lateral gene transfer as a ubiquitous force in eukaryote evolution

In a recent review, Keeling evaluated the body of literature reporting LGT into eukaryotic organisms and proposed a paradigm shift in the way we think about which mechanisms are most likely to be responsible for transfer of genes that become fixed in the recipient population[73]. While I do not intend to examine this argument here, the underlying discussion captures a shift in the prevailing view of the community on the frequency and importance of eukaryote LGT. To suggest re-consideration of eukaryote LGT mechanisms requires a general acceptance that eukaryote LGT *does* take place and that the community has ideas about how it might unfold, a fact which Keeling acknowledges. Indeed, while authors in the past have made concerted efforts to convince the eukaryote genomics community that LGT is unimportant in eukaryote evolution[1–4], no such argument has been published in the last six years, and arguments *for* the importance of eukaryote LGT have been rather more convincing[74]. In the absence of such vehement resistance, attention has turned from demonstrating *if* LGT appreciably influences eukaryote evolution to demonstrating how much it does, and in what ways, an idea which was reviewed thoroughly by Van Etten and Bhattacharya[75].

With respect to how LGT shapes eukaryote evolution, a selection of well-known examples can be quite illustrative. The transition of eukaryotes to anaerobiosis has become the flagship example due to its clear and dramatic phenotypic effect, and the ability to identify some of the crucial laterally transferred genes that made it possible[76–84]. Lateral gene transfers have been observed from bacteria to insects in complex symbiotic relationships[85,86], and from plants to insects as well[87]. The acquisition of nutrients via transporters and secreted enzymes has been facilitated by LGT in multiple cases,

including in the parasitic Microsporidia, and in the fungal-like oomycetes[88–91]. In plastid acquisition and algal evolution, there is of course a substantial contribution from EGT, but it appears that additional lateral transfer from bacteria had important contributions to establishing the primary and higher order symbioses, and a similar pattern is also seen in the integration of the cyanobacterium-derived chromatophore of *Paulinella chromatophora*[92–97]. These diverse examples demonstrate some of the possibilities for innovation in the evolution of eukaryotes via lateral gene transfer, and two recent reviews explore this topic in even greater detail than the summary presented here[98,99].

1.1.6 The advent of third generation sequencing

Genome projects from the first two generations of DNA sequencing technology could arguably be described as two extremes on a spectrum, where one end represents maximum throughput and the other represents maximum structural resolution. Next-generation sequencing can be found on the end of maximum throughput, with massive parallelization but very short reads that retain very little information about their genomic context. On the end of maximum structural resolution falls the first generation of genome sequencing (i.e. Sanger sequencing) and assembly. Of course, this is not intrinsic to the sequencing method itself; Sanger sequencing typically only achieved read lengths of hundreds to a thousand bases. However, during the era of genomics that relied on this sequencing technology, additional time- and labour-intensive efforts were made to finish genome assemblies. These efforts used methods like chromosome walking (iteratively sequencing from a known sequence into unknown adjacent sequence) and chromosome mapping (using methods like Southern hybridization to determine which genes or

scaffolds belong to the same chromosome) to sequence missing regions and provide structural context[100].

In theory, the same finishing methods could be done as a part of next-generation genome sequencing projects, but these often trended toward draft genome sequences that maximized the number of genes that could be analyzed, at the expense of investing in acquiring structural information about the genome[101,102]. Next-generation sequencing reads were simply not long enough to unambiguously assemble through any repeat content found in the target genome, especially in eukaryotes[103]. These draft genomes often ended up in hundreds to thousands of contigs. Not only does this make it difficult to gain a full structural understanding of the genome, but there are other issues such as the possibility of missing genes and the difficulty of distinguishing contamination from genuine genomic sequence.

In the mid-2010s, third generation sequencing technologies became available to consumers and over the following years, revolutionized eukaryote genomics. Third generation sequencing technologies currently comprise nanopore sequencing platforms developed by Oxford Nanopore and Single-Molecule Sequencing in Real Time (SMRT) developed by Pacific Biosciences. Each of these technologies offers average sequence read lengths in the thousands of bases, with nanopore sequencing reads having reached the megabase scale, while still being a massively parallel sequencing method. This combination of read length and throughput has enabled relatively straightforward assembly of highly contiguous, often near-chromosome-scale genome sequences from across the eukaryote tree of life, unlocking the ability to study genome structure, phase haplotypes, distinguish LGT from contamination, and have the most complete set of gene

annotations possible, all while being relatively accessible to most genomics researchers[104].

1.1.7 Thesis overview

While initially aimed at experimental characterization of eukaryote LGT, the goal of my thesis project ultimately was to amass as much information as possible about *Acanthamoeba* genome biology and evolution with two outcomes in mind. The first objective was to propel *Acanthamoeba* down the path toward becoming a fully established model organism of the sort that has been developed in plant and fungal systems. The second was to apply findings from *Acanthamoeba* to better understand eukaryote genome biology as a whole, especially on topics such as lateral gene transfer, genome organization, and dynamic life cycles. To this end, my research efforts have been dedicated to four distinct elements of *Acanthamoeba* genome biology and evolution.

In Chapter 2 I present high quality, chromosome-scale reference genome sequences of *Acanthamoeba castellanii* strain Neff and *A. castellanii* C3. This work was a collaborative effort involving the Archibald lab, Dr. Michael Gray from Dalhousie University, the Spatial Regulation of Genomes team and the Biology of Intracellular Bacteria team at the Institut Pasteur and the Franz Lang lab at Université de Montréal. These reference sequences allowed for a suite of analyses that investigated the nature of these genomes from different angles. A gene content comparison between the two strains revealed strain-specific sets of genes enriched in different functions, many of which are predicted to have facilitated adaptation to distinct environments. The analysis of Hi-C data provided a wealth of structural and organizational information about these genomes

that previously eluded researchers, such as demonstrating a haploid karyotype of 35 chromosomes, determining the genomic locus of the rDNA, characterizing chromatin loops and domains, and correlating these chromatin regions with elevated gene expression.

Chapter 3 unpacks a complex investigation into the fate of transgenes in *Acanthamoeba castellanii* strain Neff. Following artificial transformation experiments with circular and linear DNA, this chapter brings to bear data from several nanopore and Illumina sequencing runs and the results from a variety of molecular biology experiments to elucidate how the amoeba genome responds to the transforming DNA. Any given result in this study was too ambiguous to draw conclusions from; not until all the different threads were considered together could I hypothesize the most likely explanation for how *Acanthamoeba* maintains its transgenes. Early sequencing experiments revealed the presence of the plasmid and that it often formed tandem arrays of up to 11 copies, but no more information. A subsequent Southern blot experiment demonstrated that it existed on DNA species greater than 20 Kbp in length. I developed a single cell isolation protocol for the purpose of sequencing monoclonal transformant populations, which initially appeared to reveal hundreds of putative plasmid integration sites across the genome. However, this ended up being a cautionary tale about the frequency of read chimerism in nanopore sequencing experiments. While PCR experiments did confirm one genuine integration of the plasmid into the genome of one of the clonal lines, revisiting sequence data provided an additional, complementary hypothesis for its maintenance; within the sequence data from each clone was evidence for tandem arrays of plasmid sequence flanked by telomeres on each end. This ultimately

became the best explanation for all the data generated in this investigation. It appears that *Acanthamoeba* creates a minichromosome-type structure of plasmid concatemers flanked by telomeres that maintain transgenes extrachromosomally, with occasional chromosomal integration of the plasmid as well. This finding could be interpreted as a pathway that expands the window of opportunity for foreign DNA to become integrated in an LGT event, and could also be exploited as a molecular biology tool.

In Chapter 4, I co-opt the large amount of sequence data generated for Chapters 2 and 3 to expand on our understanding of *Acanthamoeba* ploidy. While these datasets were not suitable for absolute quantification of genome or chromosome copy numbers, allele frequencies of SNPs and structural variants from both nanopore and Illumina data were interpreted to reveal apparent aneuploidy in all studied isolates. In addition to appreciable variation in the ploidy signal across chromosomes within an isolate, there was also variation across isolates, even among clones isolated from the same population. This suggests that the regulation of the genome and karyotype in *Acanthamoeba* is quite dynamic and begs the question of what mechanisms might be responsible for such variation, but also what mechanisms might keep it in check so that the appropriate genetic material is inherited by all daughter cells. Cytogenetic experiments, e.g., chromatin staining and fluorescence in situ hybridization, will likely be required to fully characterize the ploidy in this organism and how it is managed.

Finally, in Chapter 5 I deviate from my more mechanistic inquiry into genome biology and take inventory of past events, namely instances of lateral gene transfer into both *Acanthamoeba* strains sequenced in Chapter 2, using the updated predicted proteomes from that work. I performed a comprehensive phylogenetic screen of all

predicted proteins from the Neff and C3 strains to detect putative LGTs and infer the donors and functions of these genes. I demonstrate overrepresentation of particular functional categories among the putative LGT genes of both strains, revealing a strong trend toward expanding metabolic capability and flexibility, as well as highlighting a few interesting acquisitions beyond these categories. Through a case study of a subset of the putative LGT genes, I was able to illustrate examples of how *Acanthamoeba* is likely to access foreign genes in its environment and exploit them to achieve greater ecological success.

While the focus of these chapters has been somewhat disparate, they are unified by their contribution to advancing our understanding of genome biology and evolution in *Acanthamoeba*. In its own right, this pursuit has been valuable in helping to develop this ubiquitous, medically and ecologically important amoeba into a model organism. However, beyond the confines of the *Acanthamoeba* genus, the lessons learned from this thesis project will influence the way we think about eukaryote genome biology and evolution as a whole by expanding our understanding of what is normal or possible in eukaryote genomes, and accelerating the discovery and characterization of such processes in other eukaryote lineages.

CHAPTER 2 A HIGH-QUALITY REFERENCE GENOME SEQUENCE OF *ACANTHAMOEBA CASTELLANII*

Portions of this chapter contain work presented in Matthey-Doret, Colp et al. 2022[105]. The results, analyses and discussion that relate to my contributions to that paper are clearly delineated throughout the chapter.

2.1 Introduction

An *Acanthamoeba castellanii* strain Neff reference genome sequence was published in 2013 by Clarke et al.[20], who presented an assembly, predicted proteins, and analyses of lateral gene transfer and the genetic complement involved in processes such as cell signalling, environmental sensing and response, cell adhesion, microbial recognition, antimicrobial defense, metabolism, and transcription regulation. While thorough and comprehensive, this genome project was limited by the sequencing technologies available at the time, namely Illumina and 454 next-generation sequencing which generate short reads, as well as some Sanger sequencing to provide some slightly longer reads. As a result, the assembly is quite fragmented relative to the expected number of chromosomes; past pulsed-field gel electrophoresis experiments [106] have estimated the number of chromosomes to be around 20, while the genome assembly contains 3192 contigs, which were placed into 384 scaffolds.

To pursue my present goal of investigating *Acanthamoeba* genome biology and its interaction with foreign DNA, such an assembly is not contiguous enough. Factors that are expected to be important for understanding these topics cannot be observed and analyzed with such a fragmented assembly. For example, transposable elements are thought to be important drivers of lateral gene transfer in eukaryotes due to their mobility, and there are examples supporting this hypothesis[107–114115 for review]. These elements can be

several kilobases in length and are often repetitive. The combination of these two factors poses a challenge when trying to resolve the position of transposable elements in the genome using short reads, which are a few hundred bases long at maximum, and by extension, poses a challenge to identifying any interaction between laterally transferred genes and transposable elements. A fragmented assembly also fails to provide a complete structural understanding of the genome. Some hypotheses on lateral gene transfer in eukaryotes include speculation about the relationship between genome structure and integration of foreign DNA; for example, recombination rates in several mammals, birds, and yeast are known to be higher in distal regions of chromosomes (nearer to telomeres)[116–122]. Higher levels of recombination are thought to facilitate acquiring foreign DNA, potentially causing laterally transferred genes to prefer chromosome ends over the rest of the genome, which is also supported empirically by some genome studies[107]. Fragmentation of the assembly makes it impossible to accurately identify regions of the genome that are near relevant structural features, therefore making it impossible to investigate whether they have an affinity for the integration of foreign DNA.

Third-generation sequencing technologies, namely nanopore sequencing offered by Oxford Nanopore and single-molecule real time sequencing offered by PacBio, have become mainstream over the period spanning from roughly 2017 to the present year, 2024. These technologies permit the sequencing of much longer reads (up to the megabase range) with a minor penalty to base pair-level accuracy, which has permitted a drastic improvement in our ability to resolve genomes at the structural level. While the very latest iterations of these sequencing methods and the associated bioinformatic tools

have largely overcome the deficit in base pair-level accuracy, it is still standard practice to also sequence the genome of interest with Illumina as well to perform fine scale correction on the structurally impressive assembly. To this end, I used the Oxford Nanopore MinION sequencing platform to produce a high-quality reference genome assembly for *Acanthamoeba castellanii* strain Neff.

During my early efforts to assemble the genome, I learned that another consortium was also sequencing this genome, for unrelated reasons. We then combined our efforts to eventually produce a near-chromosome-level reference genome sequence for *Acanthamoeba castellanii* strains C3 and Neff, with a number of interesting analyses to go along with it. These eventual collaborators comprised two groups from the Institut Pasteur: the Spatial Regulation of Genomes team, and the Biology of Intracellular Bacteria team. These teams were studying how spatial organization of the *Acanthamoeba* genome changes during infection by one of its best-known symbionts, the human pathogen *Legionella pneumophila*. The basis of their investigation was the use of the chromosome conformation capture method Hi-C[123] to characterize the genome-wide chromatin organization of *Acanthamoeba*, including chromatin loops, domains, and long-range intrachromosomal contacts as well as interchromosomal contacts. The chromatin state could be similarly determined at different time points post-infection and compared. An additional layer on top of this comparison was the identification of genes that were located where there were changes in chromatin conformation (especially loops and domains) and to correlate changes in gene expression with proximity to changes in chromatin organization. Altogether, this may elucidate the role of chromatin organization in the biology of *Legionella* infection.

Chromosome conformation capture methods originated with the 3C method invented in 2002[124]. This method and its successors are based on chemical crosslinking of genomic DNA, followed by a restriction digestion reaction, and then a ligation reaction, such that DNA fragments that were in close physical proximity in situ become ligated to one another. The crosslinking is subsequently reversed such that the nascent chimeric DNA fragments can be analyzed.

In the original 3C method, quantitative PCR is used to target particular sequences of interest and measure the frequency with which they are ligated to one another. Crosslinking frequency is known to be inversely correlated with the distance between two loci on the primary DNA sequence, so deviations from this pattern can indicate a spatial relationship of those two loci in situ. For example, yeast telomeres are known to cluster, so different chromosome ends are observed to form ligation products with one another more often than expected from their separation in the primary sequence. Hi-C combines the same underlying principles with next generation sequencing technology to greatly increase the throughput of these experiments, while also avoiding the need for locus-specific qPCR primers. The first generation of Hi-C was described in the literature in 2009[125] and incorporates biotinylation into the fragmentation and ligation process. This means ligation products can be subsequently pulled down to enrich the sequencing library for informative fragments. Two newer iterations of the Hi-C method have been described since: Hi-C 2.0[126] in 2017 and Hi-C 3.0[127] in 2021. These updated protocols serve to improve the resolution of the method by adjusting the choice of restriction enzymes, the choice of crosslinking reagents, the handling of samples, and the treatment of DNA fragment ends during labelling and ligation. The Hi-C method used in this study falls

within the Hi-C 2.0 generation of protocols.

To enable their planned experiments, our collaborators generated Oxford Nanopore long read data, Illumina short read data, and Hi-C data for *Acanthamoeba castellanii* strains Neff and C3. Both strains were used because Neff is the reference strain, while C3 is the one used for studying *Acanthamoeba-Legionella* interactions. Therefore, when we began our collaboration, I pooled my nanopore and Illumina data for the Neff strain with that of our collaborators, and a joint assembly was produced. Additional collaborators of ours from the Université de Montréal generated gene models for the assemblies. Our collaborators from the Institut Pasteur used this for their originally planned investigation, while, with assistance from Bruce Curtis in the Archibald lab, I contributed some assembly and gene model curation, as well as a high-level comparative genomic analysis.

2.2 Methods

2.2.1 Strains and growth conditions

The cultures maintained in the Archibald lab were grown at room temperature in Neff base medium with additives (ATCC Medium 712; 0.75% yeast extract, 0.75% proteose peptone, 2 mM KH_2PO_4, 1 mM $MgSO_4$, 1.5% glucose, 0.1 mM ferric citrate, 0.05 mM $CaCl_2$, 1 µg/mL thiamine, 0.2 µg/mL D-biotin, and 1 ng/mL vitamin B_{12}).

A. castellanii strains Neff and C3 were grown at the Institut Pasteur on amoeba culture medium (2% Bacto tryptone, 0.1% sodium citrate, 0.1% yeast extract), supplemented with 0.1 M glucose, 0.1 mM $CaCl_2$, 2.5 mM KH_2PO_4, 4 mM $MgSO_4$, 2.5 mM Na_2HPO_4, and 0.05 mM $Fe_4(P_2O_7)_3$ at 20 °C. The *L. pneumophila* strain Paris was

grown at the Institut Pasteur for 3 d on N-(2-acetamido)-2-amino-ethanesulfonic acid (ACES)-buffered charcoal-yeast (BCYE) extract agar at 37 °C.

2.2.2 Hi-C

Our collaborators also performed the requisite experiments to generate libraries for Hi-C and then sequenced them. Cell pellets were suspended in 1.2 mL H_2O and transferred to CK14 Precellys tubes. Cells were broken with Precellys (six cycles: 30 sec on/30 sec off) at 7500 rpm and transferred into a tube. All Hi-C libraries for *A. castellanii* strains C3 and Neff were prepared using the Arima kit and protocol with only the *Dpn*II restriction enzyme. Libraries were sequenced to produce 35-bp paired-end reads on an Illumina NextSeq machine. Statistics regarding Hi-C libraries are described in Supplementary Table 2.1.

2.2.3 Short-read sequencing

Short reads were generated for the purpose of polishing the long-read-derived assemblies. Illumina libraries associated with the data submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRX12218478 and SRX12218479 were prepared from *A. castellanii* strains C3 and Neff genomic DNA, respectively, and sequenced by Novogene at 2 × 150 bp on an Illumina NovaSeq 6000 machine.

To generate the data associated with SRA accession number SRX4625411, genomic DNA samples were obtained from A. *castellanii* strain Neff using an SDS-based lysis method, followed by digestion with RNase A, then Proteinase K, and then a phenol-chloroform-based extraction. A PCR-free library was prepared and sequenced by Génome Québec from purified *A. castellanii* strain Neff genomic DNA. The library was

barcoded and run with other samples on an Illumina HiSeq X Ten instrument, producing 150-bp paired-end reads.

2.2.4 RNA-seq

A. castellanii strain C3 RNA-seq libraries were prepared by our collaborators using the stranded mRNA TruSeq kit from Illumina and sequenced in single-end mode at 150 bp on an Illumina NextSeq machine. These libraries were poly(A)-selected and were prepared from purified *A. castellanii* total RNA.

For *A. castellanii* strain Neff (SRA: SRX7813524), total RNA was obtained using a standard TRIzol adherent cell protocol. A poly-A selected library was prepared and sequenced by Génome Québec. The library was barcoded and run with other samples on an Illumina NovaSeq 6000 instrument, producing 300-bp paired-end reads.

2.2.5 Oxford Nanopore sequencing

Our collaborators generated Oxford Nanopore long read datasets for Neff and C3 (SRA accession numbers SRX12218489 and SRX12218490). DNA was extracted from *A. castellanii* strains Neff and C3 using the Qiagen blood and cell culture DNA kit following the specific recommendations detailed by Oxford Nanopore Technologies in the information sheet entitled “High- molecular-weight gDNA extraction from cell lines (2018)” in order to minimize DNA fragmentation by mechanical constraints. Oxford Nanopore libraries were prepared with the ligation sequencing kit SQK-LSK109, flow cell model MIN106D R9. Base-calling was performed using Guppy v2.3.1-1 (https://nanoporetech.com/community).

I prepared additional Oxford Nanopore long read libraries for Neff, generating several sequence read sets. Genomic DNA samples were obtained from A. *castellanii* strain Neff using an SDS-based lysis method, followed by digestion with RNase A, then Proteinase K, and then a phenol-chloroform-based extraction. DNA samples were cleaned with Qiagen G/20 genomic clean-up columns using the manufacturer's protocol, but with double the number of wash steps. Four different libraries were prepared, using the SQK-RAD003 rapid sequencing kit (SRA: SRX4620962), the SQK-LSK308 1D2 ligation sequencing kit (SRA: SRX4620963), the SQK-RAD004 rapid sequencing kit (SRA: SRX4620964), and the SQK-LSK108 ligation sequencing kit (SRA: SRX4620965). The SQK-LSK308 and SQK-RAD003 libraries were sequenced on FLO-MIN107 flow cells, and the SQK-LSK108 and SQK-RAD004 libraries were both sequenced on a FLO-MIN106 flow cell. All four libraries were base-called with Albacore 2.1.7 (https://hub.docker.com/r/genomicpariscentre/albacore tags), as they were sequenced before the release of Guppy. Adapters were removed from the base-called reads using Porechop v0.2.3 (https://github.com/rrwick/Porechop).

2.2.6 Genome assembly

Oxford Nanopore reads were filtered using filtlong v0.2.0 (rrwick/Filtlong; https://github.com/rrwick/Filtlong [accessed July 12, 2022]) with default parameters to keep the best 80% of reads according to length and quality. Illumina shotgun libraries were used as the reference for filtering. A *de novo* assembly was generated from the raw (filtered) Oxford Nanopore long reads using flye[128] v2.3.6 with the default three iterations of polishing. The resulting assembly was polished using both Oxford Nanopore and

Illumina reads with HyPo[129] v1.0.1. Contigs appearing to be of mitochondrial origin were separated from the rest of the assembly to help avoid their inclusion into the nuclear genome during scaffolding. These contigs were identified as probably mitochondrial if >60% of their length aligned to the reference mitochondrial genome, or if the contig was at least 51% identical to the mitochondrial genome across its full length. Nuclear contigs were scaffolded with Hi-C reads using instaGRAAL[130] v0.1.2 with default parameters. instaGRAAL-polish was then used to fix potential errors introduced by the scaffolding procedure, and the final assembly was polished with the Illumina short read data using two rounds of Pilon[131] polishing. The resulting 'Neff-v2' assembly was curated manually to remove spurious insertion of mitochondrial contigs other contaminants into the scaffolds. The final assembly was polished again using Pilon with Rcorrector-corrected reads[132]. Minimap2[133] v2.17 was used for all long read or pairwise genome alignments, and Bowtie 2[134] v2.3.4.1 was used for short read alignments. Circos plots were generated using Circos[135]. To identify telomeric repeats, the ends of chromosome-scale scaffolds were manually inspected, and basic text searches for telomeric repeat sequences were also performed.

2.2.7 Genome annotation

Genome annotation for these two new high quality genome assemblies (i.e. C3 and Neff-v2) was performed by Matt Sarrasin, using a pipeline developed in the B. Franz Lang lab at the Université de Montréal, described briefly here. The structural genome annotation pipeline used here was implemented similarly as described previously[136]. Then, RNA-seq reads were mapped to the genome assembly using STAR[137] v2.7.3a,

followed by both *de novo* and genome-guided transcriptome assembly by Trinity[138] v2.12.0. Both runs of Trinity were performed with Jaccard clipping to mitigate artificial transcript fusions. The resulting transcriptome assemblies were combined and aligned to the genome assembly using PASA[139] v2.4.1. Protein sequences were aligned to the genome using Spaln[140] v2.4.2 to recover the most information from sequence similarity. The *ab initio* predictors used were AUGUSTUS[141] v3.3.2, Snap[142], Genemark[143] v4.33, and CodingQuarry[144] v2.0. Finally, the PASA assembly, Spaln alignments, and the AUGUSTUS, Snap, and CodingQuarry gene models were combined into a single consensus with EVidenceModeler[145] v1.1.1. I performed manual validation of the gene models in GenomeView, using mapped RNA-seq as a guide, and manually transposed gene models from Neff-v1 that could be detected in Neff-v2 but had not been automatically annotated.

Functional annotations were added by our Institut Pasteur collaborators using funannotate v1.5.3 (https://doi.org/10.5281/zenodo.1471785 [accessed July 12, 2022]). Repeated sequences were masked using RepeatMasker[146]. Predicted proteins were fed to Interproscan[147] v5.22, Phobius[148] v1.7.1, and Eggnog-mapper[149] v2.0.0 to generate functional annotations. Ribosomal RNA genes were annotated separately using RNAmmer[150] v1.2 with HMMER[151] 2.3.2. As described in the "Data access" section, the funannotate-based script "func_annot_from_gene_mod" used to add functional annotations to existing gene models is provided in the Zenodo record and on the associated GitHub repository.

2.2.8 Analysis of sequence divergence

To compute the proportion of substituted positions in aligned segments between the C3 and Neff strains, the two genomes were aligned by our Institut Pasteur collaborators using minimap2 with the map-ont preset and -c flag. The gap-excluded sequence divergence (mismatches/[matches+ mismatches]) was then computed in each primary alignment, and the average of divergences (weighted by segment lengths) was computed. This is implemented in the script "04_compute_ seq_divergence.py" available in the genome analyses repository listed at Zenodo (see "Data access").

2.2.9 Orthogroup inference

I used the predicted proteomes of both the Neff and C3 strains to infer orthogroups, with three other amoebozoans, *Dictyostelium discoideum*, *Physarum polycephalum*, and *Vermamoeba vermiformis*, as outgroups to improve the accuracy of orthogroup inference. The outgroup predicted proteomes were retrieved from PhyloFisher[152]. Both Broccoli[153] and OrthoFinder[154] were run with default settings for orthogroup inference.

2.2.10 Gene content comparison of Neff and C3 strains

I used custom Python scripts to retrieve genes unique to each *A. castellanii* strain, as well as orthogroups that were shared between the two strains. Genes were only determined to be strain-specific or shared if both Broccoli and OrthoFinder assigned them as such; genes were excluded from the analysis if these tools did not agree. For both strains, functional assignments for each gene ID were extracted from funannotate output

and tabulated. The tabulated assignments and strain-specific gene IDs were fed into the R package topGO (https:///bioconductor.org/packages/ topGO/ [accessed July 12, 2022]) to analyze GO term enrichment in the strain-specific genes. A Fisher's exact test with the weight algorithm was implemented in topGO for the Neff- and C3-specific genes for each of the three ontologies (biological process, cellular component, and molecular function). When building the GOdata objects for these three ontologies, nodeSize was set to 10 for both the biological process and molecular function ontologies and to five for the cellular component ontology to better suit the lower number of GO terms in this ontology.

2.2.11 Mannose-binding protein comparison

To determine whether the mannose-binding protein, a known *Legionella* cell entry receptor, was responsible for an observed difference in infectibility between Neff and C3, I compared the sequences of this protein from a handful of *Acanthamoeba* strains. I retrieved mannose-binding protein (MBP) amino acid sequences from three strains of *A. castellanii* (Neff, C3, and MEEI 0184) and one strain of *A. polyphaga*, aligned them using MAFFT-linsi[155] v7.475, and visualized them in Jalview[156] v2.11.1.3. The MEEI 0184 strain sequence was retrieved from the NCBI Protein database (accession number AAT37865.1), and the Neff and C3 sequences were retrieved from the predicted proteomes generated in this study with the MEEI 0184 sequence as a BLASTP[157] query. The *A. polyphaga* genome does not have a publicly available predicted proteome, so its MBP protein sequence was manually extracted from several contigs in the genome sequence (NCBI accession GCA_000826345.1) using TBLASTN with the MEEI 0184

sequence as a query (the sequence encoding the first eight amino acids of the protein could not be found in the genome as a result of a truncated contig). 2.2.13

2.2.12 Hi-C analyses

All Hi-C data were generated and processed by our collaborators at the Institut Pasteur. Reads were aligned with Bowtie 2 v2.4.1, and Hi-C matrices were generated using hicstuff v3.0.1 (https://zenodo.org/record/4066363 [accessed July 12, 2022]). For all comparative analyses, matrices were down-sampled to the same number of contacts using cooltools (https://www.github.com/mirnylab/cooltools), and balancing normalization was performed[158,159]. Loops and domain borders were detected using Chromosight[160] v1.6.1 using the merged replicates at a resolution of 2 kb. Intensity changes were measured in Chromosight scores during infection using pareidolia v0.6.1 (https://zenodo.org/record/5362241/export/json [accessed July 12, 2022]) on three pseudoreplicates generated by sampling the merged contact maps, as described previously[161]. This was performed to account for contact coverage heterogeneity across replicates. The 20% threshold used to select differential patterns amounts to 1.2% false detections for loops and 2.3% for borders when comparing pseudoreplicates from the same condition.

2.2.13 Data access

The sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA599339 and PRJNA487265. All processed data generated in this study, as well as

the assemblies and annotations used in this work, have been submitted to Zenodo repository under accession number https://zenodo.org/record/5507417. The code used to perform the analysis is packaged into the following Snakemake pipelines available as Supplemental Code and at GitHub: hybrid genome assembly, https://github.com/cmdoret/Acastellanii_hybrid_assembly; functional annotation of *A. castellanii*, https://github.com/cmdoret/Acastellanii_genome_annotation; analyses of genomic features in *A. castellanii*, https://github.com/cmdoret/Acastellanii_genome_analysis; and changes during infection by *Legionella*, https://github.com/cmdoret/Acastellanii_legionella_infection.

2.3 Results

2.3.1 Chromosome-level genome assembly

We used a combination of Illumina short reads, Oxford Nanopore long reads, and Hi-C to assemble the Neff and C3 genomes to chromosome scale, with 90% of the Neff genome contained within 28 scaffolds. This estimate is somewhat higher than the approximately 20 chromosomes inferred using pulsed-field gel electrophoresis[106]. For both the Neff and C3 strains, we first generated a raw *de novo* assembly using Oxford Nanopore long reads. To account for the error-prone nature of long reads, our collaborators polished the first draft assemblies with paired-end shotgun Illumina sequences using HyPo[129]. In the case of the Neff assembly, the long and short read sequence data I generated was combined with that of our collaborators prior to the assembly and polishing, while the C3 data came purely from our collaborators. The polished assemblies were then scaffolded with long-range Hi-C contacts using the probabilistic program instaGRAAL, developed by our collaborators at the Institut

Pasteur. This program exploits a Markov chain Monte Carlo algorithm to swap DNA segments until the most likely scaffolds are achieved[130,162]. Following the post-scaffolding polishing step of the program[130], the final genome assemblies displayed better contiguity (Table 2.1), completion, and mapping statistics than the previous versions, with the cumulative scaffold lengths quickly reaching a plateau (Fig. 2.1A). The latter observation is useful because it provides a relatively clear visualization of assembly contiguity that is more intuitive than N50 metrics and also more information rich. While N50 identifies a single break-point in cumulative length, the aforementioned plot captures the entire range of cumulative scaffold length. The assemblies of both strains are slightly longer than the existing Neff reference assembly ('Neff-v1'), with a smaller number of contigs (Fig. 2.1B). The BUSCO[163] completeness scores for both assemblies are also improved, with 90.6% (Neff-v2) and 91.8% (C3) complete eukaryotic universal single-copy orthologs, compared with 77.6% for Neff-v1 (Fig. 2.1C). An increase in the proportion of properly paired Illumina reads, from 71% for Neff-v1 to 84% for our new Neff assembly, Neff-v2, suggested a reduced number of short mis-assemblies. We found putative eukaryotic telomeric repeats ("TTAGGG") at the extremities of the Neff-v2 scaffolds and, to a lesser extent, at the extremities of C3 scaffolds, suggesting that some of these scaffolds indeed correspond to full-length chromosomes. Our collaborators demonstrated that this presence of TTAGGG repeats at the scaffold extremities is statistically significant (Supplementary Fig. 2.1). From this point, all subsequent analyses used Neff-v2, so this latest version of the Neff genome sequence will simply be referred to as 'Neff' for the rest of the Results section.

Table 2.1 Summary statistics of the two *Acanthamoeba castellanii* reference genomes produced in this study compared to the original Neff strain genome

Assembly	Neff-v1 (Clarke et al. 2013)	Neff-v2 (this study)	C3
Genome size (Mbp)	42.0	43.8	46.1
No. of scaffolds	384	111	174
No. of Ns (Mbp)	2.6 (6.1%)	0 (0%)	0 (0%)
N50 (Mbp)	0.3	1.3	1.4
Largest scaffold (Mbp)	2.0	2.5	2.4
GC%	57.90	58.44	58.64
No. of protein-coding genes	14,974	15,497	16,837

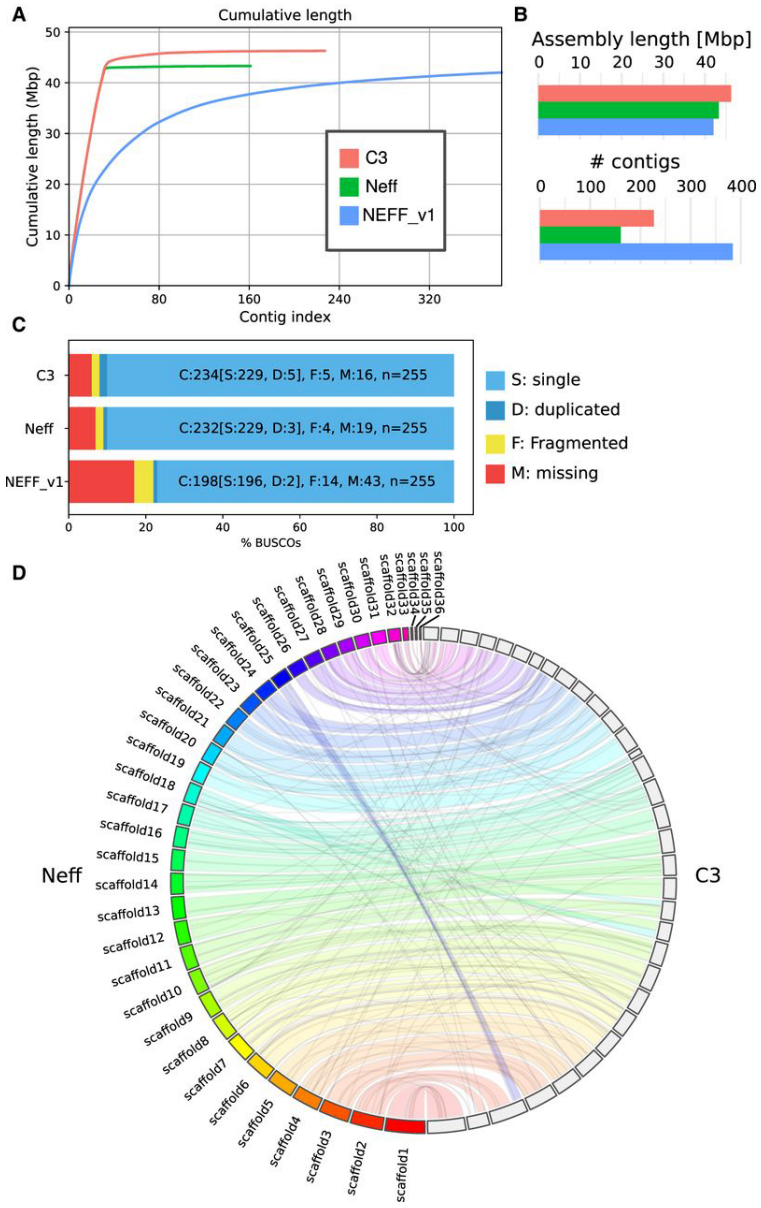


Figure 2.1 Assembly statistics for *A. castellanii* genomes. Comparison of genome assemblies for strains C3 and Neff-v2 versus the previous Neff-v1 genome assembly[20] (A) Cumulative length plot showing the relationship between the number of contigs (largest to smallest) and length of the assembly. (B) General continuity metrics. (C) BUSCO statistics showing the status of universal single-copy orthologs in eukaryotes for each assembly. (D) Circos plot showing syntenic blocks for all scaffolds of *A. castellanii* strains C3 and Neff assemblies >50 kb.

Our collaborators also took advantage of Hi-C contact maps to search for large mis-assemblies in these genome sequences, as demonstrated by Marie-Nelly et al.[162] Although this allowed them to manually address major unambiguous mis-assemblies, a number of visible mis-assemblies remained in complex regions such as repeated sequences near telomeres and ribosomal DNAs (rDNAs). These mis-assemblies could not be resolved with the data generated for this study. In the C3 assembly, there are a few (at least five) inter-chromosomal mis-assemblies that appear to be heterozygous and cannot be resolved without a phased genome. They also found read coverage depth to be highly heterogeneous between scaffolds, which is suggestive of aneuploidy (Supplementary Fig. 2.2).

The generation of chromosome-scale genome assemblies for two different *A. castellanii* strains afforded the first opportunity to compare and contrast their sizes and coding capacities. Despite their obvious relatedness (the 18S rDNA sequences of the two strains are 97% identical) (Supplementary Fig. 2.3), the C3 assembly was found to be 2.3 Mbp longer than the Neff-v2 assembly. We investigated this discrepancy by extracting all C3-specific regions from pairwise genome alignments with Neff sequences, which summed up to 3.2 Mbp. This comprised 5,072 different sequences ranging from 200 to 22,153 bp in length, with a median of 483 bp and a mean of 637 bp. These regions were scattered across the C3 scaffolds (Supplementary Fig. 2.4). Using BLASTX[157] against the NCBI Protein database, we found that the majority (>90%) of these sequences had strong hits to *Acanthamoeba* proteins.

The average gene length in the two strains (~2.3 kb), combined with the number of additional genes in C3 relative to Neff-v2 (1,307), supports the idea that the larger

number of genes in C3 explain most of the extra assembly length relative to Neff. Although the retrieved C3-specific sequence does exceed the difference in genome size between C3 and Neff-v2, the Neff-v2 genome has strain-specific genes of its own that likely account for this discrepancy. Despite the differences in gene content and genome size, our collaborators were able to demonstrate that the Neff and C3 genomes are highly syntenic (Fig. 2.1D), and the additional DNA in C3 does not appear to result in large structural rearrangements. They further tested this by aligning the Hi-C reads from C3 to the Neff genome and generating a contact map binned at 20 kb. Only three translocations were identified, in agreement with the Circos[135] plot shown in Figure 2.1D.

2.3.2 Spatial organization of the *A. castellanii* genome

Our collaborators determined the overall spatial organization in these two amoebae using their expertise in chromatin biology and organization. The following results are as reported by those authors, but they are valuable to include here as they significantly add to the expanded understanding of *Acanthamoeba* genome biology and provide context for my own analyses.

At the time of publication of our analyses, no Hi-C contact maps had been published from species of Amoebozoa. Although a Hi-C library was generated on *Entamoeba histolytica* for the purpose of genome scaffolding[164], its quality was too poor to yield a reliable contact map. Since then, at the time of writing this thesis, Hi-C analyses have been performed in a few additional amoebozoans, but in a targeted fashion and not on the scale of whole genome contact maps[165,166]. One of these targeted examples focused on contacts between the *Dictyostelium* genome and its endogenous plasmids,

while the other sought contacts between Redondoviruses and the *Entamoeba gingivalis* genome. Here, the Hi-C reads our collaborators used to generate the chromosome-scale scaffolding of the two *A. castellanii* genomes offered them the opportunity to reveal typical genome folding patterns in a species of this clade. Hi-C reads were re-aligned along the new assemblies of both the C3 and Neff strains to generate genome-wide contact maps. Visualizing the Hi-C contact maps of both genomes shows that *A. castellanii* chromosomes are well resolved in our assemblies (Fig. 2.2A).

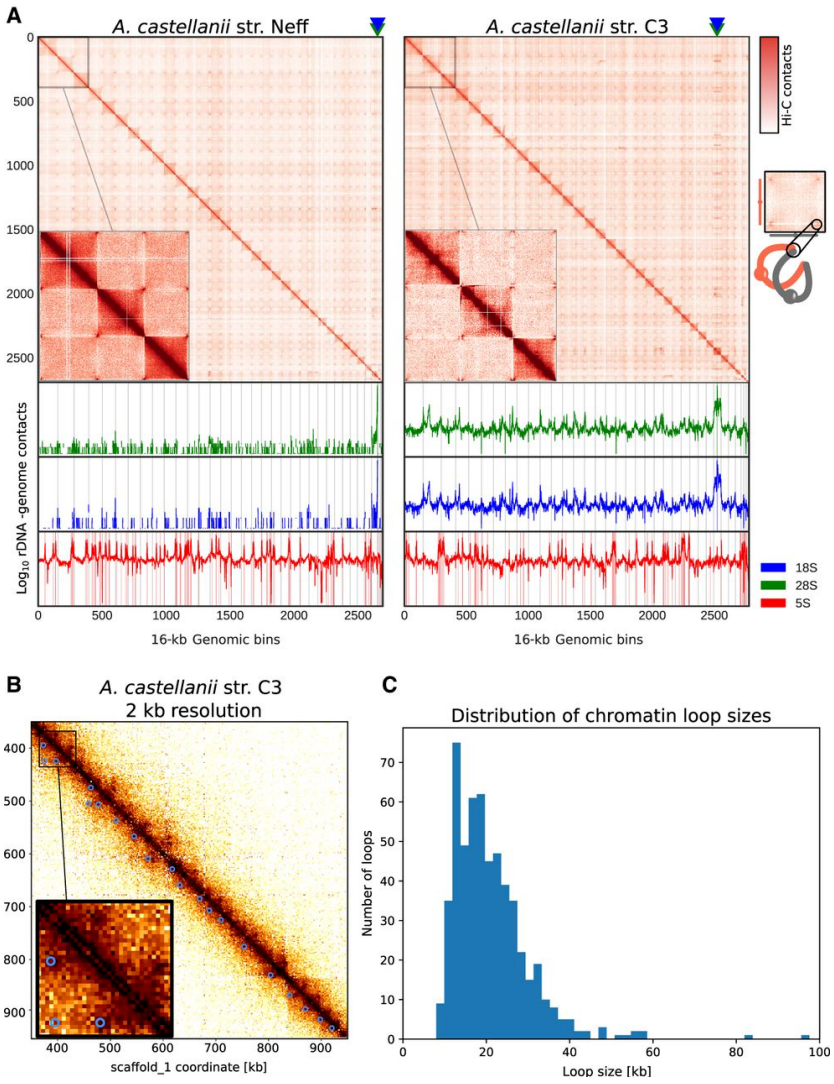


Figure 2.2 Spatial organization of the *A. castellanii* genome. (*A*, *top*) Whole-genome Hi-C contact maps of the Neff (*left*) and C3 (*right*) genomes, with a magnification of the three largest scaffolds. The genomes are divided into 16-kb bins, and each pixel represents the contact intensity between a pair of bins. Each scaffold is visible as a red square on the diagonal. In both strains, there is an enrichment of interscaffold contacts toward telomeres, suggesting a spatial clustering of telomeres, as shown on the model in the *right* margin. (*Bottom*) 4C-like representation of spatial contacts between rDNA and the rest of the genome. Scaffolds are delimited by gray vertical lines. Contacts of all rDNAs are enriched toward telomeres. The genomic position of the 18S and 28S genes is highlighted with triangles on the *top* panel, and the occurrences of 5S rDNA sequences are shown with vertical red lines on the *bottom* panel. (*B*) High-resolution contact map for a region of the C3 genome showing chromatin loops detected by Chromosight as blue circles. (*C*) Size distribution of chromatin loops detected in the C3 strain

In Neff, the highest intensity contacts are concentrated on the main diagonal, suggesting an absence of large-scale mis-assemblies. On the other hand, the C3 assembly retains a few mis-assembled blocks, mostly in the rDNA region where tandem repeats could not be resolved correctly with the data available to us. However, for both strains, the genome-wide contact maps reveal a grid-like pattern, with contact enrichment between chromosome extremities resulting in discrete dots. These contacts can be interpreted as a clustering of the telomeres, or subtelomeres, of the different chromosomes (Fig. 2.2A). Based on the presence of these inter-telomeric contact patterns, Hi-C contact maps suggest the presence of at least 35 chromosomes in both strains, ranging from ~100 kb to 2.5 Mb in length (Fig. 2.3). Additionally, we found ~100 copies of 5S rDNA dispersed across most chromosomes for both strains, and 18S/28S rDNA genes show increased contacts with subtelomeres (Fig. 2.2A). ECC finder[167], a computational tool that detects extrachromosomal circular DNA elements, was run on the Oxford Nanopore long reads to assess whether these rDNA sequences correspond to extrachromosomal circular rDNA contacting the genome at various positions. No evidence of circular DNA elements was found, suggesting that rDNA sequences in *A. castellanii* are interspersed within chromosomes. However, given the repetitive nature of these sequences, it is formally possible that some of the rDNA insertion sites are incorrect. In addition, the possible amplification of extrachromosomal palindromic linear rDNA, similar to that found in *Dictyostelium*[168], may also confound their proper analysis.

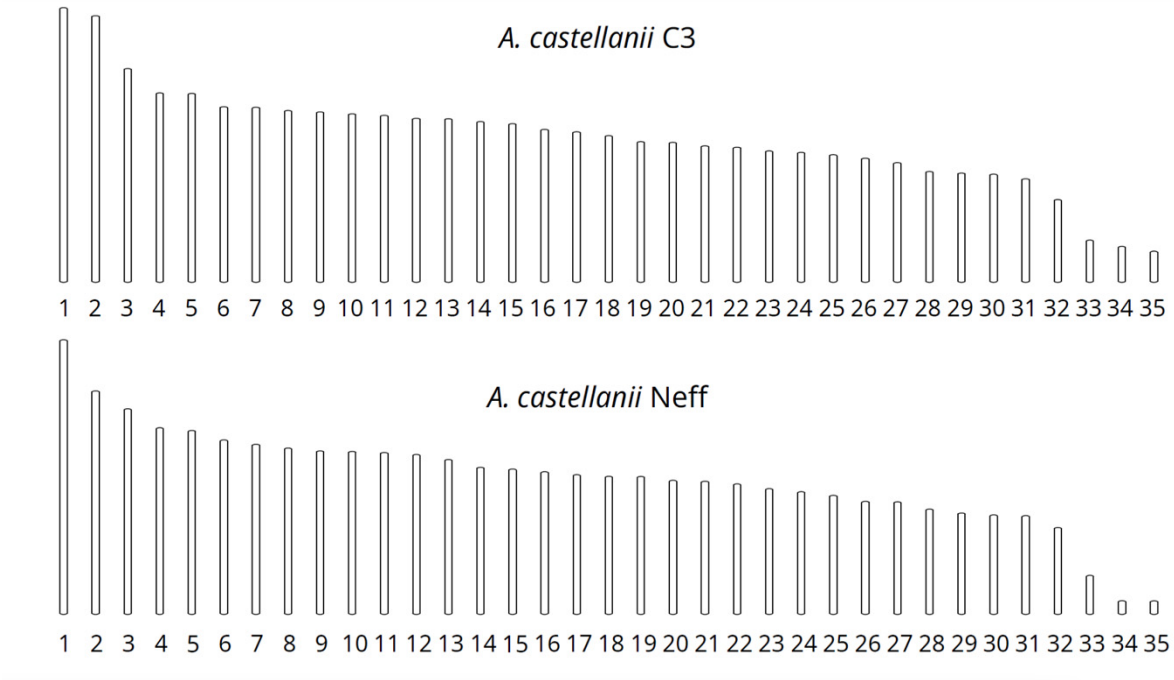


Figure 2.3 Predicted karyotypes of *A. castellanii* strains C3 and Neff
For each strain, 35 scaffolds likely to be chromosomes based on the presence of inter-telomeric contact patterns on the contact maps are ordered by size. C3 chromosomes range from an estimated 200 Kbp to 2.5 Mbp while Neff chromosomes range from an estimated 100 Kbp to 2.5 Mbp.

In addition to large, interchromosomal subtelomeric contacts, our collaborators also explored the existence of intrachromosomal chromatin 3D structures in the contact maps using Chromosight, a program that detects patterns reflecting chromatin structures on Hi-C contact maps[160]. For both strains, Chromosight identified arrays of chromatin loops along chromosomes, as well as boundaries separating chromatin domains (Fig. 2.2B). Most chromatin loops are regularly spaced, with a typical size of 20 kb (Fig. 2.2C), reminiscent of the average loop size reported along metaphase chromosomes in the budding yeast *Saccharomyces cerevisiae*[169,170]. The chromatin domains correspond to discrete squares along the diagonal (Supplementary Fig. 2.5A-C)

Our collaborators overlapped all predicted genes in the C3 genome with the domain borders detected from Hi-C data and measured their base expression using RNA-seq generated from that strain (see Methods). They selected the closest gene to each domain border and found that the genes overlapping domain boundaries are overall more highly expressed than those that do not (Supplementary Fig. 2.6C). In addition, the analysis showed that gene expression is negatively correlated with the distance to the closest domain border (Supplementary Fig. 2.6D). They performed the same comparison using chromatin loop anchors instead of domain borders. To a lesser extent, genes overlapping chromatin loops are also associated with higher expression (Supplementary Fig. 2.6A), although it is not correlated with the distance from the closest loop (Supplementary Fig. 2.6B). Altogether, these results suggest that these two types of chromatin structures (chromatin loop anchors and domain borders) are both associated with gene expression, although the association between gene expression and chromatin loop anchors is likely a result of their colocalization with domain borders (Supplementary Fig. 2.6E). Some microorganisms organize their chromosomes into microdomains whose boundaries correspond to highly expressed genes (e.g., budding yeasts and euryarchaeotes)[171,172]. Our findings in *A. castellanii* are therefore reminiscent of this type of organization.

2.3.3 The Neff and C3 genomes have partly non-overlapping gene content

I used both Broccoli[153] and OrthoFinder[154] for inference of orthologous groups (see Methods). A summary of the inferred orthogroups shared by, and specific to, the Neff and C3 strains of *A. castellanii* is presented in Figure 2.4, with orthogroup numbers

from both orthologous clustering tools included. This figure only compares Neff against C3, irrespective of orthogroup presence or absence in outgroup taxa. In this analysis, each strain-specific gene that was not assigned to an orthogroup by either program was still considered to be a single strain-specific orthogroup in order to account for the presence of genes without any orthologs across the five species (two *Acanthamoeba* strains plus three amoebozoan outgroups). Broccoli predicted more orthogroups overall and more strain-specific genes than OrthoFinder but predicted fewer shared orthogroups. Despite these differences, the overall trend is similar for the two outputs. The number of orthogroups shared by the two strains is roughly an order of magnitude greater than the number specific to either strain, whereas the C3 strain has a greater number of strain-specific orthogroups than the Neff strain as predicted by both programs.

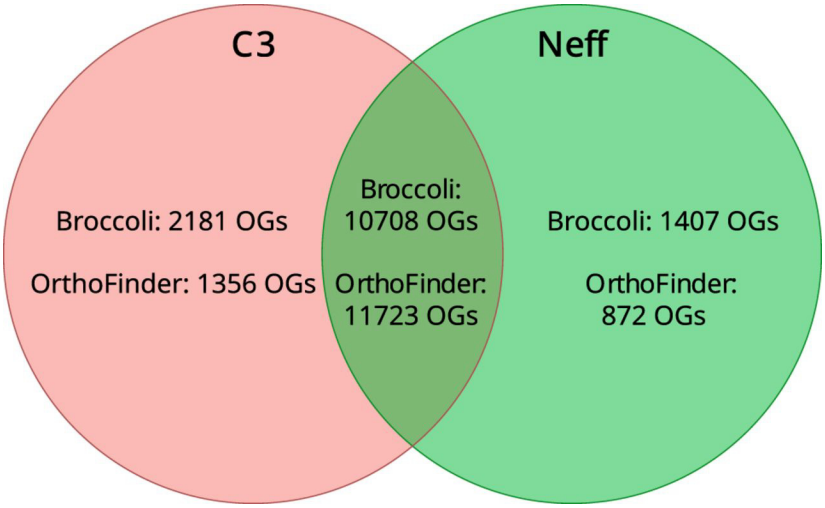


Figure 2.4 Numbers of strain-specific and shared orthologous groups in the genomes of *A. castellanii* strains C3 and Neff. Orthology inference was conducted with both Broccoli[153] and OrthoFinder[154]. *Dictyostelium discoideum*, *Physarum polycephalum*, and *Vermamoeba vermiformis* were used as outgroups to improve accuracy of orthogroup inference.

To investigate how similar the *A. castellanii* gene complement was to other members of Amoebozoa, I evaluated *A. castellanii* orthogroups for their presence in three outgroup species. Both Broccoli and OrthoFinder outputs were analyzed in this fashion. According to Broccoli, 43.5% of orthogroups shared by the two *A. castellanii* strains were not present in the other three amoebae, whereas OrthoFinder gave a figure of 48.5%. In the Neff strain, 49.3% of all orthogroups, shared or strain-specific, were not found in the three outgroup amoebae according to Broccoli compared with 51.4% as predicted by OrthoFinder. In the C3 strain, the Broccoli results indicate that 52.1% of all orthogroups are not present in the outgroup amoebae, whereas 52.8% were not found in the outgroup by OrthoFinder. This is in contrast with *A. castellanii* strain C3 sharing an estimated 83.1% (Broccoli) to 89.6% (OrthoFinder) of its orthogroups with the Neff strain, and the Neff strain sharing an estimated 88.3% (Broccoli) to 93.1% (OrthoFinder) of its orthogroups with the C3 strain.

2.3.4 *A. castellanii* accessory genes show strain-specific functional enrichment

In an attempt to gain insight into the functional significance of strain-specific genes in the C3 and Neff genomes, I identified the 30 most significantly enriched GO terms in each strain-specific set relative to the whole gene content of the corresponding strain. I used topGO (https://bioconductor.org/packages/topGO/ [accessed July 12, 2022]) for this enrichment analysis and plotted the GO terms in order of decreasing P-value for each strain/ontology combination (Figs. 2.5–2.10). Among C3-specific genes, five terms were found to be statistically significantly enriched in the “biological process” ontology

at a 95% confidence level, whereas nine terms were enriched in the “molecular function” ontology, and two were enriched in the “cellular component” ontology. Among Neff-specific genes, five terms were significantly enriched in the “biological process” ontology at a 95% confidence level, seven were enriched in the “molecular function” ontology, and one was enriched in the “cellular component” ontology. All enriched Gene Ontology (GO) terms and the corresponding P-values can be found in Table 2.2. Note that a multiple testing correction has not been applied to these P-values, on the advice of the topGO user manual[173]. The authors of that program explain that this is due to the minor role of multiple testing in the overall error rate intrinsic to the methodology, the lack of independence between GO terms due to the hierarchical organization of Gene Ontology, and the risk of underestimating enrichment if multiple testing correction is applied.

Table 2.2. Functions enriched in C3- or Neff-specific sets of genes.

	C3	
	GO term	p-value
Biological process	Macromolecule methylation	1.9 x 10-5
	Protein phosphorylation	0.00068
	Small GTPase mediated signal transduction	0.00289
	Amino acid transport	0.01822
	DNA topological change	0.02584
Molecular function	S-adenosylmethionine-dependent methyltransferase activity	0.0042
	GTP binding	0.00125
	Chromatin binding	0.00139
	Phosphotransferase activity, alcohol group as acceptor	0.00264
	Catalytic activity, acting on DNA	0.0089
	DNA topoisomerase type II (double strand cut, ATP-hydrolyzing) activity	0.01869
	Oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	0.03677
	DNA binding	0.0419
	O-acyltransferase activity	0.04753
Cellular component	Chromosome, centromeric region	0.00028
	RNA polymerase II, core complex	0.04003
	Neff	
Biological process	Protein phosphorylation	1.1 x 10-5
	Regulation of cellular process	0.00041
	DNA recombination	0.01092
	Cyclic nucleotide biosynthetic process	0.01844
	Protein homooligomerization	0.038
Molecular function	Endoribonuclease activity, producing 5‘-phosphomonoesters	4.7 x 10-5
	Protein-macromolecule adaptor activity	0.0011
	Protein kinase activity	0.0011
	Actin filament binding	0.0035
	Purine ribonucleoside triphosphate binding	0.0232
	Structural molecule activity	0.0293
	Nucleic acid binding	0.0319
Cellular component	Virion part	1.5 x 10-9

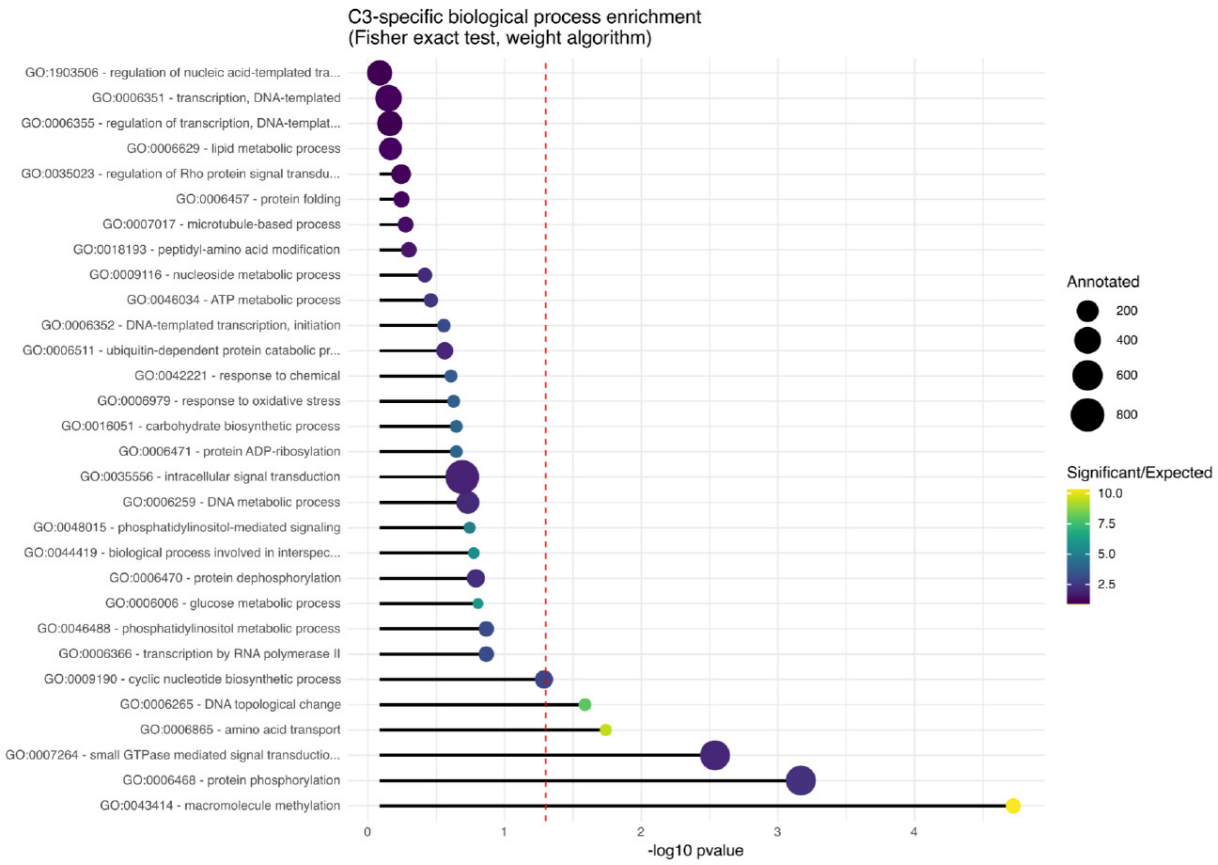


Figure 2.5 Most significant biological process GO term enrichments in genes specific to *Acanthamoeba castellanii* strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected. See Table 2.2 for names of GO terms where truncated.

