MINING THE GENOME OF CRYPTOSPORIDIUM: *Prospecting for Biomarkers*

Arthur V. Morris

A thesis submitted for the degree of

Doctor of Philosophy

at

Aberystwyth University

2021

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Candidate name: Arthur V. Morris

Signature:

Date:

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s). Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signature:

Date:

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signature:

Date:

KESS 2 Funding Statement

Knowledge Economy Skills Scholarships (KESS 2) is a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.

Abstract

Cryptosporidium is a protozoan parasite responsible for causing diarrhoeal disease in humans. Cryptosporidiosis is spread by contact with contaminated municipal water supplies or swimming pools, and can pose a serious health risk for individuals with weakened immune systems. This disease takes a massive toll on global public health, with over 200,000 deaths in children of less than two years old in Asia and Sub-Saharan Africa being attributed to it annually. Genomics can be a valuable asset in helping combat this parasite and the disease it causes. The primary focus of this project was to identify novel biomarkers around the genome of *Cryptosporidium*, using novel bioinformatics software, which can be used to furnish epidemiological surveys with high resolution data. This work necessitates generating high quality, reliable genome assemblies. Consequently, over 40 new *Cryptosporidium* genomes were sequenced and assembled. The biomarkers identified using these genomes provide a strong foundation upon which multiplicity of infection can be elucidated, using a novel *in silico* pipeline. The tools developed were designed with computational efficiency in mind, with the intention that they can be used by the Public Health Wales *Cryptosporidium* Reference Unit. This kind of computational efficiency was achieved, in part, by using alignment-free sequence analysis techniques to analyse raw read sets generated by Next-Generation sequencing projects, obviating the computationally intensive task of genome assembly. The results presented here shed light on the complex way this parasite is transmitted, and will facilitate the development of novel prevention strategies in the battle against Cryptosporidiosis.

Keywords: Cryptosporidium, Genomics, NGS, bioinformatics, alignment-free sequence analysis

Supervisors
Dr. Martin Swain
Dr. Justin Pachebat
Prof. Rachel Chalmers

Word Count: 51,592

Acknowledgements

The last four years have been a test of my creativity, endurance, and resolve. However, I sit here due to the endless support of my mentors, peers, friends, and family, without which this thesis would never have come into existence.

Firstly, to my supervisors, Martin, Justin, and Rachel. Martin, you have been an invaluable source of advice and encouragement over the last few years. You have instilled in me a passion for the world of bioinformatics and have been endlessly patient with me as I learned how to be a bioinformatician, knowing when to guide me and when to allow me to follow my nose. Justin, your endless wealth of knowledge on the experimental side of this project, and eagerness to see me flourish outside my PhD, has kept me focussed. You have consistently nourished the Biologist in me, despite me spending less time in the lab than you would like! I could not have asked for a better mentor to induct me into the circle of Crypto researchers than Rachel. Your deep understanding of this parasite has been instrumental in the development and realisation of this project. I have lost count how many times I have used the term "Rachel will know" over the last few years. Thank you for bringing your essential insight into the clinical side of this project, and hosting me down at the CRU. I owe each of you a huge debt of gratitude.

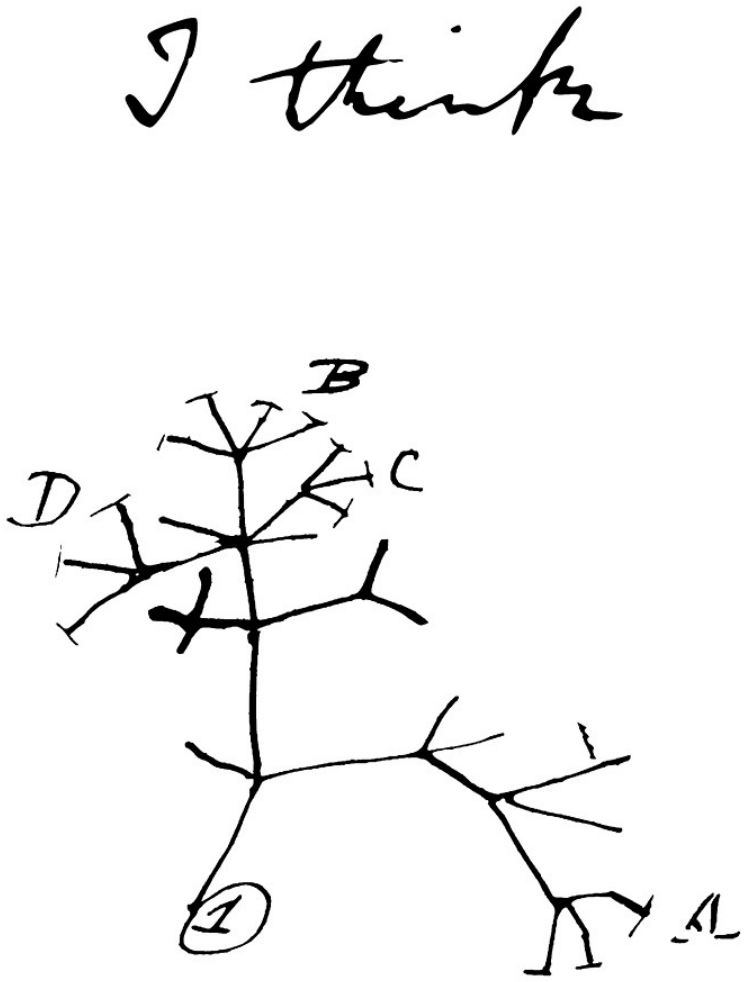
I would also like to extend my gratitude to Guy and Gregorio at the Cryptosporidium reference unit, and to the KESS team at Aberystwyth university for their consistent patience and support.

To my friends. Ross, the sheer volume of absolute rubbish we chat on a daily basis has kept me laughing for almost a decade. You are the best kind of friend. Ian, thanks for all the ddraig ddu and climbing trips to Gog. There is nothing like spending a day getting savaged on brutal Pass crimps to clear your mind of work. Lindsey, a chance meeting set up by Justin has lead to many beers and wines over the years. Thank you for being a consistent drinking partner and joining me in general lamentations, it's always nice to have someone else suffer with you. My housemates (particularly Tom, my landlord) have put up with my mood swings with great patience during this process. Your comradeship has been consistent and unfaltering, for which I am hugely grateful. To my office mates Vlad and Justyna, the office was the perfect working and complaining environment with you two in it, thank you for accepting that control of the window is mine and mine alone. To Claire, thank you for being a constant source of support to me over the last year, and for being so unfalteringly positive. Thank you to Matt, Cos, James, Richard, Swifty, Joe, and the rest of the AMC old guard. Thanks for the belays, spots, psyche, beer, and lifts over the years. And to everyone else not mentioned, cheers all, I owe you a pint.

To my Mother & Father, my sisters Jennie, Molly, Nina, Fi, and to my brothers Owen

and Digs. Despite having enough to fill this thesis, I am suddenly lost for words sufficient to thank you for everything you have done for me over thirty years. You have put up with my obsessions, and listened patiently to me ramble on about them since I was little, despite often being an unwilling audience! It was this careful nurturing of my keen interests which has lead me here. I dedicate all of this to each of you, which means you have to read it.

Finally, I feel a sense of great privilege to have been able to spend so much time climbing the rock of North Wales. I have been more elated, terrified, frustrated, satisfied, and exhausted climbing in Snowdonia than anywhere else. Diolch yn fawr iawn.



The salient points of this project are as follows:

- *Cryptosporidium* suffers from low DNA yield from clinical samples. This lack of DNA often results in WGS attempts producing genomes with very uneven depth of coverage. Due to the *Cryptosporidium* genome being small and repeat rich, misassembly was common using conventional and popular *de novo assemblers*, such as SPAdes and velvet, wherein highly uneven coverage over low complexity regions resulted in translocation of large chromosomal fragments. Consequently IDBA-UD, an assembler designed to assemble genomes with highly uneven depth of coverage, was used. This dramatically decreased the number of *in silico* translocation events that were observed during assembly. A novel pipeline utilising this assembler was developed to assemble and annotate 58 Cryptosporidium genomes.

- The Gini coefficient, along with a novel metric, Gini-granularity curves, can be used to characterise the distribution of reads across a genome assembly. These Gini-granularity curves demonstrate that WGA selectively amplifies regions of the *Cryptosporidium* genome in a biased manner.

- *Cryptosporidium* diagnostics is constrained by the convention of defining subtypes by variation in the highly repetitive region within the gene coding for a 60 kD surface glycoprotein on chromosome 6 (gp60). The single locus nature of this convention lacks the resolution of a multi locus approach. Furthermore, *Cryptosporidium* exhibits a sexual lifecyle, and has been reported to recombine at the gp60 locus. These observations may confound epidemiological research, which is reliant on subtyping using the gp60 locus. There is currently no consensus on which loci to interrogate for a MLST approach.

- Using a novel VNTR discovery and analysis pipeline (VaNTA), over 3000 polymorphic tandem repeats (VNTR's) have been identified. Around 300 of these compare favourably to gp60 in both their capacity to resolve subtype populations of *Cryptosporidium*, and the conservation of regions flanking the repeat locus, a feature essential for primer design.

- The advent of next generation sequencing has lead to an explosion in the amount of genomic data that can be produced by sequencing projects. However, this has lead to a significant bottle neck in the analysis of these data. Consequently, quick and reliable methods of analysing these data are critically important.

- BlooMine is a novel raw read mining tool developed to facilitate quick and computationally efficient local analysis of sequences captured within raw reads generated by whole or partial genome sequencing projects. It utilises Bloomfilters, a highly space efficient probabilistic data structure, to perform set membership queries and infer sequence homogeneity. Using this approach allows for analysis of genomic sequence without the need to carry out the memory and time consuming task of genome assembly.

- Using BlooMine, extensive multiplicity of infection was demonstrated within single-host clinical isolates of *Cryptosporidium parvum* across a number of VNTR loci identified using VaNTA. MOI-signature typing, a typing scheme utilising the relative abundance of alleles at a single locus within a host, appears to demonstrate superior typing resolution to conventional dominant allele schemes.

Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.

— Charles R. Darwin, *On the Origin of Species*, 1859

Contents

List of Figures

List of Tables

Glossary of Abbreviations

ABACAS algorithm based automatic contiguation of assembled sequences
AUC area under curve
bp base pairs (often accompanied by a binary prefix)
CDS coding sequence
CPU central processing unit
CRU public health Wales Cryptosporidium reference unit
EC enzyme commission number
EMBL european molecular biology laboratory
EMBOSS european molecular biology open software suite
FITC Fluorescein isothiocyanate
FTP file transfer protocol
GATK genome analysis toolkit
GO gene ontology
HIV/AIDS human immunodeficiency virus/acquired immunodeficiency syndrome
HPC high performance computing
ICORN iterative correction of reference nucleotides
IDBA-UD iterative de Bruijn graph de novo assembler for uneven depth
IGV integrative genomics viewer
IMAGE iterative mapping and assembly for gap extension
IMS immuno-magnetic separation
indel insertion/deletion event
IO input/output
LD linkage disequilibrium
MDA multiple displacement amplification
MOI multiplicity of infection
MLFT multi-locus fragment typing
MLST multi-locus sequence typing
nAUC normalised area under curve
NGS next generation sequencing
NIAID national institute of allergy and infectious diseases
ORF open reading frame

PAGIT post-assembly genome improvement toolkit
PCR polymerase chain reaction
PFAM protein family
PV parasitophorous vacuole
RATT rapid annotation transfer tool
rDNA ribosomal deoxyribonucleic acid
RAM random access memory
RFLP restriction fragment length polymorphism
RNAseq ribonucleic acid sequencing
SCID severe combined immunodeficiency
SCOP structural classification of proteins
SNP single nucleotide polymorphism
SSAHA sequence search and alignment by hashing algorithm
SSR short sequence repeat
TCA tricarboxylic acid
TR tandem repeat
TRF tandem repeats finder
VaNTA variable number tandem repeat analysis pipeline
VNTR variable number tandem repeat
WGA whole genome amplification
WGS whole genome sequencing

Glossary of Terms

Alignment the arrangement of nucleotide or protein sequences in such a way that regions of similarity are identified for the purpose of comparison
Amplicon a fragment of DNA or RNA which is the product of amplification
Annotation the elucidation of the position of genomic features across a genome assembly
Apicomplexa a Phylum of protozoans, containing many medically important parasitic genera
Assembly the construction of large, contiguous nucleotide or protein sequences from smaller fragments (read)
Biofilm a slimy extracellular matrix composing of polymers and embedded microorganisms
Bloom Filter a probabilistic data structure used to perform memory efficient set intersections queries
Count-Min Sketch a probabilistic data structure capable of representing, and answering queries of, a higher dimensional vector
Entropy (Shannon) a measure of the redundancy within a string, and by extension a measure of its informational content
Epimerete a feeding organelle of some apicomplexan life stages
FASTA a text-based format widely used to represent nucleotide and protein sequences
FASTQ a text-based format widely used to store both biological sequences and their quality scores
Genome the set of all the genetic information within an organism
Genotype the genetic constitution of an organism
Gini coefficient a measure of the inequality of a dataset
Gregarine a group of apicomplexan protozoans
Hashing a method of taking a string and representing it with an integer value by application of a hashing function
Kmer-array the set of all kmers within a genetic sequence
Kmer-spectrum a visual representation of a kmer array
Linkage Disequilibrium a measure of the association of alleles at different loci in a given population

Locus a position on a chromosome or genome
Mapping a process whereby NGS reads are aligned to a reference genome
{macro/micro}gamete sexual stages of apicomplexan parasites
Merozoite a life stage of some apicomplexans, which is involved with host cell invasion and proliferation
Meront a life stage of some apicomplexans in which schizogony occurs, producing merozoites
Multiplicity of Infection a state whereby a host is infected with multiple discrete genetic populations of a pathogen
Myzocytosis the method of nutrient extraction by apicomplexans, using the epimerete
N-space the space produced by scaffolding across regions which bear no sequence, and are therefore replaced with a sequence of N's
Oocyst the infective stage of some protozoan parasites, which is often passed in the faeces
Parasitophorous Vacuole a structure produced within a host cell by apicomplexans for the purpose of proliferation and nutrient acquisition
Polymorphism the state of a locus presenting with several different sequences
Recombination the rearrangement of a genetic sequence
Scaffolding the method of arranging assembled contigs into larger "scaffolds", usually using a reference genome
Schizogony a process found in some apicomplexans, referring to asexual reproduction by multiple fission
Syzygy the asexual exchange of genetic material by two associated protozoa
Sporozoite a life stage of some apicomplexans, which typically infiltrates host cells
Trophozoite a life stage of some protozoan parasites, which acquire nutrients from the host
Virulence the capacity of a pathogen to induce disease in a host

Chapter 1

Introduction

Protozoans belonging to the phylum Apicomplexa are of significant importance to medical and veterinary health worldwide, containing medically important genera such as *Toxoplasma*, *Babesia*, *Cryptosporidium*, and the Malaria causing parasite, *Plasmodium*. The global burden of parasitoses caused by Apicomplexans is considerable, with roughly 500,000 deaths caused by members of genus *Plasmodium* annually, between 20-50% of the world's population testing seropositive for *Toxoplasma gondii*, and over 200,000 infant deaths being attributed to *Cryptosporidium* in Asia and Sub-Saharan Africa alone each year. The huge toll on human life these parasites take necessitates the urgent development of preventative and therapeutic strategies to combat this important issue in global public health. High-throughput generation of parasite genomes has allowed for the most thorough analysis of the biology of these parasites currently possible, leading to huge advancements in our understanding of the parasites, the diseases they cause, and facilitates the development of novel therapeutics, and biomarkers for highly sensitive and reliable species and genotype identification. However, there lies a significant bottle neck in this process, wherein the generation of data has progressed at a quicker rate than methods to analyse it. In 2010, Elaine Mardis, the co-executive director of the Institute for Genomic Medicine at Nationwide Children's Hospital in Columbus Ohio, quipped on the human genome: "it's $1000 genome and a $100,000 analysis" [Mardis, 2010]. It is essential that this bottle neck is alleviated by using modern computing techniques and systems to develop methods which allow for high-throughput analysis of the enormous amount of genomic information which can be generated during genome sequencing projects.

This project aims to utilise computational genome mining techniques on next generation sequencing data to generate novel diagnostic and genotyping methods for *Cryptosporidium spp.* The data produced using the methods and tools developed during this project will aid analysis of transmission of cryptosporidiosis by furnishing epidemiological investigations carried out by the national Cryptosporidium Reference Unit, hosted by Public Health Wales.

Genome analysis will necessitate the assembly and annotation of genomes using the most recent tools and pipelines in bioinformatics. The primary aims of this project are:

- Expand the current database of assembled and annotated *Cryptosporidium* genomes available to facilitate genome analysis. This will be detailed in Chapter 2.
- Develop novel approaches to identify conserved regions within the genome of *Cryptosporidium* species and subtypes to direct species identification within clinical samples or metagenomic datasets, and new genotyping paradigms, to be used by the Cryptosporidium Reference Unit. This will be detailed in Chapter 3.
- Utilise alignment-free sequence analysis techniques which can be applied to reads to increase the speed and reduce the computational power required to analyse genomes for these genomic marker sequences. This will allow such genome mining techniques to be carried out on laptops and workstation computers, rather than HPC (High Performance Computing) clusters necessary for genome assembly. This will be detailed in Chapters 4 and 5.

This project will attempt to use comparative genomics using the methodologies developed and the NGS data generated to expand our understanding of the population genomics of genus *Cryptosporidium*, which will in turn elucidate transmission patterns, allowing for more accurate tracking of the parasite within a population. The data will be used to establish mutagenicity of loci within the genome of *Cryptosporidium* which can be used to follow transmission of the parasite back to its source, directing prevention strategies.

This chapter is split into five sections, evaluating and discussing the literature and examining the cutting edge in each section:

1. The biology, history, clinical importance, and methods of diagnosis of genus *Cryptosporidium*, with particular focus on genetic differentiation of discrete genotypes and subtypes.
2. Genome sequencing, with particular emphasis on methods employed to generate high quality complete genomes from Next Generation Sequencing (NGS) data.
3. Multiplicity of infection (MOI), approaches used to identify in-host pathogen populations and the impact it has on public health.
4. Alignment-free sequence analysis, different approaches and how the can be applied to alleviate the data analysis bottle neck.

5. Data-structures which can be used in tandem with Alignment-free sequence analysis.

1.1 Genus: *Cryptosporidium*

Genus *Cryptosporidium* is a group of parasitic protozoans within the Apicomplexa belonging to the novel clade Cryptogregorina. They can most accurately be described as facultative epicellular Apicomplexans. They are pathogens of a wide range of vertebrates, causing a mild to significant gastrointestinal condition referred to as cryptosporidiosis. The parasite has been detected in humans worldwide, but is generally more common in areas where sanitation is poor, with improper treatment of sewage and drinking water or where human faeces is used as a fertiliser ('nightsoil'). Cryptosporidiosis is exacerbated in many developing countries by the lack of available health care and malnutrition; factors which are known to effect the severity of the disease. However, the problems associated with treating water for the presence of *Cryptosporidium* oocysts as well as the proliferation of *C. parvum* in ruminants, allowing them to act as a reservoir host, means countries such as the UK, Ireland, Turkey and Mexico have ongoing problems with local infection.

1.1.1 The Life Cycle of *Cryptosporidium*

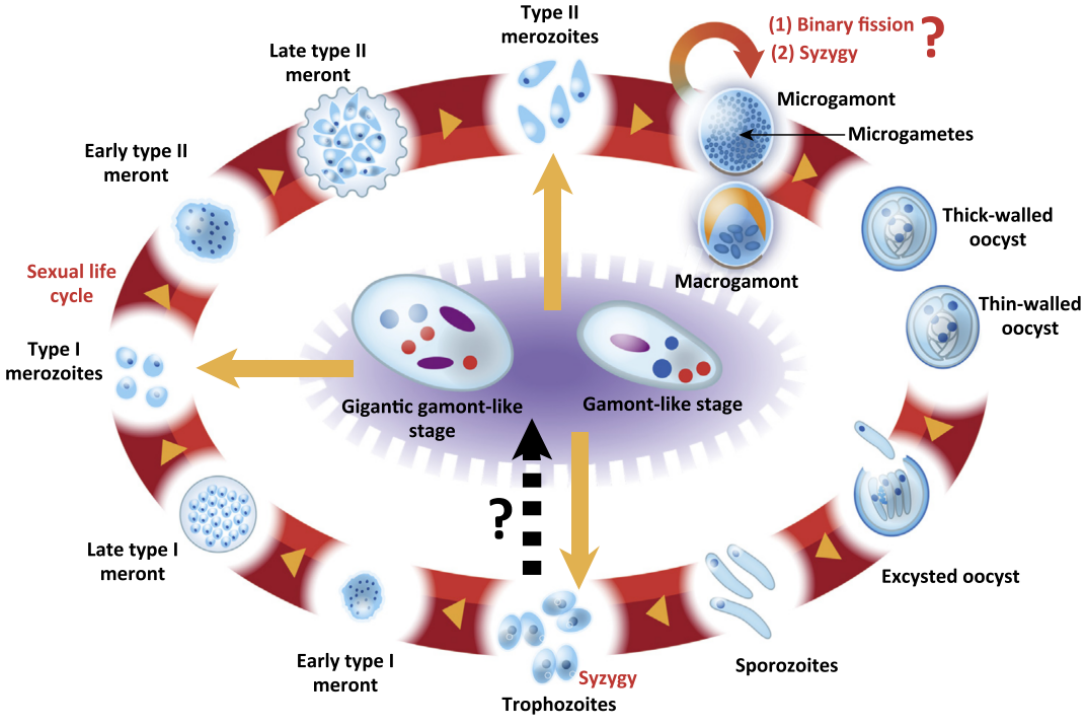


Figure 1.1: The life cycle of *C. parvum* with hypothesised roles and orgins of the gamont-like stages. Taken from Clode *et al.* [Clode et al., 2015].

Transmission of *Cryptosporidium* and its status as a parasite with outbreak potential is due almost entirely to its life cycle, and peculiarities that allow it to persist outside of a host. These observations are of paramount importance in the development of strategies to tackle cryptosporidiosis. The life cycle of *Cryptosporidium* (detailed in Figure 1.1) remains consistent with the standard characteristics of an apicomplexan life cycle; i.e. involves an infective sporozoite stage, merogony, and formation of a zygote by sexual union between gamete cells. Members of genus *Cryptosporidium* complete their full life cycle in one host. However, there are two putative life cycles, accounting for the extracellular, and epicellular observations of *Cryptosporidium*.

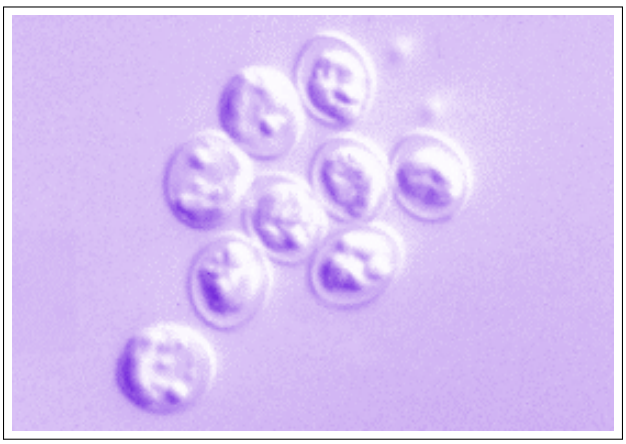


Figure 1.2: *Cryptosporidium parvum* oocysts visualised by light microscopy. Image taken from the Public Health Wales Cryptosporidium information page.

After ingestion of the oocyst (see Figure 1.2) by a suitable host, the oocyst adheres to the mucous lining of the ileum via surface lectins, whereupon excystation occurs to release four sporozoites. These sporozoites then discharge their apicle and adhere to the ileal mucous lining, mediated by mucin-like surface receptors expressed on the apical organelle (such as gp900 and gp60). The sporozoites release enzymes, breaking down the mucous lining and allowing for contact with, and infiltration of, the ileal epithelial cells mediated by receptor-ligand interaction (including Cpa135, TRAP-C1, CS, CP47 and CP12) facilitating gliding motility [Wetzel et al., 2005]. This interaction induces host-cell membrane protrusion, encapsulating the sporozoite in a parasitophorous vacuole (PV) underneath the apical membrane of the host-cell [Valigurová et al., 2008]. Within this PV, the sporozoites mature into trophozoites; the feeding form. Ultrastructural studies have shown that the PV forms radial folds at the parasite-cell interface [Valigurová et al., 2007], generating a feeder organelle referred to as an epimerete. It is this organelle that allows for extraction of nutrients from the host cell via Myzocytosis. These trophozoites subsequently undergo merogony to form type I meronts, bearing many type I merozoites. The type I merozoites are then released, forming into type II merozoites containing a number of type II merozoites, which in turn are released to undergo transformation into

the sexual cells; the microgamete (male) and the macrogamete (female). The fertilisation of a macrogamete by a microgamete forms a single thick walled oocyst, which is passed out of the host. The extracellular stages involve formation of a vacuoles around the sporozoite in the ileal lumen, facilitating transformation into the trophozoite stage. During this extracellular life cycle, trophozoites may undergo syzygy; a pairing of two trophozoite stages necessary in the formation of the enigmatic gamont-like stages (although this is a hypothetical origin), a characteristic of gregarines (in which these stages are referred to as gametocysts). The function of *Cryptosporidium* syzygy has not been conclusively elucidated, though it has been hypothesised that it it has a role in the generation of more trophozoites or merozoites, facilitating oocyst production in host-cell free environments, such as in aquatic biofilms [Borowski et al., 2008, Clode et al., 2015, Koh et al., 2013, Koh et al., 2014].

1.1.2 A Historical Perspective on *Cryptosporidium*

Cryptosporidium was first described by Tyzzer in 1907, whereupon the atypical nature of this coccidian was noted. The attachment organelles, lack of sporocysts and presumed life-cycle were of particular note and value in its taxonomic classification. The genus name was given to signify the presence of naked sporocysts (*Cryptosporidium*: literally 'hidden spores'). The morphology of the oocyst, presence of attachment organelles, the extracytoplasmic nature of the parasite, the observation that unshed oocysts can elicit autoinfection within the host and (later noted) their complete insensitivity to anticoccidial drugs made *Cryptosporidium* an atypical example of a coccidian. It was consequently placed in family Asporocystidae [Tyzzer, 1907]. Traditionally the characteristics used to identify and name novel species belonging to the Apicomplexa has been host specificity, location of endogenous life stages and morphology of the various life stages. Tyzzer applied these in forming the species nomenclature of *Cryptosporidium*, which lead to him describing *C. muris*, representing its identification within the gastric glands of mice. The study of *Cryptosporidium* following its description was considered that of curious endeavour rather than being driven by medical necessity. As is so common in classical protozoology, the convention of describing and naming new species based on the host in which they were found persisted, leading to the description of *C. agni* in sheep [Barker and Carbonell, 1974], *C. bovis* in cattle and *C. garnhami* or *C. enteritidis* in humans [Fayer, 2010]. However, because of the lack of taxonomic data, many of these novel species could not be clearly distinguished from others leading to their names becoming invalid. It became quickly clear that morphological characterisation alone was not enough to distinguish the species within genus *Cryptosporidium*, beyond observations on oocyst size, due to their extreme similarity. Because of this, from the 1970's through much of the 1990's, species distinction was based entirely on their bearing 'large' or 'small' oocysts,

and consequently only two species were consistently described; *C. muris* and *C. parvum*, bearing large and small oocysts respectively. It was generally considered that a single species could be found in the gastric glands of mice, as described by Tyzzer (*C. muris*), and that all other mammals exhibited *C. parvum*. It was not until the emergence of HIV/AIDS that *Cryptosporidium* became of medical importance, due to its pathogenicity among immunosuppressed individuals. A review by Casemore *et al.* reported 71 cases of *Cryptosporidium* in immunosuppressed individuals, and an association was established between bovine and human infection for the first time [Casemore et al., 1985]. Casemore and Jackson also proposed a hypothesis in which human cryptosporidiosis is not entirely a zoonosis, which lead to the elucidation of two distinct transmission cycles [Casemore and Jackson, 1984]. This was confirmed using molecular methodologies which eventually lead to the description of the human species; *C. hominis* as distinct from *C. parvum* [Morgan-Ryan et al., 2002]. The advent of the use of genetic and molecular data in species identification lead to the reinterpretation of the genus and its constituents, leading to *C. parvum* being split into a number of different distinct species [Nader et al., 2019].

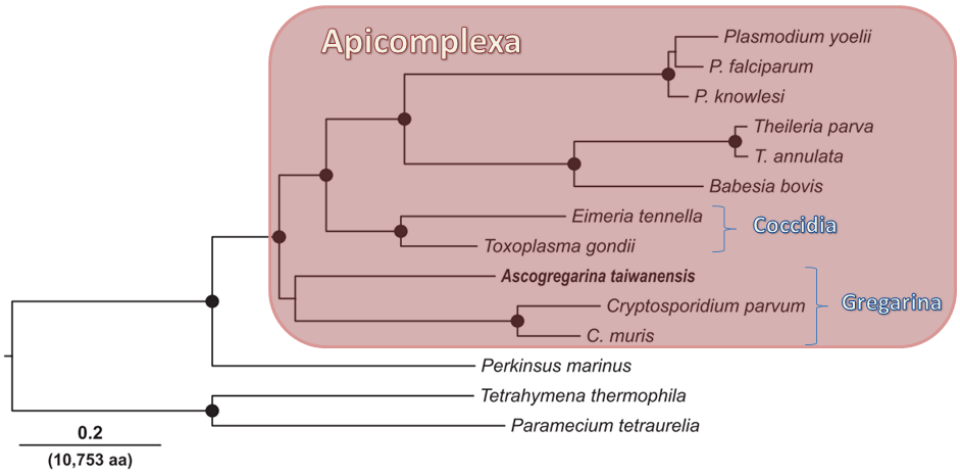


Figure 1.3: Phylogenetic relationships of the alveolates inferred from concatenated protien sequences containing 10,753 amino acid positions. Tree taken from Templeton *et al.* [Templeton et al., 2010]

Reservations in the placement of *Cryptosporidium* amongst the coccidians began to show when Bull & Chalmers *et al.* observed cross-reactivity of anti-*Cryptosporidium* monoclonal antibodies with sporocysts of the gregarine *Monocystis* from faecal specimens of small mammals in the UK [Bull et al., 1998]. Further evidence that all was not as previously thought with the membership of *Cryptosporidium* as a coccidian arose when advances in our understanding of its life cycle and morphology suggested that the may bear closer ancestry with the gregarines. Support for this was presented when molecular studies consistently illustrated a separation between the coccidia and *Cryptosporidium* [Morrison and Ellis, 1997, Carreno et al., 1998] and their position within a clade populated by the gregarines [Carreno et al., 1999, Leander et al., 2003]. Gregarine-like gamonts,

syzygy and spores bearing eight sprozoites which closely resembled those in gregarines were demonstrated in *Cryptosporidium* using ultrastructural analysis of the extracellular stages [Rosales et al., 2005]. Genetic studies have contributed a great deal of data in the elucidation of the confusing world of *Cryptosporidium* taxonomy, illustrating a disparity between codon usage in *C. parvum* and the Eimeriorines; *Toxoplasma gondii* and *Emeria tenella*, but a surprisingly high parity between that and other members of phylum Apicomplexa; *Plasmodium falciparum*, *Babesia bovis* and *Theileria parva*, as well as to the amoeba; *Entamoeba histolytica* [Char et al., 1996]. The largely intronless structure of the *C. parvum* genome also provided support for the evolutionary distinction between this genus and other members of the coccidia clade [Spano and Crisanti, 2000]. In 2014, a large scale revision of the Sporozoans by 18S rDNA analysis by Cavalier-Smith placed *Cryptosporidium* in a novel subclass; Orthogregarina, along with protozoans which are closely related to this genus (see Figure 1.3) [Cavalier-Smith, 2014]. In 2016 Ryan *et al.* discussed the re-appraisal of *Cryptosporidium* as a gregarine and the supporting evidence and its implications on the water industry [Ryan et al., 2016].

1.1.3 Cryptosporidiosis

Cryptosporidiosis is a disease of humans affecting the small intestine, leading to acute gastroenteritis. There are approximately 6000 new cases of human cryptosporidiosis reported every year in the UK, with focal points of outbreaks clustering around April and August for *C. parvum* and *C. hominis* respectively. It is a significant cause of diarrhoeal disease globally, with potentially dire consequences for young children in the third-world, where access to healthcare is poor and malnourishment rife, and immunosuppressed individuals. At least ten of the currently identified *Cryptosporidium* species have been detected in humans, with 7 species (*C. hominis, C. parvum, C. meleagridis, C. felis, C. viatorum* and *C. cuniculus*) being capable of causing disease in humans. It bears a faecal-oral route of infection which necessitates contact with the infectious agent, including an infected animal, waters (such as swimming pools and freshwater lakes or rivers), and ingestion of contaminated food or water. Contaminated water supplies constitute a major infective risk, leading to the potential for epidemics to arise. Koh *et al.* have documented the potential for *Cryptosporidium* to reside suspended in biofilms, whereby oocysts may be trapped and reintroduced into a water supply as well as acting as a nutrient rich medium within which the parasites may undergo an extracellular lifecycle and proliferate [Koh et al., 2013, Koh et al., 2014], though this has yet to be demonstrated in natural water pipes. This potential for large scale outbreaks of cryptosporidiosis makes it a public health risk of significance outweighing that implied by the small number of cases relative to other gastroenteritis' such as *Salmonella*, *Campylobacter* and *E. coli*. This is illustrated by the Milwaukee outbreak of 1994, where 403,000 people fell ill with

cryptosporidiosis, with 19 mortalities. The source was tracked to a contaminated water supply [Eisenberg et al., 2005]. Person to person infection remains a significant risk in areas where sanitation and hygiene are lacking.

The clinical presentation of cryptosporidiosis varies depending on the hosts immunocompetence and nutrition. In both cases there is diarrhoea, with the majority of immunocompetent people experiencing abdominal pain and vomiting and around half suffering from a fever. In otherwise healthy individuals, cryptosporidiosis is self limiting, usually resolving within two weeks. In immunocompromised individuals the disease is often severe, with morbidity and mortality greatly increased. Sufferers of T-cell deficiency are in the highest risk category, such as those with haematological cancers, T-cell deficiencies such as CD40 ligand deficiency and severe combined immunodeficiency (SCID), and late stage HIV patients with low CD4 lymphocyte counts. In these cases the diarrhoea is more severe, often transient and chronic in a similar fashion to cholera resulting in severe weight loss. The disease may involve the biliary tract and respiratory system, which is potentially intractable and of devastating consequence. Damage to the gastrointestinal system and bronchial tree may be extensive, with dilation of the bile duct and gall bladder, and generalised inflammation presenting as pancreatitis, cholecystitis and sclerosing cholangitis. Recrudescence of cryptosporidiosis is a possibility in these cases, as the parasite may inhabit the billiary ducts through treatment and re-emerge to cause clinical disease. With this in mind, correct treatment regimes is of paramount importance of effective clearance of the parasite. A rare complication of cryptosporidiosis in late stage HIV patients is pneumatosis cystoides intestinalise, in which gas containing cysts form within the gut wall and subsequently rupture, causing pneumoretroperitonium and pneumomediastinum [Hunter and Nichols, 2002, Sivarajah et al., 2013]. In severely affected immunocompromised individuals, standard drug treatment in the US (Nitazoxonide administration) has been shown to lack effective efficacy, which limits the clinician to attempting to improve the hosts immune condition to combat infection, an approach which has seen success. In late stage HIV patients, this takes the form of aggressive antiretroviral therapy [Chalmers and Davies, 2010]. Due to the problems associated with the treatment of the most severe cases of cryptosporidiosis, emphasis has been put on prevention, which necessitates close monitoring of incidents and data collection on the spread of the disease, as well as consistent monitoring of drinking water sources for the presence of *Cryptosporidium* oocysts.

1.1.4 Cryptosporidium Genomics: the Journey to the Genome

1.1.4.1 Sequencing the *Cryptosporidium* Genome

In the late nineties, the importance of *C. parvum* and *C. hominis* in public health lead the National Institute of Allergy and Infectious Diseases (NIAID) to fund a consortium comprising of the University of Minnesota, Virginia Commonwealth University and Tufts University to sequence the genome of these human-infective species. Molecular tools (Restriction Fragment Length Polymorphism and other genotyping methods) illustrated that human infection was caused primarily by two genotypes [Widmer and Sullivan, 2012]. Since then, the generation of whole genome sequences of members of genus *Cryptosporidium* has become more streamlined and faster due to a significant progression in the development of DNA sequencing tools and methodologies.

Attempts to sequence the genome of *Cryptosporidium* began in the early 2000's. Initial attempts involved cloning sheared fragments into plasmid vectors and Sanger sequencing. This approach resulted in > 9x coverage of the genome and yielded a fragmented assembly of 221 contigs of length > 5kbp. A more advanced sequencing project was undertaken to resolve gaps, using large *C. parvum* fragments contained within lambda DASH II libraries, and sequence missing DNA using a primer walk strategy [Widmer et al., 2002]. The completed genome of *C. parvum* (IowaII) along with a preliminary annotation was first published in 2004 by Abrahamsen *et al.* by passaging oocysts through an animal donor to produce enough parasitic material for the extraction and purification of sufficient amounts of DNA. A random shotgun sequencing approach was used, which yielded a complete genome with coverage of 13x over 18 large contigs [Abrahamsen et al., 2004]. This was shortly followed by the publication of the first draft genome of *C. hominis* (TU502) in late 2004. However, this genome proved to be much more fragmented than that of *C. parvum*, resulting in a sequence consisting of 1422 contigs [Xu et al., 2004]. In 2015, the *C. parvum* (IowaII) reference genome was reannotated, and a new *C. hominis* reference genome (UdeA01) published. This reannotation effort increased the number of putative genes from 3807 to 3865 for *C. parvum* IowaII, and predicted the presence of 3819 genes in *C. hominis* UdeA01 [Isaza et al., 2015]. In 2016, Ifeonu *et al.* reassembled and reannotated the *C. hominis* TU502 genome, along with producing new draft genomes of human isolated *C. hominis* (UKH1) and *C. meliagridis* (UKMEL1) along with the avian species *C. baileyi* (TAMU-09Q1). The *C. hominis* TU502 genome proved to be a considerable improvement on the previous 2004 version, being much more complete, and reducing the number of contigs down to 119. Annotation was facilitated by the RNAseq data generated from the oocyst stage of both *C. hominis* & *C. baileyi*, predicting the presence of 3745 protein coding genes in *C. hominis* TU502 and 3765 in *C. hominis* UKH1 [Ifeonu et al., 2016].

***Cryptosporidium* species**	**Mean oocyst dimensions (μm)**	**Major host(s)**	**Infections reported in humans**	**Genomes available (Accession no or Reference)**
C. alticolis	5.4×4.9	Voles	No	No
C. apodemi	4.2×4.0	Mice	No	No
C. andersoni	7.4×5.5	Cattle	Yes (rarely)	PRJNA354069
C. avium	6.3×4.9	Birds	No	No
C. baileyi	6.2×4.6	Birds	No	PRJNA222835
C. bovis	4.9×4.6	Cattle	Yes (rarely)	No
C. canis	5.0×4.7	Canids	Yes (occasionally)	No
C. cuniculus	5.6×5.4	Lagomorphs, Humans	Yes (occasionally)	PRJNA315496
C. ditrichi	4.7×4.2	Mice	Yes (rarely)	No
C. ducismarci	5.0×4.8	Tortoises	No	No
C. erinacei	4.9×4.4	Hedgehogs	Yes (rarely)	No
C. fayeri	4.9×4.3	Marsupials	Yes (rarely)	No
C. felis	4.6×4.0	Felids	Yes (occasionally)	No
C. fragile	6.2×5.5	Toads	No	No
C. galli	8.3×6.3	Birds	No	No
C. homai	Data not available	Guinea Pigs	No	No
C. hominis	4.9×5.2	Humans	Yes (commonly)	PRJEB10000 PRJNA13200 PRJNA252787 PRJNA222836 PRJNA222837 PRJNA307563 PRJNA253838 PRJNA253839 PRJNA253834
C. huwi	4.6×4.4	Fish	No	No
C. macropodum	5.4×4.9	Marsupials	No	No
C. meleagridis	5.2×4.6	Birds, mammals	Yes (occasionally)	PRJNA222838 PRJNA315503 PRJNA315502
C. microti	4.3×4.1	Voles	No	No
C. molnari	4.7×4.5	Fish	No	No
C. muris	7.0×5.0	Rodents	Yes (rarely)	PRJNA32283 PRJNA19553
C. occultus	5.2×4.9	Rodents	Yes (rarely)	No
C. parvum	5.0×4.5	Mammals	Yes (commonly)	PRJNA144 PRJNA320419 PRJNA439211 PRJNA253848 PRJNA253843 PRJNA253845 PRJNA253836 PRJNA253840 PRJNA253846 PRJNA253847 PRJNA320419 PRJNA315506 PRJNA437480 PRJNA315504 PRJNA315508 PRJNA315507 PRJNA315505 PRJNA13873
C. proliferans	7.7×5.3	Rodents, maybe Equids	No	No
C. proventriculi	7.4×5.7	Birds	No	No
C. rubeyi	4.7×4.3	Squirrels	No	No
C. ryanae	3.7×3.2	Cattle	No	No
C. scrofarum	5.2×4.8	Pigs	Yes (rarely)	No
C. serpentis	6.2×5.3	Reptiles	No	No
C. suis	4.6×4.2	Pigs	Yes (rarely)	No
C. testudinis	6.4×5.9	Tortoises	No	No
C. tyzzeri	4.6×4.2	Rodents	Yes (rarely)	No

Table 1.1: The currently discovered and documented human and animal Cryptosporidium species [Morris et al., 2019c].

The extraction of DNA from *Cryptosporidium* has long been a confounding issue for biologists, since the DNA yield from oocysts is low. Consequently, in order to obtain sufficient amounts of genomic DNA, a great number of oocysts must be extracted and cleaned prior to DNA extraction, or the DNA enriched. This has lead to attempts to culture oocysts *in vitro*, obviating the need for a host, which has seen little success. In addition, it was not until 2018 that a method for storing *Cryptosporidium* infective material in cryostatic conditions was developed. This method involved vitrification (ultra-fast cooling) of *Cryptosporidium* oocysts in a microcapillary, which exhibit high viability, and infectivity to interferon-γ-mice after several weeks [Jaskiewicz et al., 2018]. However, this has not been verified for longer periods of storage, and has only been demonstrated as applicable in very small quantities of oocysts, making it currently unviable for large-scale application. There are consequently no methods which allow for indefinite storage of large quantities of oocysts, which means there is a requirement for the serial passaging of isolates through host animals to propagate the parasite. Due to the sexual nature of the *Cryptosporidium* life cycle and a known propensity for certain loci (notably the gp60/cgd6_1080 locus) to recombine, this may lead to a destabilisation and variation of the subtype, a potential confounding variable if these isolates are to be used for genomic study. Hijjawi *et al.* reported success in long-term culturing (up to 25 days) of *Cryptosporidium* in cell culture using pH modification sub-culturing and gamma irradiation [Hijjawi et al., 2001]. Complete development of *Cryptosporidium in vitro* was reported by Hijjawi *et al.* in 2004, where new oocysts were reportedly present after 8 days after being inoculated into a culture devoid of host cells. This represents the first reported case in which *Cryptosporidium* has been shown to multiply, develop and complete its life cycles without the need to infiltrate and infect host cells [Hijjawi et al., 2004]. However, these results have yet to be verified independently or applied on any significant scale to aid research in this area. In 2018, a new method of culturing *Cryptosporidium* oocysts *in vitro* was detailed by Miller *et al.* This method involved the infection of COLO-680N cells, which proved a sufficient method to yield enough oocysts to inoculate further COLO-680N cell lines. The authors reported an increase in the number of oocysts from the initial inoculation, verified by microscopy and molecular techniques [Miller et al., 2018]. Despite these results, this method has also resisted attempts at independent verification.

The development of a verified, scalable, and entirely *in vitro* technique for the culture of *Cryptosporidium* oocysts would represent a major step in *Cryptosporidium* research. Such a method would allow for the application of modern molecular and cell-biology techniques, accelerating research into host-parasite interaction, virulence, and the development of novel therapeutic agents, laying the foundation for a new era of *Cryptosporidium* research.

The problem associated with DNA extraction of *Cryptosporidium* was addressed by Hawash (2014) who observed this problem in comparison to DNA recovery of cysts of other parasitic protozoans (*Entamoeba histolytica* and *Giardia sp.*) isolated from stool samples. He modified the QIAmp DNA Stool Mini Kit (Qiagen) protocol, leading to an increased sensitivity and specificity from 60% to 100% of samples testing positive for the presence of *Cryptosporidium* DNA [Hawash, 2014]. In 2015, Hadfield *et al.* developed a novel method of Immunomagnetic Seperation (IMS) of oocysts obtained from faecal samples which dramatically increased the quality of *Cryptosporidium* DNA obtained from stool samples, reducing the amount of contaminant DNA [Hadfield et al., 2015]. This allowed for extensive comparative genomic analysis of clinical isolates, which was not previously possible. The problems associated with poor DNA yield from clinical samples were tackled in 2015 by Guo *et al.*, who used whole genome amplification (WGA) to enrich Cryptosporidium DNA from 6 discrete species/genotypes extracted from 24 faecal samples from humans and animals. The results were encouraging, showing that *Cryptosporidium* DNA was significantly enriched, allowing for coverage of $> 94\%$ of the *Cryptosporidium* genome [Guo et al., 2015a].

Recently, there have been attempts to generate sequences using long read technology, such as MinION by Oxford Nanopore, and Pacific Biosciences. There exist a few draft genomes (published and unpublished) from long reads generated by PacBio. In 2018, Gilchrist *et al.* explored the genetic diversity of *C. hominis* using long reads generated by PacBio sequencing [Gilchrist et al., 2018]. The Kissinger Research group at Georgia university are also attempting to produce a high quality PacBio assembly, and the Welcome Trust, Sanger Institute generated a set of PacBio reads which are as yet unpublished. Currently, there have been no successful attempts at sequencing the genome using the MinIon platform. This is likely due to the large amount of DNA required to generate such reads using this technology, which is a known difficulty associated with *Cryptosporidium* genome sequencing.

With the advent of these novel methodologies to generate high quality genomes from clinical samples, there comes an opportunity to greatly expand the number of genomes available. The online *Cryptosporidium* genomics resource http://cryptodb.org/cryptodb/, provdes access and analytical tools to *C. hominis* TU502 and UdeA01 and other *C. hominis* genomes, *C. parvum* Iowa II and human-adapted *C. parvum*, other zoonotic species including *C. meleagridis*, and host-adapted species rarely found in humans (*C. muris*, *C. andersoni*, *C. baileyi*, and *C. tyzzeri*). To date, genome assemblies from 39 isolates of human and animal *Cryptosporidium* and are available on GenBank in various states of assembly and annotation (see Table 1.1). 68 paired end read libraries are also available

with 62 of these having been produced by The Welcome Trust Sanger Institute during a pilot study sequencing *Giardia* and *Cryptosporidium* (Accession: PRJEB3213).

1.1.4.2 The Cryptosporidium Genome

	C. parvum Iowa II	*C. hominis* UdeA01
Number of genes	3865	3819
Mean gene length (bp) excluding introns	1783	1784
Percent coding	75.7	75.4
Genes with introns (%)	10.8	10.9
exons		
Number	4553	4503
Mean length (bp)	1514	1514
Mean number per gene*	2.6	2.6
G + C content (%)	32	32
introns		
Number	688	684
Mean length (bp)	99	100
G + C content (%)	22	22
Intergenic regions		
G + C content (%)	25	25.2
Mean length (bp)	552	563
tRNAs	45	45
Annotation		
Genes with EC code	502	498
Number of unique EC code	207	206
Genes with GO terms	2030	2026
Number pf unique GO terms	967	968
Numner of unique PFAM domains	1969	1982
Number of unique SCOP superfamilies	1667	1677

Table 1.2: Summary of general statistics for the *C. parvum* Iowa II and *C. hominis* UdeA01 reference genomes taken from Isaza *et al.* 2015. *C. parvum* and *C. hominis* comprise the vast majority of cases of human cryptosporidiosis, and therefore are considered the most important species in global public health.

The genome of *Cryptosporidium parvum* (according to the Iowa II reference genome [Abrahamsen et al., 2004]) and *Cryptosporidium hominis* (according to the UdeA01 reference genome [Xu et al., 2004]) are approximately 9.1 mbp in size, arranged into 8 chromosomes, and with a genomic GC content of roughly 30%. *C. parvum* Iowa II exhibits 3,865 genes with a mean length of 1783 (excluding introns), of which 75% are coding. *C. hominis* UdeA01 exhibits 3,819 genes with a mean length of 1784 (excluding introns) of which 75.7% are coding. The dramatic absence of introns within the genome of *Cryptosporidium* is of particular note, with only 10.9% of genes bearing introns in the Iowa II *C. parvum* reference genome, and 10.9% in *C. hominis* UdeA01 [Isaza et al., 2015].

As can be seen in Table 1.2, there is little difference between the genomes of *C. parvum* and *C. hominis*. They exhibit 95-97% DNA sequence identity; with 11 protein-coding sequences identified only in *C. hominis* and 5 in *C. parvum*, and no large indels or rearrangements apparent [Widmer and Sullivan, 2012]. *Cryptosporidium* has, like the other gregarines, lost its apicoplast. *C. parvum* and *C. hominis* both exhibit a degenerate 'mitosome' in place of mitochondrion and consequently lack a mitochondrial genome and nuclear genes for many mitochondrial proteins, including but not limited to those required for the tricarboxylic acid (TCA) cycle, oxidative phosphorylation and fatty acid oxidation [Ryan and Hijjawi, 2015]. Genomes released by Abrahamson *et al.* and Xu *et al.* indicate that the genes for *de novo* biosynthesis of amino acids, nucleotides, sugars and mechanisms for gene silencing and RNA splicing are all absent [Abrahamsen et al., 2004, Xu et al., 2004]. It appears to be this loss of genes associated with metabolic biosynthesis that influences the parasitic nature of *C. parvum* and *C. hominis*, necessitating their reliance on acquiring nutrients from the host. The high conservation in the *C. hominis* genomes generated from European samples compared to the much more polymorphic *C. parvum* does not appear to be expressed in general observations on structure and base representation as illustrated in Table 1.2, suggesting that the variation is primarily exhibited within the sequence itself. This further illustrates the importance of large scale sequence comparison of *Cryptosporidium* species to elucidate potentially exploitable variation. Widmer *et al.* identified a number of highly divergent genes by comparison of the genomes of *C. parvum* gp60 subtype IIc and the Iowa reference. Furthermore they reported that within this set of genes, transporters were highly over-represented, which lead them to conclude that these transporters may have a significant role in the ability to establish infection in the host species, potentially due to their importance in controlling the movement of metabolites between the parasite and the host cell during the epicellular developmental stages [Widmer and Sullivan, 2012, Widmer et al., 2012].

1.1.4.3 Mining the Cryptosporidium Genome for Diagnostic Purposes

The differentiation of *Cryptosporidium* spp. isolates has been attempted using a variety of methodologies, including single locus sequence typing, single nucleotide polymorphism, restriction fragment length polymorphism, amplification of random loci within the genome, conformational polymorphism, simple sequence repeats and DNA polymorphism. Using the data provided by these genomic methodologies, a great deal of information was yielded pertaining to the taxonomy of the genus, leading to the characterisation of two separate species within populations of anthropopathogenic isolates: *Cryptosporidium parvum* and *Cryptosporidium hominis*.

Gp60 typing is the most widely accepted and applied method with which to identify discrete *Cryptosporidium* spp. subtypes. This typing approach involves the interrogation of the partial sequence of a polymorphic sporozoite surface glycoprotein gene (the gp60 locus). Specifically it is based on the number of 'TCA' repeat units within the gp60 locus and the sequence downstream of this tandem repeat region. The naming convention involves differentiating species by a roman numeral (I for *C. hominis* and II for *C. parvum*), a lower case letter denoting a discrete genotype within the species, upper case alphanumeric strings denoting the number of copies of the repeated sub-sequence within the gp60 tandem repeat locus [Widmer and Lee, 2010]. For example, the subtype IIaA15G2R1, refers to a *C. hominis* subtype (II), belonging to genotype 'a', and bearing 15 TCA sequences (A15), 2 TCG sequences (G2), and 1 further non TCA/TCG sequence denoted by 'R1'.

Thus far, gp60 subtyping has been the most effective and productive approach to subtyping in *C. parvum* and has lead to the characterisation of new species, reevaluation and resolution of phylogenetic relationships and elucidation of cryptosporidiosis epidemiology. However, it is known that *C. hominis* does not effectively segregate according to gp60 locus sequence typing, and is therefore not viable as an alternative subtyping methology to multi-locus subtyping (MLST) (Widmer & Lee, 2010). Despite these advances, this approach remains a single locus genotyping methodology which is made less reliable in light of fact that *Cryptosporidium* populations of distinct subtypes are known to recombine, particularly around the gp60 region, which may be related to its relevance as a virulence factor [Guo et al., 2015b]. In 2018, Gilchrist *et al.* carried out a study of the genetic diversity of *C. hominis* in slum dwelling infants in Dhaka, Bangladesh, over a two year period. They found that *C. hominis* was more abundant during the monsoon periods, and exhibited very high levels of diversity at 7 discrete loci, including the gp60 locus. They also detected high levels of recombination, evidenced by linkage disequilibrium decay. The genetic diversity of *C. hominis* encountered in this study was found to be far greater than that seen in Europe, where the *C. hominis* genome is conserved. This study reveals both the capacity of *C. hominis* to recombine at the gp60 locus, and the importance of high-throughput, wide scale genomic sequencing and analysis in elucidating the complex population structure of this parasite worldwide [Gilchrist et al., 2018].