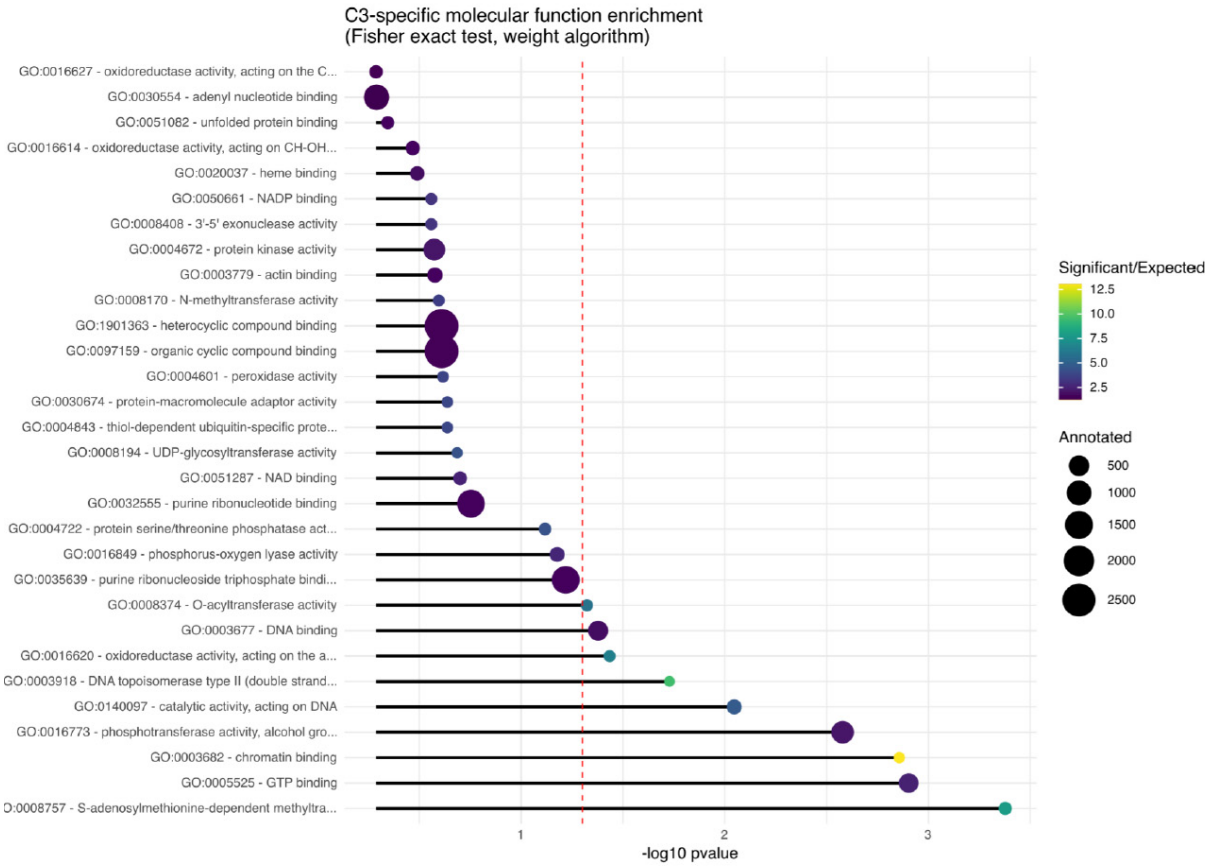


Figure 2.6 Most significant molecular function GO term enrichments in genes specific to *Acanthamoeba castellanii* strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected. See Table 2.2 for names of GO terms where truncated.

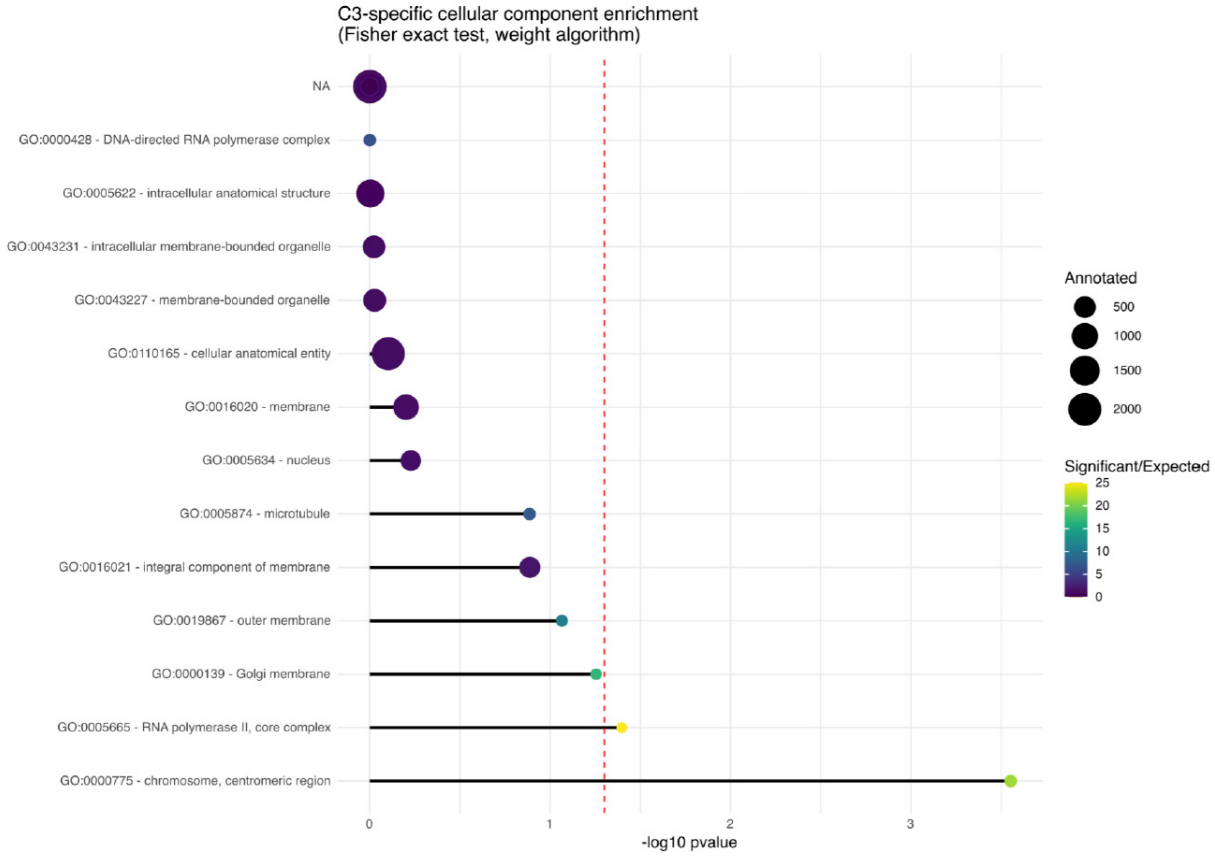


Figure 2.7 Most significant cellular component GO term enrichments in genes specific to *Acanthamoeba castellanii* strain C3. Enrichment was determined using topGO , with nodeSize set to 5 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

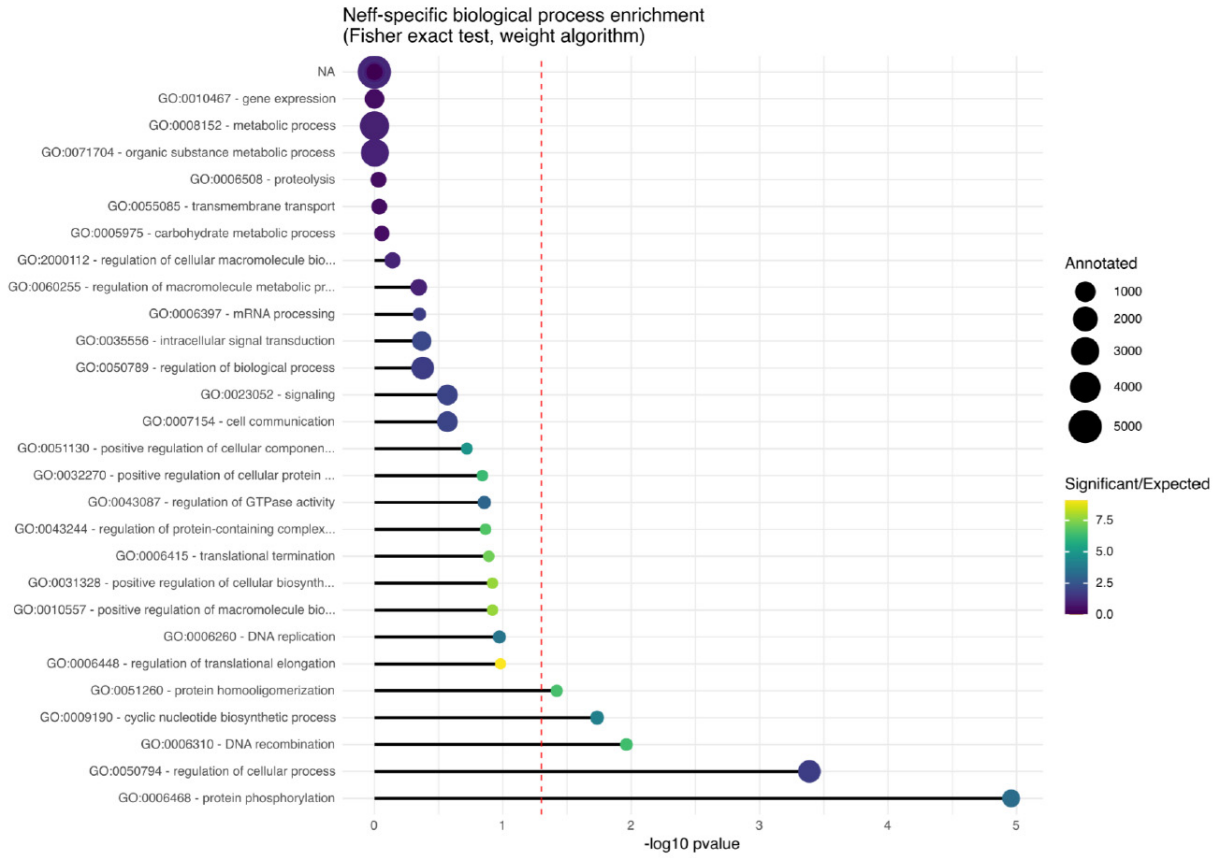


Figure 2.8 Most significant biological process GO term enrichments in genes specific to *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected. See Table 2.2 for names of GO terms where truncated.

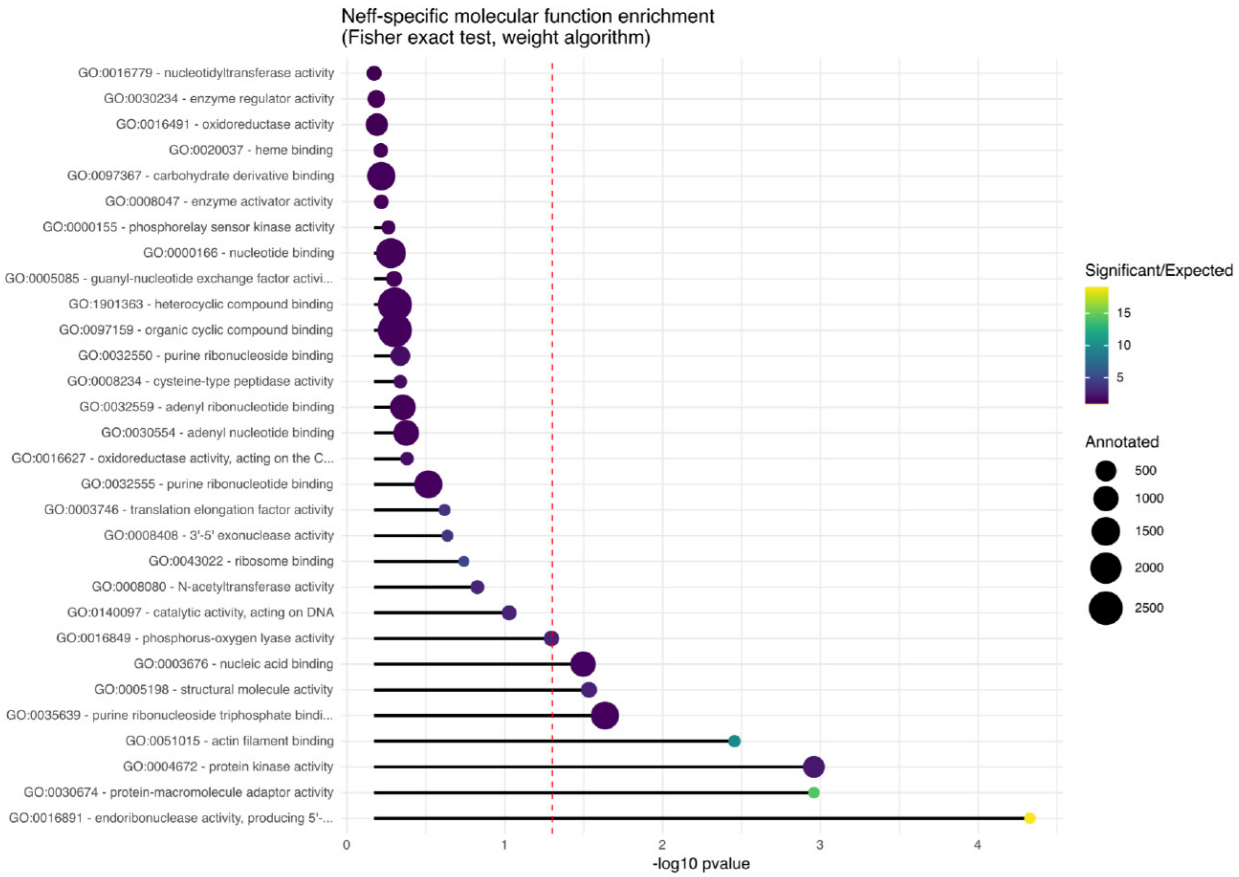


Figure 2.9 Most significant molecular function GO term enrichments in genes specific to *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected. See Table 2.2 for names of GO terms where truncated.

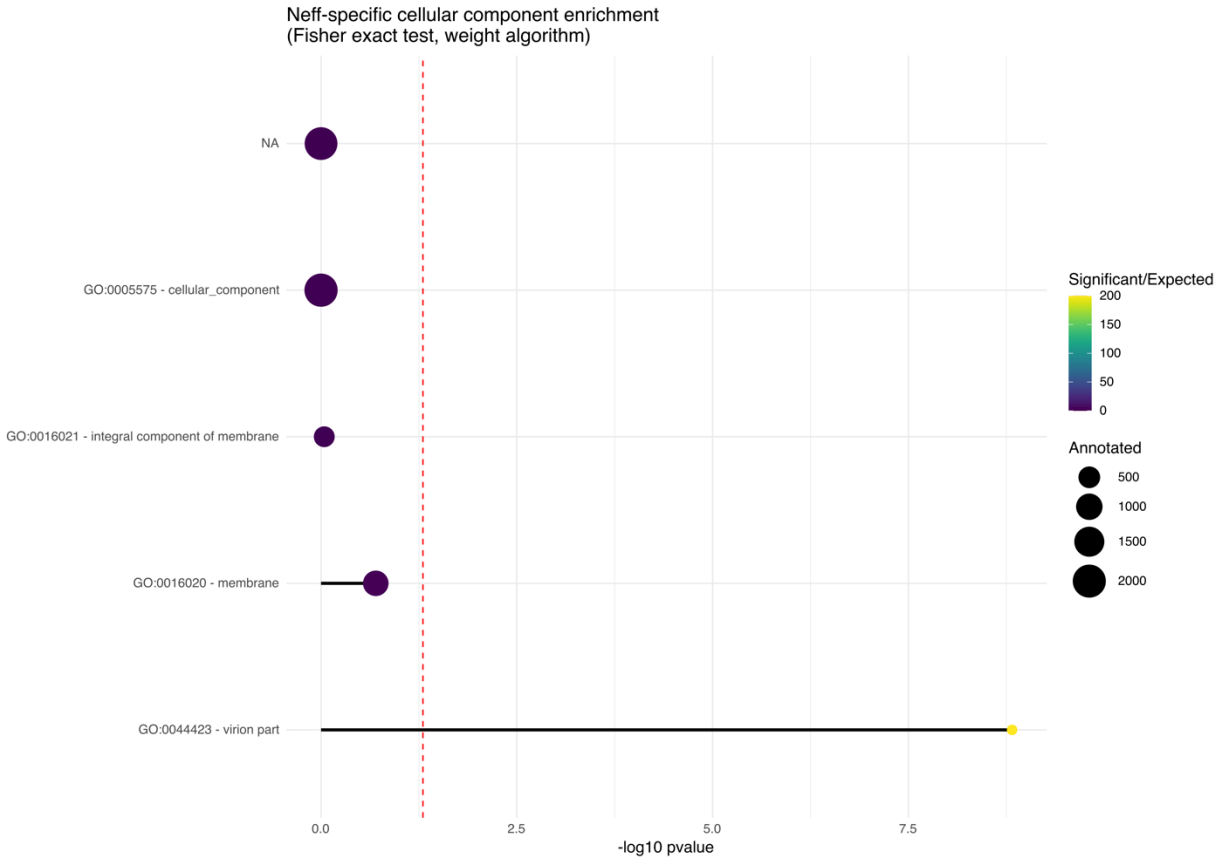


Figure 2.10 Most significant cellular component GO term enrichments in genes specific to *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 5 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

For some of the enriched functional categories, the strain-specific genes contributing to the enrichment showed a relatively cohesive signal in terms of best hits when searched against the nr database with BLASTP[157]. The Neff genes annotated as "virion parts" were the same ones responsible for the enrichment in "structural molecule activity." These genes had best BLAST hits to the major capsid protein from various nucleocytoplasmic large DNA viruses (NCLDVs). Those genes responsible for "protein

homo-oligomerization" enrichment had their best BLAST hits to K^+ channel tetramerization domains, whereas all those contributing to "DNA recombination" enrichment had best BLAST hits to IS607 family transposases, those contributing to "cyclic nucleotide biosynthetic process" enrichment had best BLAST hits to serine/threonine kinases, and those contributing to enrichment of both "protein-macromolecule adaptor activity" and "actin filament binding" had best BLAST hits to fascin-like proteins. In C3, there were fewer enrichment categories in which BLAST hits gave such a cohesive signal, but some examples are "amino acid transport" enrichment, in which the associated genes had best BLAST hits to serine/threonine kinases; "DNA topological change," in which the best BLAST hits were to DNA topoisomerase 2; and "oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor," in which the best BLAST hits of all associated genes were to Ars-J-associated glyceraldehyde-3-phosphate dehydrogenases.

2.3.5 The Neff strain has a divergent mannose-binding protein

Our collaborators noted differences in the entry of *L. pneumophila* upon infection between the C3 and Neff strains. For example, when both amoeba strains were infected with a multiplicity of infection (MOI) ratio of 0.1, they observed that *L. pneumophila* enters in higher numbers in C3 compared with Neff (Supplementary Fig. 2.7A). They also used different MOIs for infection of the Neff strain, which revealed that an MOI = 10 was required to obtain comparable numbers of bacterial entry to an MOI = 0.1 using C3. In addition, when bacterial counts were normalized to the number of bacteria that already entered the host cell, *Legionella* appeared to replicate faster in C3 compared with

the Neff strain (Supplementary Fig. 2.7B). These experiments confirmed previous empirical observations that led to the adoption of the C3 strain as a preferred model with respect to infection. Therefore, they were interested in looking for any possible explanations for these differences in the genome sequences we generated.

One gene of particular interest encodes a mannose-binding protein (MBP), which is known to be used as a receptor for cell entry by *Legionella* in some *A. castellanii* strains[174]. I decided to compare this protein across a few *A. castellanii* strains to see if it may be implicated in the difference in infectibility between the strains. The MEEI 0184 strain of *A. castellanii*, an isolate from a human corneal infection, was used as a reference sequence, because it is the only strain in which the MBP was biochemically characterized[175,176]. The orthologs from C3, Neff, and *Acanthamoeba polyphaga* were retrieved, and all four sequences were aligned (Supplemental Fig. 2.8). The percentage identity of each sequence to the reference was calculated over the sites in the alignment in which the *A. polyphaga* sequence was not missing (Table 2.3). The C3 homolog was found to be 99.5% identical to the MEEI 0184 homolog, whereas the Neff and *A. polyphaga* proteins were more divergent, sharing 91.6% and 97.2% identity to MEEI 0184, respectively. Despite being of the same species as the reference, the Neff strain homolog was found to be much more divergent than the *A. polyphaga* sequence is from the other two *A. castellanii* strains.

Table 2.3. Identity of mannose-binding proteins from *A. polyphaga* and *A. castellanii* strains Neff and C3 to their homolog in *A. castellanii* strain MEEI 0184

Strain	Identity	Gaps
Neff	757/826 (91.6%)	1/826 (0.12%)
C3	821/825 (99.5%)	0
A. polyphaga	802/825 (97.5%)	0

The first 46 sites of the alignment were excluded from the calculation because the 5' end of the gene in *A. polyphaga* was missing due to a truncated contig.

We propose that the phenotype of differential infectibility results partly from impaired receptor-mediated entry by *Legionella* into Neff cells due to differences in the receptor encoding gene (a hypothesis that our collaborators intend to test in subsequent investigations) and also from other differences in amoeba physiology between C3 and Neff strains.

2.4 Discussion

2.4.1 Chromosome-level genome assembly

Generation, analysis, and comparison of the genome sequences of two *A. castellanii* strains revealed heterogeneous coverage across scaffolds (Supplementary Fig. 2.2), which is consistent with previous findings[10] that *A. castellanii* has a high but variable ploidy of approximately 25*n* (see Chapter 4). Previous estimates of the *A. castellanii* Neff karyotype using pulsed-field gel electrophoresis estimated 17 to 20 unique chromosomes ranging from 250 kbp to just over 2 Mbp[106], whereas the Hi-C analysis from our collaborators produced an estimated karyotype of at least 35 unique

chromosomes with a similar size range of 100 kb to 2.5 Mb. The discrepancy between the number of bands in the electrophoretic karyotype and our estimate may result from chromosomes of similar size comigrating on the gel, which our collaborators were able to resolve using sequence- and contact-based information.

It was previously estimated that *A. castellanii* has 24 copies of 5S rDNA per haploid genome[177]. The Hi-C data show that both strains contain four times as many copies as originally thought. Considering features of the nuclear biology of *A. castellanii*, such as suspected amitosis[178] and probable aneuploidy, our finding that 5S rDNA is dispersed across all chromosomes may serve to ensure a consistent copy number of 5S rDNA in daughter cells.

It was also observed that at first glance, the contact maps show a clustering of subtelomeric regions, but do not display a Rabl conformation, in which centromeres cluster to the spindle-pole body[179]. However, the precise positions of centromeres are needed to see whether these chromosomes are acrocentric, which could lead to an overlap of the contact signal between centromeres with the contacts between subtelomeres and could mask centromere clustering.

2.4.2 *A. castellanii* accessory genes may permit environmental adaptation

Among the large number of genes predicted to be strain specific in *A. castellanii*, I found several functions to be significantly enriched in either the Neff or C3 strain-specific gene sets. Of these, the most biologically interesting is the enrichment of the "small GTPase-mediated signal transduction" and "GTP binding" genes in C3. The enrichment of these two GO terms, along with the enrichment of protein phosphorylation,

suggests that the C3 strain may have expanded its environmental sensing capabilities and associated cellular responses by expanding the gene families involved in signal transduction. Given the extensive gene repertoire of *A. castellanii* dedicated to cell signaling, environmental sensing, and the cellular response[20], which is thought to help the amoeba navigate diverse habitats and identify varied prey, it seems likely that alterations of this gene repertoire in C3 may have enabled further environmental adaptations.

Another notable enrichment is the "virion parts" in the *A. castellanii* Neff strain. This enrichment includes a single gene with a best BLAST hit for the major capsid proteins in various NCLDVs, including a very strong hit to *Mollivirus sibericum*. Many NCLDVs, including *Mollivirus*, are known viruses of *Acanthamoeba* spp.[67]. Although no phylogenetic analysis was performed at this time to investigate the origin of this major capsid protein gene in the Neff genome, it seems plausible that it was acquired by lateral gene transfer during an NCLDV infection, perhaps by *Mollivirus* or a closely related virus.

Some of the remaining enriched functions seem related, but we are unable to speculate on their broader biological significance. For example, we observe enrichment of "macromolecule methylation," "DNA topological change," "chromatin binding," "DNA binding," "catalytic activity, acting on DNA," and "DNA topoisomerase II (double-strand cut, ATP-hydrolyzing) activity" in the C3 strain, all of which appear to be related to DNA modification and maintenance. There are fewer such cases in the Neff strain: One can imagine possible connections between "protein phosphorylation," "protein kinase activity," and "purine ribonucleoside triphosphate binding," as well as between "endoribonuclease activity, producing 5′-phosphomonoesters" and "nucleic acid

binding" and, finally, between "actin filament binding" and "protein-macromolecule adaptor activity." Other enrichments beyond these examples have no obvious biological significance. They could well be nonadaptive, having been generated by gene duplication, differential loss in the other amoebae strain studied, or lateral gene transfer, without conferring any significant selective advantage. An improved understanding of the cellular and molecular biology of *Acanthamoeba* is needed to make sense of the genetic enrichment data presented here. The nature of GO term enrichment analysis also tends to involve quite broad categories, so the significance of some of these enrichments is difficult to ascertain.

2.4.3 Substitutions in the Neff MBP may inhibit *Legionella* entry

Alignment of the three *A. castellanii* MBPs and the *A. polyphaga* homolog may help explain the difference in susceptibility to *Legionella* infection between the Neff and C3 strains. The C3 strain MBP is highly similar to its counterpart in strain MEEI 0184, which was first to be biochemically characterized. The Neff strain MBP, however, is markedly more divergent than even the *A. polyphaga* MBP, which is not known to participate in *Acanthamoeba–Legionella* interactions[180]. Infection studies comparing infection phenotypes depending on the MOI confirmed that *L. pneumophila* enters 10 times less into Neff cells than into C3 cells in our in vitro infection system. These results are consistent with the hypothesis that the Neff strain of *A. castellanii* is a poor host for infection by *Legionella* partly due to an accumulation of amino acid substitutions in its MBP, substitutions that may prevent *Legionella* from binding to this protein during cell entry. It is worth noting that the Neff strain has been in axenic culture since 1957, so it

may be that the relaxed selective pressure on this protein, combined with repeated population bottlenecking during culture maintenance, has allowed for mutations in the Neff strain MBP gene to accumulate. Even considering these factors, though, this would imply quite a high number of nucleotide substitutions. At the present time, without available genome data for strains more closely related to the Neff strain, it cannot be determined whether these mutations arose in nature or in culture. However, given that the divergence of the *A. polyphaga* ortholog to the MEEI 0184 strain is much less than that of the Neff strain, despite all four strains having similar lifestyles in nature, evolution of the Neff strain since being deposited in the culture collection seems likely.

Two studies[181,182] published since the publication of this work in 2022 have addressed *Acanthamoeba* adhesins, specifically the mannose binding protein and the laminin binding protein, but these have investigated the role of these proteins in *Acanthamoeba* pathogenicity in animals, not *Legionella* infection of *Acanthamoeba.* These studies did, however, acknowledge that these proteins appear to exist across pathogenic and non-pathogenic *Acanthamoeba* isolates alike, and likely have a role in prey detection, while subsequently being co-opted for the parasitic lifestyle that leads to pathogenesis.

2.5 Conclusions

This project improved our understanding of the *Acanthamoeba* genome, and by extension, amoebozoan genomics as a whole. At the core of these contributions is the generation of a new high-quality reference genome sequence for *Acanthamoeba* strain Neff, as well as an additional high-quality reference genome sequence within this genus. Even ignoring any additional results that followed, there is a demand for reliable genomic

resources across the tree of eukaryotes to facilitate the study of protists and their evolution. The greater contiguity and completeness of our assemblies has led to a fuller and more accurate inventory of the gene content in these two *Acanthamoeba* strains, as well as a vastly expanded structural understanding of their genomes, and a much-reduced likelihood that any genes or features of interest found in these assemblies are artefacts derived from contamination or mis-assembly. In my view, the value of more accurately determining gene content and avoiding artefacts that plague more fragmented assemblies will be immediately obvious to most researchers in the genomics sphere. However, I think that the value of determining genome structure is somewhat underappreciated and will be crucial if the field moves in the direction of studying the mechanistic side of eukaryotic genome biology and evolution.

Beyond the genomic resources produced in this project, we were able to determine some of the fundamental features of the genome that had previously eluded researchers. Having near complete resolution of the karyotype is a crucial contribution for being able to study processes that act at the chromosome level, like the generation and maintenance of aneuploidy and polyploidy, potential endoreplication, potential depolyploidization, and whether meiosis is occurring. In Chapter 4 I demonstrate how the newly resolved chromosomes can be analyzed independently to explore ploidy in this organism. The contact information from the Hi-C analysis also allowed the chromosomal locus of the rDNA operon to be identified, which had previously not been achieved.

Additionally, the comparative genomic analyses performed here have added data points to the ongoing and rapidly expanding discussion surrounding eukaryote lateral gene transfer, eukaryote pangenomes, and fine-scale (i.e., intrageneric or intraspecific)

genomic variation in eukaryotic lineages. The gene content comparison of Neff and C3 provides another clear example of appreciable variation between two strains that appear quite closely related based on structural and nucleotide-level similarity, which advances the recently prevailing view that such variation is not rare or difficult to achieve across microbial eukaryote diversity[99]. Furthermore, our functional enrichment analysis of the strain-specific genes in this study provides another data point to support hypotheses that have been long held in the field about the types of genes that tend to make up the 'core genome' and those that tend to fall into the 'accessory genome'. More specifically, genes required for the basic functioning of a eukaryotic cell, like those involved in information processing and core cellular and metabolic processes, tend not to be strain-specific (i.e., accessory), while genes that could permit niche adaptation, such as those encoding the capacity for enhanced environmental sensing and signal transduction, or a broader range of metabolic substrates, are found more often than expected in the accessory genome[183–185]. The presence of genes in the accessory genome known to be specifically involved in viral processes also provides a hint at the mechanisms that may be involved in shaping this intraspecific variation. This topic will be explored more thoroughly in Chapter 4 through a phylogenomic analysis of the LGT contributions to both the Neff and C3 genomes.

CHAPTER 3 THE FATE OF ARTIFICAL TRANSGENES IN *ACANTHAMOEBA CASTELLANII*

3.1 Introduction

The ability to express transgenes in eukaryotic cells has become a staple of the contemporary molecular genetics toolkit, facilitating a wide array of functional studies that answer genetic, biochemical, and cell biological questions[186,187]. These experiments often seek to tag, overexpress, or suppress proteins of interest and make inferences about that protein's role in the organism's biology. Sophisticated transformation systems and the genetic toolkits built upon them are available for most model organisms, and for most new species we wish to characterize on a molecular and cell biological level, researchers strive to develop a corresponding system for transformation and genetic manipulation (ref. 188 reports on such efforts and their importance for a range of marine protists[188]). However, while the introduction and expression of transgenes is ubiquitous in molecular biology, relatively few researchers seek to determine the fate of the introduced DNA within the host cell. This is not unjustified; for most experiments of this nature, the fate of the transgene is irrelevant to the research question. For the study of genome biology, though, this knowledge gap can provide information on how eukaryotic organisms maintain their own genomes and respond to the introduction of foreign DNA. To this end, I employed a variety of molecular biological and sequence-based methods to characterize the interaction between *Acanthamoeba castellanii* and artificially introduced transgenes.

My motivation for this investigation was multifaceted. The first and most obvious motivation was simply that I had performed the transformation experiments in preparation for detecting LGT so I should see what happened at the genomic level, independent of any of my pre-existing interests. Beyond that, I expect there are at least

some similarities between the genome biology of acquiring stable transgenes artificially and acquiring them 'naturally' in cases of lateral gene transfer with no human intervention. I was hoping to leverage my findings on the patterns and mechanisms of artificial transgene integration to better understand how lateral gene transfer may take place in eukaryotic genomes.

The transformation protocol for this organism was developed by Peng, Omaruddin, and Bateman[189], with further refinement by Bateman[190]. The development of this protocol was largely a proof of principle with respect to transformation and protein expression in *A. castellanii,* so the plasmids were constructed to contain a selectable marker, neo^R, conferring neomycin and geneticin resistance, as well as a reporter gene, enhanced green fluorescent protein (EGFP), both expressed from endogenous *A. castellanii* promoters. The needs of this study were no more complex, so the same set of plasmids were used for my experiments.

In transformation experiments, there are generally thought to be two potential options for maintaining the transgenes; an additional possibility is transient transformation where the transgenes are expressed and then lost within a short time[187]. The first option for maintenance is to be integrated into one or more loci on the host chromosomes, where the integrated DNA is now faithfully replicated from generation to generation as a part of the genome. It is formally possible that the DNA could be integrated into an organellar genome, but in experiments where selection relies on expression from nuclear promoters, organellar genome integrants may not persist[191]. The second hypothesized option for stably maintaining transgenes is for the cell to replicate them as extrachromosomal elements, often called episomes, where they remain physically

separate from the host genome, but undergo the same replication and segregation processes over time. Evidence exists for both possibilities in different species, and the two are not mutually exclusive, even in the same cell. For chromosomally integrated transgenes, additional questions arise, such as the mechanism for integration and its proximity to any particular genomic features that may have influenced the process or frequency of integration.

Historically, answering these questions required a mostly experimental approach, relying heavily on polymerase chain reaction, Southern blotting, restriction fragment analysis, and some degree of Sanger sequencing[192,193]. However, the advent of long-read, single molecule sequencing technology affords us the opportunity to screen genomes in a high throughput fashion. In this study, I used Oxford Nanopore sequencing to perform a broad search for transgenes in *A. castellanii*, as well as for more targeted and detailed investigation. These results were then supported by molecular biological methods for confirmation and additional characterization.

3.2 Methods

3.2.1 Culturing

All cultures involved in this study are *Acanthamoeba castellanii* strain Neff, hereafter referred to as *"Acanthamoeba"*. They were grown at room temperature in Neff base medium with additives (ATCC Medium 712; 0.75% yeast extract, 0.75% proteose peptone, 2 mM KH_2PO_4, 1 mM $MgSO_4$, 1.5% glucose, 0.1 mM ferric citrate, 0.05 mM $CaCl_2$, 1 µg/mL thiamine, 0.2 µg/mL D-biotin, and 1 ng/mL vitamin B_{12}). Transformed culture media also contained the antibiotic G418 at a concentration of 10 µg/mL during

initial culture establishment after transformation, and 50 μg/mL for full-strength selection and long-term maintenance.

3.2.2 Transformation

All transformation experiments were based on the method described by Peng, Omaruddin, and Bateman[189], and further developed by Bateman[190]. My general adaptation of this protocol is available online at protocols.io as well (dx.doi.org/10.17504/protocols.io.s4regv6). The plasmid pGAPDH-EGFP, described by Bateman, was used for these experiments. This plasmid expresses the neomycin resistance marker from the endogenous *Acanthamoeba* TATA-box binding protein (TBP) promoter, and the enhanced green fluorescent protein (EGFP) gene from the endogenous *Acanthamoeba* glyceraldehyde-3-phosphate dehydrogenase (GAPDH) promoter. Upon receipt of the plasmid in the Archibald lab, I sent it to the Integrated Microbiome Resource at Dalhousie University for sequencing on an Illumina MiSeq instrument to confirm the sequence. As a plasmid sequence is roughly analogous in a bioinformatic sense to a transcript, Trinity[194] (2014-07-17 release) was used to assemble the reads rather than a program oriented toward assembling whole genomes. SnapGene Viewer was used to generate a plasmid map (Fig 3.1).

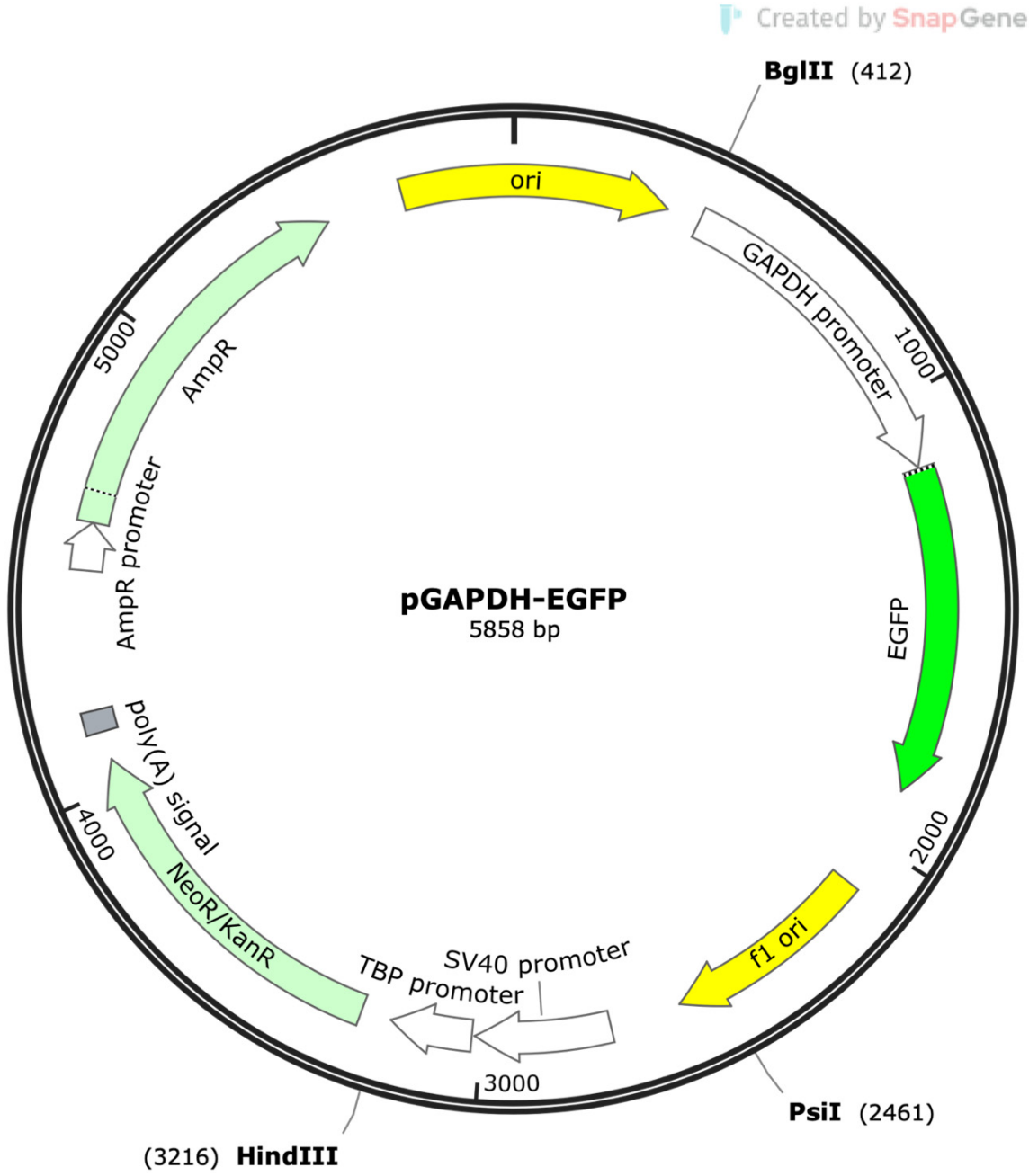


Figure 3.1 The plasmid pGAPDH-EGFP constructed by Bateman[190] was used for these experiments. The plasmid has an ampicillin resistance marker for propagation in *E. coli,* neomycin resistance marker for selection in *Acanthamoeba*, and enhanced green fluorescent protein (EGFP) as a reporter gene. The neomycin resistance marker and EGFP are expressed from endogenous *Acanthamoeba* promoters. Expression of the resistance marker is driven by the TATA-box binding protein (TBP) promoter, while EGFP expression is driven by the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) promoter. Restriction sites used in this study are marked on the plasmid map.

Briefly, for each transformation reaction, an aliquot of wild-type *Acanthamoeba* strain Neff (ATCC-30010) cell culture containing roughly 5×10^5 cells was centrifuged at 1000 *g* for 5 minutes at 4 °C to pellet the cells. The culture media was decanted from the cell pellet and replaced with 500 µL encystment medium (20 mM Tris–HCl [pH 8.8], 100 mM KCl, 8 mM $MgSO_4$, 0.4 mM $CaCl_2$, 1 mM $NaHCO_3$), then the cells were resuspended. The cell suspension was transferred to a small (3 – 6 cm) Petri dish or one well of a 6-well culture plate. 4 µg of the plasmid of choice contained in a 3 to 6 µL volume was diluted in 100 µL encystment medium, then 20 µL of the QIAGEN SuperFect reagent was added to the plasmid suspension and mixed by pipetting. This mixture was incubated for 10 minutes at room temperature to allow for the plasmid DNA to bind to the transfection reagent, then it was further diluted in another 600 µL encystment medium. This entire volume of transfection mixture was added to the vessel containing the cells to be transformed, and the cells were incubated with the mixture for 3 hours at room temperature.

Following this incubation, all liquid was gently pipetted out of the dish, leaving the adhered cells, and 4 mL of their standard growth medium was added. The cells were left to recover for 24 hours, then G418 was added to a concentration of 10 µg/mL. This antibiotic concentration was maintained until proliferation of the transformed population was observed, then the concentration was increased to 50 µg/mL. Typically this occurred after about a week, but had to be judged visually under the microscope. When the cell density appeared to be about half that of its maximum density in culture, it was an appropriate time to increase the selection. After this point, cultures could be maintained as usual, with a continuing antibiotic concentration of 50 µg/mL.

Fluorescence microscopy was used as a second line of evidence that the transformation had been successful, as the plasmid used expresses enhanced green fluorescent protein. A slide was prepared from a small aliquot of live cells directly from the culture and viewed with epifluorescence microscopy using a filter set designed for FITC observation.

3.2.3 Nanopore sequencing of transformed culture

Once the transformed cells had recovered, I obtained genomic DNA from the culture using the same method as described in Chapter 2. DNA samples were cleaned with QIAGEN G/20 genomic clean-up columns using the manufacturer's protocol, but with double the number of wash steps.

I prepared a sequencing library for the Oxford Nanopore MinION using the SQK-LSK108 ligation sequencing kit and sequenced on a FLO-MIN106 flow cell. The reads were basecalled with Albacore v2.1.7, as this sequencing experiment pre-dated the release of the Guppy basecaller.

3.2.4 Searching sequence reads for evidence of transgenes

To search for reads representing potential genomic integrations of the plasmid, BLASTn[157] was used to identify and retrieve all reads from the sequencing experiment with hits to the plasmid sequence. Then, this set of reads was compared against the wild-type genome assembly, again using BLASTn, to search for plasmid-genome junctions. After conducting this search, a simple text search was also performed against the plasmid-containing reads to search for telomeric repeats.

3.2.5 Southern blot analysis to locate transforming DNA

The following samples were prepared ahead of the Southern blot experiment: undigested genomic DNA from the transformed *Acanthamoeba* culture, as well as two aliquots from this gDNA that were digested with *Bgl*II and *Hin*dIII, respectively, and undigested pGAPDH-EGFP plasmid, as well as two aliquots digested with the same restriction endonucleases. A 1% agarose gel was prepared and these samples were run alongside GeneRuler 1 kb plus ladder at 90 V for 1 hour, then the DNA was visualized with ethidium bromide. The DNA was transferred to a positively charged nylon membrane following the Southern blotting protocol published by ThermoFisher (https://assets.fishersci.com/TFS-Assets/LSG/manuals/MAN0013296_Southern_Blotting._Dot_Blotting_UG.pdf), which is a typical alkaline transfer method.

The probe for this experiment was generated using the Thermo Scientific Biotin DecaLabel DNA labelling kit, which biotinylates an amplicon of the user's choosing. In this case, the probe was generated from a 345-bp segment of the neo^R gene on the pGAPDH-EGFP plasmid.

Hybridization of this probe to the membrane also followed the ThermoFisher Southern blotting protocol. This involved preparing a pre-hybridization solution of 6X saline sodium citrate (SSC) blotting buffer (0.9 M NaCl, 0.09 M sodium citrate), 5X Denhardt's solution (0.1% bovine serum albumin, 0.1% Ficoll, 0.1% polyvinylpyrrolidone), 50% formamide, and 0.5% sodium dodecyl sulfate (SDS). Salmon sperm DNA was denatured by heating at 96 °C for 5 minutes, then placing on ice. The denatured DNA was added to the pre-hybridization solution to a concentration of 50

μg/mL, and the membrane was incubated in this mixture for 2 hours at 42 °C with agitation. The probe was then denatured at 96 °C for 5 minutes and placed on ice, then added to fresh pre-hybridization solution with no salmon sperm DNA to a concentration of 50 ng/mL to create the hybridization solution. The pre-hybridization solution with salmon sperm DNA was discarded and the membrane was incubated in the hybridization solution for 12 hours at 42 °C with gentle agitation. The membrane was then washed twice with 2X SSC + 0.1 % SDS for 10 minutes each time at room temperature, and washed for high stringency in 0.1X SSC + 0.1% SDS for 10 minutes each time at 65 °C.

The probe was detected using the Thermo Scientific Biotin Chromogenic Detection kit. This kit differs from traditional detection methods in not needing to develop a film; a coloured product is deposited directly onto the membrane wherever probe is present. Briefly, it involves binding a streptavidin-alkaline phosphatase conjugate to the biotinylated probe, and immersing the membrane in a substrate solution that is converted to a coloured precipitate by the alkaline phosphatase. The detection reaction was performed according to the manufacturer's protocol. The protocol allows for a short 30-minute detection reaction or an overnight reaction for more intense colour; the overnight reaction was used in this experiment.

3.2.6 Single cell isolation to establish clonal transformant lines

I developed a method to isolate single amoebal cells and grow monoclonal cultures from them, which was inspired by the experiments of Neff during the initial isolation and characterization of this strain[27]. First, Page's Amoeba Saline (PAS) was prepared, and agar was added to a concentration of 1%, then the mixture was autoclaved

and PAS-agar plates were poured. Overnight cultures of *E. coli* K-12 were grown in liquid LB medium, and then the *E. coli* were heat-killed by incubating the cultures in a 65 °C water bath for 20 minutes. These heat-killed bacteria were then pelleted by centrifugation at 2500 *g* for 5 minutes at 8 °C. The media was decanted and replaced with an equal volume of PAS, and the cells were resuspended by gentle vortexing. Then, a 500 µL aliquot of the *E. coli* suspension was added to the surface of each plate to be used for the experiment and was spread until dry. A 1 µL aliquot of *Acanthamoeba* culture was added to the centre of each plate and left for 3 days at room temperature.

The plates were checked with an inverted microscope at this point to determine how far cells had migrated from the centre. Visualization is aided by trails left through the bacterial lawn behind the amoebae. Once cells had migrated such that the furthest ones from the centre were not within 1 cm of any other cells, typically within 3 to 5 days of inoculating the plates, 1 cm^2 squares were cut out of the agar bearing one cell on each, and these agar squares were placed in wells of a 12-well culture plate. The wells of the plate were filled with enough Neff medium to cover the agar squares and the isolated cells were left to grow at room temperature in these covered plates. Once the monoclonal populations had grown dense enough to see cells distributed across the entire area of the bottom of the wells, the wells could be scraped with a cell scraper and the contents transferred into 75 cm^2 tissue culture flasks for typical culture maintenance. Polymyxin B was added at a concentration of 10 µg/mL to ward off the growth of any *E. coli* that may have survived heat-killing. The cells being isolated were transformed, but selection was not resumed in the 12-well plates immediately following isolation. Selection was instead restarted once growth was established in their standard growth conditions. In this

experiment, I isolated seven clones from the previously generated population of transformants for subsequent analysis.

3.2.7 Nanopore sequencing clonal isolates of transformants

Three of the seven clones were selected for nanopore sequencing, based simply on which three reached the appropriate density first. They are hereafter referred to as ‘Clone 1’, ‘Clone 5’, and ‘Clone 8’, based on their original labels at the time of single cell isolation. DNA was extracted from these clones using the same method as described above. A barcoded sequencing library was prepared using the ligation sequencing kit (SQK-LSK109) and the native barcoding kit (EXP-NBD104) with barcodes 2, 5, and 8. The library was run on a FLO-MIN106 flow cell and then basecalled with Albacore v2.3.3.

A subsequent sequencing run included only Clone 1 and was sequenced with the ligation sequencing kit (SQK-LSK109) on a FLO-MIN106 flow cell. The raw output was basecalled with Guppy v3.1.5 using the HAC (high accuracy) basecalling model.

Plasmid-containing reads from each of these sequencing datasets were identified using BLASTn and retrieved. They were mapped against the wild-type genome using minimap2[195] v2.24, with soft clipping allowed. Then, mapped reads were visualized in IGV to locate putative chromosomal integrations of the plasmid, and BLASTn was used to confirm that the soft-clipped portion of these reads was plasmid-derived.

3.2.8 Polymerase chain reaction to verify chromosomal integration of transgenes

Four putative integration loci were selected for verification with PCR. These PCR experiments sought to amplify across the plasmid-genome junctions on each end of a putative integration, but would not attempt to span any integrations. Where possible, the long read representing the putative integration was used as a template for primer design, but the genomic and plasmid segments of the read were exchanged for the corresponding sequences from the genome and plasmid assemblies, respectively. This was intended to prevent problems with the primers due to inaccuracies in the raw nanopore read sequence. At least one long read was available as template for both junctions of two of the chosen loci, and only one junction of the other two loci. For junctions that were not represented in sequence data but were predicted to exist at the opposite end of a putative integration, multiple primers were designed on both the plasmid and the genome sides to account for potential small deletions at the end of either prior to integration. Primer pairs were designed to amplify between 500 and 1000 bp, with the exact length dictated by optimizing primer characteristics. The melting temperature of these primers ranged from 52.1 °C to 58.5 °C. For the predicted junctions with no pre-existing sequence data, there were three primers designed for each side spaced roughly 50 to 100 bp from one another, while still attempting to respect an overall amplicon length of 500 to 1000 bp.

Early PCR experiments used NEB One *Taq* polymerase in its Quick-Load 2X master mix with standard buffer. These experiments started with the following PCR program: 96 °C for 5 minutes, then 25 cycles of denaturation at 96 °C for 30 seconds, annealing at 50 °C for 30 seconds, and extension at 72 °C for 1 minute, then a final 7-

minute elongation step at 72 °C. To adjust for poor amplification, the number of cycles was then increased to 35. Then, to reduce the generation of non-specific product, the annealing temperature was increased to 56 °C.

One of the putative transgene integration loci showed promising results and PCR was further optimized using ThermoFisher Platinum II *Taq* Hot-Start DNA polymerase. First, the standard recipe recommended in the documentation of this polymerase was used without the optional GC enhancer reagent. This polymerase is designed for a universal annealing temperature of 60 °C and extension at 68 °C, so the PCR program was as follows: an initial denaturation step of 2 minutes at 94 °C, followed by 35 cycles of denaturation at 94 °C for 15 seconds, annealing at 60 °C for 15 seconds, and extension at 68 °C for 15 seconds. There was no long final extension step in this protocol. It was then found that yield was greatly improved by including the Platinum GC enhancer reagent included with this polymerase, so it was used for the final reactions.

Bands of interest were gel-extracted using a Macherey-Nagel NucleoSpin Gel and PCR Clean-up kit, and samples were sent to GeneWiz, South Plainfield, New Jersey for Sanger sequencing.

3.2.9 Repeating the transformation experiment with linearized plasmid

Another transformation experiment was performed identically to the one previously described, except the plasmid was linearized by a single cut with the restriction endonuclease *Psi*I. In this experiment, monoclonal cultures were established immediately after the population of transformants recovered fully from transformation and selection. Three of these clonal isolates were sequenced and analyzed using the same

library preparation and bioinformatic approach as Clones 1, 5, and 8 described above, but basecalling was done with Guppy v5.1.13 using the SUP (super accurate) model. The three clonal isolates from this transformation experiment are hereafter referred to as ‘Clone LT6’, ‘Clone LT8’, and ‘Clone LT9’.

3.2.10 Determining the rate of artefactual read chimerism in transformant sequence data

Before estimating how much read chimerism may have affected the inference of transgene integration, the background rate of read chimerism had to be estimated from each sequencing run used to infer integration. Bruce Curtis, a research associate in the Archibald lab at the time, devised a method for estimating read chimerism, which I applied to the sequence read sets from this study. This was done using the non-plasmid-containing reads, so that the calculated rate of chimerism could be applied to assess the veracity of the plasmid+genome reads. Reads from an individual sequencing run were mapped against the wild-type reference genome sequence with minimap2[133] v2.24 and the mapping was output as a paf file rather than a SAM file as is usually done. All reads that had exactly two mappings to the reference and no plasmid-derived sequence were retained for further analysis, because depending on the genomic location of these two mappings, reads with this mapping pattern could be chimeras from two different genomic loci.

A custom Perl script was used to identify putatively chimeric reads. This script took the reads with exactly two mappings to the reference genome and assessed as follows: if the two mappings for a read were found to be greater than 500 bp apart on the genome, each was more than 100 bp away from the end of a scaffold, and the distance

between the two mappings on the genome was at least 100 bp more than the distance between the two mappings on the read, the read was considered to be chimeric. To estimate the proportion of chimerism in a given sequencing run, this number of putative chimeric reads was divided by the total number of reads fed into the chimerism identification script.

The number of putative chromosomal integrations of the plasmid that could have been artefacts due to read chimerism was estimated using this proportion. The total number of plasmid-containing reads was multiplied by the proportion of reads expected to be chimeric, and the resulting value represented the expected number of reads to artefactually appear as putative integrations due to chimerism. This number was compared to the observed number of putative integrations from the same read set. This process had to be done independently for the output of each sequencing experiment to account for differing rates of chimerism across different sequencing runs.

3.2.11 Re-sequencing and analysis of select clones with Illumina technology

Genomic DNA was extracted from wild-type *Acanthamoeba* as well as clones LT6 and LT9 and sent to the Integrated Microbiome Resource at Dalhousie for Illumina library preparation and sequencing. The three samples were sequenced as part of a larger multiplexed run on an Illumina NextSeq2000 instrument. This sequencing experiment was configured to produce 150-bp paired-end reads. Upon receiving the sequence data, the reads were processed with Trimmomatic[196] v0.36 using the following trimming parameters: HEADCROP:10, LEADING:12, SLIDINGWINDOW:4:20, MINLEN:75.

Illumina reads from Clone LT6 and Clone LT9 were used to seek confirmation of putative integrations found with nanopore sequencing. The Illumina reads from each respective clone were mapped against the putative integration-supporting long reads from the same clone using HISAT2[197] v2.2.1 with default mismatch penalties, as well as with the maximum and minimum mismatch penalties decreased from 6 and 2 to 5 and 1, respectively. These mapped reads were visualized in Integrative Genomics Viewer[198,199] to look for reads mapping across the putative plasmid-genome junctions. Illumina reads were also compared against the putative integration-supporting long reads using BLASTn to see whether any Illumina reads showed BLAST hits that spanned the junctions.

3.2.12 Assembling transformant data to search for plasmid-bearing contigs

Flye[128] v2.9 was used to assemble the Clone LT6 nanopore reads, the Clone LT9 nanopore reads, and selected subsets of each, all in independent assembly runs. The subsets for each clone were all reads containing plasmid sequence, and all reads containing plasmid sequence with telomeric repeats on at least one end. Flye was run in its standard mode and in metagenome mode for each set of reads for each clone (so the total output was 12 different assemblies). BLASTn was then used to identify contigs in each assembly containing plasmid sequence, the bounds of the plasmid sequence on those contigs, and the identity of any non-plasmid sequence, if present.

3.2.13 Southern blot analysis to detect an episome containing transgenes

For this experiment, the same protocol and choice of probe were used as above,

but the input DNA and gel were different. The gel in this experiment was 0.5% agarose to better resolve large DNA fragments. Genomic DNA from wild-type *Acanthamoeba* and from Clone LT9 were used, as well as purified pGAPDH-EGFP plasmid. For each sample, one aliquot was left untreated, while a second aliquot was digested with the restriction endonucleases *Nhe*I, *Not*I, and *Sac*I. The samples were run on the 0.5% agarose gel along with GeneRuler 1 Kb ladder. The DNA was visualized with ethidium bromide, and a Southern blot experiment was performed using the same protocol as described above.

A pulsed-field gel electrophoresis (PFGE) experiment was performed on wild-type and Clone LT9 genomic DNA, with one ladder of lambda concatemers (48.5 Kbp to 1,000 Kbp) as well as the GeneRuler 1 Kb ladder. One lane contained both ladder types. The PFGE run used a 1% agarose gel, 0.5X TBE running buffer, an 18-hour run time, a 120 ° included angle, and a switch time ramping from 1 second to 14 seconds over the course of the run. The gel was stained with ethidium bromide to visualize the DNA.

3.3 Results

3.3.1 Transformation is successful and persists under selection

In adopting this transformation method for *A. castellanii*, my first goal was naturally to verify the success of transformation. After following the recovery protocol, the transformed cell population under selection with 50 µg/mL G418 exhibited a growth rate comparable to that of a wild-type culture with no selection, based on qualitative observation. Upon examination with epifluorescence microscopy, cells could clearly be seen to be expressing enhanced green fluorescent protein (EGFP), although at different

intensities among individual cells. Figure 3.2 demonstrates clearly visible green fluorescence, including the aforementioned variability among cells. This figure demonstrates successful transformation using two different plasmids that only differ in the EGFP promoter, but only pGAPDH-EGFP, represented in the lower panels, was included in this study. In this initial transformation experiment, the circular form of the plasmid was used. There appears to be strong localization of the fluorescent protein to the nucleus. It is unknown why this is the case, or whether it has any biological impact on the transformed cells.

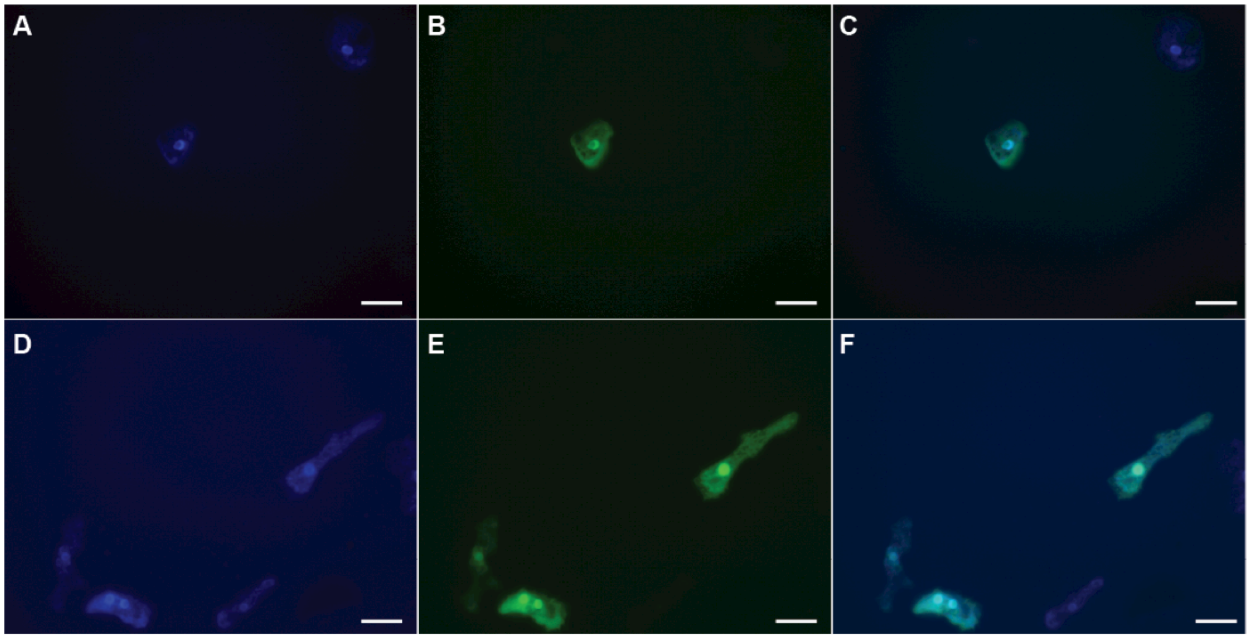


Figure 3.2 *A. castellanii* cells transfected with pTPBF-EGFP and pGAPDH-EGFP. *A-C.* Transfected with pTPBF-EGFP. *A.* Cells were imaged using fluorescence microscopy with a DAPI filter to visualize DNA stained with DAPI. *B.* Cells were imaged using fluorescence microscopy using a GFP filter to visualize EGFP expression. *C.* A merge of the DAPI and GFP images shown. *D-F.* Transfected with pGAPDH-EGFP. *D.* Cells were imaged using fluorescence microscopy with a DAPI filter to visualize DNA stained with DAPI. *E.* Cells were imaged using fluorescence microscopy using a GFP filter to visualize EGFP expression. *F.* A merge of the DAPI and GFP images shown. *A-F.* All cells were maintained under selection with 50 μg/mL G418, then fixed on slides with 4% formaldehyde, permeabilized with 0.5% Triton-X100, and stained with 300 nM DAPI for 5 minutes, then washed with phosphate-buffered saline before mounting with Fluoromount. All cells were imaged using 630X total magnification. All scale bars represent 20 μm.

3.3.2 Plasmid sequence can be detected by sequencing transformants

In an effort to exploit the advantages of nanopore sequencing as a high throughput method for molecular biology experimentation, I sequenced this culture of transformed *A. castellanii* on an Oxford Nanopore MinION device. The output of this sequencing run is summarized in Table 3.1. The goal of this sequencing experiment was to look for evidence of the transgenes that had been introduced and determine in what context they were being maintained. BLAST searches with the pGAPDH-EGFP sequence as the query allowed me to retrieve all the individual long reads that may demonstrate the fate of the internalized plasmids. In the case of chromosomal integration, I was looking for long reads containing both plasmid and genomic sequence, where at least one junction was represented. A first examination of these data revealed no such junctions. I found 9,579 reads with identifiable plasmid sequence, which were either flanked on both ends by sequence that was not recognizable, or by unrecognizable sequence on one end and telomeric repeats on the other. In total, 118 of the 9,579 had telomeric repeats on at least one end, with repeats on both ends of a single read. A much later re-analysis that employed read mapping against the reference genome did find 33 reads that appeared to contain both genomic and plasmid sequence. Notably, many of the full set of plasmid-containing reads contained tandem repeats of the plasmid sequence rather than a single copy, regardless of the flanking sequence. Arrays of up to 11 copies were observed on reads in excess of 65 Kbp; the existence of even longer arrays may have been obscured by read length limitations.

Table 3.1 Sequencing statistics for all runs performed in this study. The Clone 1 (shallow) read set was barcoded with Clone 5 and Clone 8 in the same run, while Clone 1 (deep) came from a separate run where Clone 1 was the only sample. Clones LT6, LT8, and LT9 were barcoded and sequenced together. The mixed transformant sample was the only sample in its sequencing run. All of the runs used FLO-MIN106 flow cells, and all used the SQK-LSK109 sequencing kit, except the mixed transformants which used the previous version, SQK-LSK108. The mixed transformants and Clones 1, 5, and 8 were transformed with circular plasmid, while Clones LT6, LT8, and LT9 were transformed with a linearized form of the same plasmid.

Clone	Number of reads	Total bases	Mean read length	Mean read quality	Median read length	Median read quality	Read length N50	Longest read
Mixed transformants	151,506	1,825,008,054	12,045	11	9,821	11	15,287	83,320
Clone 1 (shallow)	124,327	423,236,197	3,404	8.8	1,547	9.4	7,662	60,490
Clone 5	96,988	289,944,147	2,989	8.6	1,360	9.2	6,684	59,098
Clone 8	83,400	307,873,385	3,691	8.6	1,590	9.3	8,837	68,412
Clone 1 (deep)	708,542	7,730,398,077	10,910	10.9	5,733	11.8	22,989	393,346
Clone LT6	440,212	2,569,220,207	5,836	12	2,739	12.1	13,490	93,794
Clone LT8	748,516	4,241,519,907	5,666	11.9	4,298	12.1	9,560	154,837
Clone LT9	388,980	2,353,296,042	6,049	12.1	2,536	12.2	15,946	124,150

3.3.3 Southern hybridization identifies transgenes on high molecular weight species

To gain additional perspective on the fate of transgenes in this mixed population of transformants, experimental evidence was gathered. Genomic DNA from the mixed transformant population was extracted and run on an agarose gel in three forms: untreated, digested with *Bgl*II, and digested with *Hin*dIII. Each of these restriction endonucleases had 6 base-pair cut sites and cut a single time within the plasmid sequence. This gel was subsequently used for a Southern blot with a probe against the neomycin resistance marker gene, neo^R. The idea behind this experiment was to visualize whether any small molecule (under the ~30 Kbp resolution limit of a 1% agarose gel) was apparent in the undigested DNA sample that would clearly indicate episomal maintenance of the transgenes, and to provide an additional line of evidence that the transgenes are represented in the genetic complement of the cells. The results are shown in Figure 3.3. Briefly, there are no DNA species in the undigested lane that can be resolved from the chromosomal DNA; a single band containing all the high molecular weight DNA appears just higher than the top 20 Kbp ladder band, and the probe hybridizes to this band. In the two genomic DNA digest lanes, a variety of restriction products appear on the gel, but the Southern hybridization shows very strong signal to a band consistent with the digested plasmid control, with additional weak signal to bands slightly larger and smaller. This result is consistent with the presence of tandem arrays of the plasmid, with the slightly larger and smaller bands possibly corresponding to the ends of the array where one of the cut sites is somewhere near the array in the host genomic sequence.

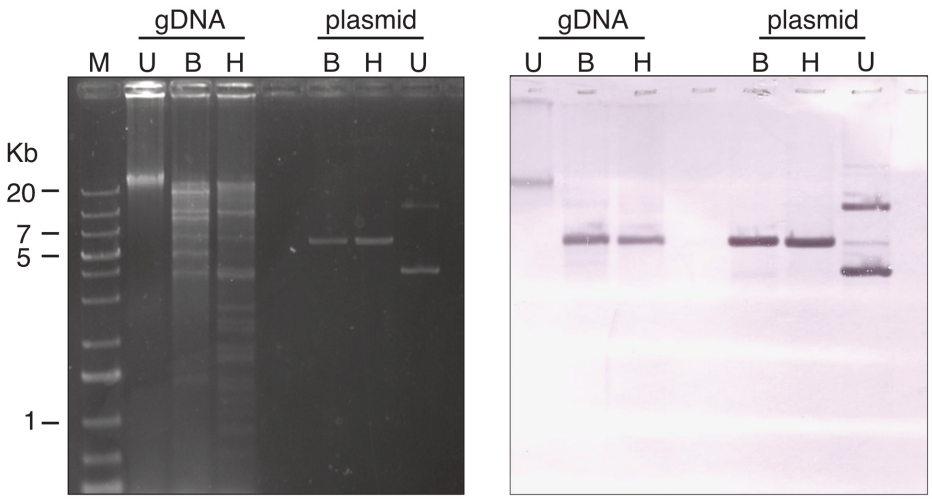


Figure 3.3 A Southern blot to detect the neoR gene in transformed *Acanthamoeba castellanii* genomic DNA. The agarose gel on the left was used for the Southern blot on the right. Genomic DNA (gDNA) from *Acanthamoeba* transformed with pGAPDH-EGFP was run undigested and two aliquots from the same sample were digested with each of two different restriction endonucleases, *Bgl*II (B) and *Hin*dIII (H). These endonucleases each have a 6 base pair recognition site that occurs once within the plasmid. The purified pGAPDH-EGFP plasmid was run as a control, with the same three treatments. The ladder in the leftmost lane is GeneRuler 1 Kb plus DNA ladder, with select band sizes marked alongside. The DNA was transferred to the nylon membrane on the right through a standard alkaline transfer and the neoR probe was hybridized to it. The probe was visualized directly on the membrane with the Thermo Scientific Biotin Chromogenic detection method.

3.3.4 Monoclonal cultures can be established for further sequencing

To eliminate the uncertainty that may arise from having a heterogeneous population of transformants in the same culture, a method was developed to isolate single cells and establish monoclonal cultures. This involved pouring agar plates with a lawn of heat-killed *E. coli* spread on the surface and adding a 1 μL suspension of amoebal cell culture in the centre of the plate. After three to five days, the plates were checked under an inverted microscope and cells had migrated outward from the centre of the plate, leaving visible trails in the bacterial lawn as they went (Fig 3.4). These trails were followed to the point where individual amoebae had become isolated from the rest and a 1 cm^2 square bearing a single cell could be cut out of the agar and placed in culture media in 12-well culture plates. After about a week, cells had reached late log phase or stationary phase and could be transferred into their usual 75 cm^2 tissue culture flasks and maintained as monoclonal strains.

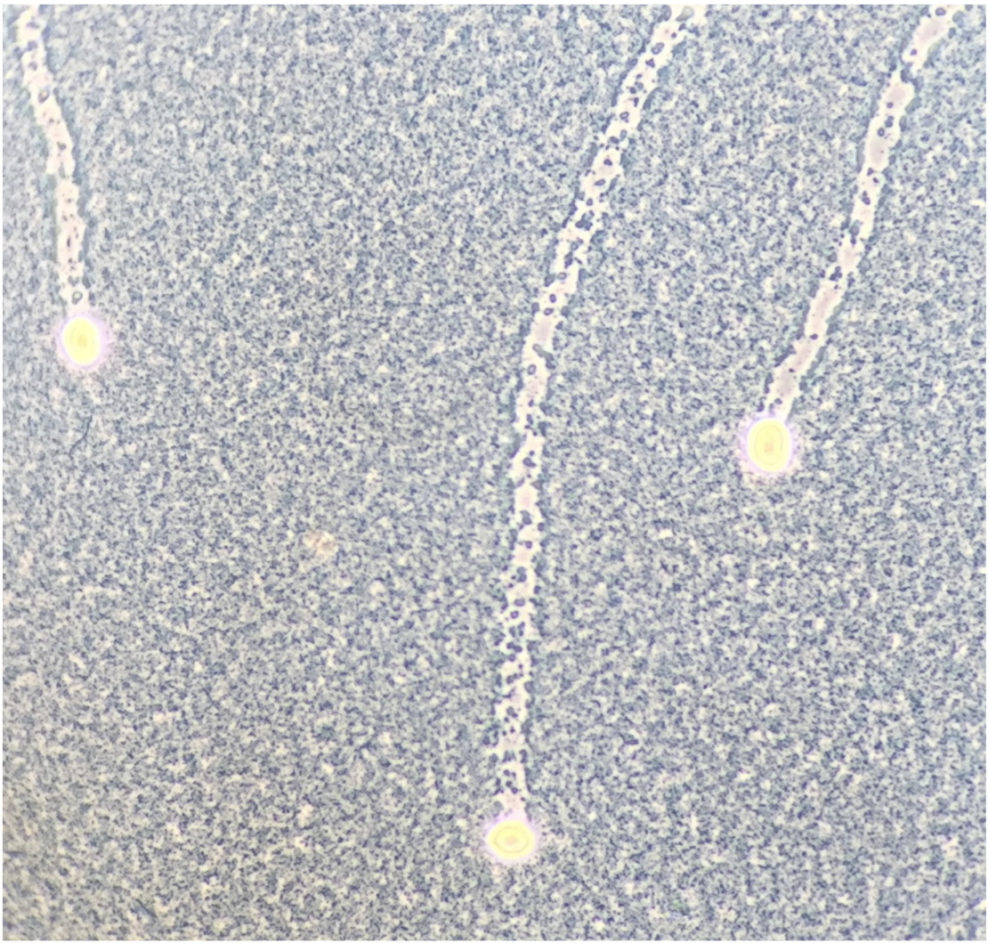


Figure 3.4 *Acanthamoeba* cells leaving trails as they traverse a lawn of *E. coli* on the surface of an agar plate. An overnight culture of *E. coli* K-12 was heat-killed and 1 mL was spread on a Page's amoeba saline-agar plate. A 1 µL aliquot of *Acanthamoeba* culture was placed in the centre of the plate and left to migrate for 3 days. The cells pictured are near the leading edge as the population migrates outward over time. The photo was taken down the eyepiece of an inverted microscope at 200X magnification with phase contrast.

3.3.5 Nanopore sequencing monoclonal transformant lines

Three of the monoclonal transformant lines were selected for DNA extraction and nanopore sequencing, based simply on which three reached the appropriate cell density first as determined by visual inspection. These three samples were barcoded and sequenced together in a single MinION run. The output of this sequencing experiment is summarized in Table 3.1. To screen for potential chromosomal integrations of the transgenes, all nanopore reads with hits to the full plasmid sequence were mapped against the wild-type reference genome. Soft clipping was allowed when mapping these reads, which means if a partial alignment could be found between a read and the genome, that region of the read would be mapped and the remainder would be ignored, rather than discarding the read entirely. This allowed genome-plasmid junctions to be observed in a high-throughput fashion by visualizing the mapped reads in Integrative Genomics Viewer[198] and then verifying that the clipped part of the read was actually plasmid-derived. One of these monoclonal isolates, ‘Clone 1’, was chosen for further investigation. Due to the low sequencing depth of coverage from the first sequencing experiment, DNA was extracted from this clone again and nanopore sequenced without other barcoded samples. The output of this sequencing experiment was much better than the previous run, summarized in Table 3.1. The same approach as above was used to detect putative genome-plasmid junctions in this more substantial dataset. One particular putative integration locus stands out; it is completely spanned by seven reads, which extend into the unique genomic sequence on both sides. The integrated DNA is only a small, 365 bp fragment of the plasmid, coming from the protein coding sequence for EGFP.

Notably, many of the reads demonstrating putative integrations contained tandem repeats of the plasmid sequence rather than a single copy. In the read set, the copy number of these tandem arrays ranged from two copies to 11 copies, although longer arrays may exist but be obscured by read length limitations.

3.3.6 PCR confirms a single putative transgene integration

The PCR experiments reported here were planned by me and Cédric Blais, who was an Honours student in the Archibald lab at the time. Under my supervision, Cédric performed all the benchwork involved in these PCR experiments. Of the four putative integration loci selected for PCR amplification, only one was successfully confirmed, and only on one end of the integrated plasmid array. This locus was the one covered by seven long reads that fully spanned the integration. Amplicons from this successful reaction were sent to GeneWiz for Sanger sequencing which confirmed their identity and therefore supported this integration, at least in part. For the plasmid-genome junctions of the other end of this locus and both ends of the other three loci, PCR experiments either showed no amplification or non-specific products. This tended to depend on how stringent or permissive the conditions were for a given reaction; we started with stringent parameters to target only our desired locus but no amplification was observed, and as we relaxed the stringency by lowering annealing temperatures, increasing the number of cycles, and adding more input DNA, non-specific product began to appear.

3.3.7 Transformation and sequencing are successful with linearized plasmid

In an effort to test whether rolling-circle replication was involved in the formation of plasmid concatemers, the transformation experiment was repeated with plasmid that had been linearized by the restriction endonuclease *Psi*I, which cuts a single time in the pGAPDH-EGFP plasmid. Monoclonal isolates were derived from the population of successful transformants in the same manner as before. DNA was extracted from three clones, and all three samples were barcoded and sequenced together in the same MinION sequencing experiment. The output from this sequencing experiment is summarized in Table 3.1. This time, the barcoded sequencing run was successful enough to use data from all three clones to infer potential chromosomal integrations of the plasmid. The number of inferred integrations from the deeply sequenced clone transformed with circular plasmid and the three clones transformed with the linearized plasmid are presented in Table 3.2. The number of inferred integrations in all four isolates was surprisingly large, with 1,251 putative integrations in the clone transformed with circular plasmid, and 280, 187, and 109 inferred integrations in the three clones transformed with linearized plasmid.

3.3.8 Determining the rate of chimerism in nanopore sequencing experiments

Read chimerism, where the sequences of DNA molecules that were not physically contiguous *in situ* appear together in a single read, is a known artefact in nanopore sequencing data reported to affect roughly 2 to 8% of reads (ref. 200 and refs within [200])

compared to an estimated 0.5 to 2% of chimeric reads in Illumina sequence data[201]. To evaluate how many putative plasmid-genome junctions were the result of read chimerism, Bruce Curtis and I estimated the background rate of read chimerism in transformant sequencing datasets and applied it to the total number of plasmid-containing reads. This analysis showed that a strikingly high number of the putative chromosomal integrations of the transgenes could potentially be artefacts generated by read chimerism. Of the 1,251 inferred integrations in Clone 1, 936 could be attributed to read chimerism, in Clone LT6 where 280 integrations were inferred, read chimerism was calculated to explain up to 336, in Clone LT8 up to 438 putative integrations could be explained by chimerism while only 187 were inferred, and in Clone LT9, 185 putative integrations could be chimeric, while only 109 were inferred. The number of putative integrations that could be attributed to chimerism was an extrapolation based on the estimated rate of chimerism and the number of plasmid-containing reads in total. This means that no specific putative integrations could be identified as the artefactual ones, and is why the expected number of artefactual integrations can exceed the number of integrations that were inferred.

To assess the veracity of these putative integrations from another angle, I looked for putative integration loci that were supported by two or more nanopore reads. In accordance with the read chimerism findings, there were very few putative integrations with more than one read supporting them. These investigations are summarized in Table 3.2. In Clones 1, LT6, LT8, and LT9 respectively, there were 25, 0, 2, and 1 inferred integrations with 2 or more reads supporting them.

Table 3.2 Summarized nanopore evidence for the fate of transforming DNA from four *Acanthamoeba* clones. Plasmid topology refers to whether the clone was transformed with circular pGAPDH-EGFP or pGAPDH-EGFP linearized by *Psi*I. Apparent integrations are loci where one or more nanopore reads contain genomic sequence from that locus joined to plasmid sequence. The predicted number of artefactually chimeric junctions is determined by calculating the background rate of chimeric reads in a sequencing run, and applying this rate to the number of plasmid containing reads to estimate how many may be chimeric. The reads with one or two telomeres refer to reads where one or more plasmid copies are found in a read with telomeric repeats on one or both ends.

Clone	Plasmid topology	Estimated genome coverage	Inferred integrations	With 2 reads	With >2 reads	# of artefactual chimeras	1 telomere reads	2 telomere reads
C1	circular	100X	1251	25	1	936	3	288
LT6	linear	50X	280	0	0	336	3	85
LT8	linear	100X	187	2	0	438	1	32
LT9	linear	65X	109	1	0	185	0	45

3.3.9 Illumina sequencing transformed clones reveals extremely high plasmid abundance

Given the uncertainty surrounding nanopore-based evidence for putative integrations, the decision was taken to deeply sequence clones LT6 and LT9 with Illumina short read sequencing. For Clone LT6, 233,564,430 reads were obtained, totalling 30.2 Gbp of sequence, and for Clone LT9, 449,678,294 reads were obtained, totalling 56.2 Gbp of sequence. This sequence data was used to investigate and verify the putative integrations from the nanopore data. Two strategies were used to query the putative integrations with Illumina reads. First, the set of long reads representing putative integrations were collected into a single file for each clone, and the corresponding Illumina read set was mapped against these putative integrations. Then, the mapping was visualized and investigated for short reads mapping across putative plasmid-genome

junctions. The second method involved BLAST comparisons of the plasmid, the putative LT6 and LT9 integration reads, and the total set of LT6 and LT9 Illumina reads.

Neither method identified any Illumina reads that supported the putative integrations of plasmid DNA into *Acanthamoeba* chromosomes. With the concern that the Illumina dataset may be unexpectedly depleted in plasmid sequence, leading to the observed lack of evidence for integration, I mapped the LT6 and LT9 Illumina reads onto the sequence of the pGAPDH-EGFP plasmid. Table 3.3 summarizes the finding of this mapping experiment. It is clear that the plasmid was present in some form in extremely high abundance in both LT6 and LT9 cells, but there was a striking absence of evidence for how it was being maintained.

Table 3.3 Illumina and nanopore read coverage depth in Clones LT6 and LT9 when mapped against the wild-type *Acanthamoeba castellanii* Neff genome and against the pGAPDH-EGFP sequence. Coverage values have been rounded to the nearest 5 for genome coverage and to two significant figures for the plasmid coverage for ease of comparison.

Clone	Genome Nanopore	Plasmid Nanopore	Genome Illumina	Plasmid Illumina
LT6	50X	12,000X	485X	150,000X
LT9	65X	18,000X	910X	270,000X

To try to explain this finding, I revisited a hypothesis that had been set aside earlier in the investigation because it had not been the best fit for the evidence at the time. I searched the nanopore reads from all the transformant clones for reads that had one or more plasmid copies flanked on both ends with telomeres, such that they represented a transgene-containing ‘minichromosome’. I found at least one of these in each of the deeply sequenced transformant clones, ranging from about 30 Kbp to 60 Kbp in length,

which suggests roughly 5 to 9 plasmid copies in tandem. Since the counts of telomere-flanked reads were very low, I identified all plasmid arrays with telomeres on at least one end to account for the possibility that most reads were simply not long enough to capture both telomeres. The tabulated counts of these reads for each clone are presented in Table 3.2.

3.3.10 Molecular search for potential transgene minichromosome

Having found that the potential minichromosomes in these clonal isolates ranged from 30 to 60 Kbp, above the expected limit of resolution of a 1% agarose gel, I decided to revisit this experiment using a lower percentage agarose gel. While 0.3% agarose is often used to resolve large fragments, it was found to be practically infeasible due to the gel being too soft to manipulate. A 0.5% agarose gel was used for these experiments instead.

The approach to restriction digestion for this experiment varied somewhat compared to the earlier Southern blot experiment as well. Rather than two different digests, each using a single six-cutting restriction enzyme with one cut site in the plasmid, the idea was to degrade away any wild-type genomic sequence as much as possible without cutting up any plasmid arrays. For this, *Sac*I, *Nhe*I, and *Not*I were chosen due to each one having at least a few thousand cut sites in the wild-type Neff genome, and none in the pGAPDH-EGFP sequence. More specifically, *Nhe*I cuts a predicted 2,565 times in the 35 chromosome-scale scaffolds of the wild-type Neff genome, for an expected average of roughly 17 Kbp between cut sites, while *Not*I cuts a predicted 3,790 times in those scaffolds for an expected average of roughly 11 Kbp between cuts, and *Sac*I cuts a predicted 24,049 times in those 35 scaffolds for an expected average of roughly 2 Kbp

between cuts. The gel was run with wild-type Neff gDNA, LT9 gDNA, and pGAPDH-EGFP. Each sample was run in undigested form as well as digested with the cocktail of restriction enzymes described above. A Southern blot was performed with this gel, with a probe against the neo^{R} gene from the plasmid. The gel and blot are visualized in Figure 3.5.

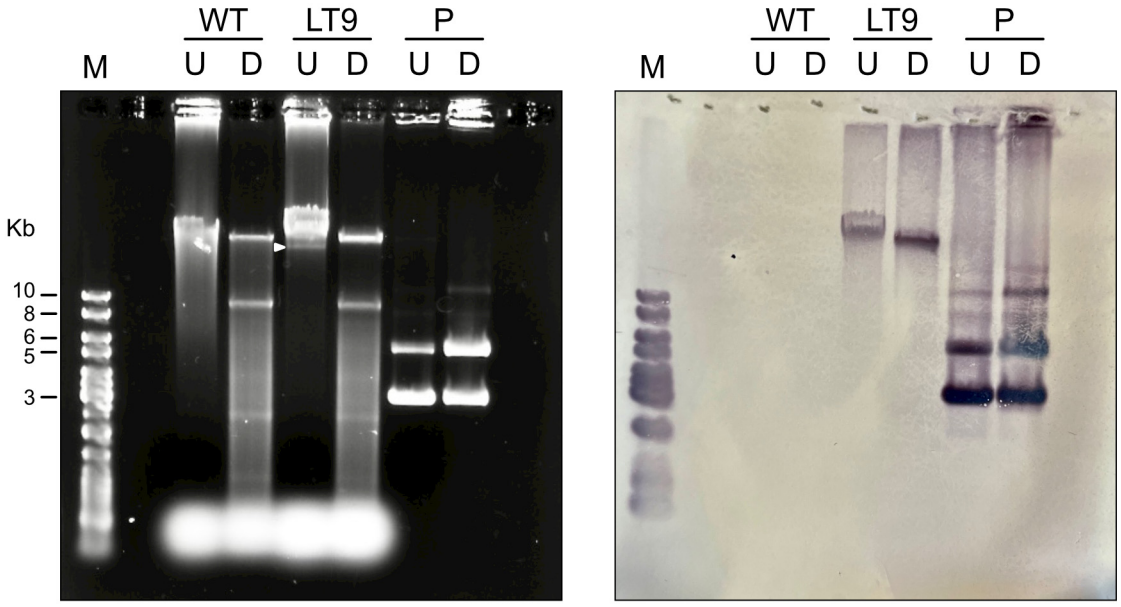


Figure 3.5 A Southern blot to look for extrachromosomal transgenes in transformed *Acanthamoeba* genomic DNA. The agarose gel on the left was used for the Southern blot on the right. Genomic DNA from wild-type (WT) *Acanthamoeba castellanii* strain Neff and Clone LT9 was run undigested (U) and digested with *Sac*I, *Nhe*I, and *Not*I (D). These endonucleases cut frequently within the wild-type genome and do not cut within the plasmid. The purified pGAPDH-EGFP plasmid (P) was run as a control, with the same treatments. The ladder in the leftmost lane is GeneRuler 1 Kb plus DNA ladder, with select band sizes marked alongside. An arrowhead marks a band that appears in undigested LT9 but not wild-type. The DNA was transferred to the nylon membrane on the right through a standard alkaline transfer and the neoR probe was hybridized to it. The probe was visualized directly on the membrane with the Thermo Scientific Biotin Chromogenic detection method.

From the gel, the WT and LT9 digested lanes appear to have approximately the same banding pattern. Interestingly, in the LT9 undigested gDNA lane there is a distinct extra band appearing just below the combined high MW DNA band. Given the

expectation for the transformants to have a 30 to 60 Kbp molecule bearing the transgenes, this distinct extra band appears to be a candidate, but the Southern blot shows no hybridization of the neo^R probe to this extra band, instead hybridizing to the high molecular weight DNA band. This means that the extra band appears not to be the hypothesized linear episome, but its identity is not clear.

3.3.11 Tandem arrays of the transforming plasmid are assembled but not in context

Given the large amount of nanopore sequence data generated for Clone LT9, I attempted to identify the new band by producing a genome assembly of the total Clone LT9 sequence data. I had previously not assembled the transformant data due to concerns that integrations may have had unforeseen structural impacts on the genome. However, genome assembly seemed like a reasonable approach to see if there was signal in the long read data for the band in question from the gel. Since the band appeared strongly on the gel, I looked for assembled contigs between 10 Kbp and 50 Kbp in length that were covered in multiple times more depth than the rest of the genome. Standard genome assembly with Flye[128] generated two contigs meeting these criteria. One was roughly 41 Kbp in length and covered 58 times more deeply than the nuclear genome but corresponded to the mitochondrial genome. The other was 5,849 bp, almost exactly the length of the transforming plasmid, which it was found to match. Using Flye in metagenome mode, which is better equipped to deal with large differences in coverage across different sequences, generated five contigs meeting the above criteria. The mitochondrial genome was once again represented, as well as three contigs of roughly 16 Kbp, 21 Kbp, and 31 Kbp, comprising plasmid arrays of various sizes. One contig of

roughly 24 Kbp contained two copies of the *Acanthamoeba* 18S sequence.

In a separate investigative effort, I employed Flye[128] to identify any contigs assembled from the Clone LT6 and Clone LT9 nanopore sequencing data that may provide more information on the fate of the transforming plasmid. Using Flye in its standard assembly mode did not produce any output with the full set of plasmid-containing reads from either clone, but Flye in metagenome mode produced an array of just over four copies of the plasmid with no other flanking sequence in LT9, and two partial plasmid contigs that together represented the full plasmid sequence in LT6. Using only the plasmid reads with terminal telomeric repeats, standard Flye produced a contig of roughly two concatenated plasmid copies in LT6, and one of roughly three concatenated plasmid copies in LT9, neither of which had any other flanking sequence. The same reads assembled in metagenome mode produced a contig of roughly four concatenated plasmid copies in LT6, and one of roughly three plasmid copies in LT9, once again with no additional flanking sequence.

3.4 Discussion

The goal of this study was to determine the fate of transgenes in *Acanthamoeba* after separate artificial transformation experiments with circular and linearized forms of the same plasmid. I used nanopore and Illumina sequencing, PCR, Southern blotting, and PFGE in order to determine whether the transgenes were episomal or chromosomally integrated, and if any modifications to them had taken place. On balance, the results of this investigation seem to favour the episomal maintenance of the transgenes in *Acanthamoeba* on a linear molecule that contains a tandem array of plasmid sequence

flanked by telomeric repeats. The evidence is not strong enough to conclude this with certainty, but of the hypotheses I have formulated, it is the best supported.

The results of this study unambiguously demonstrate that *Acanthamoeba* maintains the experimental transgenes in tandem arrays that probably range somewhere between 5 and 10 copies on average. It is not clear whether those bounds would consistently apply to all transformation experiments on this organism, or to natural acquisition of foreign DNA in either circular or linear form. Distinguishing between chromosomal or episomal maintenance of the transgenes is where the results become less clear. Using the size of the transgene-containing DNA species as a diagnostic criterion has been surprisingly inconclusive, especially with respect to the most common DNA sizing methods like agarose gel electrophoresis. The size of a putative episome, in the range of 30 Kbp to 60 Kbp based on sequence data, is much different than the size of most of the nuclear chromosomes in *Acanthamoeba* (reported in Chapter 2 of this thesis) although a couple of chromosomes at the lower end of the size range reach down to around 100 Kbp. However, the estimated size range is still above what can easily be resolved with standard agarose gel electrophoresis. Therefore, while Southern blot experiments have placed the transgenes in the same band on an agarose gel as the genomic DNA, the episome may not be expected to resolve as a separate band on these gels. Peng, Omaruddin, and Bateman[189] arrived at a similar conclusion when they first published the transformation protocol for *Acanthamoeba*, which used the circular form of the plasmid. They performed a Southern blot and concluded that the transgenes must be on a molecule of at least 12 Kbp and present in several copies, but they felt unable to differentiate between chromosomal integration, the formation of a linear episome, or a

circular molecule composed of multiple plasmid concatemers.

Using pulsed-field gel electrophoresis to resolve an episomal band from the genomic DNA did not provide additional insight; no clear band was observed in the size range that would be expected based on the sequence data (data not shown). A future experiment using a similar gel as substrate for Southern blots with probes against telomeres and the neo^R resistance marker may clarify the situation.

Overall, sequence-based analysis has arguably shifted the balance of probabilities in favour of episomal maintenance rather than strictly chromosomal integration through a combination of a few different findings. The first is that the evidence for chromosomal integration has always remained somewhat unconvincing. The discrepancy between the read support of any given putative integration and the overall genome coverage demands explanation; depending on the specific integration and the strain, this could be a 25- to 100-fold discrepancy, as the putative integrations had 1 or 2 reads supporting them, and the genome coverage of the clones ranged from 50X to 100X. The explanation held throughout much of this investigation was that the suspected polyploidy of *Acanthamoeba*, estimated to be ~25*n* by Byers[10], could account for this discrepancy if an integration took place into only one copy of a chromosome. However, finding a single supporting read for an integration with 100X genome coverage does stretch the bounds of what is thought to be known about *Acanthamoeba*'s polyploidy (see Chapter 4).

This explanation was weakened much further when the abundance of plasmid sequence in these cells was revealed. Both Illumina and nanopore sequence data showed that the coverage of a single copy of the plasmid was about 250 to 300 times higher than the genome coverage in clones LT6 and LT9. Even giving the integration hypothesis

maximum ‘benefit of the doubt’ does not account for this finding; assuming each of the 280 putative integrations in LT6 had an array of 11 plasmid units, the longest array I found in any of the data sets, explains 3,080 total copies of the basic plasmid unit, which is only 62 times higher than the overall genome coverage in the corresponding read set, falling short of the observed 250 to 300. This hypothesis is further weakened when the read chimerism analysis is considered. Based on the background rate of chimeric reads estimated in the LT6 nanopore sequencing run, chimerism could explain the appearance of up to 336 reads artefactually appearing to be junctions between plasmid and genomic sequence, which exceeds the number actually observed. It should also be noted that the criteria used to identify chimeric reads were relatively stringent; one could propose situations of true chimerism that would be rejected by the criteria implemented. This means that the estimated rate of chimeric reads, if inaccurate, would likely be an underestimate.

With reason to doubt the hypothesis of chromosomally integrated transgenes, revisiting the long read sequencing data of the transformed clones provided an alternative hypothesis. Finding telomere-flanked plasmid arrays in the reads from multiple transformed clones presented evidence for a molecule that could reasonably serve as an episome to bear the transgenes and promote their replication. This hypothesis does share some of the weaknesses of the integration hypothesis, but to a less fatal extent. Finding three or fewer reads in each clone where a plasmid array is flanked by telomeres on both ends suffers from the same issue of sparse representation in the sequence data. Even when accounting for the reads with telomeric repeats on only one end, the counts are quite low compared to the total plasmid abundance. However, the issue of read

chimerism is less problematic for this hypothesis. Some simple calculations would suggest that, for chimerism to account for all plasmid-telomere junctions observed in each clone, each chromosome would be expected to have telomeres in excess of 100 Kbp on each end. For comparison, mammalian telomeres are typically from 10 to 30 Kbp, and yeast telomeres are only a few hundred base pairs in length. This would also make each of *Acanthamoeba*'s telomeres at least the same length as the total predicted non-telomeric sequence of its smallest chromosome. Therefore, it seems more difficult to reject the existence of a linear episome composed of plasmid concatemers and flanked by telomeres than it is to reject most of the putative chromosomal integrations inferred from the sequence data.

All this said, there is a single chromosomal integration of transforming DNA in Clone 1 that clearly cannot be rejected. This is the only case where long reads fully span what were previously thought to be putative integrations, and is also supported by several times more reads than the others. It is also the only example where experimental support could be obtained. It is highly likely this is a genuine chromosomal integration. However, it is also unlike any of the other plasmid sequence found in the four clones in the fact that only a 365-bp fragment of the plasmid is integrated. This sequence comes from the EGFP gene and does not contain either of the termini required for proper expression, so it seems unlikely to have any biological influence on the cell, which includes not contributing to resistance against the selecting antibiotic in culture. Extrachromosomal copies of the transforming plasmid must have been responsible for antibiotic resistance in this clone, as is expected in the others. This integrated DNA would also not have been detected by the probe against the neo^R gene, but this particular clone was not probed experimentally

anyway.

There is compelling precedent for the type of linear episome I propose to exist in *Acanthamoeba*. A strikingly similar situation was observed in the fungal pathogen *Histoplasma capsulatum*. A series of experiments published by Woods and Goldman in 1992[202] and 1993[203] demonstrated the fate of transforming plasmids in *H. capsulatum* to be either chromosomally integrated or modified into a linear plasmid. In the case of the linear plasmid, the modifications closely mirror the ones I observe in *Acanthamoeba*. Woods and Goldman report duplication of the transforming plasmid and the addition of telomeric repeats followed by maintenance of this episome at high copy number. However, they also report instances of chromosomal integration with tandem amplification of the transforming plasmid, and transformation with circular or linear plasmid does not appear to influence these fates. Although only one chromosomal integration was observed in my study, the overall findings are nearly identical to those in *Histoplasma*.

While the example of *H. capsulatum* most strongly resembles that of *Acanthamoeba*, similar outcomes have been described across the fungal genetics literature, and even in other protists. This approach to maintaining transgenes depends on two apparent truths of the genome biology in this collection of organisms. The first is that while telomerase is recognized to function more efficiently when adding repeats to existing telomeres, it can also spuriously add repeats to other free ends of DNA. The second is that these organisms must either have sufficiently permissive replication systems such that no specific replication origin is needed, or they can use telomeres themselves as a replication origin; the literature suggests that one or the other of these

options can be true depending on the system.

Among fungi and beyond *H. capsulatum* (above), other species observed to form this type of linear episome include *Cryptococcus neoformans*[204,205] and *Fusarium oxysporum*[206]. The ciliate *Paramecium tetraurelia*[207] is observed to append telomeres to the ends of plasmids injected into its macronucleus. Of these examples, none exhibit tandem duplication of the transforming plasmid as clearly as *Acanthamoeba* or *Histoplasma*. In *F. oxysporum*, there appears to be partial duplication of the transforming DNA, as well as rearrangement and inversion of some fragments. The authors of that study believe tandem duplication could have facilitated those outcomes due to the instability of tandem repeats in fungi during mitosis. In *Cryptococcus neoformans* and *Paramecium tetraurelia*, the transforming DNA appeared to remain unit-sized in the autonomously replicating episome, with no suggestion of duplication. The addition of telomeric repeats to the transforming DNA in all three of *F. oxysporum*, *C. neoformans* and *P. tetraurelia* has been confirmed and appears not to require any particular motif such as existing telomeric repeats to template this modification. It is not clear how priming by the telomerase RNA occurs in these cases.

Among the fungal systems discussed here, all three exhibited the ability to maintain transforming DNA through both chromosomal integration and the formation of autonomous linear episomes. In fact, in each of *Cryptococcus neoformans* and *Fusarium oxysporum*, at least one clone was obtained where transforming DNA was found to be present both on a linear episome and integrated into a chromosome, while in *Histoplasma capsulatum*, several clones observed to have formed linear episomes from the transforming DNA were found to later integrate it into a chromosome. These findings

support the inference that ‘Clone 1’ in my study has generated a linear episome from the transforming DNA while also exhibiting one bona fide chromosomal integration. Whether this integration occurred independently of linear episome formation or whether a linear episome served as a substrate for this integration is unknown.

Another unknown is the mechanism for duplicating the transforming plasmid. While the authors that report duplication in fungal systems acknowledge the potential downstream effects of this duplication when it interacts with recombination machinery, they do not comment on which processes may have given rise to the tandem repeats in the first place. This is also a question for which there is not a clear answer in my study. However, due to the non-random orientation of the plasmid units within a given array, it is unlikely that random ligation of several transforming plasmids gave rise to these arrays. It seems to be the case that the transforming plasmid was concertedly duplicated in some way.

Surprisingly, the telomeres do not appear to join the plasmid array at the position where the original transforming plasmid was linearized. I attempted to use the nucleotide position on the plasmid where telomeres were joined as a proxy to estimate the diversity of episomes within each transformed clone. In each respective clone, there appeared to be one dominant form of the episome that comprised the plurality of the total pool of episomes, with a frequency between 20% and 50% depending on the clone, then one to a few minor forms with frequencies between 5% and 20%, and the remainder of the observed episome-like reads could be described as ‘singletons’ in the data. This would suggest each clone has a few different versions of the episome that are replicated somewhat more efficiently than the rest, with a large diversity of rare versions. It is

unclear whether this diversity of forms arose from a single episome due to instability, or whether multiple episomes were able to form at the time of transformation due to the large quantity of transforming DNA, and some are now held at higher copy number than others. The latter scenario seems plausible; the difference in frequency of the different forms may be caused by differences in replication efficiency due to unknown factors.

To explain why there are so many observed positions for telomeres to be joined to the plasmid, there are also at least two hypotheses. Perhaps the instability known to be intrinsic to tandemly duplicated sequences caused sufficient rearrangement to create this variation in plasmid-telomere junction sites. Examination of the episome-representing reads provides support for this hypothesis; the arrays are often composed of neatly duplicated plasmid copies but there are cases where there is significant fragmentation of the plasmid units within these episomes, presumably due to recombination. Another explanation could be that degradation of the ends of arrays, whether by passive or enzyme-mediated processes, caused the ends to vary until telomeres were added, protecting the plasmid arrays and preserving the structure that existed at that moment in time. Of course, these explanations would probably not be mutually exclusive, and both could contribute to the variation in plasmid-telomere junctions.

If the hypothesis favoured here is true — i.e., that *Acanthamoeba* maintains transforming DNA as an autonomous linear plasmid — this would have evolutionary implications as well as implications for molecular biology experimentation in this system. It would be prudent at this point to distinguish between *autonomy* and *stability* of these extrachromosomal elements. Autonomously replicating plasmids can be recognized and propagated by the endogenous DNA replication machinery, but this does not guarantee

their longevity across generations, which would be characterized as their stability. In non-integrated elements, selection must typically be maintained to prevent loss of the plasmid over time[208,209].

As a molecular biology tool, extrachromosomal elements are already employed by fungal geneticists, who have found that they can append telomeres to a variety of genetic constructs and have the cell maintain them autonomously. In many common laboratory fungal models such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and the pathogen *Candida albicans*, specific 'ARS' (autonomously replicating sequence) motifs must be present on plasmids to ensure maintenance[210–213]. In the other fungal examples discussed above, it is not clear if telomeres are sufficient for autonomous replication, if the replication machinery is simply permissive enough to operate on these elements anyway, or if the necessary sequence is present but still unidentified. In some studies, researchers found autonomous replication even when there was no endogenous fungal sequence on the plasmid aside from the telomeric repeats, suggesting that an ARS was not required[203]. Regardless, in all these systems, researchers have found that autonomously replicating plasmids provide great utility for genetic experiments by virtue of high transformation efficiency and high copy number; transforming DNA that cannot replicate autonomously has to be chromosomally integrated which is comparatively much less efficient. Some yeasts can maintain circular plasmids autonomously provided they have an ARS, but in many of these fungi, the addition of telomeres to linear plasmids creates highly efficient genetic constructs[203–205]. This knowledge could be used to develop genetic tools for highly efficient transformation in *Acanthamoeba* to study its biology more intensively. The same concept could possibly be applied to other nascent protist

model systems where transformation is difficult or not yet achieved.

The first evolutionary implication is that it would suggest a broader distribution of this genetic capability across eukaryotes than previously thought. With the exception of *Paramecium tetraurelia*, the formation of telomere-bearing linear episomes has only been observed in fungi. While Amoebozoa is the closest supergroup to Obazoa, which contains Fungi[214], this should still prompt molecular biologists working on protists from other parts of the eukaryote tree to be on the lookout for similar phenomena. The observation of this type of molecular biology in the two major groups of Amorphea, as well as in an alveolate, thought to be on the opposite side of the eukaryote root[215], could indicate that it is relatively widespread among eukaryotes.

The second evolutionary implication of this finding, especially if it did turn out to be eukaryote-wide, is the role it may play in fine-scale genome evolution. Many transformation protocols serve to facilitate access of the transforming DNA to the nucleus, but do not provide any clear impetus for the fate of the DNA once it has arrived there. There are exceptions to this, such as *Agrobacterium*-mediated transformation, where the bacterium encodes machinery for integrating the DNA[216], or transformations paired with gene editing systems like CRISPR-Cas9. The salient point, though, is that, given exogenous DNA has arrived in the nucleus, many transformation protocols may not differ from this point onward from the acquisition of exogenous DNA in a natural environment that may serve as substrate for lateral gene transfer (LGT). The spontaneous generation of autonomously replicating plasmids described in this study could allow exogenous DNA to wait in a sort of ‘holding pattern’ in the nucleus, which would greatly expand the window of opportunity for other requisite steps of LGT, such as chromosomal

integration and gene expression, to occur.

One can even imagine some of the steps involved in expressing the newly acquired genes to take place prior to integration; in fact, the unstable and recombinogenic nature of the tandemly duplicated transforming DNA in this study could facilitate the acquisition or invention of promoters and other necessary elements, if this tandem duplication also happened to naturally acquired DNA. Aside from providing a standing source of foreign DNA to later integrate, these autonomous plasmids could also be a substrate for innovation. They may even begin expressing genes that provide a selective advantage before chromosomal integration, encouraging maintenance of the new genetic material until it can be incorporated more permanently. This hypothetical model could help tip the scales in what is sometimes thought to be a highly improbable series of events.

3.5 Conclusions

This study has served as a productive exploration of both methodology and biology. On the methodological side, I have demonstrated how long-read, single molecule sequencing can serve as a high throughput readout for molecular biological experimentation. This allows for an exploratory approach to experiments that are traditionally designed to make relatively small genomic changes and receive narrow, targeted outputs. Excepting the issue of read chimerism, each of the thousands of long reads represents a specific DNA fragment that existed in situ and was detected by the sequencer. This allows for a variety of observations that would previously require molecular biology experiments to make, such as the existence of extrachromosomal

molecules, structural changes including chromosomal integration, and relatively unbiased stoichiometric comparison of different species. While still not trivial, the range of possibilities can be explored in these sequence read sets much more quickly than through experimental means.

However, the temptation to entirely replace molecular biology at the bench with sequencing should be resisted. In fact, I would strongly argue that classic molecular biology methods are underused and it is essential to accompany exploratory sequencing projects of the sort described above with robust experimental results to have confidence in the findings. While the sequence data presented here served as a rich substrate for generating hypotheses, this study clearly demonstrates how it also leaves room for ambiguities that are best resolved with more tangible wet lab experimentation. To bring my proposed model of the molecular biology involved from being the best supported of several plausible hypotheses to being the hypothesis that remains supported to the exclusion of all others, it would take deeper and more rigorous experimentation that directly falsifies the alternative explanations, rather than hoping enough additional sequence data would eventually contain the falsifying evidence somewhere within.

Moving from methodology to biology, this study's impact is founded directly on the finding that *Acanthamoeba* appears to duplicate transforming DNA into a tandem array and add telomeres to the ends such that it can be autonomously replicated in the nucleus. Before considering all other implications, this is a newly discovered genetic capability in this organism; the fate of transforming DNA was previously unknown, and there was no prior speculation of the ability to create any autonomously replicating elements, nor to add telomeres to extrachromosomal DNA. This would suggest that

Acanthamoeba's telomerase is capable of acting on substrates without existing telomeric repeats, and that its replication machinery, by some means, is able to replicate non-chromosomal DNA, potentially by recognizing the newly added telomeres. The results of this study have also not ruled out the possibility of chromosomally integrating exogenous DNA; in fact, they have shown *Acanthamoeba* has the genetic flexibility to retain foreign DNA either on autonomously replicating elements or by chromosomal integration, possibly simultaneously.

These capabilities also hint at the existence of a molecular gateway to facilitate bona fide lateral gene transfer. While the uptake of foreign DNA would differ from the way it is introduced in a transformation experiment and could follow many trajectories to the nucleus, the formation of linear episomes from this foreign DNA could capture it for exploitation by the cell in the same way the transforming DNA was captured here. Even if there was not chromosomal integration early in this process, extrachromosomal maintenance could extend the window for integration to occur. One practical implication is the possibility for this system to be exploited by molecular biologists to improve the efficiency and flexibility of genetic manipulation in this organism, which can help accelerate future exploration into all aspects of its biology.

The findings of this study can also nucleate a broader evolutionary discussion on the ubiquity of these genetic mechanisms and their role in genome evolution across eukaryotes. The similarity of the processes observed here to those characterized in model organisms from other major eukaryote groups could reflect a widespread or even universal mechanism for sampling extrachromosomal DNA in the nucleus and potentially retain it more permanently. It could be hypothesized that such a system would be useful

for retaining endogenous DNA that has somehow become freed from its chromosomal home and would otherwise no longer be replicated, but it could also serve as a major mechanism facilitating eukaryote lateral gene transfer across the diversity of eukaryotes in a similar way to how it is proposed above in *Acanthamoeba*. Any previously underappreciated genetic capabilities such as this could go a long way toward explaining how lateral gene transfer increasingly seems like an important and ubiquitous driving force in eukaryote evolution, diversification, and ecological success[99,217].

This study would benefit from targeted investigation moving forward to enrich and verify its conclusions, but it has served to reveal an interesting molecular biological mechanism that has a range of implications. It advances our understanding of this complicated but important organism and provides an opportunity to exploit the new findings to improve our genetic tools. In turn, the possibility that this phenomenon of linear episome formation is much more widespread than we previously thought could indicate this is an important mechanism in the flow of genes into and throughout eukaryotes.

CHAPTER 4 COMPLEX AND VARIABLE PLOIDY IN *ACANTHAMOEBA CASTELLANII*

4.1 Introduction

For many reasons, *Acanthamoeba* is attractive as an experimental model organism. Practically, it is easy to grow axenically to high density in the laboratory and has a large enough cell size to facilitate microscopic observation. Biologically, several features of its cell biology are thought to be reasonable proxies for those same features in eukaryote cells more broadly. In contrast, genome organization in *Acanthamoeba* is complex and poorly understood. This somewhat limits its potential as a cell biology model; functional studies that require genetic manipulation may be difficult to implement, and its genomic complexity may also obscure processes that would be informative for our understanding of eukaryote genome biology in general.

Researchers have made inroads into understanding *Acanthamoeba* nuclear biology, but these studies did not definitively characterize any of its elements. Studies published over time have placed boundaries on the realm of possibilities. In 1972, Pussard attempted to use cytological methods to count chromosomes, but found they were too small and tightly clustered to do this accurately[218]. Byers measured nuclear DNA content, proposing a now-widely-cited estimate of 25*n* for *Acanthamoeba*'s ploidy in 1986[10]. Two studies employed pulsed-field gel electrophoresis (PFGE) to explore *Acanthamoeba* karyotypes. The first of these used the Neff strain and resolved 17 bands on the gel[106]. The authors inferred that this was likely a lower bound on the number of chromosomes. The intensity of some bands suggested they may represent more than one co-migrating chromosome, and Southern blot experiments detected some known nuclear-

encoded genes in the DNA that had not migrated out of the well into the gel, suggesting that the largest chromosomes were not resolved. The second PFGE study compared electrophoretic karyotypes of 28 *Acanthamoeba* strains across 12 species, and demonstrated that the karyotypes across these strains were quite heterogeneous, even within species[219]. Based on variable intensity of the bands within many of the strains, the authors also proposed the possibility of aneuploidy in *Acanthamoeba* genomes (i.e., that not all chromosomes are present in equal copy number).

The measure of DNA content and the electrophoretic karyotype experiments each bring important evidence to bear in solving the puzzle of *Acanthamoeba*'s genome biology, but they do not directly enrich or contextualize each other. Measuring nuclear DNA allows quantification on a per-cell basis but is blind to its identity. Conversely, PFGE provides structural insight into the genome and identifies at least part of the complement of chromosomes, but quantitative information is relatively limited and can only be discerned from the relative intensities of the bands. The obstacles to Pussard's[218] cytogenetics efforts are unfortunate, because those results would have bridged the gap between the capabilities of the experiments that followed. A cytogenetics approach would provide per-cell quantitative information of the total number of chromosomes, and potentially also provide information on the karyotype, if the chromosomes could be sufficiently differentiated. This would also provide valuable information on the copy number of each chromosome to address the question of aneuploidy. As it stands, present data in hand indicate that the genome biology of *Acanthamoeba* is atypical, but does not allow any concrete conclusions to be drawn.

With the use of Hi-C technology that allowed near-complete resolution of all

Acanthamoeba chromosomes into scaffolds as a part of the genome project described in Chapter 2 of this thesis, an opportunity was afforded to revisit the question of ploidy in this organism. These assembled chromosomes serve as a predicted karyotype to a level of detail that eluded researchers in PFGE experiments, while the amplification-free library preparation method allows relative quantitation of the chromosomes; the absolute per-cell counts cannot be determined from these data, but ratios of sequence depth of coverage can hint at relative abundances of different chromosomes in an aneuploid situation. With high throughput sequencing, analysis of allele frequencies across chromosomes has also become possible, bringing additional, distinct lines of evidence to the question. Long read data provides two independent types of allele frequency data, the more common single nucleotide polymorphism (SNP) allele frequencies and structural variant allele frequencies. Short read data provides an additional set of SNP frequencies for cross-referencing. With the rich amount of sequence data from the wild-type Neff and C3 genomes sequenced in Chapter 2 of this thesis and the transformed clones sequenced for Chapter 3 of this thesis, I extracted and analyzed these three new data types with the goal of shedding light on the genome organization and ploidy in *Acanthamoeba*.

4.2 Methods

4.2.1 Plotting single nucleotide polymorphism allele frequencies from nanopore reads

All nanopore reads analyzed in this chapter were those generated for the studies presented in Chapter 2 and Chapter 3 of this thesis. To briefly recap, there were nanopore read sets of the wild-type Neff strain from the Archibald lab and the Spatial Regulation of Genomes and Biology of Intracellular Bacteria teams at the Institut Pasteur, as well as a

read set from the Institut Pasteur of the C3 strain. I also sequenced one clonal isolate from the Neff culture at Dalhousie that had been transformed with circular pGAPDH-EGFP, and three clonal isolates that had been transformed with linearized pGAPDH-EGFP.

First, wild-type Neff nanopore reads were mapped against the entire wild-type Neff assembly with minimap2[133] v2.24 and I plotted the allele frequencies of single nucleotide polymorphisms using ploidyNGS[220] v3.0.0. Then, the genome assembly was divided into its individual scaffolds, and the wild-type nanopore reads I generated in the Archibald lab were mapped against each of the 35 largest scaffolds independently. PloidyNGS plots were then generated for each of the scaffolds. Other than dividing up the scaffolds, the mapping and plotting were performed as before.

To explore the stability of the observed patterns over laboratory timescales (across wild-type Neff and the clonal isolates) and evolutionary timescales (between Neff and C3), nanopore reads from the Pasteur Institut Neff culture, ‘Clone 1’, ‘Clone LT6’, ‘Clone LT8’, and ‘Clone LT9’ were also mapped against the separated wild-type Neff scaffolds and ploidyNGS plots were generated for each scaffold using the reads from each of these data sets. The reads from the Pasteur Institut C3 culture were naturally mapped against the C3 wild-type genome assembly instead, but the 35 largest scaffolds were still analyzed independently.

4.2.2 Plotting single nucleotide polymorphism allele frequencies from Illumina reads

Illumina reads had been generated for a subset of the isolates mentioned above for the purposes of polishing nanopore data. There were Illumina reads from the Archibald lab and the Institut Pasteur for the wild-type Neff culture, and from the Institut Pasteur for

the C3 culture. Among the transformed cultures, ‘Clone LT6’ and ‘Clone LT9’ were sequenced with Illumina. For these isolates, Illumina reads were mapped against the reference genome using HISAT2[197] v2.2.1 and SNP allele frequencies were plotted with ploidyNGS[220] v3.0.0. As before, all read sets were mapped against the wild-type Neff assembly except for the C3 reads which were mapped against the C3 assembly.

4.2.3 Structural variant calling and allele frequency plotting from nanopore reads

The structural variant caller Sniffles2[221] was used to detect structural variants in the nanopore read mapping data described above. Sniffles was configured to detect insertions or deletions of at least 50 bp. Sniffles plot was used to visually summarize variant calling data, including plotting allele frequencies of structural variants.

4.3 Results

Upon finalizing the *Acanthamoeba castellanii* strain Neff assembly (Chapter 2), I undertook the relatively routine step of using ploidyNGS[220] v3.0.0 to estimate its ploidy. This program plots a histogram of the absolute counts for each possible allele frequency (in steps of 0.01) detected in sequencing read sets mapped to the reference genome, and the overall shape of the histogram can be inspected to make inferences about ploidy. However, with a very high level of background noise across the possible range of allele frequencies and no clear peaks, it was apparent that this approach would not be adequate for this genome. Suspicious of potential aneuploidy, I repeated the analysis by dividing the assembly into each individual scaffold and performing the same mapping and plotting analyses for each scaffold individually. This approach produced meaningful signal on the

individual scaffold level, which notably varies between scaffolds (the largest of which approximately correspond to chromosomes) of the same assembly (Fig. 4.1).