These weaknesses highlight the need to develop a multilocus subtyping methodology for the genus. However, attempts to do so have not lead to a consensus on which target loci to use for interrogation [Widmer and Lee, 2010]. The shortfallings of gp60 locus variation as a subtyping paradigm highlight the need for further analysis of the genome of multiple species and isolates of *Cryptosporidium* to identify novel genomic biomarkers.

Multi-locus typing approaches are acknowledged to improve resolution over single locus subtyping, particularly in the case of *Cryptosporidium*, where potentially confounding genetic recombination due to sexual reproduction is commonplace [Guo et al., 2015b]. A meta-analysis by Robinson & Chalmers (2012) investigated published literature to identify the most informative markers as candidates for the development of a standardised multi locus fragment size-based typing (MLFT) scheme for the purpose of augmenting epidemiological analysis. They collated 31 MLFT studies which reported 55 markers, of which 45 were applied to both *C. parvum* and *C. hominis*. Within the identified studies, the critical number of markers to identify 95% of all MLFTs was elucidated, illustrating that in both species, there was some marker redundancy (*C. hominis*: 40%, *C. parvum*: 27%). The authors then investigated the most informative theoretical combinations by ranking the markers. Two sets of markers were developed (one for each species), which would be evaluated for viability as MLFT schemes [Robinson and Chalmers, 2012].

In 2010, Widmer & Lee investigated whether multi-locus genotypes of *C. parvum* and *C. hominis* cluster according to the gp60 subtype. It was found that interrogation of the gp60 locus alone (gp60 subtyping) was not a suitable replacement for a MLST approach in all cases for *C. hominis* and the majority of cases for *C. parvum* [Widmer and Lee, 2010]. However, since the gp60 locus is associated with host cell invasion, and therefore can be considered a virulence factor, it may still be an appropriate target for interrogation as a phenotype determining biomarker.

During outbreaks of *C. hominis* in Europe, isolates bearing the gp60 allele IbA10G2 predominate cases of clinical infection [Hadfield et al., 2015]. MLST has been shown to be, as yet, entirely ineffective due to a lack of suitable polymorphism in loci between isolates. In contrast, there appears to be a great deal of variation between European isolates of *C. parvum*, supporting the use of a Multi-locus genotyping scheme. Consequently, much more importance has been put on *C. parvum* genotyping, driving research into genetic diversity of this species and its role in transmission.

In 2015 Widmer & Caccio investigated the relationship between sequence and length polymorphism within a set of biomarkers in the *Cryptosporidium* genome. They compared genetic distances of sequence and length polymorphism, finding that there was a weak correlation between the two distance measures. Their results also indicated that the resolution of *Cryptosporidium* population structure was dependant on the genotyping method used [Widmer and Cacciò, 2015].

The state of Cryptosporidium genotyping is far from resolved, and these publications

indicate that there is still a large amount of work to be done regarding the discovery, assessment, and selection of suitable biomarkers and genotyping conventions.

1.2 Multiplicity of Infection: An Extra Dimension in Epidemiology

Multiple populations of pathogen species within a single host is a well documented observation within the natural world. It has been reported that the majority of infections consist of multiple parasite species or discrete genetic lineages. In 1999, Lord *et al.* reported that the majority of adults infected with *Plasmodium falciparum* were host to more than five distinct strains [Lord et al., 1999]. Grinberg *et al.* reported the presence of a number of sub-populations within single isolates of *Cryptosporidium parvum* by cloning PCR amplicons of two loci which are known to be highly polymorphic (gp60 & HSP70), and utilising Next Generation Sequencing (NGS). They demonstrated the presence of two HSP70 and 10 gp60 alleles within their two isolate dataset [Grinberg et al., 2013]. Similarly, Troell *et al.* attempted to elucidate these putative intra-isolate sub-populations by using a combination of single-cell sequencing and DNA enrichment by whole genome amplification. Using this protocol they sequenced 10 single oocysts, resulting in assemblies of 49.4 - 91.8% of the size of the *C. parvum* IowaII reference genome [Abrahamsen et al., 2004]. Using these genomes, the authors detected variation at multiple loci between the assembled genomes, verifying the presence of discrete populations within the isolate [Troell et al., 2016]. It is worth noting, however, that each oocyst contains 4 sporozoites, and therefore 4 copies of the genome. In oocyst genetic heterogeneity has been suspected [Grinberg et al., 2013], and adds another layer of ambiguity to the results of whole genome analysis and epidemiological surveys.

Multiple infections have a significant impact on the incidence and spread of parasites, along with their virulence. These factors have been acknowledged to be of enormous importance to public health. An incomplete understanding of the subtypes and genotypes of a parasite within a host leads to potentially inaccurate assumptions about the clinical presentation and outlook of the infection. Efforts have been made which illustrate that there exist general patterns when ecological and evolutionary theory are applied to within-host dynamics, leading to the generalization that "basic ecological rules govern the outcome of co-infection across a broad spectrum of parasite taxa" [Graham, 2008]. Graham *et al.* carried out a meta-analysis of helmith-microparasite coinfections, which demonstrated that, as in ecology, the abundance or limitation of key resources (nutrients) were a key factor governing in-host population size, as did host immune response [Graham, 2008].

The virulence (defined here as 'the capacity of the parasite to reduce host fitness') experienced by a host infected by multiple subtypes/genotypes of a parasite (often referred to as 'overall virulence') is a result of the interactions between the different sub-populations within the host, resulting in an overall virulence of anywhere from greater than the most virulence to less than the least virulent, depending on various biological factors [Seppälä et al., 2009, Seppälä et al., 2012]. It has been argued, by Alizon *et al.*, that overall virulence cannot be used to predict virulence evolution. Instead, they suggest that detailing how the presence of multiple sub-populations affects the duration of infection, and transmission of the genetically discrete sub-populations is far more of a driving factor in virulence evolution. The evolution of virulence is a complex issue, governed by a wide array of host-parasite and in-host parasite-parasite interactions, and between-host transmission of discrete parasite populations [Alizon et al., 2013].

Because of a lack of scalable, high-throughput techniques to investigate and characterise Multiplicity of Infection (MOI), our understanding of the evolutionary consequences of this phenomenon is poor. Previously, this has lead to many public & animal health policies underestimating, or failing to take into account the impact of MOI [Read and Taylor, 2001, Balmer and Tanner, 2011].

1.3 Genome Analysis

1.3.1 Whole Genome Sequencing, Assembly, Improvement and Annotation

Sequence assembly can be split into two major categories: *de novo* assembly, and mapped assembly.

1.3.1.1 De novo Assembly

De novo assembly is a manner by which a sequence is assembled from NGS reads by the application of an algorithm, without a reference genome. Two of the most widely used assembly algorithms are the de Bruijn graph (DBG), and overlap layout consensus (OLC).

De Bruijn graph assembly involves a graph algorithm based on leveraging the kmer content of each read. Using overlapping kmers, a walk can be carried out to find overlapping reads, which in turn can be used to assembly longer contiguous sequences (contigs). This approach of digesting reads into smaller sequences (kmers) solves problems introduced by differing read lengths.

Overlap Layout Consensus is another graph based algorithm which utilises, as the name suggests, three major phases: overlap, layout and consensus. The initial phase involves calculation of sequence overlap between each read. These are then arranged according to these calculations to form contiguous sequences. During the consensus phase, each base within a read is assigned a quality score dependant on the depth of coverage for that base in the read library, i.e. if that base is sequenced 60 times (x60 depth) and 59 of those times the base is sequenced as Adenine, we can say that this base is Adenine with a high degree of certainty. Using this logic, OLC assembly algorithms form contigs by calculating the consensus for each base at each position of the genome.

There are a multitude of other assembly algorithms, such as the Greedy algorithm, which finds optimally overlapped reads to produce contigs, and Hybrid techniques, which utilise a mixture of other algorithms.

The unbiased nature of de novo assembly may pose a problem if the objective is to assemble the genome of a single organism, since contaminant DNA will also be assembled into contigs in an identical manner. It is therefore necessary either that 'cleaning' should be done on either the samples pre DNA extraction (*in vitro* using processes like immunomagnetic separation and bleaching, which has been used by Hadfield *et al* (2015) to clean *Cryptosporidium* oocysts extracted from stool samples [Hadfield et al., 2015]) or the reads prior to assembly (*in silico* using tools such as Kontaminant [Leggett et al., 2013, Daly et al., 2015]) to reduce the amount of contaminant sequence. Popular *de novo* assemblers include SPAdes [Bankevich et al., 2012], Velvet [Zerbino and Birney, 2008], MaSuRCA [Zimin et al., 2013], SOAPdenovo [Li et al., 2010] and ABySS [Simpson et al., 2009].

1.3.1.2 Mapped Assembly

Mapped assembly involves an alignment-consensus algorithm, whereby WGS reads are initially aligned to a (putatively similar) reference genome. This alignment is used to guide the assembly of the new genome. It is much faster than *de novo* assembly due to it obviating the need to carry out the intensive overlap step. Scaffolding also becomes unnecessary due to the reference guiding the structure of the new genome. However, problems arise when the reference genome is divergent from, or there is recombination or restructuring of, the genome being assembled.

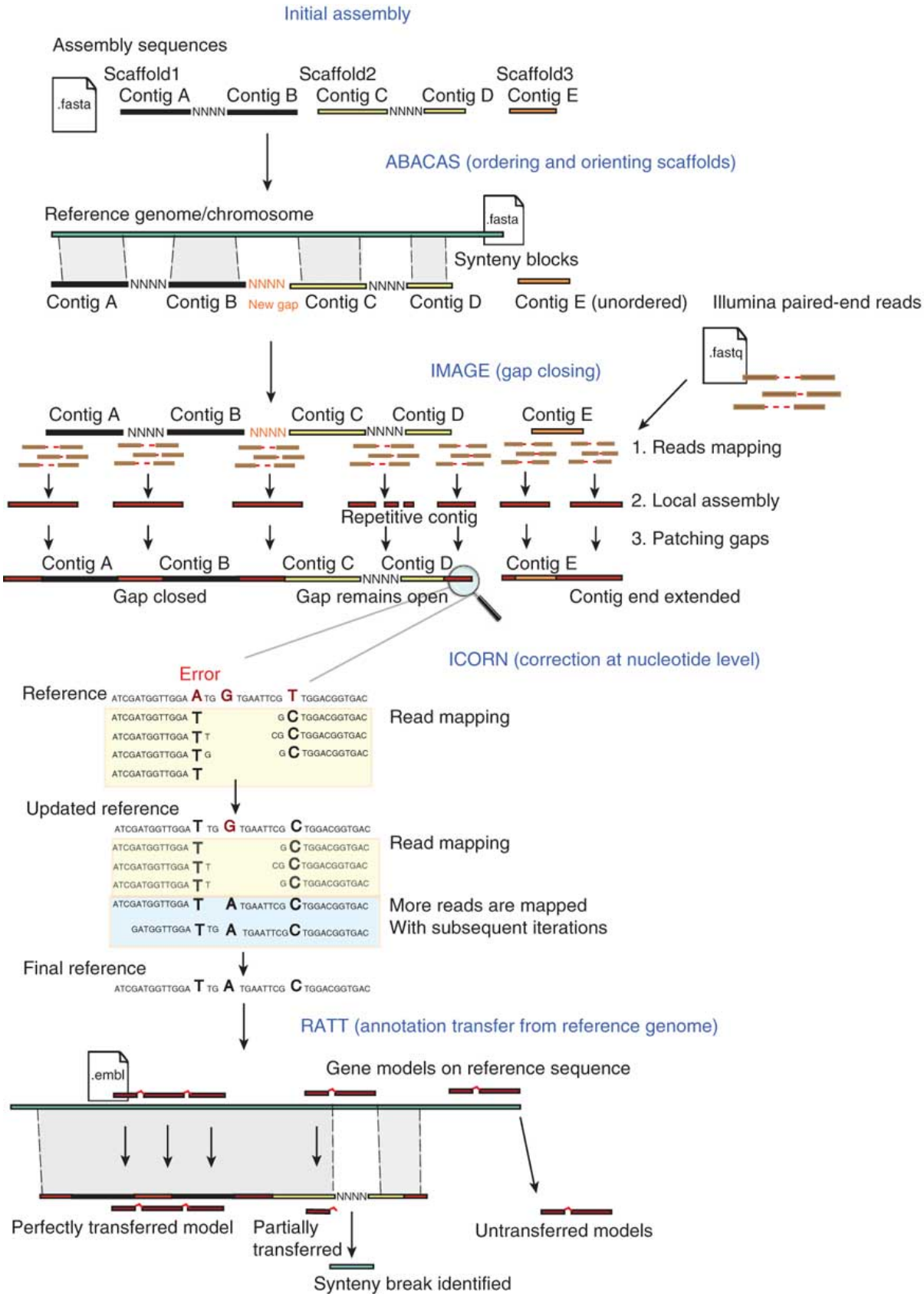


Figure 1.4: The workflow of genome improvement in the PAGIT pipeline, illustrating the function of the four components: ABACAS, IMAGE, ICORN & RATT [Swain et al., 2012].

1.3.1.3 Genome Improvement

Genome improvement is a fundamental aspect of assembly. In the past, curation of genomes of great importance (such as the human genome, *Escherichia coli* or *Saccharomyces cerevisiae*) has been undertaken manually by dedicated teams of researchers, involving a great deal of time and effort. However, this is an intensive, costly and time consuming process that is not viable for the majority of genome projects. Consequently automated improvement has become of great importance to attempt to improve the quality of assemblies in an attempt to bring it up to the standard of manually curated genomes. The main objectives of genome improvement involve gap closure, error correction and utilising reference guided assemblers and highly similar genomes to improve scaffolding and annotation. The Post-Assembly Genome-Improvement Toolkit (PAGIT) [Swain et al., 2012], developed by The Welcome Trust Sanger Institute, utilises a number of tools to improve the quality of and annotate assemblies. PAGIT consists of four components: ABACAS, IMAGE, ICORN and RATT. The function of these components are genome reference based genome structuring, gap closure, error (indel) correction, and annotation respectively (see Figure 1.4). ABACAS is used to map the contigs produced by *de novo* assembly to a reference genome of high similarity (95-99%), where it will exploit the structure and conformation of the reference assembly to produce a similarly structured assembly, whilst plugging gaps within the new assembly with 'N'-spaces. The fastq paired end reads are then used to attempt to close these gaps by aligning them in such a fashion that they overhang and extend into the N-space. ICORN utilises iterative mapping of reads onto the consensus sequence to identify putative errors (SNP's or indels of up to 3bp). It then measures the change in read coverage that would result from the correction, and if there is no reduction it is presumed that the new sequence is correct and the correction is incorporated into the assembly. ICORN utilises SSAHA to perform read mappings and the SSAHA pileup pipeline to identify SNP's and indels <=3bp [Ning et al., 2001]. Provisional corrections are then evaluated using SNP-o-matic [Manske and Kwiatkowski, 2009]. In 2016, they updated PAGIT and made it available as a web server for the pseudochromosome contiguation (using ABACAS2: an updated version of the previously described ABACAS), annotation and analysis of parasite genomes [Steinbiss et al., 2016].

1.3.1.4 Genome Annotation

Gene annotation takes place using two major approaches: *ab initio* annotation, which detects open reading frames (ORF's) by identifying regions which are flanked by start and stop codons; and reference guided annotation transfer, which uses gene orthology inference between an assembly and a similar annotated reference assembly. Rapid Annotation

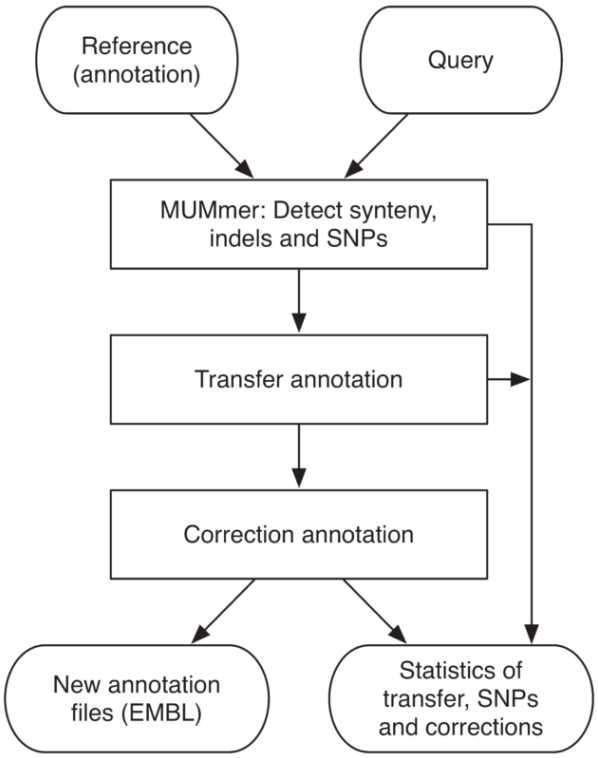


Figure 1.5: Workflow of RATT. Taken from Otto *et al* [Otto et al., 2011].

Transfer Tool [Otto et al., 2011] (RATT) is a reference based annotator which utilises this approach of orthology inference. In RATT, NUCmer (a tool from the MUMmer tookit [Kurtz et al., 2004]) is used to identify regions of sequence synteny (with a 40% nucleotide sequence identity threshold). RATT was designed to transfer annotations in three scenarios: between successive versions of the same assembly, between the genomes of closely related species, and between the genomes of closely related strains (see Figure 1.5. RATT employs various techniques of gene inference that are common in *de novo* annotation, such as ORF distinction; whereby start and stop codons are detected and allowed to influence annotation, allowing for gene variability between the reference annotation and the assembly to be annotated. Annotation of genomes has played a large part in VNTR discovery during this project, since it allows for the quick recovery of genes for comparison and analysis. However it may be necessary to attempt *de novo* annotation on the more phylogenetically distant species due to the polymorphic nature of the *C. parvum* genome leading to incomplete annotation transfer.

1.4 Alignment-free Sequence Analysis

Sequence comparison by alignment is subject to a major assumption that there is conservation of contiguity between nucleotide or protein sequences which are functionally and/or evolutionarily related (i.e. significant rearrangement is rare and often results in

functional change). This is not always the case due to genetic recombination and restructuring. Because of this, alternative approaches where developed due to the necessity of analysing the similarity of DNA sequences whilst attempting to avoid these principles of genetic rearrangement that confound alignment based sequence analysis approaches [Vinga and Almeida, 2003, Zielezinski et al., 2017]. These approaches have garnered greater interest due to the increase in the size of datasets to analyse [Vinga and Almeida, 2003]. In general, these approaches are split into two major categories: genetic 'word' (or 'k-mer') frequency approaches, and information theory based approaches [Zielezinski et al., 2017]. These approaches have been extensively developed, and therefore have rigorous mathematical foundations.

1.4.1 K-mer Analysis and Applications

Because of the dramatic increase in genomic data being produced in next generation sequencing studies in recent years, there is an elevated requirement for these genomic data to be analysed. Analysis of genomic data generally necessitates *de novo* assembly, homology searches from raw reads or mapping of data to a reference genome, which are all time consuming and have the potential of introducing internal errors. Genome analysis methods involving interpretation of the spectra of k-mers have attracted interest over recent years due to fact that they do not necessitate the genome to be assembled, analysis can be performed directly on sequencing reads. This dramatically speeds up the process of analysis for a genome [Zielezinski et al., 2017].

1.4.1.1 K-mer Counting

The first stage of k-mer spectrum (also referred to as a k-mer list) analysis involves the formation of a k-mer database from the query genome and counting the frequency (incidence) of each k-mer within the database. This database is then generally ordered by frequency to facilitate further analysis [Kurtz et al., 2008].

If we consider a sequence, X, of length n, as a string of n elements from a finite, defined alphabet, A, of size r. A k-mer (w_k) is defined as a substring of k elements of X where $k \leq n$, and can be generated using a sliding frame approach over X (detailed in Eq. 1.1). r^k possible k-mers can be generated from A. The set W_k contains all (S) k-mers that may be generated from sequence X:

$$W_k = \{w_{k,1}, ..., w_{k,S}\} \tag{1.1}$$

$$S = n - k + 1 \tag{1.2}$$

The k-mers within X can then be counted computationally. Word frequency ($f^X_{k,i}$) can be calculated in a similar manner. A vector of k-mer probabilities can be generated by calculating the probability ($p^X_{k,i}$) of encountering each k-mer within X:

$$p^X_k = \{p^X_{k,1}, p^X_{k,2}, ..., p^X_{k,S}\} \tag{1.3}$$

Likewise, the vector of frequencies can be defined as the relative abundance of each k-mer:

$$f^X_k = \frac{c^X_k}{\sum_{j=1}^{S} c^X_j} \Leftrightarrow \frac{c^X_{k,i}}{S} \tag{1.4}$$

There are many k-mer counting tools available, many of which function as part of larger programs. One of the most efficient and widely used is Jellyfish [Marçais and Kingsford, 2011], which is optimised for counting <=31bp k-mers based on a multi-threaded, lock-free hash table. Jellyfish is commonly used as a fast k-mer counting tool in higher level software or pipelines, due to it's flexible and robust nature. However, in some use-cases, a more bespoke tool may be necessary. Tallymer is a tool built specifically for the purpose of counting k-mers from large eukaryotic genomes containing large proportions of low complexity repetitive sequence [Kurtz et al., 2008]. This is of great benefit when working with plant genomes, which are both extremely large, often polyploid, and highly repetitive. Assumptions of particular genome characteristics may be vital to downstream analysis when counting k-mers within a dataset. Counting k-mers in large datasets may be an extremely computationally intensive task, requiring large amounts of memory and disk space. A rough calculation demonstrates that, given a genome (a string of elements from an alphabet of size 4) of size s (the number of nucleotides), the theoretical upper bound of required memory to store a k-mer array would be $k(s - k + 1)$ bytes, where $4^k \geq s - k + 1$, else 4^k bytes. In real terms, this means that the theoretical upper bound of required memory for storing a 15-mer array from a 1Gb genome would be 15Gb, assuming optimal uncompressed storage. In an attempt to alleviate such requirements, DSK uses a defined amount of memory and disk space by writing temporary k-mer tables onto disk storage, allowing it to run on systems with low memory [Rizk et al., 2013]. The application of probabilistic data structures has also been used to reduce the memory consumption of k-mer counting methods, since they are much more memory efficient than k-mer counting algorithms that use exact data structures. Such is the case with Khmer: a k-mer counting tool which uses a simple probabilistic data structure, called a Count-Min Sketch, which permits online retrieval and updating

of k-mer counts in memory [Zhang et al., 2014].

1.4.1.2 K-mer Spectrum Analysis

In the last 10 years, there has been interest in analysis of the distribution of k-mers across a genome. Particular interest has been directed towards the extremes of this spectrum, that is, k-mers that are poorly represented (low frequency k-mers) and k-mers that are abundant (high frequency k-mers). This has lead to various hypotheses attempting to understand the purpose of these over and under represented k-mers.

After a k-mer count has been undertaken on a genome, analysis of this spectrum must be carried out. Analysis of k-mer spectra has been used applied to tasks such as genome size estimation, de novo repeat detection, measurement of gene expression [Patro et al., 2014], read matching between metagenomic data sets, and identification of organisms within metagenomic data sets using DNA species specific DNA motifs [Wood and Salzberg, 2014].

It has been postulated that high frequency k-mers are due to their appearance in structural, mobile and regulatory elements, although it has also been suggested that the high frequency of these motifs represent unrecognised or little understood biological phenomena [Rigoutsos et al., 2006].

A paper by Csuros et al (2008) proposed that this distribution of k-mers across a genome can be primarily characterised by a double paretolognormal distribution. This would explain the observation that all genomic k-mer spectrum distributions appear to exhibit lognormal and power law features. However, the authors also stated that careful analysis of k-mer spectra should be made against random sequence to elucidate whether over-representation of k-mers is a product of the character of the genome, or of probability, in which case it will be exhibited in random sequence [Csuros et al., 2007]. This observation could be applied in k-mer spectra analysis as a method of quickly identifying exceptional k-mers within sets of k-mer spectra, indicating either highly conserved regions which could be examined as potential drug targets or genetic biomarkers, or highly divergent regions to interrogate as highly specific biomarkers. The results indicate that the analysis of motif frequency within a genome should be treated with care due to the heavy tail within the k-mer distribution spectrum, particularly if this result is also exhibited within random text [Csuros et al., 2007].

1.4.1.3 Applications of K-mer Analysis

The use of k-mers in bioinformatics has typically been associated with de-novo genome assembly and genomic and metagenomic annotation [Compeau et al., 2011, Edwards et al., 2012]. K-mer based genome assembly typically involves the construction of De Brujn graphs which are then used to map a set of reads for assembly into a longer contiguous sequence by representing k-mers as nodes and edges as single base pair differences between the k-mers. Contigs are then assembled by calculating a path through the graph [Compeau et al., 2011].

K-mer analysis can also be used to facilitate quality assurance of next-gen sequencing datasets. KmerStream is a k-mer streaming algorithm for estimating the number of unique k-mers that occur within NGS data sets. The authors report that its primary applications are in error and genome size estimation within high throughput sequencing data [Melsted and Halldorsson, 2014]. Likewise, Kontaminant is a tool developed at The Genome Analysis Centre (TGAC) for filtering sequencing reading before assembly. The tool involves the generation of a k-mer library from a reference. The reads are then filtered by comparison against these reference k-mers and subsequently discarded based on whether the read bears a number of k-mers in common of less than a threshold value (which is highly dependant on the value of k) [Ramirez-Gonzalez, 2014].

There has been prolonged use of the analysis and measurement of frequency and distribution of k-mers for the purpose of identification. The collection of k-mer frequency distributions makes for an effective way of generating unique 'barcodes' for genomes to aid in species (and intraspecies) identification. A paper by Zhou *et al* (2008) investigated the k-mer frequency distribution throughout the genomes of five prokaryotic organisms to develop barcoding schemes, and their use in solving challenging genome analysis problems. The authors used their barcoding scheme to address two applicational problems: metagenome binning problems and identification of horizontally transferred genes. The study was based on the premise that each genome has a highly stable distribution of the combined frequency for each k-mer within fragments of $> 1kb$ [Zhou et al., 2008]. The frequency distributions are used to generate a unique barcode for the genome. Toolkits such as GenomeTester4 can facilitate such analyses, allowing the user to generate and perform a range of operations on k-mer lists. In particular, GListCompare is a novel tool within this toolkit which can be used to perform union, intersection and compliment set operations. This allows for basic pairwise comparison of k-mer lists [Kaplinski et al., 2015].

The problem of correct genome identification within fragmented metagenomic data

has long been one of particular importance in metagenomic projects. Consistent and accurate separation of these fragments within a metagenomic dataset would have a significant implication on the fields of genomics and metagenomics [Zhou et al., 2008]. Kraken is a very fast and powerful tool that uses known k-mer lists to identify reads in metagenomic datasets. It classifies k-mers using a phylogenetic approach and then, by performing list comparison against k-mer lists generated from known species it estimates the origin of the organisms by closest match [Wood and Salzberg, 2014]. The development of the database in Kraken is based on a set of k-mer lists and the lowest common ancestor for all organisms. K-mers are queried to the database, whereby the lowest common ancestor which contains all of the k-mers queried is calculated. The database is developed using a list of genomes specified by the user, which dramatically decreases k-mer lookup time.

Due to the troubling nature of development of robust subtype differentiation in *Cryptosporidium*, the application of k-mers to this problem may be particularly fruitful. Identification of k-mers specific to different isolates allows for a novel and direct approach to solving this problem, since these k-mers act as isolate specific motifs which could be detected by PCR amplification in a clinical laboratory.

1.4.2 Information Theory Based Analysis and Applications

The main concept behind information theory based methods is to recognise and compute the intersection between the informational content of two sequences of interest. If one treats amino acid and nucleotide sequences simply as strings of symbols (a common practice), then their organisation becomes interpretable, and therefore their informational content more accessible. There are many examples of informational content, but the most important ones in bioinformatics are ones such as complexity and entropy. The calculation of the complexity and entropy of a DNA sequence has been the focus of a great deal of research, spanning many disciplines.

1.4.2.1 Sequence Entropy

Entropy in the context of informational content of DNA or amino acid sequence is not comparable to entropy as referenced in thermodynamics, and should be treated as an entirely different concept. The main concept behind entropy (termed 'Shannon entropy', after Claude Shannon, the father of information theory) is in investigating redundancy within a string. For example, the meaning of the sentence "the sky and the sea are blue" would still be inferable if one were to remove the words "the", "and", and "are". This idea of redundancy can be applied to DNA sequence, where rarely encountered 'words' (k-mers) are more important to sequence analysis than more common ones. Shannon

entropy, H, is described using the formula:

$$H = -\sum_i p_i \log_b p_i$$

where p_i is the probability of word number i appearing in in a stream of words. In the context of sequence analysis, this would refer to a k-mer within a k-mer array.

Using this concept of Shannon entropy, Kullback and Leibler (1951) introduced a measure of relative entropy, termed Kullback-Leibler divergence, allowing for the pairwise comparison of DNA or amino acid sequence [Leibler and Kullback, 1951], involving the calculation of word frequencies and summation of entropies within the sequences.

1.5 Data Structures used in K-mer Array Analysis

There are a number of discrete data structures which are utilised in the analysis of the k-mer spectrum of DNA sequence. In principle, any data structure with functionality to track the instance of events within a data-stream, and facilitate queries upon that set about the presence or count of that event within the data-stream, may be applied to sequence comparison using k-mer arrays. However, there are a select few with popular application in sequence analysis. This is likely due to their memory scalability, allowing these data structures to store large amounts of data in a memory efficient manner, and lookup time. This is essential when dealing with large datasets, where simply storing the set in memory and querying it directly is; at best, highly time and memory inefficient; and at worst, not possible.

1.5.1 Hash Tables

Hashing is a technique used to efficiently store an array of elements and perform lookup in $O(i)$ time. The technique relies upon the input of elements within an arbitrarily large array A into a hash table T, which is an empty array of fixed size. In this example we define $A \subset \mathbb{N}$. The problem lies in the fact that an element within A maps to a location within T, and in this example where we map using a 'one to one' system, this element may be arbitrarily large. This forces $|T|$ to be arbitrarily large, which may be highly space inefficient. The solution to this is to implement hash-functions h which utilise a 'many to one' system, meaning in this context that the value of an element within A does not necessarily refer to an index within T.

If we redefine $A = \{2, 5, 9, 12, 50\}$, and $h(x) = x \bmod |T|$ then $|T| \geqslant |\{h(x) \mid x \in A\}|$. This is because when we hash $x \in A$ using h, we produce a hashed array $h_A = \{2, 5, 9, 2, 0\}$

which we can use as index positions within T to map information to. However, as can be seen, 2 and 12 hash to the same value using h (2) and therefore there exists a collision within our hash table, where information from two elements within A have been mapped to the same index in T. This can be resolved using a variety of approaches, usually divided into open hashing and closed hashing approaches.

1.5.1.1 Open Hashing

Chaining is a method by which hash collisions can be resolved. In chaining, elements are not stored in the hash table itself, instead the hash table stores linked lists which store elements and keys. Upon a collision, a key element will be inserted to the index position to form a chain of nodes. The initial node therefore stores the initial element which was mapped to that index location within the hash table, and a key pointing to the linked node containing the second element which caused the collision. Searching this linked list is therefore performed in linear time.

1.5.1.2 Closed Hashing

Linear Probing is a method by which hash collisions are resolved by simply mapping the element to the next empty index position within the hash table. This is achieved by introducing a second function, $f(i) = i$, which incrementally increases the output from h until an empty position within T is found. This adjusts h to $h(x) = (h(x) + f(i)) \bmod 10$ where $i = \{0, 1, 2, ...\}$. When a lookup is performed, the query element is hashed as usual and a linear search (defined by f) is performed from the index position within T until the query is encountered. If an empty index is encountered first, the queried element is presumed to not belong to the hash table.

The major issue with linear probing is that it tends to induce clustering of data within the hash table. This clustering results in a greater number of elements within the hash table needing to be checked for each query, increasing lookup time. To resolve this, **Quadratic Probing** may be implemented. Quadratic probing is identical to linear probing, with the exception that $f(i) = i^2$. This has the effect of evening out data mapping throughout the hash table.

1.5.2 Bloom Filters

A Bloom filter is a memory efficient probabilistic data structure used to infer set membership. Bloom filters bear resemblance to hash tables, but differ in that they do not store elements themselves, but adjust a fixed length common bit array according to the element observed within a set. They achieve this by using a number of hash functions to generate a bit signature for each element, and adjusting the bit array at positions dictate

by this bit signature (see figure 1.6). This is a fast and highly efficient process, but due to its probabilistic nature it can only *infer* set membership to a certain degree of accuracy defined by a false positive rate. Because Bloom filters are of a fixed size, it can represent a set of an arbitrarily large size, unlike a hash table. However, the false positive rate is adjusted every time a new element is added, and therefore defining the length of the bit array requires care.

To build a Bloom filter, an empty bit array of fixed length is initialised. n hash functions are then utilised to hash elements of a set (S). When an item, x where $x \in S$, is added to the filter, the bits at n indices $h^1(x), h^2(x), ...h^n(x)$ are adjusted immutably to 1, where the indices are defined by the output from the hash functions. False positive results can be controlled by adjusting the size of the Bloom filter, wherein a larger bit array relative to the set being stored reduces the chance of false positive results. However, this would also increase latency when elements are being added to and screened against the Bloom filter. The probability of false positive results can therefore be calculated using the size of the bit array, the size of the set, and the number of hash functions. Let m be the size of the bit array, k be the number of hash functions utilised, and n be the number of elements in the set used to generate the Bloom filter, the false positive rate (P) can be calculated as:

$$P = (1 - [1 - \frac{1}{m}]^{kn})^k$$

The size of the bit array can be calculated if the expected number of elements to be added to the Bloom filter, and the false positive rate are known. This can be calculated as:

$$m = -\frac{n \ln P}{(\ln 2)^2}$$

The optimum number of hash functions (H) can be calculated as:

$$H = \frac{m}{n} \ln 2$$

In the context of sequence analysis, Bloom filters can be used in conjunction with k-mer arrays to infer sequence similarity between a given sequence and the sequence which was used to generate the Bloom filter. Figure 1.6 illustrates the method by which a Bloom filter is generated from a sequence. A k-mer array is generated from a given sequence. This k-mer array is used as the set from which a Bloom filter can be built. Each k-mer within the array is hashed using a number of hash functions, and the bit array adjusted according to this output. After every k-mer has been added to the Bloom filter, screening can commence using a k-mer array (using an identical k value, otherwise no intersections will be detected) generated from a query sequence (see figure 1.7). Each k-mer from the query sequence is hashed using the same hash functions that were used to generate the

Bloom filter, and the subsequent output (termed the "bit-signature") compared against the Bloom filter to detect intersections. Where intersections are detected, a counter is incremented by 1. This hit counter can be used to determine the level of similarity between the sequence used to generate the Bloom filter and the query sequence.

1.5.3 Count-Min Sketch

A count-min sketch is a probabilistic data structure capable of representing, and answering queries of, a high-dimensional vector. It is in essence a frequency table, storing events within a data-stream. In effect, it is a counting Bloom filter. Rather than storing a one-dimensional bit array, a count-min sketch C is an array of counters of width w and depth d, $C[1,1]...C[d,w]$. w and s are fixed when the sketch is created. Upon being presented with a new event i, each row j of the table is updated using the corresponding hash function to obtain the column index $k = h_j(i)$. The counter at row j column k is incrementally increased by one.

Upon the creation of this sketch, various queries on the data stream can be made. The most apparent of which is simply a query of the count of an event. This is slightly more problematic than it may seem, due to a number of positions within the sketch being incremented by a single event, defined by the number of hash functions used in its generation. Therefore due to different events potentially generating identical outputs when consumed by different hash functions (in effect, hash collisions), the count at each position generated by each hash function may not be the same. To resolve this, the minimum count is taken as the actual count for the number of observations of an event within a data-stream. Because of this, in a similar manner to Bloom filters presenting with no type-2 error at the expense of manageable type-1 error, a count-min sketch may overestimate the number of events within a data-stream, but can never underestimate it.

1.6 Conclusion

The importance of examining and resolving the current bottle neck between the generation and analysis of NGS data cannot be overstated. The wealth of information within publicly available genomes which have yet to be analysed necessitates the development of novel approaches to genome analysis and high-throughput application of the subsequently developed tools, reducing the amount of manual work that is required for large scale biological data analysis. Analysis of sequencing reads from raw read files is a potential solution to this problem, which would dramatically speed up analysis time due to avoiding the necessity of genome assembly and therefore obviate the requirement for costly equipment such as a High-Performance Computing (HPC) cluster, allowing for

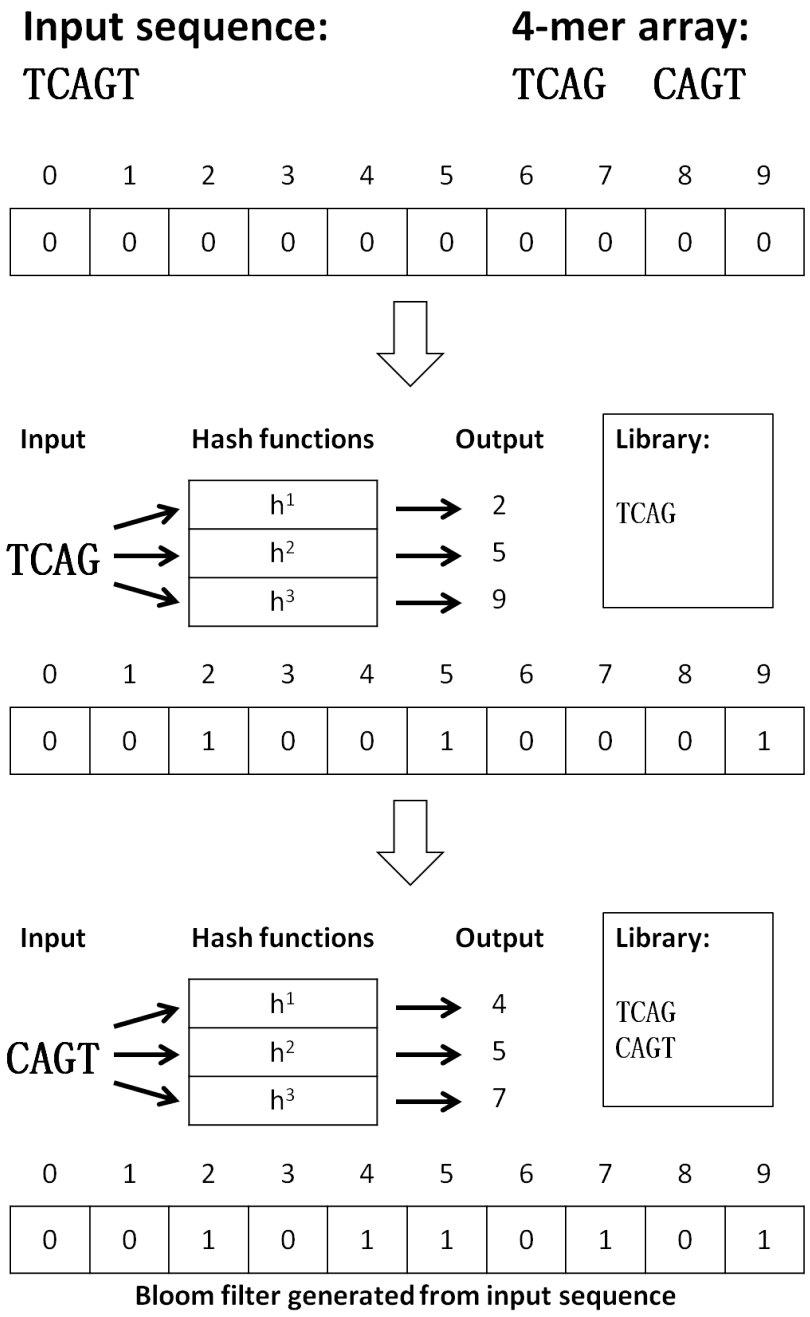


Figure 1.6: A graphical representation of utilising a k-mer array to generate a Bloom filter. An input sequence is k-merised and the k-mers hashed using a number of hash functions. The output of the hash functions refers to index positions on the bit array, which are then adjusted immutably to 1.

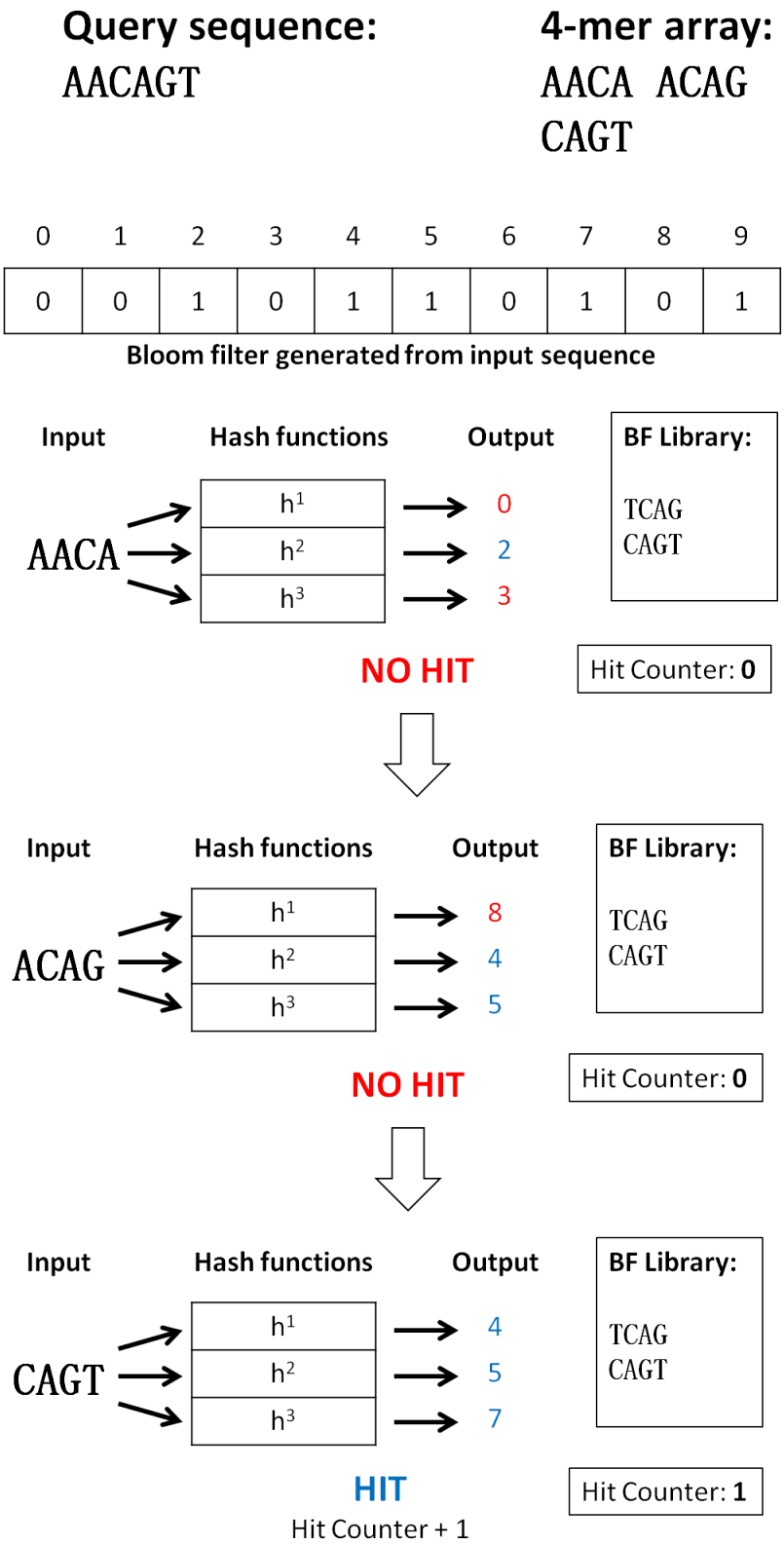


Figure 1.7: A graphical representation of the process of screening a set of k-mers generated from a query sequence against a Bloom filter (see figure 1.6). Each k-mer within the query array is hashed using the same hash functions that were used to generate the Bloom filter, and the output used to infer intersections. If an intersection is detected, the hit counter is increased by 1.

analysis to be carried out on the standard workstation computer or laptop. The application of k-mer analysis tools to this area could be a potentially valuable approach in the discovery of biomarkers for species identification or small conserved sequences associated with gene regulation which could be examined as targets for novel therapies.

Chapter 2

From Gastroenteritis to Genome: Generating Genome Assemblies from Clinical Samples of *Cryptosporidium*

The sequencing and assembly of whole or partial genomes has become an essential tool in modern science, facilitating research in almost every area of biology. A primary concern in *Cryptosporidium* research is extracting from clinical samples sufficient amounts of high quality, low contaminant DNA for sequencing. Without this, sequencing may result in low coverage sequence, variable sequencing depth and poor quality genome assemblies [Morris et al., 2019b, Morris et al., 2019a]. In the area of Cryptosporidiosis research, the impact of genomics has been limited by the need to propagate the parasite in animals to generate enough oocysts from which to extract DNA of sufficient quantity and purity for analysis [Abrahamsen et al., 2004]. In 2015 this problem was overcome through an approach that now allows genomic *Cryptosporidium* DNA suitable for whole genome sequencing to be prepared directly from human stool samples [Hadfield et al., 2015]. Hadfield *et al.* (2015) applied their method to the whole genome sequencing of eight *C. parvum* and *C. hominis* isolates. Presently, the *Cryptosporidium* genomics resource, CryptoDB [Puiu et al., 2004], currently gives access to 13 complete genomes, with a total of 10 available from the NCBI.

Currently clinical diagnosis of Cryptosporidiosis relies on conventional genotyping tests. The availability of whole *Cryptosporidium* genome sequences provides much higher resolution information for genotyping. In addition, the genomes can be used to study a wide array of aspects of pathogen biology, such as identity, taxonomy in relation to other pathogens, sensitivity or resistance to drugs, development of novel therapeutic agents, virulence, and epidemiology. My interest is to build on current genotyping tests by developing a standardised multi-locus typing scheme. This will allow sources of contamination and routes of transmission to be characterized and compared in a cost- and time-efficient

manner [Perez-Cordon et al., 2016, Chalmers et al., 2017]. Here variable-number of tandem-repeats (VNTR) are used, with recent investigations concluding that additional loci need to be identified and validated [Chalmers et al., 2017]. Our work is building on that of Perez-Cordon *et al.* (2016), who used Tandem Repeats Finder [Benson, 1999] to identify polymorphic VNTR's around the genome of *C. parvum*, and analysed them for variation across the eight genomes sequenced by Hadfield *et al.* (2015). I aim to use whole genome sequencing of additional isolates and species to help achieve this goal, but this work is hampered by the quality of available genome sequences [Perez-Cordon et al., 2016].

This chapter is structured as follows. First, I explain the quality issues associated with genome sequences extracted from clinical stool samples containing *Cryptosporidium* oocysts. Then I describe our methods, including the data sets used, the novel utilisation of Gini and Gini-granularity curves to measure the distribution of read depth in a set of sequenced reads, the process of assembly with the identification of misassemblies, and the effect Whole Genome Amplification (WGA) has on coverage distribution. In the results and discussion sections, I summarise properties of the sequenced reads, show how they can lead to misassemblies, and give evidence of the types of misassembly encountered. I also describe how using the Gini coefficient, and analysis of Gini-granularity curves can explain some of these assembly errors and characterise the distribution of reads across a genome. I then give an outline of the strategy used to generate genome assemblies of sufficient quality to use for the discovery of novel VNTR's in *Cryptosporidium*, and use it to generate genome assemblies for an extended dataset of *C. parvum* and *C. hominis*. Finally I discuss the value of WGA in generating high quality assemblies clinical samples of *Cryptosporidium*, and how it affects the coverage across a genome.

Data within this chapter was published in the *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS* under the title *Identifying and Resolving Genome Misassembly Issues Important for Biomarker Discovery in the Protozoan Parasite , Cryptosporidium.* A further publication is currently (as of October 2019) in peer review, to be published in a special edition Springer CCIS series book under the title *Generating Reliable Genome Assemblies of Intestinal Protozoans from Clinical Samples for the Purpose of Biomarker Discovery.*

2.1 The Problem

Although it is possible to derive high quality *Cryptosporidium* DNA by culturing the parasite in donor animals [Abrahamsen et al., 2004], this is expensive and time consuming,

and is not appropriate for clinical samples, since sequence identity may be modified during asexual reproduction or sexual recombination. Sustaining sequence identity of clinical samples is of great importance in epidemiological surveys. Sequencing *Cryptosporidium* from clinical samples suffers from three major problems:

1. The yield of oocysts from clinical samples is low.
2. The oocysts are extracted directly from faeces, necessitating extensive cleaning and purification before DNA extraction.
3. The DNA yield per oocyst is low.

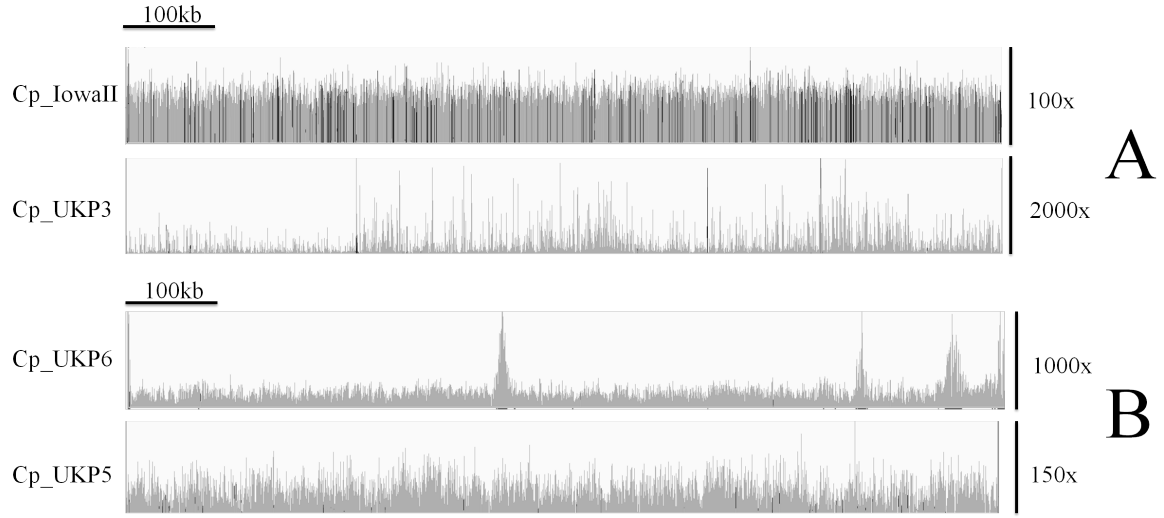


Figure 2.1: **A**: Read depth of coverage across chromosome 7 of the *C.parvum* IowaII reference genome (top track) and *C.parvum* UKP3 (bottom track) genomes to illustrate the extreme coverage inequality of the UKP3 isolate genome (UKP3 $Gini = 0.5489$, IowaII $Gini = 0.112$). Image produced using IGV. Note that the IowaII DNA sequences were derived from an animal model, and have low or "normal" read depth variation, whereas UKP3 is more typical of DNA sequences extracted from clinical samples. **B**: The coverage over chromosome 1 for 2 genomes: UKP6 ($G \mid W_1 = 0.255$, $nAUC = 0.921$) and UKP5 ($G \mid W_1 = 0.278$, $nAUC = 0.884$) with $G \mid W_1$ across chromosome 1 alone of 0.262 and 0.264 respectively. UKP6 illustrates large, broad peaks (blocking) in a number of areas. These do not appear to correspond to any obvious features, however, it is presented here as a useful example of a type of coverage inequality. See Section 2.2.4 for an explanation of the annotation.

These three problems commonly result in sequenced data sets with very uneven depth of coverage, see Figure 2.1 for examples. The reasons for uneven depth of coverage are unclear; in this chapter I have attempted to elucidate some of the issues. Uneven sequencing depth has been identified in datasets obtained from published and unpublished paired end read libraries generated by different groups, and which were prepared using the standard Nextera XT DNA sample preparation kit. Moreover, many groups use

WGA to increase the quantity of extracted DNA. This may have additional impact on the depth of coverage. WGA has been touted as a potential solution to samples which yield low levels of DNA or for which little DNA exists [Lasken and Egholm, 2003, Zhang et al., 2006, Hosono et al., 2003]. There has, however, been limited rigorous research into coverage bias introduced by such DNA enrichment techniques. Moreover, laboratory kits and equipment have been demonstrated as a potential source of confounding contamination, which is potentially greatly exacerbated by the amplification of DNA using such enrichment techniques [Salter et al., 2014, Mohammadi et al., 2005, Van Der Horst et al., 2013]. Uneven sequencing depth may lead to genome misassembly, and I have identified this an issue with a number of popular *de novo* assemblers. Poor quality genome assemblies can find their way into public repositories of genome sequence and this can confound the development of novel prevention strategies, therapeutics, and diagnostic approaches.

2.1.1 Datasets

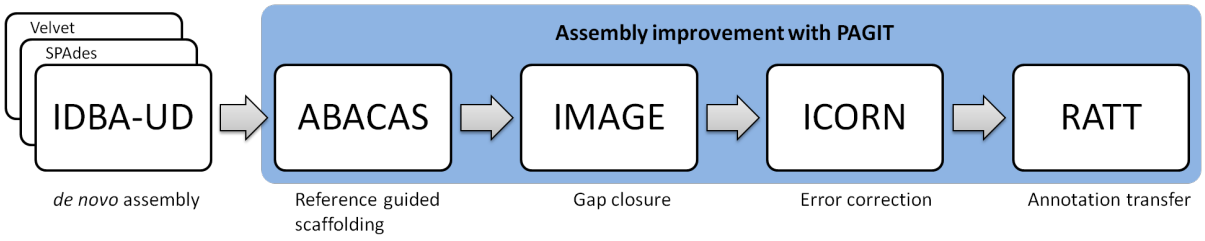


Figure 2.2: The workflow for assembly, adapted from that used by Hadfield *et al.* for the assembly of genomes with high coverage depth inequality.

Two main datasets were used within this chapter. These datasets a differentiated by their origin: published by Hadfield *et al.* (Dataset 1 as a whole), or generated using methods described in this chapter (Dataset 2).

2.1.1.1 Dataset 1: Hadfield *et al.* Data

Dataset 1 refers to the dataset generated by Hadfield *et al*[Hadfield et al., 2015]. This dataset is split into two subsets based on how they were assembled (Datasets 1.1 and 1.2).

2.1.1.2 Dataset 1.1: The Published Hadfield *et al.* Genome Assemblies

Dataset 1.1 refers to the assemblies generated by Hadfield *et al.*, using the 18 fragment *C. parvum* IowaII assembly [Abrahamsen et al., 2004].

2.1.1.3 Dataset 1.2: The Reassembled Hadfield *et al.* Genomes

Dataset 1.2 refers to the assemblies carried out here. These assemblies used the updated *C. parvum* IowaII reference genome for *C. parvum*, and the *C. hominis* TU502 genome for *C. hominis*, which bear non-fragmented chromosomes (totalling all 8 large chromosome scale scaffolds).

Reference to 'Dataset 1' refers to either the isolates in a general sense or the raw data used to generate the assemblies in Datasets 1.1 and 1.2. The 18 fragment Hadfield *et al.* assemblies in Dataset 1.1 were used to investigate the misassemblies caused by uneven depth of coverage. These were then reassembled using the pipeline detailed in Figure 2.2, to generate Dataset 1.2. These isolates were NOT subjected to DNA enrichment by WGA prior to sequencing. The genomes in Dataset 1 were used because they currently represent the largest collection of published *Cryptosporidium* draft genomes from clinical isolates.

2.1.1.4 Dataset 2: New Genome Assemblies

The assembly workflow detailed in this chapter (see Figure 2.2) was used to generate 29 new *C. parvum* and 19 new *C. hominis* assemblies. These isolates were subjected to DNA enrichment using WGA (by multiple displacement amplification) prior to sequencing. This new dataset of assemblies will hereafter be referred to as 'Dataset 2'. The isolation of oocysts, purification of DNA and subsequent WGA of Dataset 2, was done by myself as detailed in this chapter.

All datasets were used to investigate the correlation between genes transferred to chimeric regions and Gini, read distribution, and the effect of WGA on coverage.

2.2 Sequencing and Read Analysis Methodology

2.2.1 Oocyst purification and DNA preparation

Purification of *Cryptosporidium* Oocysts from Dataset 2 was carried out according to Hadfield *et al.*. *Cryptosporidium* oocysts where harvested from 1 - 2 ml of each faecal sample, using a saturated salt solution as a first pass purification approach. Oocyst numbers where then quantified by staining with FITC-labelled anti- *Cryptosporidium* monoclonal antibody (Crypto-Cel, Cellabs, Australia) and visualised under light microscopy using a Neubauer improved haemocytometer (C-Chip, Peqlab, Sarisbury Green, UK) to count.

The oocysts where then subjected to a second pass purification step using an Isolate®

IMS kit (TCS Biosciences, Botolph Claydon, UK) according to manufacturers standard documentation of use. Oocysts where dissociated from the ferromagnetic beads during IMS purification and decanted into 1.5ml microfuge tubes (Ekay, Basingstoke, UK), for surface sterilisation by iterative bleaching and washing steps. One cohort of UKP2 being isolated using an alternative Caesium chloride (CsCl) gradient centrifugation protocol. Oocysts where enumerated by microscopy prior to and following surface sterilisation.

Oocysts were shattered using an iterative liquid nitrogen freeze thaw protocol.

2.2.2 DNA library Preparation and Sequencing

Barcoded paired-end libraries were prepared for each isolate using the Nextera XT DNA sample preparation kit (Version C protocol, Illumina). Amplicons exceeding 500bp in size were selected during the post-PCR purification stages (however, this was not subjected to verification). DNA quantification by Qubit HS DNA assay demonstrated that, for the majority of the samples, the Nextera XT libraries were very low concentration. Consequently, these libraries were pooled to give 0.1nM concentration, and subjected to SpeedVac to dramaticaly improve concentration. In the pilot phase, the libraries were sequenced on the Illumina MiSeq platform using 2 x 151 bp reads, and on the Illumina HiSeq 2500 platform using a 2 x 151 bp rapid run during the main phase [Hadfield et al., 2015].

2.2.3 Raw read analysis

The reads were mapped to a reference genome (*C. parvum* IowaII for *C. parvum* and *C. hominis* TU502 [Xu et al., 2004] for *C. hominis*) using Bowtie2 v2.3.3.1 [Langmead et al., 2009]. Coverage analysis was then performed using Samtools v1.5 [Li and Durbin, 2009].

Depth of coverage over each assembly was calculated using the 'depth' tool within the samtools package [Li and Durbin, 2009]. The Gini coefficient is a metric used to measure the inequality within a dataset. It is commonly used in economics to measure the distribution of income within a population, where it is represented by a value between 0 and 1, with 0 representing perfectly even distribution, and higher values representing higher inequality of distribution. Here I have applied this coefficient to measure inequality of depth of coverage across a genome. For each of the genomes, I calculated the Gini coefficient of read depth. The Gini coefficient is calculated as:

$$G = A/(A + B)$$

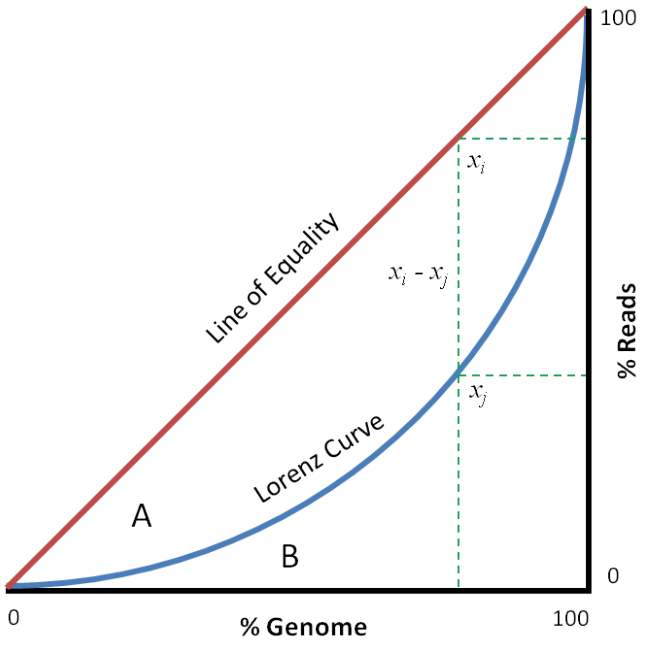


Figure 2.3: Graphical representation of the Gini coefficient. In this graph, the Gini coefficient can be calculated as $A/(A + B)$, which represented area under the Lorenz curve (blue) inversely proportional to the line of equality (red). The green dotted lines denote the percentage of reads which cover 80% of a genome used to generate the Lorenz curve (unequal coverage depth) as compared to a perfectly equal distribution of reads.

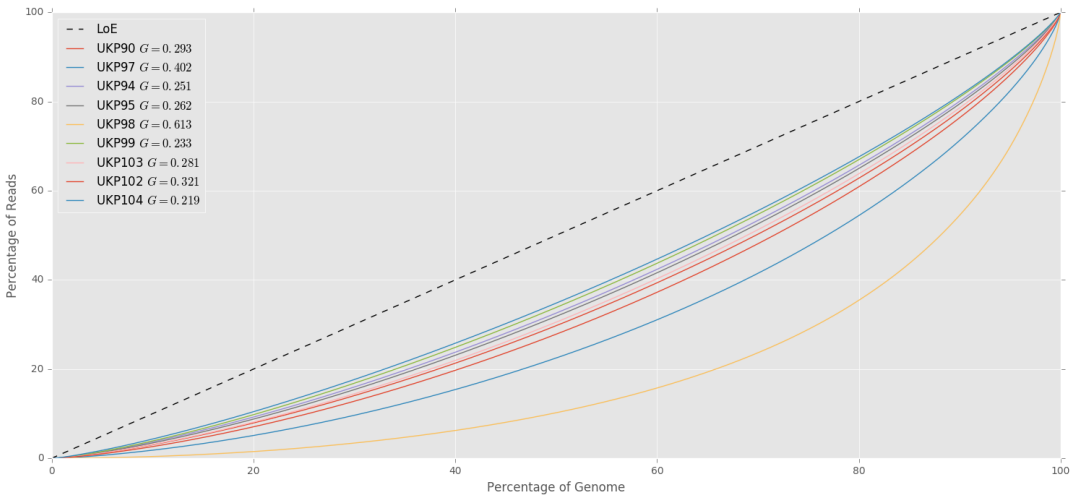


Figure 2.4: Gini curves for a selected number of genome assemblies from Dataset 2. LoE refers to the Line of Equality, wherein theoretic perfect equality of the coverage from a single isolate is represented, achieving a Gini of 0. These Gini curves were generated using a window size of 500.

where A is the area under the line of equality, and B the area under the Lorenz curve, on the graph of distribution inequality (see Figure 2.3). The green dotted lines (marked at 80% on the x axis) in Figure 2.3 gives an example of how, in the dataset used to generate the Lorenz curve, 80% of the genome is covered by only 40% of reads (the value at the position of collision of the green dotted line on the y axis), whereas in a perfect distribution it would be covered by 80% of reads.

The algorithm for calculating a genome's Gini coefficient of read depth coverage involves first calculating the mean depth of coverage of 1bp windows ($W = 1$) over the genome. These windows are ordered according to their depth of coverage values, and these values rescaled between 0 and 100. This ordered set of read depth values is used to generate the Lorenz curve, L, where the value at every position i on the curve represents the sum of all values at positions $\leq i$. A line of equality, E, was generated to represent perfectly even distribution of reads across a genome. The difference between the values at each position on E and L is then calculated and the summed inverse proportional difference (the Gini coefficient) of these values calculated.

Here, the Gini coefficient of coverage across a genome is calculated as half the relative mean absolute difference. The mean absolute difference of a population is the average of the distance (absolute difference) between all given pairs across a population. Take a population, x, the mean absolute difference of this population, $\bar{D}$, is

$$\bar{D} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{n^2} \tag{2.1}$$

where x_i and x_j are positions i and j in population x. In the context of coverage analysis, this refers to the mean coverage across window i and j within the ordered set of all windows. To get the relative mean absolute difference, this must be divided by the average of x, $\bar{x}$.

It therefore follows that the Gini coefficient can be calculated as

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \tag{2.2}$$

which is mathematically equivalent to $G = A/(A + B)$.

The Gini coefficient for each genome represents the unevenness of read depth across the genome sequence (see Figure 2.1A for an example of uneven coverage across chromosome 7 of UKP3 as compared to Iowa II).

2.2.4 Gini-Granularity Curves

To further investigate the read distribution throughout each assembly, the Gini coefficient was calculated across window sizes of 1-10,000 nucleotides. Furthermore, these curves were normalised such that the Gini at maximum data granularity (which is obtained by calculating G at window size of 1) is adjusted to 1 for the purpose of simplified comparison. For each of these normalised and un-normalised curves, the area under the curve (AUC) was calculated. These curves are hereafter referred to as Gini-granularity curves.

More formally, consider a nucleotide sequence s where the array of depth of coverage for each position in s is c. A partitioned array of c is generated using window size w, forming N partitions where $N = \frac{|c|}{w}$:

$$P_w^c = \{c_{[j,j+w]} \mid j = iw,\ 0 \leq i < N,\ i \in \mathbb{N}\} \tag{2.3}$$

The mean coverage over each partition is consequently:

$$C_w^c = \{\bar{n} \mid n \in P_w^c\} \tag{2.4}$$

where $\bar{n}$ is the mean depth of coverage of partition n. Taking G_w^c as the Gini of C_w^c (see Equation 2.2 for the calculation of the Gini coefficient), an array of Gini values over a range of partition sizes is, $r = [i, j]$ where $r \subset \mathbb{N}$ is:

$$G_r^c = \{G_w^c \mid w \in r\} \tag{2.5}$$

A_r^c is the area under the curve generated by G_r^c. The normalised area under the curve, $normA_r^c$ is calculated as the area under $normG_r^c$ where:

$$normG_r^c = \{G_i^c + (1 - G_1^c) \mid 0 < i \leq |G_r^c|,\ i \in \mathbb{N}\} \tag{2.6}$$

The area and the normalised area under the Gini granularity curve for a sample will hereafter be referred to as AUC and $nAUC$ respectively.

2.2.5 DNA Enrichment using Whole Genome Amplification

Due to the low DNA yield of *Cryptosporidium* isolated from clinical samples, WGA by Multiple Displacement Amplification (MDA) was utilised to enrich the DNA for sequencing. The protocol was followed as documented in the protocol: 'Amplification of Purified Genomic DNA using the REPLI-g Mini Kit' by Qiagen. This was carried out as follows: 5μl Buffer D1 was added to 5μl template DNA, vortexed to mix, and briefly centrifuged.

These were then incubated at room temperature for 3 minutes. During this time a master mix was prepared using 29μl REPLI-g Mini Reaction Buffer and 1μl REPLI-g Mini DNA Polymerase, to a total of 30μl. 10μl Buffer N1 was added to the samples and mixed by vortexing, and centrifuged briefly. The master mix was then added to 20μl of this denatured DNA, and incubated at 30^{o}C for 16 hours. After this incubation period, the REPLI-g Mini DNA Polymerase was inactivated by heating the sample at 65^{o}C for 3 minutes.

2.2.6 Sequencing Bias Analysis in WGA datasets

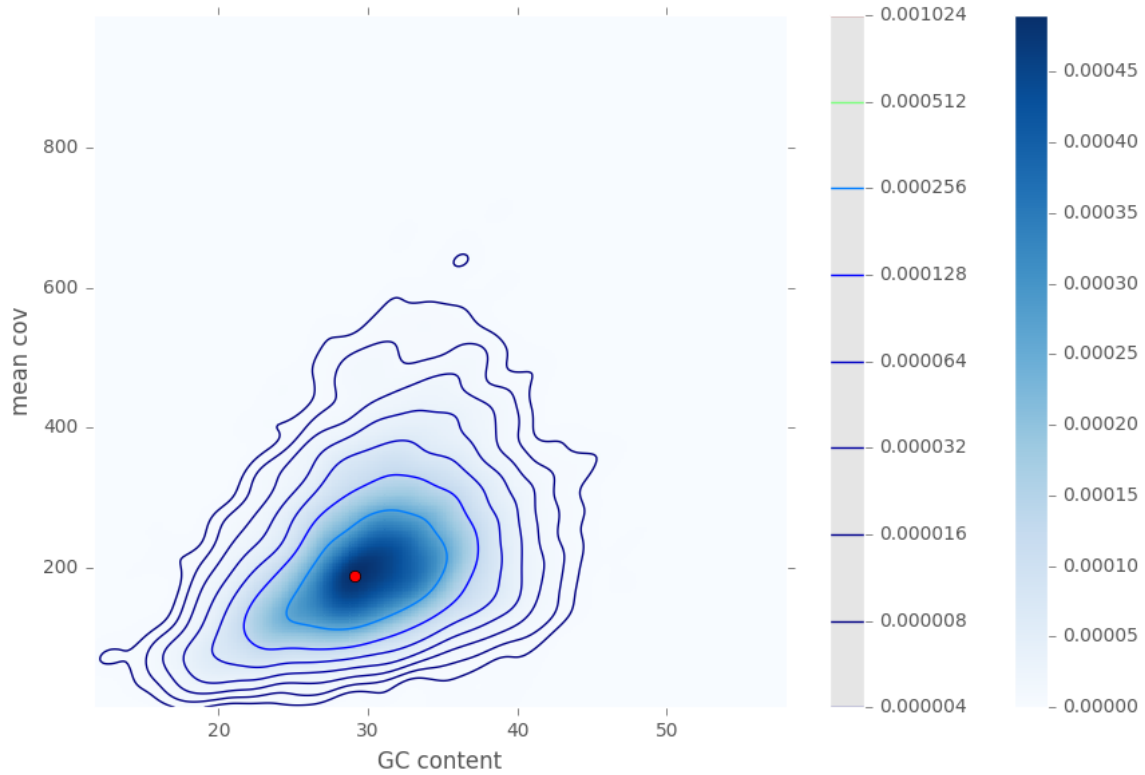


Figure 2.5: % GC content plotted against mean coverage over window sizes of 1000bp of *C. parvum* UKP99. This isolate was subjected to DNA enrichment by WGA (see Section 2.2.5). This plot was generated using kernel density estimation, and overlaid with contour lines, equating to thresholds (t). The red dot marks the calculated centre of mass of the graph object (see Equations 2.9 and 2.10 in Section 2.2.6.1). $R^2 = 0.365$ $G = 0.238$

To investigate bias which may exist in how DNA is enriched using WGA, or sequenced, I used the `depth` tool within the Samtools package. Bespoke python scripts were written to analyse the relationship between coverage and GC content across windows of various sizes. Kernel density estimation was carried out on these datasets to investigate the relationship between coverage and genomic content using the SciPy package in Python v3.6 [Jones et al., 2001].

The angular momentum of the graph objects generated by the distribution seen within these plots was calculated as described in Section 2.2.6.1. Angular momentum was used as it is a metric which considers the mass, shape, and size of an object. This makes it ideal for comparing the effect of WGA on a graph object, and by extension, a genome. Given a matrix K generated by the kernel density estimation of a dataset of % GC content against mean depth of coverage of windows over a genome, a graph object $A \subseteq K$ is defined as:

$$A = \{x \mid x \in K,\ x \geq t\} \tag{2.7}$$

where t is some threshold. Figure 2.5 shows an example of these objects, where a contour line is a threshold, and therefore defines the outer limits of a graph object.

2.2.6.1 Angular Momentum of a Rigid Body

When analysing graphs of mean coverage vs GC content of a genome assembly (such as that seen in Figure 2.5), there were two major variables which were important in interpretation: the size of the distribution, and the shape of the distribution. The angular momentum of the graph objects (as defined in Equation 2.7) were therefore calculated, since angular momentum is a product of these two variables. Consequently, the effective shape of the distribution can be characterised by this single metric.

Consider a two dimensional system (the rigid body) of n particles of masses $m_1, m_2, ..., m_n$, whos coordinates are $(x_1, y_1), (x_n, y_n), ..., (x_n, y_n)$ respectively, situated at distances $r_1, r_2, ..., r_n$ from a centre of mass, cm. We can define the total mass of the system as

$$M = \sum_{i=1}^{n} m_i \tag{2.8}$$

and the coordinates for the centre of mass of the system as

$$x_{cm} = \frac{m_1 x_1 + m_2 x_2 + ... + m_n x_n}{m_1 + m_2 + ... + m_n} \tag{2.9}$$

and

$$y_{cm} = \frac{m_1 y_1 + m_2 y_2 + ... + m_n y_n}{m_1 + m_2 + ... + m_n} \tag{2.10}$$

Let $v_1, v_2, ..., v_n$ be the linear velocities of the particles respectively, then the linear momentum of the first particle is $L_1 = m_1 v_1$. Since $v_n = r_n \omega$ and the linear momentum of a particle $= m_n(r_n \omega)$, the moment of linear momentum of the particle is therefore $= (m_n r_n \omega) \times r_n$, and the angular momentum $= m_n r_n^2 \omega$.

The angular momentum of a rotating rigid body is calculated as the sum of the moment of the linear momenta of all particles within the body:

$$L = m_1 r_1^2 \omega + m_2 r_2^2 \omega + ... + m_n r_n^2 \omega \tag{2.11}$$

$$= \omega(m_1 r_1^2 + m_2 r_2^2 + ... + m_n r_n^2) \tag{2.12}$$

$$= \omega \left(\sum_{i=1}^{n} m_i r_i^2 \right) \tag{2.13}$$

$$\therefore L = \omega I \tag{2.14}$$

where $I = \sum_{i=1}^{n} m_i r_i^2$, the moment of inertia of the rotating rigid body around the centre of mass.

2.3 Assembly and Post-Assembly Improvement Methodology

Assembly was initially carried out using the raw reads from Dataset 1 (published by Hadfield *et al.*), and the updated (8 chromosome) reference genome assemblies for reference guided scaffolding. This generated Dataset 1.2. The pipeline used to produce the reliable genome assemblies from the Dataset 1.2 (see Figure 2.2) was then used to generate assemblies from Dataset 2.

2.3.1 *De novo* Assembly

First *de novo* assembly was undertaken in the same manner as those reported by Hadfield *et al.* (2015). SPAdes v3.7.1 [Bankevich et al., 2012] *de novo* assembler was used to construct scaffolds from paired end read files. K-mer sizes of 23, 33, 55, 65, 77 & 89 were used in the assembly, with 1 iteration used for error correction, repeat resolution was enabled and the coverage cut off set to 'off'. Various k-mer sizes, coverage cut-offs, repeat masking, and a reference guided assembly approach were used in an attempt to improve assembly quality.

Velvet v1.2.10 assembler [Zerbino and Birney, 2008] was run in two parts, as is standard. The hashing program, `velveth` was run using a maximum k-mer length of 31, and options of `-fastq`, `-separate`, and `-short`. The De Bruijn graph construction program, `velvetg` was run with coverage cutoff set to `auto`, alignments exported for assembly analysis, long nodes which where eliminated by coverage filters exported, and `-coverage_mask` set to 2.

A third assembly was undertaken using IDBA-UD [Peng et al., 2012], to resolve low coverage regions whilst attempting to prevent generation of chimeric fragments during assembly and scaffolding. Read files in .fastq format were merged using `fq2fa` prior to assembly with IDBA-UD. IDBA-UD was run using these merged read files, and using default parameters.

2.3.2 Post Assembly Processing

The assemblies were improved using the Post Assembly Genome Improvement toolkit (PAGIT) [Swain et al., 2012]: a pipeline consisting of four standalone tools with the aim of improving the quality of genome assemblies. The tools are, in suggested order of execution: ABACAS [Assefa et al., 2009], IMAGE [Tsai et al., 2010], ICORN [Otto et al., 2010], & RATT [Otto et al., 2011].

The workflow of this assembly pipeline can be found in Figure 2.2. The differences between this workflow and the one used by Hadfield *et al.* is in the *de novo* assembler used (IDBA-UD rather than SPAdes) and the updated reference genome assemblies, of which the *C. parvum* IowaII reference genome was in 8 chromosomes rather than the 18 fragment version used in their approach.

2.3.2.1 ABACAS: Algorithm Based Automatic Contiguation of Assembled Sequences

ABACAS is a contig-ordering and orientation tool which is driven by alignment of the draft genome against a suitable reference. Suitability of the reference is defined by amino acid similarity of at least 40%. Alignment is performed by NUCmer or PROmer from the MUMmer package [Kurtz et al., 2004]: a tool designed for large scale genome alignment. Contigs from the draft assembly are positioned according to alignment to the reference genome, with spaces between the contigs being filled with 'N's, generating a scaffold of the draft assembly.

ABACAS was executed using the updated (All 8 chromosomes resolved) *C. parvum* IowaII [Abrahamsen et al., 2004] reference genome with default parameters.

2.3.2.2 IMAGE: Iterative Mapping and Assembly for Gap Extension

IMAGE uses Illumina paired end reads to extend contigs by closing gaps within the scaffolds of the draft genome assembly. IMAGE uses read pairs where one read aligns to the end of a contig and the other read overhangs beyond the end of the contig into the gap. This gap can then be partially closed using the overhanging sequence and by extending

the contig.

IMAGE was run in groups of three iterations at k-mer sizes of 91, 81, 71, 61, 51, 41, & 31, totalling 21 iterations. Scaffolding was then performed with a minimum contig size of 500, joining contigs with gaps of 300 N's.

2.3.2.3 ICORN: Iterative Correction of Reference Nucleotides

ICORN was developed to identify small errors in the nucleotide sequence of the draft genome, such as those which may occur due to low base quality scores. It was designed to correct small erroneous indels, and is not suitable for, or capable of, correcting larger indels or misassemblies.

ICORN was run using 8 iterations and a fragment size of 300.

2.3.2.4 RATT: Rapid Annotation Transfer Tool

RATT is an annotation transfer too used to infer orthology/homology between a reference genome and a draft assembly. This is achieved by utilising NUCmer from the MUMmer package to identify shared synteny between annotated features within the reference genome, and sequence within the draft assembly. Annotation files (EMBL format) are produced which contain regions which are inferred to be common features. The regions are filtered and transferred dependant on whether the transfer is between strains (Strain, similarity rate of 50-94%), species (Species, similarity rate of 95-99%), or different assemblies (Assembly, similarity rate of >=99%).

RATT was run using IowaII annotations in EMBL format, downloaded from CryptoDB, as a reference. The Strain parameter was used to transfer feature annotations to the draft assembly.

2.3.3 Analysis of Draft Genomes

VNTR's around the reference and draft genomes were identified for the purpose of VNTR comparison and polymorphism analysis. Tandem Repeats Finder v4.09 [Benson, 1999] was used to identify VNTR's around the *C. parvum* IowaII reference genome using a matching weight of 2, mismatch and indel penalties of 5, match and indel probabilities of 80 and 10 respectively, minimum score of 50 and maximum period size of 15. The number of VNTR's per gene is included as a heat map in Figure 2.13.

2.3.4 Identification of Misassembly

The observation that only low coverage and low complexity regions formed the interface between chimeric contigs indicated that the chromosomal translocation events were a result of misassembly rather than an actual biological signal. Consequently, experimental validation was considered to be unnecessary.

The draft genomes were analysed in two ways (1) by transferring gene annotations from the reference genome to the drafts using RATT, and (2) by aligning the contigs (from IDBA-UD) or scaffolds (from SPAdes/Velvet) from the draft assemblies to the IowaII reference genome. RATT was used to identify the number of genes which were transferred between genomes: it provided a convenient way of identifying putative chimeric regions i.e. regions on a draft chromosome that contained genes from 2 or more reference chromosomes. NUCmer was then used to investigate these putative chimeric regions by performing whole genome alignments. NUCmer (from the MUMmer package [Kurtz et al., 2004]) was used with a minimum length of match set to 100, preventing the report of small regions of similarity, a maximum gap of 90, and a minimum cluster length of 65.

2.3.5 Quality assessment with Gini

The Gini coefficient for each isolate was calculated using the average coverage over windows of 1000 and plotted against the number of genes transferred to chimeric regions (detailed in section 2.3.4). The coefficient of determination (R^2) was used to calculate the amount of variance in the number of genes transferred to chimeric regions explained by the Gini coefficient.

2.3.6 Data Visualisation

The *C. parvum* assemblies (UKP2-8) were visualised alongside the *C. parvum* IowaII reference genome using the Circos package v0.69 [Krzywinski et al., 2009]. Mapped reads were visualised using Integrative Genomics Viewer v2.4.16 [Thorvaldsdóttir et al., 2013].

2.4 Results and Discussion for Sequencing and Read Analysis

Isolate	Proportion of reads mapping to Ref.	Fraction of Ref. covered	Mean Depth of Coverage	nAUC	$G \mid W_1$
Non-WGA					
UKP2*	0.93	1.00	51.8	0.892	0.223
UKP3*	0.89	0.99	166.42	0.815	0.556
UKP4*	0.89	0.99	192.48	0.805	0.566
UKP5*	0.85	0.99	26.86	0.884	0.277
UKP6*	0.82	0.99	104.83	0.921	0.255
UKP7*	0.89	0.99	77.85	0.804	0.555
UKP8*	0.84	0.98	174.39	0.796	0.566
UKP10	0.89	0.99	94.87	0.783	0.55
UKP11	0.89	0.99	145.87	0.783	0.57
UKP12	0.98	0.78	34.7	0.810	0.432
UKP13	0.95	0.70	73.87	0.783	0.499
UKP14	0.99	0.94	65.55	0.790	0.532
UKP15	0.93	0.98	134.890	0.827	0.402
UKP16	0.96	0.99	157.880	0.828	0.398
UKH3*	0.90	0.98	34.710	0.888	0.236
UKH4*	0.85	0.96	209.170	0.786	0.601
UKH5*	0.81	0.96	201.920	0.777	0.584
WGA					
UKP90	0.96	1.00	111.850	0.969	0.297
UKP94	0.95	1.00	116.179	0.976	0.255
UKP95	0.96	1.00	183.517	0.968	0.267
UKP97	0.75	1.00	122.642	0.972	0.406
UKP98	0.28	0.98	179.686	0.968	0.615
UKP99	0.96	1.00	224.938	0.970	0.238
UKP102	0.89	1.00	204.001	0.972	0.324
UKP103	0.95	1.00	125.109	0.964	0.286
UKP104	0.97	1.00	249.685	0.971	0.224
UKP106	0.96	1.00	103.495	0.971	0.242
UKP107	0.78	1.00	123.906	0.964	0.501
UKP118	0.97	1.00	123.549	0.966	0.25
UKP119	<0.01	0.17	123.305	0.945	0.638
UKP120	0.96	1.00	171.077	0.969	0.272
UKP121	0.90	1.00	127.770	0.972	0.313
UKP122	0.78	1.00	106.596	0.968	0.433
UKP123	0.11	0.66	140.564	0.961	0.765
UKP124	0.38	0.98	186.368	0.965	0.59

Continued on next page

Table 2.1 – continued from previous page

Isolate	Proportion of reads mapping to Ref.	Fraction of Ref. covered	Mean Depth of Coverage	nAUC	$G \mid W_1$
UKP125	0.95	1.00	119.095	0.969	0.243
UKP126	0.09	0.79	121.587	0.958	0.667
UKP127	0.53	0.99	121.092	0.969	0.502
UKP128	0.90	1.00	117.884	0.971	0.349
UKP129	0.53	0.99	106.458	0.969	0.452
UKP130	0.98	1.00	129.992	0.966	0.222
UKP131	0.97	1.00	229.840	0.964	0.219
UKP132	0.03	0.43	122.373	0.957	0.749
UKP133	0.40	0.98	126.124	0.967	0.556
UKP134	0.92	1.00	123.355	0.975	0.277
UKP135	0.16	0.80	124.486	0.963	0.733
UKH51	0.97	1.00	227.015	0.970	0.215
UKH55	0.97	1.00	207.896	0.971	0.218
UKH56	0.97	1.00	258.286	0.969	0.217
UKH57	0.90	0.99	123.245	0.970	0.227
UKH58	0.80	0.94	162.412	0.935	0.514
UKH59	0.90	0.99	149.380	0.970	0.224
UKH60	0.90	0.99	179.545	0.970	0.214
UKH61	0.92	0.99	199.525	0.969	0.212
UKH62	1.00	1.00	221.506	0.969	0.213
UKH63	1.00	1.00	232.824	0.976	0.492
UKH64	1.00	1.00	232.824	0.976	0.492
UKH65	1.00	1.00	207.611	0.975	0.329
UKH66	1.00	1.00	208.255	0.972	0.234
UKH67	0.95	1.00	184.459	0.975	0.355
UKH68	0.96	1.00	214.868	0.976	0.265
UKH69	0.96	1.00	203.663	0.974	0.248
UKH70	0.96	1.00	198.205	0.976	0.285
UKH71	0.96	1.00	205.866	0.973	0.254
UKH72	0.96	1.00	220.255	0.977	0.391

Table 2.1: Bowtie2 mapping statistics for *C. parvum* and *C. hominis* reads. The Gini coefficient is included in this table as an indication of uneven depth of coverage (IowaII=0.112). *Cryptosporidium* reads were mapped to appropriate reference genomes for each species: **C. parvum* IowaII, †*C. hominis* TU502. Included is the the area under the normalised Gini granularity curves as an indication of read distribution (see Section 2.2.4). *Isolates which were originally published by Hadfield *et al.* and updated using the described workflow. All other genomes were newly assembled and analysed. Isolates highlighted in red cover an insufficient portion of the reference genome, and are therefore considered to have failed in the objective of sequencing the entire genome.

Table 2.1 indicates high depth of coverage inequality throughout the genomes, represented by relatively high Gini coefficient values in comparison to that exhibited by the *C.*

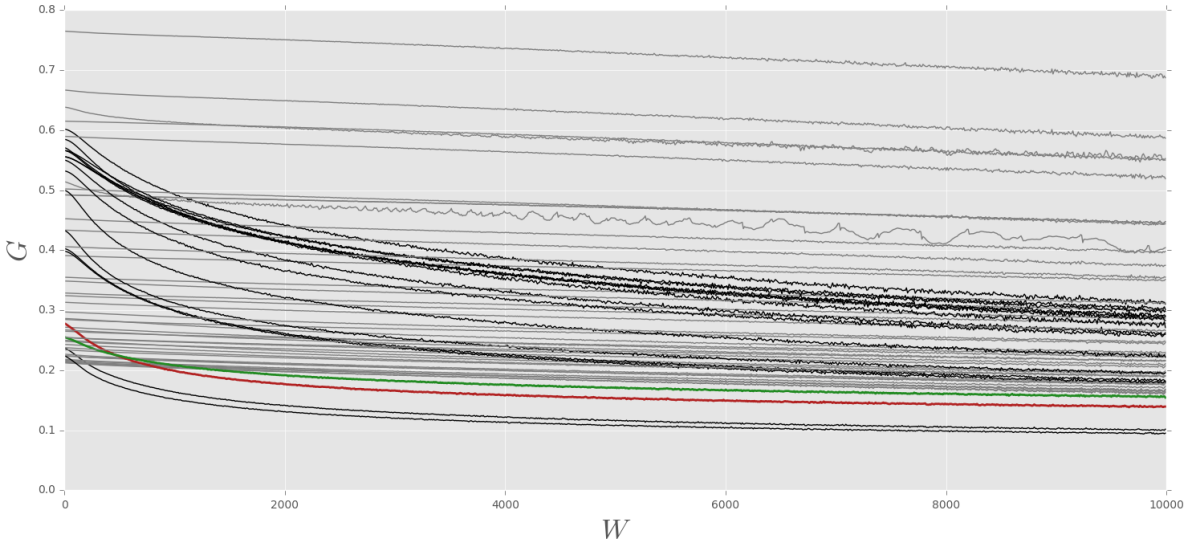


Figure 2.6: Gini granularity curves generated from Gini values (G) calculated using different window sizes (W) for the genome assemblies presented in this chapter. Black samples were not subjected to DNA enrichment prior to sequencing, grey samples were subjected to DNA enrichment by WGA (see Section 2.2.5). Samples of interest are colourised as: red = UKP5, green = UKP6.

parvum Iowa II reference genome ($G \mid W_1 = 0.112$), which the mean depth and breadth of coverage (fraction of the reference covered) will not indicate. This appears to be a common issue when sequencing intestinal protozoans from human clinical samples. Paired end read libraries accessed from GenBank, sequenced by the Welcome Trust Sanger Institute (Bioproject PRJEB3213), and those published by Troell *et al.* (2016) (Bioproject PRJNA308172), who was attempting to generate whole genome sequences from single cells using whole genome amplification [Troell et al., 2016], also suffered from very high Gini coefficients, indicating that this problem is not restricted to a single research team. See Figure 2.1 for an example of how the Gini value corresponds to actual read depth variation.

Gini granularity curves are presented here as a more complete indication of the coverage over a genome. The premise behind this is based on two shortcomings of using the Gini coefficient alone as a measure of depth of coverage inequality:

- Two genomes with identical ordered coverage arrays will produce identical Lorentz curves, and therefore an identical Gini. This does not take into account the distribution of depth of coverage across the genome.
- The Gini coefficient is known to be confounded by data granularity [Monfort, 2008].

Curves generated from Gini granularity analysis are found in Figures 2.6 and 2.7. These curves show a similar set of characteristics, which can be defined by two phases:

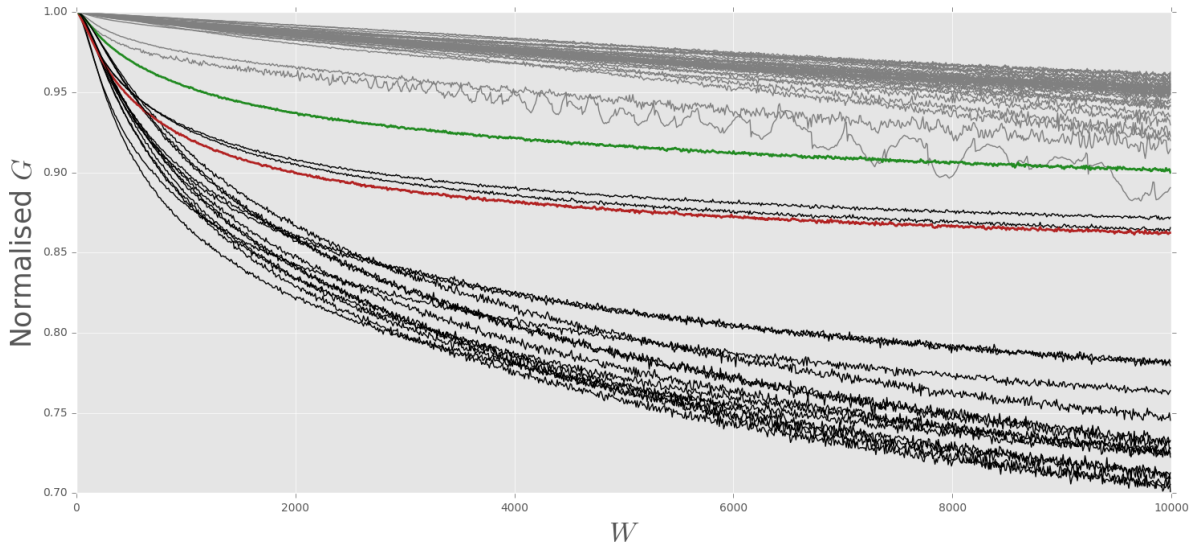


Figure 2.7: Normalised Gini granularity curves generated from normalised Gini values (G) calculated using different window sizes (W) for the genome assemblies presented in this chapter. Black samples were not subjected to DNA enrichment prior to sequencing, grey samples were subjected to DNA enrichment by WGA (see Section 2.2.5). Samples of interest are colourised as: red = UKP5, green = UKP6.

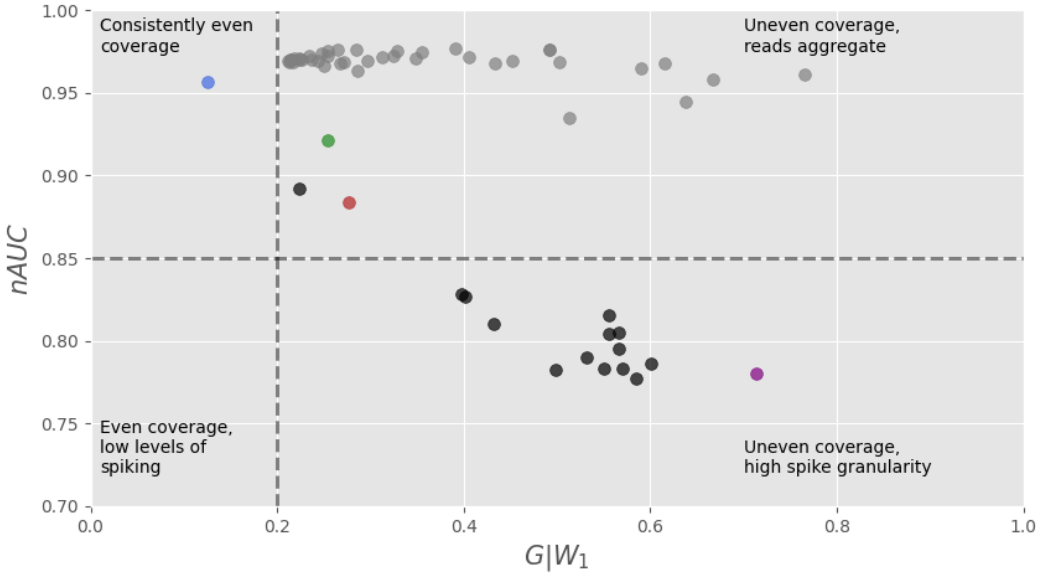


Figure 2.8: $G \mid W_1$ plotted against $nAUC$. Four different cases are presented in each corner as an indication of how this graph can be interpreted. Black samples were not subjected to DNA enrichment prior to sequencing, grey samples were subjected to DNA enrichment by WGA (see Section 2.2.5). Samples of interest are colourised as: red = UKP5, green = UKP6. Blue and purple samples belong to *Cyclospora cayetanensis* genomes, and are included to demonstrate that these areas of the graph can be populated.

1. Decline phase: The Gini value (G) decreases quickly as the window size (W) increases.

2. Perturbation phase: The Gini value plateaus, and perturbation increases, as window size increases.

The two phases of the Gini granularity curves (Figures 2.6 and 2.7) may be indicative of characteristics of each dataset, and the method by which they were generated. The magnitude of the drop exhibited during the decline phase appears to vary considerably, with some isolates presenting a large decrease in G over smaller window size increases (e.g. UKH5), and others presenting with very little drop. The initial drop in G may be as a result of the window size being less than the insert size of fragments used to generate these reads. Variation in the rate of drop during the decline phase may indicate higher coverage spiking[1]. There appears to be some amount of variation in the characteristics of the curves generated for each species. For example, the *Cryptosporidium* data (seen in grey) differentiates into two discrete groups in Figure 2.8, when using $G \mid W_1$ (the Gini calculated using a window size of 1, and therefore at maximum granularity). This is also seen in Figure 2.4:

1. UKH4, UKP3, UKP4, UKP7, & UKP8 exhibiting $G = \ 0.55 - 0.60$

2. UKH3, UKH5, UKP2, UKP5, & UKP6 exhibiting $G = \ 0.22 - 0.28$

The magnitude and length of the decline phase appears to vary depending on the Gini value calculated over single base windows ($W = 1$ or W_1). Greater $G \mid W_1$ values appear to exhibit an extended decline phase (see Table 2.1). Likewise, the perturbation phase differs between the two groups, where group 1 (higher G) levels off at a much slower rate, and shows large levels of G perturbation, and group 2 (lower G) levels off at a quicker rate and shows lower levels of G perturbation. The variation in the characteristics of the perturbation phase may be as a result of a number of factors, such as the level of noise within the dataset, genome incompleteness, and the number of contigs within the final assembly. These results indicate that the analysis of these Gini curves hints at coverage features which are lost by considering a single Gini value alone.

Within high Gini isolates, a lower area under the normalised curve indicates uniform spiking, whereas a higher area under the normalised curve indicates aggregation of reads throughout the genome (see Figure 2.7). Analysis of these curves allows for a more comprehensive analysis of problematic genomes with high depth of coverage inequality,

[1]'Spiking' and 'blocking' are used here to describe the density of peaks and troughs in coverage across a sequence, wherein spiking refers to a high spike granularity and a larger number of peaks and troughs within a sequence, and blocking refers to low spike granularity and the aggregation of reads into areas of high coverage, resembling blocks. See Figure 2.1B for examples of spiking and blocking.

wherein spiking indicates a general problem with sequencing, and high read aggregation indicates problems sequencing particular regions. Colourised are curves of particular interest, such as green (UKP6) and red (UKP5) which represent the differences which may be exhibited by two genomes with similar $G \mid W_1$. These curves suggest two different distribution types, due to UKP5 bearing a more pronounced decline phase than UKP6, and therefore bearing a lower area under the normalised Gini-granularity curve, suggesting greater levels of spiking. Figure 2.1B shows the coverage over chromosome 1 for both of these genomes to be very different in character, despite there being only a 0.002 different in Gini at absolute granularity. Coverage over UKP6 appears to present as relatively even, but with localised spikes of coverage reaching and exceeding 1000x ('blocking'). In contrast UKP5 presents with non-localised homogeneous pronounced spiking, with little width (high spike-granularity), and reach similar depth. This should serve as a clear example of how the difference in the $nAUC$ of genomes which bear similar $G \mid W_1$ related to the distribution of read coverage across a genome.

Figures 2.6 and 2.7 illustrate that WGA significantly alters the distribution of coverage over a genome (seen in the grey samples, which were subjected to WGA). It appears to effectively remove the decline phase in Gini-granularity curves using these window sizes, increasing the $nAUC$ of a sample. This results in an upward drift on the $nAUC$ vs $G \mid W_1$ plot seen in Figure 2.8. The position of the isolate within Figure 2.8 indicates the nature of the read distribution across a genome, wherein $nAUC$ describes the amount of spiking or blocking (aggregation) of the reads, and $G \mid W_1$ describes the unevenness of coverage. Consequently, the ideal coverage profile of an isolate would be high read aggregation (low $nAUC$, indicating blocking) and low inequality of coverage (low $G \mid W_1$). This would indicate the read coverage across the genome is evenly distributed. $nAUC$ becomes more descriptive of the read distribution across a genome as $G \mid W_1$ increases, where it will indicate the level to which reads aggregate (blocking), or whether unevenness of coverage is universal across the entire genome (spiking). Figure 2.9 is a graphical representation of the how $nAUC$ and $G \mid W_1$ characterises the distribution of data (in this case, reads) within a dataset (in this case, across a genome). One criticism of this figure is that it is not clear how the bottom left (represented by even coverage and high spiking) can be populated, since as $G \mid W_1$ trends to 0, the $nAUC$ is likely to trend to 1, since the change in in $G \mid W_n$ across values of n will also decrease. Assemblies with perfect coverage equality could be interpreted as: all reads aggregate to a single block which spans the entire assembly, rather than organising into discrete blocks across the genome.

WGA acts to amplify large ($>$ 10kb) regions of the genome. These results suggests a bias in which regions are amplified, which results in significant coverage aggregation

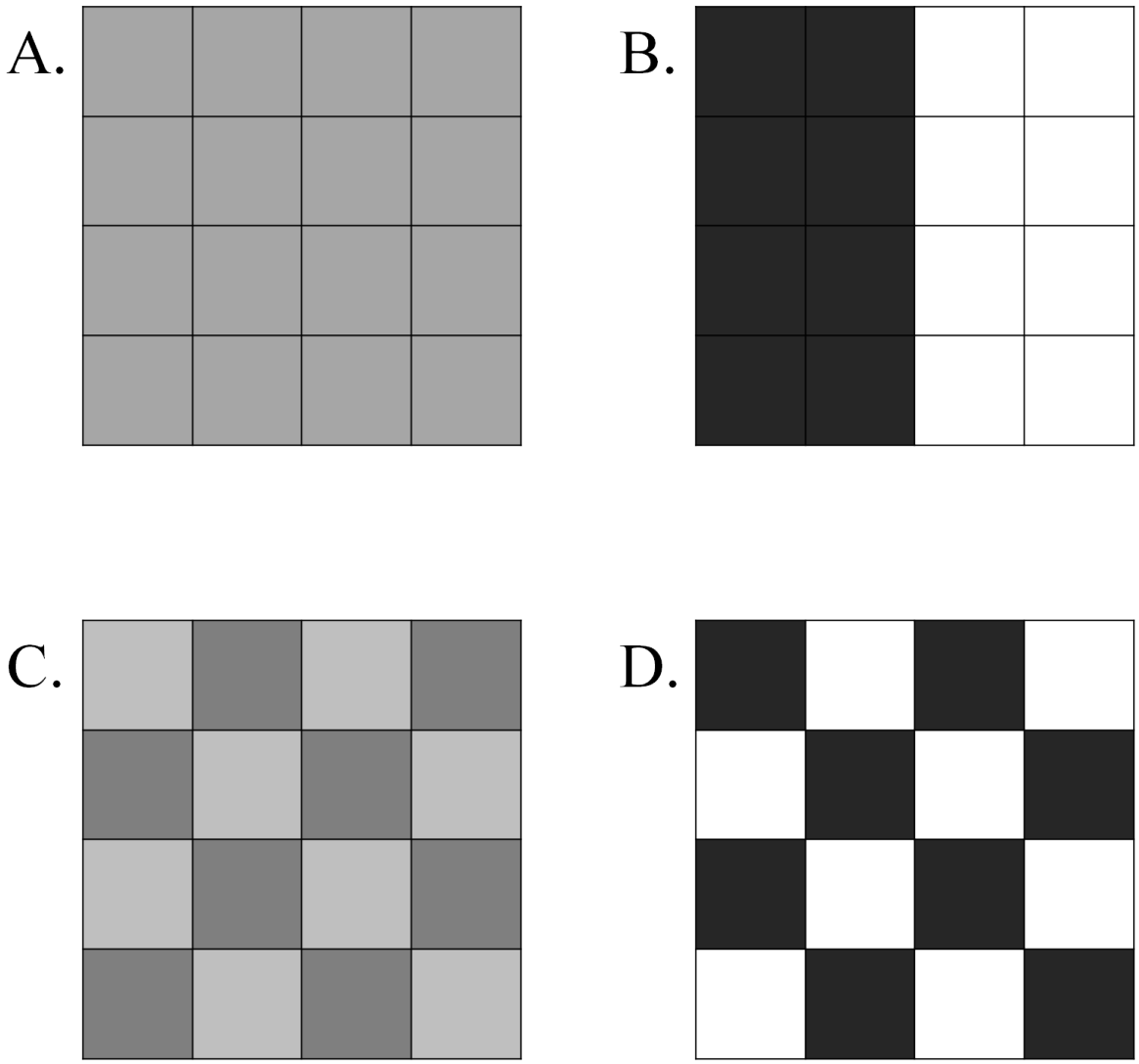


Figure 2.9: A graphical representation of the types of distribution of data characterised by comparing $nAUC$ and $G \mid W_1$, wherein the darker the tone of the cell, the more data (as a proportion of the whole) that cell contains. Cases A, B, C & D are equivalent in position to the four cases seen in Figure 2.8. Set A ($G \mid W_1 = 0.00$) represents high $nAUC$ and low $G \mid W_1$, wherein data is evenly distributed and aggregated (though in this case, the location at which the data aggregates represents the whole set). Set B ($G \mid W_1 = 0.50$) represents a high $nAUC$ and high $G \mid W_1$, wherein data is highly aggregated and unevenly distributed, resulting in high aggregation of data within the set. Set C ($G \mid W_1 = 0.25$) represents a low $nAUC$ and a low $G \mid W_1$, wherein data is moderately evenly distributed, but some spiking is present. Set D ($G \mid W_1 = 0.50$) represents a low $nAUC$ and a high $G \mid W_1$, wherein data is not aggregated and not evenly distributed, leading to high spiking. Gini values here are approximate and used only as an example.

(blocking, indicated by low $nAUC$) in large regions which are preferentially amplified during this process. This would lead to a decrease in $nAUC$, as seen in these data. Since G was calculated over window sizes of between 1-10000 bases, the decline phase would be effectively eliminated. However, increasing the window size beyond the average size of the regions amplified during the WGA process may resolve these decline phases.

Considering the effects of WGA, Figure 2.10 shows kernel density estimation plots from 4 isolates of *C. parvum*, where GC content is plotted against mean coverage over windows of 1000 nucleotides (see Section 2.2.6). The isolates within this dataset can be split into two cohorts:

- Genomes which were sequenced from DNA extracted and purified from clinical isolates without any further processing to enrich DNA prior to sequencing.
- Genomes which were sequenced from DNA extracted and purified from clinical isolates and then enriched using WGA, as detailed in section 2.2.5.

These results show that in isolates where the Gini score is high, indicating low depth of coverage equality across the genome, coverage over regions exhibiting GC content of 30% $\geq$ is significantly higher than that of regions exhibiting GC content of 30% $\leq$ (Figure 2.10a). In instances where Gini is low, indicating high depth of coverage equality across the genome, there is a less clear increase, or none whatsoever (Figure 2.10b). Gini scores from these genomes are variable, indicating variable levels of depth of coverage inequality (see Table 2.1).

The results illustrate much more dispersion in the graphs generated from enriched genomes than those generated from un-enriched genomes. These results indicate that the distribution, and subsequent shape of the graph, is at least partially associated with the Gini score. There appear to be two major distinct types of distribution based in the shape of the distribution and the position of the centre of mass (see Section 2.2.6 for a description of how the centre of mass was calculated) of the distribution:

Type I A very tight distribution with a greater than linear distribution, exhibiting a defined increase in coverage at 30% GC content. An example of this type of distribution can be seen in Figure 2.10a.