From these histograms, ploidy can be inferred based on where along the x-axis peaks form. For example, if the plot has peaks near allele frequencies of 0 and 1 with little representation of the allele frequencies in between, this is indicative of haploidy. If, however, peaks are observed near 0 and 1 along with a peak around 0.5, this suggests diploidy (the 0.5 peak corresponds to the presence of two — and only two — different alleles in the individual reads). Peaks near allele frequencies of 0, 0.33. 0.67, and 1 indicate possible triploidy. This reasoning can be extended to higher ploidy levels, but at signals higher than pentaploid, it becomes difficult to visually resolve the peaks on the plot.

In this study, it is important to recognize the difference between what is being called “ploidy signal” and the actual ploidy. Ploidy signal here refers to the lowest ploidy level that would be consistent with the pattern of peaks on the plot, essentially the least common denominator. For example, peaks at allele frequencies 0, 0.5, and 1 will be called diploid signal here, but it is formally possible that 10 copies of a chromosome could give this signal if there were two identical sets of five. Since I do not have absolute quantitation of any of these chromosomes, I defaulted to identifying the lowest ploidy level consistent with the plots as the ‘ploidy signal’, while recognizing the possibility that the true copy number is a multiple of the one indicated. This is also why, while Byers[10] inferred a ploidy of 25*n* from nuclear DNA content, my inferred ploidy signals are lower.

In Figure 4.1, ploidy signal can be observed for the 12 largest scaffolds of the wild-type Neff assembly (ranging from 1.47 to 2.54 Mbp in size). Real data do not

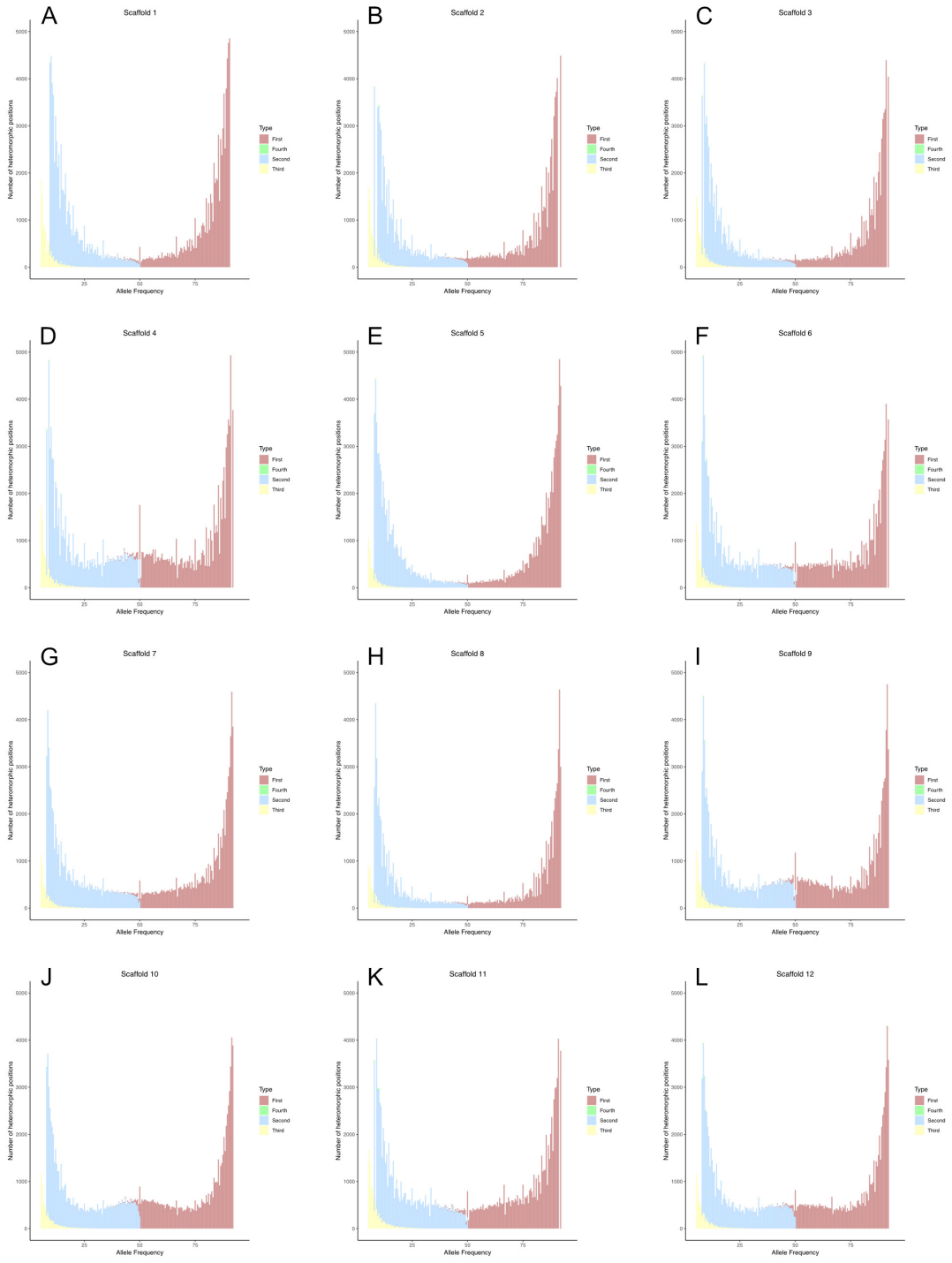
always cleanly conform to the above-mentioned patterns expected for different ploidy levels, but most of these plots can be interpreted. The plots for scaffolds 1, 2, and 3 (Fig. 4.1A-C) appear to represent haploid signal, with peaks near the extremes of the plot but not between them. The peaks for Scaffold 1 (Fig 4.1A) do tail out further toward the middle of the allele frequency range, but no distinct peaks can be observed along those curves that would change the interpretation. For Scaffolds 5, 7, and 8 (Fig. 4.1E, G, H), a haploid signal is also the best inference; again, the peaks at the extremes of some of the plots do taper off a bit slowly but there are no distinct 'shoulders' along their curve that could be indicative of another set of allele frequencies.

The plots for Scaffolds 4, 9, 10, and 12 (Fig. 4.1D, I, J, L) all closely resemble what one would expect for a diploid chromosome, with peaks at the extremes and a smooth peak centering around an allele frequency of 0.5. It is a bit more difficult to interpret the plots for Scaffold 6 and Scaffold 11 (Fig. 4.1F and K). There appears to be allele frequencies represented somewhere between the two peaks at the extremes, but these additional peaks are not well resolved in the plots. In the plot of Scaffold 11, it appears that as the peaks from the extremes of the plot tail down toward the centre, there may be a small inflection in the corresponding curve around allele frequencies 0.33 and 0.67 which would represent a triploid signal. The plot of Scaffold 6 is the least clear; there could be additional peaks near 0.33 and 0.67 once again, or potentially 0.4 and 0.6 which may correspond to pentaploidy.

Regardless of the exact interpretations at this point, these examples illustrate how these plots can be used to infer ploidy signal in general and across different scaffolds within a genome, and why it appeared that aneuploidy was a possibility in

Acanthamoeba. It also demonstrates that with real data and ploidy levels higher than diploid, this is still a useful and informative approach, but it is difficult to make precise inferences.

Figure 4.1 PloidyNGS SNP frequency plots from the 12 longest scaffolds of wild-type *Acanthamoeba castellanii* strain Neff (Archibald lab line). Plots were generated using nanopore reads mapped to the reference genome with minimap2. The ploidy signal is clearly varied across these 12 scaffolds from the same strain. A. Scaffold 1. B. Scaffold 2. C. Scaffold 3. D. Scaffold 4. E. Scaffold 5. F. Scaffold 6. G. Scaffold 7. H. Scaffold 8. I. Scaffold 9. J. Scaffold 10. K. Scaffold 11. L. Scaffold 12. Colours in the histograms correspond to the ranking of a given allele at its respective site based on frequency. Red represents the 'major' allele with the highest frequency, blue represents the 'minor' or second allele, with the second highest frequency, and small amounts of green and yellow represent the third and fourth alleles.


A
Scaffold 1
B
Scaffold 2
C
Scaffold 3
D
Scaffold 4
E
Scaffold 5
F
Scaffold 6
G
Scaffold 7
H
Scaffold 8
I
Scaffold 9
J
Scaffold 10
K
Scaffold 11
L
Scaffold 12
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third

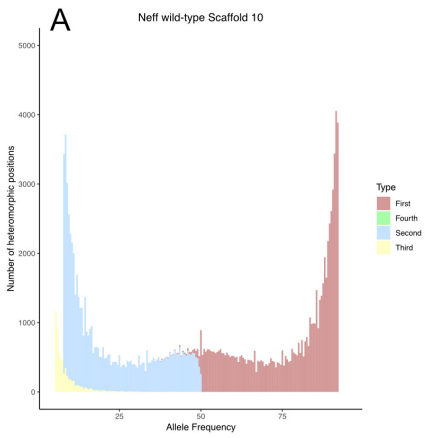
Having found support in the wild-type long read sequencing data for aneuploidy in this genome, several other independent read sets from different cultures were analyzed in the same way to ask several related evolutionary questions. These additional cultures included: another wild-type culture of Neff, maintained and sequenced separately by collaborators at the Pasteur Institute, a wild-type culture of *A. castellanii* C3, also maintained and sequenced at the Pasteur Institute, and four monoclonal isolates of Neff derived from my wild-type culture, transformed by the same plasmid, and then individually bottlenecked into clonal cultures. Three of these were transformed with linearized pGAPDH-EGFP in the same experiment, while one was transformed with a circular form of the same plasmid in a separate experiment. Interestingly, variation appeared in the profile of inferred ploidy across all chromosome-level scaffolds from these different datasets generated from the different cultures. The transformed clonal isolates often varied in ploidy signal among one another, as well as from the wild-type culture from which they were derived. This is demonstrated using Scaffold 10 as an example in Figure 4.2.

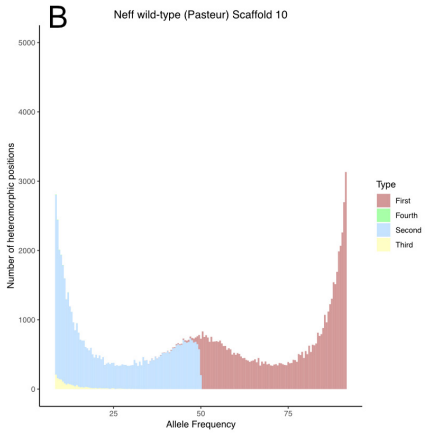
In this particular example, the two different Neff cultures (Fig. 4.2A and B) have the same diploid signal for this scaffold, while in C3 (Fig. 4.2C) the signal appears more likely to be triploid. The four clonal isolates, derived from wild-type Neff, show additional variation for this scaffold. Tetraploid signal is indicated by peaks at allele frequencies 0.25 and 0.75 in Clones LT8 and LT9 (Fig. 4.2F and G), while triploid signal appears in Clone LT6 (Fig. 4.2E). At first glance, the signal in ‘Clone 1’ (Fig. 4.2D) appears similar to triploid, but closer inspection reveals that the peaks are closer to allele frequencies of 0.4 and 0.6, corresponding to a pentaploid signal. While there are not

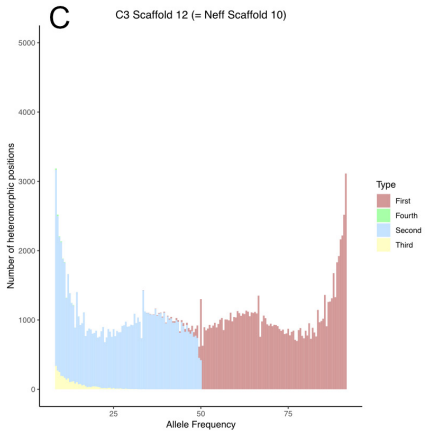
always so many different signals across the isolates for each scaffold, this example is a clear illustration of the type of variability in ploidy signal that is seen on many of the more variable scaffolds, especially in and among the clonal isolates when compared to the wild-type cultures.

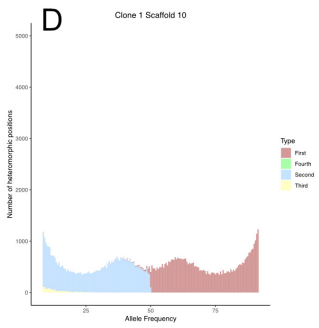
The two wild-type Neff datasets are also different in inferred ploidy signal for a number of chromosomes, despite both having been acquired first-hand from ATCC under the same strain ID by the respective laboratories at Dalhousie and the Pasteur Institute prior to sequencing (see Table 4.1 below). Neither were bottlenecked into monoclonal isolates prior to the genome project. Finally, variation in ploidy signal between the two wild-type Neff cultures and the C3 strain were observed, and the C3 strain was more different from either Neff culture than they were from each other. While evolutionary time and distance does appear to correlate with the degree of variation in effective ploidy signal, it is perhaps not a very linear correlation.

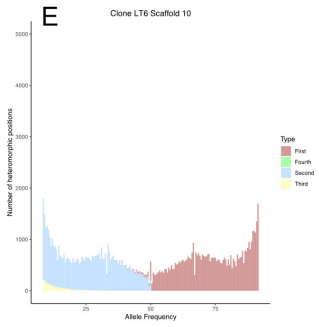
Figure 4.2 PloidyNGS SNP frequency plots for the same scaffold across seven *Acanthamoeba castellanii* isolates. Plots were generated from long reads mapped to the Neff reference genome with minimap2. A. Scaffold 10 with reads mapped from the wild-type culture of the Neff strain in the Archibald lab. B. Scaffold 10 with reads mapped from the wild-type culture of the Neff strain sequenced at the Institut Pasteur. C. Scaffold 12 from the C3 strain, homologous to Scaffold 10 in Neff, with reads mapped from C3. D. Scaffold 10 with reads mapped from Clone 1. E. Scaffold 10 with reads mapped from Clone LT6. F. Scaffold 10 with reads mapped from Clone LT8. G. Scaffold 10 with reads mapped from Clone LT9. Colours in the histograms correspond to the ranking of a given allele at its respective site based on frequency. Red represents the 'major' allele with the highest frequency, blue represents the 'minor' or second allele, with the second highest frequency, and small amounts of green and yellow represent the third and fourth alleles.

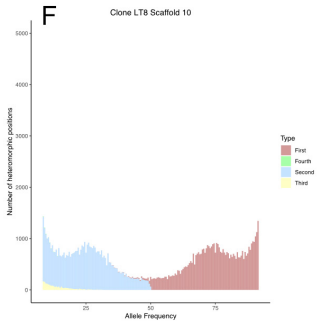

A
Neff wild-type Scaffold 10
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third

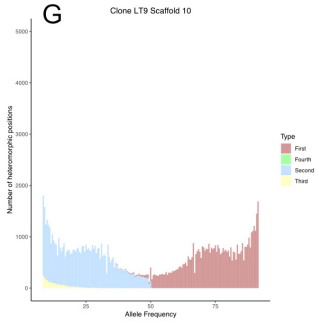

B
Neff wild-type (Pasteur) Scaffold 10
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third


C
C3 Scaffold 12 (= Neff Scaffold 10)
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third


D
Clone 1 Scaffold 10
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third


E
Clone LT6 Scaffold 10
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third


F
Clone LT8 Scaffold 10
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third


G
Clone LT9 Scaffold 10
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third

Two additional lines of evidence were introduced to support the allele frequency-based inferences from the long read data: ploidyNGS allele frequency plots from more accurate Illumina short read data, and structural variant allele frequency information from the long read data. When visualizing long read mapping to the reference genome, clear instances of structural variation could be observed in the mapped reads and coverage plots, where some fraction of the reads were missing a segment of DNA ranging from roughly a hundred to a few thousand base pairs, with a corresponding cleanly delineated drop in coverage accompanying this. Structural variants on the same chromosome were found to share the same allele frequency (where alleles are presence or absence of the indel) when they came from the same dataset, suggesting that at least some of the signal is biological. The long read structural variant calling program Sniffles 2 was used to make structural variant calls on all chromosomes independently, using each distinct long read dataset, such that each SNP-based ploidyNGS plot across the range of chromosomes and datasets had a corresponding structural variant-based Sniffles 2 histogram of allele frequencies.

To generate the ploidyNGS plots using Illumina data, an analogous approach was used as for the long reads, with appropriate adjustments, such as the choice of read mapping program. For a subset of the isolates, this meant the possibility of comparing three different effective ploidy estimates to arrive at what appeared to be the best inference for each chromosome in each of those isolates. Given the noise or uncertainty in some of these plots by themselves, having additional lines of evidence proved valuable. Figure 3 illustrates four different isolate-scaffold pairs where all three types of data were available, albeit with differing degrees of agreement between the three.

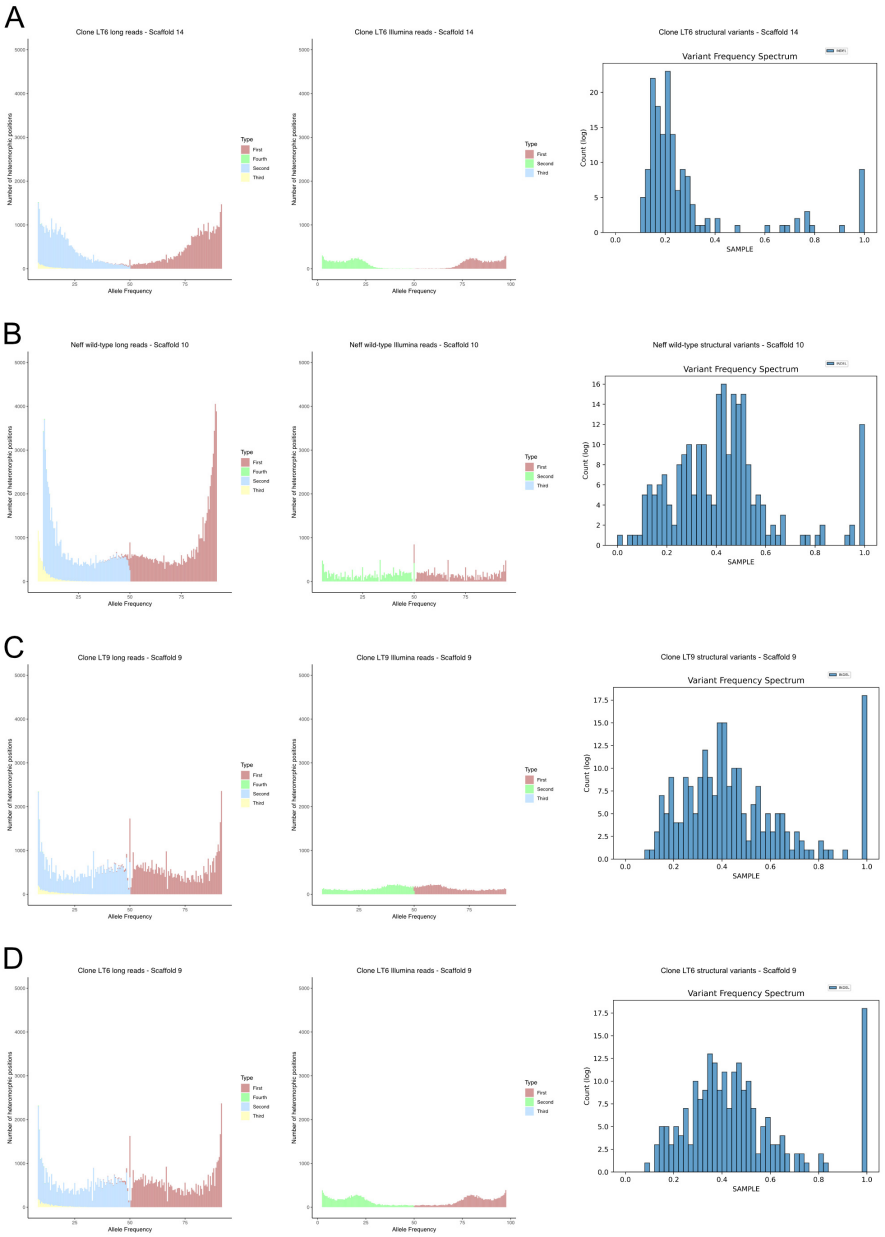
The first two examples, from Scaffold 14 in Clone LT6 (Fig. 4.3A) and Scaffold 10 in the wild-type Neff culture from the Archibald lab (Fig. 4.3B), demonstrate cases where inference was already relatively straightforward; here the three types of data collectively add confidence to my initial inferences. In the first case, peaks around allele frequencies of 0.2 and 0.8 can be clearly observed from the SNP frequency data (Fig. 4.3A), with the more accurate Illumina data providing much tidier plots. The structural variant allele frequency plot is consistent with the other two, with a peak around an allele frequency of 0.2 for the variant. These three plots all converge on an inferred pentaploid signal for Scaffold 14 in Clone LT6. In the next set of plots, Scaffold 10 from Neff wild-type from the Archibald lab (Fig. 4.3B), the SNP frequency plot for the long read data has a well-defined peak at allele frequency 0.5. The plot for the Illumina data is not quite as clean, but the signal is consistent with a peak at allele frequency 0.5. The structural variant plot also has its peak centered around allele frequency 0.5, meaning the three different types of plots agree on a diploid signal for Scaffold 10 in the wild-type Neff culture from the Archibald lab.

The third example, Clone LT9 Scaffold 9 (Fig. 4.3C), is one where one or more of the plots would have been misleading or difficult to interpret, but the combination of the three allows an inference to be made. The SNP frequency plot from the long-read data roughly resembles a diploid signal with a peak at an allele frequency of 0.5, but there is a bit of a dip directly in the centre of the peak. The tidier SNP frequency plot of the Illumina data reveals that there are actually peaks on either side of 0.5, probably at allele frequencies of 0.4 and 0.6, and the structural variant allele frequency plot has a peak centered around 0.4, which confirms the apparent signal from the Illumina SNP

frequency plot. Therefore, combining the information from all three plots reveals that the signal is probably pentaploid.

Finally, there are cases where the three plots clearly do not agree with respect to ploidy signal. In the fourth example, Clone LT6 Scaffold 9 (Fig 4.3D), the long read SNP frequency plot looks quite similar to the one from the previous example, where the peak is generally centered around 0.5 with a small dip in the middle. The structural variant allele frequency plot, also like the previous example, demonstrates that the signal probably actually peaks at allele frequencies of 0.4 and 0.6. However, it is quite clear from the Illumina SNP frequency plot that the peaks are at 0.2 and 0.8. It is worth noting that these both correspond to pentaploid signal but the ratios of one haplotype to the other are different between the two samples that gave rise to these plots, despite both samples coming from the same clonal culture within only a few months of one another.

Figure 4.3 SNP and structural variant allele frequency plots for four different scaffold-isolate combinations to illustrate how multiple lines of evidence influence interpretation. A. PloidyNGS plots using long and short reads from Clone LT6 mapped to Scaffold 14, and a structural variant allele frequency plot generated from the same mapped long reads. B. PloidyNGS plots using long and short reads from the Archibald lab wild-type Neff culture mapped to Scaffold 10, and a structural variant allele frequency plot generated from the same mapped long reads. C. PloidyNGS plots using long and short reads from Clone LT9 long reads mapped to Scaffold 9, and a structural variant allele frequency plot generated from the same mapped long reads. D. PloidyNGS plots using long and short reads from Clone LT6 mapped to Scaffold 9, and a structural variant allele frequency plot generated from the same mapped long reads. Colours in the histograms correspond to the ranking of a given allele at its respective site based on frequency. Red represents the 'major' allele with the highest frequency for both long and short reads. For the long read ploidyNGS plots, the secondary allele is blue, while it is green for the short read plots. The remaining colours represent the low-frequency third and fourth alleles.


A
Clone LT6 long reads - Scaffold 14
Clone LT6 Illumina reads - Scaffold 14
Clone LT6 structural variants - Scaffold 14
Variant Frequency Spectrum
Number of heteromorphic positions
Allele Frequency
Type
First
Fourth
Second
Third
Count (log)
SAMPLE
B
Neff wild-type long reads - Scaffold 10
Neff wild-type Illumina reads - Scaffold 10
Neff wild-type structural variants - Scaffold 10
C
Clone LT9 long reads - Scaffold 9
Clone LT9 Illumina reads - Scaffold 9
Clone LT9 structural variants - Scaffold 9
D
Clone LT6 long reads - Scaffold 9
Clone LT6 Illumina reads - Scaffold 9
Clone LT6 structural variants - Scaffold 9

Based on these three different lines of evidence, I compiled inferences of the ploidy signal for the 30 largest chromosomes across all of the isolates for the Neff strain, although some isolates had more information available than others. I stopped at the 30 largest because for the five smallest chromosome-scale scaffolds, there was too little data to populate the plots densely enough to make any inferences. For the C3 strain, I determined which regions were homologous to the top 30 Neff scaffolds based on the Circos plot data generated in Chapter 2. In many cases there was not an unambiguously correct answer to infer from the plots so the best guess was recorded and in cases where there were clear discrepancies between the long and short reads, I defaulted to the ploidy level indicated by the long reads; the goal was not necessarily to precisely define the ploidy of each chromosome across all isolates, but to get the best possible representation of the variability in ploidy signal across chromosomes and isolates. As an extension of this caveat, the table does appear to contain significant octoploid signal, but this is not necessarily biologically meaningful; there were several plots where the peaks were somewhere in the range from 0.1 to 0.2 and 0.8 to 0.9 respectively, and an octoploid signal seemed like the most reasonable inference, but it was difficult to be precise. The ploidy signals were inferred and recorded on the basis of all of the available information but not recorded for each plot individually because there were many cases where two or three plots were required to be able to make sense of the data. All inferred ploidy signals for the top 30 chromosomes of each isolate are presented in Table 4.1, which has been colour-coded in heat map form to more easily visualize the trends.

Table 4.1. Inferred ploidy signal for the top 30 scaffolds of six *Acanthamoeba castellanii* strain Neff isolates, and for the homologous scaffolds in *Acanthamoeba castellanii* strain C3. The colour scale ranges from lowest ploidy signal – white to highest ploidy signal – red.

Scaffold	C3	Neff Dal	Neff Pasteur	Clone 1	LT6	LT8	LT9
1	1	8	8	1	1	1	1
2	5	5	1	1	1	1	1
3	4	4	1	1	1	1	1
4	4	2	5	2	5	2	2
5	8	5	8	5	4	5	5
6	2	5	5	3	4	3	10
7	3	3	5	4	3	3	4
8	1	1	1	1	1	1	1
9	5	2	5	5	5	2	5
10	5	2	2	5	3	4	10
11	1	3	5	5	3	2	5
12	1	2	5	2	5	2	5
13	3	5	2	5	5	5	3
14	5	3	4	5	5	8	8
15	2	3	5	8	4	3	4
16	1	3	3	4	5	4	10
17	1	5	3	2	5	3	5
18	4	5	3	2	5	2	5
19	5	8	8	8	8	5	8
20	4	1	1	1	1	1	1
21	1	8	8	10	4	3	4
22	1	3	3	5	3	8	4
23	1	5	4	4	5	3	4
24	4	5	5	2	5	5	5
25	5	5	10	5	3	4	5
26	2	5	2	5	10	3	4
27	1	2	3	5	5	2	5
28	1	8	1	1	1	1	1
29	4	4	4	4	4	4	3
30	2	2	5	3	10	5	3

4.4 Discussion

The goal of this study was to use a wealth of new sequence data from multiple *Acanthamoeba* isolates to expand our understanding of its ploidy as much as possible. To that end, the allele frequency-based information used here was not able to provide absolute quantification of the copy number of any chromosomes, but it was able to show patterns of variation in effective ploidy signal across isolates, as well as between chromosomes within a given isolate. The information gathered helped me to assess the range of possible effective ploidy levels for individual *Acanthamoeba* chromosomes, as well as estimate roughly how dynamically these can vary.

To start, it is helpful to acknowledge the limitations of this approach, such that it is more clear which conclusions can and cannot be drawn from the data. The ability to resolve the peaks on the SNP frequency plots can be challenging, and as the signal moves toward a higher ploidy level, it becomes difficult to distinguish heterozygous alleles with a 1:N-1 ratio from homozygous alleles. As a result, precise inferences cannot be made at any ploidy level, but the ploidy levels may also be systematically underestimated because without extremely high resolution, estimating the signal from a plot will tend toward less granular interpretations of exactly where peaks appear. Another issue is that the three wild-type datasets are not from clonal isolates, so inferences may be less precise or entirely inaccurate due to mixed signals in the data (corresponding to signal from diverging clones within the same culture). This concern did manifest in the ploidyNGS plots, where peaks from clonal isolates were cleaner and easier to identify than those from non-clonal isolates. Therefore, this study can use the overall trends in the data to learn about the dynamic nature of *Acanthamoeba* ploidy, but not pinpoint ploidy levels

precisely.

The first axis of variation to consider here is the variation in ploidy signal across chromosomes within the same isolate. The isolates in this study had as many as seven different ploidy levels inferred across their 30 largest scaffolds, and the issues with resolving peaks may be masking even higher levels of ploidy. Inferences ranging from haploid to decaploid signal were made across all the isolates and chromosomes investigated. Clearly to have seven different ploidy levels inferred within a single isolate, there must be considerable variability in the copy number of the chromosomes. It is formally possible that all of the chromosomes are present in equal copy number but some are more heterogeneous than others. For example, if a hypothetical isolate was found to have only haploid, diploid, and tetraploid signal across its chromosomes, it could be the case that all chromosomes are present in four copies. However, to explain all the different ploidy signals inferred in this study through that logic, the copy number would have to be divisible by all of those inferred ploidy levels, making it unreasonably high.

If it is accepted that there genuinely is a substantial amount of variation in copy number across the chromosomes of any given *Acanthamoeba* cell, it must be asked whether there is a biological advantage to this. It is possible that gene dosage plays a role in the varied copy number of the chromosomes, but the patterns are so complex that it would seem absurd for this to be a key mechanism of regulating gene expression. Additionally, to jump ahead for a moment, the variation across isolates somewhat debunks this hypothesis because it seems fair to assume that the gene expression needs across the isolates would be relatively consistent; their ploidy signal clearly is not. One hypothesis for the benefits of polyploidy in asexual amoebae is that it is advantageous to

combat Muller's ratchet[222]. However, the variation in copy number across chromosomes also seems unlikely to play a role if this hypothesis were true. The simplest explanation for the varied copy number is that the differential amplification of different chromosomes is stochastic and has no overarching importance. However, in this case, there must be some short- to medium-term stability within an isolate after the variable copy numbers become established, because within clonal isolates the ploidy signal seems relatively stable based on the relative lack of conflict in the clonal ploidyNGS plots.

The second axis of variation detected in this study is the difference in effective ploidy signal of any given chromosome across isolates. Despite the short- to medium-term stability observed within a given isolate, there is clear variation in signal across isolates (Table 4.1). This hints at 'infrequent' events that somehow alter or reset chromosome copy number, with relative stability being maintained between such events. At first glance, it also appears that there is some correlation between evolutionary relatedness and ploidy signal, as careful inspection of Table 4.1 reveals that C3 is qualitatively more different from the other isolates than they are from one another. However, many of the differences involve either C3 or the different Neff isolates having a haploid signal while the other does not. Rather than reflecting a divergence in the actual maintenance of chromosome copy number, this could be explained by a difference in heterogeneity of the chromosomes where C3 is an outlier. The haploid signal in these cases could simply result from having very few polymorphisms on that particular chromosome to infer ploidy signal from, perhaps as a result of a selective sweep, while it is more heterogeneous in the opposite species. It is important to recognize that stochastic amplification and bottlenecking of chromosome copy numbers could easily result in the

loss of allelic diversity in these populations, which, in extreme cases, could result in highly homogeneous chromosome copies for which no ploidy signal but haploid can be distinguished.

One explanation for the short- to medium-term stability observed within isolates but variation across them could be that there is some loosely controlled endoreplication-type process that occurs somewhat infrequently, or, in terms relative to the generation time of *Acanthamoeba*, very infrequently. This process could amplify the different chromosomes to different levels, after which each one would be faithfully replicated and segregated during each round of mitosis. A hint from my data at this type of mechanism is the apparent existence of two underlying haplotypes for each chromosome. SNP frequency plots with an effective ploidy of tetraploid or higher, such as in Figure 2F, never appear with the full range of possible peaks. For example, if a chromosome was permanently pentaploid, it may be expected that there would be at least some alleles with each of the possible 1:4 and 2:3 ratios, but only one or the other seems to appear. An argument could be made that it is much more likely that a single point mutation occurs at a given locus than two of the same mutation in two copies, so a 1:4 ratio would not be particularly surprising, but there are a non-trivial number of 2:3-dominant plots as well. This phenomenon could be explained by having a genome defaulted to effective diploidy, but when endoreplication occurs, each haplotype is differentially amplified.

Another piece of evidence in favour of this hypothesis comes from some Sanger sequencing results collected by Dudley Chung, a postdoctoral fellow in the Archibald lab. During the course of developing CRISPR-Cas9 gene editing in *Acanthamoeba*, Dudley amplified and Sanger sequenced a few genomic loci from the genome, acquiring five

Sanger reads of each locus (Chung, pers. commun.; data not shown) Only two major haplotypes appeared within each set of five, with minor heterogeneity in some reads on top of this.

If there truly are only two underlying haplotypes in *Acanthamoeba*, the proposed endoreplication process to generate variation in chromosome copy number would also require a mechanism to return the genome back to its diploid state. This would be difficult to achieve with canonical meiosis due to the complexity of the aneuploidy in these cells. In microbial eukaryotes with characterized cycles of endoreplication and reduction, there are a few known mechanisms that could achieve this. Foraminifera employ a striking method known as genome segregation where cells with one highly polyploid nucleus undergo many consecutive rounds of mitosis to produce a great number of haploid gametes[49,50]. *Amoeba proteus* is also known to participate in cyclic polyploidy[223,224], but until recently, its approach to genome reduction was not well understood. It has now been shown that it uses a phenomenon called 'chromatin extrusion' to reset its ploidy just prior to mitosis by expelling the excess DNA from the nucleus to be degraded[225]. Another amoebozoan, *Entamoeba*, undergoes endomitosis to vary the number of nuclei in a cell, and thus also vary the genome copy number[11].

Of these examples, the mechanisms described in foraminifera and *Entamoeba* do not seem to fit the case of *Acanthamoeba*. If gametogenesis occurs in *Acanthamoeba*, it has not been observed or characterized, and while I can anecdotally claim to have observed *Acanthamoeba* cells with 2 to 4 nuclei on occasion, a mitotic mechanism for cyclic polyploidy would not explain the aneuploidy as well. The approach of *Amoeba proteus* seems compatible with the hypothetical endoreplication and reduction cycle I

propose in *Acanthamoeba,* but if the genome does reduce to a diploid rather than a haploid state, there must be some mechanism to retain one copy of each haplotype. To my knowledge there is currently no direct evidence to support a chromatin extrusion-type reduction method in *Acanthamoeba*. Another caveat to this is that *A. proteus* reduces its nuclear DNA content once per cell cycle, while the hypothetical reduction in *Acanthamoeba* must happen less frequently.

The parasitic kinetoplastid genus *Leishmania* is an informative example of how to think about *Acanthamoeba* aneuploidy moving forward. In this organism it is thought that aneuploidy is a constitutive feature, but the genome is fundamentally 2*n*, like what I propose for *Acanthamoeba*[226]. In general, aneuploidy in *Leishmania* is referred to as 'mosaic aneuploidy', referring to the fact that a given population can have a diversity of karyotypes represented at any given time. Some data suggests that upon founding of a population, there is an initial diversification of karyotypes followed by a shift toward a subset that becomes more dominant[227]. There appears to be a correlation between gene expression and chromosome copy number in *Leishmania*; under drug treatment or environmental stress, there appears to be an increase in copy number of the chromosomes bearing the genes that are upregulated in response to the change in conditions[228]. This suggests that perhaps this is a reasonable explanation, at least in part, for the observed aneuploidy in *Acanthamoeba*.

One study of aneuploidy in *Leishmania* used FISH to observe the karyotype dynamics of individual cells directly[226]. The authors observed that the aneuploidy state was variable among clones and strains, and that even after isolating a clone, mosaicism reappeared within the population, although the ploidy state generally resembled the

parent strain. If aneuploidy functions similarly in *Acanthamoeba*, this could suggest that even in the clonal isolates I analyzed, which showed a less convoluted signal than the wild-type strains that were not bottlenecked, there could be a small amount of mosaicism emerging that was not strong enough to be detected by the methods I employed.

The authors of the FISH study also observed a high rate of asymmetric chromosome allocations during mitosis, where, for example, three copies of a given chromosome were allocated to one daughter cell, and two copies to another. They hypothesized that this resulted from poor control over DNA replication, where one chromosome might be replicated an extra time, or one may fail to be replicated, resulting in one daughter cell receiving an extra copy or lacking a copy, respectively. Such a process would also be able to explain the apparent underlying 2*n* nature of the *Acanthamoeba* genome, because there would only be two major haplotypes to be segregated asymmetrically. If the replication defects were infrequent, it would also explain the apparent short-term stability of the ploidy state. In *Leishmania,* it was generally observed that chromosomes only reached trisomy, that is, three copies of a given chromosome, with very rare tetrasomy, while in *Acanthamoeba,* I infer higher copy numbers for many chromosomes. To achieve this result, presumably the asymmetric replication and segregation of chromosomes would have to persist over more generations and/or be biased toward synthesis of extra chromosome copies rather than failure to replicate existing ones.

Overall, the aneuploidy studies of *Leishmania* serve as a useful template for how to proceed with understanding *Acanthamoeba* aneuploidy, given their superficial similarity. Importantly, establishing reliable FISH protocols in *Acanthamoeba* would

provide a very powerful tool to make more sense of its ploidy state and its dynamic nature.

4.5 Conclusions

While there is still much to learn about *Acanthamoeba* aneuploidy, polyploidy, and whether the dynamic state of its ploidy is actively regulated, this study was able to capitalize on a wealth of sequence data to add significant depth to our previous understanding. These analyses were unable to shed more light on the question of just how polyploid *Acanthamoeba* may be, but they did reveal what appears to be quite prominent and dynamic aneuploidy, with appreciable variation in apparent ploidy state across the chromosomes in a given clone, as well as variation across clones. There is also some evidence for an underlying diploid state of the genome where each haplotype can be differentially amplified. The data analyzed here were not suitable for making any mechanistic inferences about these ploidy states. However, comparing these observations with those from *Leishmania* provides some hypotheses for mechanisms that generate aneuploidy in *Acanthamoeba,* and also provides examples of experiments that may be effective for better understanding this system.

CHAPTER 5 PHYLOGENOMIC SURVEY FOR PAST LATERAL GENE TRANSFER IN THE GENOMES OF *ACANTHAMOEBA CASTELLANII* STRAINS NEFF AND C3

5.1 Introduction

The potential for lateral gene transfer (LGT) to serve as an efficient and powerful driver of evolution and adaptation is undeniable. The existence of this phenomenon turns biological communities, especially microbial ones, into a vast and diverse standing pool of genes that fortunate recipients can sample and relatively easily add to their existing repertoire to gain additional capabilities. There are, of course, some prerequisites along the way; the genes of another organism must first be encountered, must enter the cell, become genomically integrated, and be expressed by the gene expression machinery of their new host. However, success at all these junctures provides the recipient with genes that have already evolved to perform some function, which, if sufficiently advantageous (and not toxic to the existing machinery), may become fixed in this new lineage. This is in contrast with the alternative where new functions can be acquired only after functional gene expression motifs and protein coding sequence have evolved *de novo* from previously non-functional DNA, or an existing gene has been duplicated and undergone subfunctionalization or neofunctionalization[229–231].

It is an uncontested fact in the present day that LGT is a major driving force in prokaryote evolution and diversification. This phenomenon was first reported by Tatum and Lederberg in 1947[232], who observed that nutritional deficiencies in *E. coli* mutants could be reversed by exposure to strains encoding the missing enzymes or pathways. However, it was the advent of genome sequencing decades later that really shone a light

on the scale at which LGT was shaping prokaryotic genomes, as illustrated when the whole genome sequence of the thermophilic bacterium *Thermotoga maritima* was published in 1999[233]. At the time of publication, it was only the 17th published whole prokaryotic genome, but phylogenetic inference was able to demonstrate species relationships among those sequenced prokaryotes, and similarity searching revealed a staggering 24% of the protein coding genes in *T. maritima* were more similar to archaeal genes than to its closest bacterial relatives. This finding was the prelude to many similar studies turning up lateral gene transfer across prokaryote diversity. These transfers were often predicted to spread capabilities such as antibiotic resistance, virulence, and expanded accessory metabolism[234]. As molecular biology moved further into the genomics era, the extent of prokaryote LGT revealed was such that some authors believe it may be a more important evolutionary force on some prokaryote genomes than gene duplication or point mutation[235,236].

While the greater size and complexity of eukaryote genomes has caused a lag, sequencing technology has now advanced to the point that genome sequencing of many microbial eukaryotes is now almost trivial. This wave of eukaryote genome sequences has revealed the ubiquity of LGT among eukaryotes as well. Although the scale is not quite so great that it is the dominant force in eukaryote evolution as some believe to be the case in prokaryotes, it is now clear that some of the most important and dramatic transitions in eukaryote diversification, such as the evolution of anaerobiosis, have been facilitated by LGT[237]. It is also becoming clear when researchers look at a finer scale, such as within a family or genus, that less extreme but still ecologically advantageous variation in microbial eukaryotes is facilitated by LGT. Accordingly, it is now

commonplace to include a phylogenomic survey for LGT in genome projects of microbial eukaryotes, and equally commonplace to find it. One such survey was performed with the first publication of the *Acanthamoeba castellanii* strain Neff genome by Clarke et al. in 2013[20], who inferred 2.9% of the predicted proteome to have been acquired by LGT, mostly from Bacteria with minor contributions from Archaea and viruses.

Just over a decade on from the work of Clarke et al., the present study was intended to update the known LGT footprint in *A. castellanii*, capitalizing on several factors that came into play over time. With continual discovery of new lineages, and an accelerating rate of sequencing such lineages as well as known but previously unsequenced ones, taxon sampling in public databases is much better, especially among eukaryotes and viruses. A lesser but still relevant improvement has been an increase in the sophistication and accuracy of phylogenetic models; this probably has less bearing on inferring the evolutionary history of single genes than it does on deep phylogenomic inference, but improved modelling is nonetheless positive[238–241]. One main motivation to repeat this type of investigation on *A. castellanii* is having the opportunity to compare two closely related strains and determine whether lateral gene transfer has shaped their evolution differently. To this end, phylogenetic trees were estimated for each protein in the predicted proteomes of the Neff and C3 strains and screened for putative LGT. Then, the functions and donors of each putative LGT gene were inferred to explore the possible biological significance of these acquisitions.

5.2 Methods

5.2.1 Homolog search, alignment, and phylogenetic inference

The set of query sequences for this investigation comprised the full predicted proteomes from *Acanthamoeba castellanii* strains Neff and C3, as generated from the genome project described in Chapter 2. DIAMOND[242] v2.1.8.162 BLASTp was used to identify homologous sequences for each protein in the Neff and C3 predicted proteomes. Searches were performed independently against the nr database, the MMETSP database[243], and the EukProt database[244]. DIAMOND BLASTp was run using an e-value threshold of 10^{-10}, --more-sensitive alignment parameters, 500 hits retrieved per query sequence, and a maximum of 3 HSPs per hit retrieved.

For each query protein, custom Python scripts were used to retrieve the full amino acid sequence of each hit, and the multi-FASTA files containing the hits from each database were merged. Any protein for which fewer than 9 homologs were retrieved was excluded from the analysis due to the anticipated difficulty of interpreting phylogenies with so few taxa.

To reduce the number of taxa in each tree below 200, CD-HIT[245] v4.7 was used to cluster sequences that were close in sequence identity, reducing redundancy. Prior to using CD-HIT, any multi-FASTA files already containing fewer than 200 sequences were set aside to proceed to the next steps. Then, CD-HIT was applied iteratively, reducing the clustering threshold each time. After each round, those files dropping below 200 sequences were set aside to move forward, while for those that did not, the clustering was performed again on their original set of retrieved homologs at a lower clustering threshold. Clustering started at an 80% identity threshold which was decreased by 5%

each round, but any alignments that had not dropped below 200 in number after clustering at 45% identity were carried forward without further reduction in redundancy. Word size was decreased alongside clustering threshold according to the CD-HIT user guide (word size 5 for 70-100% identity, 4 for 60-70%, 3 for 50-60%, and 2 for 50% and below).

The sequences were aligned using mafft[155] v7.520 with the --auto setting, and then BMGE[246] v1.12 was used for trimming and block removal using the BLOSUM30 substitution matrix, a maximum of 50% gaps permitted per site, a word size of 3, and an entropy score cut-off of 0.5.

All of the initial set of single protein trees were inferred using IQ-TREE[247] v2.2.2.7 under the LG+F+Γ evolutionary model with 1000 ultrafast bootstraps for branch support.

5.2.2 LGT detection

Candidate LGTs were detected using the method applied by Žársky et al.[248] to perform a similar survey in *Mastigamoeba balamuthi*. The program used for LGT detection is written to identify prokaryote-to-eukaryote transfers, but for this study it was also modified slightly to search for virus-to-eukaryote transfers. Trees with a Directionality Score and a Non-ancestral Score both greater than 0.5 were flagged as candidate LGT trees and investigated further.

After visual inspection of the resulting trees and the trimmed alignments from which they were generated, some of the sequences were observed to contain more than 70% gaps in the alignment while also falling in a location in the tree clearly incongruous

with known species phylogenies. This was addressed by removing all sequences in candidate LGT alignments that contained more than 70% gaps. Note that the preconceived bias of known species phylogeny led me to notice the presence of very gappy sequences, but it was not taken into consideration in removing them; the 70% threshold was applied uniformly across all sequences in all of the candidate LGT alignments.

To account for homologous proteins found within and between the two *Acanthamoeba* genomes from this study, a custom Python script referred to the orthogroups inferred by Broccoli in Chapter 2 to merge the pre-alignment multi-FASTA files of candidate LGT proteins that came from the same orthogroup. Other proteins from those Broccoli-inferred orthogroups that were not independently identified as LGT candidates were not included in the merged datasets. Since the newly merged files often now contained well in excess of 200 proteins, the same redundancy-reduction approach described above was repeated on this subset of the proteins, followed by also repeating the alignment and trimming steps described above, and removing sequences containing 70% gaps or more.

For this new set of alignments containing multiple *Acanthamoeba* homologs, as well as the existing alignments that did not need to be merged, tree inference was repeated as described above, except the more sophisticated LG+C20+Γ model of evolution was used this time.

5.2.3 Candidate tree inspection and curation

All of the resulting trees were manually inspected at this point. This step was used

to screen out apparent false positive LGT predictions and to infer putative donor lineages for the inferred LGTs. Two types of false positive trees were manually filtered out: those where there were too few non-eukaryotic sequences to make any inference about transfer, and those where the diversity of eukaryotes appeared to be fully represented (or nearly so) and the eukaryotic proteins appeared to have a shared evolutionary history. In practice, the former type of false positive was defined as trees where there were two or fewer prokaryote sequences and the rest were eukaryotic. The latter type was defined as trees where the *Acanthamoeba* sequence of interest was found in a clade of eukaryotes with at least one representative from each of Obazoa, Amoebozoa, Archaeplastida, Rhizaria, Stramenopila, Alveolata, Cryptista, Haptista, and Discoba, which was a proxy for full eukaryote diversity. Due to the unique circumstances surrounding viruses and their interactions with cellular life, trees containing viral sequences were exempt from the above criteria and all were assessed.

Due to the less constrained nature of viral evolution, virus-containing trees were screened differently. For the purposes of this study, only transfers from viruses to *Acanthamoeba* were sought. This presented as trees where the sequence of interest was found in a clade of eukaryotic sequences that was less diverse or numerous than the viral group from which it emerged. For example, a clade of nine paralogous proteins from *Acanthamoeba* branching within seven diverse representatives of Nucleocytoviricota would be inferred to be a virus to amoeba transfer, whereas a clade of six *Pandoravirus quercus* proteins branching within a group of single proteins from *Acanthamoeba, Luapelamoeba, Dracoamoeba,* and *Endostelium* would not be inferred as a transfer from viruses into amoebae. Cases where viral and eukaryotic representation was roughly

equivalent were rejected due to lack of information, as were cases where a tree had several *Acanthamoeba* homologs and homologs from a few different viruses, but the viral homologs did not branch together. Cases where *Acanthamoeba* branched within a large and diverse prokaryotic clade, and one or a few closely related viral sequences branched sister to or within the *Acanthamoeba* sequence(s) were assumed to be a transfer from prokaryotes to *Acanthamoeba* to viruses and not included.

5.2.4 Inferring LGT donors

To infer the donor of each transfer, I started at the candidate LGT gene in each tree and worked backward node-by-node until the following two criteria were satisfied: I stopped at a node with at least 90% ultrafast bootstrap support, and the subtree defined by that node had at least one non-eukaryote sequence. Then, I determined the least inclusive taxonomic group that included all of the non-eukaryotes in that subtree, which became the taxon I inferred to be the donor lineage.

5.2.5 Functional enrichment analysis

Gene ontology enrichment analysis was performed on the set of putative laterally transferred genes using topGO the R package topGO (https:///bioconductor.org/packages/topGO/ [accessed July 12, 2022]). The analysis was performed for each strain independently. Functional annotations for each predicted protein in both genomes were retrieved from the funannotate output generated as part of Chapter 2. Gene IDs and their associated GO terms were retrieved using a custom Python script. To expand the amount of information available for the analysis, InterPro and

PFAM IDs were retrieved as well. The program pfam2go (v1.1.2) was used to convert retrieved pfam IDs into corresponding GO terms. The pfam-GO conversions in this program come from a list manually constructed by the curators of InterPro[249] (https://current.geneontology.org/ontology/external2go/pfam2go. Similarly, the InterPro team has created an InterPro to GO conversion table (https://www.ebi.ac.uk/GOA/InterPro2GO) which I processed with a custom Python script to convert the retrieved InterPro IDs to GO terms. There were a number of cases where no direct conversion was established, so these InterPro and PFAM IDs were ignored. After the inclusion of GO terms converted from InterPro and PFAM where possible, the proteins that still did not have any annotated GO terms were excluded from the analysis.

A Fisher's exact test with the weight algorithm was implemented in topGO for the laterally transferred genes from each strain, for each of the three ontologies (biological process, cellular component, and molecular function). When building the GOdata objects for these three ontologies, nodeSize was set to 10 for both the biological process and molecular function ontologies and to five for the cellular component ontology to better suit the lower number of GO terms in this ontology.

5.2.6 Targeted InterProScan searches

Proteins with genus- or species-level donor identifications were searched against InterPro v100.0 using InterProScan[147] and results were visualized in the InterProScan web interface. All non-redundant function information for each protein was recorded, including the presence or absence of a signal peptide.

5.3 Results

5.3.1 Total LGT contribution into each strain

In total, 447 genes in Neff (2.9% of the predicted proteome) and 268 in C3 (1.6% of the predicted proteome) were inferred to have an LGT origin. Of the 447 predicted LGTs in Neff, 205 (46%) had an ortholog in C3 as predicted by Broccoli, while of the 268 predicted LGTs in C3, 181 (68%) had an ortholog in Neff as predicted by Broccoli. For each LGT, a donor lineage was then inferred to the most precise taxonomic level possible by working back through the nodes of the tree to find those that were sufficiently well supported and considering the taxonomic composition of the subtree defined by that node. In many trees, the internal nodes were not well supported, so while some trees looked at face value like a more precise donor could be inferred, the taxonomic representation within the final subtree was very broad. As a result, a large proportion of the putative laterally transferred genes could only be traced back to a kingdom- or domain-level taxon. There were many trees where the donor lineage appeared to be prokaryotic, but a distinction could not be made between Bacteria and Archaea, so while these domains together represent a paraphyletic group, it was still useful to identify some laterally transferred genes as simply prokaryotic in origin.

5.3.2 LGT genes are predominantly bacterial in origin

Donors were inferred from each of Bacteria, Archaea, and viruses in varying amounts and encompassing differing fractions of the total diversity of each (Fig. 5.1A). In this section, the combination of all predicted LGT genes in both strains are discussed together. There were 593 putative LGT genes assigned to Bacteria in total. Of the four

bacterial kingdoms, Pseudomonadati, Bacillati, Thermotogati, and Fusobacteriati, three were inferred as the originating lineage of at least one LGT in this study. A large number of transfers were traced back to Bacteria more broadly. There were also several transfers attributed to the relatively diverse phylum-level lineage *Candidatus* Patescibacter, commonly known as the candidate phyla radiation[250]. Where transfers could be attributed to a bacterial kingdom, there appeared to be a strong bias toward Pseudomonadati; Pseudomonadati was the inferred source of 188 genes, Bacillati was inferred to have donated 52, and only one was attributed to Thermotogati. A phylum-level overview of the LGTs attributed to bacterial taxa is presented in Figure 5.2.

Archaeal representation among the donor lineages was somewhat less complete than for Bacteria; only 19 of the candidate LGTs could be traced back to Archaea unambiguously, and only six originated from validly published archaeal phyla, two from Thermoproteota and four from Methanomicrobiota. The rest were attributed to the candidate phyla *Candidatus* Asgardarchaeota, *Candidatus* Heimdallarchaeota, *Candidatus* Lokiarchaeota, *Candidatus* Thorarchaeota, *Candidatus* Woesearchaeota, and *Candidatus* Pacearchaeota, or to Archaea more broadly (Fig. 5.1B).

All of the 20 inferred LGTs from viruses were attributed to the phylum Nucleocytoviricota, and most could be assigned to the genus level (Fig. 5.1C). Interestingly, none of the transfers were found to have come from the order Imitervirales, best known as the home of the family Mimiviridae. The majority of transfers were from the order Pandoravirales, including transfers more precisely attributed to Pandoravirus, Mollivirus, or Medusavirus. The remaining seven were inferred to have come from Pithovirus. Many of these phylogenies did have numerous homologs from Imitervirales,

but they were not found to be the closest relatives of the candidate LGT genes.

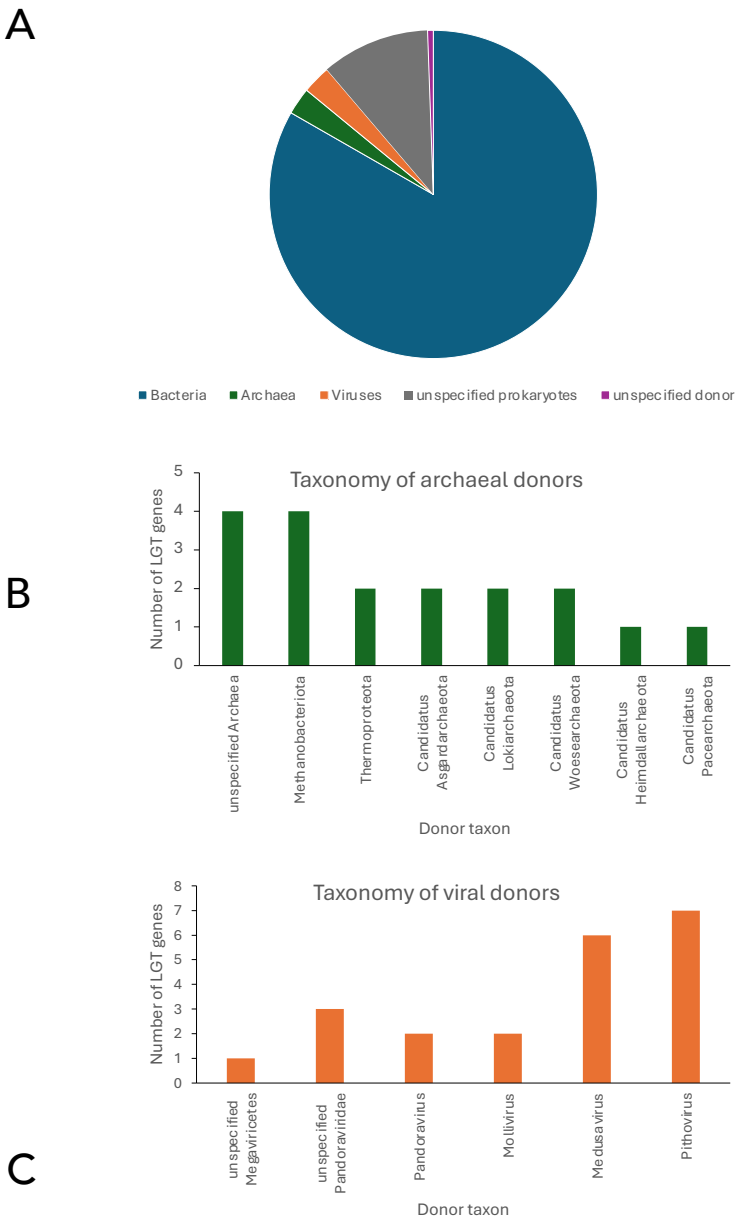


Figure 5.1 Taxonomic breakdown of inferred donors for LGT-derived genes in *Acanthamoeba castellanii* strains Neff and C3. Donors were inferred based on the least inclusive subtree defined by a node with at least 90% ultra-fast bootstrap support that contained the putative LGT sequence. A. Domain-level breakdown of the inferred donors. B. Number of inferred LGTs originating from each archaeal phylum (for which at least one donor was inferred). C. Number of inferred LGTs originating from each viral genus (for which at least one donor was inferred). For each of these representations, some ambiguous donor inferences at a higher taxonomic level are included. A contains LGTs for which a distinction between an archaeal and bacterial origin could not be made, and LGTs for which no donor identification could be made. B contains LGTs that were inferred to be archaeal in origin but could not be assigned to a phylum. C contains LGTs that could not be assigned to a taxon below class Megaviricetes, and LGTs that could be assigned to Pandoraviridae but not to a genus.