Type II A radially dispersed distribution with no clear trend, and a centre of mass at 30% GC content. An example of this type of distribution can be seen in Figure 2.10b.

These types account for the majority of all distributions seen within the dataset. However, there also appears to be two minor types of distribution which do not easily fit into the two major types:

Type III A linear distribution with a fairly dense centre of mass and some dispersion. An example of this type of distribution can be seen in 2.10d.

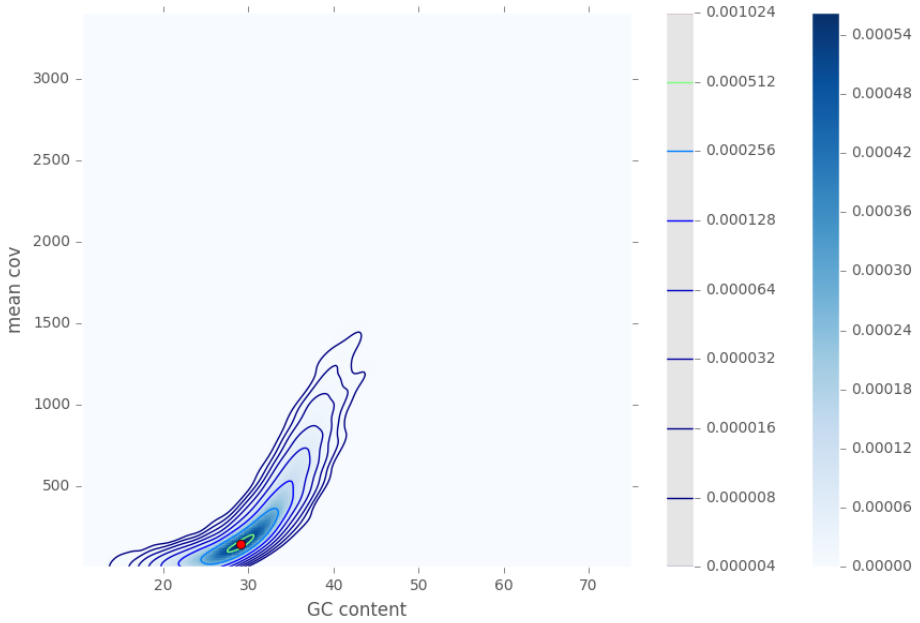
Type IV A large and widely dispersed distribution, with some windows exhibiting very high coverage relative to the centre of mass. Low density centre of mass, which lies at 30% GC content and low coverage. An example of this type of distribution can be seen in Figure 2.10c, where coverage over some windows exceeds 4000x. This may be considered to be a variation on type II dispersion, as when a maximum coverage cut-off is applied, the dispersion appears to be type 2.

The centre of mass within these kernel density arrays indicates the mean % GC content and coverage across the genome. Since the plots here described are all generated using *Cryptosporidium* genome assembles, the centre of mass is expected to be at a similar location (30% on the x-axis), however, the mean coverage (y-axis) is expected to vary. The shape of the graph object (see Section 2.2.6 for a working definition of 'graph object') is therefore dictated by the distance of all points with intensity greater than a given threshold within the kernal density array from the centre of mass. The angular momentum will therefore vary dependant on the shape and density of a graph object, and consequently the relationship between % GC content and coverage over the genome.

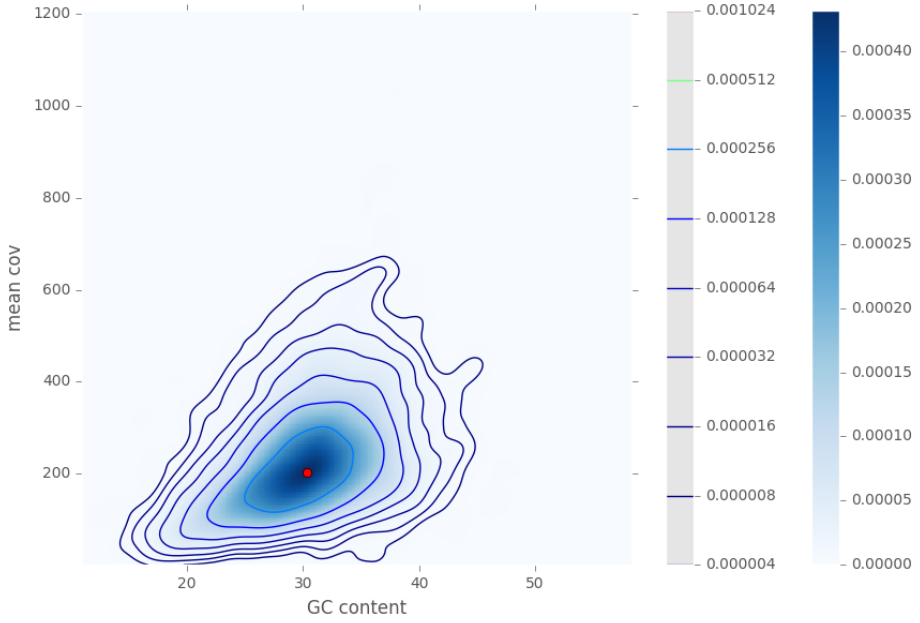
The results detailed in Figure 2.10 illustrate that the implementation of WGA as a means to enrich DNA prior to sequencing significantly alters the coverage over the genome. Rather than increasing the mean depth of coverage over the genome uniformly, however, more our analysis highlights that it selectively amplifies certain sequences. Due to the existing GC bias which has been reported within Illumina sequencing data [Benjamini and Speed, 2012] (illustrated by Type I distribution, as seen in UKP4 in Figure 2.10), this has the effect of **compensating for this bias**, resulting in a much more radially dispersed distribution with a far less clear positive correlation between coverage and GC content. These results highlight the value of using DNA enrichment by WGA for generating high quality, reliable genome assemblies from clinically isolated *Cryptosporidium* samples. However, it also hints at a complex relationship between depth of coverage and sequence content across enriched genomes.

Figure 2.11 shows Gini content plotted against angular momentum of distribution objects exhibited in GC content vs coverage plots, such as those shown in Figure 2.10. Angular momentum is calculated for graphed distribution objects with density cut-offs (shown above the plot). These cut-offs are highlighted as contour lines in plots shown in Figure 2.10. As the cut-off threshold (t) increases, the correlation between Gini and angular momentum for both the WGA (red) and non-WGA (blue) datasets increase, with the highest correlation being seen at $t = 0.000256$, which exhibits R^2 values of 0.872 and

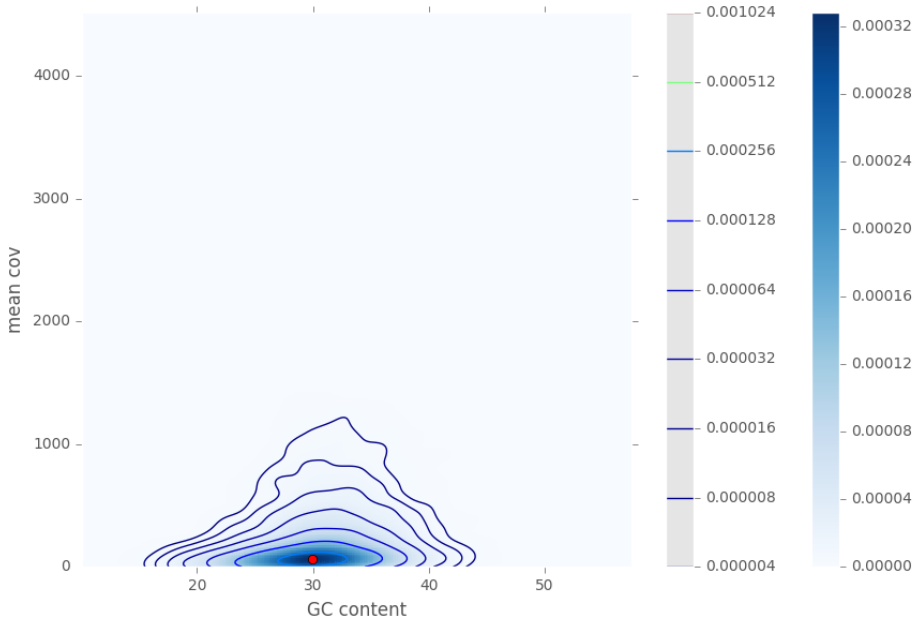
0.931 for WGA and non-WGA datasets respectively. The R^2 scores then drop considerably at $t = 0.000512$. Thresholds of higher than this exceeded the maximum density of some of the isolates, which resulted in data loss. There is a strong negative correlation between the Gini scores and the angular momentum of distributions in genomes which are subjected to DNA enrichment using WGA, and a similarly strong positive correlation in genomes which were not subjected to WGA. There also appears to be some clustering based on distribution types, with type I (circle) showing loose clustering at the upper-right area (high angular momentum and high Gini), type II (triangle) showing tight clustering primarily in the lower-middle/lower-right areas (moderate to high angular momentum and low Gini), type III (square) only being represented by a single datum in the lower left area (low angular momentum and Gini), and type IV (cross) moderately clustering in the upper left portion (low angular momentum and high Gini). These results indicate that the optimal cut-off threshold for elucidating structure within these data is $t = 0.000256$.



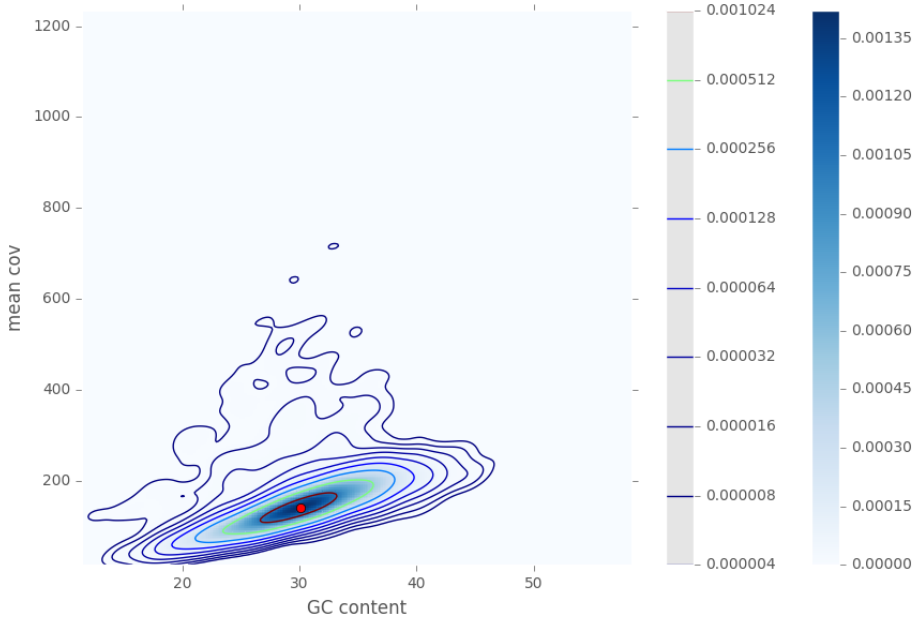
(a) UKP4 (non-WGA) $R^2 = 0.823$ $G = 0.4693$



(b) UKP94 (WGA) $R^2 = 0.396$ $G = 0.2539$



(c) UKP98 (WGA) $R^2 = 0.182$ $G = 0.6145$



(d) UKP6 (non-WGA) $R^2 = 0.455$ $G = 0.2508$

Figure 2.10: Coverage vs GC contents plotted within 1000bp windows for 4 UK isolated *C. parvum* genomes. UKP3 and UKP6 were not subjected to enrichment by a Whole Genome Amplification (WGA) process. DNA of UKP94 and UKP98 were enriched by a WGA process (ϕ29) prior to sequencing. The plots were generated using kernel density estimation, overlaid with contour lines, equating to thresholds (t). The red dot marks the calculated centre of mass of the graph object (see Equations 2.9 and 2.10 in Section 2.2.6.1).

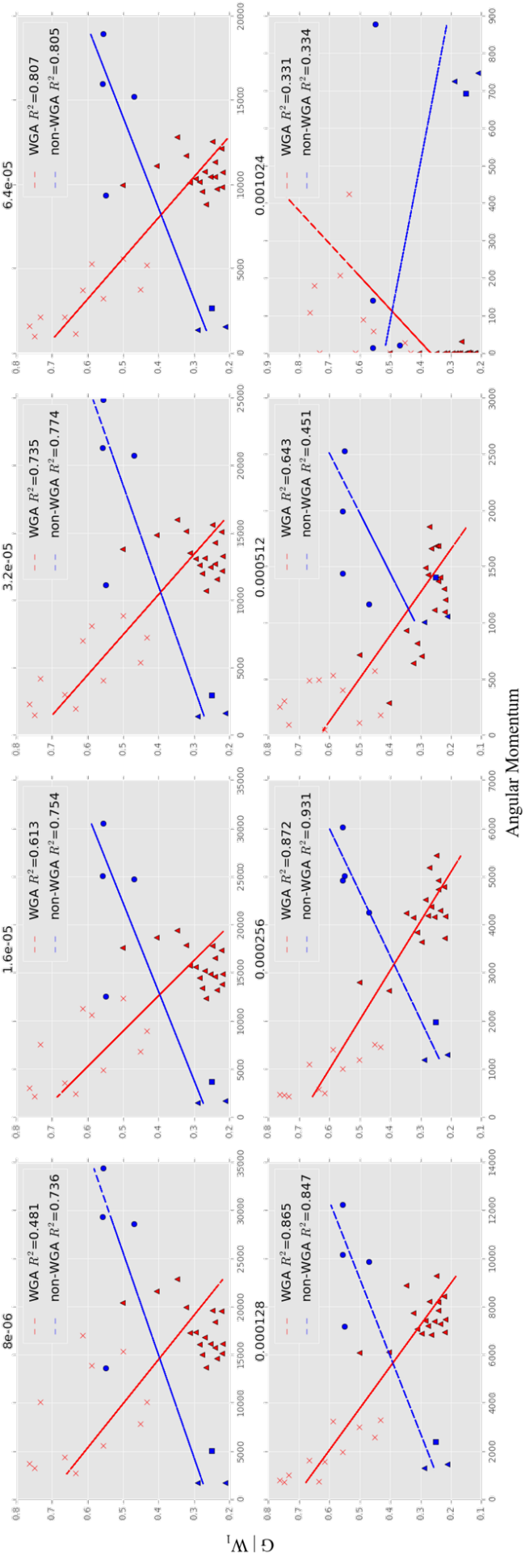


Figure 2.11: Angular momentum of distributions seen in GC-content/mean-coverage plots, taken as rigid bodies, against their corresponding Gini. Each marker refers to a single isolate. Angular momentum is calculated as detailed in Section 2.2.6.1. Each graph refers to angular momentum calculated using the stated density threshold. Blue markers refer to isolates which were not subjected to WGA. Red markers refer to isolates which were subjected to DNA enrichment using WGA. Marker shapes denote distribution type: circle - Type I, triangle - Type II, square - Type III, cross - Type IV. Marker colour and shape were manually annotated. Linear regressions for WGA and non-WGA datasets are included, with R^2 values.

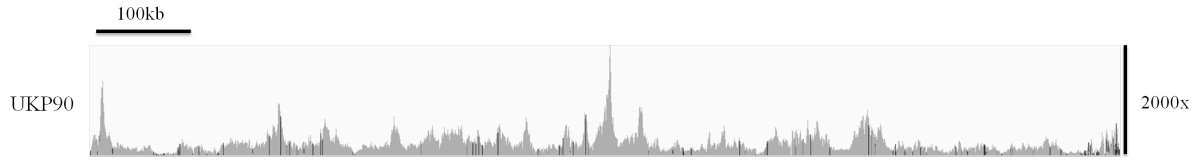


Figure 2.12: Read coverage across chromosome 4 of UKP90 (WGA) ($nAUC = 0.969$, $G \mid W_1 = 0.297$) showing high levels of read aggregation.

The results in Figure 2.11 demonstrate that there is a complex relationship between the Gini and the angular momentum of isolate genomes. The nature of this relationship depends on whether the DNA extracted from the isolate has been enriched using whole genome amplification (see Section 2.2.5 for a description of this process). Isolates which have been subjected to WGA appear to have angular momentum which bears strong negative correlation with their Gini score. Furthermore, the structure of the distribution graph is influenced strongly, where WGA appears to result primarily in a type II distribution, with some exhibiting a type IV distribution. This distribution correlates very strongly with low Gini coefficient. Furthermore, isolates which have not been subjected to WGA prior to sequencing illustrated a strong positive correlation between Gini and angular momentum. These non-WGA isolates exhibit primarily type I distribution, with some type II. Type I distribution (typical of non-WGA samples) clearly shows a bias towards coverage at higher GC content, characterised by a clear increase in coverage at 30% GC content (e.g. 2.10a), indicating that the process of DNA extraction and high-throughput Illumina sequencing biases higher GC content sequences. This GC content bias within Illumina sequencing data has been reported since 2008 [Dohm et al., 2008], with subsequent attempts to characterise and address this bias [Benjamini and Speed, 2012, Ross et al., 2013].

The results detailed in Figure 2.11 illustrate that enriching DNA using WGA prior to sequencing significantly alters the coverage over the genome. Rudimentary analysis suggests that it merely increases the mean depth of coverage over the genome, however, more thorough analysis indicates that it selectively amplifies certain genomic regions to a higher degree than others. Due to the existing GC bias which has been reported within Illumina sequencing data (illustrated by Type I and type IV distribution, as seen in Figures 2.10a and 2.10d), this has the effect of compensating for this bias, resulting in a much more radially dispersed distribution with a far less clear positive correlation between coverage and GC content. The amplification bias of WGA which was previously discussed and indicated in Figures 2.6, 2.7 and 2.8, may be responsible for compensating for the Illumina GC bias by preferentially amplifying large ($>$ 10Kb) regions of the genome [Alsmadi et al., 2009] (as seen in Figure 2.12) by some other set of criteria, leading to alteration in the type of distribution seen, and consequently the angular momentum of the

Isolate	Pre-PAGIT stats			Assembly size post-PAGIT (kb)	Gaps closed by IMAGE	Genes transferred: all (erroneously)
	No.	N50	Av. (kb)			
UKH3	168	149.9	54.0	9293	12	3792 (401)
UKH4	522	57.4	17.5	9594	95	3791 (467)
UKH5	463	54.6	19.6	9357	92	3787 (496)
UKP2	157	216.0	58.2	9254	23	3720 (356)
UKP3	270	109.8	33.7	9336	23	3688 (453)
UKP4	235	175.2	38.7	9226	22	3770 (349)
UKP5	447	70.7	20.3	9271	51	3800 (430)
UKP6	689	332.6	14.1	9826	13	3731 (96)
UKP7	521	62.6	17.3	9257	19	3797 (475)
UKP8	369	93.0	24.7	9473	26	3803 (518)

Table 2.2: The assembly statistics (SPAdes and post-PAGIT) include the number of scaffolds (No.), scaffold N50 metric, scaffold mean length (Av.), and the total size of the final assembly. Gene annotations were transferred by RATT out of a total of 3805 gene annotations in the reference assembly. Genes erroneously transferred refers to genes transferred to regions which have been identified as chimeric (and therefore misassemblies). Within *C. hominis*, the erroneous transfers are putative, due to differences between *C. parvum* and *C. hominis*.

isolates read distribution. Such bias has been documented previously, but not investigated thoroughly, and is often characterised as random [Alsmadi et al., 2009, Börgstrom et al., 2017]. Further investigation into this bias must be carried out to characterise this in more detail, along with the relationship between the Illumina %GC bias and WGA.

2.5 Results and Discussion for Assembly and Post-Assembly Processing

Table 2.2 shows the results of assembly using SPAdes. The results from assembly with Velvet were comparable to that of SPAdes, and therefore are not shown here. Table 2.3 shows the results of assembly using IDBA-UD. The results shown in these tables indicate that SPAdes produced assemblies with longer and fewer contigs than IDBA-UD, highlighting the differences between the assembly approaches adopted by the assemblers.

Both the assemblies were then run through the PAGIT pipeline to make the improvements described in the methods section, including gap closing and the transfer of gene annotations. The results can be found in Tables 2.2 and 2.3. The SPAdes assemblies required fewer gaps to be closed by IMAGE. The mean percentage of genes transferred by RATT to the improved SPAdes assemblies is >99%. The mean percentage of genes transferred to chimeric regions is 10.6%.

Isolate	IDBA-UD assembly stats			Assembly size post-PAGIT (kb)	Gaps closed by IMAGE	Genes transferred: all (erroneously)
	No.	N50	Av. (kb)			
UKH3	419	52.9	21.5	9102	104	3757 (0)
UKH4	627	39.7	14.3	9212	229	3688 (44)
UKH5	619	38.7	14.5	9197	247	3699 (32)
UKP2	360	63.9	25.2	9143	241	3776 (0)
UKP3	563	47.8	16.0	9168	312	3767 (1)
UKP4	509	53.7	17.7	9154	292	3772 (0)
UKP5	1830	11.2	4.8	9273	1791	3552 (1)
UKP6	768	51.4	12.1	9135	105	3702 (2)
UKP7	829	32.0	10.7	9184	288	3775 (6)
UKP8	614	40.7	14.7	9177	293	3756 (0)

Table 2.3: Statistics for draft genomes assembled using IDBA-UD as per Table 2.2. An extended table including assembly stats from the extended *C. parvum* and *C. hominis* dataset can be found in the Appendix (Table 1).

Table 2.3 shows the results of assembly using IDBA-UD, and subsequent improvement and annotation using PAGIT. These genomes benefited greatly from gap closure by IMAGE over those produced by SPAdes (see Tables 2.2 and 2.3), since gaps in intragenic repetitive regions were much more common, potentially confounding VNTR analysis. The mean percentage of genes transferred by RATT to the improved IDBA-UD assemblies is 98%. The mean percentage of genes transferred to chimeric regions is 0.2%. In the IDBA-UD assemblies, the *C. hominis* genomes performed slightly worse, with 0, 44, and 32 genes transferred to chimeric regions respectively across UKH3, UKH4, and UKH5.

The dramatic decrease in the number of genes transferred to chimeric regions indicates significantly fewer misassemblies in improved genomes generated by IDBA-UD than in those of SPAdes, marking a significant improvement. This indicates the effectiveness of using ABACAS to identify gaps within the IDBA-UD assemblies, and IMAGE to close them, which SPAdes would resolve during assembly.

NUCmer, from the MUMMER package was used to identify misassembly, as detailed in section 2.3.3. Figure 2.13 shows the extent of misassembly in the isolate genomes, denoted by coloured bars corresponding to which chromosomes regions belong to according to NUCmer. Extensive misassembly was identified in all of the genomes, to varying degrees. The most consistently misassembled chromosome is chromosome 7, with a consistent chromosome 8 misassembly. The most misassembled isolates where UKP3 and UKP8, with 8 misassemblies of larger than 10kb. These two isolates have very high Gini scores (see Table 2.1), of 0.5489 and 0.5570 respectively.

Figure 2.14 illustrates a moderate correlation ($R^2 = 0.41$) between the Gini coeffi-

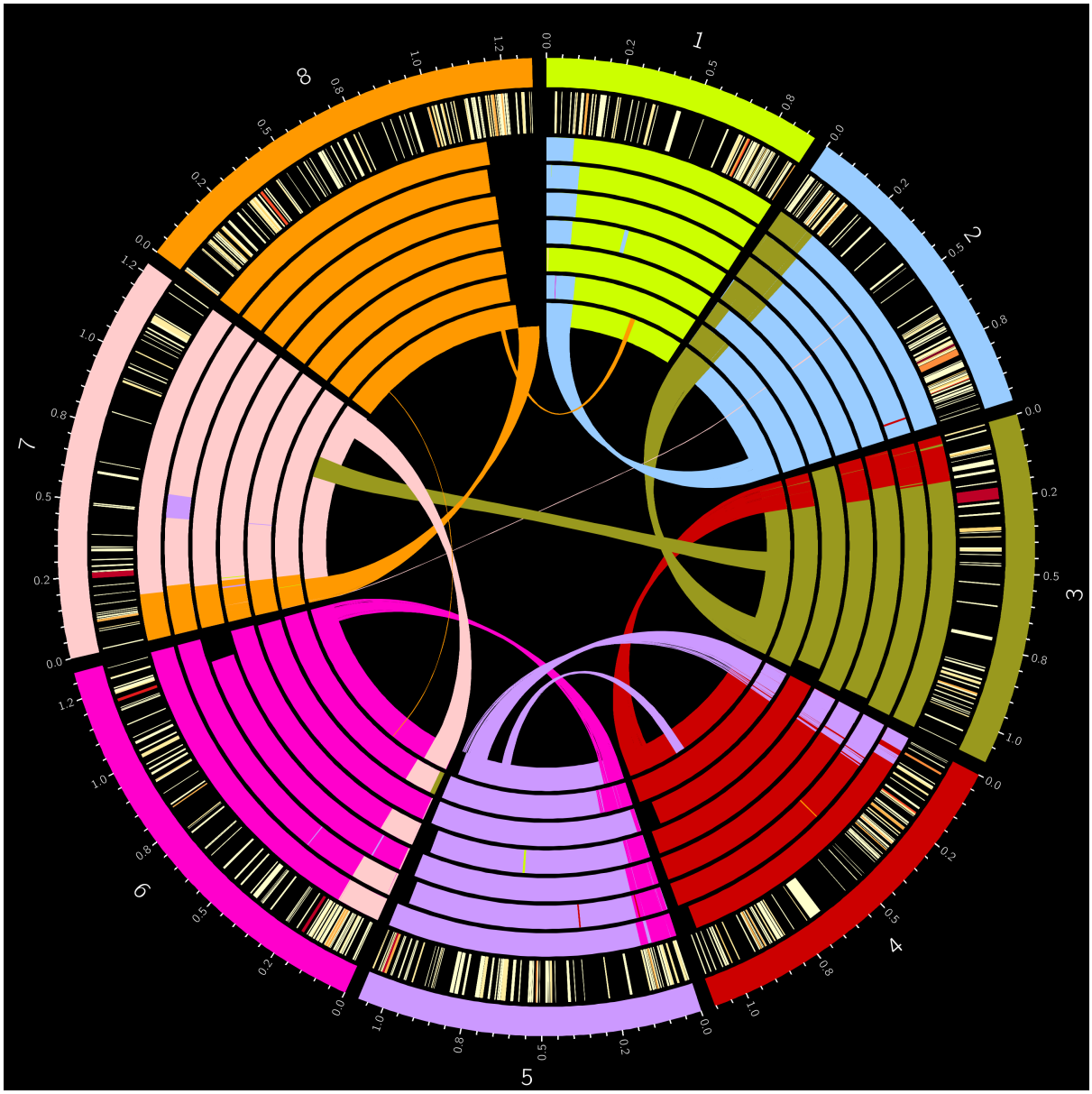


Figure 2.13: Misassembled regions on each SPAdes assembled Hadfield *et al. C. parvum* genome. Regions are colour coordinated by which chromosome of the *C. parvum* IowaII reference genome (represented by the outer track) they map to. From outermost to innermost, the inner tracks represent the genomes of each isolate from UKP2-8. The innermost track (UKP8) also includes a linkage map showing precisely where the regions map to in the IowaII reference genome. The second from outer track shows a heatmap of genes bearing Tandem Repeats (TR's), from light yellow denoting a single VNTR within the gene to dark red indicating many TR's within the gene. TR's were identified using Tandem Repeats Finder (see section 2.3.3).

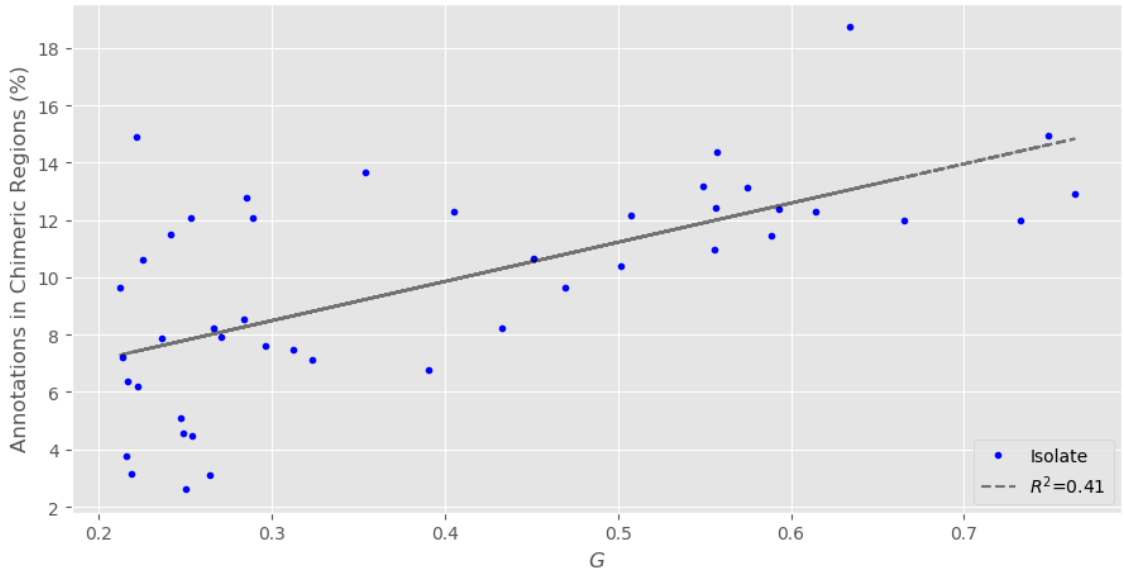


Figure 2.14: The percentage of genes transferred to chimeric (misassembled) regions against Gini coefficient of coverage for 45 isolates of *C. parvum* and *C. hominis*. R^2 is the coefficient of determination.

Isolate	VNTR regions missing before IMAGE	VNTR regions missing post-IMAGE
UKP2	48	7
UKP3	56	12
UKP4	63	10
UKP5	209	33
UKP6	62	13
UKP7	62	8
UKP8	67	13

Table 2.4: The number of VNTR regions missing within the IDBA-UD assemblies pre and post gap closing with IMAGE.

cient and number of misplaced genes within misassembled chromosomal regions across 45 isolates of *C. parvum* and *C. hominis*.

Table 2.4 shows the number of VNTR regions that were missing from the IBDA-UD assemblies before and after gap closure with IMAGE. These results show that a large amount of VNTR regions were resolved using IMAGE, indicating the importance of post-assembly genome improvement in the generation of accurate and reliable genome assemblies.

Figure 2.13 shows putatively misassembled regions (translocations) within the *C.parvum* UKP2-8 [Hadfield et al., 2015] PAGIT-improved SPAdes assemblies. A heatmap showing the number of VNTR's per coding sequence (CDS) is included. Every genome assembly

within the dataset exhibits significant misassembly across all chromosomes, particularly at the terminal end.

Whole genome alignments were used to identify *in silico* translocation events (considered putative misassemblies), as detailed in section 2.3.3. Figure 2.13 illustrates that translocation occurred in a similar fashion throughout each of the assemblies, with the same areas being merged into similar chimeric genomes, as can be seen in chromosome 7, where the initial 120kb region has merged into the end of chromosome 8 throughout all of the genomes. It is interesting to note that only on UKP3 was a 70kb area from chromosome 5 seen starting at 500kb on chromosome 7. Similarly only in UKP8 was a unique 70kb translocated region seen in chromosome 7 from chromosome 3. These two genomes bear high Gini coefficients, as detailed in Table 2.1, which may contribute to this. A peculiarity of these misassemblies is the observed trend of chimeric chromosomes being a result of the native chromosome being flanked upstream by 80kb of the downstream extreme portion of the subsequent chromosome. This is illustrated very clearly in Figure 2.13.

Taxonomic evaluation carried out by Hadfield *et al.* utilising the gp60 marker show that there are five gp60 subtypes within the *C. parvum* dataset. This variation within the Hadfield *C. parvum* isolates is supported by Perez-Cordon *et al.* [Perez-Cordon et al., 2016] which shows clear variation across 28 VNTR loci, suggesting a number of genetic lineages. The very low likelihood of similar translocation occurring across different populations of *C. parvum* indicates that these events are as a result of misassembly by SPAdes, rather than a biological observations.

Examination of one such chimeric contig (the chr8-chr7 chimeric region at 0-0.14Mb of UKP3 on Figure 2.13) revealed that the region has very low depth of coverage, with no single read spanning the chromosomal fragments. Moreover, the sequences from different chromosomes are joined using a simple "AT" repetitive region with only three reads spanning the repeat region and no reads pairing across it (see Figure 2.15). This was observed in a number of other chimeric interface regions. Due to the low complexity, high repeat rich nature of the *Cryptosporidium* genome, coupled with the difficulties associated with DNA extraction and sequencing of this parasite, there is insufficient evidence to suggest that this represents true biological variation. Instead, it may be attributed to a misassembly by the SPAdes software. This kind of assembly error was also typical of the assemblies produced by using Velvet *de novo* assembler.

Unlike SPAdes, the IDBA assembler leaves these sequence fragments unjoined, with the result that significantly less chimeric regions are seen in the IDBA assemblies. This

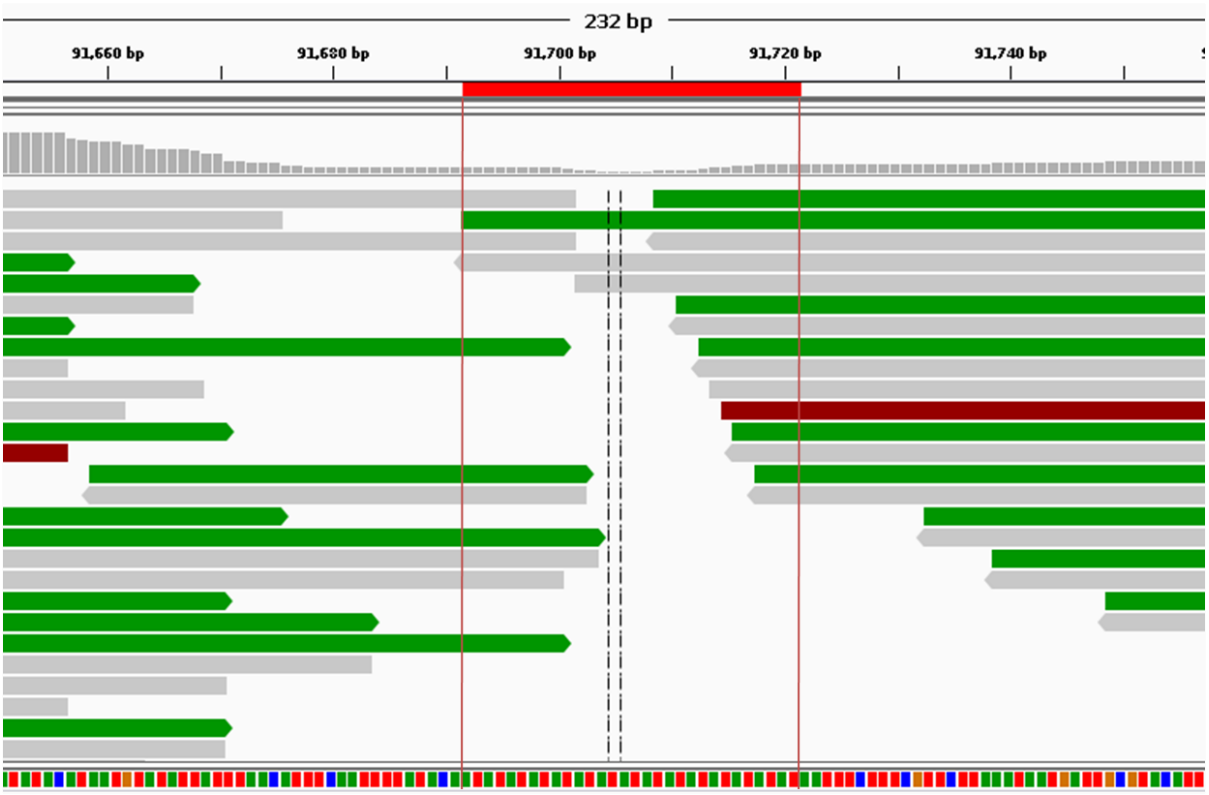


Figure 2.15: The misassembly interface between fragments from chromosomes 8 and 7 on the chimeric chromosome 7 of UKP3. Single reads are shown, as is a colourised sequence track (A = Green, T = Red, C = Blue, G = Orange) at the bottom where the repeat region implicated in the formation of this chimeric contig can be seen. Image produced using IGV.

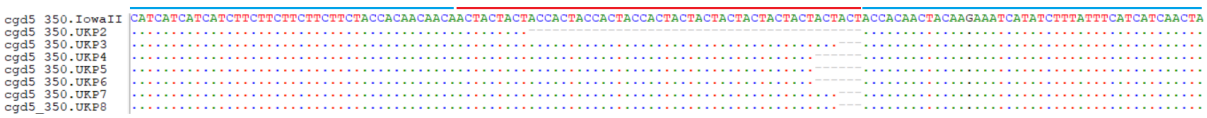


Figure 2.16: A multiple alignment of a VNTR region within cgd5_350 in the *C. parvum* dataset utilised in this chapter that was resolved within IDBA-UD assemblies using IMAGE. Four alleles are seen within this alignment, defined by variation in the number of 'ACC' and 'ACT' codons present within the region. Differences such as this within a VNTR region can be used to define distinct genotypes, used for diagnostic evaluation. This VNTR was identified by VaNTA, as discussed in Chapter 3.

is because IDBA is designed for the task of assembling genomes of highly uneven depth of coverage. Although IDBA-UD did not create so many chimeric contigs, the low complexity regions were often left unassembled, with the result that CDS regions contained gaps. Unfortunately, these gaps often included the VNTR's that require for a multi-locus subtyping scheme.

Both SPAdes and Velvet (data from Velvet not shown) produced full, ungapped CDS regions (see Table 2.2). Thus the IDBA assemblies were not suitable for VNTR analysis and further biomarker identification without significant improvement. PAGIT was used to improve the genomes from all assemblers (see section 2.3.2), and this improved the resolution of low complexity regions within the IDBA-UD assemblies. Within PAGIT, ABACAS performs scaffolding on the genome assemblies and introduces gaps across the unassembled regions, the IMAGE tool then performs gap closure on these regions, resulting in high quality intragenic VNTR's for biomarker analysis.

An example of a region resolved by IMAGE can be seen in Figure 2.16, which shows a multiple alignment of the cgd5_350 gene from each of the Hadfield *C. parvum* assemblies. This region exhibits 4 distinct alleles, and can therefore be used to define specific genotypes: an essential tool of clinical diagnostics. The number of gaps closed within the IDBA-UD assemblies was significantly higher than within the SPAdes assemblies. This difference in gaps closed was expected, as IDBA-UD was designed for the purpose of assembling genomes which suffer from poor depth of coverage equality, and is therefore more conservative in extending reads across regions with shallow coverage.

The *C.parvum* assemblies produced by IDBA-UD and PAGIT exhibited very few misassemblies compared to the SPAdes assemblies. However, the *C. hominis* genomes suffered from a greater amount of putative misassemblies within the IDBA-UD genomes, as measured by the number of genes being transferred between chromosomes. Note that, genes are transferred from the *C. parvum* IowaII reference genome, which is as different, albeit similar species, and so some biological changes may be expected. Further

analysis is required to fully eliminate assembly error as a cause of these chromosomal translocations. Table 2.4 shows that IMAGE is essential within this workflow for the resolution of repetitive regions which are not resolved during assembly with IDBA-UD. The results show a five to six-fold decrease in the number of VNTR regions missing within the assemblies.

2.6 Conclusion

In this chapter I have performed a detailed analysis of 10 *Cryptosporidium* genomes (Dataset 1) assembled with 3 popular assemblers. The results of this analysis was then used to direct the re-assembly of Dataset 1.1 to generate Dataset 1.2, and to assemble a further 48 *Cryptosporidium* genomes (Dataset 2: 29 *C. parvum* and 19 *C. hominis*) isolated from UK clinical samples.

To investigate sequencing depth and breadth of coverage, I have developed a novel approach that uses the Gini coefficient to determine coverage inequality. I also present a novel technique which allows for further investigation of depth of coverage inequality by generating Gini-granularity curves. I demonstrate how these curves characterise the distribution of reads across a genome and relate this to the quality of subsequent genome assemblies. Distribution of reads across a genome are characterised by the aggregation of reads, leading to coverage 'blocking' and 'spiking', describing the appearance of read coverage. This aggregation of reads is quantified by the use of Gini-granularity curves.

I have demonstrated that the use of WGA to enrich DNA within clinical samples is a viable way of increasing read coverage. However, these results also demonstrate that there is a complex relationship between the selectivity of amplified DNA during WGA, and its sequence content, which is characterised by non-uniform amplification of large (tens of kilobases) regions of the genome, resulting in a read coverage bias (see Figure 2.12 for an example of this preferential amplification of large regions, leading to read aggregation). The results suggest that this bias introduced by WGA compensates for the Illumina GC content bias.

I found the SPAdes and Velvet assemblies to be problematic on our datasets. This led to misassemblies across low coverage, low complexity regions, resulting in the creation of chimeric chromosomes: up to 15% of all genes were being placed within these chimeric chromosomes. Although the assemblies generated by IDBA-UD did not suffer from the problem of chimeric sequences, they were problematic due to a different assembly approach, leading to a large number of gaps, particularly in repetitive regions. This is a significant issue because these gaps often contained the VNTR sequences that are impor-

tant to us for developing new clinical genotyping strategies. However, the IMAGE gap closing tool from the genome improvement pipeline, PAGIT, was able to resolve these missing low complexity regions. Using this strategy, of assembly with IDBA followed by gap closing with IMAGE, it will be possible to perform more in depth VNTR analysis with the intention of identifying biomarkers that will facilitate the development of novel prevention strategies in the fight against Cryptosporidiosis.

In the following Chapter, I will discuss the use of these high-quality, reliable genomes in identifying novel biomarkers. I will introduce a novel bioinformatic tool which automates the process of identifying Variable Number Tandem Repeat (VNTR) regions around a CDS dataset generated using these genomes.

Chapter 3

VaNTA: Automated Discovery and Variant Analysis of VNTR's Within Coding Regions.

3.1 Introduction

Due to the advent of high throughput next generation sequencing (NGS) approaches, the amount of genomic data able to be produced has significantly increased. However, a bottle neck exists in the analysis of these genomic data. The fast and cheap nature of NGS in modern biology has allowed for sequencing of multiple populations of the same species, which opens the door for comparative genomic analysis. The current state of sequencing in biology necessitates the development of high throughput, lightweight bioinformatic tools and pipelines, which automate the process of genome analysis in a cost, and time effective manner.

Tandem Repeats (TR's) are low complexity regions within DNA sequence consisting of a single motif or subsequence repeated a number of times. The presence of these tandem repeats has been implicated in disease, and may act as regulatory and epigenetic elements within a genome. They have been interrogated and their variation exploited for the purpose of laboratory diagnostics, whereby their presence or absence, and/or their variation has been used to define groups of organisms or infer evolutionary lineage.

Tandem repeats are a common feature of many genomes, with 10% of the human genome comprising of this type of sequence. In particular, telomeric regions are highly repetitive in nature, which serves a number of biological and evolutionary purpose, such as acting as expendable sequence during DNA replication.

TR regions are subject to high rates of mutation. This is thought to be largely due

to a process called slipped strand mispairing, whereby DNA strands are displaced during DNA replication, resulting in mispairing of complementary bases. Depending on the nature of the slippage, this can result in the expansion or retraction of a repetitive region. Further point mutations, insertions or deletions of this repetitive sequence may result in further sequence variation, resulting in approximate copies of a repeat subunit being exhibited, which are then subject to the same replication errors. These processes can produce highly polymorphic sequences within a population over relatively few generations. Due to their high mutagenicity, particularly in repeat copy number, they are ideal candidates for defining genotypes. When repeat copy number within a TR is identified as polymorphic within a population, it is considered a Variable Number Tandem Repeat (VNTR).

Within the context of this paper, we shall consider a VNTR to be a TR region that exhibits > 1 allele within a population, and can therefore be used to define subpopulations.

It is common for research groups to develop and use their own 'in house' pipelines when carrying out TR analysis and detection. A typical example of this is reported by Perez-Cordon *et al.*, who used Tandem Repeats Finder (TRF) to detect VNTR regions within the genome of Cryptosporidium parvum Iowa II isolate and aligned them to homologues within a dataset of genomes generated by Hadfield *et al.* [Perez-Cordon et al., 2016, Benson, 1999, Hadfield et al., 2015].

The Garner lab uses a method which also utilised TRF to identify VNTR regions around the human genome, followed by alignment of reads against the VNTR reference set using a bespoke set of scripts driven by BLAST and BWA [McIver et al., 2013, McIver et al., 2011, Altschul et al., 1990, Li and Durbin, 2009].

There are a limited number of tools available for VNTR analysis. The tools which are available come under two general classes: TR discovery, and variant detection. lobSTR is a Short Tandem Repeat (repetitive regions of 2-6 nucleotide repeat subunits) profiler, which uses known STR's within a genome detected by TRF and aligns sequencing reads to them. However, it is known to be limited to the number of microsatellite loci it can call, and is not able to call monomers [McIver et al., 2013, Gymrek et al., 2012]. RepeatSeq is a tool for genotyping microsatellite repeats, using a Bayesian model selection guided by an error model incorporating sequence and read properties. Reads are mapped to a reference region and the mapped reads realigned using the GATK IndelRealigner tool, with reads not fully spanning the region being discarded. Genotypes are assigned using a fully Bayesian approach, considering the reference length of the repeat, the repeat

subsequence size, and the average base quality [Highnam et al., 2013, Alkan et al., 2011].

As should now be clear, many of these tools use very similar approaches, being repeat discovery using TRF, followed by variant detection by read mapping onto these repeat reference loci. Because of this, they are subject to similar weaknesses, specifically the usage of reads to detect variation, which may be confounded by reads generated by unsupported sequencing technologies, poor read coverage or quality, or contaminant sequence, and scales poorly with the size of the .fastq file and query dataset. They also do not take advantage of the wealth of assembled DNA sequence which is far more readily available and easily accessible over the internet, as well as annotation data.

Here we present the Variable Number Tandem Repeat Analysis Pipeline (VaNTA), a bioinformatic pipeline designed to automate in silico discovery and variability analysis of VNTR's within a query data set of coding regions (CDS'), using a TR reference library generated using a suitable reference genome. CDS regions are interrogated for VNTR's due to the higher selective pressure being exerted on these regions, resulting in a higher probability of variation at these loci. The approach utilised in the VaNTA pipeline avoids the usage of raw reads in favour of fully assembled and annotated genomes due to the intention of identifying VNTR's in large datasets of multiple genomes, which would necessitate significant time and computational resources to achieve by processing raw reads. Furthermore, it was designed with the intention of utilising the enormous amount of assembled genomic sequence and annotation data available online to streamline and automate a typical pathogen biomarker identification workflow, which provide essential theoretical research backing to guide clinical molecular diagnostics and genotyping schemes.

3.2 Method

3.2.1 Data Sets

Two data sets (assemblages) consisting of protozoans belonging to the apicomplexa were assembled (Table 3.1). The data sets were constructed with the objective of identifying VNTR biomarkers that can be used to define genotypes. VaNTA was run using each assemblage as a dataset, totalling two separate runs.

3.2.2 .vff Format and Construction

Annotation files in .embl or .gff format are reformatted into .vff format; a bespoke annotation format developed for the purpose of storing all information required by VaNTA.

Assemblage	Reference	Query data set
A1	*C. parvum* Iowa II [Abrahamsen et al., 2004]	*C. parvum* (n=7): UKP2-8 [Hadfield et al., 2015]
A2	*P. falciparum* 3D7 [Gardner et al., 2002]	*P. falciparum* (n=7): Pf2004, Pf7G8, PfDD2, PfE5, PfHB3, PF-IT, PfNF54

Table 3.1: The contents of assemblages 1 and 2.

This format has both annotation and concatenated sequence data to streamline file parsing. The format is structured in a linear fashion similar to .gff format. A single line is structured as follows:

`feature_ID feature_class start_cat end_cat strand chromosome_ID start_chr end_chr`

Where:
`feature_ID`: The designated name of the annotation feature (e.g. the name of the a gene).
`feature_class`: The class of the annotated feature (e.g. CDS, gene, mRNA, exon etc.).
`start_cat`: The first base of the feature within the concatenated sequence at the end of the .vff file.
`end_cat`: The last base of the feature within the concatenated sequence at the end of the .vff file.
`strand`: The strand the feature can be found on ('+' or '-').
`chromosome_ID`: The designated name of the chromosome the feature can be found on (e.g. chr_1).
`start_chr`: The first base of the feature within the chromosomal sequence alone.
`end_chr`: The last base of the feature within the chromosomal sequence alone.

The annotation lines are followed by a single line, reading '##FASTA##', denoting the start of the full concatenated genome sequence. This comprises all chromosomal sequences merged together into a single unbroken whole genome sequence.

These .vff files were generated using .gff annotation files of the assemblages (TTable 3.1) downloaded from CryptoDB (A1), and the Welcome Trust: Sanger FTP server (A2 & A3). A tool for reformatting .embl or .gff format files into .vff can be found in the VaNTA github repository or on the Galaxy platform.

3.2.3 VaNTA Input and Arguments

The input for VaNTA necessitates the generation of an arguments file for command line usage. An example of this argument file can be found in figure 3.1. This file can be autogenerated using a VaNTA wrapper script, whereby only the essential arguments are

```
##VaNTA_config_file##

~#script_dir:    /home/user/VaNTA/script/
~#ref_vff: /home/user/VaNTA/ref/ref.vff
~#query_vff_array_dir:      /home/user/VaNTA/vff_array/
~#ref_outdir:    /home/user/VaNTA/ref
~#CDS_library_array_dir:    /home/user/VaNTA/CDS_library/
~#feature_type: CDS
~#TRF_path:      /home/user/VaNTA/script/trf409.linux64
~#VNTR_lib_gen_outdir:      /home/user/VaNTA/trf_out/
~#flank_size:    41
~#VNTR_multifasta_out_prefix:    ref.VNTR
~#flank_outdir: /home/user/VaNTA/flank_seq/
~#EMBOSS_water_path:  /home/user/VaNTA/script/EMBOSS-
6.6.0/emboss/water
~#single_locus_multifasta_outdir:
     /home/user/VaNTA/single_locus_MF/
~#alignment_outdir:   /home/user/VaNTA/trf_out
~#csv_outdir:    /home/user/VaNTA

##ADVANCED OPTIONS##
## IMPORTANT: indicate FALSE if you do not wish to use these options

~#feature_tag:   FALSE
~#assembly_dir:  FALSE

##ENDFILE##
```

Figure 3.1: The layout of the VaNTA argument file.

provided, and a unique temporary directory structure is created in the local `/tmp/` directory and used as a working directory for that run only. The directory structure is as follows:

```
VaNTA.XXXXXX
├── ref
├── vff_array
├── single_locus_multifasta
├── CDS_array
└── trf_out
```

Where the root `/tmp/VaNTA.XXXXXX` is generated using the `mktmp` command, and `.XXXXXX` refers to a unique string to identify this run. All output will be deposited into this temporary working directory structure.

Table 3.2 details the parameters used to run Tandem Repeats Finder and EMBOSS-water local aligner. Parameters used for TRF were selected to favour minisatellites and smaller microsatellite repeats which had little variation in period (repeat subunit) sequence. Parameters used for EMBOSS-water local aligner were generic, which was suitable for flank conservation and TR variation analysis.

TRF Argument	Value	EMBOSS-Water Argument	Value
Matching Weight	2	Gap Penalty	10
Mismatch Penalty	5	Gap Extension Penalty	0.5
Indel Penalty	5		
Match Probability	80		
Indel Probability	10		
Min Score	50		
Max Period Size	15		
Options	-h		

Table 3.2: Arguments and options used for Tandem Repeats Finder [Benson, 1999] and EMBOSS-water [Rice et al., 2000].

3.2.4 Step 1: CDS and TR Reference Library Construction

Feature libraries may be constructed from any feature present in the .vff annotation files (e.g. exons, intron, mRNA etc.), however, this tool was designed to identify VNTR's within CDS', due to the higher selective pressure exerted upon them in comparison to non-coding sequence.

CDS libraries were constructed using the .vff format annotation files. A TR reference library generator script driven by Tandem Repeats Finder (TRF) [Benson, 1999] is then used to identify tandem repeats within the reference CDS library, which are then isolated, along with flanking regions, to form a library of reference TR's (Figure 3.2A). The flanking region is user defined, and accomplished two tasks, the first being to guide pairwise alignment, since VNTR regions are, by their nature, highly polymorphic and therefore subject to misalignment. The second being to allow regions flanking the TR to be checked for conservation throughout the isolates for the purpose of primer design.

3.2.5 Step 2: TR and Flank Alignment

Using this TR reference library, pairwise alignments are carried out between the reference TR's and query homologs using a script driven by EMBOSS-Water local aligner [Rice et al., 2000]. EMBOSS-Water local aligner was used due to its efficacy in aligning regions with high sequence similarity.

Flanking regions are then isolated from the TR loci and individually subjected to a similar pairwise alignment to establish upstream and downstream flank conservation.

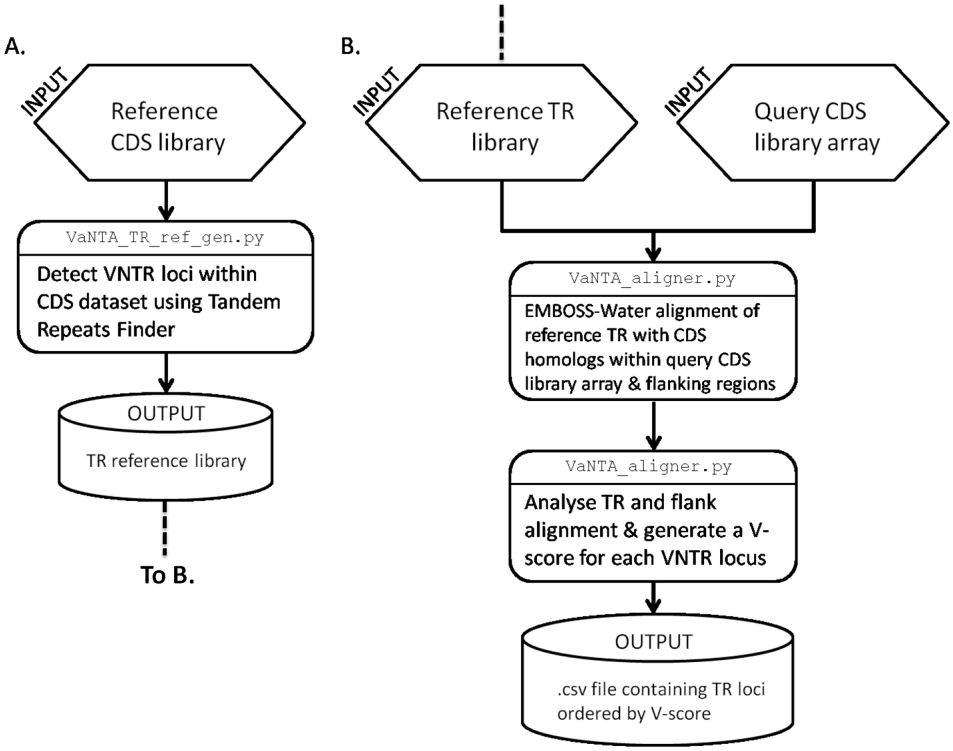


Figure 3.2: Workflow of VaNTA VNTR discovery pipeline. A. TR reference library construction. B. TR and flank alignment and analysis.

Pairwise TR alignment occurs on a one against all basis, in which all query homologs are aligned to a single reference TR and subjected to indirect evaluation by score comparison, which results in an analysis time of $O(N)$ where N is the number of query CDS libraries (Figure 3.2B).

3.2.6 TR Evaluation

Each TR locus is then evaluated by calculating a score based on how polymorphic the locus is, the level of flank conservation and the length of the repeat region (Figure 1B). The V-score is calculated as follows.

Take f^1 and f^2 to be the sets of alignment similarity percentages of flanks 1 and 2 respectively. These were generated by aligning TR locus flanks from each of the isolates against the reference TR locus. The arithmetic mean of these sets is calculated as

$$F^j = \frac{1}{n} \sum_{n}^{i=0} f_i^j$$

where $j \in \{1, 2\}$. Also take p to be the length of the TR period (e.g. $p = 3$ in a TR consisting of TCA repeats), and l to be the length of the TR locus in the reference genome. The V-score is calculated as

$$V\_score = 100v + F^1 + F^2 - 10p - l$$

This score is weighted to bias smaller repeat regions with small repeat unit size, high flank conservation and high repeat unit number polymorphism. The premise behind selecting for smaller repeat regions is to select for copy number variation, rather than sequence variation. In smaller regions, the impact that a single repeat unit variation has on the relative sequence difference of the region is larger. Likewise, smaller repeat units bear a lower probability of non-repeat insertion or SNP events, which may confound analysis by undesirably exaggerating the variability of the TR locus.

3.3 Results

Running VaNTA on assemblage 1 (A1) using TRF and EMBOSS-Water options detailed in figure 2, 889 TR's were identified and aligned in at least one of the query CDS libraries across 532 genes. Of these, 283 were identified as VNTR's (2 or more alleles identified) across 192 genes.

Table 3.4 (see next page) shows the comparison between results obtained by Perez-Cordon *et al.* and VaNTA. Of the 25 coding regions identified by Perez-Cordon *et al.*, 19 were identified and aligned by VaNTA, with similar results in number of alleles identified at each locus within the dataset. Non-coding regions were not analysed during this run of VaNTA. Of the 6 which were not identified, 4 were as a result of the period size exceeding our defined max period size detailed in figure 2 (cgd2_3300_1504, cgd4_1340_1688, cgd4_3940_298, & cgd7_1010_9527) with no potential to reduce the period into a smaller repeat period as seen in cgd2_3690_5176 (cgd2_3690.P.5111-5261) and cgd6_4290_9811 (cgd6_4290.P.9810-9902). However, increasing the max period size to 18 resolved these loci. The two which were not identified by VaNTA (cgd1_3670_5956 & cgd2_3490_2029) were investigated. cgd1_3670_5956 was resolved by setting the mismatch and indel penalties to 7, to prevent it being reported as an 18 bp repeat and subsequently rejected for being >15bp, and cgd2_3490_2029 was resolved by reducing the minimum score to 30, to prevent it from being screened from the output (see figure 2).

The VNTR present at the gp60 locus (cgd6_1080.P.106-163) was identified within all 7 query CDS libraries and places at rank 3 with a V-score of 1943, exhibiting 5 alleles. The results from VaNTA are in accordance with the known gp60 subtypes of this dataset, seen within table 3.3.

Figure 3.3 shows a number of VNTR loci which were identified by running VaNTA on

C. parvum Isolate	Gp60 allele	BioProject Number	Number of features in annotation	CDS lib size (exons split)	VNTR's identified /missing
Iowa II	IIaA15G2R1	PRJNA15586	3805	4042	1154/NA
UKP2	IIaA19G1R2	PRJNA253836	3694	3963	1089/65
UKP3	IIaA18G2R1	PRJNA253840	3642	3931	1091/63
UKP4	IIaA15G2R1	PRJNA253843	3613	4333	1077/77
UKP5	IIaA15G2R1	PRJNA253845	3559	4733	1014/140
UKP6	IIaA15G2R1	PRJNA253846	3684	3939	1118/36
UKP7	IIaA17G1R1	PRJNA253847	3604	3951	1061/93
UKP8	IIdA22G1	PRJNA253848	3602	3911	1064/90

Table 3.3: Summary of the C.parvum isolates used in assemblage 1. N.B. Annotation features collapse intronic genes into a single feature, which are split into separate exons during CDS library generation.

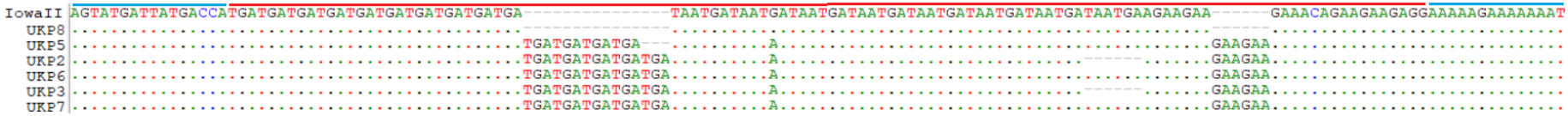
assemblage 1 with flanking and repetitive regions highlighted. Figure 3.3c illustrates a region (cgd1_3170.P.4280-4371) which exhibits a number of adjacent repeat subsequences (periods), resulting in 4 identifiable alleles. Figure 3.3d shows the gp60 VNTR locus (cgd6_1080.P.106-163) isolated by VaNTA, illustrating that all 5 expected alleles (as detailed in table 3.3) were resolved and identified.

Locus*	Locus VaNTA	PS*	CNS*	NS VaNTA	# alleles*	# alleles VaNTA	V-score	Rank
cgd1_470_1429	cgd1_470.P.1415-1488	6	TC(T/G)GAT	TTCTGA	3	2	1727	137
cgd1_3060_604	cgd1_3060.P.587-636	6	TCCTCA	CATCCT	2	2	1751	84
cgd1_3170_4182	cgd1_3170.P.4175-4215	12	TGATTCCAATTC	ATTCTGATTCCA	2	3	1760	70
cgd1_3670_5956	NF	6	GAGCCT	NF	2	NF	NF	NF
cgd2_430_451	cgd2_430.P.433-494	6	TCAAGT	CAAGTT	2	2	1739	117
cgd2_3300_1504	NF	18	CATTCTGGTAGGGGAGGA		2	NF	NF	NF
cgd2_3320_1621	cgd2_3320.P.1564-1680	12	GAACAGGAGCAT	CAAGAGCATGAA	2	2	1684	186
cgd2_3490_2029	NF	6	TCATCT	NF	2	NF	NF	NF
cgd2_3550_1474	cgd2_3550.P.1470-1503	12	TCCACTTCTGCT	TTCCACTTCTGC	2	2	1767	53
cgd2_3690_5176	cgd2_3690.P.5111-5261	18	GAAAAGGAGGAGAAAGAG	GAGAAAGAG	2	2	1650	523
cgd3_3620_1036	cgd3_3620.P.933-980	6	AAAGA(C/T)	AAAGAC	3	3	1853	12
cgd4_1340_1688	NF	21	GGTACTAAAATTAC(C/T)AATACC		2	NF	NF	NF
cgd4_2350_796	cgd4_2350.P.795-1001	15	CC(T/C)GGTATGGG(T/C)CC(A/G)	CCTGGTATGGGTCCA	5	5	1494	906
cgd4_3450_4336	cgd4_3450.P.4333-4357	6	TCTGAA	AATCTG	2	2	1775	32
cgd4_3630_880	cgd4_3630.P.847-959	12	CCAAGTAG(C/G)(A/G)CT	CAAGTAGCACTC	2	2	1588	805
cgd4_3940_298	NF	18	GAAAGCGATTCTGATAGT		2	NF	NF	NF
cgd4_3970_1525	cgd4_3970.P.1519-1554	6	ATGCCT	CCTATG	2	2	1765	64
cgd5_10_310	cgd5_10.P.275-367	12	GCTCAGGAAGGA	AGGAAGGAGCTC	2	2	1708	166
cgd5_NC_3600_36	NF	18	CATCATCACCA(A/T)CATCAC		2	NF	NF	NF
cgd5_4490_2941	cgd5_4490.P.2931-2986	6	CAGAGC	CCAGAG	4	4	1745	101
cgd6_530_1561	cgd6_530.P.1518-1591	9	ACAGGAACA	CAGGAACAA	2	2	1727	138
cgd6_3930_1823	cgd6_3930.P.1746-1772	9	CAGCTCCTC	AGCTCCTCC	2	2	1774	36
cgd6_3940_688	cgd6_3940.P.647-716	6	ATGCCA	CAATGC	3	4	1331	1018
cgd6_4290_9811	cgd6_4290.P.9810-9902	27	(TCT*/TCC)TCTTCTTCCTCCTCT (TCTTCTTCC/TCCTCCTCT**)	TCT	3	2	1708	165
cgd7_420_4750	cgd7_420.P.4747-4786	6	(G/A/C)AA(C/G)AA	GAACAA	2	3	1861	9
cgd7_1010_9527	NF	18	TTGGACAGGGGTGTGGAG		2	NF	NF	NF
cgd8_NC_4440_50	NF	6	TGAGC(C/T)		7	NF	NF	NF
cgd8_NC_4990_36	NF	13	GGCGG(G/T)CAATTTT		2	NF	NF	NF

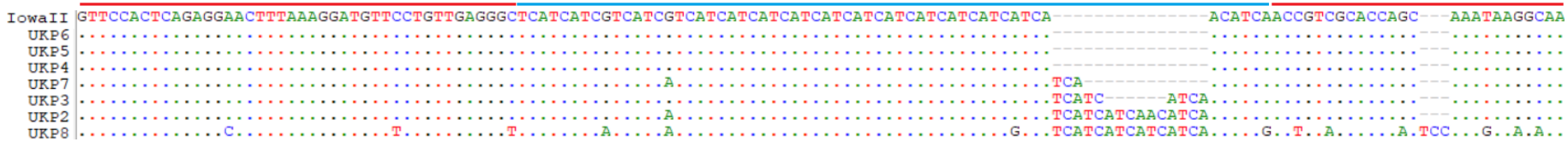
Table 3.4: Comparison of the results obtained during manual analysis by Perez-Cordon *et al.* and by VaNTA. NF - Not Found. Highlighted in red are loci which are not covered by the parameters used to run VaNTA (Non-Coding regions or a period size that exceed the max period size detailed in Figure 3.1). *Perez-Cordon *et al.*, PS = Period Size, CDS = Corrected Nucleotide Sequence, NS = Nucleotide Sequence.

(a) locus: cgd5_350.P.54-116 subseq: ACT alleles: 4

(b) locus: cgd6_4290.P.9810-9902 subseq: TCT alleles: 4



(c) locus: cgd1_3170.P.4280-4371 subseq: TGA/TAATGA/AGA alleles: 4



(d) locus: cgd6_1080.P.106-163 subseq: TCA alleles: 5

Figure 3.3: Example VNTR's identified by running VaNTA on dataset A1. Flanking and VNTR regions are denoted by blue and red bars respectively above the alignment.

Figure 3.4 shows the consistency in conservation of TR regions across the isolates included in assemblage 1 (A1). The major high density TR regions are exhibited across all isolates, such as those seen in chromosomes 3, 6 and 7. There are few TR regions which are not exhibited by all isolates, but instances of such can be seen in chromosome 1 at approximately 0.5mb, where UKP5 lacks a TR, and chromosome 3 at approximately 0.4mb where UKP5 lacks a TR, and again in chromosome 3 at approximately 1.0mb where UKP4 and UKP6 lack TR regions.

Running VaNTA on assemblage 2 (A2) using TRF and EMBOSS-water options detailed in figure 2, 11,151 TR's were identified and aligned in at least one of the query CDS libraries across 2597 genes. Of these 2160 were identified as VNTR's (2 or more alleles identified) across 959 genes.

3.4 Discussion

The results presented in table 3.4 illustrate the value of automated analysis of polymorphic VNTR identification. It is apparent that the reported repeat subunits (periods) differ between those reported by VaNTA and those reported by Perez-Cordon *et al.*, most commonly only by a single reading frame shift, such as that seen in cgd2_430.P.433-494 (cgd2_430_451) where the subsequence is reported by Perez-Cordon *et al.* to be TCAAGT, and by VaNTA to be CAAGTT. This is likely due to an artefact of tandem repeats finder reporting slightly differing scores due to different option parameters.

Missing data points such as those present in figure 3.3 were also observed in the annotation files for these isolates, which suggests lack of assembly of these regions, possibly due to poor local coverage, resulting in incomplete annotation. Table 3.3 demonstrates that the gp60 VNTR locus exhibits 5 alleles within assemblage 1. However, as seen in figure 3.4, the downstream flanking region is poorly conserved, and there are a number of point mutations which can be observed, particularly within the UKP8 isolate. In comparison, the loci illustrated in figures 3.1-3.3 each exhibit 4 alleles within assemblage 1, but with much better flank conservation, and (with exception of an A/G point mutation within cgd1_3170.P.4280-4371, exhibited in figure 3.3) polymorphism constrained to repeat subsequence copy number. If only repeat copy number is taken into consideration, gp60 exhibits 4 alleles within assemblage 1, since UKP2 and UKP8 vary only in A/G point mutations in two positions. This sequence polymorphism necessitates sequencing of the locus to identify subtypes. This difference in allelic designation is a non-trivial issue in a clinical context. Diagnostic laboratories use either sequence polymorphism or fragment length polymorphism to distinguish subtypes. The former necessitates sanger sequencing of the interrogated region, which is more expensive and time consuming than fragment

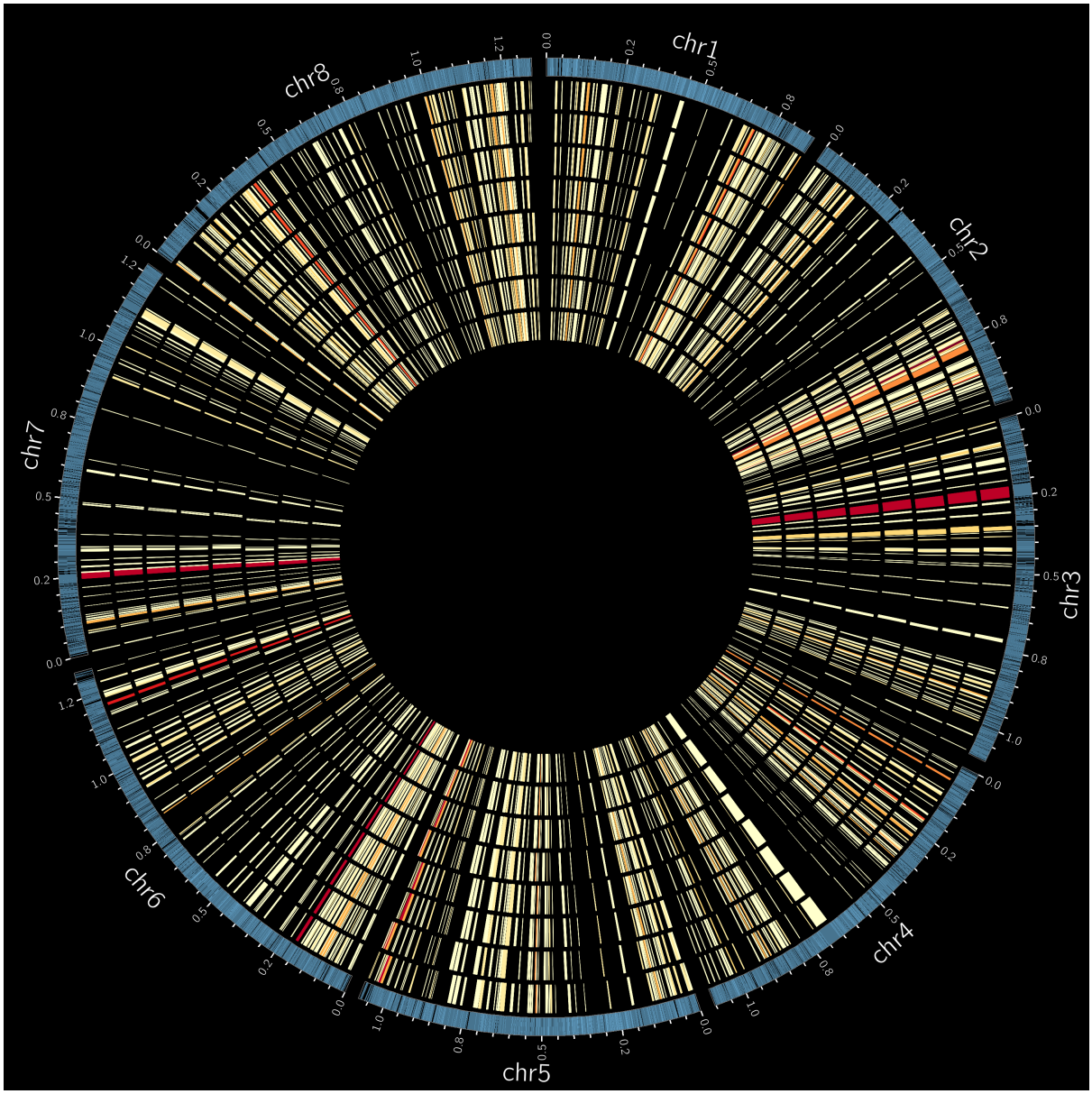


Figure 3.4: A circularised representation of VNTR density (number of VNTR's within a CDS) across the genomes of *C. parvum* isolates, from outermost to innermost: IowaII, UKP2-8. The outermost ring, representing the karyotypic ideogram of the *C. parvum* IowaII reference genome also includes gene positions represented as blue bars. The heatmap spectrum goes from light yellow (lowest density) to crimson (highest density). This figure was generated using Circos v0.69-6.

length polymorphism, which looks for differences in the length of the region and infers repeat copy number from this, but is more accurate. Therefore different results may be obtained depending on which convention is used, which may confound investigation.

The currently available tools use similar approaches of generating a set of reference tandem repeats (TR's) and map reads to them all suffer from similar weaknesses, primarily that they require a great deal of computational power to run. VNTRseek [Gelfand et al., 2014], which attempts to reduce the number of reads mapped against the reference TR dataset by running TRF on every read within a short read archive, generating a set of reads which contain repetitive elements. These low complexity reads are then mapped to the TR reference dataset and VNTR's identified. This approach is highly RAM, disk space, CPU and IO intensive. Approaches such as that employed by lobSTR, which maps reads to flanking regions for the purpose of short tandem repeat profiling are efficient and effective for STR profiling [Gymrek et al., 2012]. STRScan performs a similar task as lobSTR, but uses a more targeted approach, only profiling a set of user defined STR's, reducing redundancy of analysis [Tang and Nzabarushimana, 2017]. Such STR profiling approaches are effective for the analysis of single datasets but unsuitable for the discovery and analysis of repeats across multiple genomes. More relevant tools such as that developed by Denoeud and Vergnaud (2004) [Denoeud and Vergnaud, 2004] and implemented in an online platform, achieve the task of TR comparison within strains. However, this tool is limited to 16 bacterial species, and the orthopox viruses, and therefore lacks the scope and versatility of VaNTA. popSTR attempts to utilise the mapping approach seen in lobSTR and STRscan, but applies it to population scale datasets in a time and memory efficient manner. This is achieved by generating a reference micro-satellite profile and using mapped paired end illumina reads stored in a BAM file to identify VNTR's throughout a query dataset by positional comparison [Kristmundsdóttir et al., 2016]. However, this approach represents only half the workflow of VNTR detection, as paired end read mapping still needs to be carried out. This tool is therefore subject to the same criticisms as the lobSTR and STRscan.

Tan *et al.* (2010) performed a detailed analysis of VNTR's within the *P. falciparum* genome. They reported that 489 (9%) of the 5268 genes within the 3D7 reference bore VNTR regions when aligning against the HB3 and/or Dd2 strains [Tan et al., 2010]. In this paper we report 959 (18.2%) genes bearing VNTR's across the *P. falciparum* genome. This increase may be attributed to the fact that we used a larger dataset of 7 strains rather than 3. Furthermore, they only reported perfect or near perfect TR's of up to 12 bases repeat period size and maximum repeat size of 2000 bases, compared to our 15 base repeat period size and no limit to maximum repeat size. The majority of the repeats were <1000 bases in length, though one particularly notable repeat within the PF3D7_1038400

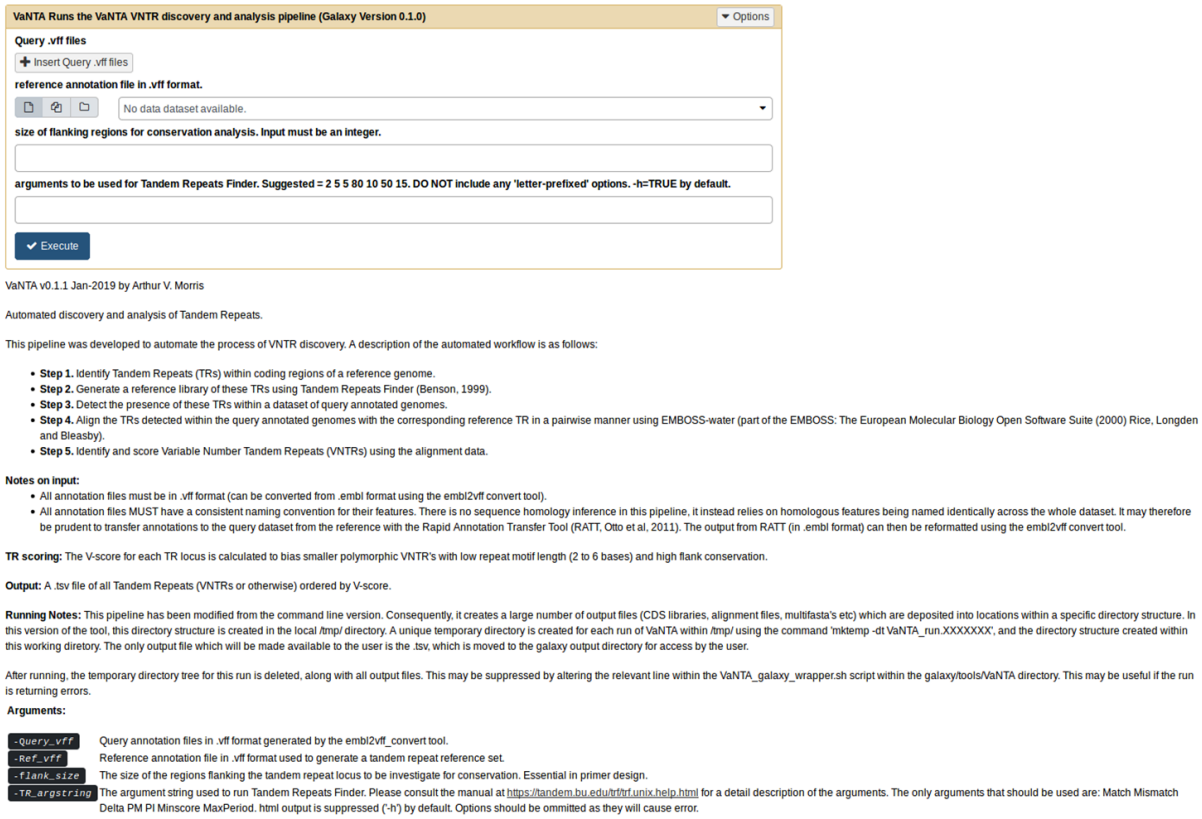


Figure 3.5: A Screenshot of the Galaxy implementation of VaNTA.

gene was reported at 24.6Kb with a period sequence of TGAAGAAGTTATACC, though it did not exhibit any variation within this dataset.

VaNTA necessitates the assembly and annotation of multiple genomes which, due to the constantly expanding number of high quality genomes being made publicly available, and the ever increasing ease with which they can be sequenced and assembled, VaNTA will become more and more valuable and applicable for large scale genomic analysis. We consider that most research teams are likely to have access to a sufficient number of high quality genomes and their corresponding annotations. These data should be utilised to their fullest extent, with particular emphasis on using annotation data to identify polymorphic TR's within coding regions, which are much more likely to be under selective pressure and therefore more likely to exhibit polymorphism that could be associated with strain virulence.

VaNTA is currently available for use via the Aberystwyth University Galaxy platform, facilitating its use by researchers who are not familiar with using tools designed to run via the Linux command line. Figure 3.5 shows a screen shot of the Galaxy implementation of VaNTA.

3.5 Conclusion

In this chapter we have presented VaNTA: a new bioinformatics pipeline which utilises an approach which is novel in its automation in detecting and analysing VNTR regions across multiple genomes. VaNTA represents the automation of a commonly used workflow which is usually carried out manually by research groups, and produces results that are comparable, allowing for identical downstream analysis and processing. We have shown that VaNTA can be used on data sets of two different apicomplexan parasites which can be accessed via online databases. However, usage of VaNTA is not limited to any organism in particular. VaNTA utilises the constantly increasing amount of high quality sequence and annotation data to perform *ab initio* VNTR identification and analysis in a time efficient and computationally un-intensive manner. To the best of our knowledge, this is the first pipeline or tool that has been made available to utilise assembled genomes and annotation data for the purpose of VNTR detection and analysis. Unlike the majority of the tools available, the VaNTA pipeline is intended to be run on personal computers and laptops in a reasonable amount of time. This allows for research groups and/or individuals who do not have access to high performance computing platforms, or belong to bodies or institutions which utilise sensitive data that cannot be submitted to tool hosting websites or for analysis on cloud computing services, such as the NHS, to carry out analysis.

In the next chapter, I present a computational tool which was developed to identify regions, such as the VNTR loci presented in this chapter, contained within reads from next-generation sequencing datasets. Using these loci, we may very quickly and accurately pull out and analyse reads which map to loci of interest.

Chapter 4

BlooMine: A Bloom filter driven tool for mining raw sequencing reads

4.1 Introduction

Alignment-free sequence analysis has the capacity to resolve the significant bottle neck between data generation and analysis which has been well documented in 'omics research, as it may be both faster and more computationally efficient than sequence alignment based approaches. This allows for the development of tools which can mine information from raw sequencing reads, obviating the significantly time consuming and computationally expensive task of assembly. Furthermore, in instances where sequences are divergent, alignment-free approaches may be more reliable than alignment based ones. The application of alignment-free sequence analysis should be treated with care, however, as it may be less accurate for some tasks. For example, it may be less sensitive than alignment based approaches if the sequences are highly conserved.

There are a great many tools which employ alignment-free methods to identify nucleotide and protein sequence similarity, developed to resolve problems in just about every discipline within Bioinformatics, including, but not limited to:

- Assembly

  de novo assembly [Zerbino and Birney, 2008, Zimin et al., 2013, Bankevich et al., 2012, Simpson et al., 2009, Rustagi et al., 2016, Peng et al., 2012]

  Read clustering [Marchet et al., 2017]

  Assembly error correction [Otto et al., 2010]

- Mapping

  Transcript quantification [Patro et al., 2014]

Variant calling [McKenna et al., 2010]

General read mapping [Li and Durbin, 2009, Langmead et al., 2009, Kojima et al., 2016]

- Metagenomics

Assembly-free phylogenomics [Pajuste et al., 2017, Wood and Salzberg, 2014]

Taxonomic profiling [Pajuste et al., 2017, Gupta et al., 2016]

However, no tools have been developed for the specific purpose of resolving multiple pathogen populations within clinical NGS data. The reason for this is that it requires a depth of read coverage which, until recently, have not been readily achievable from clinical data, where DNA yield may be low. This has lead to alternative approaches being used to attempt to identify heterogeneity of clinical samples (see Section 5.1 for a description of these approaches). Furthermore, multiple populations of pathogen species within a single host (termed Multiplicity of Infection - MOI) has been poorly documented for the vast majority of pathogens, with the majority of the research in this area being carried out on disease causing eukaryotes such as *Plasmodium falciparum* [Bell et al., 2006], *Trypanosoma brucei* [Balmer and Tanner, 2011], and pathogenic bacteria such as *Mycoplasma tuberculosis* [Gardy et al., 2018]. With modern methods of DNA isolation and purification, and the fall in price of using high-throughput NGS technologies to perform whole genome sequencing (WGS), investigating MOI by mining WGS reads has become a more feasible approach. Furthermore, it allows for unrestricted investigation of local heterogeneity at any position within the genome, something which is simply not possible to carry out using experimental methodologies.

Identifying VNTRs and MOI accurately and robustly is essential within a clinical context. The presence or absence of specific VNTRs in a clinically isolated NGS dataset may indicate disease within the patient. Diseases such as Alzheimer's have a documented relationship with the fragment length of a VNTR within ATP-Binding Cassette Subfamily A Member 7, wherein expansion of this region is strongly correlated with the development of Alzheimers disease [De Roeck et al., 2018]. Despite great advances in the use of NGS data for assembly-free genomic analysis, low concordance of variant calling pipelines has been reported, which remains a troubling issue [O'Rawe et al., 2013]. It is therefore of utmost importance that clinicians can be confident in the reliability of the bioinformatics tools and pipelines they employ to identify variants.

4.2 BlooMine

Here we present BlooMine: a novel raw read mining tool developed to facilitate quick and computationally efficient local analysis of sequences captured within raw reads generated by whole or partial genome sequencing projects. It utilises Bloom filters, a highly space efficient probabilistic data structure, to perform set membership queries and infer sequence homogeneity (see Section 1.5.2). The BlooMine package consists of three primary scripts to perform read mining:

- BlooMine_gen: Bloom filter generation
- BlooMine_screen: Bloom filter driven read screening
- BlooMine_spaln: Second-pass alignment free screening

During an analysis, scripts are called as a pipeline in this order (see Figures 4.1 an 4.3 for for a diagrammatic of the first two steps).

4.3 Bloom filter Generation: BlooMine_gen

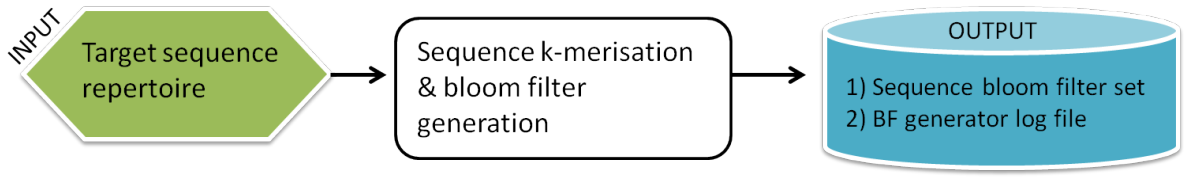


Figure 4.1: The procedure for generating a Bloom filter and control file using BlooMine_gen.

BlooMine_gen is the first script to be executed in the sequence mining pipeline. It is used to generate a Bloom filter from a target sequence in fasta format, which is necessary to screen reads against later in the pipeline. This is achieved by generating a kmer array from the target sequence, K_t^k (where K_t^k is a kmer array generated using kmers of size k from target sequence t), before passing each kmer through n hash functions, h, to generate a bit signature' for this kmer, which can be used to alter the bit array, B_t (where B_t is a bit array generated from a target sequence t). This bit array (also referred to as the Bloom filter) is output to a file in binary format (see Figures 1.6 and 1.7 for a graphical representation of how a Bloom filter is generated from a query sequence). A control file is also generated by this script, which details the arguments and target sequence used to generate the Bloom filter. The structure of the control file can be seen in Figure 4.2. This control file includes the arguments and files used to generate the Bloom filter, as well as the path to each Bloom filter generated. If a multifasta file is parsed into BlooMine_gen,

a Bloom filter will be generated for each, and the headers written to the end of the control file.

```
##BF_LOG## BlooMine control file generated 01-01-01 12:00:00

Output Path: /home/user/BlooMine/WD/
k: 7
false positive rate: 0.05
bit array length: Default

Parent FASTA path: /home/user/BlooMine/query_f1.fasta

Sequences detected in FASTA file: 1

>query_flank_1
```

Figure 4.2: The layout of the BlooMine_gen control file.

4.4 First-Pass Screening: BlooMine_screen

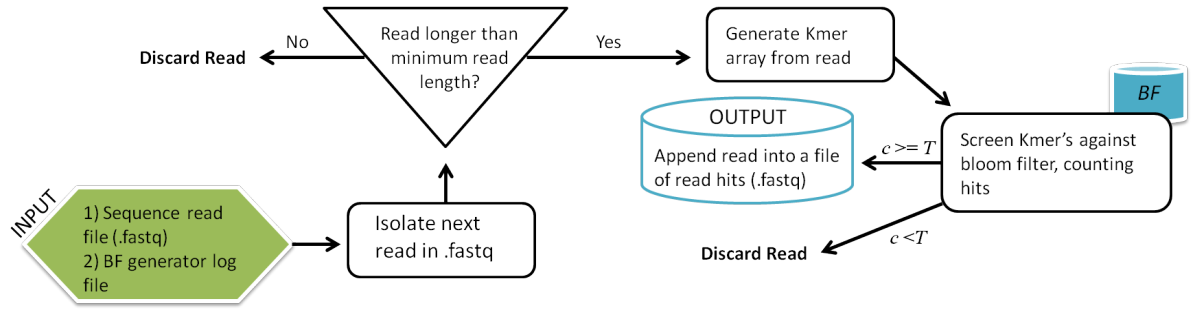


Figure 4.3: A diagrammatic representation of the BlooMine_screen algorithm. T is the threshold for minimum number of common kmers, c is the number of kmers intersection the target and read kmer arrays, inferred by Bloom filter hit events.

The control file detailed in Figure 4.2 is used as an argument when running BlooMine_screen, along with a raw read file in .fastq format. The BlooMine_screen tool constitutes the first pass screening, using the Bloom filter generated by the BlooMine_gen tool (see Algorithm 1 for a description of the Bloom filter generation process). The first pass screening tool (detailed in Algorithm 2) screens reads within the .fastq file against the Bloom filter. Each read is used to generate a kmer-array using the value of k detailed in the control file. This tool then iterates through this kmer-array and hashes each kmer using the

Algorithm 1 BlooMine_gen: Generate a Bloom filter

Input 1: a target sequence, t, in fasta format.
Arguments: integer value for k
Output: a Bloom filter and a control file.

1: $K_t^k \Leftarrow Kmerise : t, k$
2: **for all** $Kmer \in K_t^k$ **do**
3: $b_t \Leftarrow \{h^1 : Kmer, ..., h^n : Kmer\}$
4: incorporate the target sequence bit signature, b_t, into the Bloom filter bit-array
5: **end for**
6: write Bloom filter bit-array and control file to disk

Where:
t = A target sequence in .fasta format, with which a Bloom filter will be generated.
$Kmerise$ = A function which takes a sequence and generates an array of $Kmer$, K, of size k, ST is the minimum score to trigger a hit.
$h_1, h_2, ..., h_n$ = A set of n hash-functions, h.
b_t = The bit signature for a single $Kmer$ within K_t^k, generated by feeding the $Kmer$ to each hash-function. E.g. if there are 3 hash-functions, a single $Kmer$ will yield 3 hashes (one for each hash-function), and therefore a bit signature consisting of these 3 hashes (see an example in Figure 1.6).

set of hash functions which were used to generate the Bloom filter. The subsequent bit signature generated by this hashing process is screened against the Bloom filter bit-array to infer set membership. The detection of a hit increments the score by 1. If the overall score for the read exceeds the minimum score threshold, sequence similarity is inferred and the read returned and written to disk in .fastq format. This process can be seen on lines 3 to 13 of algorithm 2, and can be effectively represented (by stripping out Bloom filter functionality) as:

if $|K_t^k \cap K_r^k| \geq T$ **then**
return r
end if

Where K_n^k is a kmer array of sequence n of kmer size k, r is a read sequence, t is the target sequence, and T is the threshold for minimum number of common kmers.

Algorithm 2 BlooMine_screen First-pass screen: Bloom filter driven kmer set intersection

Input 1: a set of reads in .fastq format.
Input 2: a control file generated using BlooMine_gen.
Output: a set of reads bearing similarity to the target sequence, in .fastq format.
Require: a set of n hash functions (h^n) used to generate the $Bloomfilter$.

for all $read \in fastq$ **do**
  $K_r^k \Leftarrow Kmerise : read, k$
  for all $Kmer \in K_r^k$ **do**
    $b_{Kmer} \Leftarrow \{h^1 : Kmer, ..., h^n : Kmer\}$
    if $b_{Kmer} \in Bloomfilter$ **then**
      increment score
    else
      next $Kmer$
    end if
  end for
  if $score \geq ST$ **then**
    return $read$
  end if
  Write read hits to disc in .fastq format
end for

Where:
b_{Kmer} = A bit-array generated by hashing a $Kmer$ using the hash function array used to build the Bloom filter.
$Kmerise$ = A function which takes a sequence and generates an array of kmers, K, of size k.
ST = the minimum score to trigger a hit.

During this first pass screening procedure, the minimum score threshold is intentionally set low, allowing for a lower fraction of the target array to intersect with the read array to trigger a read hit. This results in a reduction of type-II error (FN) at the expense of a significant increase in type-I error (FP). The result of this is that the returned set of 'hit' reads contains a large number of reads which do not contain the target sequence, and triggered a hit due to simply bearing a sufficient number of similar kmers. These reads represent a subset of a larger set of reads contained within the fastq file. The intention of this first-pass filtering is to significantly reduce the number of reads which should be assessed for homology with the target sequence using the second-pass filtering tool (see Section 4.5); which performs in a very time-inefficient manner on very large datasets in comparison to the Bloom filter driven first-pass approach. This high type I error rate is resolved during second-pass screening.

4.5 Second-Pass Screening: BlooMine_spaln

Algorithm 3 BlooMine_spaln Second-pass screen: Alignment-free alignment inference

Input 1: A set of reads in .fastq format generated by the first pass screen.
Input 2: A target sequence used in the first pass screening process.
Output: A set of reads in .fastq format containing the target sequence.

1: $K_t^k \Leftarrow Kmerise : target, k$
2: **for all** $read \in fastq$ **do**
3: $K_r^k \Leftarrow Kmerise : read, k$
4: Initialise array of 0's, N, of length $|read|$
5: **for all** $Kmer \in K_r^k$ **do**
6: **if** $Kmer \in K_t^k$ **then**
7: $i =$ index position $Kmer$ maps to within K_r^k
8: $Kmer_i, ..., Kmer_{i+k-1} \mapsto N_i, ..., N_{i+k-1}$
9: **end if**
10: **end for**
11: $gap\_count = 0$
12: **for all** $p \in N$ **do**
13: **if** p is a gap **then**
14: increment gap_count
15: **else if** p is a hit **then**
16: **if** $gap\_count \geq T$ **then**
17: split N at $p - 1$
18: $gap\_count = 0$
19: **else**
20: $gap\_count = 0$
21: **end if**
22: **end if**
23: **end for**
24: $s \Leftarrow score\_alignments : N$
25: **if** $s \geq MST$ **then**
26: return $read$
27: **end if**
28: **end for**

Where:
B = A bit-array generated by hashing using a hash function array used to build the Bloom filter.
$Kmerise$ = A function which takes a sequence and generates an array of kmers, K, of size k.
p = a position within N.
T = a gap threshold, calculated as detailed in Equation 4.1.
$score\_alignments$ = a alignment scoring function, detailed in Equations 4.2 and 4.3.
MST = the minimum score to trigger a hit.

BlooMine_spaln performs the second-pass filtering using a novel 'soft-alignment' approach. The algorithm driving this soft-alignment can be seen in Algorithm 3. Given a target kmer array (line 1) and a read file in .fastq format, BlooMine_spaln will generate a kmer array for each read, retaining the relative positions of each kmer (line 3). A blank array, N, of the same length as the read is initialised (line 4), to act as a framework onto which kmers which intersect the read and the target arrays can be mapped. It will then proceed to identify sequences of kmers within the read array that intersect with those in the reference array (lines 5-7). These intersecting kmers are mapped onto N at the position the kmer hits within the read array. Alignment is inferred by identifying sequences of overlapping kmers within N (referred to here as contigs or contiguities), and scoring them in a manner similar to a standard pairwise aligner, where scores are improved by the presence of longer, or multiple contiguities, and penalised by gaps. This is achieved by iterating through the alignment array, N, and assessing each position for whether is it a hit or a gap (line 12-23). Upon the detection of a gap, *gap_count* is incremented (line 13-14). This gap count is retained until a hit is encountered, upon which, if the *gap_count* exceeds the gap threshold, T, then the alignment array, N, is split at the position prior to the encountered hit, and *gap_count* set to 0 (line 15-18). If $gap\_count < T$, *gap_count* is set to 0 and sub-alignment assessment proceeds (line 19-20). This yields an array of sub-alignments, each of which can be assessed by a scoring function, and the highest scoring sub-alignment returned (line 24). This scoring function can be seen in Equations 4.2 and 4.3. If the score of this maximum scoring sub-alignment exceed the minimum score threshold, MST (calculated as detailed in Section 4.5.1), a target hit is called and the read is considered to contain the target sequence, upon which the read is returned (line 25-26).

To identify discrete alignment chunks, a gap threshold is calculated as:

$$T = \left\lceil \frac{hk - g}{n} \right\rceil \tag{4.1}$$

Where:
h = match score.
k = kmer size.
g = gap open penalty.
n = gap extend penalty.

The function for scoring aligned chunks within BlooMine_spaln is:

$$s_c = \sum_{i=0}^{|c|} A : c_i \tag{4.2}$$

$$A : c_i = \begin{cases} s + h & | \; c_i \in K_t^k \\ s - g & | \; c_i \notin K_t^k \wedge c_{i-1} \in K_t^k \\ s - n & | \; c_i \notin K_t^k \wedge c_{i-1} \notin K_t^k \end{cases} \tag{4.3}$$

Where:
s_c = the score of the aligned chunk, c.
h = the value the score is incremented by when a hit is triggered.
g = the value the score is penalised by when a gap is opened.
n = the value the score is penalised by when a gap is extended.
c_i = a position, i, within the aligned chunk, c.

This gap threshold, T, is calculated such that a single kmer hit followed a gap equal to the gap threshold will result in a score of approximately 0. The alignment is broken at each position where a gap exists of $\geq T$. Every permutation of sequential sub-alignments are then scored, totalling $n(\frac{n}{2} + 0.5)$ where n is the number of sub-alignments.

A worked example of how an alignment array is processed and scored to identify the optimal scoring sub-alignment.

Consider N as the alignment `ACT---GACT----ACTG-----GACTGA`, and parameters of $k = 3$, $h = 4$, $g = 4$, $n = 2$. Using Equation 4.1, we calculate the gap threshold $T = 4$. Three sub-alignments will be yielded (colourised for clarity):
`ACT---GACT`
`ACTG`
`GACTGA`

This initial step is achieved by splitting N at gaps $\geq T$. The purpose of this step is to define sub-alignments which will be processed as a single unit. `ACT---GACT` has been redefined as a single sub-alignment due to the gap between `ACT` & `GACT` $< T$, and will therefore not be split during the next step. These will form six sequential sub-alignments:
`ACT---GACT` $s = 20$
`ACTG` $s = 16$
`GACTGA` $s = 24$
`ACT---GACT----ACTG` $s = 26$
`ACTG-----GACTGA` $s = 28$
`ACT---GACT----ACTG-----GACTGA` $s = 38$

This step generates every sub-alignment within N. Each of these will be assigned a score (s) using the function described in Equation 4.2 and the maximum scoring sub-alignment returned. If the score of this sub-alignment exceeds the minimum score threshold, MST, then a hit is triggered and the read returned as containing the target sequence. In this example, the highest scoring sub-alignment using these parameters is the entire alignment ($s = 38$).

This scoring algorithm reveals the optimum local alignment of the target sequence within the read using only a positional kmer intersection approach.

4.5.1 Minimum Score Threshold

Take $|s|$ to be the length of a given sequence, s, and F to be a mismatch frequency, whereby the maximum threshold for the number of mismatches across s resolves to $|s|\frac{1}{F}$. The calculation of the minimum score threshold, MST, is based on the premise that mismatches cluster, rather than present as an even distribution such as: `match match match mismatch`, since if $k \geq F$, there will be no kmers mapping to the read array and consequently will yield a score of 0. The score of an alignment where mismatches are clustered must be calculated.

To achieve this, windows of size $k + F - 1$ are utilised, within which we can approximate there to be on average one kmer from the target array, K_t^k, which does not map to the local region within the read which has triggered an alignment (this will be a region of approximately the same length as the target sequence, referred to previously as the maximum scoring sub-alignment). The number of kmers within K_t^k which must fail to map to a window in order that there are $|s|\frac{1}{F}$ base mismatches is therefore calculated as $e = \frac{k+F-1}{F}$. This method attempts to approximate the score of a sequence with a mismatch frequency of $\frac{1}{F}$. This cannot be resolved by simply calculating the minimum score as $h \cdot (|K_t^k| - m)$ (where h is the value the score is incremented by when a hit is triggered, and m is a value the score is penalised by when a mismatch is triggered) due to the dynamic nature of the scoring, whereby gap-open events are penalised differently to gap-extension events. For example, an alignment of `AT-AC-AA` will yield an event-string of [`hit hit gap-open hit hit gap-open hit`], whereas an alignment of `AT--ACAA` will yield a event-string of [`hit hit gap-open gap-extend hit hit hit hit`]. These event-strings will yield different scores if $g \neq n$, despite having identical lengths and hit counts.

The minimum score threshold (MST) is calculated using the following approach: Take

h as the value the score is incremented by when a hit is triggered, g as the value the score is penalised by when a gap is opened, n as the value the score is penalised by when a gap is extended (mismatch), and F as the frequency of mismatches, where F refers to the number of positions within the alignment which contains 1 mismatch on average.

The number of observation windows, w, across the alignment is:

$$w = \frac{|K_t^k|}{k + F - 1} \tag{4.4}$$

Where a single observation window is as a set of adjacent positions of length $k + F - 1$ within N.

The expected number of mismatch events within an observation window, e, is:

$$e = \frac{k + F - 1}{F} \tag{4.5}$$

For example, if $k = 5$ and $F = 10$, an observation window will be a stretch of 14 positions within N. This will resolve as an acceptable threshold of 1.4 errors within a single window. The total acceptable number of mismatches within the entire alignment, M, is therefore calculated as:

$$M = e(w - 1) \tag{4.6}$$

Consequently the minimum score threshold is:

$$MST = h \cdot |K_t^k| - wg + nM \tag{4.7}$$

$$= H - wg + nM \tag{4.8}$$

Where H is the score of a perfect match to the target sequence (repeats collapsed[1]), calculated as $h \cdot |K_t^k|$.

The purpose of performing these calculations across multiple windows of size $k+F-1$ is to calculate the number of match and mismatch events assuming mismatches cluster, rather than occur in a consistently distributed manner.

Therefore, the approximate observed rate of mismatch events is resolved by assuming that you will see $F - 1$ kmer hits followed by a sequence of mismatch events which will resolve to $\frac{1}{F}$ mismatches over a given alignment.

[1]Repeats were effectively collapsed by taking the kmer array for the target sequence as a set.

This alignment algorithm is novel and uses concepts from both alignment and alignment-free sequence analysis approaches for the purpose of establishing highly gapped or rearranged alignments, since regions flanking highly poymorphic target regions may be poorly conserved. Furthermore, alignments are carried out on small sequences (single reads) potentially many thousands of times, which may render conventional alignment approaches less efficient.

4.5.2 Screening Algorithm Runtime Comparison

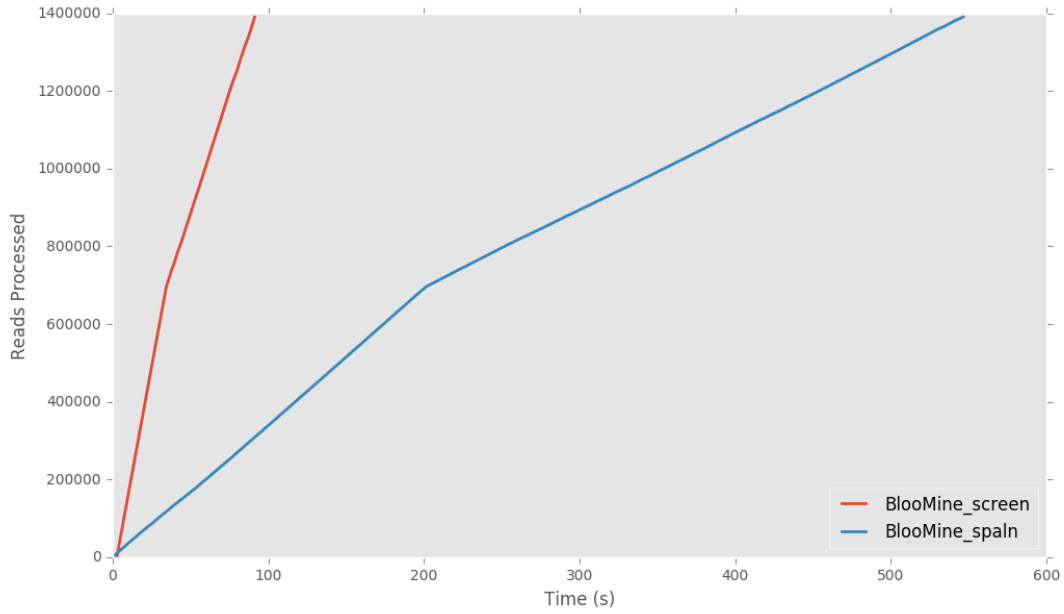


Figure 4.4: A comparison of the processing speed of BlooMine_screen vs BlooMine_spaln on a paired end read file containing reads generated by the sequencing of the clinical isolate *Cryptosporidium parvum* UKP10. The 5' gp60 flanking was used to screen these reads. Time was measured at every 1000 reads processed. The paired end read file contained a total of 1391780 reads. No minimum read length cutoffs were used in the generation of these data.

A comparison of the speed of BlooMine_screen against BlooMine_spaln can be seen in figure 4.4, which shows that BlooMine_screen is significantly faster in screening reads than BlooMine_spaln. The paired end read file was generated by the concatenation of the forward and reverse read files, to prevent the manipulation of low complexity regions that may occur during read pair merging. A greater number of read hits to the 5' gp60 flank were identified in the reverse read file. The drop in processing rate at roughly 700K reads (indicating the start of the reverse reads) can be explained by the increased amount of time read hits require to process compared to read misses.

4.6 Implementation as a Pipeline to Investigate Multiplicity of Infection

Multiplicity of Infection (MOI) is a state in which an individual is host to multiple subpopulations of a single pathogen species. It is generally used when referring to a plurality of genetically distinct populations of a pathogenic bacteria or parasite within the host. Nomenclature for viruses may vary, with the term 'quasi-species', referring to a population of viruses which diverge within a host, is roughly analogous. The implementation of this tool as a pipeline for the purpose of investigating MOI is straightforward, and is detailed in Figure 4.6. The pipeline is similar to the standard pipeline for target sequence detection, consisting of running the Bloom filter generation tool followed by the first-pass then the second-pass filtering tools. The MOI detection pipeline involves a few more steps. In this instance, the target region is presumed to be highly polymorphic and/or low complexity, which would make direct identification very difficult and could likely result in high False Positive (FP) and/or False Negative (FN) rates. Consequently, identification of more conserved and high-complexity regions directly adjacent to the target region are identified (termed 'flanks' or 'flanking regions'). With this approach, reads containing the target region can be isolated, despite the target region itself bearing large amounts of variation, which would confound a direct reference guided approach (see Figure 4.5).