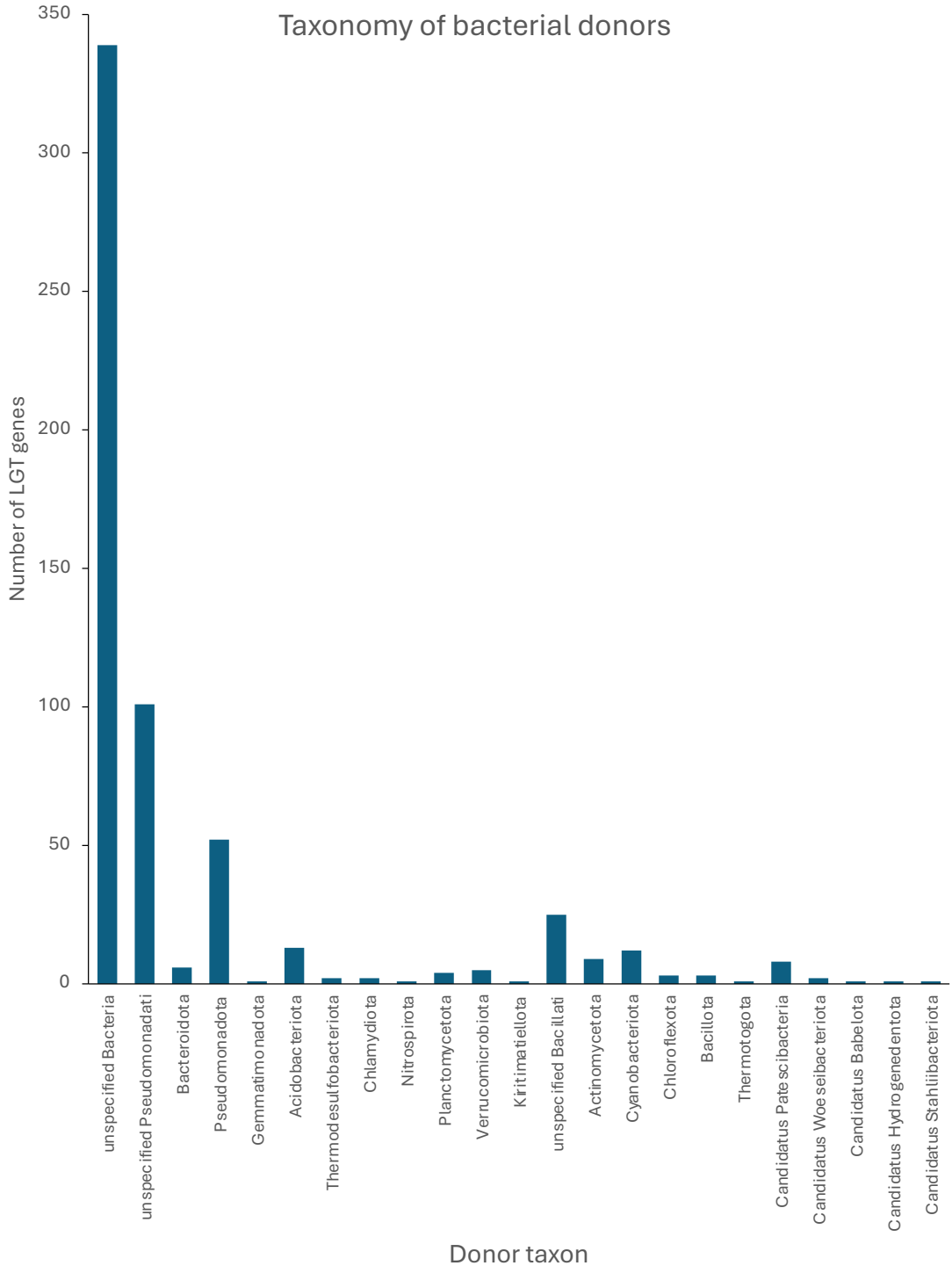


Figure 5.2 Number of LGT-derived genes in *Acanthamoeba castellanii* strains Neff and C3 assigned to each bacterial phylum. Donors were inferred based on the least inclusive subtree defined by a node with at least 90% ultra-fast bootstrap support that contained the putative LGT sequence. Some LGT-derived genes could be assigned to a kingdom (Pseudomonadati or Bacillati in this chart) but not a phylum, while some could not be assigned beyond the domain level.

5.3.3 Candidate LGT trees frequently suggest transfer to additional eukaryotes

Each of the phylogenies representing prokaryote-to-eukaryote transfers could be described by one of four major patterns with respect to which eukaryotes branch from within the prokaryotes and where. These can be summarized as follows: (i) *Acanthamoeba* is the only eukaryote in a tree of prokaryotic sequences (Fig. 5.3 for example), (ii) *Acanthamoeba* forms a clade with one or more other Amoebozoa in a tree of otherwise prokaryotic sequences (Fig. 5.4 for example), (iii) *Acanthamoeba* and any number of Amoebozoa form a clade with at least one non-Amoebozoan eukaryote in a tree of otherwise prokaryotic sequences (Fig. 5.5 for example), or (iv) *Acanthamoeba* forms a clade with or without other eukaryotes in a tree where at least one other eukaryote branches from a separate lineage of prokaryotes (Fig. 5.6 for example). The last pattern could include an unlimited number of independent eukaryotic sequences or clades branching from different positions within prokaryotes. The most common pattern was for multiple independent eukaryote groups to branch at various points within the prokaryotes, observed 56% of the time, followed by a single eukaryote clade including non-amoebozoan representatives appearing 23% of the time, *Acanthamoeba* alone emerging from prokaryotes 13% of the time, and a clade of *Acanthamoeba* with additional amoebozoans arising from within prokaryotes 7% of the time.

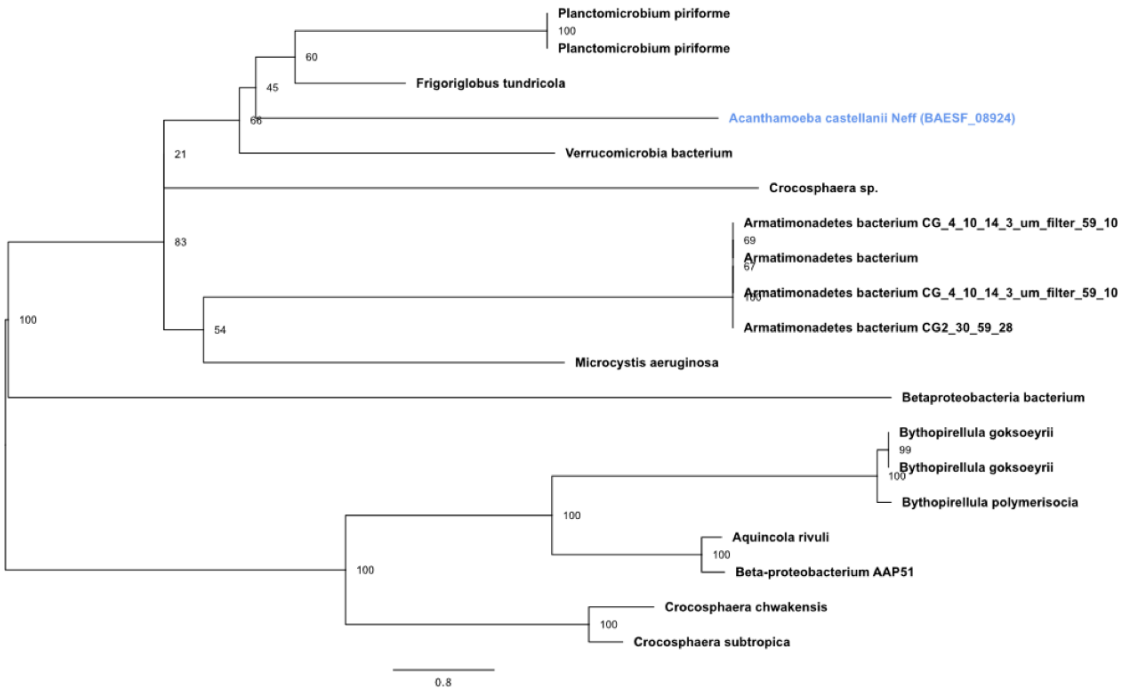


Figure 5.3 A phylogeny with a single *Acanthamoeba castellanii* LGT candidate branching within otherwise prokaryotic sequences. This candidate LGT protein is from *A. castellanii* strain Neff and has the locus tag BAESF_08924. This maximum likelihood tree was estimated under the LG+C20+Γ model of evolution using IQ-TREE. Ultrafast bootstrap values are displayed at the nodes. Label colours indicate taxonomic affiliation: black – prokaryotes, blue – Amoebozoa.

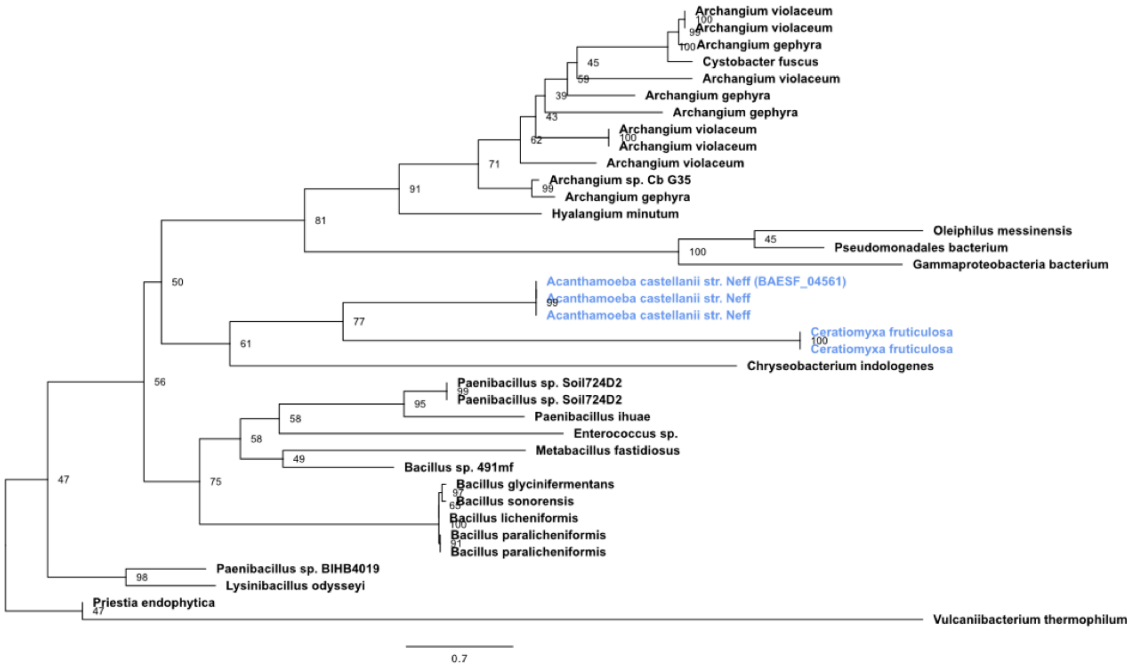


Figure 5.4 A phylogeny with an *Acanthamoeba castellanii* LGT candidate in an amoebozoan clade among otherwise prokaryotic sequences. This candidate LGT protein is from *A. castellanii* strain Neff and has the locus tag BAESF_04561. This maximum likelihood tree was estimated under the LG+C20+Γ model of evolution using IQ-TREE. Ultrafast bootstrap values are displayed at the nodes. Label colours indicate taxonomic affiliation: black – prokaryotes, blue – Amoebozoa.

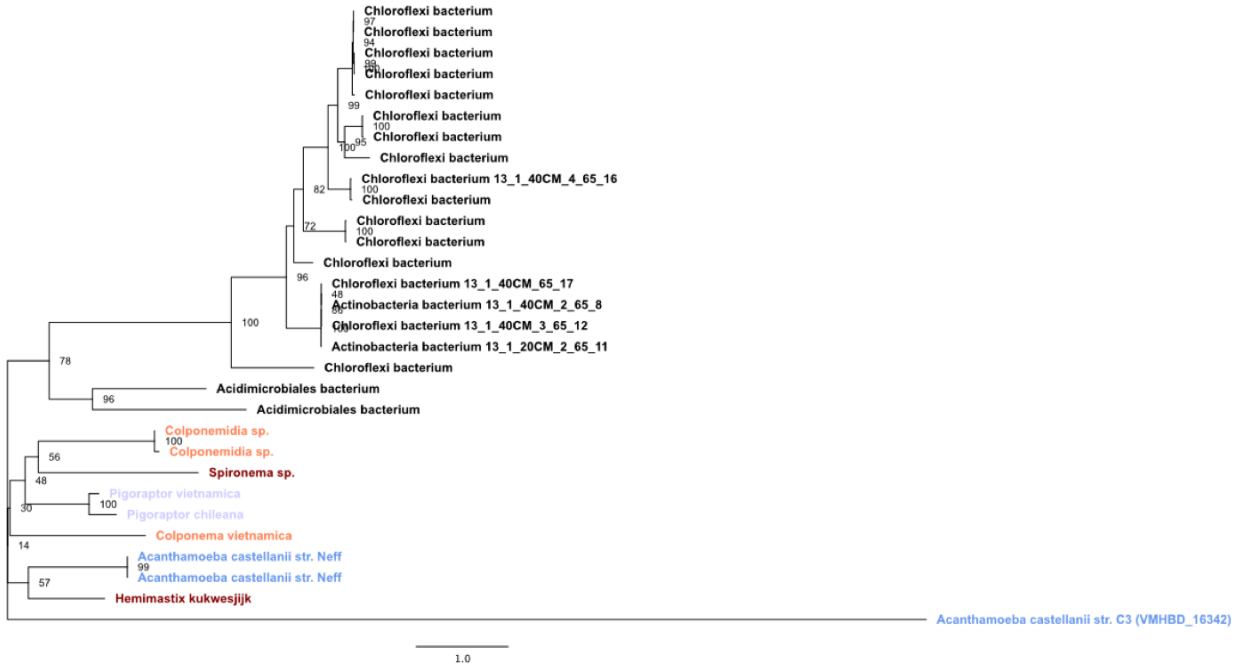


Figure 5.5 A phylogeny with an *Acanthamoeba castellanii* LGT candidate in a clade of eukaryotes among otherwise prokaryotic sequences. This candidate LGT protein is from *A. castellanii* strain C3 and has the locus tag VMHBD_16342. This maximum likelihood tree was estimated under the LG+C20+Γ model of evolution using IQ-TREE. Ultrafast bootstrap values are displayed at the nodes. Label colours indicate taxonomic affiliation: black – prokaryotes, blue – Amoebozoa, orange – Alveolata, lavender – Obazoa, dark red – Hemimastigophora.

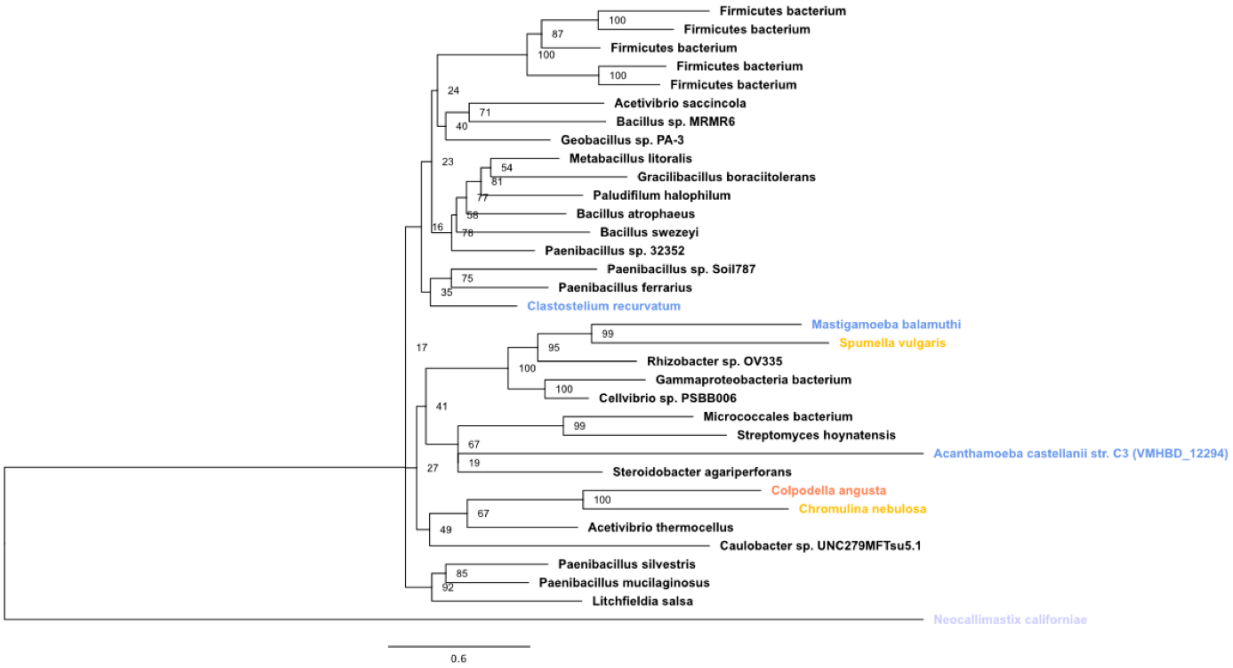


Figure 5.6 A phylogeny where an *Acanthamoeba castellanii* LGT candidate and other eukaryote sequences branch separately within prokaryotic sequences. This candidate LGT protein is from *A. castellanii* strain C3 and has the locus tag VMHBD_12294. This maximum likelihood tree was estimated under the LG+C20+Γ model of evolution using IQ-TREE. Ultrafast bootstrap values are displayed at the nodes. Label colours indicate taxonomic affiliation: black – prokaryotes, blue – Amoebozoa, orange – Alveolata, yellow – Stramenopila, lavender – Obazoa.

5.3.4 Functional enrichment analysis reveals expansion of metabolic capabilities in both strains

To determine whether these *Acanthamoeba* strains appear to have more LGTs relating to particular functions, I performed a gene ontology enrichment analysis using topGO for each of the two strains independently. In C3, the predicted LGT genes showed statistically significant enrichment of 11 terms from the 'molecular function' ontology (Fig 5.7), seven terms from the 'biological process' ontology (Fig 5.8), and two terms from the 'cellular component' ontology (Fig 5.9). In Neff, the predicted LGT genes had 13 enriched terms from the 'molecular function' ontology (Fig 5.10), 19 enriched terms

from the ‘biological process’ ontology (Fig 5.11), and two enriched terms from the ‘cellular component’ ontology (Fig 5.12). All enriched terms for C3 are tabulated in Table 5.1 while those for Neff are presented in Table 5.2.

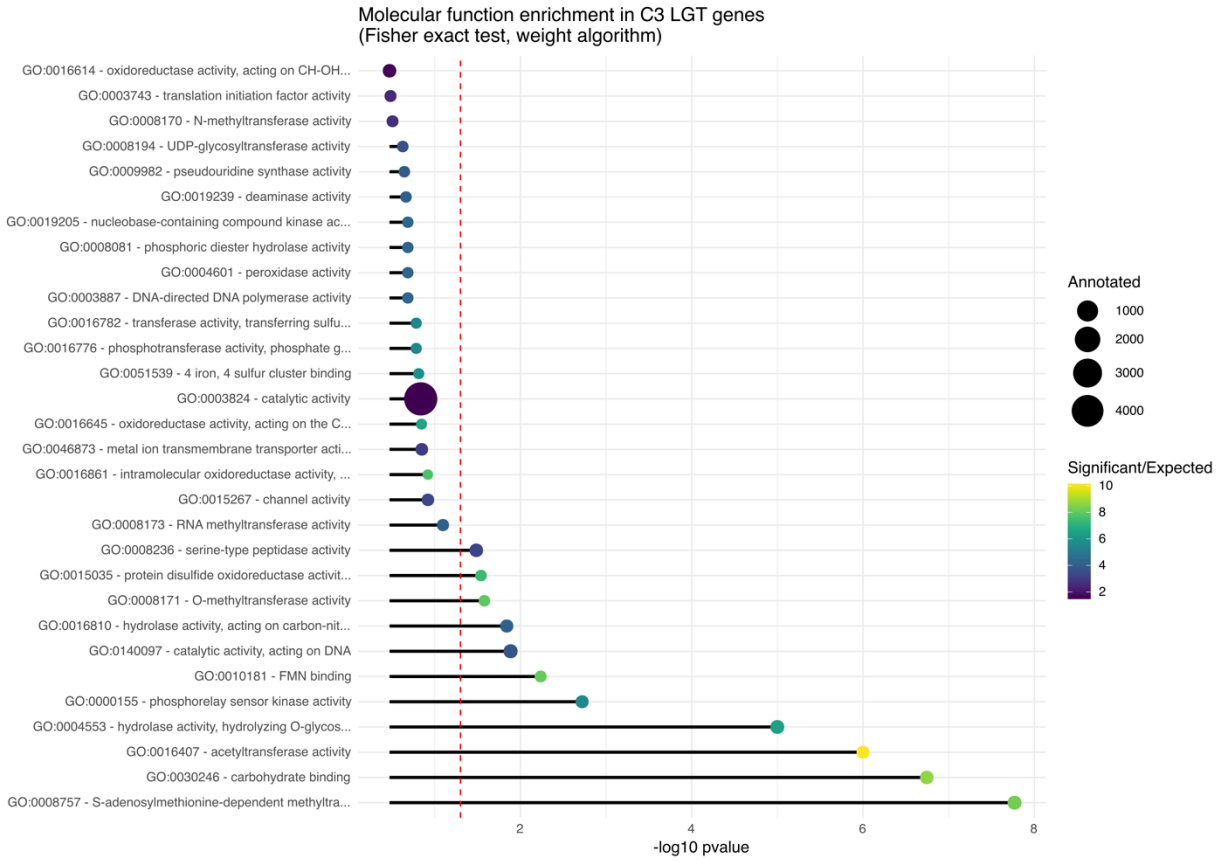


Figure 5.7 Most significant molecular function GO term enrichments among LGT-derived genes in *Acanthamoeba castellanii* strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

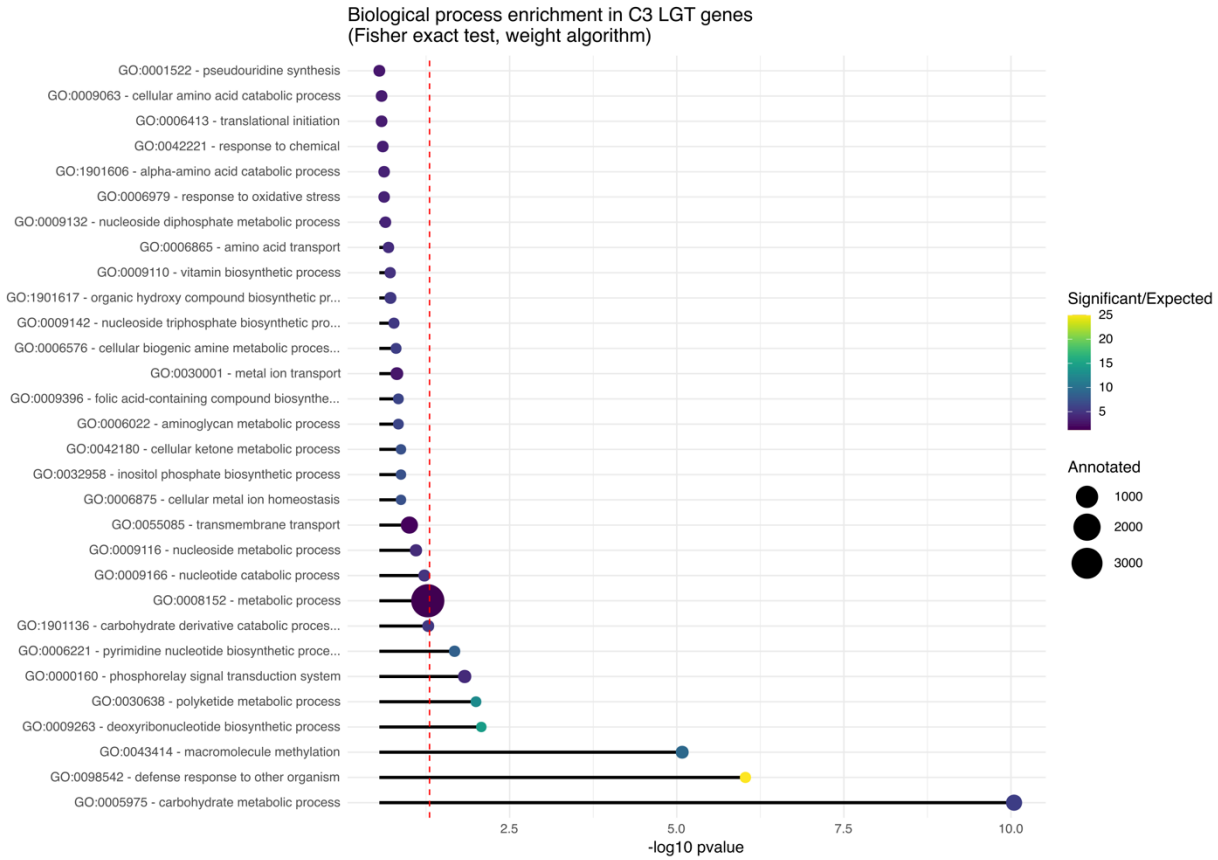


Figure 5.8 Most significant biological process GO term enrichments among LGT-derived genes in *Acanthamoeba castellanii* strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

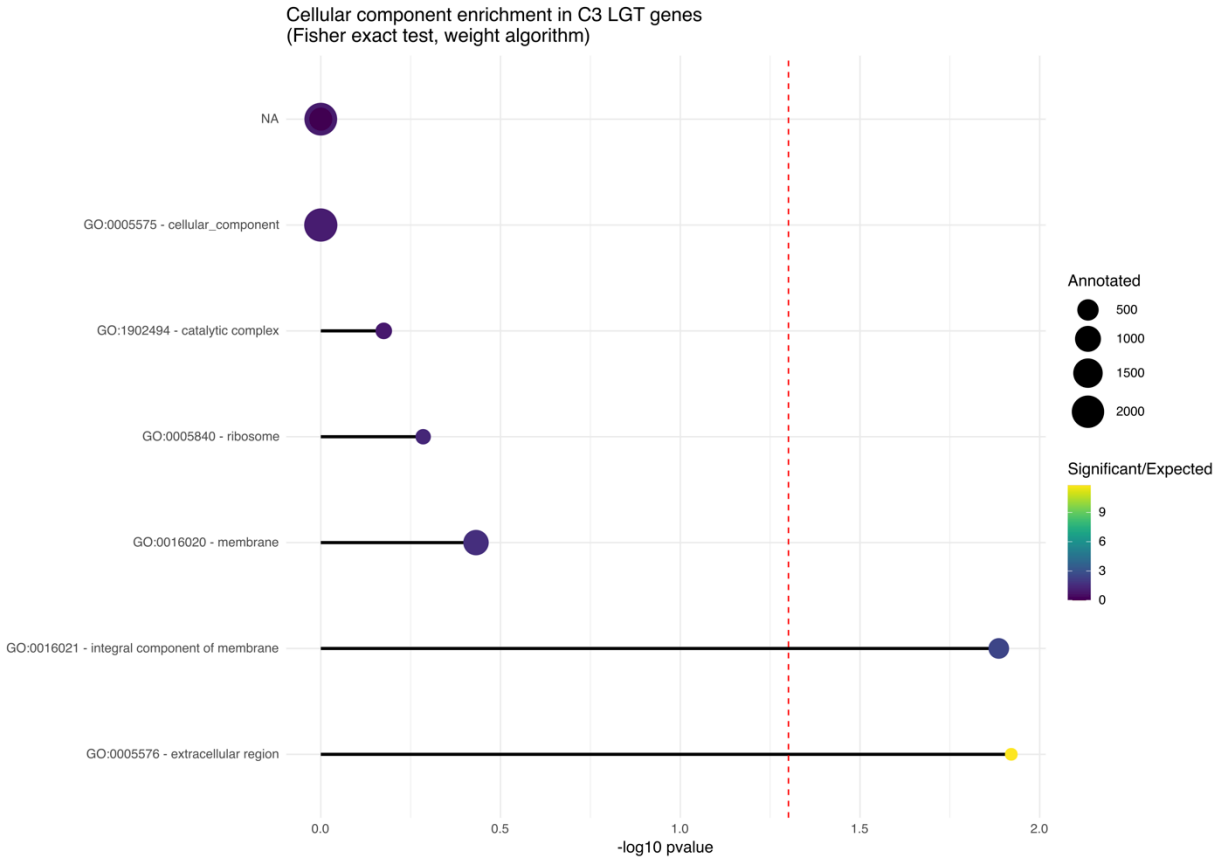


Figure 5.9 Most significant cellular component GO term enrichments among LGT-derived genes in *Acanthamoeba castellanii* strain C3. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

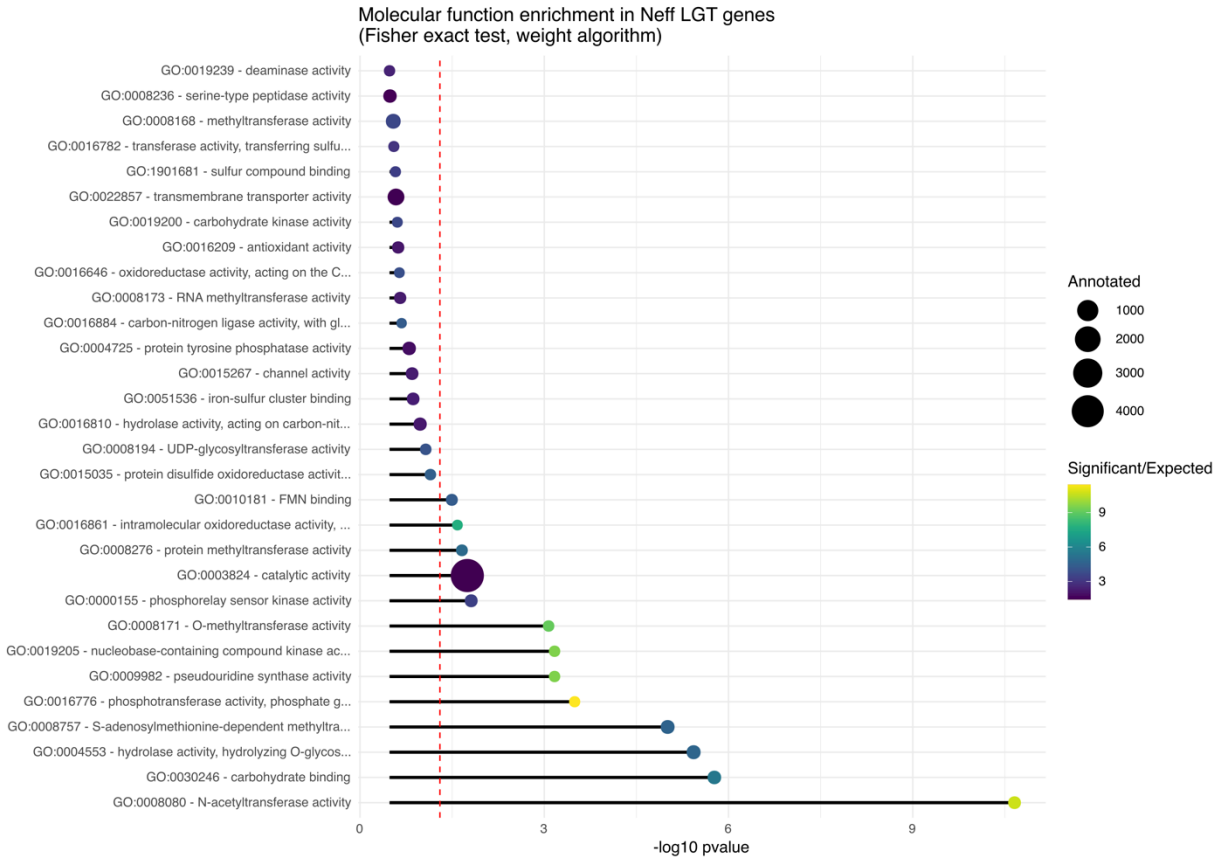


Figure 5.10 Most significant molecular function GO term enrichments among LGT-derived genes in *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

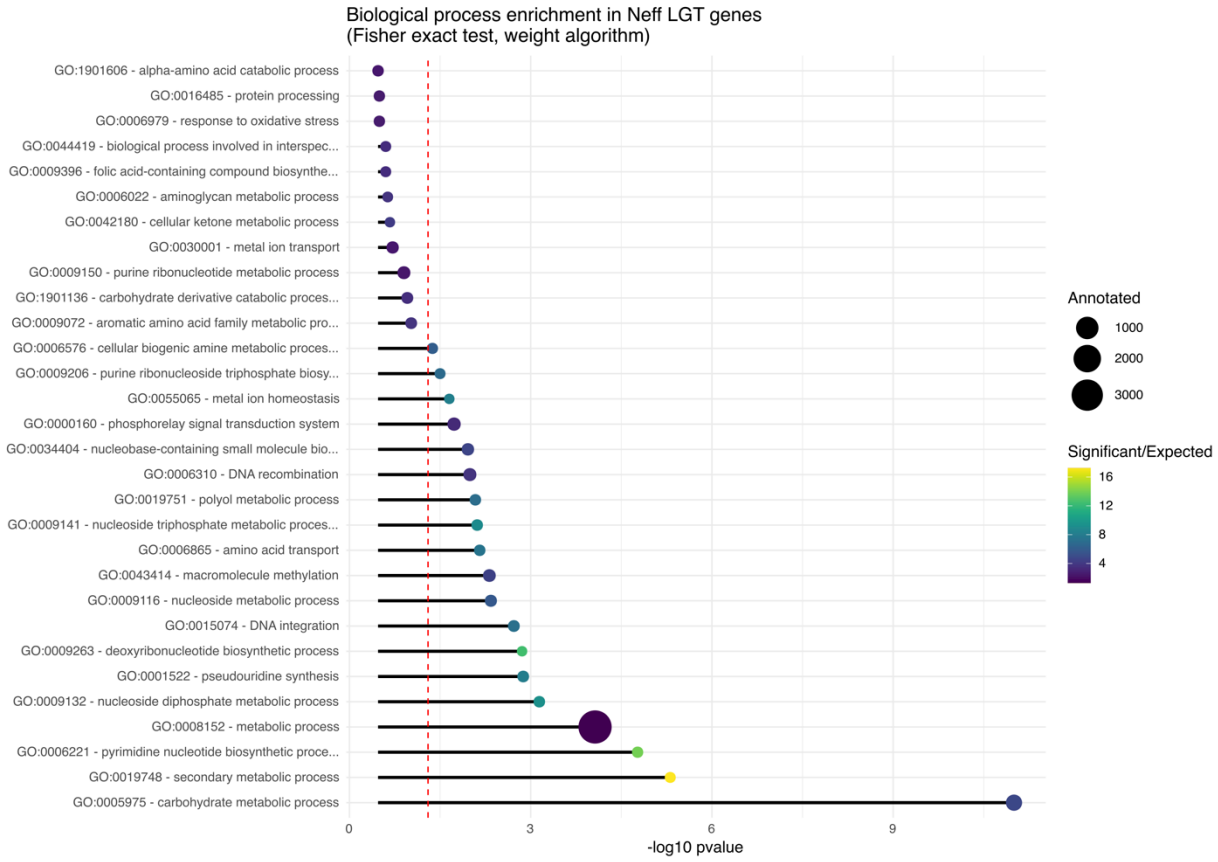


Figure 5.11 Most significant biological process GO term enrichments among LGT-derived genes in *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

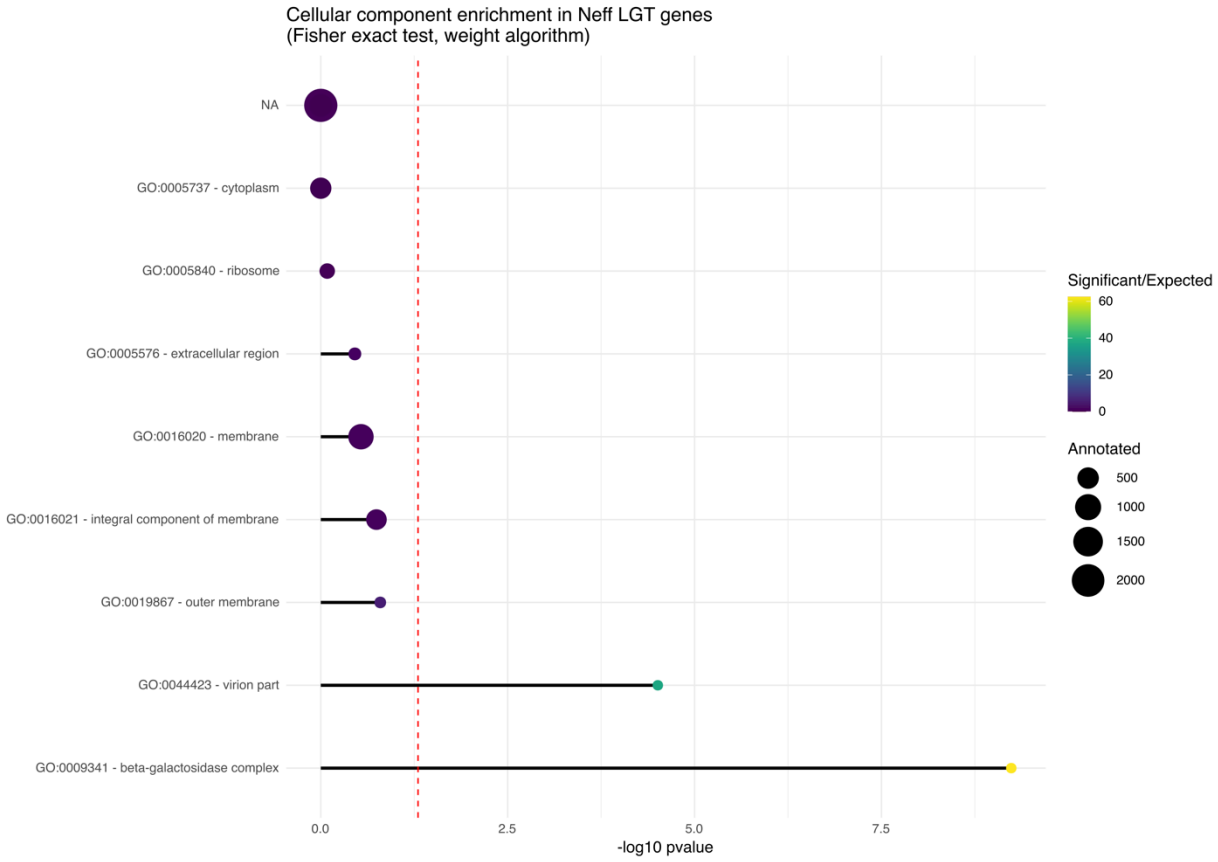


Figure 5.12 Most significant cellular component GO term enrichments among LGT-derived genes in *Acanthamoeba castellanii* strain Neff. Enrichment was determined using topGO, with nodeSize set to 10 when building the GOdata object. The size of the circles at the end of the bars represents the number of genes annotated under that GO term in the genome, and the colour scale of the circles represents the ratio of how many genes were found in the strain-specific set for that term compared to how many were expected.

Table 5.1. Enriched gene ontology terms among LGT-derived genes in *Acanthamoeba castellanii* strain C3.

Ontology	Enriched terms	p-value
Biological process	Carbohydrate metabolic process	9.0×10^{-11}
	Defense response to other organism	9.4×10^{-7}
	Macromolecule methylation	8.3×10^{-6}
	Deoxyribonucleotide biosynthetic process	0.0084
	Polyketide metabolic process	0.0101
	Phosphorelay signal transduction system	0.0149
	Pyrimidine nucleotide biosynthetic process	0.0211
Molecular Function	S-adenosylmethionine-dependent methyltransferase activity	1.7×10^{-8}
	Carbohydrate binding	1.8×10^{-7}
	Acetyltransferase activity	1.0×10^{-6}
	Hydrolase activity, hydrolyzing O-glycosyl compounds	1.0×10^{-5}
	Phosphorelay sensor kinase activity	0.0019
	FMN binding	0.0058
	Catalytic activity, acting on DNA	0.013
	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	0.0144
	O-methyltransferase activity	0.0263
	Protein disulfide oxidoreductase activity	0.0288
	Serine-type peptidase activity	0.0327
Cellular component	Extracellular region	0.012
	Integral component of membrane	0.013

Table 5.2. Enriched gene ontology terms among LGT-derived genes in *Acanthamoeba castellanii* strain Neff.

Ontology	Enriched terms	p-value
Biological process	Carbohydrate metabolic process	1.0×10^{-11}
	Secondary metabolic process	4.9×10^{-6}
	Pyrimidine nucleotide biosynthetic process	1.7×10^{-5}
	Metabolic process	8.6×10^{-5}
	Nucleoside diphosphate metabolic process	0.00072
	Pseudouridine synthesis	0.00133
	Deoxyribonucleotide biosynthetic process	0.00139
	DNA integration	0.0019
	Nucleoside metabolic process	0.00456
	Macromolecule methylation	0.00483
	Amino acid transport	0.00698
	Nucleoside triphosphate metabolic process	0.00769
	Polyol metabolic process	0.00824
	DNA recombination	0.01015
	Nucleobase-containing small molecule biosynthetic process	0.01095
	Phosphorelay signal transduction system	0.01855
	Metal ion homeostasis	0.02225
	Purine ribonucleoside triphosphate biosynthetic process	0.03164
	Cellular biogenic amine metabolic process	0.0423
Molecular Function	N-acetyltransferase activity	2.2×10^{-11}
	Carbohydrate binding	1.7×10^{-6}
	Hydrolase activity, hydrolyzing O-glycosyl compounds	3.7×10^{-6}
	S-adenosylmethionine-dependent methyltransferase activity	9.8×10^{-6}
	Phosphotransferase activity, phosphate group as acceptor	0.00032
	Pseudouridine synthase activity	0.00068
	Nucleobase-containing compound kinase activity	0.00068
	O-methyltransferase activity	0.00085
	Phosphorelay sensor kinase activity	0.01549
	Catalytic activity	0.01791
	Protein methyltransferase activity	0.02193
	Intramolecular oxidoreductase activity, interconverting aldoses and ketoses	0.02595
	FMN binding	0.03201
Cellular component	Beta-galactosidase complex	5.8×10^{-10}
	Virion part	3.1×10^{-5}

There is some degree of overlap in the functions found to be enriched among laterally transferred genes between Neff and C3. For both strains, there is enrichment of ‘carbohydrate metabolic process’, ‘macromolecule methylation’, ‘deoxyribonucleotide biosynthetic process’, ‘phosphorelay signal transduction system’, and ‘pyrimidine nucleotide biosynthetic process’ from the molecular function ontology. From the biological process ontology, ‘carbohydrate binding’, ‘flavin mononucleotide binding’, ‘phosphorelay sensor kinase activity’, ‘hydrolase activity, hydrolyzing O-glycosyl compounds’, and ‘O-methyltransferase activity’ are the common enrichments. The enriched cellular component terms are different between the two strains.

Within the functional enrichment results, some patterns could be observed. In Neff, eight of the 19 enriched biological process GO terms were related to processes involving nucleotides, nucleosides, or nucleobases. This was also reflected in the enriched molecular function terms, where three of the 13 enriched terms also relate to that category of molecules, while at least one other enriched term has ‘child’ terms that would be relevant. It is not clear if those particular child terms are contributing to the enrichment here. The LGT genes from both strains also showed an appreciable representation of metabolic biological processes and molecular functions among their enriched GO terms, especially with respect to carbohydrate metabolism. The two enriched cellular component terms in Neff do not have an obvious connection, but in C3, ‘extracellular region’ and ‘integral component of membrane’ could both apply to cell surface proteins.

A few other enriched terms were interesting in their own right. In Neff, biological processes relating to ‘DNA integration’ and ‘DNA recombination’ were enriched, and in

C3, biological processes relating to ‘defense against other organisms’ were enriched. Like in the strain-specific enrichment analysis from Chapter 2 of this thesis, ‘virion parts’ were enriched in the cellular component ontology, also due at least in part to the acquisition of major capsid protein genes.

5.3.5 A case study of select LGT genes and their donors recapitulates broader trends

I took the opportunity to use some of the LGTs with genus- or species-level donor inferences as case studies of how LGT may shape *Acanthamoeba* biology, including where it is encountering these donor organisms and what capabilities the new genes might be conferring. The inferred genus- and species-level donors, their general habitats and lifestyles, and a description of the LGT gene(s) from each donor are summarized in Table 5.3. A cursory glance at the lifestyles of these bacteria shows that most inhabit environments where interactions with *Acanthamoeba* are plausible. Of the 17 examples, 11 of them are known to inhabit soil environments or the plant rhizosphere, where *Acanthamoeba* is abundant. A few more of these bacteria inhabit various mesophilic freshwater environments, which are also known *Acanthamoeba* habitats. Two of these donor species are known to be marine; *Acanthamoeba* have been observed in marine habitats in some studies[251]. Finally, *Caedibacter* species are known endosymbionts of microbial eukaryotes, and while the taxonomy is somewhat unclear, they may have an evolutionary history of symbiosis with *Acanthamoeba*[252].

Table 5.3. A selection of LGTs predicted in *Acanthamoeba castellanii* for which a donor genus or species could be inferred. See refs. 253-268 for information on the lifestyle of these donors[253–268].

Donor	Lifestyle	Signal peptide?	Description
Acidisarcina polymorpha	acidic soil, degrades complex polysaccharides	yes	zinc-binding dehydrogenase; quinone oxidoreductase
Acidobacterium sp. S8	soil acidophile	no	cysteine-rich CPCC domain
Ammoniphilus sp.	rhizosphere/plant detritus, oxalotrophic	no	universal stress protein A-like
Aquimonas voraii	warm spring water	no	has cytosolic, transmembrane, extracellular domains
Azospirillum sp.	plant rhizosphere, nitrogen fixer	no	unknown function
Caedibacter sp.	Endosymbionts of paramecia, acanthamoebae	yes	protein binding, host-pathogen enzyme activity, nitrous oxide reductase, periplasmic binding
...	...	yes	protein binding, leucine binding, host-pathogen enzyme activity, periplasmic binding
...	...	yes	lipid metabolic process, secreted monoacyl-/diacylglycerol lipase
Cystobacter ferrugineus	soil	yes	polycystic kidney disease domain
Escherichia coli	enteric and soil	no	ATP binding, ATP hydrolysis activity, bacterial antiviral defense nuclease
Nostoc sp.	phototroph, terrestrial and aquatic	no	catalytic activity, carbohydrate metabolic process, glycosyl hydrolase, alpha-amylase
Opitutus sp.	soil anaerobe	yes	glycosyl hydrolase, arabinanase/levansucrase/invertase
Pseudoalteromonas luteoviolacea	marine	no	phenylacetic acid degradation-related, 1,4-dihydroxy-2-naphthyl-CoA thioesterase
Pseudocolwellia agarivorans	marine, agar-digesting, facultative anaerobe	yes	bacterial immunoglobulin
Rhodomicrobium lacus	alkali-tolerant freshwater phototroph	no	glycosyltransferase (extracellular), transmembrane domain, cytosolic domain
Sorangium cellulosum	aerobic soil/detritus environments, sociality	no	GAF-like domain
Streptomyces sp.	soil, possible symbiont of plants, animals, fungi	yes	carbohydrate metabolic process, gylcosyl hydrolase (alpha-L-iduronidase)
...	...	no	proline oxidase/dehydrogenase, FAD-linked oxidoreductase
Vulgatibacter incomptus	heterotrophic soil aerobe	yes	aspartyl/glutamyl-tRNA amidotransferase subunit B-like
Xanthomonas sp.	plant pathogen	no	carbohydrate metabolic process, catalytic activity, 6-hairpin glycosidase

To supplement the GO terms already assigned to the proteomes of our two *Acanthamoeba* strains, I used InterProScan[147] to search for additional functional information on the candidate LGT genes for this case study. Not only was this useful to predict functional domains, but it also revealed which of these proteins have a predicted signal peptide and may be secreted. Overall, the functions of these proteins proved to be consistent with the results of the enrichment analysis. There were several proteins predicted to perform metabolic reactions, especially on carbohydrate molecules. Some other noteworthy examples are the detection of a universal stress protein A-like domain transferred from *Ammoniphilus*, host-pathogen enzyme activity predicted for a few of the proteins transferred from *Caedibacter*, and a bacterial antiviral defense nuclease domain encoded by the gene transferred from *E. coli*. Of the 20 different transferred genes, nine encode a predicted signal peptide, and five of those were predicted to be either a hydrolase or an oxidoreductase.

5.4 Discussion

The aim of this study was to revisit the LGT analysis performed by Clarke et al.[20] in the first *Acanthamoeba castellanii* strain Neff genome paper to see how much more could be learned with more modern methods and increased data availability. This study used the predicted proteome of the updated Neff assembly presented in Chapter 2 of this thesis, as well as the predicted proteome of C3, for which no genome was available in 2013[20]. These two new predicted proteomes afforded the possibility of detecting more LGTs by virtue of being more complete than previous proteomes, and the greatly expanded taxon sampling in public databases was expected to improve the phylogenetic

inference and sampling of potential donors required for LGT detection.

It can be seen immediately that purely in the total number of laterally transferred genes in the Neff strain, this study and the 2013 study arrived at a similar number. The 2013 analysis inferred 450 genes to have been acquired by LGT while this study has detected 447. The C3 strain, not analyzed for LGT previously, was inferred to have acquired 268 genes by LGT, despite having a slightly larger total predicted proteome than the Neff strain. In Neff, both studies arrive at a similar domain-level breakdown of LGT donors. In the 2013 study, out of 450 predicted LGTs, 13 were attributed to archaeal donors and 11 to viral donors. In my study, of 447 predicted LGTs in the Neff strain, 16 were predicted to have viral donors and 13 were predicted to be archaeal in origin. However, it is possible that a larger proportion of the predicted LGTs in this study could be viral or archaeal due to the few genes for which no donor lineage could be pinpointed, and the 76 that could not be distinguished between being bacterial or archaeal in origin (Fig. 5.1A).

Although the new Neff predicted proteome has just over 600 more proteins than the predicted proteome from 2013, I predicted 10 fewer genes to be acquired by LGT. There are several differences between the two studies that could help explain this, but none are exceedingly obvious. First, the screening step for candidate LGTs in my study used a 90% ultrafast bootstrap support cutoff for considering nodes during LGT detection, while the screening step in the 2013 analysis used 75% real bootstrap support. Given the general understanding of how ultrafast bootstrap support values compare to real bootstrap support values[269], those cutoffs are likely to be similarly strict in practice, but it is formally possible that this could contribute to the minor difference in number of

predicted LGTs.

Another consideration is the method used for candidate LGT screening. In the 2013 analysis, candidates were defined by having a node in the tree where an amoeba protein was with bacterial, archaeal, or viral proteins with no more than two other protists, and at least 75% bootstrap support. However, the LGT detection method in the present study considered the taxonomic breakdown at all nodes in the tree that met the 90% ultrafast bootstrap support threshold. Perhaps by narrowing the focus to only the part of the tree closest to the amoebal protein, the 2013 method was slightly more permissive of candidate LGTs that could be rejected by considering the whole tree topology.

Finally, an intuitive expectation for the effect of greater taxon sampling and better phylogenetic modelling might be that more LGTs would be predicted because it would be easier to retrieve homologs from donors, and fewer candidates would have to be rejected due to poor support values. However, one can imagine both advancements having the opposite effect as well, lowering false positive rates due to more accurate phylogenetic inference and the potential of adding more sequences that branch between the query protein and homologs from what would otherwise appear to be donors. It is hard to predict whether these forces would push the results toward more or fewer predicted LGTs, and perhaps it varies on a case-by-case basis, but it could be part of the explanation for why this study inferred slightly fewer LGTs than the previous one. Ultimately, the numbers are not that different and based on how strong the patterns in the data are, such as in the functional enrichment analysis, perhaps there is not much difference in the high-level biological insight that can be gained.

With respect to the two strains analyzed in the present study, the methodology for

inferring LGT was the same so it is possible to explore the difference in inferred LGTs from a biological perspective. While C3 has 1340 more predicted proteins than Neff, it only had 268 predicted to be acquisitions by LGT, compared to 447 in Neff. For C3, 181 of these are shared with Neff, and for Neff, 205 are shared with C3. The difference in number is due to a greater number of paralogs arising in Neff after the two strains diverged. This result raises questions about the differences between these strains. First is the question of why C3 might have fewer LGTs than Neff.

This question is difficult to answer given what we know about these two strains. The Neff strain was isolated from soil in North America in 1957[27], while the C3 strain was isolated from a water reservoir in Europe in the 1990s[270], but beyond that information, any lifestyle differences are unclear. As described in Chapter 2 of this thesis, it is known that in laboratory experiments, C3 is more susceptible to *Legionella* infection than Neff, but it would make more sense for less infection to lead to fewer opportunities for LGT, and this is only one possible intracellular bacterium either strain could have encountered in its evolutionary history. Similarly, differences between Neff and other *Acanthamoeba* strains are often explained by its extensive time spent in laboratory culture, but it would not make sense for axenic culturing to be responsible for *more* genes derived from LGT. Given what is known about the two strains, an inference cannot be made at the moment about why their LGT content differs in scale, but some possibilities include differences in food uptake, molecular biology machinery, or effects of culturing beyond which strain may have been exposed to LGT donors most recently. Unpublished studies by Cédric Blais in the Archibald lab have found fewer genes of putative viral origin in C3 than Neff, and viral genes in Neff are also often found in clusters, which is

less commonly observed in C3 (Blais, pers. commun.). These findings are consistent with the lower number of inferred LGTs into C3.

5.4.1 *Acanthamoeba* prey selection may influence the source of its LGT genes

The bias toward acquiring LGTs from Pseudomonadati over other bacterial kingdoms, especially Bacillati, is puzzling at first. Seeking an explanation for this result in the biological and ecological characteristics of the different bacterial lineages is not especially compelling; both are common in the same environments as *Acanthamoeba*, and perhaps Bacillati even more than Pseudomonadati. The diverse metabolic capabilities among members of Pseudomonadati could perhaps be argued to provide more opportunity for niche adaptation if soil protists were to sample their genetic material, leading to more frequent fixation of genes transferred from this kingdom, but this feels like a tenuous explanation, especially when comparing at such a broad taxonomic level. At the time of writing, there were roughly 1.6 times as many proteins in the NCBI Protein database from Pseudomonadati taxa than from Bacillati taxa, so taxon sampling likely has some influence on this result. However, the inferred number of LGTs in this study from Pseudomonadati donors is roughly 3.6 times as many as from Bacillati donors, so other factors are probably also involved.

A key piece of information for unravelling this pattern is that soil amoebae appear to strongly prefer grazing on Gram-negative bacteria[271]. Under the ‘you are what you eat’ hypothesis of LGT[272], this would suggest a bias toward acquisition of genes from Gram-negative bacteria. The two major Gram-positive-containing phyla are Actinomycetota and

Bacillota, both members of the kingdom Bacillati. Therefore, if phagocytosis is a major route for LGT, a bias toward acquiring genes from Pseudomonadati over Bacillati would be expected.

5.4.2 Genes that are laterally transferred tend to be acquired independently across eukaryotes

The patterns described above from the trees demonstrating prokaryote-to-eukaryote transfer can provide some insight into the different possible trajectories of genes being transferred from prokaryotes to eukaryotes. Each pattern can be approximately equated with a different hypothetical evolutionary scenario. A tree with *Acanthamoeba* appearing alone with many prokaryotic sequences is consistent with a single, relatively recent transfer from one donor into a recent ancestor of *Acanthamoeba*. Expanding this to *Acanthamoeba* and additional amoebozoans would still suggest a single transfer, but a less recent one such that more divergence has occurred since the transfer of the gene in question. A clade of *Acanthamoeba* or multiple amoebozoans with additional eukaryotes formally has two possible explanations depending on the identity of the members of that clade. An even earlier transfer would be seen in Amoebozoa and closely related lineages, but a clean example of this was not seen in the trees generated for this study. The other explanation for such topologies is that the gene was transferred a single time from a prokaryote to a eukaryote and was subsequently spread among eukaryotes in one or more eukaryote-to-eukaryote transfers. Finally, a tree where eukaryotes branch from within prokaryotes at multiple positions suggest that the gene in question has been independently transferred from prokaryotes to eukaryotes on multiple occasions with different donor and recipient lineages.

The frequency of each of these patterns in the data generated here may hint at more general patterns of eukaryote LGT. It is noteworthy that not only are cases of multiple independent transfers observed more frequently than any of the other patterns, they actually represent more than half of the total observations. One implication of this finding is that prokaryote-to-eukaryote LGT must be relatively frequent; even when limiting the scope to predicted LGTs from a single species, the phylogenies of those genes still reveal hundreds of examples where there were likely independent transfers into other eukaryotic lineages. These results could also be interpreted to support the idea that particular types of genes are more frequently involved in lateral gene transfer than others; if genes acquired by *Acanthamoeba* are more often than not also acquired independently by unrelated eukaryotes, it suggests that there is utility in having access to the capabilities they confer, which would otherwise not be available. As such, the phylogenetic pattern that has emerged appears to be consistent with the functional enrichment analysis in demonstrating a bias toward particular genes in prokaryote-to-eukaryote LGT. Further analysis to identify the functions encoded by genes with this type of phylogenetic pattern would be a valuable step toward supporting this hypothesis, but this investigation has not been performed at this time. The functional enrichment analysis as currently presented was agnostic to the specific phylogenetic patterns within the trees, so while functional annotations are known for all trees, an additional functional analysis with more restricted genes of interest would be needed to evaluate enrichment in any particular subset of the LGT genes.

The second most common of the four described patterns, a clade of diverse eukaryotes branching from within prokaryotes, is a topology that has caused some debate

amongst LGT researchers. There has been some reluctance, even as LGT into eukaryotes increasingly appears to be common and important in their evolution, to infer an LGT from a prokaryote into a eukaryote, followed by eukaryote-to-eukaryote transfers of the same gene that distribute its function throughout the eukaryote tree of life. While this is often the best explanation for the data, and biologically the most reasonable, it can be met with suggestions that since eukaryote LGT has long appeared to be rare and the mechanisms for it are not clearly described, it is easier to infer that the distribution of such genes has resulted from differential loss across eukaryotes[1–3]. However, with 157 such trees generated from this study alone, it is hard not to invoke the so-called "Genome of Eden" problem[273], which would require an untenably large genome in the last eukaryote common ancestor to explain all these results as differential loss. Therefore, the alternative implication of these 157 trees is that prokaryote-to-eukaryote LGT followed by subsequent spread among eukaryotes may not be such a rare phenomenon, and it may not be necessary to shy away from this hypothesis when it is the best explanation for the data.