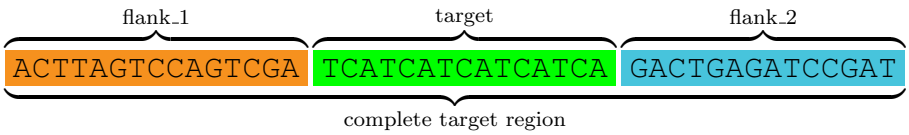


Figure 4.5: A model of a putatively polymorphic, low-complexity target region with conserved high-complexity flanking regions. Identification of both flanking regions within a single read imply the presence of the target region within this read.

Bloom filter generation using BlooMine_gen and first-pass screening using BlooMine_screen are carried out as usual, using flank_1 as the proxy target region. This first-pass screening step outputs the set of all reads which bear high kmer-content similarity to flank_1 (R_{fp}). Second-pass screening using BlooMine_spaln is split into two parts, second-pass screen I and second-pass screen II. Second-pass screen I is carried out using R_{fp}, and the target region (flank_1), outputting a read set (R_{sp1}) which consists of reads containing flank_1, where $R_{sp1} \subseteq R_{fp}$. Second-pass screening II is carried out using the corresponding flank of the pair (flank_2) as the target region, screened against the R_{sp1}. The output from second-pass screening II (R_{sp2}) is a set of reads that have captured both flanks, and therefore the target region, where $R_{sp2} \subseteq R_{sp1}$. A diagrammatic representation of this multi-pass BlooMine_spaln screening approach can be seen in Figure 4.7. This figure

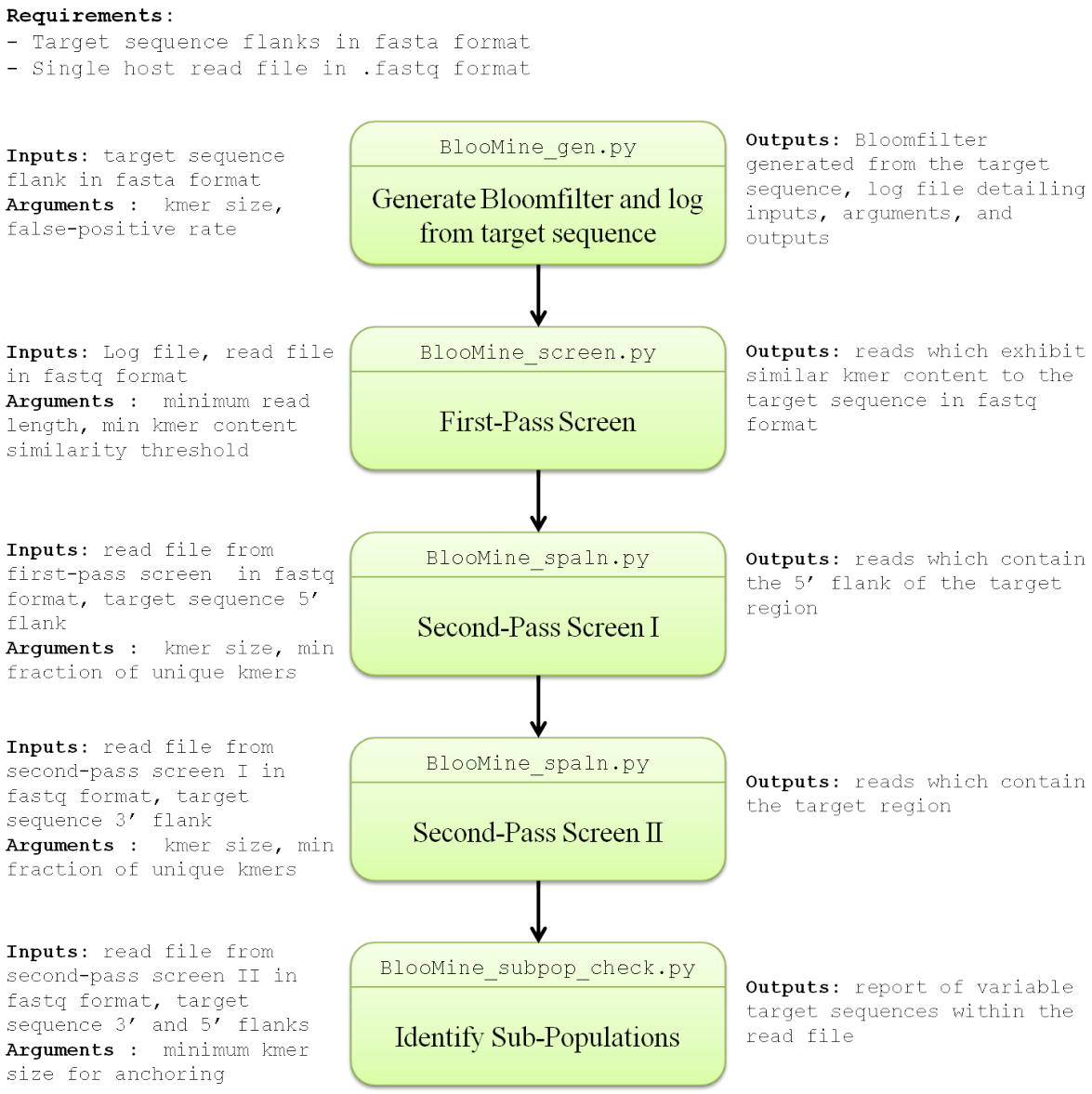


Figure 4.6: A diagrammatic representation of the pipeline implementation of BlooMine used to investigate sub-populations within a set of reads.

shows examples of different alignment outcomes, such as:

- Fragmented and weak alignment: indicating a read which bears kmer-content similarity but no homology to the flanking sequence, and therefore does not exceed the minimum scoring threshold (reads 1 and 5).

- Strong full, fragmented and partial alignment: indicating reads which contain the flanking sequence exhibiting strong full (read 3 for flank 1 and read 2 for flank 2), strong partial (read 3 for flank 2), or strong fragmented (read 2 and 4 for flank 1, and read 2 for flank 2) alignments which exceed the minimum scoring threshold.

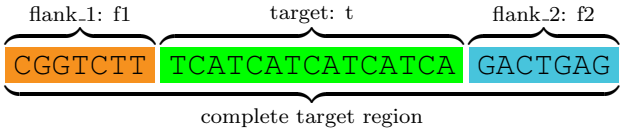
- Single flank alignment: indicating a read which exhibits only a single flanking sequence, and therefore captures only part of the target sequence (read 4).

Reads 2 and 3 of the example within the diagram have captured the target region (the region between the flanking sequences) in it's entirety, and will therefore be subject to further analysis within this pipeline.

Figure 4.7: A diagrammatic representation of the second-pass screen I & II by BlooMine_spaln, used to identify reads which capture the full target region for MOI investigation.

An example of potential erroneous local fragment resolution by target-flank interface variation using a standard alignment approach.

Take the complete target region to be:



and the read to be the sequence:

```
Sequence:    ATAAGTACGGTCTATCATCATCATCATCAAACTGAGACAAGAA
             |    |    |    |    |    |    |    |    |
Position:    0    5    10   15   20   25   30   35   40
```

Standard hard-alignment of the target flanks against the read would result in an alignment profile of:

```
Sequence:    ATAAGTACGGTCTATCATCATCATCATCAAACTGAGACAAGAA
                    ||||||.                .||||||
             -------CGGTCTT---------------GACTGAG-------
Position:    0    5    10   15   20   25   30   35   40
```

resulting in a partial alignment of flank_1 from position 7-12, and partial aligment of flank_2 from position 30-35. Positions 13 and 29 (representing the last and first base of the flanks respectively) present as mismatches, which would lead to the target region being isolated from position 13-29 and reported as ATCATCATCATCATCAA. This presents an 'off-by-one' error for both flank alignments, as the actual target region spans from position 14-28.

Isolation of this flank for the purpose of further polymorphism analysis downstream analysis would result in erroneous reporting of alleles (by fragment size) which are a result of sequence variation within the flank, rather than any variation within the target region itself.

Figure 4.8: A worked example of how using conventional hard-alignment tools to identify the start/end indices of a target region may lead to erroneous results.

Following this, a process to determine the number of alleles of the target sequence that are present within this final read set is carried out using BlooMine_subpop_check. These alleles are presumed to represent discrete sub-populations if a single copy of the target is present in the organism genome. It achieves this by performing a longest-kmer anchoring approach to identify the first base of the target region. This is necessary because if there are variations in the sequence at the interface between the flank and the target region due to either biological variation or sequencing error, a standard local aligner may

report a premature start/end of the target sequence (see Figure 4.8). If these coordinates are used to isolate the target sequence for fragment length and sequence polymorphism analysis, then it will report artefact alleles. This can be resolved, but would require further processing to identify the longest matching substring, which is precisely what the longest-kmer anchoring approach achieves without the need for an alignment step. This longest-kmer anchoring approach is detailed in Algorithm 4. It takes a set of reads in .fastq format generated by the second-pass screen II by BlooMine_spaln, and generates a kmer array for each of these reads (line 1). For each flank, $f1$ and $f2$, a 'kascade' array is generated. These are a set union of kmer-arrays from $k = k_{min}$ (where k_{min} is the minimum size a kmer must be to constitute an anchor) to $k = |t|$ (the length of the flanking region), where $t \in \{f1, f2\}$ (the set of both target flanking regions) refers to an unbroken string of elements, n, where $n \in \{A, T, C, G\}$ (lines 2 - 3). Each kmer within the kascade array, $K_t^{[k,|t|]}$, is tested for membership within the read kmer array, K_r^k, and the longest intersecting kmer used as the anchor point for calculation of the target start/end indices (lines 6 - 8). The first and last bases of the target are calculated using the position the longest kmer maps to within the read sequence (t_i) relative to the position it maps to within the flank ($Kmer_j^t$). The equations used to calculate these indices are as follows:

$$\text{target start} = f1_i + |f1| - Kmer_j^{f1} \tag{4.9}$$

$$\text{target end} = f2_i - Kmer_j^{f2} - 1 \tag{4.10}$$

Where:
$f1_i$ & $f2_i$ = the position the longest $Kmer$ maps to within the read, for flank_1 ($f1$) and flank_2 ($f2$) respectively.
$|f1|$ = length of the flank_1 ($f1$).
$Kmer_j^t$ = the position the longest $Kmer$ maps to within a flanking region, t, where $t \in f1, f2$.

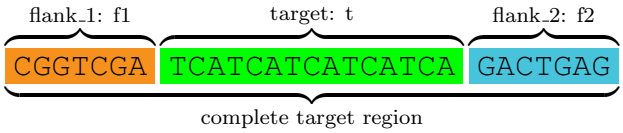
An example of this approach can be seen in Figure 4.9.

The indices produced by this algorithm can be used to very precisely isolate the target region between the provided flanking regions. This is necessary to investigate variation which is present within these target regions alone, and ignoring variation which may exist within the rest of the read.

The output from BlooMine_subpop_check is a report file in .txt format containing information about the amount of variation at the target region which is captured within the raw read file. Variation is assessed at both sequence and fragment length levels. An example of the report file can be seen in figure 4.10.

Worked Example of Longest Kmer Anchoring.

Take the complete target region to be:



and the read to be the sequence:

```
Sequence:    ATAAGTACGGTCTATCATCATCATCATCAAACTGAGACAAGAA
             |    |    |    |    |    |    |    |    |
Position:    0    5    10   15   20   25   30   35   40
```

The kascade array for flank_1 ($f1$) from $k = 3$ to $k = |f1| = 7$, $K_{f1}^{[k,|f1|]}$, is therefore:

Position	$k = 7$	$k = 6$	$k = 5$	$k = 4$	$k = 3$
0	CGGTCGA	CGGTCG	CGGTC	CGGT	CGG
1		GGTCGA	GGTCG	GGTC	GGT
2			GTCGA	GTCG	GTC
3				TCGA	TCG
4					CGA

This kascade array is then iterated over, and each $Kmer$ mapped to the read sequence. The longest $Kmer$ to map to the read sequence is CGGTC, which maps to position 0 within $f1$ ($Kmer_j^{f1} = 0$) and position 7 within the read ($f1_i = 7$). We can then calculate the start position of the target region, t, as:

$$\text{target start} = f1_i + |f1| - Kmer_j^{f1} \tag{4.11}$$

$$= 7 + 7 - 0 = 14 \tag{4.12}$$

This process is then repeated for flank_2 ($f2$), which yields the longest mapping kmer of ACTGAG, which maps to position 1 within $f2$ ($Kmer_j^{f2} = 1$) and position 30 within the read ($f2_i = 30$). We can calculate the end position of the target region, t, as:

$$\text{target end} = f2_i - Kmer_j^{f2} - 1 \tag{4.13}$$

$$= 30 - 1 - 1 = 28 \tag{4.14}$$

Which yields the target start-end coordinates as 14-30. Using these coordinates, the target sequence is identified as TCATCATCATCATCA.

Figure 4.9: A worked example of the longest Kmer anchoring approach used to identify target locus start and end points within a read.

Algorithm 4 Multiplicity of Infection Investigation: Longest-kmer anchoring

Input 1: A set of reads in .fastq format generated by second-pass screen II.
Input 2: The flanking regions of the target sequence.
Output: A report detailing sequence variation at the target location, across reads within the read file.

1: $K_r^k \Leftarrow Kmerise : r, k$
2: **for all** $t \in \{f1, f2\}$ **do**
3: $K_t^{[k,|t|]} \Leftarrow \bigcup_{k}^{|t|} Kmerise : t, k$
4: **for all** $Kmer \in K_t^{[k,|t|]}$ **do**
5: $Kmer_j \Leftarrow$ index $Kmer$ maps to in K_t^k
6: **if** $Kmer \in K_r^k$ **then**
7: $t_i \Leftarrow$ index $Kmer$ maps to in K_r^k
8: break loop
9: **end if**
10: **end for**
11: **end for**
12: yielding $f1_i$ and $f2_i$

Where:
t = a sequence, flank_1 or flank_2, flanking the target region. See Figure 4.5.
$[k, |t|] = \{k \in \mathbb{N} \mid k_{min} \leq k \leq |t|\}$.
$f1_i$ and $f2_i$ refer to the index the longest kmer of flank 1 and flank 2 respectively map to in the read array; K_r^k.
$Kmerise$ = a function which takes a sequence and generates an array of kmers, K, of size k.

```
Subpop report generated 01-05-19 15:54:38

fastq: /home/amorris/BlooMine/MWD/Cparvum/cgd6_1080.P.108-164/cgd6_
1080.P.108-164.head_cgd6_1080.P.108-164.tail_UKP3_cat.fastq
target flanks:      /home/amorris/BlooMine/MWD/Cparvum/cgd6_
1080.P.108-164/cgd6_1080.P.108-164_flanks.fa


Sequence polymorphisms:
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATC                1
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCGTCATCATCATCATCAACATC         1
TCACCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC         1
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC      2
TCATCATCGTCATCGTCATCAGCATCATCATCATCATCATCATCATCATCATCATCATCAACATC      1
TCATCATCGTCATGGTCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC         1
TCATCATCGTCATAGTCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC         1
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATTATCATCAACATC         1
TCATCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC            2
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCAACATC           20
GCATCATCGGCATCGGCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC         1
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC        139
CCATCACCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC         1
TCATCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCATCATCATCAACATC         1
TCATCATCGTCATCGTCATCATCATCATCATCATCATCATCATCATCATCATCAACATC            4

Length polymorphisms:
65      146
59      7
68      2
62      22
```

Figure 4.10: An example of a report file generated by BlooMine_subpop_check. The target sequence was cgd6_1080 at positions 108-164, referring to the VNTR within the gp60 locus. The read file mined for variation was *Cryptosporidium parvum* isolate UKP3 [Hadfield et al., 2015].

4.7 Method

4.7.1 Prediction Accuracy and Read Recovery on Simulated Data

To validate the precision of read recovery from .fastq files, a single-end read set containing artificial target sequences exhibiting multiple alleles was simulated. The target sequence consisted of a `'TCA'` repeat locus flanked by 40 nucleotide regions which was utilised as flanking sequences for target recovery. The target sequence exhibited 28 alleles within the read set, of lengths spanning a single repetition (3 nucleotides) to 10 repetitions (30 nucleotides) at increments of a single nucleotide. The purpose of varying the alleles (defined here as a variation in the target sequence) by less than a single repeat unit is to simulate challenging data to test the precision of the tool when identifying target regions as thoroughly as possible. The flanking sequence was varied, from 0 to 7 SNP's throughout the entire sequence. `N`'s were introduced within the flanking sequences at random positions, and randomly selected bases, to simulate variations which avoids the possibility of false-positive kmer hits. This totals 8 flanking sequences each flanking 27 allele sets at a depth of 30 reads for each flank-allele variation pair, for a total of 6720 simulated reads containing the target sequence in full. The non-target reads were generated as random strings of 130 nucleotides. The total number of reads within the simulated read set was 200,000. A model for the target sequence can be seen in figure 4.11.

Figure 4.11: A model for the target sequence, showing the target region (orange) consisting of a TCA repeat region, and flanking regions (blue). Where ... refers to an arbitrary number of the preceding repeat motif.

The BlooMine pipeline was ran 8 times on the simulated dataset with different kmer sizes for each screening pass in the pipeline (BlooMine_screen: first-pass, BlooMine_spaln: second-pass, & BlooMine_spaln: third-pass). The kmer sizes used are detailed in Table 4.1. Runs VI - IIX are hybrid runs, in that they use different kmer sizes for each screening-pass.

To identify optimal parameters for identifying target exhibiting reads within read sets, Receiver Operating Characteristic curves were generated by running BlooMine_screen tool on the simulated dataset using the threshold (`-t`) parameter set to 1 to 100, totalling 100 iterations. The threshold setting takes an integer for use as a minimum percentage of kmers within the target kmer array which must intersect with the read kmer array. False-Positive, True-Positive, False-Negative, and True-Negative were calculated, and metrics

BlooMine Pipeline Run	Kmer size: First-Pass Screen	Second-Pass Screen	Third-Pass Screen
I	3	3	3
II	4	4	4
III	5	5	5
IV	6	6	6
V	7	7	7
VI	7	4	4
VII	7	5	5
IIX	7	5	4

Table 4.1: The kmer sizes used for the first, second and third-pass screening processes, within the BlooMine pipeline, executed on the simulated dataset.

derived as detailed in section 4.7.2.

4.7.2 Confusion Matrix Derivations

A confusion matrix was used to derive metrics to elucidate the effectiveness of each BlooMine run in identifying target reads within a simulated dataset (see section 4.7.1). A confusion matrix is generated by calculating the False-Positive (FP), False-Negative (FN), True-Positive (TP), and True-Negative (TN). Metrics are derived from this matrix as follows:

$$\text{True-Positive Rate} = TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{4.15}$$

$$\text{True-Negative Rate} = TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{4.16}$$

$$\text{Precision} = PPV = \frac{TP}{TP + FP} \tag{4.17}$$

$$\text{Negative Predictive Value} = NPV = \frac{TN}{TN + FN} \tag{4.18}$$

$$\text{Miss Rate} = FNR = \frac{FN}{P} = \frac{FN}{FN + TP} \tag{4.19}$$

$$\text{False-Positive Rate} = FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{4.20}$$

$$\text{False-Discovery Rate} = FDR = \frac{FP}{FP + TP} \tag{4.21}$$

$$\text{False-Omission Rate} = FOR = \frac{FN}{FN + TN} \tag{4.22}$$

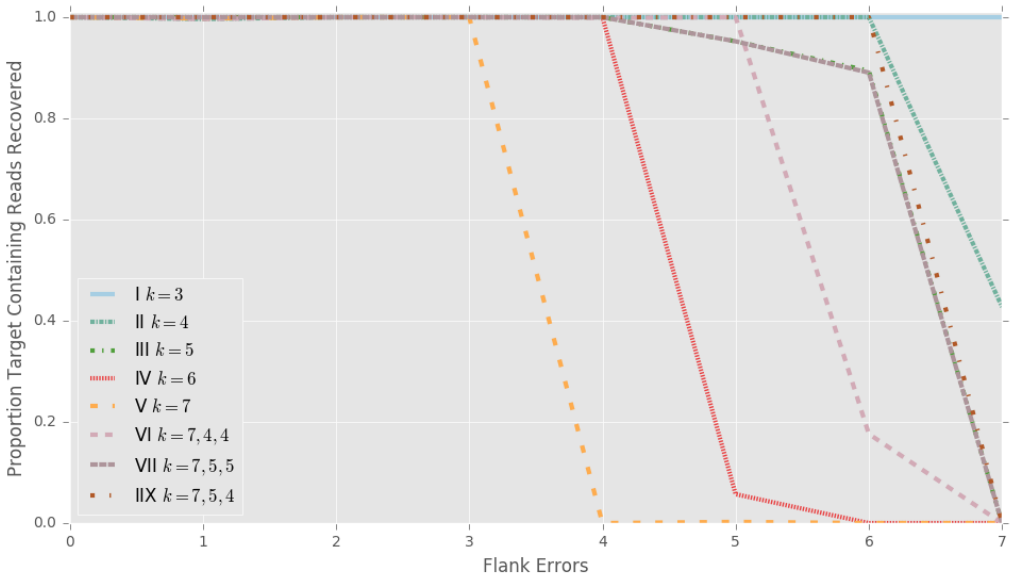


Figure 4.12: Total recovery rates from running the BlooMine pipeline on the simulated dataset at each kmer size. Proportions are shown as the total number of reads bearing targets of all allelic variations, grouped by the number of introduced flank errors.

4.8 Results

4.8.1 Results of BlooMine Simulated Dataset Mining

Figure 4.12 shows the recovery rates of target regions of varying error rates (0-7) from the simulated read set. The data illustrates that the recovery rates for run I (k=3) were 100% across all error rates, however the false-positive rates were also extremely high at this kmer size (0.9663). Run II ($k = 4$) also exhibited very high recovery rates, as well as very high false-positive (FP) rates (0.8675). Runs III-V illustrate some loss of target-exhibiting reads, where they were not identified as containing the target sequence due to flank variation. FP rates dropped as the kmer size increased, with run III ($k = 5$) exhibiting an FP rate of 0.0002, and runs IV and V ($k = 6$ and $k = 7$ respectively) showing FP rates of 0.0. The hybrid runs (VI-IIX) reported the best results in terms of overall read recovery rates and FP rates of 0.0, with run IIX (k=7, 5, 4) illustrating the highest recovery, wherein all target exhibiting reads bearing 0 - 6 errors within each flanking region were recovered. Target exhibiting reads containing flanking regions bearing 7 errors were not recovered. In flanking regions of 40 nucleotides, 7 errors would resolve at approximately 1 error per 5.71 nucleotides. If $k \geq \frac{e}{|f|}$ (where e is the number of introduced errors, and $|f|$ is the length of the region flanking the target), the read is unlikely to be returned as containing the target sequence, as there may not be a sufficient number of kmers which map to the target flanks to trigger a hit event.

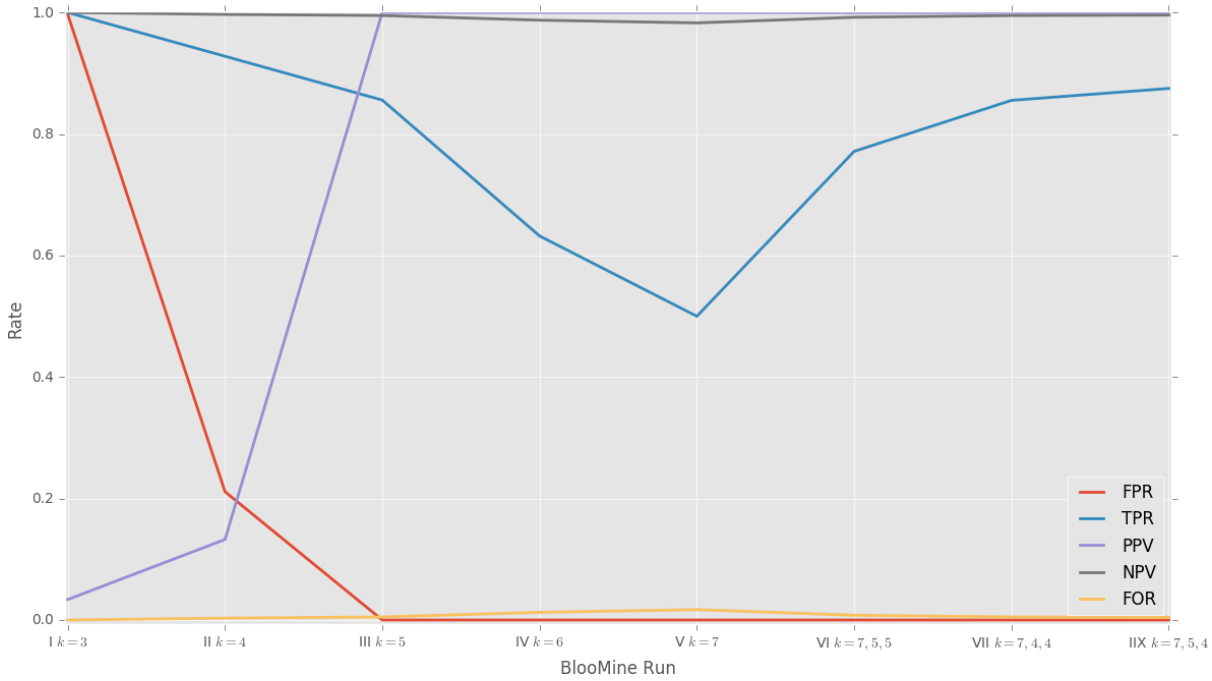


Figure 4.13: Rates of a set of metrics given as proportions between 0-1. Metrics calculated using the confusion matrix of returned reads by running the BlooMine pipeline on the simulated dataset using various kmer sizes. Metrics are calculated as detailed in section 4.7.2. See table 4.1 for a full description of the BlooMine runs shown on the x-axis. Where FPR = false-positive rate, TPR = true-positive rate, PPV = precision, NPV = negative predictive value, and FOR = false-omission rate.

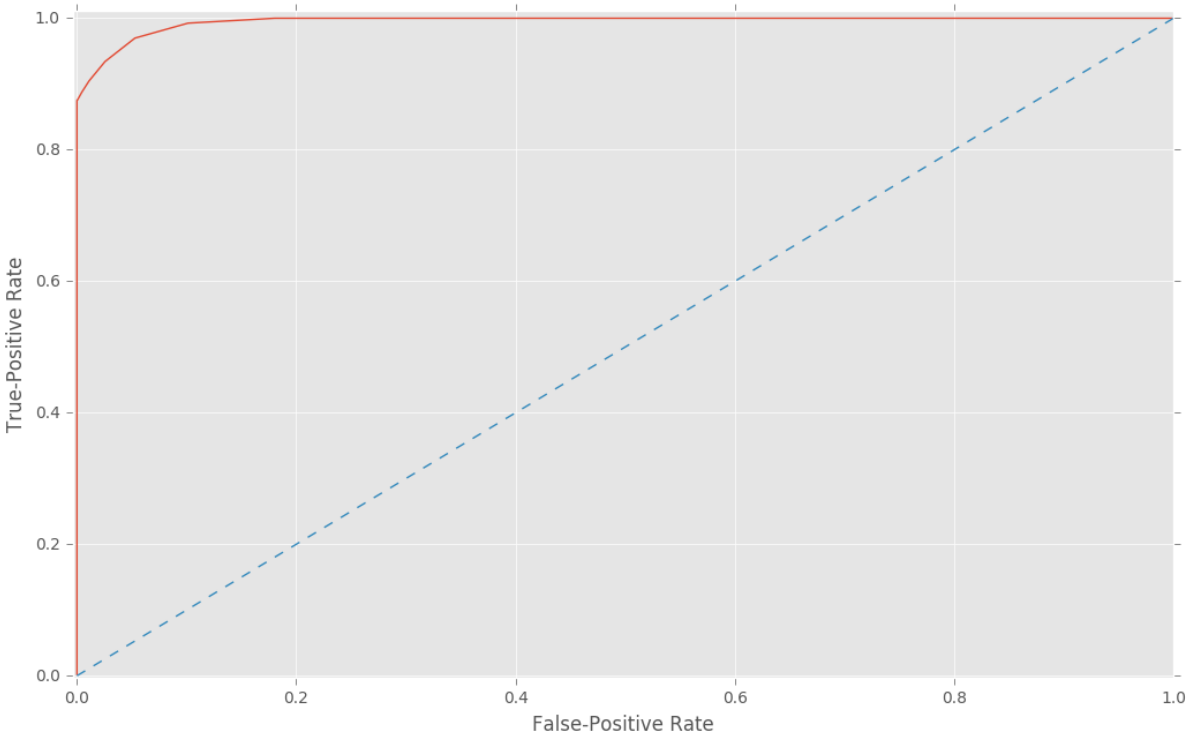


Figure 4.14: The receiver-operating characteristic (ROC) curve for BlooMine ran on the simulated dataset. False-Positive is plotted against True-Positive rates at each threshold level. The dashed line is a curve representing results which would illustrate predictive power no better than a random guess.

Figure 4.13 shows the False-Positive Rate (FPR), True-Positive Rate (TPR), Precision (PPV), Negative Predictive Value (NPV), and False-Omission Rate (FOR) for each BlooMine run. The results illustrate that hybrid kmer sizes used through the three screening procedures in the BlooMine pipeline are the most effective at identifying target-exhibiting reads. Of these, run IX $(k = 7, 5, 4)$ produced the most favourable results.

Figure 4.14 shows the Receiver Operating Characteristic curve for BlooMine_screen run iterations using minimum kmer match thresholds of between 1 and 100. The results illustrate clear distinction between type I (FP) and type II (FN) error. Figure 4.15 shows curves of all confusion matrix derivatives generated from BlooMine_screen executed at different threshold levels. The results indicate that accuracy increases sigmoidally, with maximum accuracy being achieved at a threshold of 35-40. Accuracy then decreases in an approximately linear fashion as the threshold surpasses 50. False positive rate decreases in an inverse sigmoidal fashion, with 0 being achieved at a threshold of roughly 35. True positive rate drops at a threshold of roughly 25, and plateaus at 30 before steeply declining at thresholds ≥ 50. Precision (Positive Predictive Value) increases in a sigmoidal fashion before reaching maximum at a threshold of roughly 37. False omission rate starts to increase at a threshold of roughly 50. These results indicate that the optimum threshold

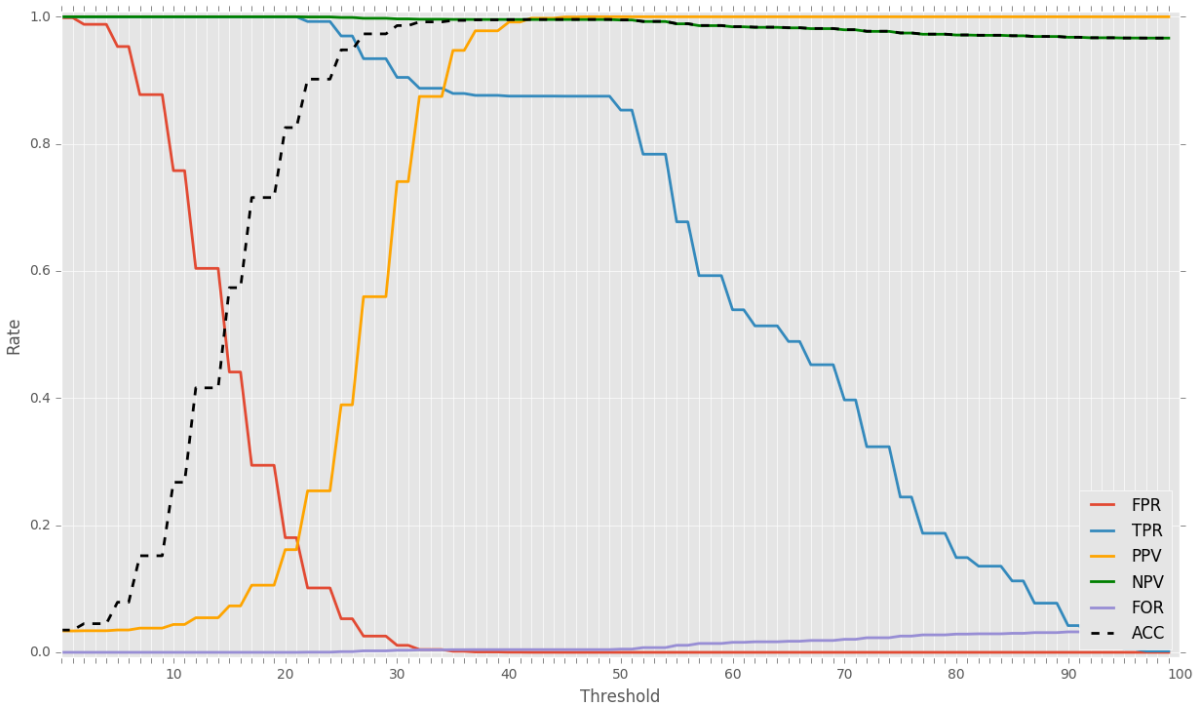


Figure 4.15: Confusion Matrix derivation curves for each threshold. See section 4.7.2 for a description of CM derivations used within this plot. Where FPR = false-positive rate, TPR = true-positive rate, PPV = precision, NPV = negative predictive value, FOR = false-omission rate, and ACC = accuracy.

to maximum TPR and TNR whilst minimising FPR and FNR is around 40.

4.9 Discussion

The number and quality of whole-genome sequenced *Cryptosporidium* samples taken from clinical samples is rapidly expanding, largely due to recent advancements the extraction, purification, and enrichment of high-quality DNA, and sequencing technologies. This facilitates the assembly-free analysis of parasite genomes, as well as the *in silico* investigation of MOI within clinical samples. The bottle neck currently observed between the generation and analysis of genomes can only be alleviated by advances in how we investigate genomic data, such as performing analysis on raw reads rather than genome assemblies. Investigation into the extent of MOI within clinical samples of *Cryptosporidium* will greatly improve our understanding of the evolution of virulence, drug efficacy, infection dynamics, pathogenesis, and transmission of *Cryptosporidium*. BlooMine is a general purpose read analysis tool which can be used for local analysis of genomes, quality assurance by investigating local read depth, and rapid determination of MOI by targeting regions flanking known variable loci. Sufficient read depth is required to establish MOI within a set of reads, particularly if the loci varies on a sequence level, rather than by fragment length. Furthermore, multiple copies of the target locus within the read set may

confound results of MOI analysis. We have shown that false-positive and false-negative results can be significantly alleviated by performing multiple stages of read screening based on different strategies, results in highly accurate and reliable detection of target sequence exhibiting reads within a read set. BlooMine has only been used with Illumina sequencing data, and as such, it is unclear as to whether it is usable with other sequencing technologies. Although k-mer based sequence analysis approaches should not vary based on the technology used to generate the sequence, aspects such as coverage uniformity and depth, error rate, and sequencing bias are known to vary, and therefore could affect results. For example, GC-bias is a documented issue observed in Illumina sequencing data [Benjamini and Speed, 2012].

Comparison of BlooMine against other tools developed to quantify MOI is difficult for a number of reasons:

1. Tools are generally developed to identify MOI in BAM files which have been generated using conventional mapping software, and are therefore subject to the weaknesses of these software.

2. These tools generally focus on SNPs and small INDELs, and are poor at identifying larger variations (such as VNTRs).

3. There are very few entirely *in silico* tools available to investigate MOI.

To my knowledge, the kmer soft-alignment algorithm implemented in the second-pass screening script is novel. There are a small number of tools which utilise positional kmer intersection approaches similar to the one we present in this script, such as that used in the Segmental Duplication mapping tool, ASGART [Delehelle et al., 2018], but none which utilise this approach to perform sequence alignment. This algorithm is highly effective at aligning small sequences with a high degree of accuracy, making it ideal for precise local analysis within reads. However, due to its method of investigating every possible sub-alignment and set of sub-alignments within a sequence to find the maximum scoring alignment, this algorithm scales poorly with sequence size. Consequently, it is unlikely to be appropriate for larger scale alignment.

A comparison of the technical functionality of BlooMine against similar tools can be found in Table 4.2. Due to fundamental differences in how BlooMine works when compared to approaches driven by conventional read mapping software (aligning small sequences against reads, as opposed to aligning reads against larger sequences), a direct comparison using a dataset was not practical.

TOOL	target organism	input format	Supported Variant Types			variation tolerance
			SNP	INDEL	VNTR	(n/v)
BlooMine	any	FASTQ	yes	yes	yes	$k+1$
estMOI	any	BAM	yes	yes	limited	mapper dependant
quasitools	virus	BAM	yes	yes	limited	mapper dependant

Table 4.2: Comparison of BlooMine against two other available *in silico* MOI/quasispecies analysis tools [Marinier et al., 2019, Assefa et al., 2014]. The error tolerance threshold is given as nucleotides per variation (n/v).

The primary strengths of BlooMine over other similar tools are its capacity to run directly on raw reads, rather than BAM files, which circumvents the error tolerance thresholds employed by many read alignment tools. Although such tools can comfortably handle smaller indels which high fidelity, these thresholds often make it more difficult to align reads which exhibit high indel-type variability, such as that seen in many tandem repeat regions [Ziemann, 2016]. Due to the method BlooMine employs, sequence variation captured in reads can be identified to a degree which is beyond the capacity of popular read mapping tools such as BWA and Bowtie2.

The memory usage of this tool is very efficient due to the use of Bloom Filters to store sets of kmers. This allows for its usage on personal desktops and laptops, obviating the necessity of using High-Performance Computing systems. The benefit of this is to allow it to be utilised in analysing datasets by research teams who do not have access to such systems. Furthermore, with some adaptation, BlooMine could be adapted for use in tandem with the new generation of highly portable sequencers, such as the Oxford Nanopore MinIon platform, wherein reads could be screened in real time as they are produced by the MinIon. The error rate currently exhibited by MinIon may affect the results of BlooMine, though the results presented in this chapter indicate that BlooMine is effective at handling relatively high error rates. This should be treated with care, however, as the recovery rate of reads with high error rates is highly sensitive to the parameters used (see Figure 4.12).

Due to the run time limitations of writing BlooMine in Python, a great deal of improvement on execution time could be achieved by rewriting parts of this tool using a compiled language, such as C, C++, or Rust. Such an improvement may decrease the execution time of first and second pass screening by around 100 times.

BlooMine still requires optimising to be able to handle multiple query sequences when assessing a .fastq read set for variation at these target regions. Presently, it can perform first stage screening (see Section 4.4) on multiple target sequences, but not second stage

screening (Section 4.5) or MOI assessment (Section 4.6).

4.10 Conclusion

In this chapter I have presented BlooMine, a novel bioinformatics tool to facilitate local analysis of sequences from raw read files. I have demonstrated its efficacy at identifying reads which contain a target sequence from large simulated read sets, which is performs to a high degree of accuracy and reliability. Furthermore, I have detailed extended functionality which allows the user to characterise Multiplicity of Infection by identifying target variation within a set of reads with a high degree of accuracy. The reliability and sensitivity of BlooMine to identify target sequences within read sets and quantify MOI will increase as sequencing technologies improve, allowing for deeper genome sequencing with more even coverage. Furthermore, as the number of genomes available increases, the population structure and relationship between populations will continue to be resolved with increasing levels of detail.

Chapter 5

Investigating Multiplicity of Infection in Cryptosporidium

5.1 Introduction

Multiplicity of Infection (MOI) is a well documented observation within the natural world. It has been reported that the majority of parasitic infections consist of multiple parasite species or discrete genetic lineages. In 1999, Lord *et al* reported that the majority of human adults infected with *Plasmodium falciparum* were host to more than five distinct strains [Lord et al., 1999]. Multiple infections have a significant impact on the incidence and spread of parasites, along with their virulence. These factors have been acknowledged to be of enormous importance to public health. An incomplete understanding of the subtypes and genotypes of a parasite within a host leads to potentially inaccurate assumptions about the clinical presentation and outlook of the infection. In particular, the virulence of experienced by a host infected by multiple subtypes/genotypes of a parasite (often referred to as 'overall virulence') is a result of the interactions between the different sub-populations within the host, resulting in an overall virulence of anywhere from greater than the most virulence to less than the least virulent, depending on various biological factors [Seppälä et al., 2009, Seppälä et al., 2012, Alizon et al., 2013]. Efforts have been made which illustrate that there exist general patterns when ecological and evolutionary theory are applied to within-host dynamics, leading to the generalization that 'basic ecological rules govern the outcome of co-infection across a broad spectrum of parasite taxa' [Graham, 2008].

It is both biologically plausible (due to unrestricted sexual recombination between sub-populations, see Figure 5.1) and there is strong evidence (described below) that infections can arise from, and give rise to, multiple sub-populations of Cryptosporidium *spp.* which will be present in single clinical samples. The current approaches of sequencing and assembly results in the resolution of a single allele at each locus which, if multiple

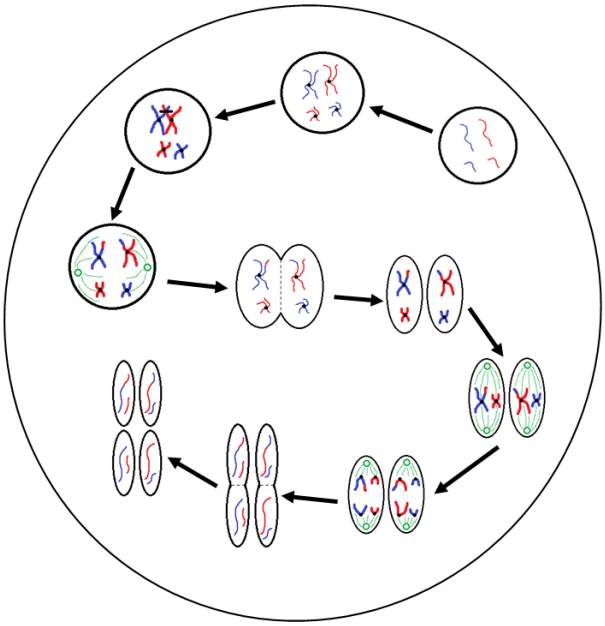


Figure 5.1: A simplified schematic of genetic recombination in Cryptosporidium, potentially generating variation between sporozoites within oocysts. In a mixed infection population, different fertilization scenarios potentially occur - between the same genotypes (resulting in identical daughter sporozoites) or between different genotypes, as in the example shown, that result in a variety of outcomes depending on the random genetic exchange, or lack of, that occurs during meiosis. For simplicity only two example chromosomes are shown with DNA from different genotypes represented by blue and red. The diploid zygote contains duplicate pairs of chromosomes, one set from each parent cell; during interphase (In) the DNA in each chromosome is replicated to produce two identical sister chromatids held together with a centromere; in prophase I (Pr I) the chromosomes start to condense and pair up with the homologous chromosome from the other parent cell, and cross-over can occur resulting in an genetic exchange; during metaphase I (Me I) the paired chromosomes line up along the center of the cell and microtubules connect the centromeres to the centrosomes (shown in green); during anaphase I (An I) each complete set of chromosomes (still paired as sister chromatids) are pulled towards each centrosome. The chromosomes from either parent are randomly combined at this phase introducing a further opportunity for recombination (a blue and a red chromosome are drawn to each centrosome in this example); in telophase I (Te I) the chromosomes start to unravel and cytokinesis starts to split the cell into two, resulting in two haploid cells; in prophase II (Pr II) the chromosomes condense again; during metaphase II (Me II) the chromosomes line up along the center of the cells and microtubules connect the centromeres to the centrosomes; this time during anaphase II (An II) the sister chromatids are separated and pulled apart towards the centrosomes, creating new daughter chromosomes; finally in telophase II (Te II) the chromosomes unravel and cytokinesis starts to split the cells, which in the case of this example due to the crossover event in prophase I, results in four genetically different haploid sporozoites. Depending upon whether random genetic exchanges take place between chromosomes from different genotype parents (either in prophase I or anaphase I) the resulting haploid sporozoites can either be all different, two pairs of identical sporozoites that are different from each parent, or two pairs of identical sporozoites that are the same as the two parents [Morris et al., 2019c].

discrete sub-populations exist within the isolate, would in effect simply represent the most populous sequence variant at each locus within the assembly. This may confound epidemiological analysis, which generally relies on the assumption that large scale genetic recombination does not occur within a host, and that in general a single host exhibits a single population. It is therefore essential that these in-host variations are well understood and accounted for in order to develop novel prevention strategies in the fight against Cryptosporidiosis and other parasitic diseases. The investigation into the impact of MOI within a host relies on the accurate and reliable detection and discrimination of discrete populations of parasites within a host.

There are a number of approaches to achieve this:

- Restriction Fragment Length Polymorphism (RFLP) analysis by PCR of variable regions.
- Cloning and sequencing key loci to detect variation.
- Isolating and sequencing single oocysts from clinical isolates.
- Investigating sequence variation among reads within short read archive generated by sequencing of clinical isolates.

These approaches attack this problem from very different angles: variable locus cloning and single cell sequences from an experimental angle, and sequence variation within reads from an *in silico* angle. This lends them unique challenges to overcome.

5.1.0.1 PCR-RFLP

In 2004, Henning *et al.* investigated the relationship between MOI and morbidity in Tanzanian children infected with *Plasmodium falciparum*. They performed PCR on the highly variable msp2 (a merozoite surface protein associated with host cell invasion) and defined heterogeneity within samples by identifying restriction fragment length polymorphism (RFLP) within this locus. Their results indicated a complex relationship between MOI, immune status, and endemicity. They found that children who are subjected to continuous infection due to multiple infective bites over a period of time, and therefore exhibit high levels of MOI, are less likely to exhibit high levels of morbidity than those who are subjected to little previous exposure, but also show high levels of MOI. They suggest that this is possibly linked to short-term immune cross-protectivity elicited by long term exposure to multiple strains of *P. falciparum* [Henning et al., 2004].

5.1.0.2 Biomarker Cloning

In 2013, Grinberg et al reported the presence of a number of sub-populations within single isolates of *C. parvum* by cloning PCR amplicons of selected loci (gp60 and HSP70) and utilising Next Generation Sequencing (NGS). They demonstrated the presence of two HSP70 and 10 gp60 alleles within their two isolate dataset. Furthermore, they reported that in both isolates there is a dominant allele, which represented the majority of the amplicons sequenced, and a number of sub-dominant ones [Grinberg et al., 2013].

To compare the results obtained from Sanger sequencing and NGS, Zahedi *et al.* compared results of Sanger and NGS sequencing of Gp60 amplicon from 11 *C. hominis*, 22 *C. parvum*, and 8 *C. cuniculus* animal samples from Australia and China. They demonstrated that NGS is more effective at resolving the presence of multiple populations of Cryptosporidium within a sample, and the extent of multiplicity of infection. There was concordance between the subtypes identified by both platforms, but additional subtypes were identified using NGS on *C. parvum* and *C. cuniculus* Gp60 amplicons, but not *C. hominis* [Zahedi et al., 2017]. In 2017, Kaupke *et al.* attempted to resolve the presence of multiple species of Cryptosporidium within pigs by using NGS (MiSeq sequencing) on amplicons containing the 18 SSU rRNA gene locus. Molecular analysis suggested the presence of Cryptosporidum species other than the pig-specific species, *C. suis* and *C. scrofarum*, however, this was not demonstrated using NGS. These results indicated that, although in-host subtype and genotype-level diversity may be complex, species-level diversity may not be. The results also further indicate the discordance between molecular and NGS based analyses when investigating MOI [Kaupke et al., 2017].

5.1.0.3 Single-Cell Sequencing

In 2016, Troell et al attempted to develop the protocol of single cell sequencing for Cryptosporidium for the purpose of elucidating these putative intra-isolate sub-populations. Using the protocol they developed, they sequenced 10 oocysts, resulting in assemblies of 49.4 - 91.8% of the size of the *C. parvum* IowaII reference genome [Abrahamsen et al., 2004]. By pooling the reads from all 10 oocysts, they generated a 94.4% complete genome. Using these genomes, the authors detected variation at multiple loci between the assembled genomes, verifying the presence of discrete populations within the isolate [Troell et al., 2016].

The major issue with these aforementioned experimental approaches is that they are extremely labour intensive and time consuming, leading to poor scalability. This leads to difficulties in generating sufficient data with which to begin to unravel the role of these parasite sub-populations in altering their ability to reduce host-fitness (virulence)

and transmission, as well as generating novel subtypes via sexual recombination and the overall impact these factors have on global public health. There is therefore a great need to develop strategies which can carry out this kind of analysis in a high-throughput manner, utilising the wealth of raw genomic data which is available for Cryptosporidium and other related parasites.

A few tools and pipelines have been developed for the purpose of identifying MOI in NGS data. These tend to utilise conventional read mapping software, in conjunction with bespoke scripts. estMOI is one such tool, which identifies local heterogeneity in read sets using smalt, and a set of perl scripts. The authors demonstrated its efficacy on clinical and cultured read sets of *P. falciparum*. However, as it is driven by smalt, it is limited to SNP and small indel identification. It may therefore be confounded by more variable, and/or lower complexity sequence, or larger levels of variation such as that seen in Tandem Repeats [Assefa et al., 2014].

There are a number of tools which are dedicated to identifying homozygous variation (local variation between a query and a reference) within a read set, which could potentially be adjusted to identify heterozygosity. Such tools include Dante, VNTRseek [Gelfand et al., 2014], lobSTR [Gymrek et al., 2012], & RepeatSeq [Highnam et al., 2013]. Each of these tools identify TR variation by mapping reads onto a reference set of TR's previously identified by running a TR discovering tool (e.g. Tandem Repeats Finder [Benson, 1999]) on a reference genome. The performance of these tools in their capacity to identify heterozygosity within a read set is dependant largely on the read mapping software that drives the tool, as some mapping software will deal with high-levels of local variation in a manner more appropriate to variant calling. Larger amounts of variation in the target locus may lead to data loss, where alignment is not achieved. Furthermore, as these tools are not designed for identification of sample heterogeneity, they may need development to support this function.

Problems associated with high levels of variation at a target locus, which could confound conventional read mapping approaches, can be circumvented by mapping to more conserved regions flanking the target sequence. However, mapping to regions which are less than the average read length is not possible. Consequently, an alternative approach may be to, in effect, map the target flanking regions to the reads. Any read which contains both flanking regions can be assumed to also contain the low complexity target region. The method by which the presence of a sequence within a read is ascertained, however, is an important factor, since there may be many millions of read pairs within an NGS read file, which necessitates very time efficient methods of sequence similarity assessment which can be repeated many millions of times in an appropriate time frame.

Isolate	Total reads	Mean read size	Mean depth of coverage	Proportion of the genome covered	$G \mid W_1$
UKP2	16,121,704	128.86	51.80x	0.93	0.223
UKP3	67,417,128	132.65	166.42x	0.89	0.556
UKP4	74,741,128	137.86	192.48x	0.89	0.566
UKP5	10,571,200	148.14	26.86x	0.85	0.277
UKP6	44,788,256	138.44	104.83x	0.82	0.255
UKP7	29,974,248	137.68	77.85x	0.89	0.555
UKP8	89,290,752	124.82	174.39x	0.84	0.566

Table 5.1: Basic statistics for the Hadfield *et al.* *C. parvum* read file dataset. The total reads includes both forward and reverse. The proportion of the genome covered was calculated using bowtie v2.3.3.1 to align the reads against the *C. parvum* IowaII reference genome. The GINI was calculated as detailed in Section 2.2.3.

5.2 Method

5.2.1 Predicting Multiplicity of Infection in Clinical Data

To investigate multiplicity of infection in clinical data from *C. parvum*, we utilised VNTR regions identified by VaNTA within both of these organisms (see Chapter 3). We used the 100 top scoring VNTR loci as targets (see Table 5.2), and the Hadfield *et al.* dataset, consisting of raw read files from the single-host clinical samples UKP2-8 [Hadfield et al., 2015]. Basic statistics for each of these read files can be seen in Table 5.1. For further analysis of the viability of MOI-signatures as a novel typing scheme, *C. parvum* isolates from Dataset 2 (see Section 2.1.1) were also investigated using 2 chosen loci. The mean read length for all isolate read files within the Hadfield *et al.* dataset is 135.49x (see Table 5.1), indicating that the total size of the target sequence and the two flanks should be around this size or less to be captured in a single read. The algorithms for target detection within BlooMine do not necessitate the entire flank being detected within the read to trigger a hit, and there is therefore some flexibility. The targets which are likely to significantly exceed the mean read length (where the size of both individual flanks is 41 nucleotides) and therefore unlikely to be detected in full are highlighted in red in Table 5.2. These target loci are included in this dataset for the purpose of providing an example as to how these larger target sequences are processed by BlooMine, as target fragments may still be identified and returned. Read pairs were not merged, due to the potential for repeat collapse, which would affect the results of sub-population identification in instances where the target is a repeat region.

Using the MOI data generated using the Hadfield *et al.* dataset, two loci were selected for further analysis. *C. parvum* isolates from Dataset 2 (see Section 2.1.1) were then analysed for MOI at these loci ($n = 29$).

The $G \mid W_1$ for UKP3, UKP4, UKP7, and UKP8 are high (0.556, 0.566, 0.555, and 0.566 respectively), which indicates a high level of coverage inequality. This suggests that some loci will be covered to a higher depth than others, which may be reflected in the level of in-host variation detected for target sequences.

Target Locus	**TR subseq**	***C. parvum* IowaII TR length**	**head flank cons. (%)**	**tail flank cons. (%)**
cgd1_140.P.2350-2463	CTC	125	100	85.714
cgd5_4480.P.1455-1500	AAT	46	85.714	100
cgd3_1250.P.634-662	AT	29	100	100
cgd8_4170.P.2518-2555	TAA	38	100	100
cgd7_420.P.4749-4787	GAACAA	39	100	100
cgd7_1190.P.2275-2313	TCTTCC	39	100	100
cgd3_1330.P.464-501	TCC	38	100	100
cgd3_3620.P.935-981	AAAGAC	47	100	100
cgd2_3540.P.5762-5808	TCA	47	100	100
cgd6_1080.P.108-164	TCA	57	85.714	85.714
cgd8_660.P.1554-1715	CCAGGA	162	100	100
cgd4_1340.P.815-878	TAA	64	100	100
cgd7_440.P.1066-1129	AAG	64	100	100
cgd6_780.P.1043-1111	GA	69	100	100
cgd6_490.P.468-543	AAT	76	100	100
cgd8_680.P.2962-3038	CACAACCAT	77	100	100
cgd7_1010.P.8357-8436	TAA	80	100	100
cgd8_700.P.747-842	ACTCCT	96	100	100
cgd8_1220.P.825-881	ACTTCT/CTT	97	100	100
cgd5_4090.P.428-544	GAATCT	117	100	100
cgd8_700.P.5284-5400	CTGGTG	117	100	100
cgd2_3590.P.5378-5402	TAC	25	100	100
cgd3_280.P.3887-3911	AAT	25	100	100
cgd4_3450.P.4333-4357	AATCTG	25	100	100
cgd6_3930.P.1748-1773	AGCTCCTCC	26	100	100
cgd5_130.P.4193-4218	TCA	26	100	100
cgd5_4190.P.231-257	GTA	27	100	100
cgd1_800.P.182-208	ACA	27	100	100
cgd8_4860.P.506-532	GGA	27	100	100
cgd7_1010.P.4475-4501	AGG	27	100	100
cgd6_520.P.300-327	AAT	28	100	100
cgd8_4170.P.2262-2289	ATA	28	100	100
cgd8_2260.P.2694-2721	AAT	28	100	100
cgd8_2250.P.3218-3246	AAT	29	100	100
cgd1_3550.P.1267-1295	TAA	29	100	100
cgd7_420.P.3445-3474	ATA	30	100	100
cgd8_4170.P.1991-2020	GAT	30	100	100
cgd8_2250.P.3970-4000	TTATAATCGGGATCC	31	100	100
cgd7_4980.P.5482-5513	TAA	32	100	100
cgd2_3700.P.4257-4288	GAGGAT	32	100	100
cgd2_3550.P.1472-1504	TTCCACTTCTGC	33	100	100
cgd7_5010.P.3301-3332	GAG	32	100	100
cgd2_3540.P.4474-4507	TAA	34	100	100
cgd4_1610.P.322-355	ATA	34	100	100
cgd4_2090.P.192-226	GAAGGAGAG	35	100	100
cgd6_3690.P.595-629	AAG	35	100	100
cgd4_3970.P.1521-1555	CCTATG	35	100	100
cgd2_400.P.329-363	TCA	35	100	100
cgd2_3590.P.18288-18323	AAT	36	100	100
cgd7_1010.P.8072-8108	TAA	37	100	100
cgd6_610.P.1783-1819	TAA	37	100	100
cgd8_970.P.3734-3771	TCTGGATCTGGA	38	100	100
cgd3_90.P.1112-1149	AAT	38	100	100
cgd3_720.P.16227-16264	CTACAACTA	38	100	100
cgd8_550.P.5068-5107	TCATAATAATGACAA	40	100	100
cgd6_830.P.7342-7445	CAA/ACAACAGCAACAACA/CAT	140	100	100
cgd1_3170.P.4177-4216	ATTCTGATTCCA	40	100	100

Continued on next page

Table 5.2 – continued from previous page

Target Locus	TR subseq	*C. parvum* IowaII TR length	head flank cons. (%)	tail flank cons. (%)
cgd4_3660.P.167-207	GAG	41	100	100
cgd6_5110.P.5014-5056	ATTCAG	43	100	85.714
cgd4_1420.P.405-447	AAT	43	100	100
cgd5_3750.P.497-608	GAT/ATA/AATAATAATAAC	146	100	100
cgd8_4940.P.1528-1573	TAG	46	100	100
cgd4_1360.P.1109-1155	TAA	47	100	100
cgd7_4500.P.633-679	CAC	47	100	100
cgd1_3450.P.494-540	GATTCT	47	100	100
cgd8_1570.P.575-622	TCT	48	100	100
cgd1_3060.P.589-637	CATCCT	49	100	100
cgd8_1380.P.662-811	GAACAA	150	100	100
cgd6_520.P.1935-1984	ATA	50	100	100
cgd6_730.P.3630-3679	AAT	50	100	100
cgd1_3270.P.668-717	TAA	50	100	100
cgd2_700.P.2705-2755	AGA	51	100	100
cgd6_4030.P.2123-2172	CCACCAGCC	50	100	100
cgd2_3870.P.4485-4548	CAAATAATAATAATT/AATAATAAAT/AAT	151	100	85.714
cgd2_680.P.5030-5079	CCACCT	50	100	100
cgd8_4830.P.581-715	ACATCCACATCT/CTACAT	152	100	100
cgd4_3630.P.1196-1247	AGCGG	52	100	100
cgd6_1430.P.373-425	TAA	53	100	100
cgd4_810.P.418-471	ATA	54	100	100
cgd5_4490.P.2933-2987	CCAGAG	55	85.714	85.714
cgd3_270.P.2731-2785	TACTGCTACTAC	55	100	100
cgd8_4510.P.486-541	CAAAACCAAAGC	56	100	100
cgd8_4840.P.6354-6611	GGAGCT	258	100	100
cgd8_2800.P.1088-1245	TAC	158	100	100
cgd4_600.P.751-809	ATA	59	100	100
cgd6_530.P.752-810	TAA	59	100	100
cgd6_830.P.6111-6169	TCT	59	100	100
cgd2_3980.P.632-691	AAT	60	100	100
cgd6_520.P.942-1001	ATA	60	100	100
cgd2_3850.P.340-499	TCCTGC	160	100	100
cgd2_430.P.435-495	CAAGTT	61	100	100
cgd8_4000.P.2102-2134	TCATCATCATCTTCT/TCATCATCTTCT	62	100	100
cgd5_350.P.56-117	ACT	62	100	100
cgd2_410.P.680-744	TCAAGG	65	100	100
cgd7_4990.P.19317-19382	AGAAGAAGGAAGAGA	66	100	100
cgd7_1010.P.10224-10288	GGAGGAATGA/AGG	66	100	100
cgd2_3340.P.137-202	GGAACA	66	100	100
cgd1_470.P.1417-1489	TTCTGA	73	100	100
cgd6_760.P.999-1073	CCAGCT	75	100	100
cgd2_1000.P.1638-1675	ACT/ACTAATACTACT	75	100	100

Table 5.2: Tandem Repeat (TR) target loci for MOI analysis. These loci represent the top 100 scoring tandem repeats generated by running VaNTA on UKP2-8 using *C. parvum* IowaII as reference. The target locus name is formatted as {gene_name}.P.{position}. The TR subseq is the most likely repeat subsequence as reported by Tandem Repeats Finder. Where multiple subsequences are reported, seperated by a forward slash, this represented target regions where there are 2 or more adjacent TR's with discrete repeat subsequences, and have therefore been reported as a single TR. *C. parvum* IowaII TR length refers to the length of the TR region alone (without flanking regions) in the reference genome. Flank conservation is reported as the percentage of the query dataset (UKP2-8) which exhibits complete sequence similarity in that flanking region. Targets highlighted in red are those which bear cumulative flank and target sequence lengths of greater than the mean read length across each dataset, and are therefore unlikely to be fully captured in a single read.

5.2.2 Target Variation Distance Calculation

To investigate target sequence and fragment-length variation within multi-dimensional datasets, Euclidean distance (D^n) was calculated using the following equation:

$$D^n = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}. \tag{5.1}$$

Where $q = (q_1, q_2, ..., q_n)$ and $p = (p_1, p_2, ..., p_n)$ are points in an nth dimensional Euclidean space, which in this context refers to a set of fragment-length/sequence variations at a target locus.

Given a dataset of the number of fragment-length/sequence variations at a target location within a set of single-host read libraries, D^n is calculated as a measure of the value of the target for interrogation to elucidate multiple sub-populations within a single host. The higher the D^n, the greater the overall variation throughout all read libraries, and therefore the more common variation of this target is seen within a host. Figure 5.2 is a graphical representation of D^n in 3 dimensions. In the context of target variation distance, each axis measures the number of alleles present within a single sample at the locus being interrogated.

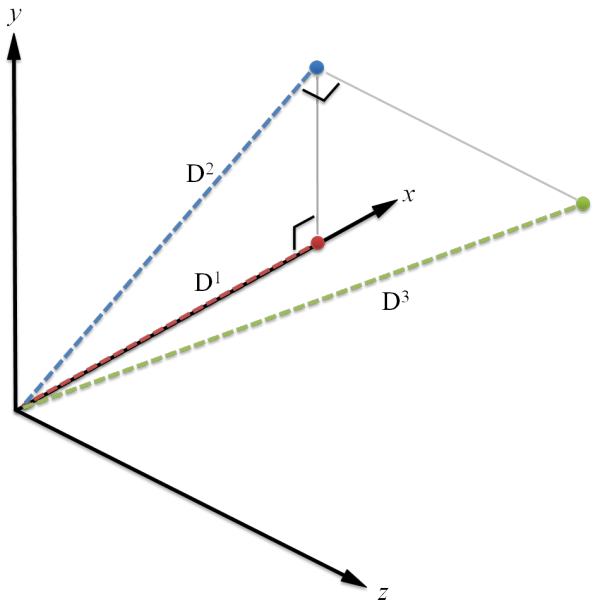


Figure 5.2: A graphical example of the D^n of a 3-dimensional dataset. The red, blue and green lines denote the D^1, D^2, and D^3 distances respectively.

5.2.3 MOI Signature Distance

To examine the efficacy MOI-signature as a high-sensitivity and reliable measure of isolate relatedness, isolates were clustered according to their MOI-signature. If we consider an isolate, I, which is a sample containing an arbitrary number of alleles at each locus. An allele, a, is defined as a variation in a target locus (in this instance, variation is dictated by fragment length alone). The set M_a contains all alleles at the interrogated locus present within I:

$$M_a = \{a_1, a_2, ..., a_n\} \tag{5.2}$$

The allele count ($c^I_{a,i}$) is defined simply as the number of reads which contain allele i within I. A vector of allele probabilities can be generated by calculating the probability ($p^I_{a,i}$) of encountering each allele within I:

$$p^I_a = \{p^I_{a,1}, p^I_{a,2}, ..., p^I_{a,n}\} \tag{5.3}$$

Likewise, the vector of frequencies can be defined as the relative abundance of each allele:

$$f^I_a = \frac{c^I_a}{\sum_{j=1}^{n} c^I_a, j} \Leftrightarrow f^I_{a,i} = \frac{c^I_{a,i}}{N} \tag{5.4}$$

where:

$$N = \sum_{i=1}^{n} c^I_{a,i} \tag{5.5}$$

This vector of frequencies can be considered the MOI-signature for I. In order that subdominant alleles within an MOI-signature are weighted to contribute more strongly to clustering, the calculation of vector frequencies was adjusted to:

$$f^I_a = \sqrt{\frac{c^I_{a,i}}{N}} \tag{5.6}$$

Euclidean distance (see Section 5.2.2) of MOI-signatures was calculated for each isolate, yielding a pairwise distance matrix. This pairwise distance matrix was then graphed to reveal population data structure, and subsequent isolate relatedness. This was compared against a phylogenetic tree of the dominant allele at this locus for each isolate to reveal whether there was concordance between the two methodologies, and whether MOI-signature distance reveals population structure which is lost by classic dominant allele subtyping. Two loci were selected for further analysis in this manner: gp60 (cgd6_1080.P.108-164) and cgd7_440.P.1066-1129, due to their capacity to resolve sub-

types of *C. parvum*.

5.2.4 Co-Occurrence Alleles Within MOI Signatures

To investigate the concordance of alleles within MOI-signatures, the coefficient of Linkage Disequilibrium (L), which is defined as

$$L_{AB} = p_{AB} - p_A p_B, \tag{5.7}$$

of each allele within the *C. parvum* gp60 MOI-signature set was calculated. Where p_A and p_B are the probabilities of encountering alleles A and B within the dataset, and p_{AB} is the probability of the co-incidence of alleles A and B within a single sample within the dataset.

For the purpose of comparing L between different pairs of alleles, values were normalised using the correlation coefficient between allele pairs, as follows:

$$r = \frac{L}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \tag{5.8}$$

Although the coefficient of Linkage Disequilibrium is used here, it is applied only as a statistical measure of the probability of co-incidence of alleles, and does not make any implication about co-occurrence of alleles within a single genome. Likewise it makes no implication of actual physical linkage. However, since the method of statistical analysis is identical and the concepts similar, the co-occurrence of alleles will be hereafter referred to as Linkage Disequilibrium.

5.2.5 Data Management and Visualisation

All data analysis was carried out using the NumPy [Oliphant, 2006, Van Der Walt et al., 2011] and SciPy [Jones et al., 2001], and visualisation in Matplotlib [Hunter et al., 2007] packages in Python v2.7.

5.3 Results

5.3.1 Results of BlooMine Read Mining on Real Data

Figure 5.3 shows the distribution of target loci (shown in Table 5.4) across the genome of *C. parvum*. Also shown are target fragment length variation and sequence variation distances (see Section 5.2.2), and TR-length. In-host TR variability appears to be uniformly distributed, with no obvious clustering of variable target loci across the genome

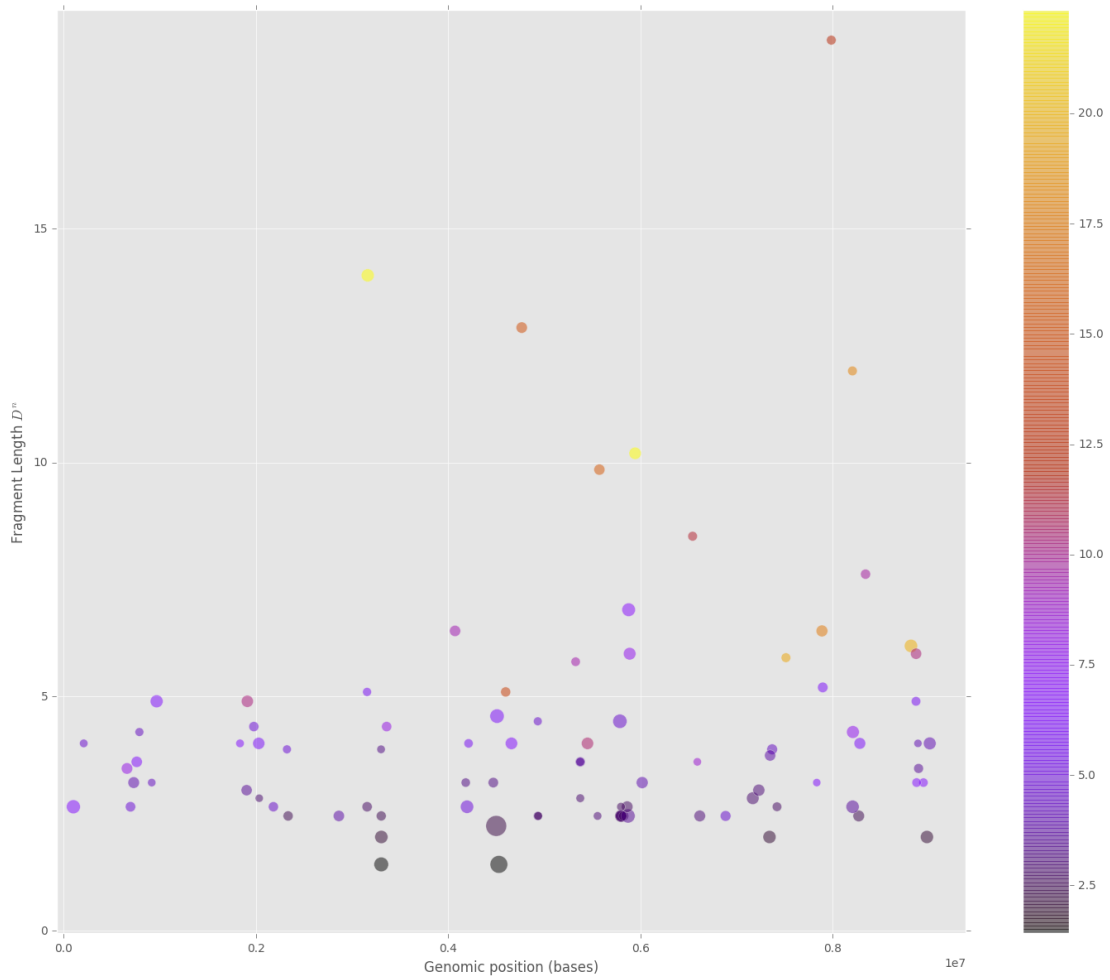


Figure 5.3: Fragment length (y axis) and sequence (marker colour) variation of a TR locus plotted against genomic position (x-axis). The diameter of the markers denotes the length of the Tandem Repeat region within the IowaII reference genome (i.e. the fragment length of the TR within the IowaII genome). Variation is presented as Euclidean distance (D^n) in n-dimensions, where n is the size of the dataset, as detailed in section 5.2.2. Here, a greater D^n indicates a greater capacity for the TR locus to differentiate isolates by either fragment length or sequence variation, as indicated by its variability in the dataset. Target fragment length and sequence variation refer to the the two kinds of variability at a locus, and can both be used to define an allele. Fragment length variation refers to a difference in the length of an allele, measured in nucleotides. Sequence variation refers to a difference in DNA sequence of the allele.

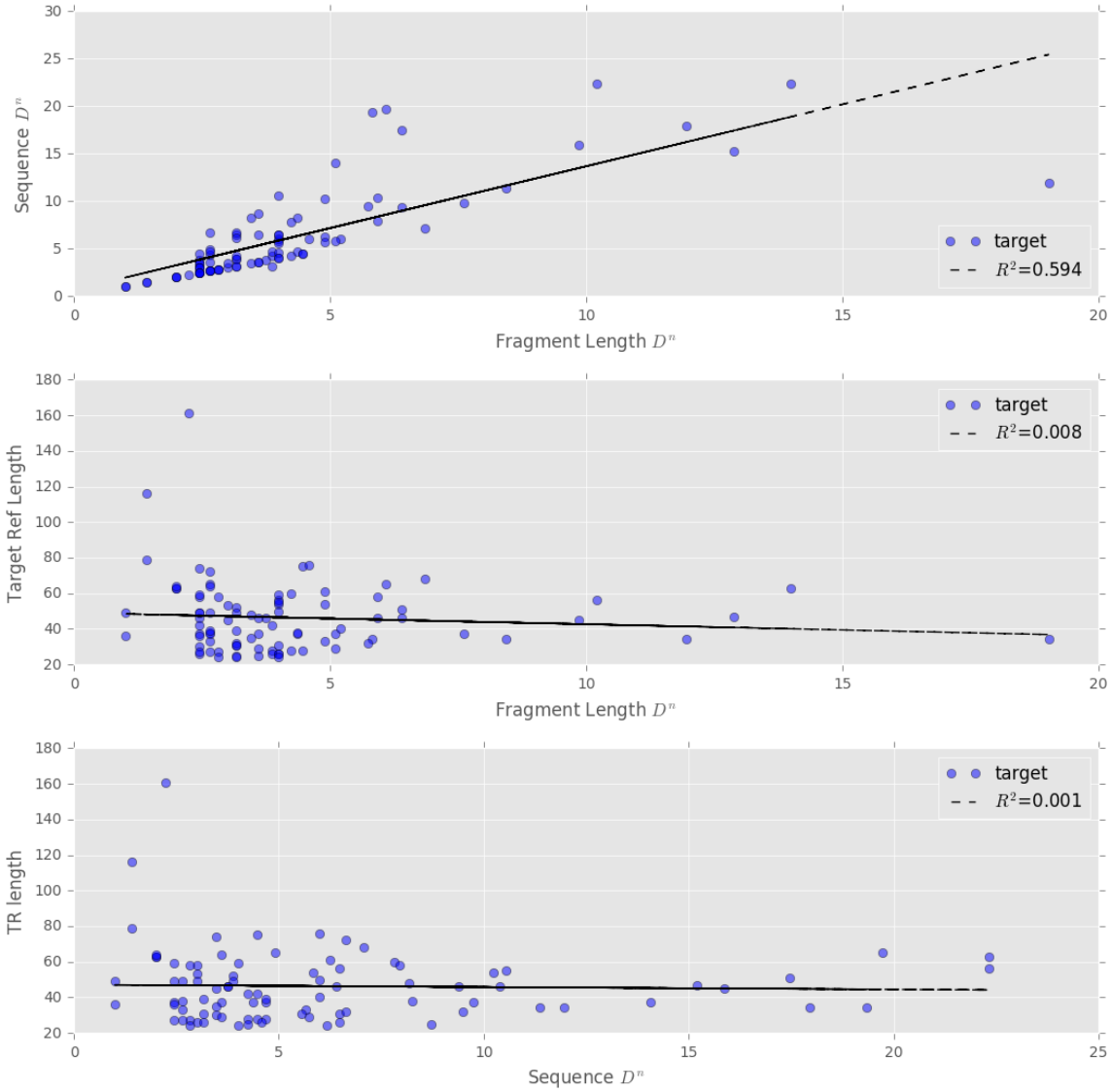


Figure 5.4: Plotted values of reference target length (TR length) and fragment-length/sequence variation (the number of alleles present in a sample defined by fragment-length and sequence variation) for each target. Distance is calculated as detailed in section 5.2.2.

being presented. There appears to be correlation between fragment-length and sequence variation within this dataset (illustrated by markers on the x-axis and marker colour). To further examine the relationships between metrics within this dataset, they were plotted against each other. Figure 5.4 shows scatter plots of target metrics within the dataset of clinical read sets, where each marker within the plot represents the length of the target within a reference genome, or the Euclidian distance of the fragment length or sequence variation of the target within all datasets from 0 (with 0 representing no variation at all), as detailed in section 5.2.2. A moderately strong positive correlation between variation of target TR sequence and fragment-length can be seen, exhibiting an R^2 of 0.594. This indicates that target variation by fragment length and sequence are mutually explicable, in that if there is variation in fragment-length of the target within a sample (indicating multiple populations), there is also likely to be variation at a sequence level. There appears to be no correlation between the length of the target TR within the reference genome and fragment-length ($R^2 = -0.008$) or sequence variation ($R^2 = -0.001$).

Table 5.4 shows results of the number of variable target alleles across the dataset. Not all datasets exhibited all target regions, which is possibly due to lack of coverage over these regions, or loss of the target loci. The .fastq read sets have been seen to have uneven coverage across the genome, represented by GINI values in Table 5.1, where higher values represent higher levels of coverage inequality. Table 5.1 also illustrates that portions of the genome are not covered by reads within the datasets. 76 of the targets exhibited multiple alleles (target variation present as either target length or sequence) within at least one of the paired-end read sets. 58 of these targets exhibited variation within 2 or more read sets, and 4 across all. 10 of the target loci detailed in table 5.2 were not captured in a single read within any of the isolate read sets. These were cgd1_140.P.2350-2463, cgd8_700.P.747-842, cgd5_4090.P.428-544, cgd6_830.P.7342-7445, cgd5_3750.P.497-608, cgd8_1380.P.662-811, cgd8_4830.P.581-715, cgd8_4840.P.6354-6611, cgd8_2800.P.1088-1245, cgd2_3850.P.340-499. Each of these missing target sequences were highlighted in table 5.2 as bearing cumulative flanking and target sequence lengths of significantly greater than the average read length across each isolate of 135.49 nucleotides.

Isolate	Unscreened Reads	Reads Post FP-screen	Reads Post SP-screen I	Reads Post SP-screen II
UKP2	4030426	472	66	2
UKP3	16854282	4866	990	177
UKP4	18685282	1291	1003	198
UKP5	2642800	49	24	11
UKP6	11197064	3658	100	23
UKP7	7493562	650	412	78
UKP8	22322688	6454	903	174

Table 5.3: The number of reads at each stage of screening: Unscreened (present in the concatenated .fastq files), First-Pass (see Section 4.4), and Second-Pass screen I and II (see Section 4.5). See Figure 4.6 for an outline of this process.

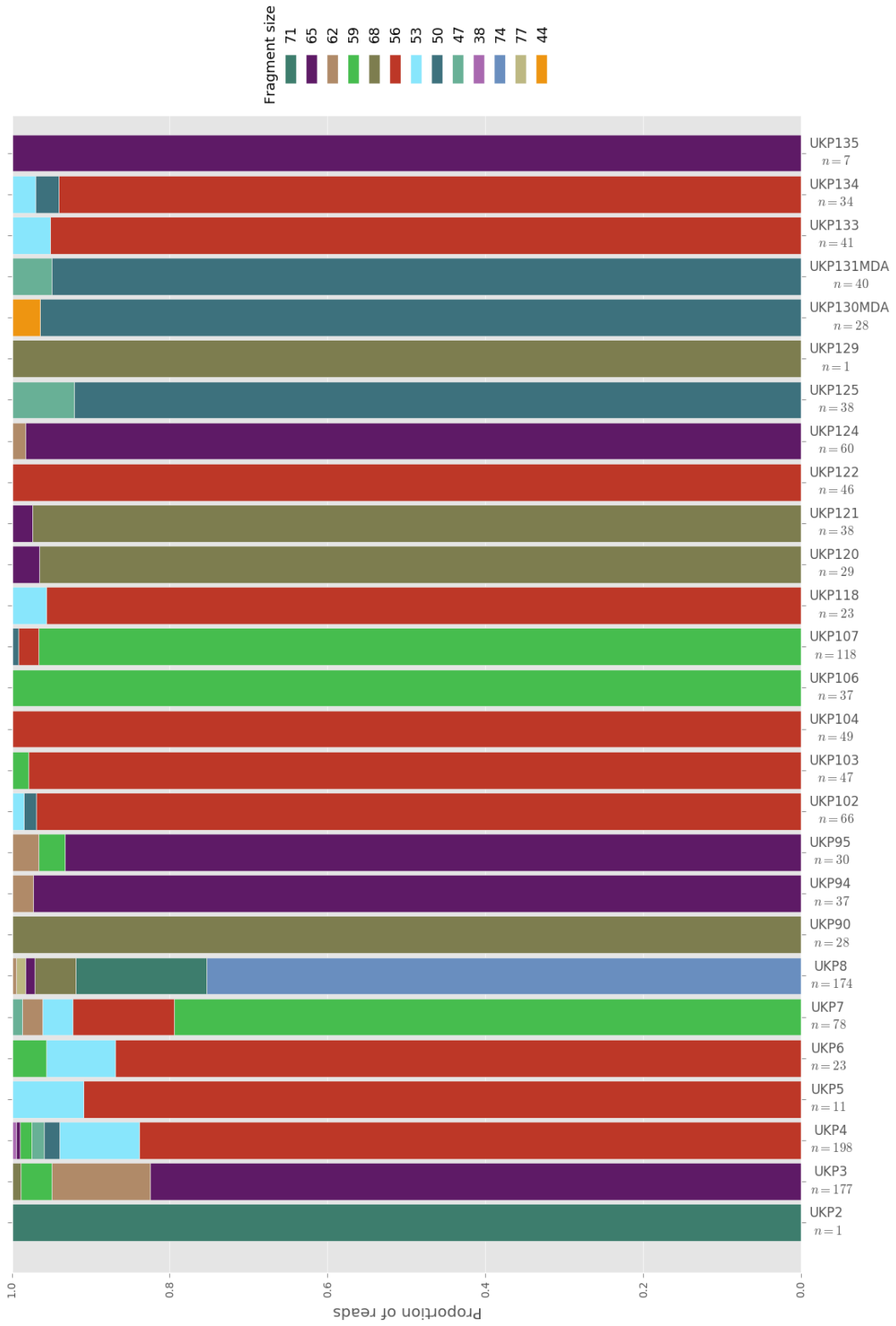


Figure 5.5: The breakdown of fragment lengths (alleles indicating the presence of discrete sub-populations defined by gp60 subtype) at the gp60 locus mined from raw read sets generated from *C. parvum* isolated from clinical samples and the fragment lengths are given in the legend. n refers to the number of reads which fully captured the gp60 region, and are therefore presented in the data.

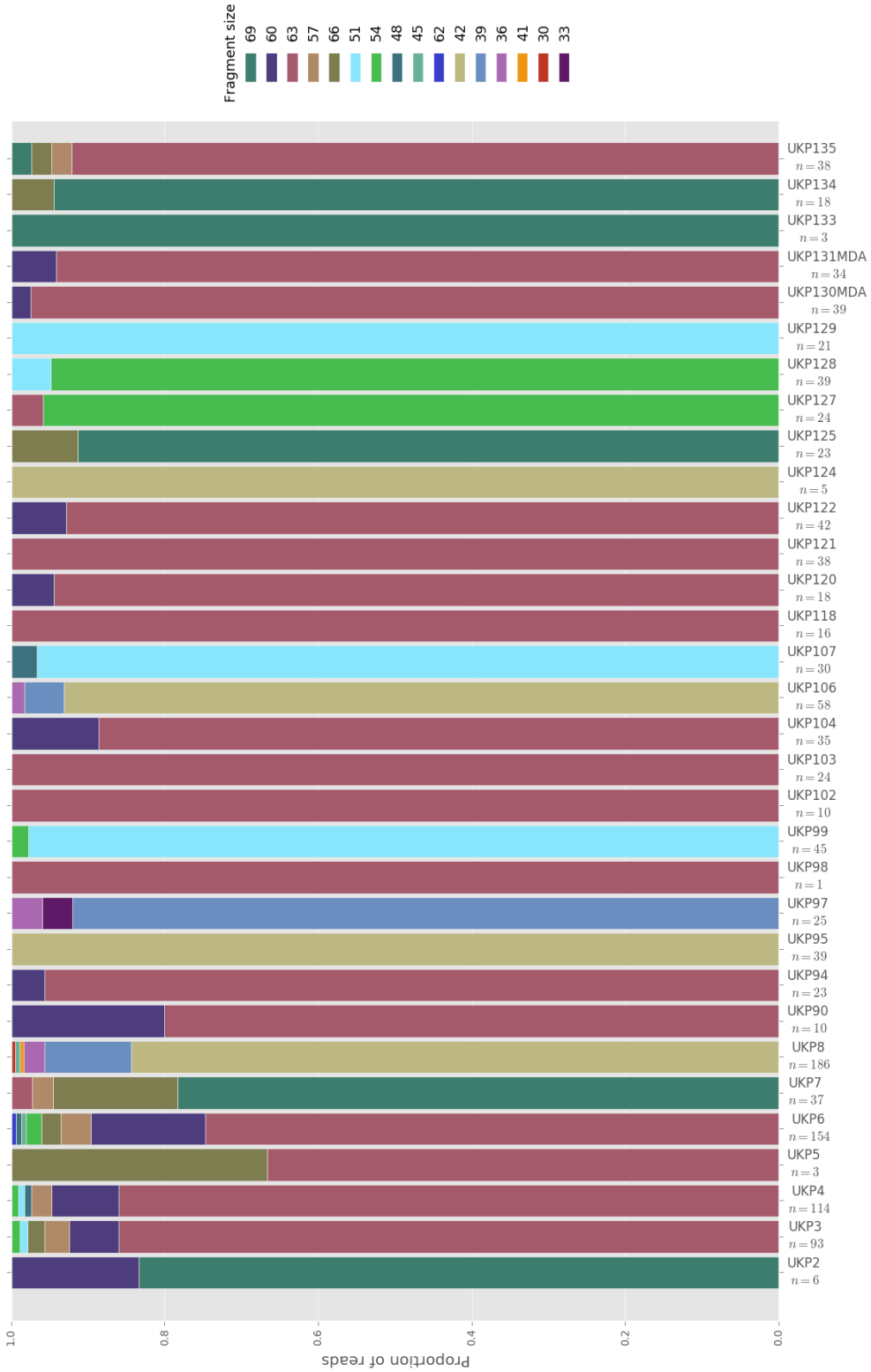


Figure 5.6: The breakdown of fragment lengths (alleles indicating the presence of discrete sub-populations) at the cgd7_440.P.1066-1129 locus mined from raw read sets generated from *C. parvum* isolated from clinical samples and the Fragment lengths are given in the legend. *n* refers to the number of reads which fully captured the local region, and are therefore presented in the data.

Figures 5.5 and 5.6 shows the breakdown for gp60 and cgd7_440.P.1066-1129 fragment lengths within a *C. parvum* dataset respectively. Table 5.3 shows the number of reads which were output from each step of the screening process using the gp60 locus as the target. The most abundant fragment length for each isolate in 5.5 represents the length of the gp60 allele detected by PCR, shown in Table 2. Most isolates exhibit a similar population structure, with the dominant allele accounting for the majority of the alleles present within the sample. Hereafter we use the allele naming convention of "allele #", where # refers to the fragment length of the allele. The results also show that UKP4, 5, 6, 118, 103, 133 and 134 each exhibit very similar gp60 MOI-signatures (defined here as the relative abundance of each allele within the sample), with the same dominant allele (allele 56) and sub-dominant allele (allele 53). Similarly, UKP3, 94, 95, and 124 have similar MOI-signatures, bearing allele 65 as dominant and allele 62 as sub-dominant. Cgd7_440.P.1066-1129 population structure appears to be dominated by allele 63, presenting as the dominant allele of 17 isolates within this dataset. Similar MOI-signatures can be seen in UKP90, 94, 104, 120, 122, 130MDA, and 131MDA, which bear allele 63 as dominant and allele 60 as the only sub-dominant allele. Similarly, further sub-dominant population structure can be seen in UKP3, 4, and 6, which each bear alleles 63 and 60 as dominant and sub-dominant respectively, but also exhibit a number of other alleles.

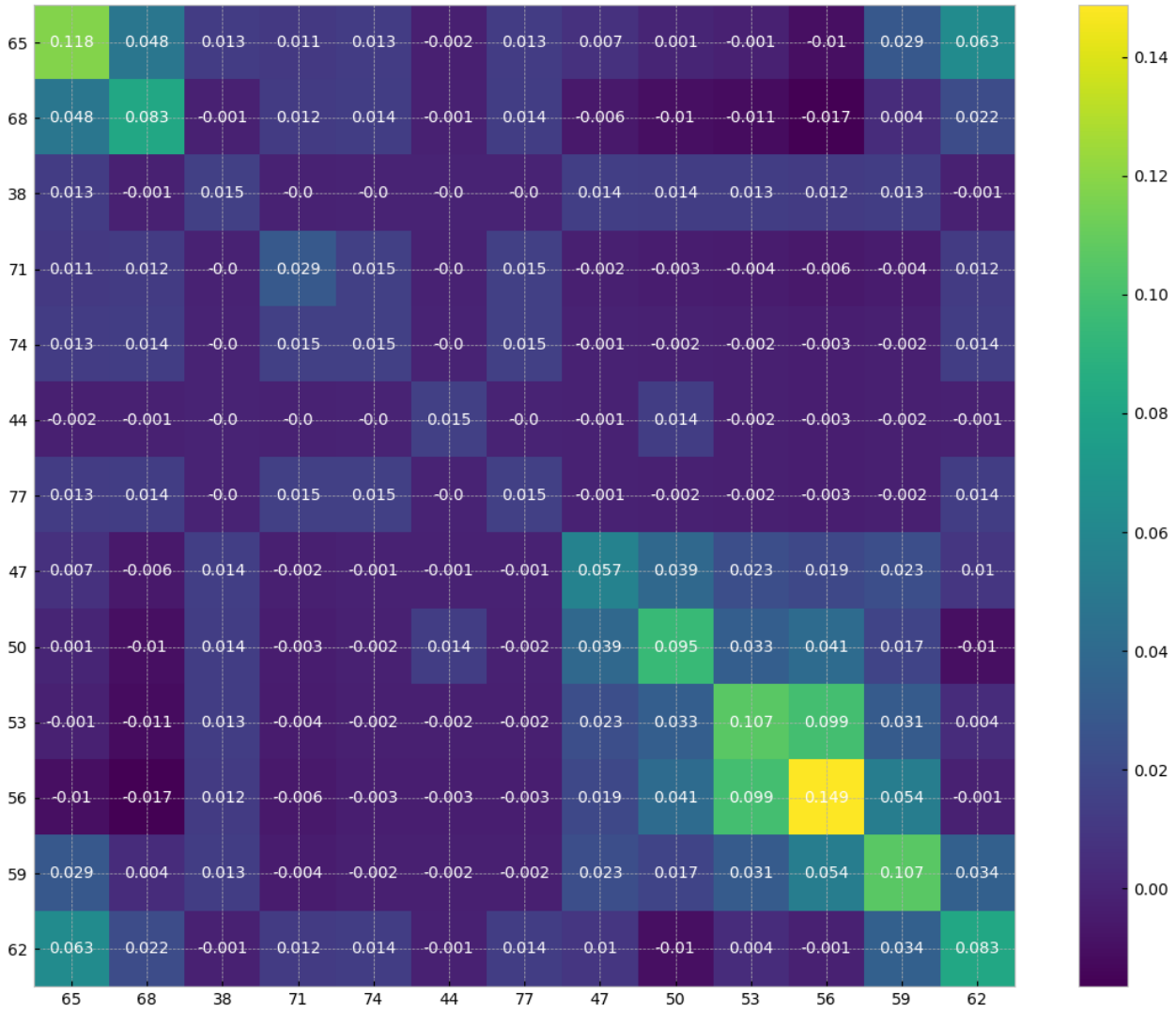


Figure 5.7: A heat map of un-normalised Linkage Disequilibrium (see Section 5.2.4) of gp60 alleles (by fragment size) present within MOI-signatures in the *C. parvum* dataset. Positive values indicate the level of co-occurrence of alleles, and negative values indicate the level of mutual exclusivity of alleles. Both the x and y axis refer to fragment sizes of the gp60 allele.

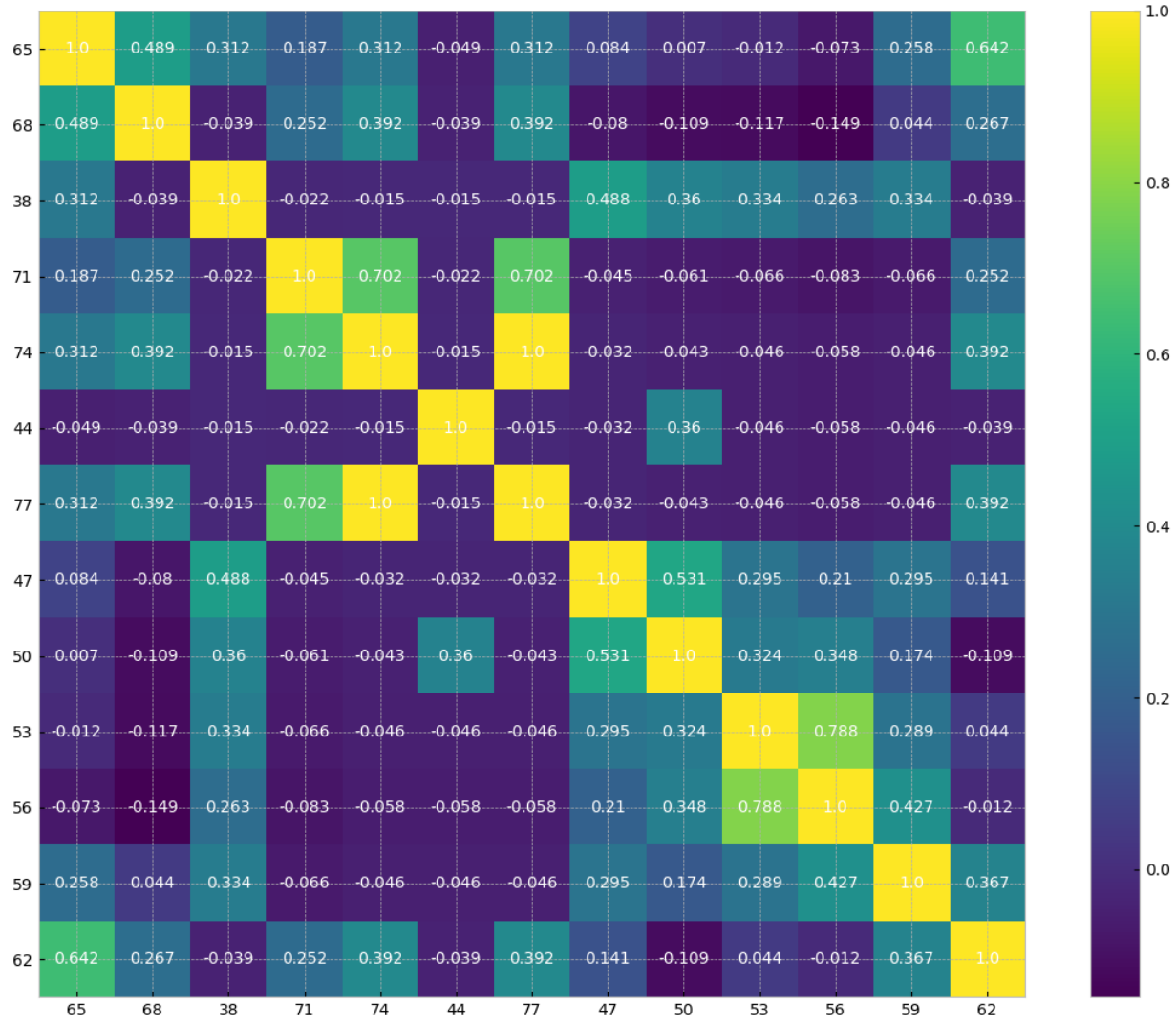


Figure 5.8: A heat map of normalised Linkage Disequilibrium (see Section 5.2.4) of gp60 alleles (by fragment size) present within MOI-signatures in the *C. parvum* dataset. Positive values indicate the level of co-occurrence of alleles, and negative values indicate the level of mutual exclusivity of alleles. Both the x and y axis refer to fragment sizes of the gp60 allele.

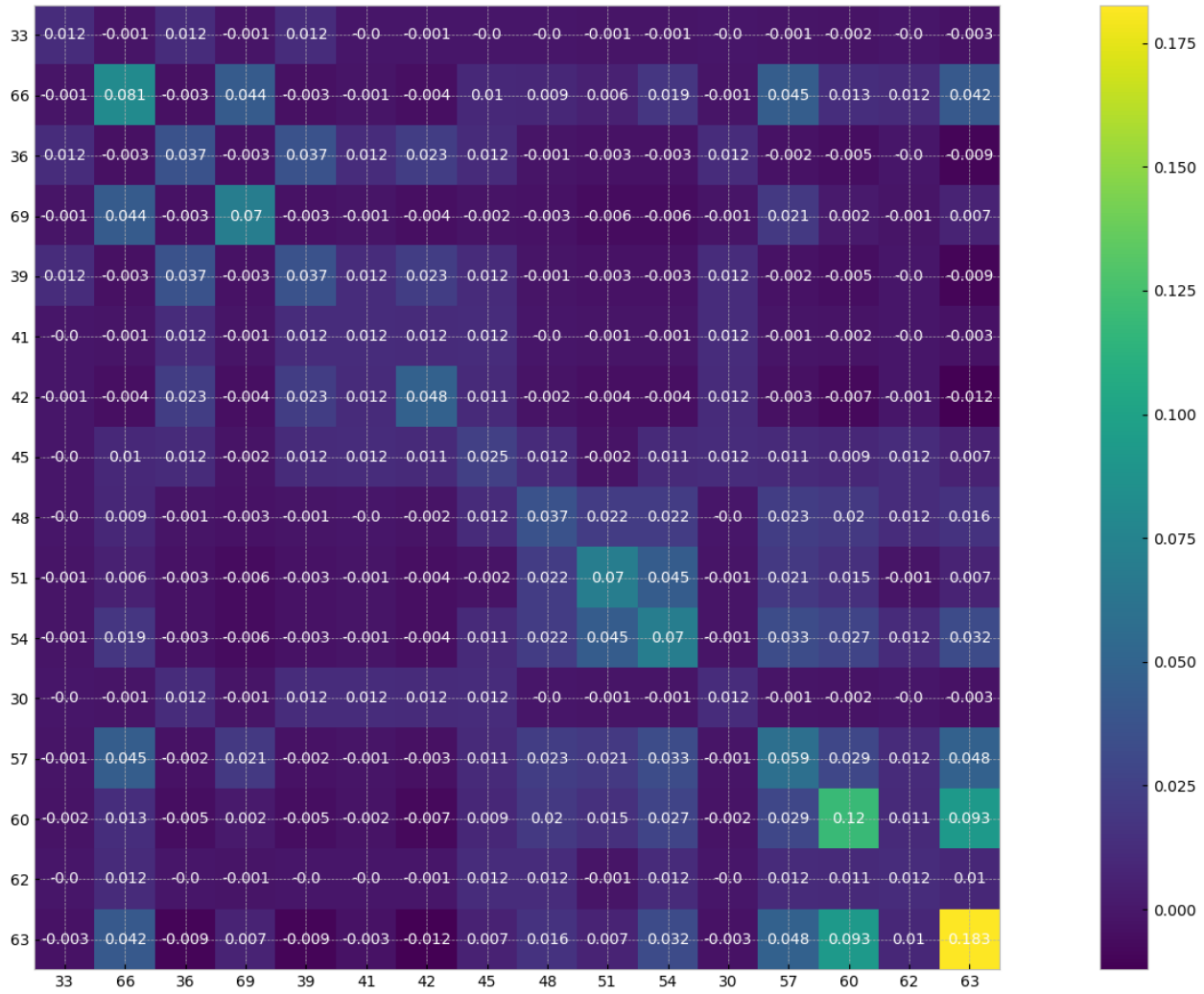


Figure 5.9: A heat map of un-normalised Linkage Disequilibrium (see Section 5.2.4) of cgd7_440.P.1066-1129 alleles (by fragment size) present within MOI-signatures in the *C. parvum* dataset. Positive values indicate the level of co-occurrence of alleles, and negative values indicate the level of mutual exclusivity of alleles. Both the x and y axis refer to fragment sizes of the cgd7_440.P.1066-1129 allele.

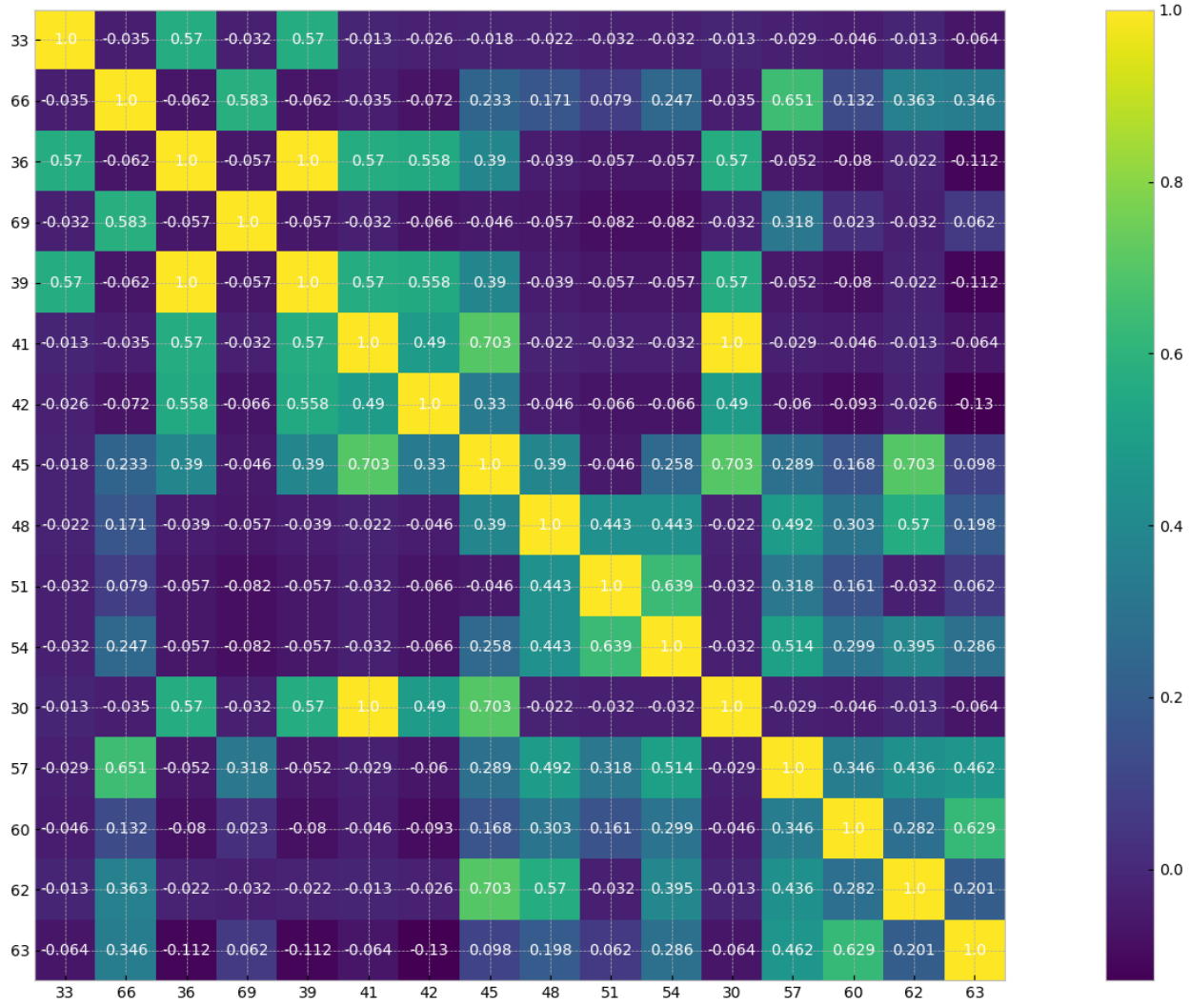


Figure 5.10: A heat map of normalised Linkage Disequilibrium (see Section 5.2.4) of cgd7_440.P.1066-1129 alleles (by fragment size) present within MOI-signatures in the *C. parvum* dataset. Positive values indicate the level of co-occurrence of alleles, and negative values indicate the level of mutual exclusivity of alleles. Both the x and y axis refer to fragment sizes of the cgd7_440.P.1066-1129 allele.

To investigate the co-incidence of fragment sizes (alleles) within these isolates, the Linkage Disequilibrium (LD) of allele pairs was calculated (see Section 5.2.4 Equation 5.7) for both loci (gp60 and cgd7_440.P.1066-1129). To clarify these data, the results were normalised using the correlation coefficient (see Section 5.2.4 Equation 5.8)

Figure 5.7 shows the result of Linkage Disequlibrium calculations on gp60 allele pairs. These results suggest non-random association of gp60 allele pairs within this dataset. The results of normalisation can be seen in Figure 5.8. This shows that there is significant positive LD between a number of read pairs, particularly alleles 71-74 (.702), 77-71 (.702), 77-74 (1.000), 53-56 (.788), and 62-65 (.642). Little negative LD was apparent, however, allele 68 accounted for the majority of the negative LD presented within this dataset. This suggests that allele 68 is strongly dominant within a population, and presents with very few or no sub-populations. This is also indicated by results seen in Figure 5.5.

Figures 5.9 and 5.10 show the raw and normalise LD values for each allele pair for cgd7_440.P.1066-1129. These results show high levels of non-random association, with pairs of alleles presenting with positive LD, such as allele pairs 30-41 (1.000), 57-66 (0.651), 30-45 (0.703), 45-62 (0.703), 45-41 (0.703), and 60-63 (0.629). In particular, allele 45 shows positive LD with 12 of the 15 alleles (0.096 to 0.703) and only slight negative LD with the remaining 3 (-0.018 to -0.046). Allele 63 show the highest rate of negative LD, exhibited across 5 allele pairing (-0.064 to -0.13) though it also exhibits moderately high levels of positive LD in the remaining allele pairs.

For both of these loci, structure in the data, such as the clear positive LD 'boxes' which can be seen at cgd7_440.P.1066-1129 alleles 39-45 and 57-63, and gp60 alleles 47-59 indicate that there may be higher levels of LD between allele pairs which are closer in fragment length.