5.4.3 Both *Acanthamoeba* strains may have increased flexibility in carbohydrate and nucleotide metabolism

The functions inferred to be enriched among the laterally transferred genes in both of these strains are broadly consistent with what we have come to expect of LGT. The conventional wisdom is that there is a bias toward acquiring functions that would facilitate niche adaptation and ecological success over functions central to the basic functioning of the cell[183,185]. This can be thought to roughly correspond with more LGT of genes that provide additional 'accessory' capabilities like accessory metabolism or

additional transporters, while there is less frequent LGT of genes involved in information processing, basic cell biology, and core metabolism[185]. Within the biological process ontology, the enriched terms in the LGT genes of both Neff and C3 are dominated by metabolic and biosynthetic processes. In terms of molecular functions, both strains have activity enrichment in hydrolases, transferases, oxidoreductases, and catalytic activity more broadly, while Neff also has enrichment in synthase activity and C3 has enrichment in peptidase activity. The enrichments observed in these two ontologies are very consistent with what has come to be expected of genes involved in lateral gene transfer[99,183–185].

Looking holistically at the enriched biological process and molecular function terms in both strains reveals some overarching patterns that may explain how *Acanthamoeba* is benefiting from the genes it has acquired by LGT. Among the many enriched nucleotide- and nucleoside-related biological process terms, several specify that the process is biosynthetic rather than catabolic. Although it is to be expected that the ancestors of *Acanthamoeba* would be able to fulfill their nucleoside and nucleotide needs without LGT, these acquisitions could provide greater flexibility in meeting those needs from a wider range of sources. Enrichment of amino acid transport in Neff also hints at an expansion of the capacity to acquire nutrients. Acquiring genes relating to carbohydrate metabolic processes also invokes the idea of more flexible energy metabolism, a hypothesis that is supported by some of the enriched molecular function terms such as ‘intramolecular oxidoreductase activity, interconverting aldoses and ketoses’, which relates to sugar metabolism. Taken together, these GO term enrichments suggest that LGT has allowed *Acanthamoeba* to be more flexible in meeting its nutritional needs.

There are a few other enriched terms worth specific consideration in each of the two strains. It is interesting to see enrichment of DNA integration and DNA recombination among the LGT genes in Neff. It almost seems self-explanatory in the sense that mobile genetic elements and viruses often encode integrases or recombinases that facilitate their mobility and integration into new genetic loci, potentially in new hosts. Therefore, the enrichment of these terms in the context of LGT could be due to genes with those functions driving their own transfer. It is also possible that after their acquisition, they can facilitate the integration of additional fragments of foreign DNA and drive additional lateral gene transfer. The enrichment in C3 of the biological process 'defense response to other organism' is interesting, but perhaps not surprising. *Acanthamoeba* often lives in very rich and complex communities that could well pose the threat of infection or predation. The range of viruses and bacterial symbionts that are known to associate with *Acanthamoeba* further illustrate why a defense response may be necessary. It is easy to imagine that *Acanthamoeba* would benefit from the ability to protect itself and manage how it interacts with other members of its community to avoid being exploited.

Each strain has just a couple of enriched terms from the cellular component ontology. In Neff, one of these, 'virion part', has already been discussed in Chapter 2 of this thesis in the context of strain-specific functional enrichment, and it is not surprising to also see it enriched here. By definition, virion-associated proteins are not native *Acanthamoeba* features, and likely became integrated after a past viral infection. In particular, the virion parts found here are mostly major capsid proteins. Further investigation is needed to determine whether the amoeba expresses and makes use of this

protein in some way or if it only has relevance to viral infection. The other enriched cellular component in Neff, ‘beta-galactosidase complex’, can in part be explained by the enrichments in carbohydrate metabolic processes and glycosyl hydrolase activity from the other ontologies; this complex participates in both, so the acquisition of a beta-galactosidase complex would be advantageous because it provides those capabilities. In C3, the enriched cellular component terms ‘extracellular region’ and ‘integral component of membrane’ could be related. One can imagine various surface proteins and receptors would be exposed to the exterior of the cell, but also be anchored within the cell membrane. However, further investigation of the precise proteins annotated with these terms would be needed to confirm this connection.

Finally, it is worthwhile to summarize the whole picture of how the functions acquired by LGT compare between strains. Overall, the biological processes and molecular functions enriched in both strains follow a similar trend of including a lot of enzymatic activity involved in metabolism, especially of carbohydrates and nucleotides. The Neff strain has a somewhat more expanded assortment of specific processes and functions that are enriched, potentially due to its higher number of predicted LGTs, but they still generally follow the same trends. Terms such as ‘DNA integration’ and ‘defense against other organisms’ that were singled out above as being interesting, but did not contribute to the prevailing trends in the functional enrichment analysis. However, these terms probably represent the most distinct differences between the functions enriched in the two strains. The cellular component enriched terms also differ entirely between strains, but there are very few enriched terms in this ontology so their effect on the biology of the two may not be very dramatic.

5.4.4 A case study on how *Acanthamoeba* may use LGT to exploit its environment

The examination of LGT from genus- and species-level donors to *Acanthamoeba* (Table 5.3) has proved to be an interesting microcosm of what appears to be the overarching trend among the transfers predicted in this study. This selection of LGT genes and their donors does include LGTs that are specific to either Neff or C3, and several that are shared. However, my vision for this case study is to connect each example with the overarching trends from the enrichment analysis and *Acanthamoeba*'s lifestyle in the environment, which does not strictly require each example to connect to the others. This means that these specific examples generally remain illustrative of the overall picture, despite not all of these acquired genes and their functions being applied simultaneously in the same organism.

While this particular subset of LGTs is somewhat more taxonomically restricted than the full range of predicted donors, it still recapitulates the trend of a small fraction of donors having an intimate association with *Acanthamoeba* in their capacity as endosymbionts or viruses, while the remainder occupy diverse physical and ecological niches where they may be members of the same communities as *Acanthamoeba*. The story is similar with respect to the functions assigned to this subset of LGTs. Like in the overall enrichment analysis, carbohydrate metabolism and nucleotide metabolism are well represented, and there are glycosyl hydrolases, oxidoreductases, transferases, and esterases, covering a wide range of different types of reactions (Table 5.3).

Looking more closely into the identities and lifestyles of these putative donors, it is easy to imagine opportunities for interaction with *Acanthamoeba*. Most are known to

broadly inhabit soil environments, or more specifically to be part of the plant rhizosphere, where *Acanthamoeba* is known to graze and play an important role in shaping the microbial community[274]. Under the 'you are what you eat' hypothesis of lateral gene transfer[272], *Acanthamoeba* would have a wealth of potential gene donors available in these environments, including the donors mentioned here. *Caedibacter* in particular is an endosymbiont of ciliates but is thought to have coevolved with acanthamoebae at some point in the past, providing a very clear window of opportunity for genetic exchange[252].

Considering the functions conferred by this selection of LGTs allows a picture to be painted of exactly how *Acanthamoeba* could better exploit its environment with these additional capabilities. Several of the enzymes are predicted to have a signal peptide and may be secreted, which could allow *Acanthamoeba* to favourably alter its immediate environment in various ways. In the case of oxidoreductases, perhaps this is a means of detoxification; for example, there is a quinone oxidoreductase with a predicted signal peptide that may serve to protect against reactive oxygen species.

There are also several glycosidases and a lipase in this set of genes that are predicted to have a signal peptide. Osmotrophy is an obvious possibility for the employment of these enzymes; they could be secreted to digest macromolecules in the environment such that the products can be absorbed for nutrition. However, there is another less direct possibility for the use of glycosyl hydrolases in acquiring nutrition. *Acanthamoeba* is known to graze on biofilms, where its prey organisms are enveloped in a thick extracellular matrix[275]. Polysaccharides are one of the polymers that often contributes to these biofilms. Perhaps *Acanthamoeba* could secrete hydrolases into the extracellular matrix of biofilms to digest them and more easily access and phagocytose

prey cells. While energy metabolism was mentioned above as the most straightforward explanation for an enrichment in carbohydrate metabolism, this case study and the knowledge that some of the hydrolases are secreted adds another, possibly complementary, explanation for this enrichment. The presence of an amylase in this set of proteins that is not predicted to be secreted does support the earlier hypothesis that utilizing a wider range of macromolecules for energy is one of the capabilities involved in the expansion of carbohydrate metabolism.

Of the remaining proteins in this set that have not yet been directly discussed, there are some where the benefit is not immediately obvious, or the functional prediction is not very detailed. This includes a bacterial immunoglobulin, a cysteine-rich CPCC domain, a polycystic kidney disease domain, and a GAF-like domain. There are still some interesting proteins to mention, though. The relatively straightforward explanation for acquiring a protein with the universal stress protein A-like domain is simply that it provides additional capacity to handle environmental stresses. It is interesting that there is a transfer predicted from the endosymbiotic *Caedibacter* that has a 'host-pathogen enzyme activity' GO term annotation; presumably this would have been used natively by the bacterium as an effector to control the host, but it is not clear how this may be repurposed by *Acanthamoeba*. The predicted transfer of a bacterial antiviral defense nuclease from *E. coli* is interesting. Given *Acanthamoeba*'s known interactions with viruses, it may have repurposed this enzyme for its own protection against viral infection.

Overall, investigating this subset of LGT genes has been an interesting case study to neatly illustrate how *Acanthamoeba* may be taking advantage of the resources available within its ecological community and applying them to its own advantage. While

only a handful of genes and putative donors have been covered here, it seems reasonable to expect that this sample is at least partly representative of the more extensive acquisition and employment of laterally transferred genes that has been revealed by this study as a whole.

5.5 Conclusions

This study aimed to build on the LGT survey performed by Clarke et al.[20] by leveraging methodological advances, improved taxon sampling in databases, and an updated, chromosome-scale reference genome sequence to more thoroughly investigate LGT into *Acanthamoeba castellanii* strain Neff. In addition, a new high quality reference genome sequence was also published for *A. castellanii* C3, which afforded the opportunity to compare the influence of LGT on these two strains. A similar number of LGTs into Neff were identified as previously, while C3 was found to have around 60% as many genes acquired via LGT as Neff. This study did provide a more comprehensive functional analysis of the genes acquired by LGT in both strains, revealing a profound tendency toward acquiring genes involved in expansion of metabolic capabilities.

Like in the previous study, the vast majority of LGTs were predicted to come from Bacteria, with minor contributions from Archaea and viruses. The inferred donors spanned most of bacterial diversity, but there was a bias toward acquisitions from the kingdom Pseudomonadati over the kingdom Bacillati, which may have implications for supporting the idea that phagotrophy is a driver of eukaryote LGT; Bacillati is home to the Gram-positive bacteria, which are not preferred prey of *Acanthamoeba*. The inferred donors from Archaea and viruses were more taxonomically restricted. All transfers from

viruses came from the Nucleocytoviricota, which is not surprising as the viruses known to infect *Acanthamoeba* predominantly come from this lineage.

A case study on a small subset of the LGT genes allowed me to infer a trend that was generally consistent with the higher level findings of the study. Investigating this subset of genes demonstrated that *Acanthamoeba* lives in diverse microbial communities where the collective community members possess a wide range of metabolic and other capabilities that allow them to exploit their environment and protect their own interests. Through grazing, symbiosis, and likely other mechanisms, *Acanthamoeba* is able to sample the rich diversity of coding capacity in its environment to acquire some of those capabilities for its own use, allowing for improvements in self-defense, resistance to stress, acquiring nutrition, and more. Referencing the overall enrichment analysis and breakdown of donor taxa inferred in this study suggests that this narrative is representative of the entire repertoire of genes *Acanthamoeba* has acquired by LGT, albeit from a wider range of donors that also include Archaea and viruses. Overall, this study demonstrates that, while LGT is not always strikingly (and obviously) transformative for its recipients, it can provide a consistent source of genetic material to allow microbial eukaryotes to thrive in their respective environments.

CHAPTER 6 OVERALL CONCLUSIONS

In this thesis, I cast a wide net with the goal of compiling as much information as possible across all aspects of *Acanthamoeba* genome biology and evolution. The intention was to serve the research community on two different levels. The first level recognizes the medical and ecological impact of *Acanthamoeba*, its utility as a cell biological model, and its status as one of the better-developed free living protozoan model systems. I aimed to lean into these facts to provide an even deeper understanding of genome biology in this organism with the goal of maximizing its experimental utility, while potentially becoming a more generalized model protist for studying biology within Amoebozoa and eukaryote-wide. The second level on which I hoped to provide value was by providing data and hypotheses that can inform how the community interprets and investigates eukaryote genome biology as a whole. I hope that the findings of this thesis can prime other researchers to notice, seek out, or simply not dismiss the kinds of phenomena I have described here, and that synthesizing my findings with those of others from other systems will shape the way 'normal' eukaryote genome biology is conceptualized. To this end, the distinct chapters of this thesis focus on expanding the knowledge and thinking around different aspects of *Acanthamoeba* biology.

In Chapter 2, assembling chromosome-level reference genome sequences for strains Neff and C3 added value in several ways. The completeness afforded by this high level of contiguity has allowed an accurate inventory of all genes present in both strains. The chromosome-scale resolution of the genomes also greatly improves our structural understanding, allowing the inference of a karyotype and providing a foundation for studying processes and mechanisms of genome biology that act at the chromosome level.

Having predicted proteomes from two related strains also allowed me to investigate fine-scale genome variation between the two. All of these outcomes are clearly valuable for better understanding *Acanthamoeba* but their impacts can also be extended more broadly. High quality genomic resources for microbial eukaryotes are needed in general for comparative genomics and phylogenetics studies, among others. Chromosome-level processes revealed by a highly contiguous assembly can inform the research of such processes in other systems. Finally, patterns detected in the fine-scale genome variation between these two strains can be considered when determining patterns common to all instances of fine-scale genome variation, as this becomes a more integral part of studying eukaryote evolution.

Chapter 3 provided both biological and methodological insights from its study of the fate of transgenes. On the biological side, an unexpected mechanism appears to be responsible for maintaining transgenes in *Acanthamoeba*. The transforming plasmid, linear or circular, becomes tandemly duplicated several times and has telomeres added to the ends, where it can be maintained like a mini-chromosome. This phenomenon has been observed in other eukaryotes but this finding suggests it is more widespread than previously thought. My results also indicate that episomal maintenance and chromosomal integration are not mutually exclusive, and the former may facilitate the latter by allowing DNA to linger in the nucleus for longer. This may have implications for facilitating LGT. On the methodological side, this study has demonstrated the application of long-read sequencing technology as a high-throughput tool of molecular biology, albeit one that may need some additional verification. The trade-off for the need to validate, though, is being able to screen more easily for particular molecules in a sample. This study also

highlights the importance of molecular biologists continuing to have wet lab methods in their repertoire. At the same time, the findings in this chapter provide insights on how we may improve the genetic tools for *Acanthamoeba* in the future by exploiting the properties of its linear episome.

The results from Chapter 4 add significant depth to our knowledge on *Acanthamoeba* ploidy. While they could not provide absolute quantitation of chromosome number per cell, they revealed the prominent aneuploidy that appears to be an intrinsic feature of *Acanthamoeba* genome biology. Further study of this aneuploid state may reveal some functional or adaptive properties that it confers. There is a hint at a possible cryptic diploid stage in this organism as well, but further study will be needed to support this hypothesis. Moving forward on the topic of *Acanthamoeba* ploidy, cytogenetic methods may be the most direct route to more answers about this system, and researchers can take lessons from similar studies in *Leishmania* to pursue this avenue of investigation.

Chapter 5 moves away from a focus on present-day processes and mechanisms and looks at the legacy those processes have left in the form of past lateral gene transfers in *Acanthamoeba*. This study serves as an update to the LGT survey conducted during the original strain Neff genome project[20], while also bringing C3 into the mix. The number of predicted LGTs in Neff is similar to what was found previously, and the addition of inferring LGTs in C3 is direct value added on top of this. A synthetic analysis and summary provides some additional takeaways. Where donors could be inferred, bacteria dominate as a source of LGTs into *Acanthamoeba,* with a bias in taxonomic distribution that may hint at a role in feeding and prey selection when it comes to LGT.

Unsurprisingly, the viral LGT contributions are from Nucleocytoviricota, a taxon which contains the majority of known *Acanthamoeba* viruses. Functional enrichment analysis reveals a strong trend toward acquisition of metabolic functions, which aligns with contemporary expectations of lateral gene transfer. A deeper look at a subset of the LGTs illustrates how microbial eukaryotes can constantly fine-tune their fitness in their present environment, and LGT can influence their evolution without always being accompanied by a rapid and striking transition.

The work I have presented in this thesis brings a wealth of data, analysis, and interpretation to bear on several aspects of *Acanthamoeba* genome biology and evolution, broadening and filling out our understanding of each one of them. When I started developing *Acanthamoeba* as an experimental model with the goal of studying lateral gene transfer, I was oblivious to the genetic and genomic complexities I would face. However, that is what makes the body of work presented in this thesis so valuable. From being such a 'black box' that I could not pursue my originally intended experiments, *Acanthamoeba* is now understood to be very complex in terms of genome biology. The precise details of this complexity are still not fully understood, but crucially, we do understand generally in what *ways* it is complex. My hope, and belief, for the future impact of this thesis work is that *Acanthamoeba* researchers will be able to develop the experimental systems and perform the type of investigations I had originally envisioned as a result of the substantial foundation I have laid here, while also further exploring the complexities of its genome biology through bioinformatic and molecular biology methods. Beyond *Acanthamoeba*, I hope and expect this work to contribute to an understanding in the scientific community that eukaryote genome biology and evolution

is more complex and dynamic than previously understood, due to an underappreciation of microbial eukaryote genome biology.

References

1. Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F.L., Lockhart, P.J., Bryant, D., Hazkani-Covo, E., McInerney, J.O., Landan, G., and Martin, W.F. (2015). Endosymbiotic origin and differential loss of eukaryotic genes. Nature *524*, 427–432. https://doi.org/10.1038/nature14963.

2. Ku, C., and Martin, W.F. (2016). A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. BMC Biol *14*, 89. https://doi.org/10.1186/s12915-016-0315-9.

3. Martin, W.F. (2017). Too Much Eukaryote LGT. BioEssays *39*. https://doi.org/10.1002/bies.201700115.

4. Martin, W.F. (2018). Eukaryote lateral gene transfer is Lamarckian. Nat Ecol Evol *2*, 754–754. https://doi.org/10.1038/s41559-018-0521-7.

5. Stegemann, S., Hartmann, S., Ruf, S., and Bock, R. (2003). High-frequency gene transfer from the chloroplast genome to the nucleus. Proceedings of the National Academy of Sciences *100*, 8828–8833. https://doi.org/10.1073/pnas.1430924100.

6. Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The New Tree of Eukaryotes. Trends Ecol Evol *35*, 43–55. https://doi.org/10.1016/j.tree.2019.08.008.

7. Raĭkov, I.B. (Igor′ B. (1982). The protozoan nucleus, morphology and evolution (Springer-Verlag).

8. Friz, C.T. (1968). The biochemical composition of the free-living Amoebae Chaos chaos, amoeba dubia and Amoeba proteus. Comp Biochem Physiol *26*, 81–90. https://doi.org/10.1016/0010-406X(68)90314-9.

9. Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., et al. (2005). The genome of the social amoeba Dictyostelium discoideum. Nature *435*, 43–57. https://doi.org/10.1038/nature03481.

10. Byers, T.J. (1986). Molecular Biology of DNA inAcanthamoeba, Amoeba, Entamoeba, and Naegleria. In, pp. 311–341. https://doi.org/10.1016/S0074-7696(08)61430-8.

11. Lohia, A. (2003). The cell cycle of Entamoeba histolytica. Mol Cell Biochem *253*, 217–222. https://doi.org/10.1023/A:1026055631421.

12. Hofstatter, P.G., Brown, M.W., and Lahr, D.J.G. (2018). Comparative Genomics Supports Sex and Meiosis in Diverse Amoebozoa. Genome Biol Evol *10*, 3118–3128. https://doi.org/10.1093/gbe/evy241.

13. Bloomfield, G., Skelton, J., Ivens, A., Tanaka, Y., and Kay, R.R. (2010). Sex Determination in the Social Amoeba *Dictyostelium discoideum*. Science (1979) *330*, 1533–1536. https://doi.org/10.1126/science.1197423.

14. BLASKOVICS, J.C., and RAPER, K.B. (1957). ENCYSTMENT STAGES OF DICTYOSTELIUM. Biol Bull *113*, 58–88. https://doi.org/10.2307/1538802.

15. WALLACE, M.A., and RAPER, K.B. (1979). Genetic Exchanges in the Macrocysts of Dictyostelium discoideum. J Gen Microbiol *113*, 327–337. https://doi.org/10.1099/00221287-113-2-327.

16. Loftus, B., Anderson, I., Davies, R., Alsmark, U.C.M., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R.P., Mann, B.J., et al. (2005). The genome of the protist parasite Entamoeba histolytica. Nature *433*, 865–868. https://doi.org/10.1038/nature03291.

17. Tanifuji, G., Cenci, U., Moog, D., Dean, S., Nakayama, T., David, V., Fiala, I., Curtis, B.A., Sibbald, S.J., Onodera, N.T., et al. (2017). Genome sequencing reveals metabolic and cellular interdependence in an amoeba-kinetoplastid symbiosis. Sci Rep *7*, 11688. https://doi.org/10.1038/s41598-017-11866-x.

18. Žárský, V., Klimeš, V., Pačes, J., Vlček, Č., Hradilová, M., Beneš, V., Nývltová, E., Hrdý, I., Pyrih, J., Mach, J., et al. (2021). The *Mastigamoeba balamuthi* Genome and the Nature of the Free-Living Ancestor of *Entamoeba*. Mol Biol Evol *38*, 2240–2259. https://doi.org/10.1093/molbev/msab020.

19. De Jonckheere, J.F. (1991). Ecology of Acanthamoeba. Clinical Infectious Diseases *13*, S385–S387. https://doi.org/10.1093/clind/13.Supplement_5.S385.

20. Clarke, M., Lohan, A.J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T.R., et al. (2013). Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol *14*, R11. https://doi.org/10.1186/gb-2013-14-2-r11.

21. Lorenzo-Morales, J., Khan, N.A., and Walochnik, J. (2015). An update on *Acanthamoeba* keratitis: diagnosis, pathogenesis and treatment. Parasite *22*, 10. https://doi.org/10.1051/parasite/2015010.

22. Leger, M.M., Gawryluk, R.M.R., Gray, M.W., and Roger, A.J. (2013). Evidence for a Hydrogenosomal-Type Anaerobic ATP Generation Pathway in Acanthamoeba castellanii. PLoS One *8*, e69532. https://doi.org/10.1371/journal.pone.0069532.

23. Hall, J., and Voelz, H. (1985). Bacterial Endosymbionts of Acanthamoeba sp. J Parasitol *71*, 89. https://doi.org/10.2307/3281982.

24. HORN, M., and WAGNER, M. (2004). Bacterial Endosymbionts of Free-living Amoebae [1]. Journal of Eukaryotic Microbiology *51*, 509–514. https://doi.org/10.1111/j.1550-7408.2004.tb00278.x.

25. Schmitz-Esser, S., Toenshoff, E.R., Haider, S., Heinz, E., Hoenninger, V.M., Wagner, M., and Horn, M. (2008). Diversity of Bacterial Endosymbionts of Environmental *Acanthamoeba* Isolates. Appl Environ Microbiol *74*, 5822–5831. https://doi.org/10.1128/AEM.01093-08.

26. Iovieno, A., Ledee, D.R., Miller, D., and Alfonso, E.C. (2010). Detection of Bacterial

Endosymbionts in Clinical Acanthamoeba Isolates. Ophthalmology *117*, 445-452.e3. https://doi.org/10.1016/j.ophtha.2009.08.033.

27. Neff, R.J. (1957). Purification, Axenic Cultivation, and Description of a Soil Amoeba, Acanthamoeba sp. J Protozool *4*, 176–182. https://doi.org/10.1111/j.1550-7408.1957.tb02505.x.

28. PAGE, F.C. (1967). Re-Definition of the Genus *Acanthamoeba* with Descriptions of Three Species. J Protozool *14*, 709–724. https://doi.org/10.1111/j.1550-7408.1967.tb02066.x.

29. Meyers, L.A., and Levin, D.A. (2006). ON THE ABUNDANCE OF POLYPLOIDS IN FLOWERING PLANTS. Evolution (N Y) *60*, 1198–1206. https://doi.org/10.1111/j.0014-3820.2006.tb01198.x.

30. Rieseberg, L.H., and Willis, J.H. (2007). Plant Speciation. Science (1979) *317*, 910–914. https://doi.org/10.1126/science.1137729.

31. Otto, S.P. (2007). The Evolutionary Consequences of Polyploidy. Cell *131*, 452–462. https://doi.org/10.1016/j.cell.2007.10.022.

32. BRETAGNOLLE, F., and THOMPSON, J.D. (1995). Gametes with the somatic chromosome number: mechanisms of their formation and role in the evolution of autopolyploid plants. New Phytologist *129*, 1–22. https://doi.org/10.1111/j.1469-8137.1995.tb03005.x.

33. Ramsey, J., and Schemske, D.W. (1998). PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. Annu Rev Ecol Syst *29*, 467–501. https://doi.org/10.1146/annurev.ecolsys.29.1.467.

34. Otto, S.P., and Whitton, J. (2000). POLYPLOID INCIDENCE AND EVOLUTION. Annu Rev Genet *34*, 401–437. https://doi.org/10.1146/annurev.genet.34.1.401.

35. Leggatt, R.A., and Iwama, G.K. (2003). Occurrence of polyploidy in the fishes. Rev Fish Biol Fish *13*, 237–246. https://doi.org/10.1023/B:RFBF.0000033049.00668.fe.

36. Ohno, S., Muramoto, J., Christian, L., and Atkin, N.B. (1967). Diploid-tetraploid relationship among old-world members of the fish family Cyprinidae. Chromosoma *23*, 1–9. https://doi.org/10.1007/BF00293307.

37. Drauch Schreier, A., Gille, D., Mahardja, B., and May, B. (2011). Neutral markers confirm the octoploid origin and reveal spontaneous autopolyploidy in white sturgeon, Acipenser transmontanus. Journal of Applied Ichthyology *27*, 24–33. https://doi.org/10.1111/j.1439-0426.2011.01873.x.

38. Cannatella, D.C., and de Sa, R.O. (1993). Xenopus Laevis as a Model Organism. Syst Biol *42*, 476–507. https://doi.org/10.1093/sysbio/42.4.476.

39. Wang, M.-J., Chen, F., Lau, J.T.Y., and Hu, Y.-P. (2017). Hepatocyte polyploidization and its association with pathophysiological processes. Cell Death Dis *8*, e2805–e2805. https://doi.org/10.1038/cddis.2017.167.

40. Albertin, W., and Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication. Proceedings of the Royal Society B: Biological Sciences *279*, 2497–2509. https://doi.org/10.1098/rspb.2012.0434.

41. Todd, R.T., Forche, A., and Selmecki, A. (2017). Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. Microbiol Spectr *5*. https://doi.org/10.1128/microbiolspec.FUNK-0051-2016.

42. Phillips, N., Kapraun, D.F., Gómez Garreta, A., Ribera Siguan, M.A., Rull, J.L., Salvador Soler, N., Lewis, R., and Kawai, H. (2011). Estimates of nuclear DNA content in 98 species of brown algae (Phaeophyta). AoB Plants *2011*. https://doi.org/10.1093/aobpla/plr001.

43. Coyer, J.A., Hoarau, G., Pearson, G.A., Serrão, E.A., Stam, W.T., and Olsen, J.L. (2006). Convergent adaptation to a marginal habitat by homoploid hybrids and polyploid ecads in the seaweed genus *Fucus*. Biol Lett *2*, 405–408. https://doi.org/10.1098/rsbl.2006.0489.

44. Varela-Álvarez, E., Loureiro, J., Paulino, C., and Serrão, E.A. (2018). Polyploid lineages in the genus Porphyra. Sci Rep *8*, 8696. https://doi.org/10.1038/s41598-018-26796-5.

45. Chepurnov, V.A., Mann, D.G., Vyverman, W., Sabbe, K., and Danielidis, D.B. (2002). SEXUAL REPRODUCTION, MATING SYSTEM, AND PROTOPLAST DYNAMICS OF *SEMINAVIS* (BACILLARIOPHYCEAE) [1]. J Phycol *38*, 1004–1019. https://doi.org/10.1046/j.1529-8817.2002.t01-1-01233.x.

46. Ioos, R., Andrieux, A., Marçais, B., and Frey, P. (2006). Genetic characterization of the natural hybrid species Phytophthora alni as inferred from nuclear and mitochondrial DNA analyses. Fungal Genetics and Biology *43*, 511–529. https://doi.org/10.1016/j.fgb.2006.02.006.

47. Parfrey, L.W., Lahr, D.J.G., and Katz, L.A. (2008). The Dynamic Nature of Eukaryotic Genomes. Mol Biol Evol *25*, 787–794. https://doi.org/10.1093/molbev/msn032.

48. Goodkov, A. V., Berdieva, M.A., Podlipaeva, Y.I., and Demin, S.Yu. (2020). The Chromatin Extrusion Phenomenon in *Amoeba proteus* Cell Cycle. Journal of Eukaryotic Microbiology *67*, 203–208. https://doi.org/10.1111/jeu.12771.

49. Goetz, E.J., Greco, M., Rappaport, H.B., Weiner, A.K.M., Walker, L.M., Bowser, S., Goldstein, S., and Katz, L.A. (2022). Foraminifera as a model of the extensive variability in genome dynamics among eukaryotes. BioEssays *44*. https://doi.org/10.1002/bies.202100267.

50. Timmons, C., Le, K., Rappaport, H.B., Sterner, E.G., Maurer-Alcalá, X.X., Goldstein, S.T., and Katz, L.A. (2024). Foraminifera as a model of eukaryotic genome dynamism. mBio *15*. https://doi.org/10.1128/mbio.03379-23.

51. Mungroo, M.R., Siddiqui, R., and Khan, N.A. (2021). War of the microbial world: Acanthamoeba spp. interactions with microorganisms. Folia Microbiol (Praha) *66*, 689–699. https://doi.org/10.1007/s12223-021-00889-7.

52. Proca-Ciobanu, M., Lupascu, Gh., Petrovici, Al., and Ionescu, M.D. (1975). Electron

microscopic study of a pathogenic Acanthamoeba castellani strain: The presence of bacterial endosymbionts. Int J Parasitol *5*, 49–56. https://doi.org/10.1016/0020-7519(75)90097-1.

53. Hall, J., and Voelz, H. (1985). Bacterial Endosymbionts of Acanthamoeba sp. J Parasitol *71*, 89. https://doi.org/10.2307/3281982.

54. HORN, M., and WAGNER, M. (2004). Bacterial Endosymbionts of Free-living Amoebae [1]. Journal of Eukaryotic Microbiology *51*, 509–514. https://doi.org/10.1111/j.1550-7408.2004.tb00278.x.

55. Schmitz-Esser, S., Toenshoff, E.R., Haider, S., Heinz, E., Hoenninger, V.M., Wagner, M., and Horn, M. (2008). Diversity of Bacterial Endosymbionts of Environmental *Acanthamoeba* Isolates. Appl Environ Microbiol *74*, 5822–5831. https://doi.org/10.1128/AEM.01093-08.

56. Greub, G., and Raoult, D. (2004). Microorganisms Resistant to Free-Living Amoebae. Clin Microbiol Rev *17*, 413–433. https://doi.org/10.1128/CMR.17.2.413-433.2004.

57. Molmeret, M., Horn, M., Wagner, M., Santic, M., and Abu Kwaik, Y. (2005). Amoebae as Training Grounds for Intracellular Bacterial Pathogens. Appl Environ Microbiol *71*, 20–28. https://doi.org/10.1128/AEM.71.1.20-28.2005.

58. Thomas, V., McDonnell, G., Denyer, S.P., and Maillard, J.-Y. (2010). Free-living amoebae and their intracellular pathogenic microorganisms: risks for water quality. FEMS Microbiol Rev *34*, 231–259. https://doi.org/10.1111/j.1574-6976.2009.00190.x.

59. Heinz, E., Kolarov, I., Kästner, C., Toenshoff, E.R., Wagner, M., and Horn, M. (2007). An *Acanthamoeba* sp. containing two phylogenetically different bacterial endosymbionts. Environ Microbiol *9*, 1604–1609. https://doi.org/10.1111/j.1462-2920.2007.01268.x.

60. Müller, A., Walochnik, J., Wagner, M., and Schmitz-Esser, S. (2016). A clinical Acanthamoeba isolate harboring two distinct bacterial endosymbionts. Eur J Protistol *56*, 21–25. https://doi.org/10.1016/j.ejop.2016.04.002.

61. Wang, Z., and Wu, M. (2017). Comparative Genomic Analysis of Acanthamoeba Endosymbionts Highlights the Role of Amoebae as a “Melting Pot” Shaping the Rickettsiales Evolution. Genome Biol Evol *9*, 3214–3224. https://doi.org/10.1093/gbe/evx246.

62. Scola, B. La, Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.-M., and Raoult, D. (2003). A Giant Virus in Amoebae. Science (1979) *299*, 2033–2033. https://doi.org/10.1126/science.1081867.

63. Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J.-M. (2004). The 1.2-Megabase Genome Sequence of Mimivirus. Science (1979) *306*, 1344–1350. https://doi.org/10.1126/science.1101485.

64. Colson, P., De Lamballerie, X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D.K., Cheng, X.-W., Federici, B.A., Van Etten, J.L., Koonin, E. V., et al. (2013). “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol *158*,

2517–2521. https://doi.org/10.1007/s00705-013-1768-6.

65. Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., et al. (2013). Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. Science (1979) *341*, 281–286. https://doi.org/10.1126/science.1239181.

66. Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., et al. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proceedings of the National Academy of Sciences *111*, 4274–4279. https://doi.org/10.1073/pnas.1320670111.

67. Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., Alempic, J.-M., Ramus, C., Bruley, C., Labadie, K., et al. (2015). In-depth study of *Mollivirus sibericum* , a new 30,000-y-old giant virus infecting *Acanthamoeba*. Proceedings of the National Academy of Sciences *112*. https://doi.org/10.1073/pnas.1510795112.

68. Maumus, F., and Blanc, G. (2016). Study of Gene Trafficking between *Acanthamoeba* and Giant Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses. Genome Biol Evol *8*, 3351–3363. https://doi.org/10.1093/gbe/evw260.

69. Gallot-Lavallée, L., and Blanc, G. (2017). A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. Viruses *9*, 17. https://doi.org/10.3390/v9010017.

70. Filée, J. (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: The visible part of the iceberg? Virology *466–467*, 53–59. https://doi.org/10.1016/j.virol.2014.06.004.

71. La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., et al. (2008). The virophage as a unique parasite of the giant mimivirus. Nature *455*, 100–104. https://doi.org/10.1038/nature07218.

72. Diesend, J., Kruse, J., Hagedorn, M., and Hammann, C. (2018). Amoebae, Giant Viruses, and Virophages Make Up a Complex, Multilayered Threesome. Front Cell Infect Microbiol *7*. https://doi.org/10.3389/fcimb.2017.00527.

73. Keeling, P.J. (2024). Horizontal gene transfer in eukaryotes: aligning theory with data. Nat Rev Genet *25*, 416–430. https://doi.org/10.1038/s41576-023-00688-5.

74. Leger, M.M., Eme, L., Stairs, C.W., and Roger, A.J. (2018). Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). BioEssays *40*. https://doi.org/10.1002/bies.201700242.

75. Van Etten, J., and Bhattacharya, D. (2020). Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends in Genetics *36*, 915–925. https://doi.org/10.1016/j.tig.2020.08.006.

76. Stairs, C.W., Roger, A.J., and Hampl, V. (2011). Eukaryotic Pyruvate Formate Lyase and Its Activating Enzyme Were Acquired Laterally from a Firmicute. Mol Biol Evol *28*,

2087–2099. https://doi.org/10.1093/molbev/msr032.

77. Stairs, C.W., Eme, L., Brown, M.W., Mutsaers, C., Susko, E., Dellaire, G., Soanes, D.M., van der Giezen, M., and Roger, A.J. (2014). A SUF Fe-S Cluster Biogenesis System in the Mitochondrion-Related Organelles of the Anaerobic Protist Pygsuia. Current Biology *24*, 1176–1186. https://doi.org/10.1016/j.cub.2014.04.033.

78. Nývltová, E., Stairs, C.W., Hrdý, I., Rídl, J., Mach, J., Pačes, J., Roger, A.J., and Tachezy, J. (2015). Lateral Gene Transfer and Gene Duplication Played a Key Role in the Evolution of Mastigamoeba balamuthi Hydrogenosomes. Mol Biol Evol *32*, 1039–1055. https://doi.org/10.1093/molbev/msu408.

79. Stairs, C.W., Leger, M.M., and Roger, A.J. (2015). Diversity and origins of anaerobic metabolism in mitochondria and related organelles. Philosophical Transactions of the Royal Society B: Biological Sciences *370*, 20140326. https://doi.org/10.1098/rstb.2014.0326.

80. Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S.C., Petrželková, R., Eme, L., Novák, L., Žárský, V., Barlow, L.D., Herman, E.K., et al. (2016). A Eukaryote without a Mitochondrial Organelle. Current Biology *26*, 1274–1284. https://doi.org/10.1016/j.cub.2016.03.053.

81. Leger, M.M., Kolisko, M., Kamikawa, R., Stairs, C.W., Kume, K., Čepička, I., Silberman, J.D., Andersson, J.O., Xu, F., Yabuki, A., et al. (2017). Organelles that illuminate the origins of Trichomonas hydrogenosomes and Giardia mitosomes. Nat Ecol Evol *1*, 0092. https://doi.org/10.1038/s41559-017-0092.

82. Stairs, C.W., Eme, L., Muñoz-Gómez, S.A., Cohen, A., Dellaire, G., Shepherd, J.N., Fawcett, J.P., and Roger, A.J. (2018). Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. Elife *7*. https://doi.org/10.7554/eLife.34292.

83. Vacek, V., Novák, L.V.F., Treitli, S.C., Táborský, P., Čepička, I., Kolísko, M., Keeling, P.J., and Hampl, V. (2018). Fe–S Cluster Assembly in Oxymonads and Related Protists. Mol Biol Evol. https://doi.org/10.1093/molbev/msy168.

84. Gawryluk, R.M.R., and Stairs, C.W. (2021). Diversity of electron transport chains in anaerobic protists. Biochimica et Biophysica Acta (BBA) - Bioenergetics *1862*, 148334. https://doi.org/10.1016/j.bbabio.2020.148334.

85. Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M., and Fukatsu, T. (2008). *Wolbachia* genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes. Genome Res *18*, 272–280. https://doi.org/10.1101/gr.7144908.

86. Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A.C.C., et al. (2013). Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis. Cell *153*, 1567–1578. https://doi.org/10.1016/j.cell.2013.05.040.

87. Gilbert, C., and Maumus, F. (2023). Sidestepping Darwin: horizontal gene transfer from

plants to insects. Curr Opin Insect Sci *57*, 101035. https://doi.org/10.1016/j.cois.2023.101035.

88. Dean, P., Sendra, K.M., Williams, T.A., Watson, A.K., Major, P., Nakjang, S., Kozhevnikova, E., Goldberg, A. V., Kunji, E.R.S., Hirt, R.P., et al. (2018). Transporter gene acquisition and innovation in the evolution of Microsporidia intracellular parasites. Nat Commun *9*, 1709. https://doi.org/10.1038/s41467-018-03923-4.

89. Milner, D.S., Attah, V., Cook, E., Maguire, F., Savory, F.R., Morrison, M., Müller, C.A., Foster, P.G., Talbot, N.J., Leonard, G., et al. (2019). Environment-dependent fitness gains can be driven by horizontal gene transfer of transporter-encoding genes. Proceedings of the National Academy of Sciences *116*, 5613–5622. https://doi.org/10.1073/pnas.1815994116.

90. Richards, T.A., Soanes, D.M., Jones, M.D.M., Vasieva, O., Leonard, G., Paszkiewicz, K., Foster, P.G., Hall, N., and Talbot, N.J. (2011). Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proceedings of the National Academy of Sciences *108*, 15258–15263. https://doi.org/10.1073/pnas.1105100108.

91. Savory, F., Leonard, G., and Richards, T.A. (2015). The Role of Horizontal Gene Transfer in the Evolution of the Oomycetes. PLoS Pathog *11*, e1004805. https://doi.org/10.1371/journal.ppat.1004805.

92. Archibald, J.M., Rogers, M.B., Toop, M., Ishida, K., and Keeling, P.J. (2003). Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proceedings of the National Academy of Sciences *100*, 7678–7683. https://doi.org/10.1073/pnas.1230951100.

93. Suzuki, K., and Miyagishima, S. -y. (2010). Eukaryotic and Eubacterial Contributions to the Establishment of Plastid Proteome Estimated by Large-Scale Phylogenetic Analyses. Mol Biol Evol *27*, 581–590. https://doi.org/10.1093/molbev/msp273.

94. Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., et al. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature *492*, 59–65. https://doi.org/10.1038/nature11681.

95. Qiu, H., Price, D.C., Weber, A.P.M., Facchinelli, F., Yoon, H.S., and Bhattacharya, D. (2013). Assessing the bacterial contribution to the plastid proteome. Trends Plant Sci *18*, 680–687. https://doi.org/10.1016/j.tplants.2013.09.007.

96. Nowack, E.C.M., Price, D.C., Bhattacharya, D., Singer, A., Melkonian, M., and Grossman, A.R. (2016). Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. Proceedings of the National Academy of Sciences *113*, 12214–12219. https://doi.org/10.1073/pnas.1608016113.

97. Singer, A., Poschmann, G., Mühlich, C., Valadez-Cano, C., Hänsch, S., Hüren, V., Rensing, S.A., Stühler, K., and Nowack, E.C.M. (2017). Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba Paulinella chromatophora. Current Biology *27*, 2763-2773.e5.

https://doi.org/10.1016/j.cub.2017.08.010.

98. Husnik, F., and McCutcheon, J.P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol *16*, 67–79. https://doi.org/10.1038/nrmicro.2017.137.

99. Sibbald, S.J., Eme, L., Archibald, J.M., and Roger, A.J. (2020). Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. Trends Parasitol *36*, 927–941. https://doi.org/10.1016/j.pt.2020.07.014.

100. International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945. https://doi.org/10.1038/nature03001.

101. Levy, S.E., and Boone, B.E. (2019). Next-Generation Sequencing Strategies. Cold Spring Harb Perspect Med *9*, a025791. https://doi.org/10.1101/cshperspect.a025791.

102. Giani, A.M., Gallo, G.R., Gianfranceschi, L., and Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J *18*, 9–19. https://doi.org/10.1016/j.csbj.2019.11.002.

103. Slatko, B.E., Gardner, A.F., and Ausubel, F.M. (2018). Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol *122*. https://doi.org/10.1002/cpmb.59.

104. Athanasopoulou, K., Boti, M.A., Adamopoulos, P.G., Skourou, P.C., and Scorilas, A. (2021). Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life *12*, 30. https://doi.org/10.3390/life12010030.

105. Matthey-Doret, C., Colp, M.J., Escoll, P., Thierry, A., Moreau, P., Curtis, B., Sahr, T., Sarrasin, M., Gray, M.W., Lang, B.F., et al. (2022). Chromosome-scale assemblies of Acanthamoeba castellanii genomes provide insights into Legionella pneumophila infection-related chromatin reorganization. Genome Res *32*, 1698–1710. https://doi.org/10.1101/gr.276375.121.

106. Rimm, D.L., Pollard, T.D., and Hieter, P. (1988). Resolution of Acanthamoeba castellanii chromosomes by pulsed field gel electrophoresis and construction of the initial linkage map. Chromosoma *97*, 219–223. https://doi.org/10.1007/BF00292964.

107. Gladyshev, E.A., Meselson, M., and Arkhipova, I.R. (2008). Massive Horizontal Gene Transfer in Bdelloid Rotifers. Science (1979) *320*, 1210–1213. https://doi.org/10.1126/science.1156407.

108. Hill, T., and Betancourt, A.J. (2018). Extensive exchange of transposable elements in the Drosophila pseudoobscura group. Mob DNA *9*, 20. https://doi.org/10.1186/s13100-018-0123-6.

109. El Baidouri, M., Carpentier, M.-C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S.A., and Panaud, O. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res *24*, 831–838. https://doi.org/10.1101/gr.164400.113.

110. Peccoud, J., Loiseau, V., Cordaux, R., and Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. Proceedings of the National Academy of Sciences *114*, 4721–4726. https://doi.org/10.1073/pnas.1621178114.

111. Reiss, D., Mialdea, G., Miele, V., de Vienne, D.M., Peccoud, J., Gilbert, C., Duret, L., and Charlat, S. (2019). Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLoS Genet *15*, e1007965. https://doi.org/10.1371/journal.pgen.1007965.

112. Ivancevic, A.M., Kortschak, R.D., Bertozzi, T., and Adelson, D.L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol *19*, 85. https://doi.org/10.1186/s13059-018-1456-7.

113. Paganini, J., Campan-Fournier, A., Da Rocha, M., Gouret, P., Pontarotti, P., Wajnberg, E., Abad, P., and Danchin, E.G.J. (2012). Contribution of Lateral Gene Transfers to the Genome Composition and Parasitic Ability of Root-Knot Nematodes. PLoS One *7*, e50875. https://doi.org/10.1371/journal.pone.0050875.

114. McDonald, M.C., Taranto, A.P., Hill, E., Schwessinger, B., Liu, Z., Simpfendorfer, S., Milgate, A., and Solomon, P.S. (2019). Transposon-Mediated Horizontal Transfer of the Host-Specific Virulence Protein ToxA between Three Fungal Wheat Pathogens. mBio *10*. https://doi.org/10.1128/mBio.01515-19.

115. Sibbald, S.J., Eme, L., Archibald, J.M., and Roger, A.J. (2020). Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. Trends Parasitol *36*, 927–941. https://doi.org/10.1016/j.pt.2020.07.014.

116. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat Genet *31*, 241–247. https://doi.org/10.1038/ng917.

117. Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas, M.A., Haussler, D., and Jacob, H.J. (2004). Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. Genome Res *14*, 528–538. https://doi.org/10.1101/gr.1970304.

118. Serre, D., Nadon, R., and Hudson, T.J. (2005). Large-scale recombination rate patterns are conserved among human populations. Genome Res *15*, 1547–1552. https://doi.org/10.1101/gr.4211905.

119. Barton, A.B., Pekosz, M.R., Kurvathi, R.S., and Kaback, D.B. (2008). Meiotic Recombination at the Ends of Chromosomes in *Saccharomyces cerevisiae*. Genetics *179*, 1221–1235. https://doi.org/10.1534/genetics.107.083493.

120. Groenen, M.A.M., Wahlberg, P., Foglio, M., Cheng, H.H., Megens, H.-J., Crooijmans, R.P.M.A., Besnier, F., Lathrop, M., Muir, W.M., Wong, G.K.-S., et al. (2009). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res *19*, 510–519. https://doi.org/10.1101/gr.086538.108.

121. Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., Webster, M.T., Öst, T., Schneider, M., Kempenaers, B., et al. (2010). The recombination landscape of the zebra finch *Taeniopygia guttata* genome. Genome Res *20*, 485–495. https://doi.org/10.1101/gr.101410.109.

122. Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E.M., Bowden, R., Aneas, I., Broxholme, J., et al. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science (1979) *336*, 193–198. https://doi.org/10.1126/science.1216872.

123. van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. Journal of Visualized Experiments. https://doi.org/10.3791/1869.

124. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science (1979) *295*, 1306–1311. https://doi.org/10.1126/science.1067799.

125. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (1979) *326*, 289–293. https://doi.org/10.1126/science.1181369.

126. Belaghzal, H., Dekker, J., and Gibcus, J.H. (2017). Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods *123*, 56–65. https://doi.org/10.1016/j.ymeth.2017.04.004.

127. Lafontaine, D.L., Yang, L., Dekker, J., and Gibcus, J.H. (2021). Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture. Curr Protoc *1*. https://doi.org/10.1002/cpz1.198.

128. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol *37*, 540–546. https://doi.org/10.1038/s41587-019-0072-8.

129. Kundu, R., Casey, J., and Sung, W.-K. (2019). HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. bioRxiv, 2019.12.19.882506. https://doi.org/10.1101/2019.12.19.882506.

130. Baudry, L., Guiglielmoni, N., Marie-Nelly, H., Cormier, A., Marbouty, M., Avia, K., Mie, Y.L., Godfroy, O., Sterck, L., Cock, J.M., et al. (2020). instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. Genome Biol *21*, 148. https://doi.org/10.1186/s13059-020-02041-z.

131. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One *9*, e112963. https://doi.org/10.1371/journal.pone.0112963.

132. Song, L., and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. Gigascience *4*, 48. https://doi.org/10.1186/s13742-015-0089-y.

133. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

134. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

135. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Res *19*, 1639–1645. https://doi.org/10.1101/gr.092759.109.

136. Gray, M.W., Burger, G., Derelle, R., Klimeš, V., Leger, M.M., Sarrasin, M., Vlček, Č., Roger, A.J., Eliáš, M., and Lang, B.F. (2020). The draft nuclear genome sequence and predicted mitochondrial proteome of Andalucia godoyi, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC Biol *18*, 22. https://doi.org/10.1186/s12915-020-0741-6.

137. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

138. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644–652. https://doi.org/10.1038/nbt.1883.

139. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res *31*, 5654–5666. https://doi.org/10.1093/nar/gkg770.

140. Gotoh, O. (2008). A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res *36*, 2630–2638. https://doi.org/10.1093/nar/gkn105.

141. Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics *7*, 62. https://doi.org/10.1186/1471-2105-7-62.

142. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59. https://doi.org/10.1186/1471-2105-5-59.

143. Lomsadze, A., Burns, P.D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res *42*, e119–e119. https://doi.org/10.1093/nar/gku557.

144. Testa, A.C., Hane, J.K., Ellwood, S.R., and Oliver, R.P. (2015). CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics *16*, 170. https://doi.org/10.1186/s12864-015-1344-4.

145. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell,

C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

146. Smit, A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0. Preprint at http://www.repeatmasker.org.

147. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res *33*, W116–W120. https://doi.org/10.1093/nar/gki442.

148. Kall, L., Krogh, A., and Sonnhammer, E.L.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res *35*, W429–W432. https://doi.org/10.1093/nar/gkm256.

149. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol *38*, 5825–5829. https://doi.org/10.1093/molbev/msab293.

150. Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160.

151. Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res *39*, W29–W37. https://doi.org/10.1093/nar/gkr367.

152. Tice, A.K., Žihala, D., Pánek, T., Jones, R.E., Salomaki, E.D., Nenarokov, S., Burki, F., Eliáš, M., Eme, L., Roger, A.J., et al. (2021). PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. PLoS Biol *19*, e3001365. https://doi.org/10.1371/journal.pbio.3001365.

153. Derelle, R., Philippe, H., and Colbourne, J.K. (2020). Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. Mol Biol Evol *37*, 3389–3396. https://doi.org/10.1093/molbev/msaa159.

154. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol *20*, 238. https://doi.org/10.1186/s13059-019-1832-y.

155. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

156. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

157. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

158. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., and Mozziconacci, J. (2012). Normalization of a chromosomal contact map. BMC Genomics *13*, 436. https://doi.org/10.1186/1471-2164-13-436.

159. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods *9*, 999–1003. https://doi.org/10.1038/nmeth.2148.

160. Matthey-Doret, C., Baudry, L., Breuer, A., Montagne, R., Guiglielmoni, N., Scolari, V., Jean, E., Campeas, A., Chanut, P.H., Oriol, E., et al. (2020). Computer vision for pattern detection in chromosome contact maps. Nat Commun *11*, 5795. https://doi.org/10.1038/s41467-020-19562-7.

161. Yang, T., Zhang, F., Yardımcı, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., and Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res *27*, 1939–1949. https://doi.org/10.1101/gr.220640.117.

162. Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D.P., Syan, S., Guillén, N., Margeot, A., Zimmer, C., et al. (2014). High-quality genome (re)assembly using chromosomal contact data. Nat Commun *5*, 5695. https://doi.org/10.1038/ncomms6695.

163. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

164. Kawano-Sugaya, T., Izumiyama, S., Yanagawa, Y., Saito-Nakano, Y., Watanabe, K., Kobayashi, S., Nakada-Tsukui, K., and Nozaki, T. (2020). Near-chromosome level genome assembly reveals ploidy diversity and plasticity in the intestinal protozoan parasite Entamoeba histolytica. BMC Genomics *21*, 813. https://doi.org/10.1186/s12864-020-07167-9.

165. Girard, F., Even, A., Thierry, A., Ruault, M., Meneu, L., Adiba, S., Taddei, A., Koszul, R., and Cournac, A. (2023). Anchoring of parasitic plasmids to inactive regions of eukaryotic chromosomes through nucleosome signal. bioRxiv, 2023.10.04.558402. https://doi.org/10.1101/2023.10.04.558402.

166. Keeler, E.L., Merenstein, C., Reddy, S., Taylor, L.J., Cobián-Güemes, A.G., Zankharia, U., Collman, R.G., and Bushman, F.D. (2023). Widespread, human-associated redondoviruses infect the commensal protozoan Entamoeba gingivalis. Cell Host Microbe *31*, 58-68.e5. https://doi.org/10.1016/j.chom.2022.11.002.

167. Zhang, P., Peng, H., Llauro, C., Bucher, E., and Mirouze, M. (2021). ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. Front Plant Sci *12*. https://doi.org/10.3389/fpls.2021.743742.

168. Sucgang, R., Chen, G., Liu, W., Lindsay, R., Lu, J., Muzny, D., Shaulsky, G., Loomis, W., Gibbs, R., and Kuspa, A. (2003). Sequence and structure of the extrachromosomal

palindrome encoding the ribosomal RNA genes in Dictyostelium. Nucleic Acids Res *31*, 2361–2368. https://doi.org/10.1093/nar/gkg348.

169. Costantino, L., Hsieh, T.-H.S., Lamothe, R., Darzacq, X., and Koshland, D. (2020). Cohesin residency determines chromatin loop patterns. Elife *9*. https://doi.org/10.7554/eLife.59889.

170. Dauban, L., Montagne, R., Thierry, A., Lazar-Stefanita, L., Bastié, N., Gadal, O., Cournac, A., Koszul, R., and Beckouët, F. (2020). Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners. Mol Cell *77*, 1279-1293.e4. https://doi.org/10.1016/j.molcel.2020.01.019.

171. Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108–119. https://doi.org/10.1016/j.cell.2015.05.048.

172. Cockram, C., Thierry, A., Gorlas, A., Lestini, R., and Koszul, R. (2021). Euryarchaeal genomes are folded into SMC-dependent loops and domains, but lack transcription-mediated compartmentalization. Mol Cell *81*, 459-472.e10. https://doi.org/10.1016/j.molcel.2020.12.013.

173. Alexa, A., and Rahnenfuhrer, J. (2022). topGO: Enrichment Analysis for Gene Ontology. Preprint.

174. Declerck, P., Behets, J., De Keersmaecker, B., and Ollevier, F. (2007). Receptor-mediated uptake of Legionella pneumophila by Acanthamoeba castellanii and Naegleria lovaniensis. J Appl Microbiol *103*, 2697–2703. https://doi.org/10.1111/j.1365-2672.2007.03530.x.

175. Garate, M., Cao, Z., Bateman, E., and Panjwani, N. (2004). Cloning and Characterization of a Novel Mannose-binding Protein of Acanthamoeba. Journal of Biological Chemistry *279*, 29849–29856. https://doi.org/10.1074/jbc.M402334200.

176. Garate, M., Cubillos, I., Marchant, J., and Panjwani, N. (2005). Biochemical Characterization and Functional Studies of *Acanthamoeba* Mannose-Binding Protein. Infect Immun *73*, 5775–5781. https://doi.org/10.1128/IAI.73.9.5775-5781.2005.

177. Yang, Q., Zwick, M.G., and Paule, M.R. (1994). Sequence organization of the *Acanthamoeba* rRNA intergenic spacer: identification of transcriptional enhancers. Nucleic Acids Res *22*, 4798–4805. https://doi.org/10.1093/nar/22.22.4798.

178. GICQUAUD, C., and TREMBLAY, N. (1991). Observations with Hoechst Staining of Amitosis in *Acanthamoeba castellanii*. J Protozool *38*, 221–224. https://doi.org/10.1111/j.1550-7408.1991.tb04432.x.

179. Rabl C (1885). Uber zelltheilung. In Morphol. Jahrbuch., Gegenbauer C, ed. (Willhelm Engelmann), pp. 214–330.

180. Harb, O.S., Venkataraman, C., Haack, B.J., Gao, L.-Y., and Kwaik, Y.A. (1998). Heterogeneity in the Attachment and Uptake Mechanisms of the Legionnaires' Disease Bacterium, *Legionella pneumophila* , by Protozoan Hosts. Appl Environ Microbiol *64*, 126–132. https://doi.org/10.1128/AEM.64.1.126-132.1998.

181. Corsaro, D. (2022). Acanthamoeba Mannose and Laminin Binding Proteins Variation across Species and Genotypes. Microorganisms *10*, 2162. https://doi.org/10.3390/microorganisms10112162.

182. Wang, Y., Jiang, L., Zhao, Y., Ju, X., Wang, L., Jin, L., Fine, R.D., and Li, M. (2023). Biological characteristics and pathogenicity of Acanthamoeba. Front Microbiol *14*. https://doi.org/10.3389/fmicb.2023.1147077.

183. Domingo-Sananes, M.R., and McInerney, J.O. (2021). Mechanisms That Shape Microbial Pangenomes. Trends Microbiol *29*, 493–503. https://doi.org/10.1016/j.tim.2020.12.004.

184. Brockhurst, M.A., Harrison, E., Hall, J.P.J., Richards, T., McNally, A., and MacLean, C. (2019). The Ecology and Evolution of Pangenomes. Current Biology *29*, R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012.

185. McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have pangenomes. Nat Microbiol *2*, 17040. https://doi.org/10.1038/nmicrobiol.2017.40.

186. Gaikani, H.K., Stolar, M., Kriti, D., Nislow, C., and Giaever, G. (2024). From beer to breadboards: yeast as a force for biological innovation. Genome Biol *25*, 10. https://doi.org/10.1186/s13059-023-03156-9.

187. Fus-Kujawa, A., Prus, P., Bajdak-Rusinek, K., Teper, P., Gawron, K., Kowalczuk, A., and Sieron, A.L. (2021). An Overview of Methods and Tools for Transfection of Eukaryotic Cells in vitro. Front Bioeng Biotechnol *9*, 701031. https://doi.org/10.3389/fbioe.2021.701031.

188. Faktorová, D., Nisbet, R.E.R., Fernández Robledo, J.A., Casacuberta, E., Sudek, L., Allen, A.E., Ares, M., Aresté, C., Balestreri, C., Barbrook, A.C., et al. (2020). Genetic tool development in marine protists: emerging model organisms for experimental cell biology. Nat Methods *17*, 481–494. https://doi.org/10.1038/s41592-020-0796-x.

189. Peng, Z., Omaruddin, R., and Bateman, E. (2005). Stable transfection of Acanthamoeba castellanii. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *1743*, 93–100. https://doi.org/10.1016/j.bbamcr.2004.08.014.

190. Bateman, E. (2010). Expression plasmids and production of EGFP in stably transfected Acanthamoeba. Protein Expr Purif *70*, 95–100. https://doi.org/10.1016/j.pep.2009.10.008.

191. Mileshina, D., Koulintchenko, M., Konstantinov, Y., and Dietrich, A. (2011). Transfection of plant mitochondria and in organello gene integration. Nucleic Acids Res *39*, e115–e115. https://doi.org/10.1093/nar/gkr517.

192. Takano, M., Egawa, H., Ikeda, J.-E., and Wakasa, K. (1997). The structures of integration sites in transgenic rice. The Plant Journal *11*, 353–361. https://doi.org/https://doi.org/10.1046/j.1365-313X.1997.11030353.x.

193. Hadi, M.Z., McMullen, M.D., and Finer, J.J. (1996). Transformation of 12 different plasmids into soybean via particle bombardment. Plant Cell Rep *15*, 500–505. https://doi.org/10.1007/BF00232982.

194. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol *29*, 644–652. https://doi.org/10.1038/nbt.1883.

195. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

196. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

197. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

198. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26. https://doi.org/10.1038/nbt.1754.

199. Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178–192. https://doi.org/10.1093/bib/bbs017.

200. Yang, C., Lo, T., Nip, K.M., Hafezqorani, S., Warren, R.L., and Birol, I. (2023). Characterization and simulation of metagenomic nanopore sequencing data with Meta-NanoSim. Gigascience *12*. https://doi.org/10.1093/gigascience/giad013.

201. Martin Cerezo, M.L., Raval, R., de Haro Reyes, B., Kucka, M., Chan, F.Y., and Bryk, J. (2022). Identification and quantification of chimeric sequencing reads in a highly multiplexed <scp>RAD</scp> -seq protocol. Mol Ecol Resour *22*, 2860–2870. https://doi.org/10.1111/1755-0998.13661.

202. Woods, J.P., and Goldman, W.E. (1992). *In vivo* generation of linear plasmids with addition of telomeric sequences by *Histoplasma capsulatum*. Mol Microbiol *6*, 3603–3610. https://doi.org/10.1111/j.1365-2958.1992.tb01796.x.

203. Woods, J.P., and Goldman, W.E. (1993). Autonomous replication of foreign DNA in Histoplasma capsulatum: role of native telomeric sequences. J Bacteriol *175*, 636–641. https://doi.org/10.1128/jb.175.3.636-641.1993.

204. Varma, A., Edman, J.C., and Kwon-Chung, K.J. (1992). Molecular and genetic analysis of URA5 transformants of Cryptococcus neoformans. Infect Immun *60*, 1101–1108. https://doi.org/10.1128/iai.60.3.1101-1108.1992.

205. Edman, J.C. (1992). Isolation of Telomerelike Sequences from *Cryptococcus neoformans* and Their Use in High-Efficiency Transformation. Mol Cell Biol *12*, 2777–2783. https://doi.org/10.1128/mcb.12.6.2777-2783.1992.

206. Powell, W.A., and Kistler, H.C. (1990). In vivo rearrangement of foreign DNA by

Fusarium oxysporum produces linear self-replicating plasmids. J Bacteriol *172*, 3163–3171. https://doi.org/10.1128/jb.172.6.3163-3171.1990.

207. Gilley, D., Preer, J.R., Aufderheide, K.J., and Polisky, B. (1988). Autonomous Replication and Addition of Telomerelike Sequences to DNA Microinjected into *Paramecium tetraurelia* Macronuclei. Mol Cell Biol *8*, 4765–4772. https://doi.org/10.1128/mcb.8.11.4765-4772.1988.

208. San Millan, A., Peña-Miller, R., Toll-Riera, M., Halbert, Z. V, McLean, A.R., Cooper, B.S., and MacLean, R.C. (2014). Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat Commun *5*, 5208. https://doi.org/10.1038/ncomms6208.

209. Murugan, S., M, T.E., H, S.D., and R, C.D. (2011). Selection Pressure Required for Long-Term Persistence of bla CMY-2-Positive IncA/C Plasmids. Appl Environ Microbiol *77*, 4486–4493. https://doi.org/10.1128/AEM.02788-10.

210. Struhl, K., Stinchcomb, D.T., Scherer, S., and Davis, R.W. (1979). High-frequency transformation of yeast: autonomous replication of hybrid DNA molecules. Proceedings of the National Academy of Sciences *76*, 1035–1039. https://doi.org/10.1073/pnas.76.3.1035.

211. Cannon, R.D., Jenkinson, H.F., and Shepherd, M.G. (1990). Isolation and nucleotide sequence of an autonomously replicating sequence (ARS) element functional in Candida albicans and Saccharomyces cerevisiae. Mol Gen Genet *221*, 210–218. https://doi.org/10.1007/BF00261723.

212. Liachko, I., and Dunham, M.J. (2014). An autonomously replicating sequence for use in a wide range of budding yeasts. FEMS Yeast Res *14*, 364–367. https://doi.org/10.1111/1567-1364.12123.

213. Clyne, R.K., and Kelly, T.J. (1995). Genetic analysis of an ARS element from the fission yeast Schizosaccharomyces pombe. EMBO J *14*, 6348–6357. https://doi.org/10.1002/j.1460-2075.1995.tb00326.x.

214. Brown, M.W., Sharpe, S.C., Silberman, J.D., Heiss, A.A., Lang, B.F., Simpson, A.G.B., and Roger, A.J. (2013). Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proceedings of the Royal Society B: Biological Sciences *280*, 20131755. https://doi.org/10.1098/rspb.2013.1755.

215. Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., Lang, B.F., and Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. Proceedings of the National Academy of Sciences *112*. https://doi.org/10.1073/pnas.1420657112.

216. Gelvin, S.B. (2003). *Agrobacterium* -Mediated Plant Transformation: the Biology behind the "Gene-Jockeying" Tool. Microbiology and Molecular Biology Reviews *67*, 16–37. https://doi.org/10.1128/MMBR.67.1.16-37.2003.

217. Hirt, R.P., Alsmark, C., and Embley, T.M. (2015). Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. Curr Opin Microbiol *23*, 155–162. https://doi.org/10.1016/j.mib.2014.11.018.

218. Pussard, M. (1972). MORPHOLOGICAL COMPARISON OF 4 STRAINS OF ACANTHAMOEBA OF GROUP ASTRONYXIS-COMANDONI. JOURNAL OF PROTOZOOLOGY *19*, 557–563.

219. Matsunaga, S., Endo, T., Yagita, K., Hirukawa, Y., Tomino, S., Matsugo, S., and Tsuruhara, T. (1998). Chromosome Size Polymorphisms in the Genus Acanthamoeba Electrokaryotype by Pulsed-Field Gel Electrophoresis. Protist *149*, 323–340. https://doi.org/10.1016/S1434-4610(98)70039-2.

220. Augusto Corrêa dos Santos, R., Goldman, G.H., and Riaño-Pachón, D.M. (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. Bioinformatics *33*, 2575–2576. https://doi.org/10.1093/bioinformatics/btx204.

221. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. https://doi.org/10.1038/s41587-023-02024-y.

222. Maciver, S.K. (2016). Asexual Amoebae Escape Muller's Ratchet through Polyploidy. Trends Parasitol *32*, 855–862. https://doi.org/10.1016/j.pt.2016.08.006.

223. Makhlin, E.E., Kudryavtseva, M.V., and Kudryavtsev, B.N. (1979). Peculiarities of changes in DNA content of Amoeba proteus nuclei during interphase. Exp Cell Res *118*, 143–150. https://doi.org/10.1016/0014-4827(79)90592-5.

224. Afon'kin, S.J. (1986). Spontaneous "depolyploidization" of cells in Amoeba clones with increased nuclear DNA content. Arch Protistenkunde *131*, 101–112.

225. Goodkov, A. V., Berdieva, M.A., Podlipaeva, Y.I., and Demin, S.Yu. (2020). The Chromatin Extrusion Phenomenon in *Amoeba proteus* Cell Cycle. Journal of Eukaryotic Microbiology *67*, 203–208. https://doi.org/10.1111/jeu.12771.

226. Sterkers, Y., Lachaud, L., Crobu, L., Bastien, P., and Pagès, M. (2011). FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. Cell Microbiol *13*, 274–283. https://doi.org/10.1111/j.1462-5822.2010.01534.x.

227. Black, J.A., Reis-Cunha, J.L., Cruz, Angela.K., and Tosi, Luiz.R.O. (2023). Life in plastic, it's fantastic! How Leishmania exploit genome instability to shape gene expression. Front Cell Infect Microbiol *13*. https://doi.org/10.3389/fcimb.2023.1102462.

228. Dumetz, F., Imamura, H., Sanders, M., Seblova, V., Myskova, J., Pescher, P., Vanaerschot, M., Meehan, C.J., Cuypers, B., De Muylder, G., et al. (2017). Modulation of Aneuploidy in *Leishmania donovani* during Adaptation to Different *In Vitro* and *In Vivo* Environments and Its Impact on Gene Expression. mBio *8*. https://doi.org/10.1128/mBio.00599-17.

229. Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res *20*, 1313–1326. https://doi.org/10.1101/gr.101386.109.

230. Schlötterer, C. (2015). Genes from scratch--the evolutionary fate of de novo genes. Trends Genet *31*, 215–219. https://doi.org/10.1016/j.tig.2015.02.007.

231. Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. Nat Rev Genet *12*, 692–702. https://doi.org/10.1038/nrg3053.

232. Tatum, E.L., and Lederberg, J. (1947). Gene Recombination in the Bacterium *Escherichia coli*. J Bacteriol *53*, 673–684. https://doi.org/10.1128/jb.53.6.673-684.1947.

233. Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., et al. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature *399*, 323–329. https://doi.org/10.1038/20601.

234. Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304. https://doi.org/10.1038/35012500.

235. Treangen, T.J., and Rocha, E.P.C. (2011). Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLoS Genet *7*, e1001284. https://doi.org/10.1371/journal.pgen.1001284.

236. Vos, M., Hesselman, M.C., te Beek, T.A., van Passel, M.W.J., and Eyre-Walker, A. (2015). Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends Microbiol *23*, 598–605. https://doi.org/10.1016/j.tim.2015.07.006.

237. Stairs, C.W., Leger, M.M., and Roger, A.J. (2015). Diversity and origins of anaerobic metabolism in mitochondria and related organelles. Philos Trans R Soc Lond B Biol Sci *370*, 20140326. https://doi.org/10.1098/rstb.2014.0326.

238. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol *32*, 268–274. https://doi.org/10.1093/molbev/msu300.

239. Le, S.Q., Dang, C.C., and Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. Mol Biol Evol *29*, 2921–2936. https://doi.org/10.1093/molbev/mss112.

240. Susko, E., Lincker, L., and Roger, A.J. (2018). Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. Mol Biol Evol *35*, 1266–1283. https://doi.org/10.1093/molbev/msy026.

241. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

242. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods *18*, 366–368. https://doi.org/10.1038/s41592-021-01101-x.

243. Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the

functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol *12*, e1001889. https://doi.org/10.1371/journal.pbio.1001889.

244. Richter, D.J., Berney, C., Strassert, J.F.H., Poh, Y.-P., Herman, E.K., Muñoz-Gómez, S.A., Wideman, J.G., Burki, F., and de Vargas, C. (2022). EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. . Peer Community Journal *2*, e56.

245. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

246. Criscuolo, A., and Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol *10*, 210. https://doi.org/10.1186/1471-2148-10-210.

247. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.

248. Žárský, V., Klimeš, V., Pačes, J., Vlček, Č., Hradilová, M., Beneš, V., Nývltová, E., Hrdý, I., Pyrih, J., Mach, J., et al. (2021). The *Mastigamoeba balamuthi* Genome and the Nature of the Free-Living Ancestor of *Entamoeba*. Mol Biol Evol *38*, 2240–2259. https://doi.org/10.1093/molbev/msab020.

249. Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res *43*, D213–D221. https://doi.org/10.1093/nar/gku1243.

250. Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature *523*, 208–211. https://doi.org/10.1038/nature14486.

251. Mohd Hussain, R.H., Abdul Ghani, M.K., Khan, N.A., Siddiqui, R., and Anuar, T.S. (2022). *Acanthamoeba* species isolated from marine water in Malaysia exhibit distinct genotypes and variable physiological properties. J Water Health *20*, 54–67. https://doi.org/10.2166/wh.2021.128.

252. Beier, C.L., Horn, M., Michel, R., Schweikert, M., Görtz, H.-D., and Wagner, M. (2002). The Genus *Caedibacter* Comprises Endosymbionts of *Paramecium* spp. Related to the *Rickettsiales* (*Alphaproteobacteria*) and to *Francisella tularensis* (*Gammaproteobacteria*). Appl Environ Microbiol *68*, 6043–6050. https://doi.org/10.1128/AEM.68.12.6043-6050.2002.

253. Belova, S.E., Ravin, N. V, Pankratov, T.A., Rakitin, A.L., Ivanova, A.A., Beletsky, A. V, Mardanov, A. V, Sinninghe Damsté, J.S., and Dedysh, S.N. (2018). Hydrolytic Capabilities as a Key to Environmental Success: Chitinolytic and Cellulolytic Acidobacteria From Acidic Sub-arctic Soils and Boreal Peatlands. Front Microbiol *9*.

254. Zaitsev, G.M., Tsitko, I. V, Rainey, F.A., Trotsenko, Y.A., Uotila, J.S., Stackebrandt, E., and Salkinoja-Salonen, M.S. (1998). New aerobic ammonium-dependent obligately oxalotrophic bacteria: description of Ammoniphilus oxalaticus gen. nov., sp. nov. and Ammoniphilus oxalivorans gen. nov., sp. nov. Int J Syst Bacteriol *48 Pt 1*, 151–163. https://doi.org/10.1099/00207713-48-1-151.

255. Saha, P., Krishnamurthi, S., Mayilraj, S., Prasad, G.S., Bora, T.C., and Chakrabarti, T. (2005). Aquimonas voraii gen. nov., sp. nov., a novel gammaproteobacterium isolated from a warm spring of Assam, India. Int J Syst Evol Microbiol *55*, 1491–1495. https://doi.org/10.1099/ijs.0.63552-0.

256. Lang, E., and Reichenbach, H. (2013). Designation of type strains for seven species of the order Myxococcales and proposal for neotype strains of Cystobacter ferrugineus, Cystobacter minus and Polyangium fumosum. Int J Syst Evol Microbiol *63*, 4354–4360. https://doi.org/10.1099/ijs.0.056440-0.

257. Nautiyal, C.S., Rehman, A., and Chauhan, P.S. (2010). Environmental Escherichia coli occur as natural plant growth-promoting soil bacterium. Arch Microbiol *192*, 185–193. https://doi.org/10.1007/s00203-010-0544-1.

258. Chin, K.J., Liesack, W., and Janssen, P.H. (2001). Opitutus terrae gen. nov., sp. nov., to accommodate novel strains of the division "Verrucomicrobia" isolated from rice paddy soil. Int J Syst Evol Microbiol *51*, 1965–1968. https://doi.org/10.1099/00207713-51-6-1965.

259. Alker, A.T., Delherbe, N., Purdy, T.N., Moore, B.S., and Shikuma, N.J. (2020). Genetic examination of the marine bacterium Pseudoalteromonas luteoviolacea and effects of its metamorphosis-inducing factors. Environ Microbiol *22*, 4689–4701. https://doi.org/10.1111/1462-2920.15211.

260. Xu, Z.-X., Zhang, H.-X., Han, J.-R., Dunlap, C.A., Rooney, A.P., Mu, D.-S., and Du, Z.-J. (2017). Colwellia agarivorans sp. nov., an agar-digesting marine bacterium isolated from coastal seawater. Int J Syst Evol Microbiol *67*, 1969–1974. https://doi.org/10.1099/ijsem.0.001897.

261. G, S., Kumar, D., Uppada, J., Ch, S., and Ch V, R. (2020). Rhodomicrobium lacus sp. nov., an alkalitolerent bacterium isolated from Umiam lake, Shillong, India. Int J Syst Evol Microbiol *70*, 662–667. https://doi.org/10.1099/ijsem.0.003813.

262. Gerth, K., Pradella, S., Perlova, O., Beyer, S., and Müller, R. (2003). Myxobacteria: proficient producers of novel natural products with various biological activities—past and future biotechnological aspects with the focus on the genus Sorangium. J Biotechnol *106*, 233–253. https://doi.org/https://doi.org/10.1016/j.jbiotec.2003.07.015.

263. Seipke, R.F., Kaltenpoth, M., and Hutchings, M.I. (2012). Streptomyces as symbionts: an emerging and widespread theme? FEMS Microbiol Rev *36*, 862–876. https://doi.org/10.1111/j.1574-6976.2011.00313.x.

264. Yamamoto, E., Muramatsu, H., and Nagai, K. (2014). Vulgatibacter incomptus gen. nov., sp. nov. and Labilithrix luteola gen. nov., sp. nov., two myxobacteria isolated from soil in

Yakushima Island, and the description of Vulgatibacteraceae fam. nov., Labilitrichaceae fam. nov. and Anaeromyxobacteraceae fam. nov. Int J Syst Evol Microbiol *64*, 3360–3368. https://doi.org/10.1099/ijs.0.063198-0.

265. Boch, J., and Bonas, U. (2010). Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu Rev Phytopathol *48*, 419–436. https://doi.org/10.1146/annurev-phyto-080508-081936.

266. Beier, C.L., Horn, M., Michel, R., Schweikert, M., Görtz, H.-D., and Wagner, M. (2002). The genus Caedibacter comprises endosymbionts of Paramecium spp. related to the Rickettsiales (Alphaproteobacteria) and to Francisella tularensis (Gammaproteobacteria). Appl Environ Microbiol *68*, 6043–6050. https://doi.org/10.1128/AEM.68.12.6043-6050.2002.

267. Steenhoudt, O., and Vanderleyden, J. (2000). Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects. FEMS Microbiol Rev *24*, 487–506. https://doi.org/10.1111/j.1574-6976.2000.tb00552.x.

268. Mollenhauer, D., Bengtsson, R., and Lindstrøm, E.-A. (1999). Macroscopic cyanobacteria of the genus Nostoc: a neglected and endangered constituent of European inland aquatic biodiversity. Eur J Phycol *34*, 349–360. https://doi.org/10.1080/09670269910001736412.

269. Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. Mol Biol Evol *30*, 1188–1195. https://doi.org/10.1093/molbev/mst024.

270. Michel, R., and Hauröder, B. (1997). Isolation of an Acanthamoeba Strain with Intracellular Burkholderia pickettii Infection. Zentralblatt für Bakteriologie *285*, 541–557. https://doi.org/10.1016/S0934-8840(97)80116-8.

271. Foster, R.C., and Dormaar, J.F. (1991). Bacteria-grazing amoebae in situ in the rhizosphere. Biol Fertil Soils *11*, 83–87. https://doi.org/10.1007/BF00336368.

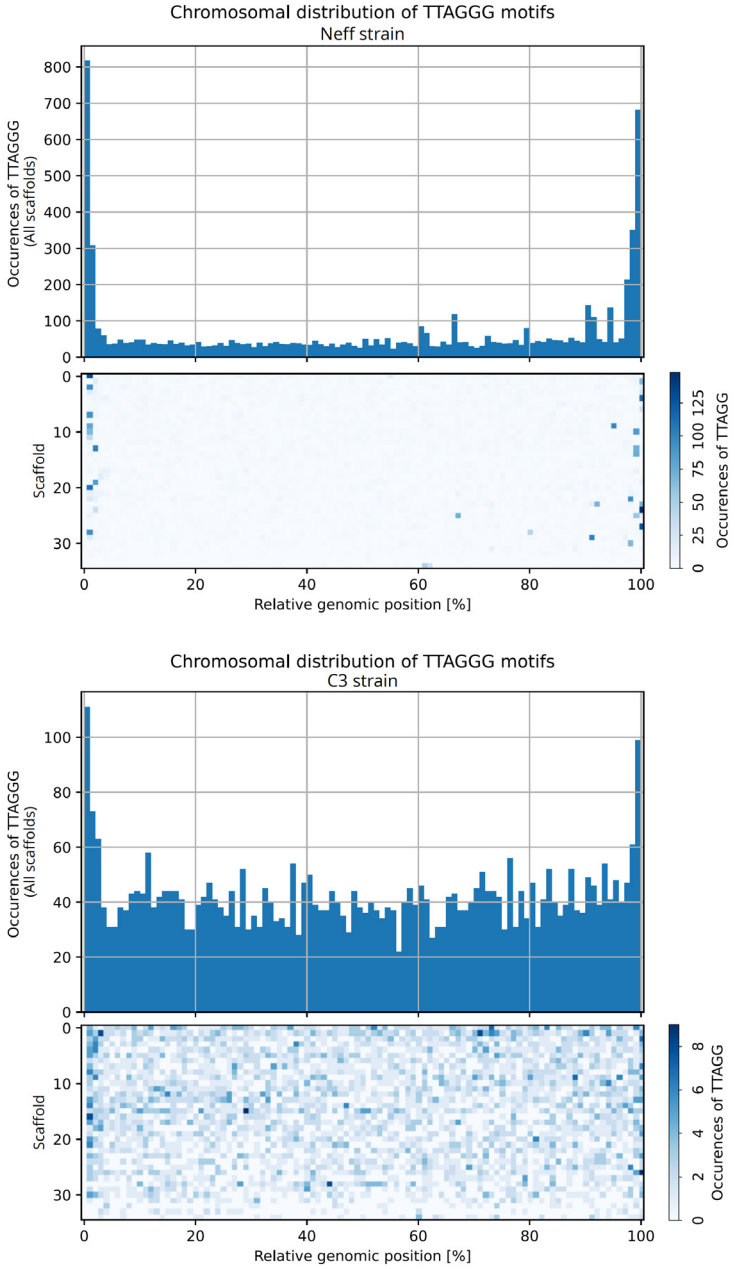
272. Ford Doolittle, W. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends in Genetics *14*, 307–311. https://doi.org/10.1016/S0168-9525(98)01494-2.

273. Doolittle, W.F., Boucher, Y., NesbØ, C.L., Douady, C.J., Andersson, J.O., and Roger, A.J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci *358*, 39–58. https://doi.org/10.1098/rstb.2002.1185.

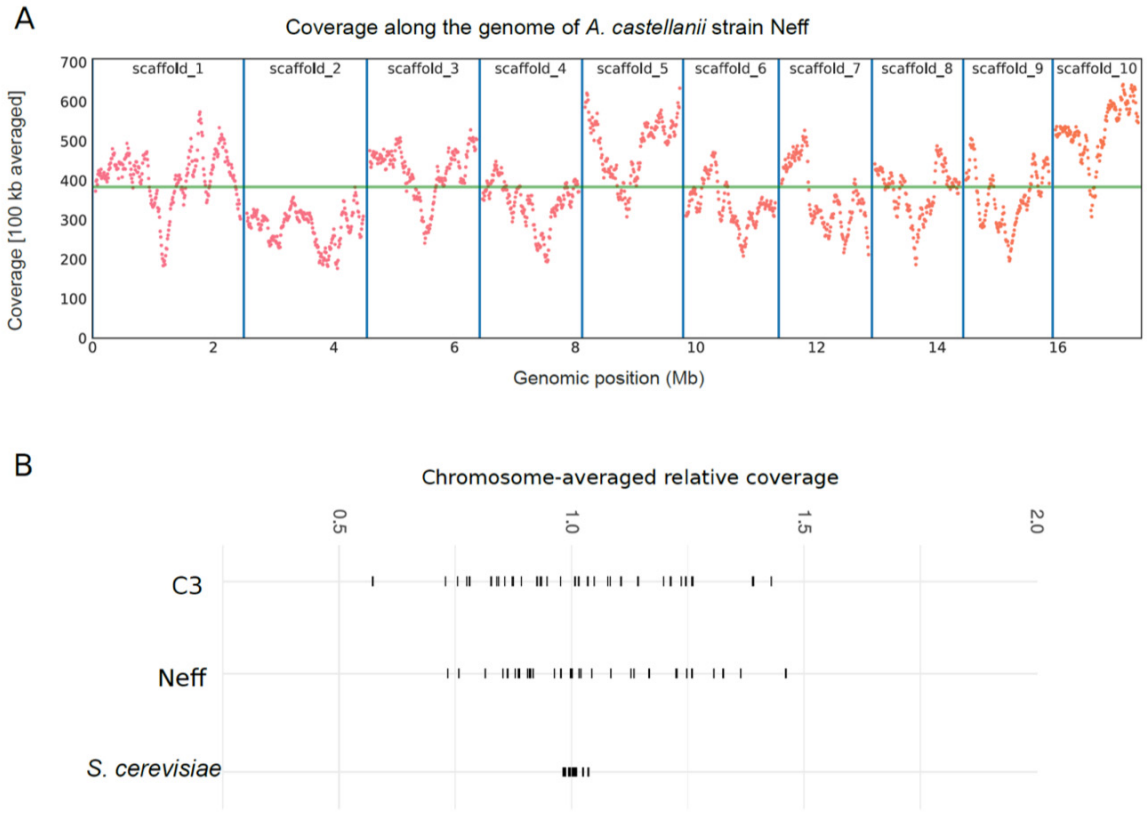
274. Rosenberg, K., Bertaux, J., Krome, K., Hartmann, A., Scheu, S., and Bonkowski, M. (2009). Soil amoebae rapidly change bacterial community composition in the rhizosphere of *Arabidopsis thaliana*. ISME J *3*, 675–684. https://doi.org/10.1038/ismej.2009.11.

275. Pinto, L.F., Andriolo, B.N.G., Hofling-Lima, A.L., and Freitas, D. (2021). The role of Acanthamoeba spp. in biofilm communities: a systematic review. Parasitol Res *120*, 2717–2729. https://doi.org/10.1007/s00436-021-07240-6.

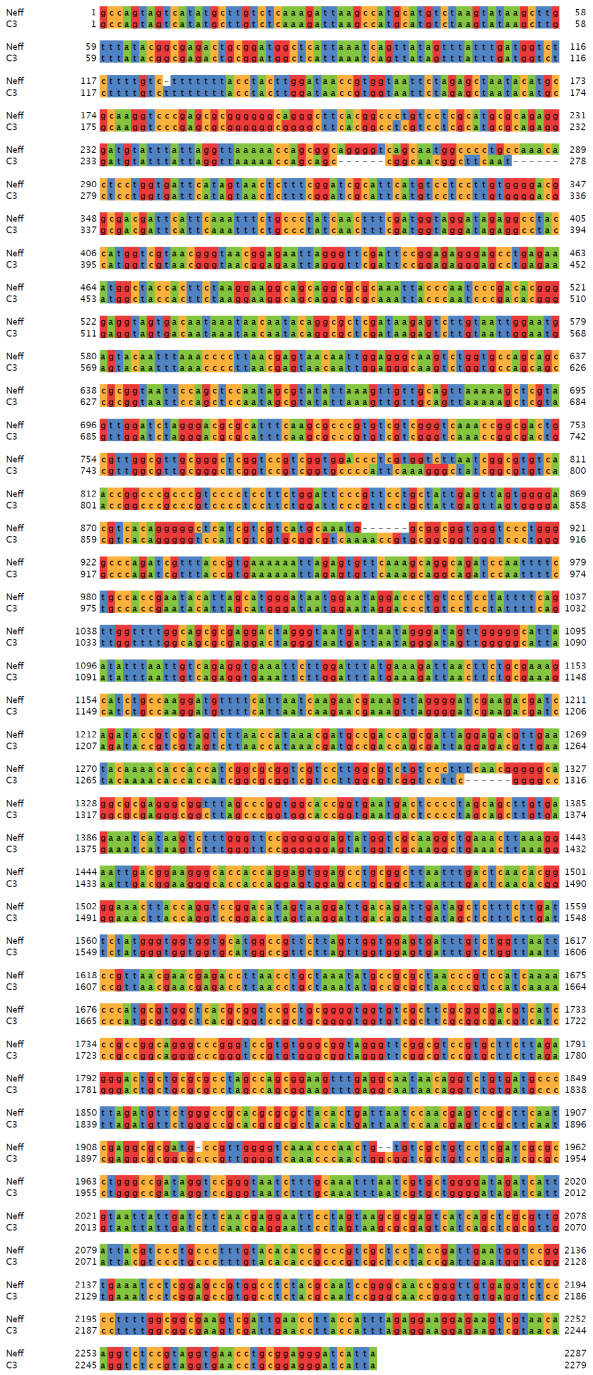
Appendix A



Supplementary Figure 2.1 Genomic distribution of subtelomeric repeats along *Acanthamoeba castellanii* scaffolds. Relative position of "TTAGGG" subtelomeric repeats, from the beginning (0%) to the end (100%) of scaffolds. In *A. castellanii* strains Neff (top) and C3 (bottom). The histograms show the total repeat density summed across the 35 largest scaffolds, while the heatmaps show the distribution for each scaffold on separate rows. Both the heatmap and histogram are binned at 1% of relative scaffold length.

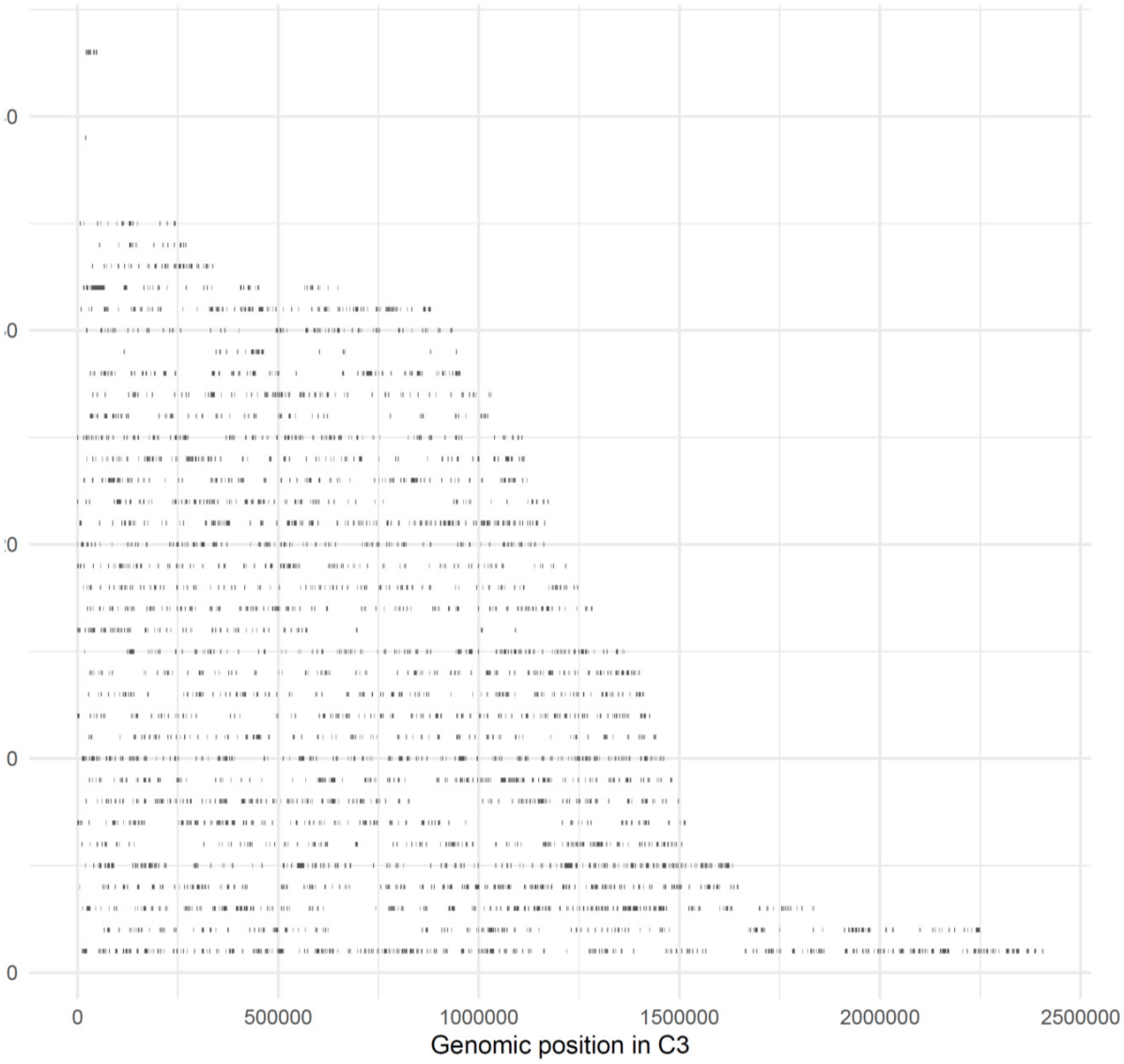


Supplementary Figure 2.2 Read coverage across scaffolds of *A. castellanii*.
A, Illumina short-reads coverage along the 10 largest scaffolds of *A. castellanii* Neff in a 100 kb sliding
window, with the horizontal green line showing genome median coverage.
B, Variability of median coverage per chromosome (relative to genome median) for *A. castellanii* strains C3 and Neff, and asynchronous *Saccharomyces cerevisiae* haploid strain BY4741. For *S. cerevisiae,* library SRR1569870 was used.



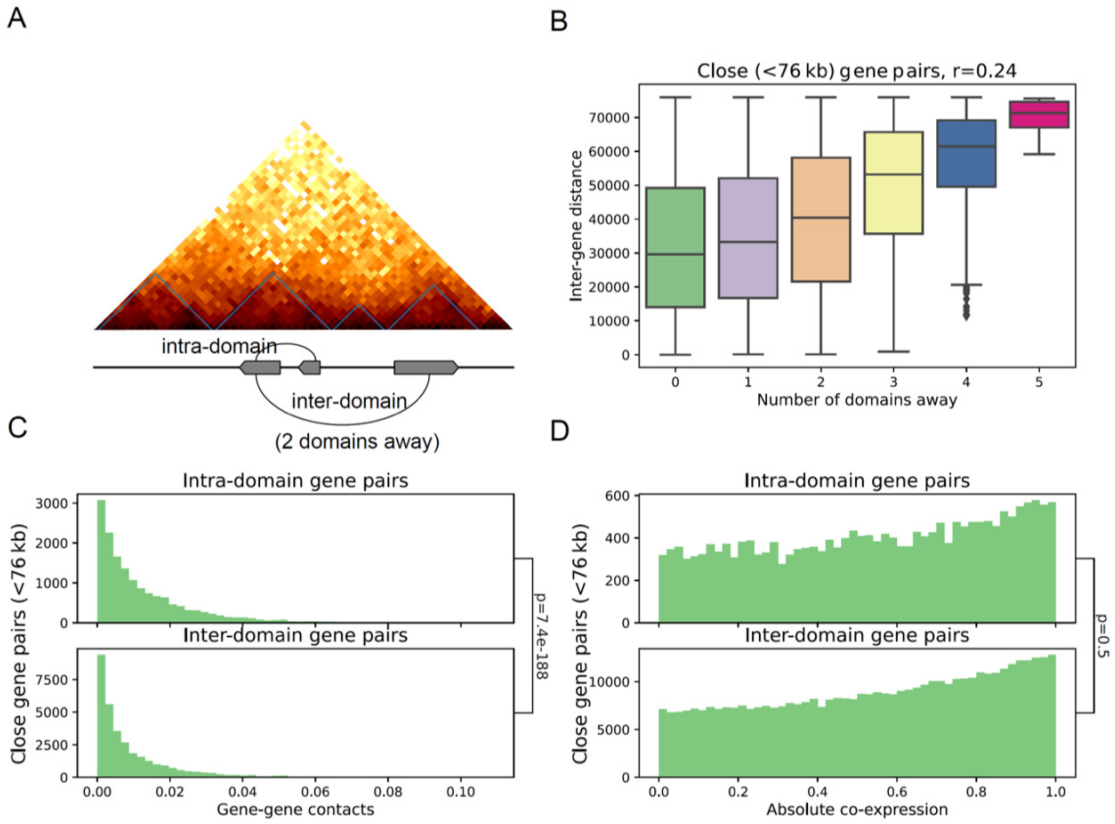
Supplementary Figure 2.3 Comparison of 18S rDNA sequences from *Acanthamoeba castellanii* C3 and Neff strains.

The *A. castellanii* 18S sequence from NCBI was used to retrieve the 18S rDNA sequences from the C3 and Neff assemblies. These retrieved sequences were aligned using MAFFT[155] v7.4 with auto setting and visualized in Jalview[156] v2.1.1.3.

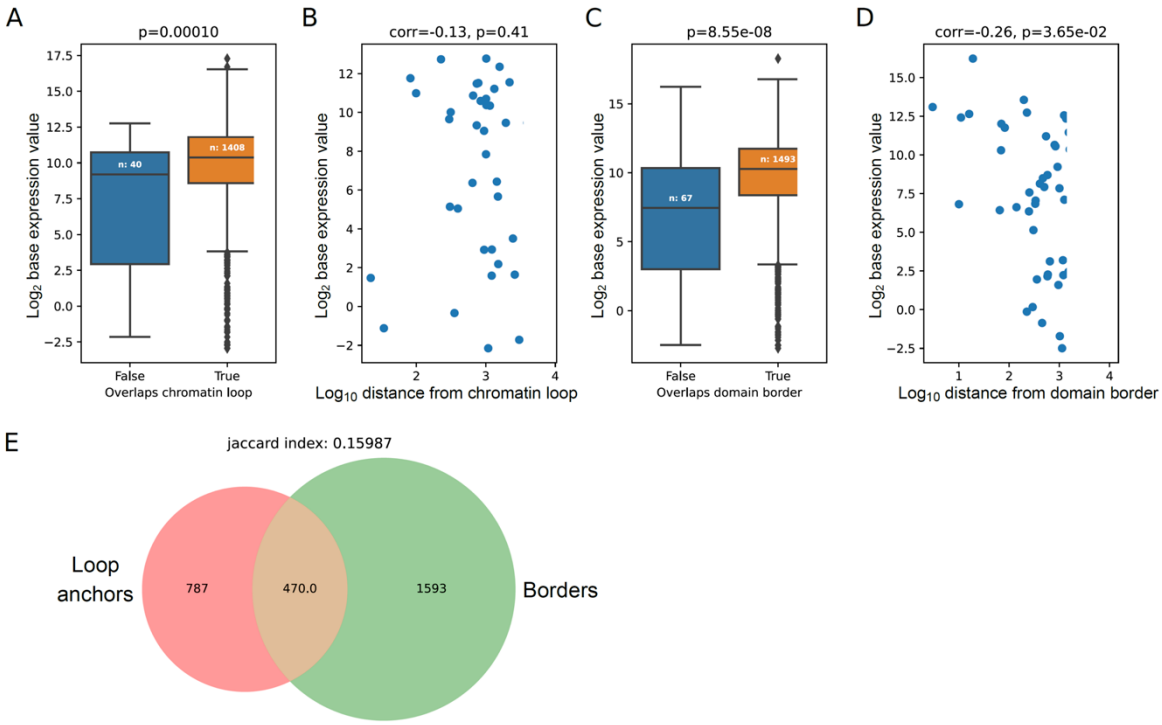


Supplementary Figure 2.4 Genomic distribution of *Acanthamoeba castellanii* strain C3 regions with no mapping in Neff.

The C3 genome was aligned to the Neff assembly using Minimap2[133] v2.1, and the coordinates of C3 regions with no mapping in Neff were retrieved. The figure shows C3-specific regions in black along the 50 largest C3 scaffolds.



Supplementary Figure 2.5 Relationship between genes and self-interacting domains.
A, Example of domains detected using Chromosight in the C3 strain, with hypothetical genes indicated for
the purpose of illustration.
B, Relationship between inter-gene distance and the number of domains separating them. **C**, Distribution of mean inter-gene contacts according to domain separation status.
D, Distribution of gene-pairs co-expression according to domain separation status. For all panels, only gene pairs separated by less than the median domain size are selected.



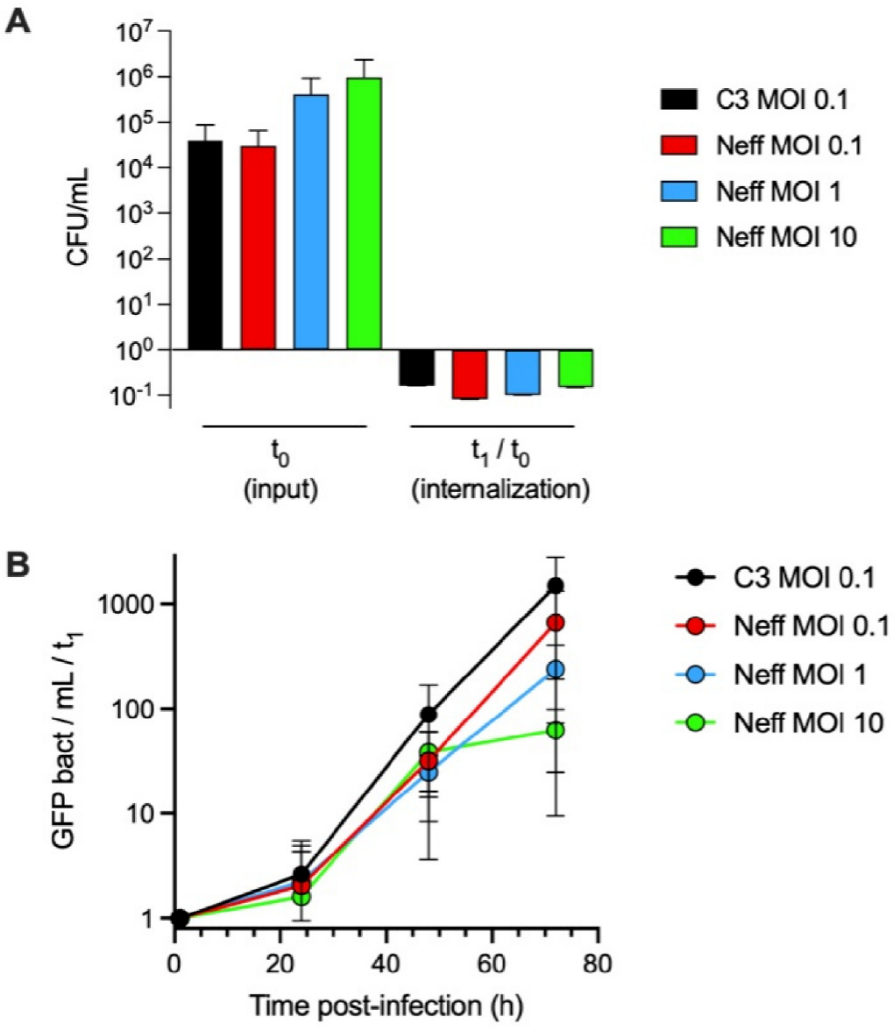
Supplementary Figure 2.6 Gene expression according to position relative to chromatin loop.

Expression of the closest gene to each loop anchors versus **A**, overlap status with chromatin loops and **B**, distance to closest loop.

Expression of the closest gene to domain borders versus **C**, overlap status with domain borders and **D**, distance to closest border.

P-values reported for overlap comparisons are obtained using Mann-Whitney U test, correlation coefficients and associated p-values are computed using Spearman 's correlation test.

E, Overlap between chromatin loop anchors and domain borders represented as a Venn diagram.

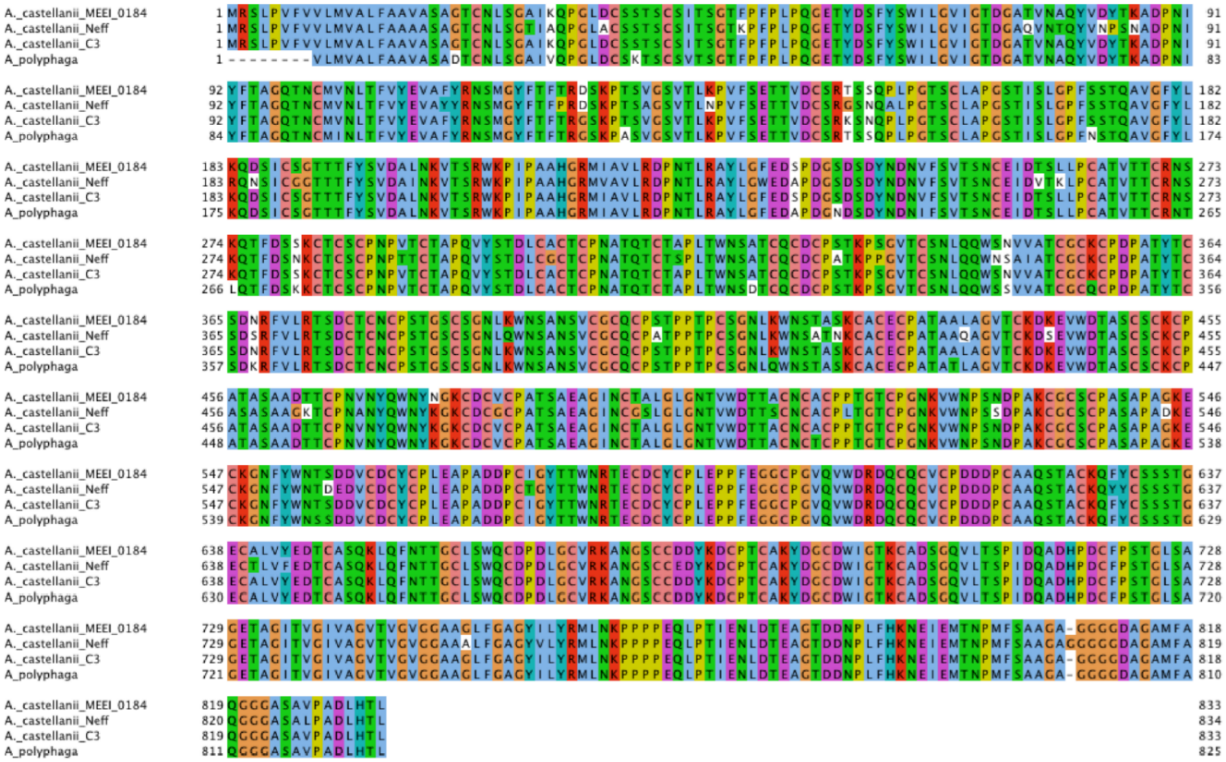


Supplementary Figure 2.7 Comparative entry and replication of *L. pneumophila* in C3 and Neff strains.

Acanthamoeba castellanii strains C3 and Neff were infected with *Legionella pneumophila* strain Paris constitutively expressing GFP. Strain C3 was infected at an MOI=0.1 and strain Neff was infected at MOI=0.1, MOI=I or MOI=IO.

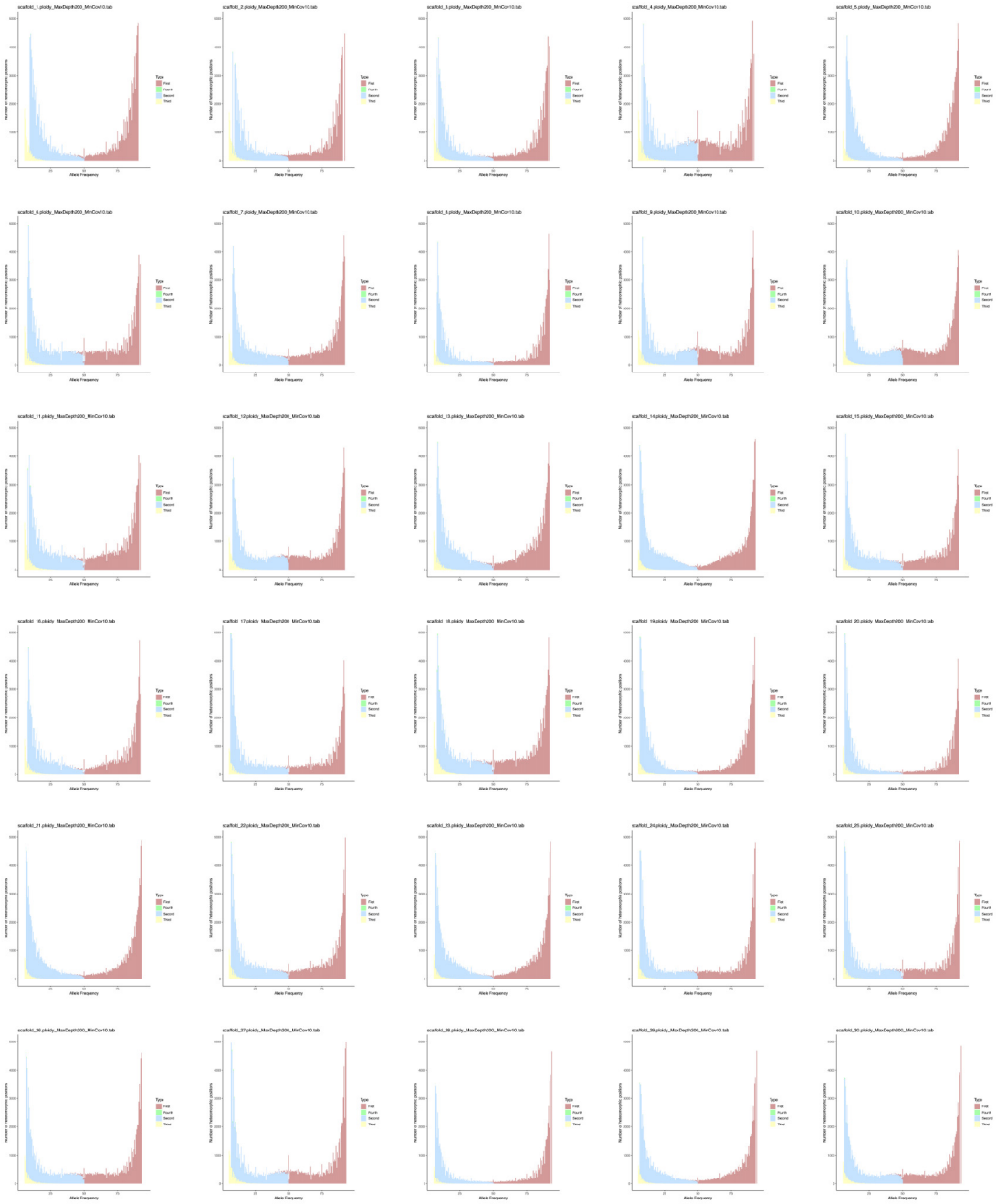
A, At t=O hours post infection (hpi) and at t=l hpi , 300 µL of infected amoeba were removed from the infection cultur e, amoebae were lysed and 100 µL were plated on BCYE plates. CFU were counted to determine bacterial numbers used for infection (input , *t0)* and numbers of internalized bacteria (internalization, ti /t0).

B, At t=O, 1, 24 , 48 and 72 hpi, 300 µL of infected amoeba were removed from the infection culture, amoeba were lysed and 100 µL were placed in 96 well plates in duplicates and analyzed by Flow Cytometry. Graph shows absolute numbers of GFP bacteria per mL. Data were normalized to ti (n=3).

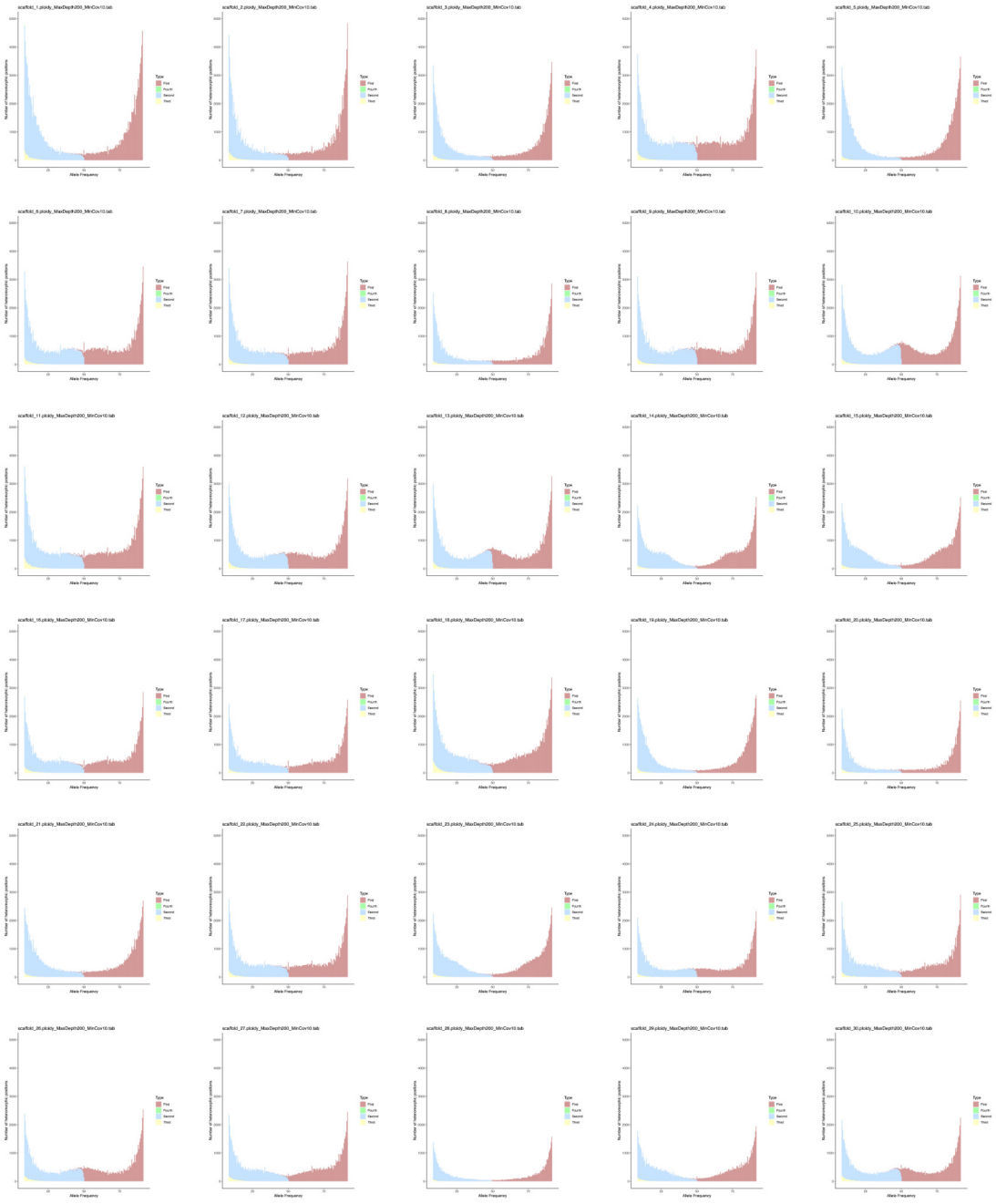


Supplementary Figure 2.8 Multiple sequence alignment of mannose binding protein orthologs across three strains of *Acanthamoeba castellanii* and one strain of *Acanthamoeba polyphaga*.

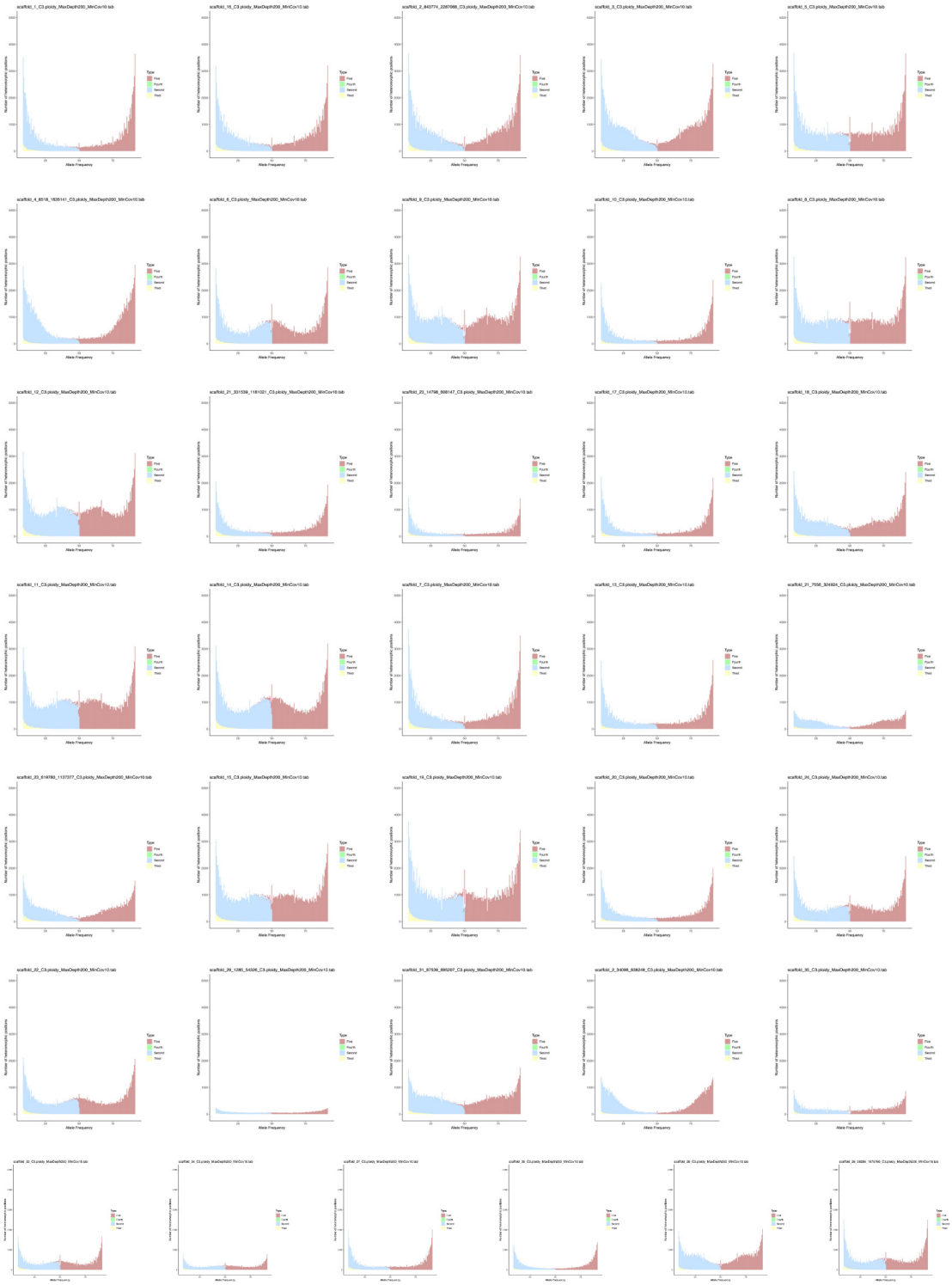
Sites are coloured according to the Clustalx colour scheme and residues differing from the consensus at any given site are not coloured. The alignment was generated with MAFFT- linsi, and was viewed and coloured in Jalview.



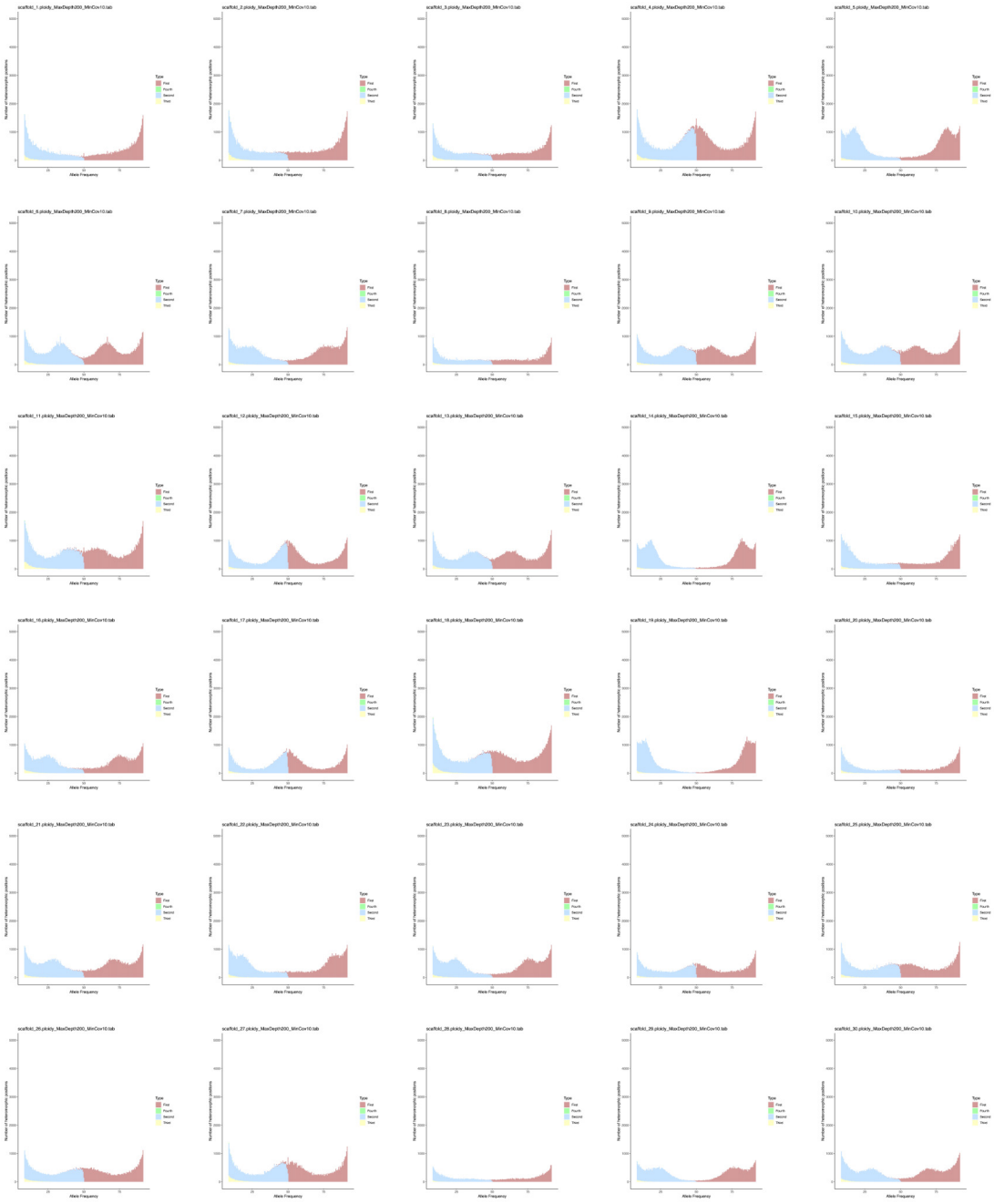
Supplementary Figure 4.1 PloidyNGS plots generated from wild-type Neff long reads from the Archibald lab, mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



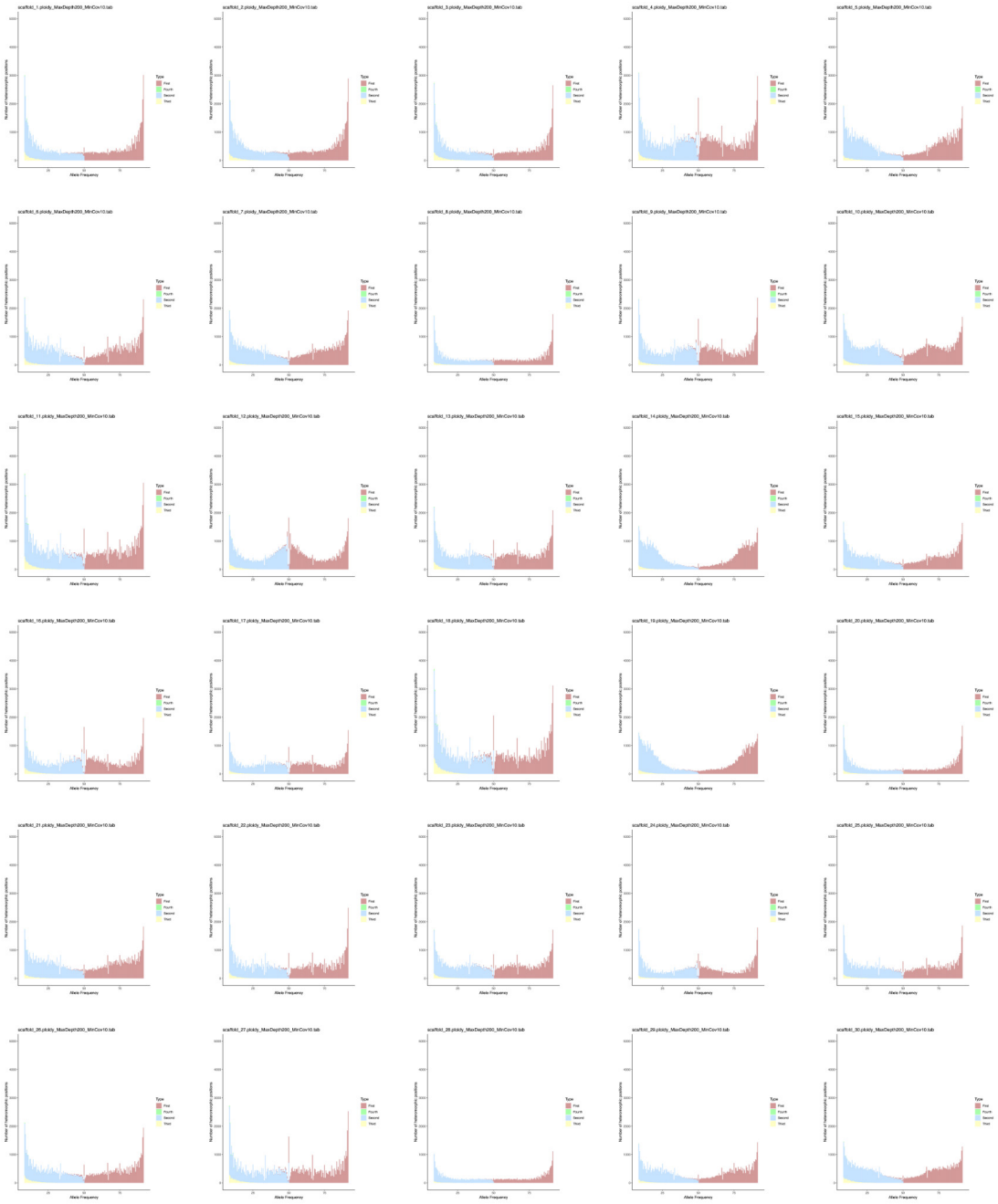
Supplementary Figure 4.2 PloidyNGS plots generated from wild-type Neff long reads from the Institut Pasteur, mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



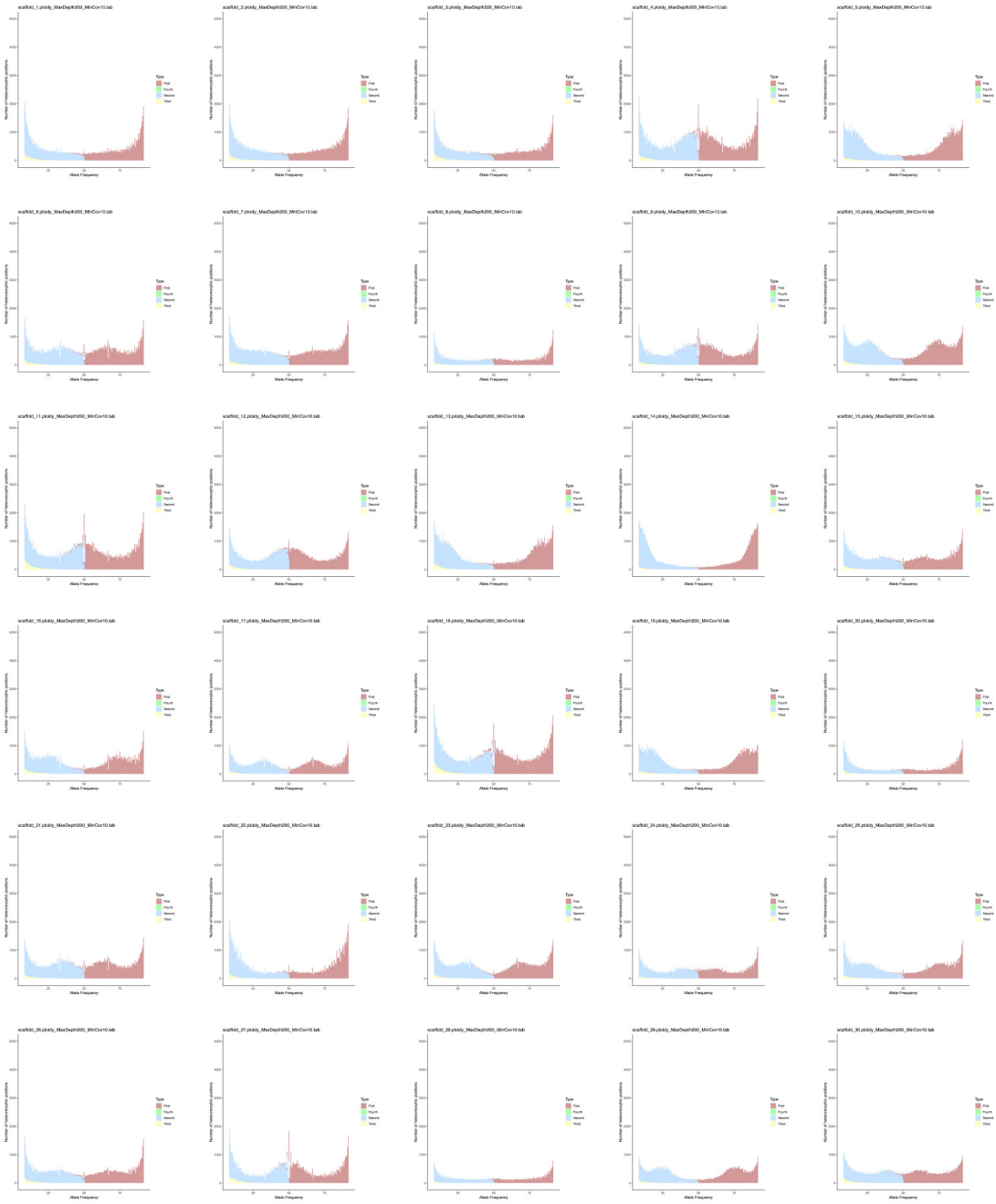
Supplementary Figure 4.3 PloidyNGS plots generated from wild-type C3 long reads, mapped to the C3 regions homologous to the 30 largest scaffolds of the Neff assembly. Plots are ordered according to the Neff scaffold they are homologous to, from largest to smallest.



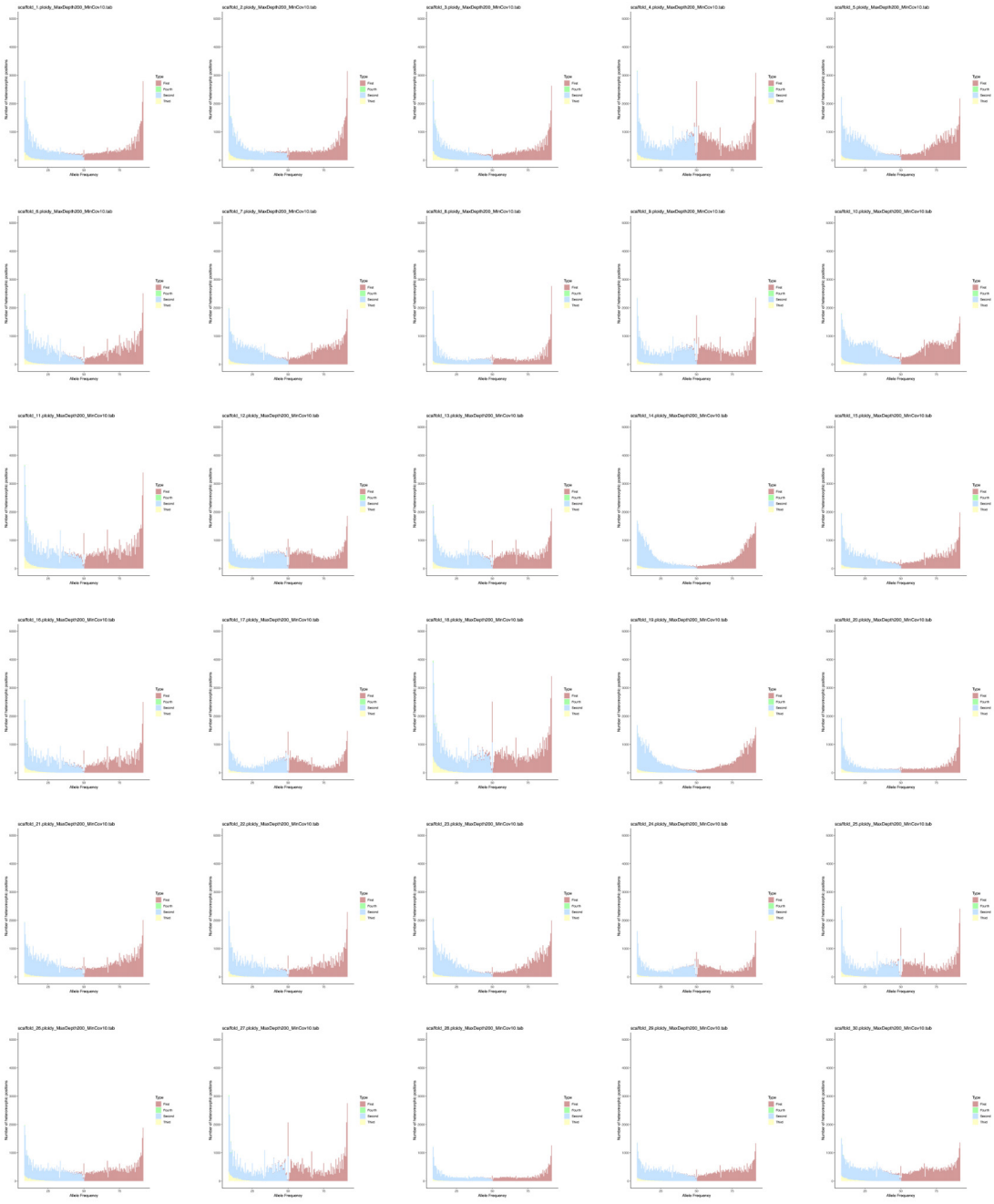
Supplementary Figure 4.4 PloidyNGS plots generated from Clone 1 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



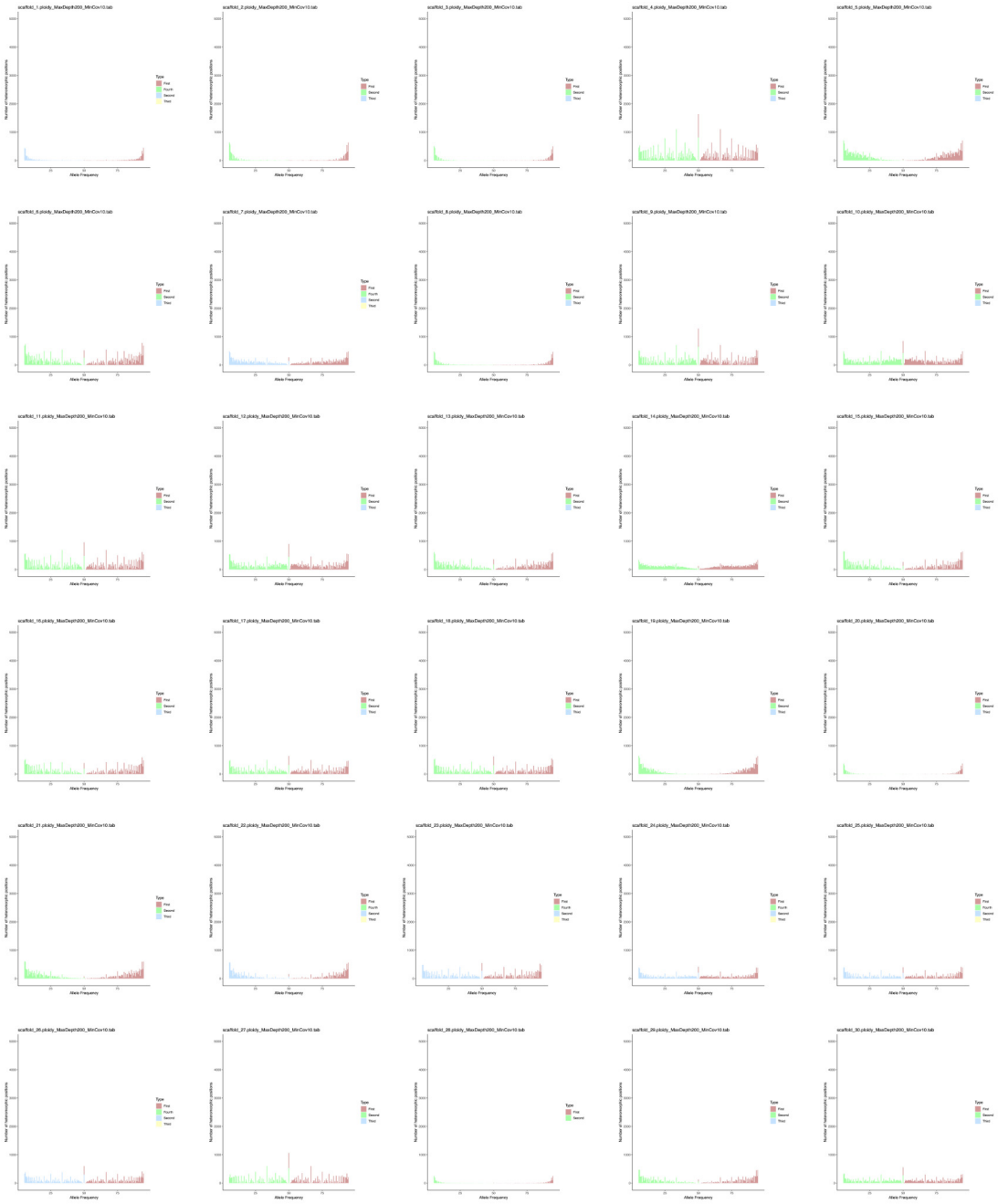
Supplementary Figure 4.5 PloidyNGS plots generated from Clone LT6 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



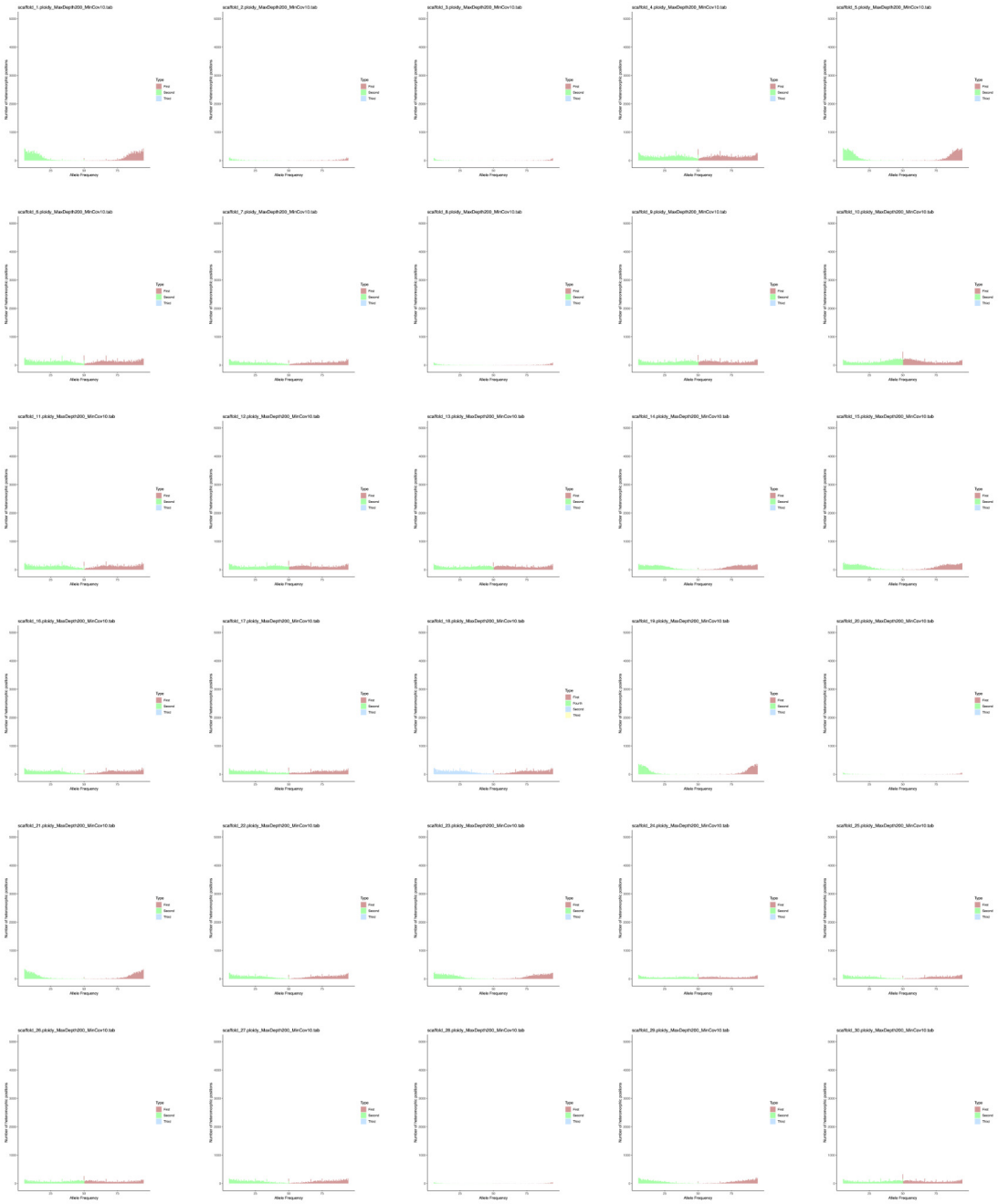
Supplementary Figure 4.6 PloidyNGS plots generated from Clone LT8 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



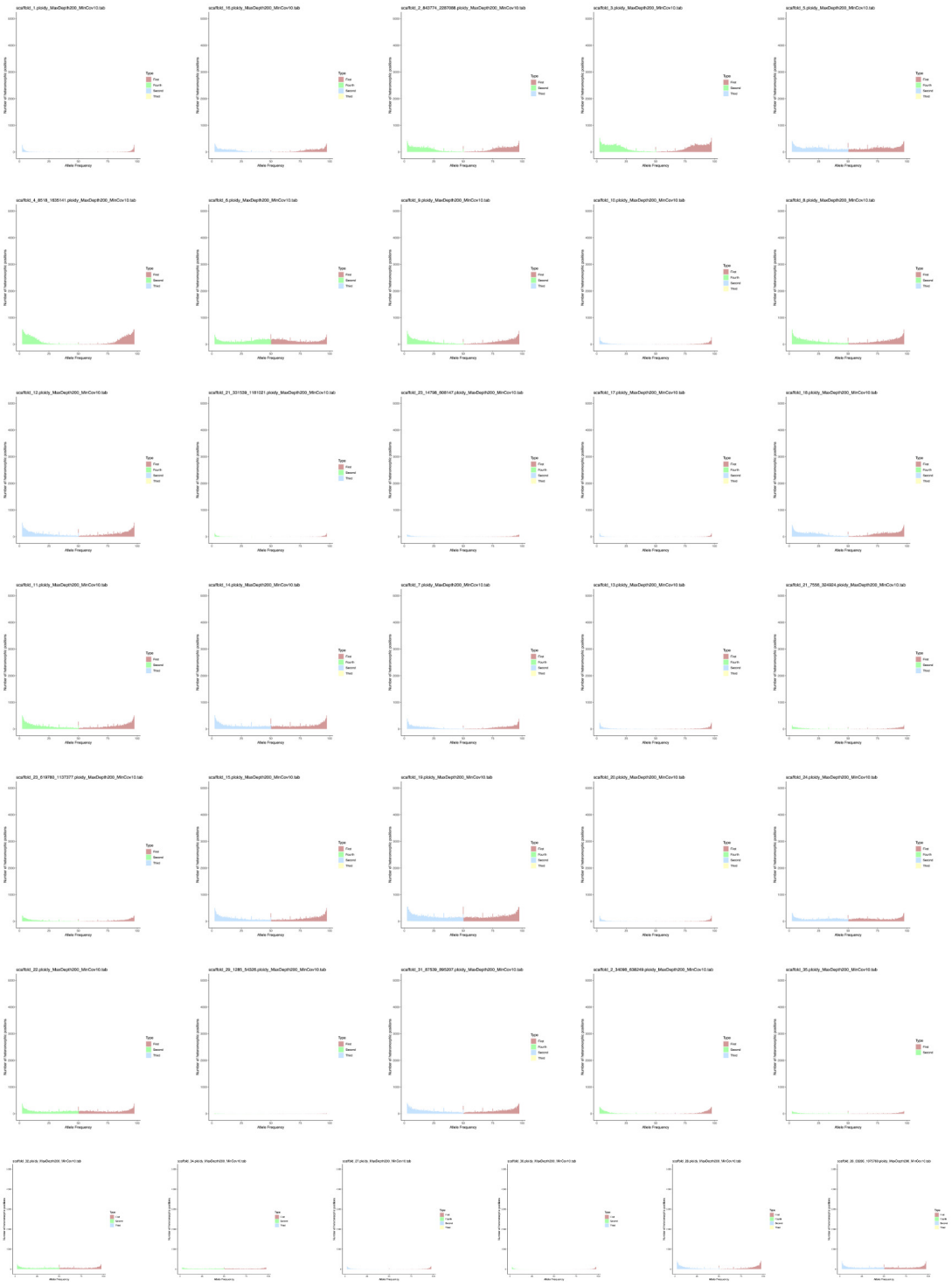
Supplementary Figure 4.7 PloidyNGS plots generated from Clone LT9 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



Supplementary Figure 4.8 PloidyNGS plots generated from wild-type Neff short reads from the Archibald lab, mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



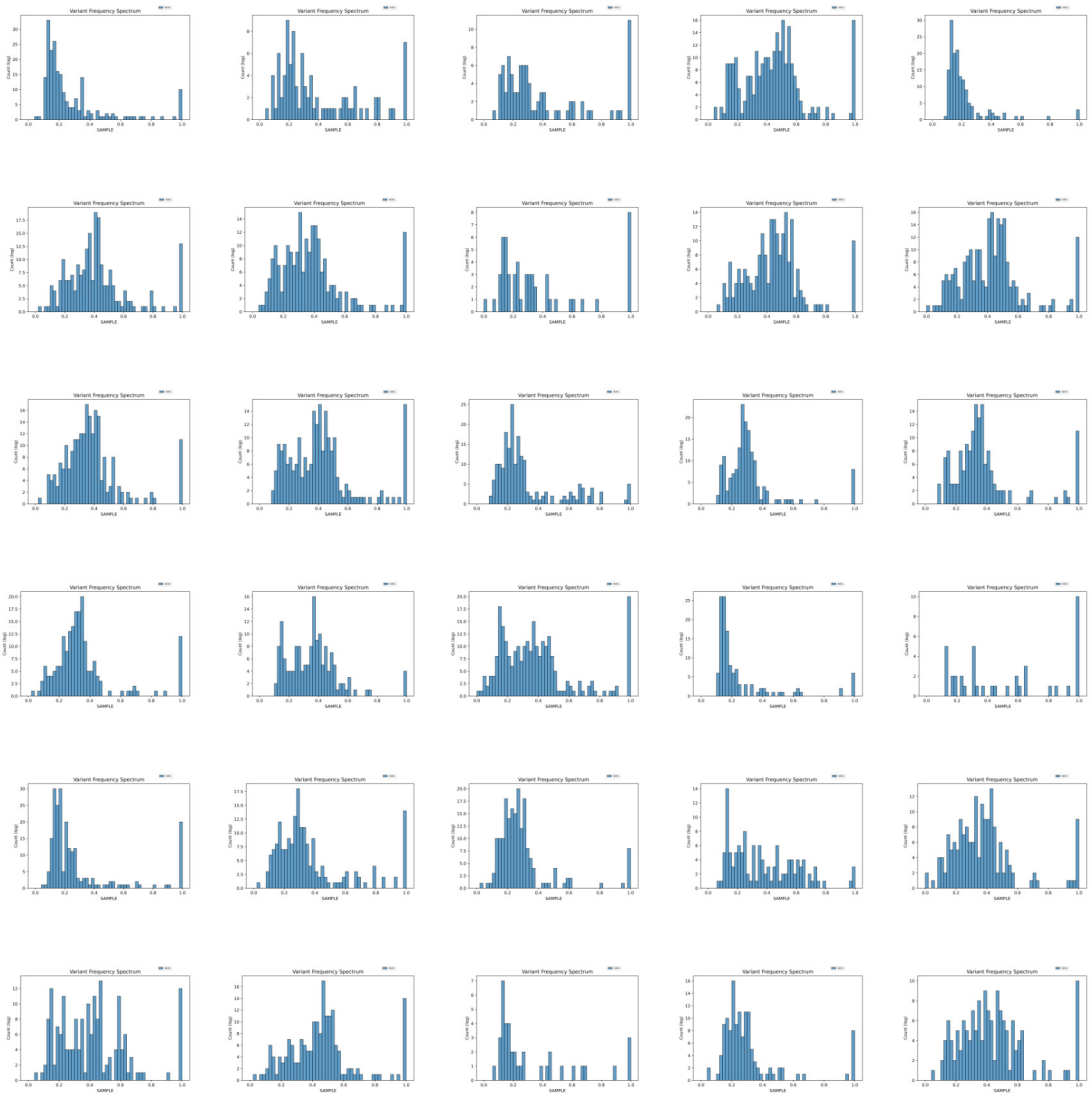
Supplementary Figure 4.9 PloidyNGS plots generated from wild-type Neff short reads from the Institut Pasteur, mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



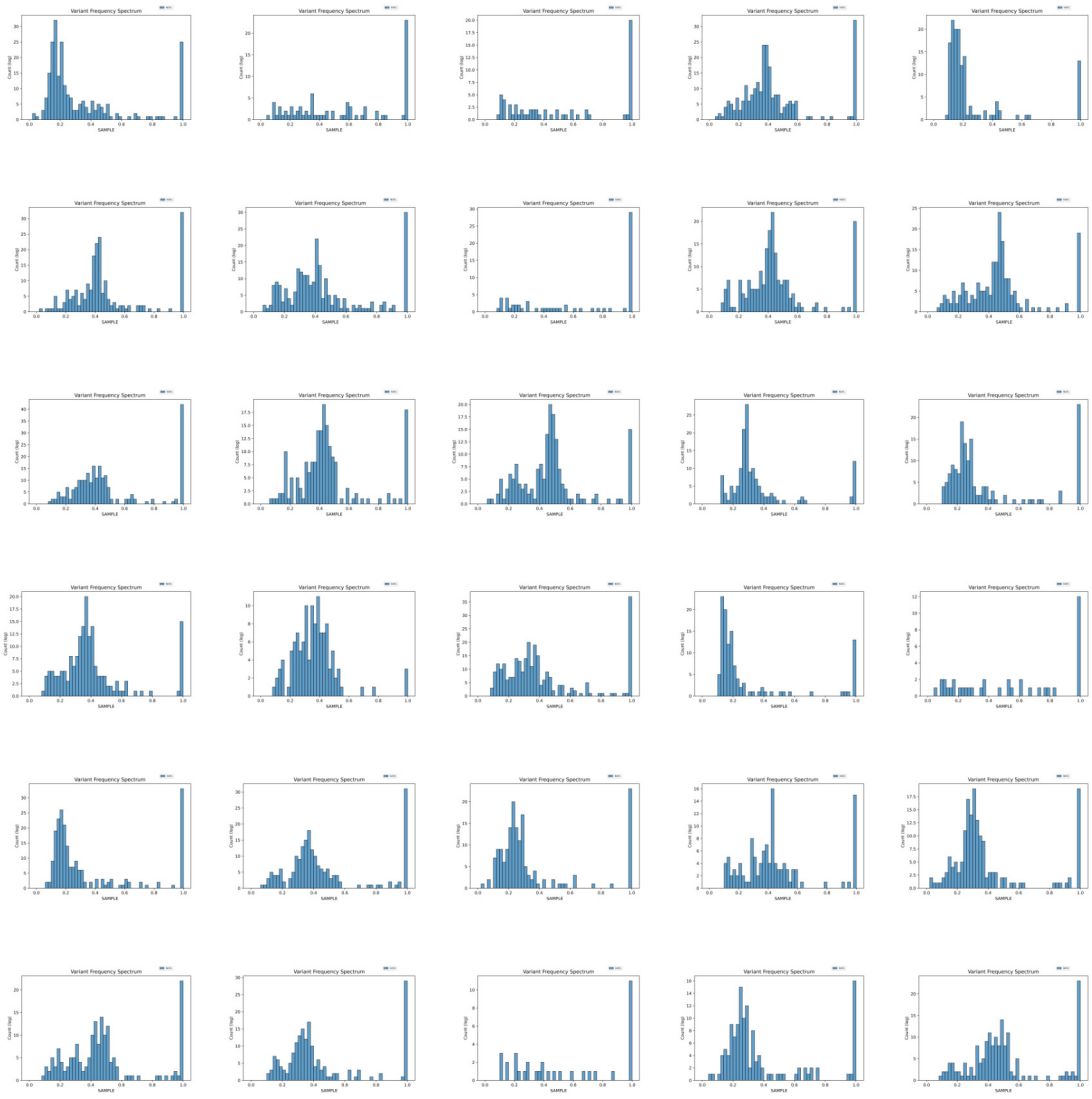
Supplementary Figure 4.10 PloidyNGS plots generated from wild-type C3 short reads, mapped to the C3 regions homologous to the 30 largest scaffolds of the Neff assembly. Plots are ordered according to the Neff scaffold they are homologous to, from largest to smallest.

Supplementary Figure 4.11 PloidyNGS plots generated from Clone LT6 short reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.

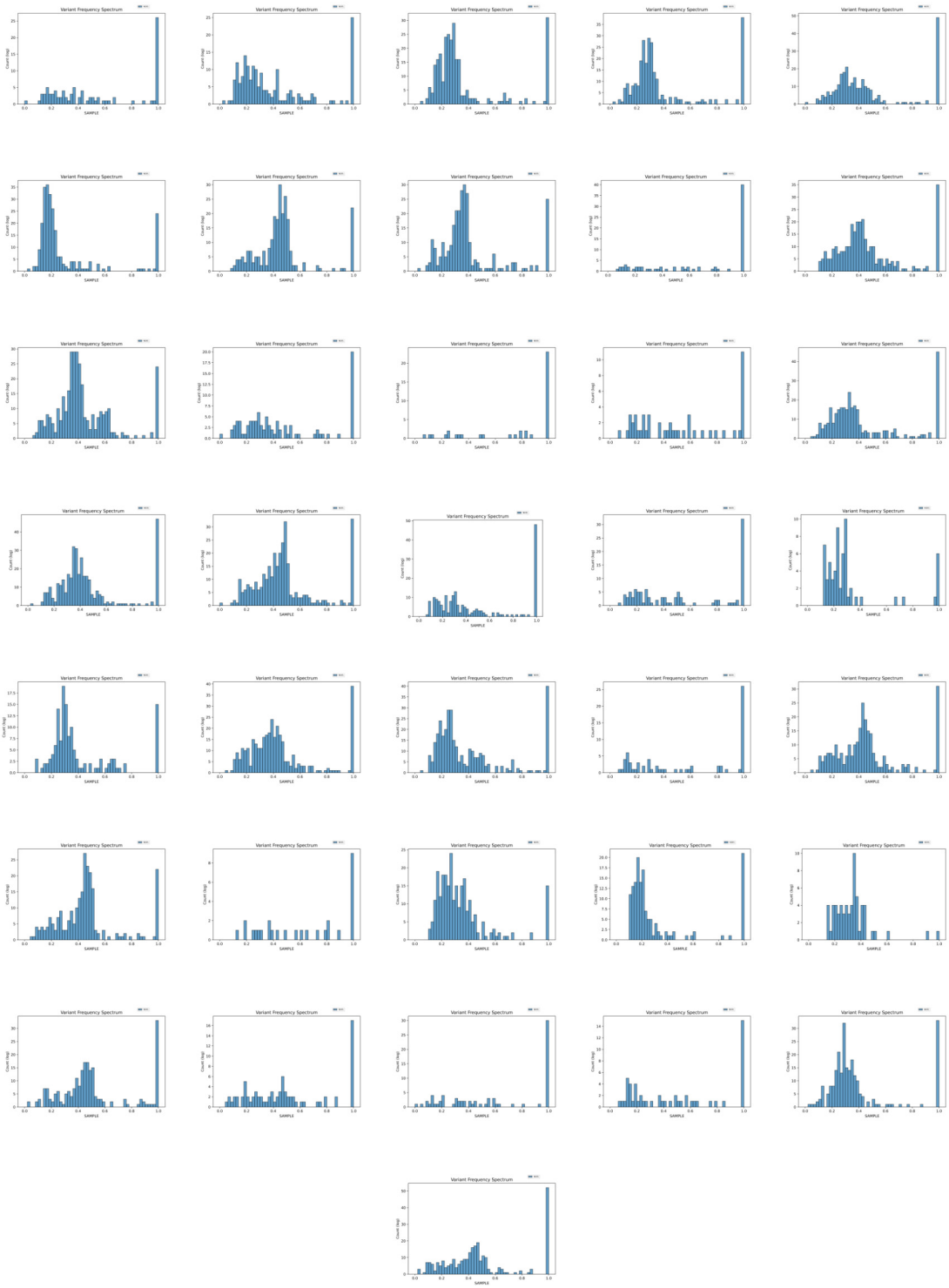
Supplementary Figure 4.12 PloidyNGS plots generated from Clone LT9 short reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



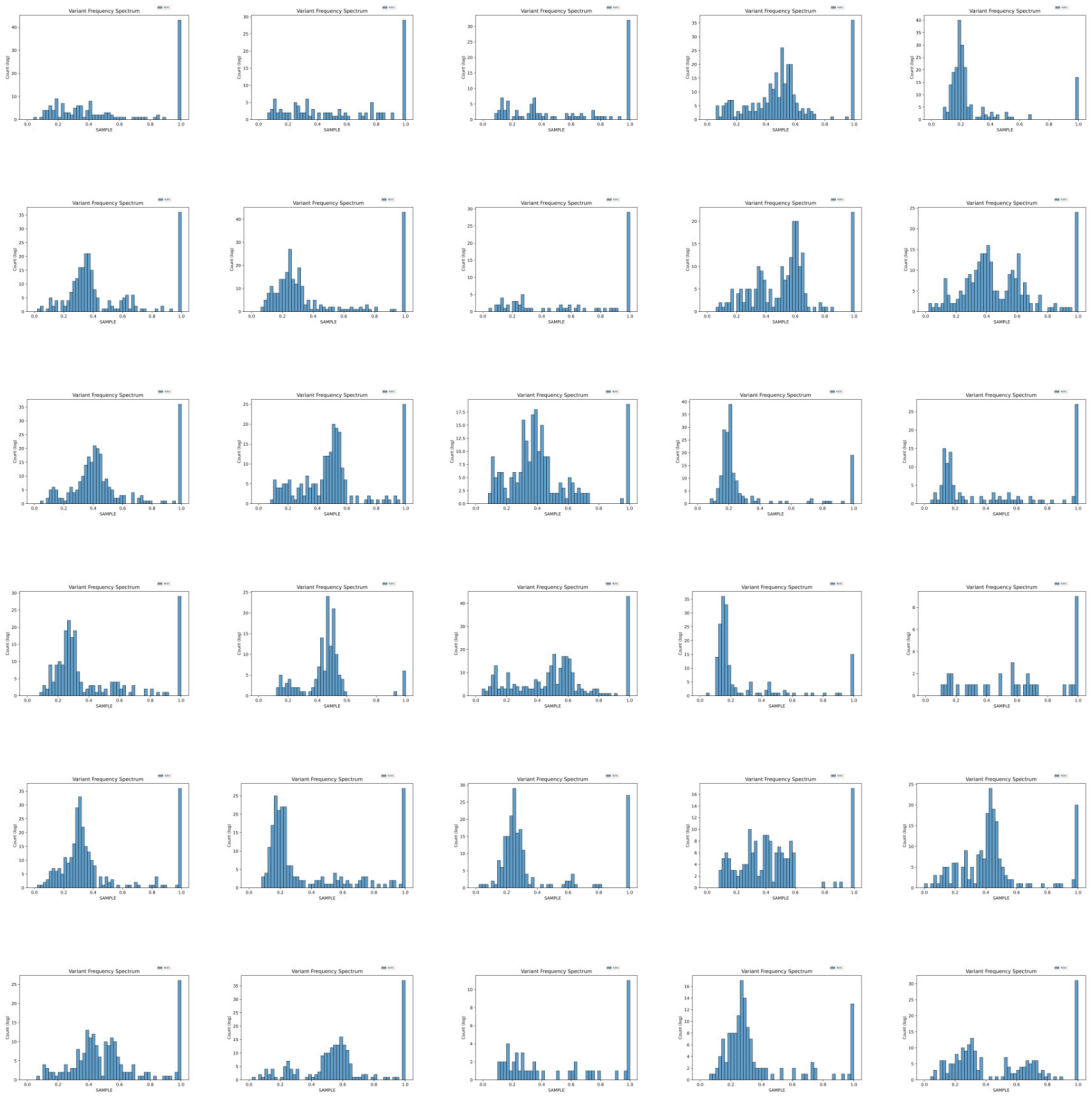
Supplementary Figure 4.13 Structural variant allele frequency plots generated from wild-type Neff long reads from the Archibald lab, mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



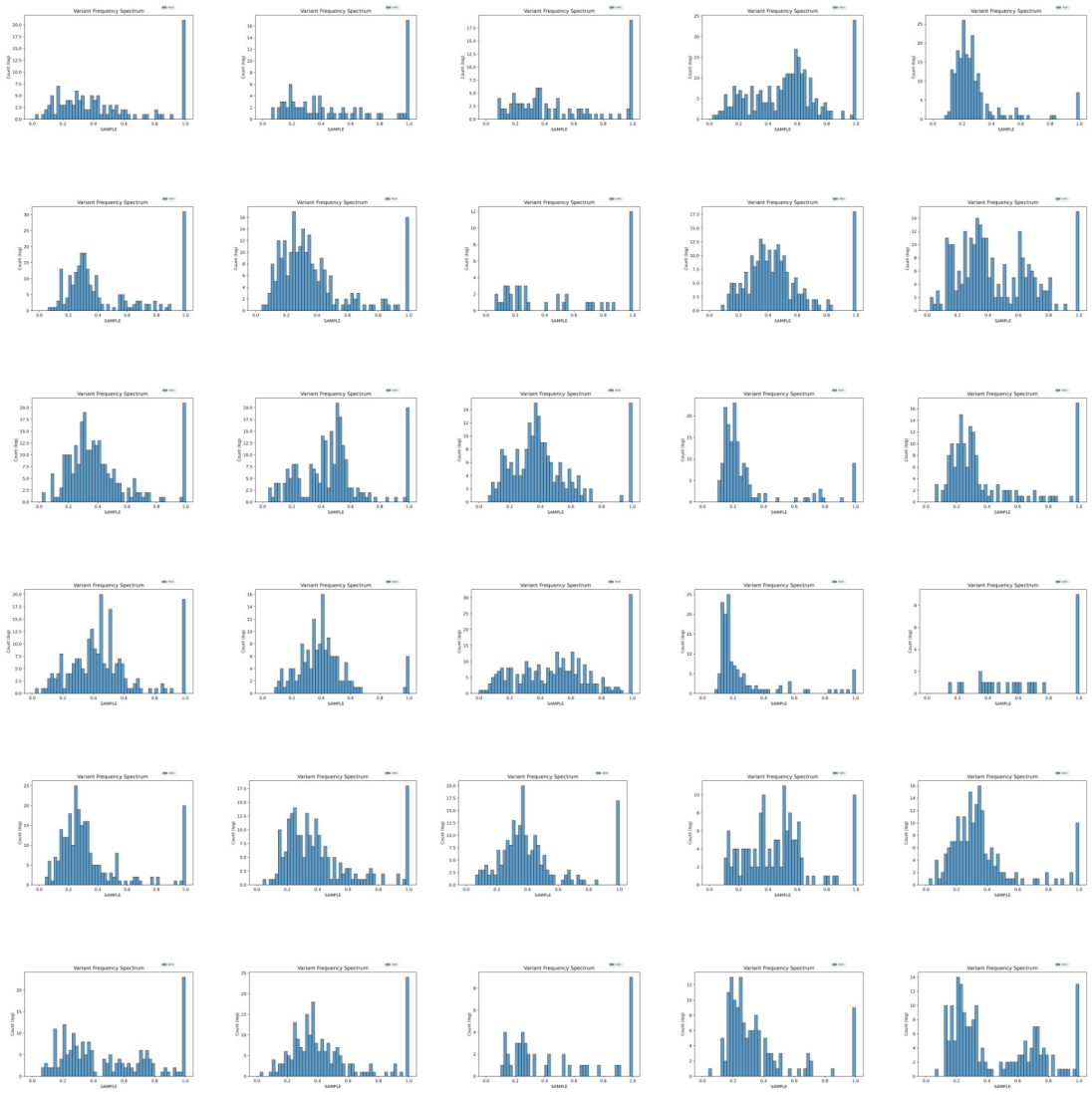
Supplementary Figure 4.14 Structural variant allele frequency plots generated from wild-type Neff long reads from the Institut Pasteur, mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



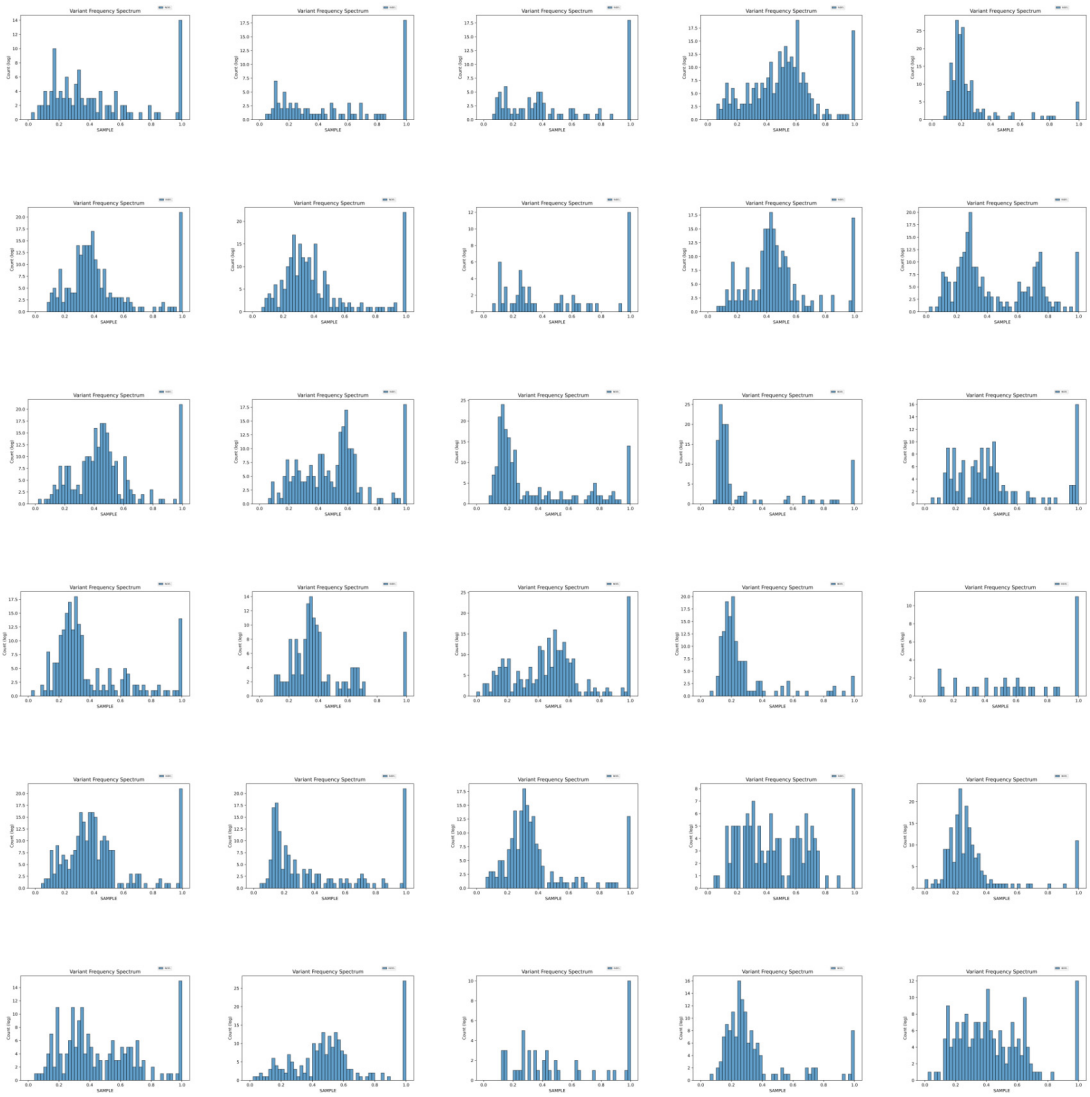
Supplementary Figure 4.15 Structural variant allele frequency plots generated from wild-type C3 long reads, mapped to the C3 regions homologous to the 30 largest scaffolds of the Neff assembly. Plots are ordered according to the Neff scaffold they are homologous to, from largest to smallest.



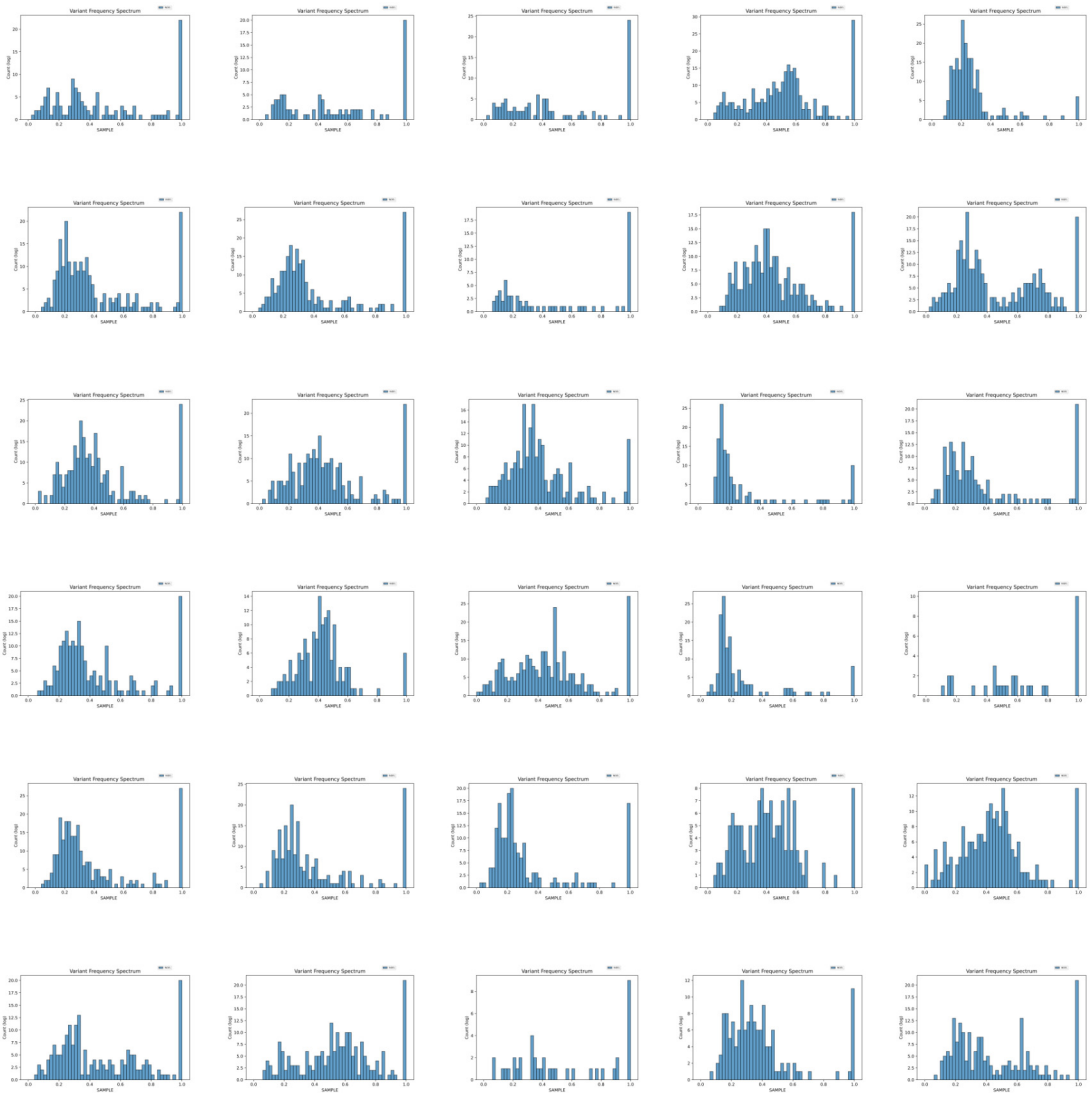
Supplementary Figure 4.16 Structural variant allele frequency plots generated from Clone 1 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



Supplementary Figure 4.17 Structural variant allele frequency plots generated from Clone LT6 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



Supplementary Figure 4.18 Structural variant allele frequency plots generated from Clone LT8 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.



Supplementary Figure 4.19 Structural variant allele frequency plots generated from Clone LT9 long reads mapped to the 30 largest scaffolds of the Neff assembly. Plots are ordered from largest scaffold to smallest.

Appendix B

Samples					Alignment				Event types					
Used for	Strain	Condition	Library	Mn. pairs	No	Multi	Single	MQ30	Discard	Intra	Inter	Dups.	Mn. used	% in matrix
Infection	C3	infected	AT418	102.5	23.4%	14.2%	62.3%	63%	91.1%	8.8%	2.8%	23%	3.8	3.7%
		uninfected	AT419	94.0	13.9%	19.7%	66.4%	68%	97.9%	2.1%	0.6%	32%	0.8	0.8%
		infected	AT420	94.4	19.1%	15.9%	64.9%	66%	94.9%	5.0%	1.6%	26%	2.0	2.1%
		uninfected	AT421	125.4	30.1%	16.1%	53.7%	55%	94.0%	5.9%	1.6%	44%	2.0	1.5%
Assembly		uninfected	AT337	50.8	6.8%	65.2%	27.8%	30%	23.9%	76.1%	23.7%	5%	6.5	12.9%
		infected	AT407	115.5	19.6%	14.5%	65.8%	64%	88.1%	11.8%	5.4%	7%	6.8	5.9%
		uninfected	AT408	112.2	27.1%	14.2%	58.6%	58%	85.8%	14.1%	6.8%	8%	6.9	6.2%
		infected	PM106	87.3	8.5%	17.0%	74.4%	73%	80.3%	19.6%	8.8%	18%	8.7	10.0%
	Neff	uninfected	AT338	40.0	60.5%	5.5%	33.8%	26%	10.0%	89.9%	37.7%	4%	2.5	6.3%

Supplementary Table S1 Read statistics for *Acanthamoeba castellanii* Hi-C libraries.
The first columns describe each library's sample: For what type of analysis the library was used (infection or genome assembly), what *A. castellanii* strain it contains, its ID and the number of read pairs sequenced in millions. The next columns describe alignment statistics: The percentage of reads which did not align to the reference, aligned more than once or a single time, as well as the total percentage of reads with a mapping quality above 30 (MQ30). The "Event types" columns describe the proportion of different Hi-C events relative to single-aligned reads that passed the MQ30 threshold: discarded events represent undigested restriction fragments or religation on the same fragment, while intra and inter represent valid Hi-C contact within- or between-scaffolds. The remaining columns show general statistics of the libraries, such as the proportion of PCR duplicates, millions and percentage of read pairs used in the final Hi-C contact maps. Despite showing higher percentages of retained reads, libraries AT337, AT407, AT408 and PM106 were not used for infection analysis because they were prepared in separate batches and presented technical variations.