Figure 5.11 shows the Euclidean Distance (see Section 5.2.2) for MOI-signatures generated from interrogation of the gp60 locus within the clinical *C. parvum* dataset. The data suggests clustering is occurring according to MOI-signature. To further explore this, a nearest neighbour tree was generated from the MOI-signature pairwise distance matrix. Figure 5.12 shows this dataset clusters into 4 distinct clades (green, red, azure, and purple) and 1 less distinct clade (yellow). Green, red, azure, and purple clades consist of isolate with dominant populations of alleles 56, 68, 65, and 50. Each of these major clades pertain to distinct gp60 fragment lengths, and are therefore equivalent to gp60 subtypes. Further discrimination is a result of the abundance of sub-populations of alleles within each isolate. Large amounts of structure has been resolved at a sub-dominant level. The yellow clade appears to be much more diverse, containing 3 dominant alleles (71, 59, and

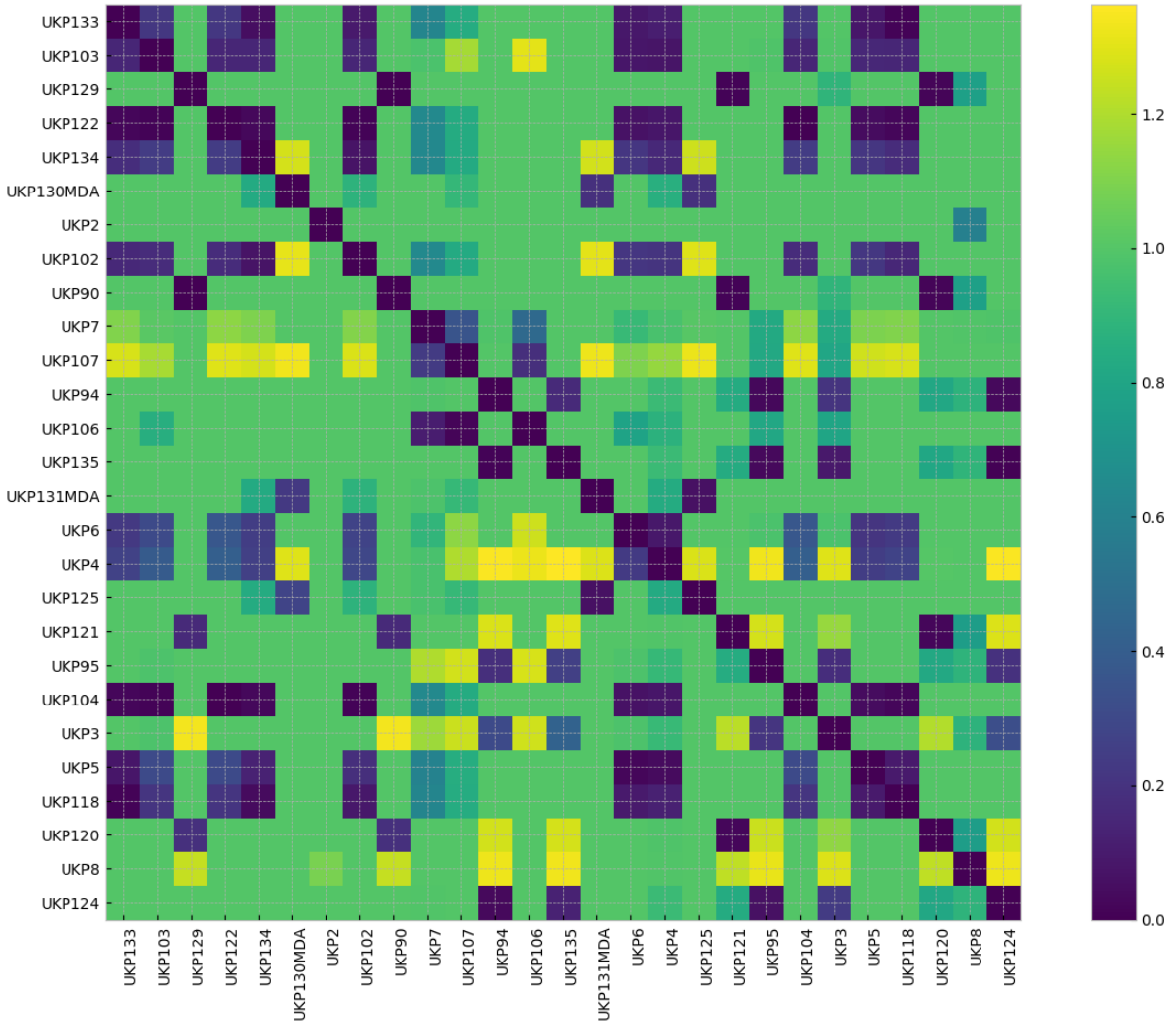


Figure 5.11: A heat map of the Euclidean Distance (see Section 5.2.2) of MOI-signatures of isolates within the *C. parvum dataset* (detailed in Section 5.2.3) at the gp60 locus. Both the x and y axis refer to *C. parvum* isolates.

53), and is consequently rooted more distantly from the leaves.

Figure 5.15 is a neighbour joining tree generated by multiple alignment of gp60 sequences from the *C. parvum* dataset. It indicates the presence of 11 discrete gp60 subtypes within this dataset, with the largest (IIaA15G2R1) accounting for 10 of the 26 isolates. UKP129 appears to be significantly distinct from the other isolates within this dataset.

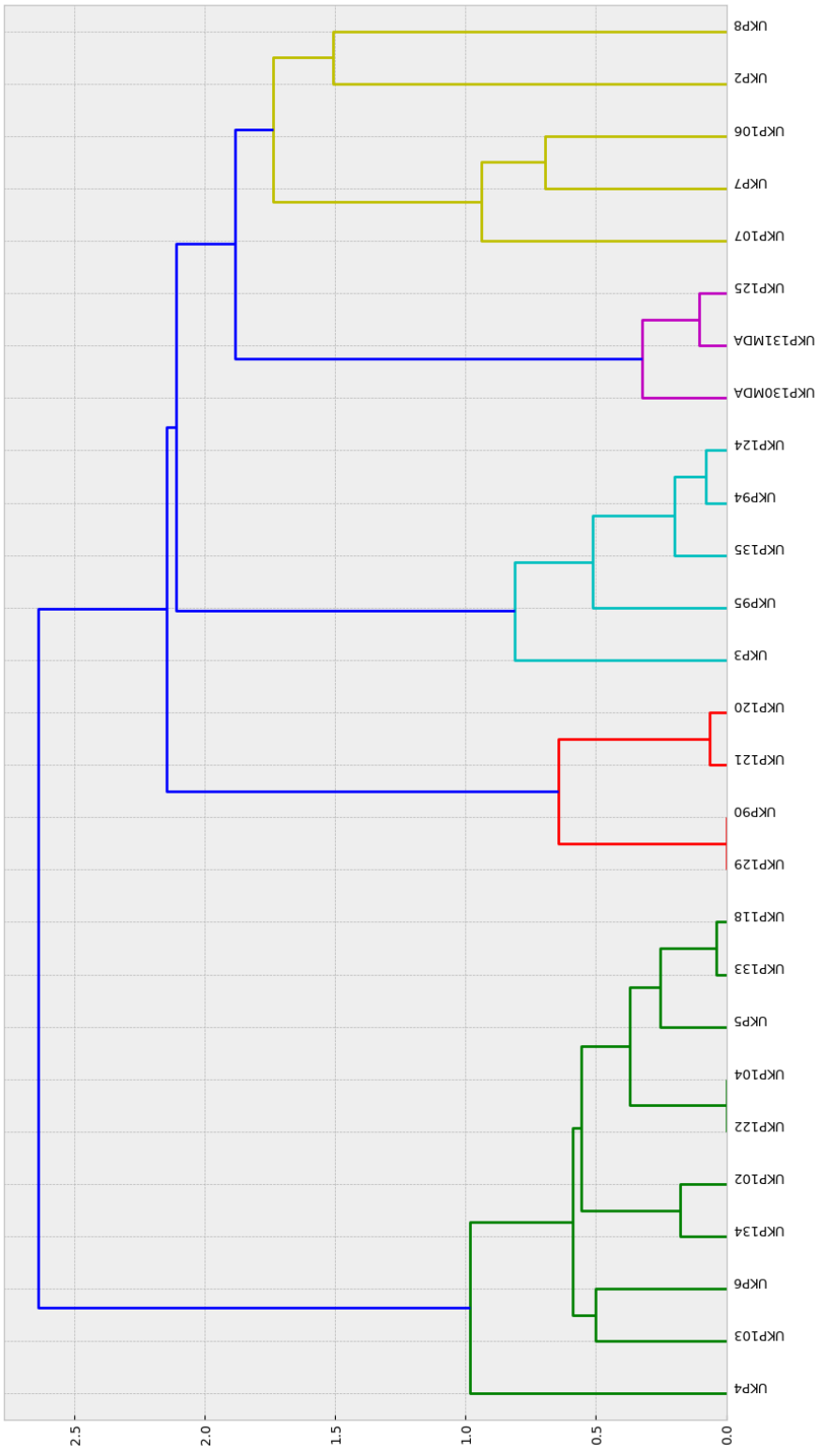


Figure 5.12: A neighbour joining tree based on the pairwise Euclidean Distance matrix (see Section 5.2.2) of MOI-signatures of isolates within the *C. parvum dataset* (detailed in Section 5.2.3) at the gp60 locus.

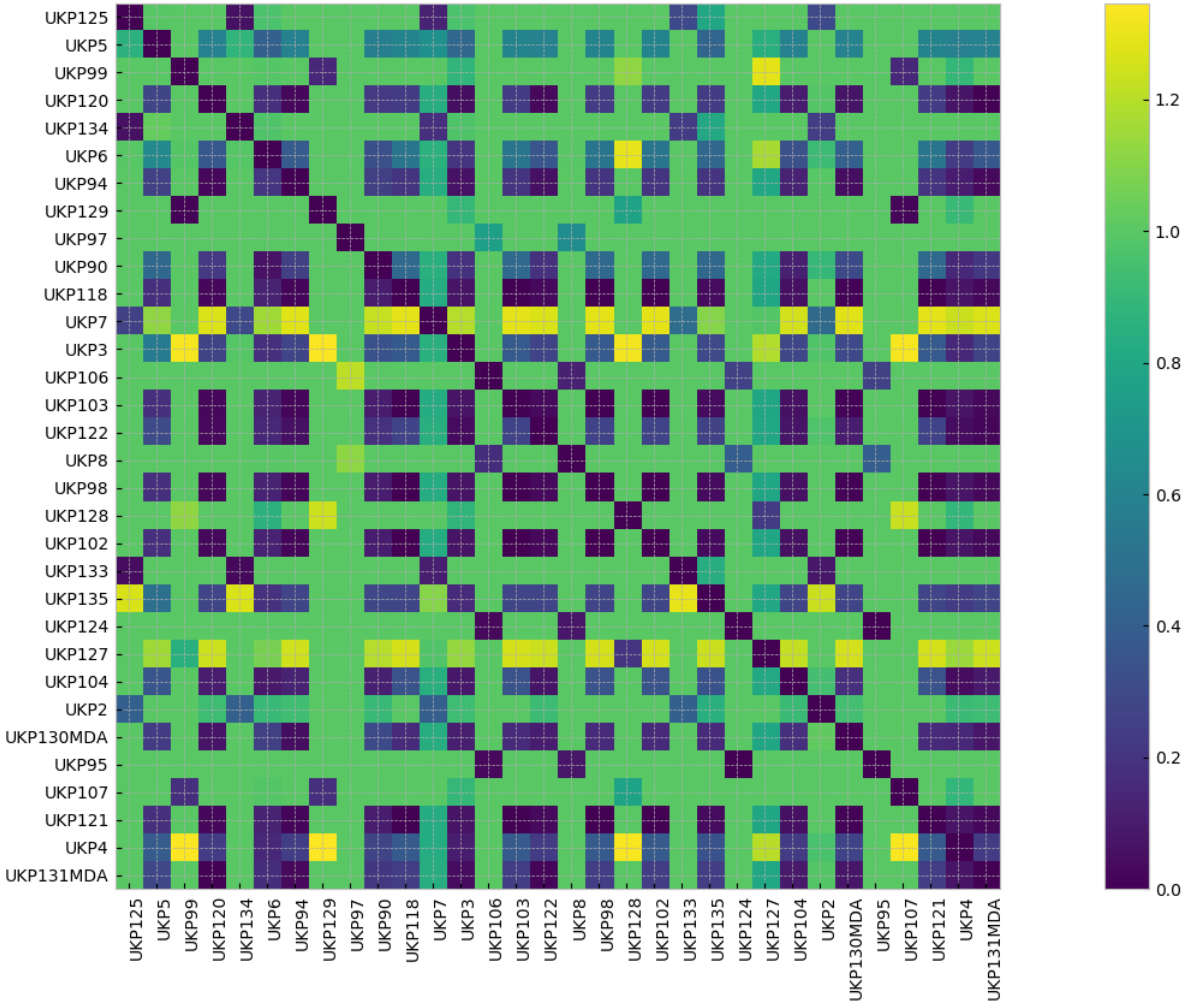


Figure 5.13: A heat map of the Euclidean Distance (see Section 5.2.2) of MOI-signatures of isolates within the *C. parvum dataset* (detailed in Section 5.2.3) at the cgd7_440.P.1066-1129 locus.

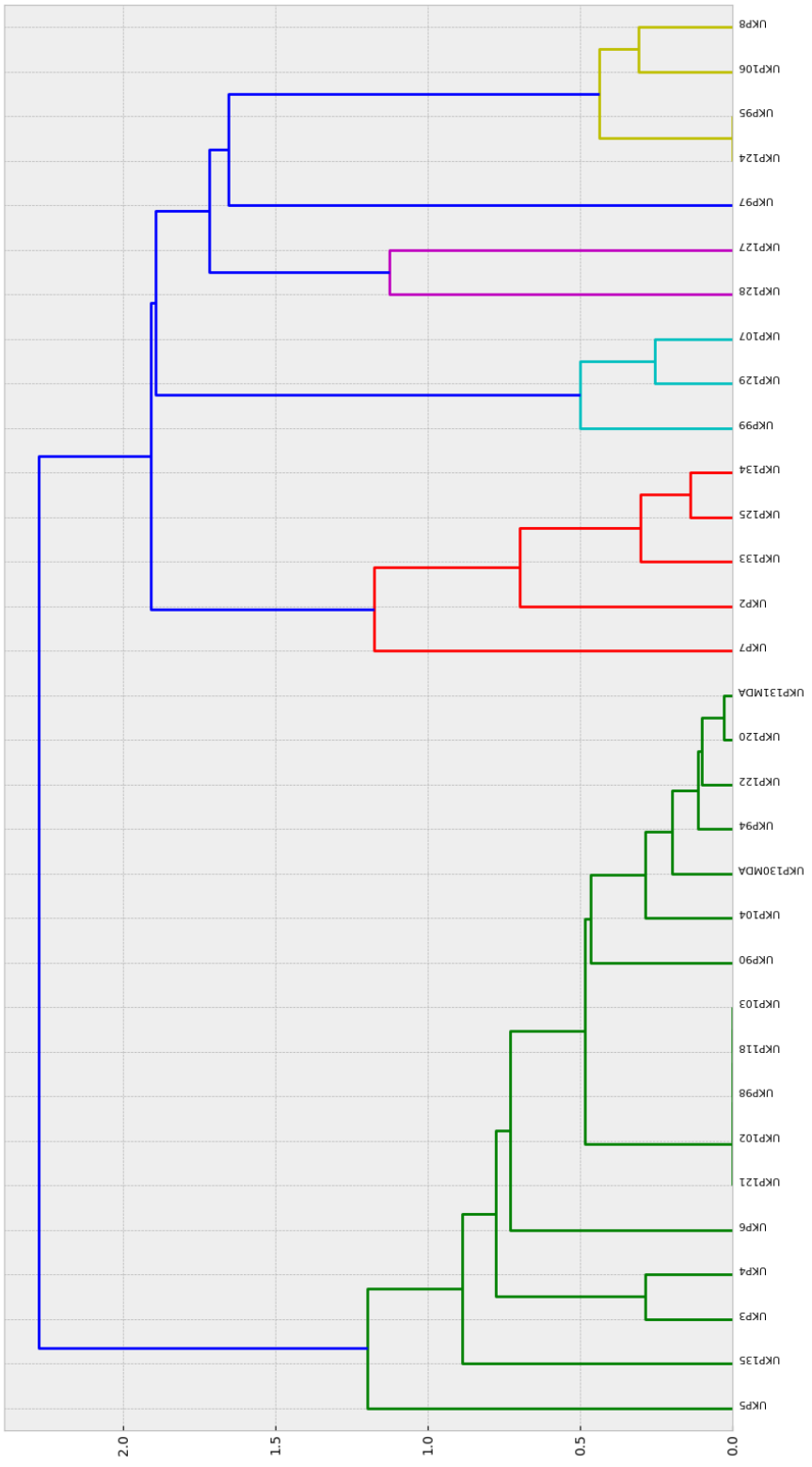


Figure 5.14: A neighbour joining tree based on the pairwise Euclidean Distance matrix (see Section 5.2.2) of MOI-signatures of isolates within the *C. parvum dataset* (detailed in Section 5.2.3) at the cgd7_440.P.1066-1129 locus.

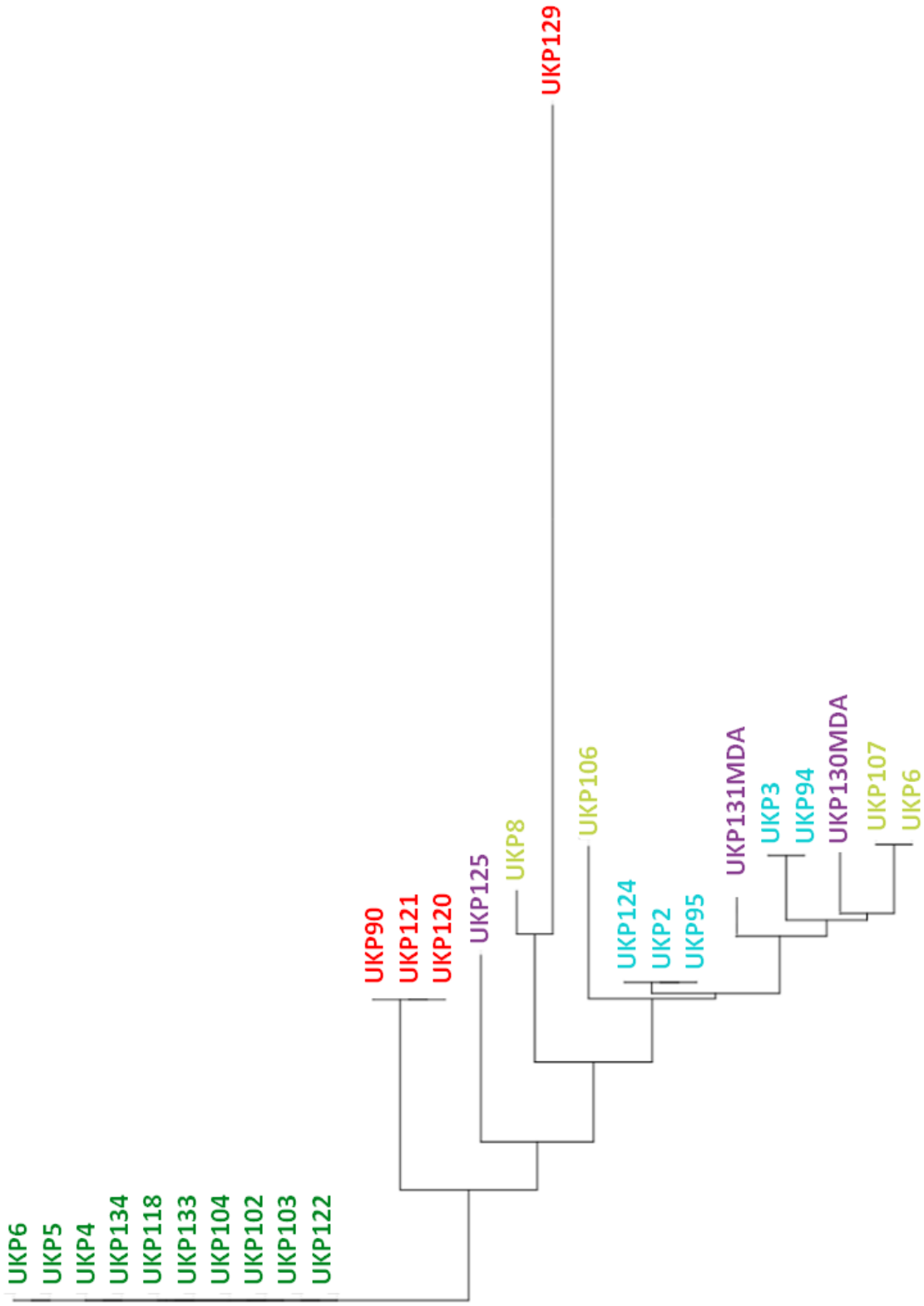


Figure 5.15: A neighbour joining tree generated from multiple alignment of the gp60 locus using ClustalW. Leaf labels are colourised in accordance with their placement in the tree seen in Figure 5.12.

Target Locus	UKP2	UKP3	UKP4	UKP5	UKP6	UKP7	UKP8	length D^n	sequence D^n
cgd4_3970.P.1521-1555	2 1	0 0	10 6	2 1	6 4	7 5	13 8	19.026	11.958
cgd7_440.P.1066-1129	2 2	6 8	6 11	2 2	8 13	4 4	6 11	14	22.338
cgd8_1570.P.575-622	3 3	4 4	6 7	2 2	7 9	6 6	4 6	12.884	15.199
cgd2_400.P.329-363	1 2	0 0	7 10	2 4	3 3	4 7	8 12	11.958	17.944
cgd6_1080.P.108-164	1 1	4 15	7 14	2 2	3 6	5 6	0 0	10.198	22.316
cgd8_4940.P.1528-1573	3 3	4 7	3 10	2 2	3 4	5 5	5 7	9.849	15.875
cgd6_3690.P.595-629	2 2	3 3	4 7	1 1	3 5	4 4	4 5	8.426	11.358
cgd2_1000.P.1638-1675	2 2	4 5	4 6	1 1	4 4	2 2	1 3	7.616	9.747
cgd6_780.P.1043-1111	0 0	1 1	4 4	1 2	4 4	2 2	3 3	6.856	7.071
cgd4_3630.P.1196-1247	1 1	2 8	3 13	1 1	1 3	3 5	4 6	6.403	17.464
cgd7_4500.P.633-679	1 2	3 5	2 5	1 1	1 1	3 4	4 4	6.403	9.381
cgd2_3340.P.137-202	1 1	3 9	3 9	1 1	2 2	2 5	3 14	6.083	19.723
cgd6_830.P.6111-6169	2 2	2 2	3 5	2 2	3 4	1 1	2 3	5.916	7.937
cgd2_3540.P.5762-5808	2 2	2 3	3 5	1 1	2 4	2 2	3 7	5.916	10.392
cgd4_2090.P.192-226	0 0	2 8	2 11	2 2	3 10	2 6	3 7	5.831	19.339
cgd8_4000.P.2102-2134	1 1	1 3	3 5	2 3	4 6	1 1	1 3	5.745	9.487
cgd4_3660.P.167-207	0 0	2 2	2 2	1 1	4 5	1 1	1 1	5.196	6
cgd8_970.P.3734-3771	1 1	3 9	3 8	1 1	1 1	1 1	2 7	5.099	14.071
cgd7_420.P.3445-3474	2 2	2 2	1 1	0 0	3 4	2 2	2 2	5.099	5.745
cgd5_350.P.56-117	1 1	1 2	3 4	1 1	2 3	2 2	2 2	4.899	6.245
cgd2_3540.P.4474-4507	1 1	2 3	1 2	2 2	3 3	2 2	1 1	4.899	5.657
cgd5_4490.P.2933-2987	1 1	0 0	2 7	1 1	1 2	1 1	4 7	4.899	10.247
cgd8_680.P.2962-3038	2 2	1 1	1 2	1 1	1 1	2 4	3 3	4.583	6
cgd6_490.P.468-543	1 1	1 1	2 2	2 2	3 3	1 1	0 0	4.472	4.472
cgd8_2250.P.3218-3246	1 1	1 1	1 1	0 0	4 4	0 0	1 1	4.472	4.472
cgd3_90.P.1112-1149	1 1	2 2	2 2	1 2	2 2	2 2	1 1	4.359	4.69
cgd7_1190.P.2275-2313	1 1	2 3	1 3	2 2	1 2	2 5	2 4	4.359	8.246
cgd2_430.P.435-495	1 1	1 6	3 3	1 1	1 2	1 1	2 3	4.243	7.81
cgd1_3550.P.1267-1295	2 2	2 2	2 2	0 0	1 1	1 1	2 2	4.243	4.243
cgd3_270.P.2731-2785	1 2	2 2	2 4	1 1	1 1	1 2	2 2	4	5.831
cgd5_4190.P.231-257	2 3	1 3	1 3	1 1	2 2	1 1	2 3	4	6.481
cgd2_700.P.2705-2755	1 2	1 2	2 3	1 1	2 1	1 1	2 4	4	6
cgd8_1220.P.825-881	1 1	2 2	1 3	1 2	1 2	2 2	2 4	4	6.481
cgd2_3980.P.632-691	1 1	2 2	1 1	2 2	2 2	1 1	1 1	4	4
cgd7_5010.P.3301-3332	0 0	1 1	2 4	0 0	1 1	3 3	1 2	4	5.568
cgd1_800.P.182-208	1 1	1 1	2 3	1 1	1 1	2 2	2 2	4	4.583
cgd8_4510.P.486-541	1 2	2 6	2 6	1 1	1 4	2 3	1 3	4	10.536
cgd2_3590.P.5378-5402	1 1	1 1	1 1	1 1	2 2	2 2	2 2	4	4
cgd4_1420.P.405-447	1 1	1 1	1 2	1 1	3 3	1 1	1 1	3.873	4.243
cgd3_1250.P.634-662	1 1	1 1	1 1	1 1	3 4	1 1	1 1	3.873	4.69
cgd7_1010.P.4475-4501	1 1	1 1	1 1	1 1	1 1	1 1	3 2	3.873	3.162
cgd4_1360.P.1109-1155	1 1	0 0	2 2	2 2	2 2	1 1	0 0	3.742	3.742
cgd8_4170.P.2518-2555	2 2	1 1	2 2	1 1	1 1	1 1	1 1	3.606	3.606
cgd6_3930.P.1748-1773	1 1	2 6	2 4	1 1	1 3	1 2	1 3	3.606	8.718
cgd8_4170.P.1991-2020	2 2	1 1	1 1	1 1	2 2	1 1	1 1	3.606	3.606
cgd1_3450.P.494-540	1 1	2 3	1 4	1 1	1 2	1 1	2 3	3.606	6.403
cgd1_3060.P.589-637	2 3	0 0	2 6	1 1	1 2	1 1	1 4	3.464	8.185
cgd2_3590.P.18288-18323	1 1	2 2	2 2	0 0	1 1	1 1	1 1	3.464	3.464
cgd8_550.P.5068-5107	2 2	1 1	1 1	1 1	1 1	1 1	1 1	3.162	3.162
cgd6_1430.P.373-425	1 1	1 1	1 1	2 3	1 1	1 1	1 1	3.162	3.873
cgd1_3270.P.668-717	0 0	1 1	1 1	2 2	2 3	0 0	0 0	3.162	3.873
cgd2_3700.P.4257-4288	1 1	2 5	1 1	1 1	1 2	1 1	1 3	3.162	6.481
cgd2_3550.P.1472-1504	1 1	1 3	1 2	1 1	1 2	1 3	2 4	3.162	6.633
cgd7_4980.P.5482-5513	1 1	1 1	1 1	1 1	1 1	1 1	2 2	3.162	3.162
cgd5_130.P.4193-4218	1 1	1 1	1 1	1 1	1 3	1 1	2 2	3.162	4.243
cgd4_3450.P.4333-4357	1 1	1 3	2 4	1 1	1 1	1 1	1 3	3.162	6.164
cgd4_810.P.418-471	1 1	1 1	1 1	0 0	1 1	2 2	1 1	3	3
cgd5_4480.P.1455-1500	1 1	1 1	1 1	0 0	1 1	2 2	1 2	3	3.464

Continued on next page

Table 5.4 – continued from previous page

Target Locus	UKP2	UKP3	UKP4	UKP5	UKP6	UKP7	UKP8	length D^n	sequence D^n
cgd4_600.P.751-809	0 0	1 1	1 1	1 1	2 2	1 1	0 0	2.828	2.828
cgd8_4170.P.2262-2289	1 1	1 1	1 1	2 2	1 1	0 0	0 0	2.828	2.828
cgd3_280.P.3887-3911	1 1	1 1	1 1	0 0	2 2	0 0	1 1	2.828	2.828
cgd7_420.P.4749-4787	1 1	1 1	1 1	1 1	1 1	1 1	1 1	2.646	2.646
cgd4_1610.P.322-355	1 1	1 1	1 1	1 1	1 1	1 1	1 1	2.646	2.646
cgd3_720.P.16227-16264	1 1	1 2	1 1	1 1	1 2	1 2	1 2	2.646	4.359
cgd7_4990.P.19317-19382	1 1	1 2	1 2	1 1	1 3	1 1	1 2	2.646	4.899
cgd6_730.P.3630-3679	1 1	1 1	1 1	1 1	1 1	1 1	1 1	2.646	2.646
cgd1_470.P.1417-1489	1 2	1 4	1 3	1 1	1 2	1 3	1 1	2.646	6.633
cgd1_3170.P.4177-4216	1 1	1 1	1 1	1 1	1 4	1 1	1 1	2.646	4.69
cgd6_520.P.300-327	1 1	1 1	1 1	1 1	1 1	1 1	1 1	2.646	2.646
cgd2_410.P.680-744	1 1	1 2	1 1	1 1	1 1	1 2	1 1	2.646	3.606
cgd3_3620.P.935-981	1 1	1 1	1 1	0 0	1 3	1 1	1 1	2.449	3.742
cgd6_530.P.752-810	1 1	1 1	1 2	0 0	1 1	1 1	1 1	2.449	3
cgd3_1330.P.464-501	1 1	0 0	0 0	0 0	2 2	0 0	1 1	2.449	2.449
cgd6_520.P.942-1001	0 0	1 1	1 1	1 1	1 1	1 1	1 1	2.449	2.449
cgd6_760.P.999-1073	1 2	1 2	1 1	0 0	1 1	1 1	1 1	2.449	3.464
cgd8_2250.P.3970-4000	0 0	1 2	1 1	1 1	1 2	1 1	1 1	2.449	3.464
cgd8_4860.P.506-532	1 1	1 1	1 1	0 0	1 2	1 1	1 1	2.449	3
cgd7_1010.P.8072-8108	1 1	1 1	1 1	1 1	1 1	0 0	1 1	2.449	2.449
cgd6_5110.P.5014-5056	1 1	1 1	1 1	1 2	1 3	0 0	1 2	2.449	4.472
cgd6_520.P.1935-1984	1 1	1 1	1 1	1 1	1 2	1 1	0 0	2.449	3
cgd2_680.P.5030-5079	0 0	0 0	2 2	0 0	0 0	1 1	1 1	2.449	2.449
cgd8_2260.P.2694-2721	1 1	1 1	1 1	0 0	1 1	1 1	1 1	2.449	2.449
cgd8_660.P.1554-1715	0 0	1 1	0 0	0 0	0 0	2 2	0 0	2.236	2.236
cgd2_3870.P.4485-4548	1 1	1 1	0 0	0 0	0 0	1 1	1 1	2	2
cgd4_1340.P.815-878	1 1	0 0	1 1	1 1	1 1	0 0	0 0	2	2
cgd7_1010.P.10224-10288	1 1	1 1	0 0	0 0	1 1	0 0	1 1	2	2
cgd8_700.P.5284-5400	0 0	1 1	0 0	0 0	0 0	0 0	1 1	1.414	1.414
cgd7_1010.P.8357-8436	0 0	0 0	1 1	0 0	1 1	0 0	0 0	1.414	1.414
cgd6_610.P.1783-1819	1 1	0 0	0 0	0 0	0 0	0 0	0 0	1	1
cgd6_4030.P.2123-2172	0 0	0 0	0 0	0 0	0 0	1 1	0 0	1	1

Table 5.4: Results from BlooMine mining on 100 target regions around the genome of *C. parvum*, across the Hadfield *et al.* dataset ordered by length D^n. The 90 target loci were fully captured in a single read within at least one of the datasets. The target locus name is formatted as {gene_name}.P.{position}. Variation count cells are structured as: `target_length_variation target_sequence_variation`. Distance (D^n) is calculated as described in Section 5.2.2.

5.4 Discussion

The results presented in Section 5.3.1 indicate extensive within-host genetic diversity of *C. parvum* across multiple loci. These results are consistent with those which have been reported in experimental studies. In 2013 Grinberg *et al.* reported high levels of within-host diversity amongst two clinical isolates of *C. parvum*, isolated from cases of Cryptosporidiosis in New Zeland, across two known variable loci (HSP70 and gp60). Their results indicated extreme within-host diversity at the the gp60 locus, demonstrating the presence of 10 distinct alleles within the two isolates. This is supported by our results, which highlights gp60 (cgd6_1080.P.108-164) as bearing high levels of within-host diversity. As can be seen in table 5.4, the gp60 locus bears the fifth highest D^n by fragment length variation, at 10.198. Troell *et al.* reported variation between single sequenced

oocysts isolated from a single clinical sample, as well as variation in sporozoites within single oocysts. The results from both of these studies indicated a consistent population structure within a host, where a single target variant (allele) is seen to be dominant, representing the vast majority of reads within the read set. This dominant variant is present along with 1 or more subdominant variants. An example of this can be seen in Figure 4.10, where 82.5% (146 of 177) of the reads are represented by a single target fragment length variant, with the remaining reads bearing 3 different target fragment length variants each represented by > 1 read. This observation is even more extreme when looking at sequence variation, where figure 4.10 illustrates the presence of 15 discrete sequence variants of the target cgd6_1080.P.108-164 (the gp60 VNTR locus) within the *C. parvum* clinical isolate, UKP3, alone.

General clinical surveys do not demonstrate heterogeneity within clinical samples. This may be explained by a lack of sensitivity to identify multiple discrete allelic populations within a sample using PCR alone, particularly in these instances where population structure is characterised by a single highly dominant population and the fragment sizes of the populations are so similar. It is probable that clinical surveys are detecting and reporting the most abundant population, and the sub-dominant populations, in effect, are being obscured by the dominant one. It has previously been reported that such PCR based typing approaches commonly underestimate MOI within clinical samples [Zhong et al., 2018]. This is particularly problematic in a clinical context, where mis-reporting of the allelic profile of a clinical sample may alter the treatment decisions the clinician may make.

The results of gp60 MOI analysis presented in Figure 5.5 clearly show relationships between each isolate by the distribution of gp60 subtypes present within the sample. For example, UKP4, UKP5 and UKP6 all present as having very similar MOI-signature, and were isolated from the same cluster case, giving confidence in these results and indicating that MOI-signature may be a novel method of high-sensitivity subtyping. This is also indicated by the clustering that can be seen in 5.12, which shows that this dataset resolves into a number of clades, each with complex structure, when clustering is performed using gp60 fragment length MOI-signature rather than gp60 subtype alone. A comparison of the gp60 MOI-signature tree (Figure 5.12) and the dominant allele sequence alignment tree (Figure 5.15) indicate some notable differences. Isolates UKP4, 5, 6, 102, 103, 104, 118, 122, 133, & 134 are demonstrated as identical by sequence similarity of the dominant allele, however, MOI-signature indicates significantly more complexity to the population structure, with 9 distinct branches. In particular, UKP4, 5, & 6 appear to be relatively distant within this tree, due to the presence of dissimilar sub-dominant allelic profiles. Diversity in the yellow clade seen in Figure 5.12 is likely a result of alleles 71 and 53 only

being represented as dominant by single isolates, UKP2 and UKP8 respectively, whereas UKP7, 106, and 107 all bear dominant allele 59. This indicates difficulties associated with clustering according to MOI-signature of isolates bearing rarer dominant alleles.

The MOI-signature tree generated with the tandem repeat locus cgd7_440.P.1066-1129 shown in 5.14 indicates the presence of two major clades, split into 6 minor ones (designated by colour). These two major clades are split by the fragment length of the dominant allele, with the green clade bearing dominant allele 63, and the other clades bearing varying dominant alleles. The green clade accounts for the majority of this dataset (17/32 isolates). However, this clade is diverse, bearing a maximum vertical distance of 1.2. The MOI tree for cgd7_440.P.1066-1129 exhibits deeper roots at each coloured clade than that of gp60 seen in Figure 5.12, indicating higher levels of diversity amongst sub-dominant alleles than that of gp60. There are some key similarities between the trees seen in Figures 5.12 and 5.14, markedly the membership and size of the largest (green) clade. There are also similarities in the placement of rarer MOI-signatures (yellow clade) which contain UKP8 and 106 in trees generated for both gp60 and cgd7_440. UKP129 is the only isolate which is mono-dominant (i.e. has only one allele present within the MOI-signature) across both loci, bearing alleles 68 and 51 for gp60 and cgd7_440.P.1066-1129 respectively. It is worth noting, however, that the MOI-signature generated using the gp60 locus for this isolate was represented by only a single read, due to low coverage at this region within this dataset.

Comparison of trees seen in Figures 5.12 and 5.15 highlight some differences in the placement of certain isolates. The green clade is shown to be identical in both trees. The placements on the rest of the tree follow a similar general structure, with the red and blue clades clustering in a similar manner. However, members of the yellow and purple clades do not cluster together. These differences are likely due to sequence variation, which is not accounted for in the MOI-signature tree. The MOI-signature tree reveals extensive structure in the red and green clades which is not revealed by the gp60 sequence tree, indicating that dominant populations within these samples bear an identical gp60 subtype, but a great deal of variation exists amongst the sub-dominant populations.

Analysis of LD among allele pairs for both loci show extensive positive LD, indicating high levels of non-random allelic association. Structure within Figures 5.10 and 5.8 indicate higher levels of positive LD between allele pairs which are closer in fragment length. This suggests that MOI-signature variation may be as a result of repeat copy number expansion or retraction, rather than as a result of multiple sources of infection, or a non-clonal infective event, as we would not expect to see correlation between LD and relative allele length were this the case. These results suggest a mechanism by which MOI

can occur, wherein a single dominant population baring a single allele replicates and, due to errors in DNA replication, variation is introduced into these regions by slipped strand mispairing, resulting in the advent of a novel allele and therefore population. The results support this mechanism of MOI diversity over non-clonal infection of a host. Further work must be carried out to verify this hypothesis.

5.5 Conclusion

In this chapter, I have presented a detailed, and entirely *in silico*, analysis of Multiplicity of Infection within a diverse *Cryptosporidium parvum* dataset isolated from clinical samples using BlooMine. Heterogeneity within these samples was extensive across two VNTR loci (gp60 and cgd7_440.P.1066-1129), which supports the findings of experimental studies into MOI within clinical sample of *C. parvum* [Grinberg and Widmer, 2016, Troell et al., 2016]. I have also presented a novel typing methodology using the diversity of alleles at a single locus within a clinical sample, termed MOI-signature typing. I demonstrated that MOI-signature typing provides more detailed information about the diversity of populations of *C. parvum* within a clinical isolate than conventional 'dominant-allele' typing approaches, and therefore presents epidemiological surveys with a new dimension of data with which to investigate transmission dynamics. Finally, I demonstrate that there is a clear association between the incidence of certain pairs of alleles. In particular, alleles which are closer in fragment size appear to be more likely to be in linkage disequilibrium, indicating that the method by which these incidences of MOI occur is variation from a single clonal population, rather than infection of the host with multiple distict populations. More work needs to be done to verify this hypothesis, and to detect these MOI events experimentally. These results represent the most thorough investigation into MOI that has been done for Cryptosporidium, and lays the groundwork for further research into this under-investigated aspect of Cryptosporidium biology. Such projects are essential in the development of preventative strategies, novel therapeutics, and our interpretation of epidemiological surveys, and are therefore of great importance in the fight against Cryptosporidiosis.

Chapter 6

Conclusions and Future Work

6.1 Project Review

Due to the huge toll Cryptosporidiosis takes on global human health annually, the importance of efforts to improve our understanding of the transmission cycles of this parasite cannot be overstated. Genome mining proves to be an essential tool in advancing our understanding of this parasite. It allows us to formulate novel methods of detection and transmission investigation, facilitating the development of novel prevention strategies which are so essential for reducing the spread of this parasite and the disease it causes.

6.1.1 Generating Cryptosporidium Genomes from Clinical Samples

The problems associated with generating high quality whole-genome assemblies from clinical samples of Cryptosporidium are largely due to the insufficient recovery of DNA from these samples. This problem is due to a number of factors, such as small sample sizes, low oocyst recovery rates, low DNA yield per oocyst, and extensive cleaning of oocysts needed prior to extraction leading to oocyst loss. Consequently, DNA enrichment using Multiple-Displacement Amplification (MDA) has been used to increase the amount of genomic DNA available for sequencing.

In Chapter 2 I discuss how low levels of DNA isolated from clinical samples can lead to extremely uneven depth of coverage across the genome, resulting in extensive misassembly in Cryptosporidium whole genome assemblies. I then use the Gini coefficient as a method of measuring the inequality of depth of coverage throughout the genome, and present a novel extension of this coefficient to resolve the issues associated with data granularity when calculating the Gini coefficient, termed Gini-granularity curves. These curves are used to investigate the distribution of reads across a genome, and characterise the nature of the inequality of coverage, i.e. whether there are large areas of read aggrega-

tion, or very high levels of spiking. Furthermore, in Section 2.4 I demonstrate that using MDA to enrich DNA for sequencing results in complex alterations of the distribution of coverage plotted against GC content. Gini-granularity curves were used to indicate that using MDA results in significant bias of large ($> 10kb$) portions of the genome, which cannot be easily characterised by using other methods of analysis, such a word frequency or GC content bias.

I demonstrate a method by which reliable, high-quality genome assemblies can be generated using a pipeline to carry out *de novo* assembly and post-assembly improvement. This pipeline was shown to be effective at resolving repetitive regions, which may be investigated for utility as novel biomarkers to interrogate for diagnostic purposes.

The work presented in this chapter represents **the first account of the wide scale misassembly which can result from of uneven read coverage across Cryptosporidium genomes** (and subsequently the first attempt to resolve these issues) [Morris et al., 2019b], and **the first attempt to characterise the way in which WGA affects the distribution of reads across the Cryptosporidium genome** [Morris et al., 2019a].

6.1.1.1 Elucidating the Criteria by which Sequences are Biased by WGA

Despite the investigation presented within this chapter, the criterion by which WGA selectively amplified certain regions over others remains elusive. We can, with a large degree of certainty, say that it biases large portions of the genome, which significantly alters the way in which reads aggregate throughout a genome, and that this poses an often obfuscated challenge to a bioinformatician. However, further investigations should be carried out to elucidate the manner in which regions are biased, as it is my opinion that it is not, as suggested in some articles, entirely random.

6.1.2 Identifying Novel Biomarkers Around the Genome of Cryptosporidium

Tackling Cryptosporidiosis presents a number of issues to medical professionals. The primary issue is that the treatment is seldom effective in cases where the patient is immunocompromised. Furthermore, in these instances, Cryptosporidiosis is life-threatening. Emphasis has therefore been placed on tackling this disease at source and preventing its spread. The development of prevention strategies are dependant on our ability to track this disease and elucidate transmission cycles, which in turn necessitates the development of highly sensitive, specific, and reliable subtyping schemes. The conventional method

of Cryptosporidium subtyping is to interrogate the VNTR locus within a region coding for a 60 Kda sporozoite surface protein: gp60 (gene cgd6_1080). This gp60 region has been shown as highly polymorphic within *C. parvum*. However, there remain a number of shortcomings of using this typing scheme:

- It is known to poorly differentiate *C. hominis*.
- Cryptosporidium has been reported to recombine at this locus [Widmer and Lee, 2010].
- As a single locus typing scheme, it lacks the reliability and resolution of multi-locus typing.

These issues necessitate the development of novel typing schemes, driven by identifying polymorphic regions around the genomes of Cryptosporidium.

In Chapter 3 I present VaNTA, a pipeline developed for the discovery of novel VNTR's *in silico*. It is used to identify novel VNTRs in coding regions around the genome of Cryptosporidium and Plasmodium, and assess them for viability as biomarkers. A number of VNTRs within Cryptosporidium were presented as **comparable in their typing power to the gp60 locus**. Furthermore, we present a large number of VNTRs discovered around the genome of *Plasmodium falciparum* which exhibit strong typing power. VaNTA can be used to identify VNTRs within any organism, making it very versatile. Consequently it will be made available for public use on the Galaxy platform, improving usability and availability.

It is the intention that VNTR regions identified here will be used in Cryptosporidium surveillance studies carried out by the Public Health Wales Cryptosporidium Reference Unit, which will aid in the development of effective prevention strategies against Cryptosporidiosis.

6.1.2.1 Experimental Validation of VNTRs Identified by VaNTA

Further work must be done to validate this biomarkers as appropriate for clinical application. Small variations in fragment length can be difficult to detect experimentally, necessitating further experimental evaluation. This will involve designing primers using the flanking sequences of these VNTR regions (which are assessed for conservation by VaNTA, for this purpose) and amplifying them using PCR. These PCR products should then be sequenced and assessed for fragment length variation which can be detected using conventional fragment length typing methodologies.

6.1.3 Mining Reads for Sequences of Interest

The bottleneck between data generation and data analysis is a major problem in genomic science. A primary reason for this bottleneck is that the technological progressions for genome sequencing outstrip the progressions of genomic data analysis. This bottleneck necessitates the development of tools which facilitate genome analysis in a cost and time efficient way. Alignment-free sequence analysis presents a method by which genomes can be analysed in a high-throughput and computationally efficient manner. Using alignment-free sequence analysis, genome analysis can be performed on raw read data, obviating the computationally expensive and time consuming task of genome assembly.

In Chapter 4 I present BlooMine, a tool developed to mine raw read sets(.fastq) for query sequences. This tool utilises Bloom Filters, a highly efficient probabilistic data structure, to perform set membership queries between kmer arrays generated from a query sequence and a read. BlooMine proves **highly reliable at pulling out reads containing query sequences from a large simulated read set**, despite variations being introduced into the query sequences of these reads (see Section 4.7.1). Furthermore, BlooMine was used to **identify variations in the target locus within a read set**, which it achieved on the simulated dataset with high fidelity and sensitivity.

6.1.4 *in silico* Detection of Multiplicity of Infection in Cryptosporidium

The effect Multiplicity of Infection has on clinically important factors such as virulence, transmission, and clinical presentation are cryptic and poorly understood, leading Alizon *et al.* to remark "Depending on the biological system, overall virulence[1] can be higher than the virulence of the most virulent parasite, lower than the virulence of the least virulent parasite, or take some intermediate value between the two" [Alizon et al., 2013]. This complex issue is made all the more difficult by the lack of reliable, high-throughput approaches to investigate it. Most studies carried out to investigate MOI do so using experimental techniques (detailed in Section 5.1) which are very labour intensive and therefore scale poorly. There is consequently a great need for bioinformatic tools which can carry out this kind of investigation in a reliable, sensitive, and scalable manner.

In Chapter 5 I use BlooMine to carry out an analysis of the MOI of *Cryptosporidium parvum* across the 100 highest scoring VNTR loci identified by VaNTA in Chapter 3. I then carry out further analysis of MOI at two loci: gp60 and cgd7_440.P.1066-1129.

[1] "Overall virulence" here is defined as the virulence experienced by the multiply infected host, which is a complex product of the interaction between the individual genotypes of the pathogen which are present in the host, and between the pathogen population and the host biological system.

The results demonstrate that there is extensive sub-dominant population structure at these two loci, indicating a **high level of *Cryptosporidium parvum* population complexity within a single host**. Finally, I present a novel typing scheme, termed MOI-signature typing. This typing scheme involves using in-host diversity at a single locus to identify the relationship between clinical isolates. Using this scheme I show a **complex population structure within this *C. parvum* dataset ($n = 36$), which is not achievable using conventional dominant allele typing schemes** (such as that used in gp60 subtyping). The characteristics of *C. parvum* population structure are in accordance with experimental results [Grinberg et al., 2013, Troell et al., 2016]. A hypothesis is also presented, suggesting that the advent of MOI in a host is more commonly as a result of divergence within the host than non-clonal infection. The results presented in this chapter are supported by epidemiological data, where available. **To my knowledge, these results represent the largest entirely *in silico* evaluation of MOI carried out on clinical isolates of a unicellular eukaryotic parasite.**

6.1.4.1 Experimental Validation of MOI-signature Typing

The discordance between clinical and *in silico* results of MOI detection has previously been documented. However, it is possible that conventional methods of clinical typing lack the sensitivity to elucidate sub-dominant populations of Cryptosporidium [Zhong et al., 2018], since we have demonstrated that there is often an order of magnitude difference in abundance between the amount of DNA present (extrapolated from coverage over individual alleles) for each allele (see Figures 5.5 and 5.6). Further work must therefore be undertaken to validate the presence of these sub-populations experimentally, as these *in silico* results, despite being supported indirectly by the observations of other teams and of epidemiological collection data, are still predictions of putative sub-populations.

Care must be taken when evaluating MOI in samples which have been subjected to DNA enrichment using WGA, as such enrichment may have an effect on the relative abundances of sub-populations, or introduce pseudo-populations by incorrect amplification. The hypothesis presented, suggesting MOI more commonly arises from variation within a host rather than as a result of non-clonal infection also warrants further experimental investigation. Finally, it is a possibility that errors introduced during sequencing may exaggerate the level of MOI reported, therefore care should be taken when interpreting the results in the absence of experimental ones.

6.2 Conclusion

Within this thesis, I have detailed the work carried out using the genomic sequencing data available for Cryptosporidium to generate high-quality and reliable whole genome assemblies, used these assemblies to prospect for VNTRs for use as novel biomarkers, then used these biomarkers to illustrate that there is extensive Multiplicity of Infection in clinical Cryptosporidiosis. The problems encountered and detailed here were resolved using a variety of tools, some widely recognised and used (such as *de novo* assemblers like spades, and aligners such as EMBOSS-water). However, some of these problems required the development of entirely novel approaches and tools, which lead to the development of VaNTA and BlooMine. The importance of developing such tools and applying them to solve these biological problems cannot be understated, nor overlooked. The power of *in silico* analysis will only increase as more 'omics data is generated with higher and higher quality, and as computational power increases, meaning such tools will only increase in their value and scope to resolve these problems and investigate biological phenomena in a high-throughput manner. It is only by a thorough understanding of the transmission of diseases such as Cryptosporidiosis, facilitated by such analyses and collaborations like the one which lead to the development of this project, that we can hope to prevent their spread and reduce the vast toll they have on human and animal health worldwide.

Bibliography

[Abrahamsen et al., 2004] Abrahamsen, M. S., Lancto, C. A., Deng, M., Liu, C., Bankier, A. T., Dear, P. H., Konfortov, B. A., Spriggs, H. F., Iyer, L., Anantharaman, V., Aravind, L., Kapur, V., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Widmer, G., Tzipori, S., Buck, G. a., and Xu, P. (2004). Complete Genome Sequence of the Apicomplexan, Cryptosporidium parvum. *Science*, 304(5669):441–445.

[Alizon et al., 2013] Alizon, S., de Roode, J. C., and Michalakis, Y. (2013). Multiple infections and the evolution of virulence. *Ecology Letters*, 16(4):556–567.

[Alkan et al., 2011] Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5):363–376.

[Alsmadi et al., 2009] Alsmadi, O., Alkayal, F., Monies, D., and Meyer, B. F. (2009). Specific and complete human genome amplification with improved yield achieved by phi29 DNA polymerase and a novel primer at elevated temperature. *BMC Research Notes*, 2:1–7.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool.

[Assefa et al., 2009] Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., and Berriman, M. (2009). ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15):1968–1969.

[Assefa et al., 2014] Assefa, S. A., Preston, M. D., Campino, S., Ocholla, H., Sutherland, C. J., and Clark, T. G. (2014). EstMOI: Estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics*.

[Balmer and Tanner, 2011] Balmer, O. and Tanner, M. (2011). Prevalence and implications of multiple-strain infections.

[Bankevich et al., 2012] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a., and Pevzner, P. a. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

[Barker and Carbonell, 1974] Barker, I. K. and Carbonell, P. L. (1974). Cryptosporidium agni sp.n. from lambs, and Cryptosporidium bovis sp.n. from a calf, with observations on the oocyst. *Zeitschrift f??r Parasitenkunde*, 44(4):289–298.

[Bell et al., 2006] Bell, A. S., de Roode, J. C., Sim, D., and Read, A. F. (2006). Within-Host Competition in Genetically Diverse Malaria Infections: Parasite Virulence and Competitive Success. *Evolution*, 60(7):1358.

[Benjamini and Speed, 2012] Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):1–14.

[Benson, 1999] Benson, G. (1999). Tandem Repeats Finder: a program to analyse DNA sequences. *Nucleic Acids Res.*, 27(2):573–578.

[Börgstrom et al., 2017] Börgstrom, E., Paterlini, M., Mold, J. E., Frisen, J., and Lundeberg, J. (2017). Comparison of whole genome amplification techniques for human single cell exome sequencing. *PLoS ONE*, 12(2):1–15.

[Borowski et al., 2008] Borowski, H., Clode, P. L., and Thompson, R. C. A. (2008). Active invasion and/or encapsulation? A reappraisal of host-cell parasitism by Cryptosporidium. *Trends in Parasitology*, 24(11):509–516.

[Bull et al., 1998] Bull, S., Chalmers, R., Sturdee, A. P., Curry, A., and Kennaugh, J. (1998). Cross-reaction of an anti-Cryptosporidium monoclonal antibody with sporocysts of Monocystis species. *Veterinary Parasitology*, 77(2-3):195–197.

[Carreno et al., 1999] Carreno, R. A., Martin, D. S., and Barta, J. R. (1999). Cryptosporidium is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. *Parasitology research*, 85(11):899–904.

[Carreno et al., 1998] Carreno, R. A., Schnitzler, B. B., Jeffries, A., Tenter, A., Johnson, A. M., and Barta, J. R. (1998). Phylogenetic analysis of coccidia based on 18S rDNA sequence comparison indicates that Isospora is most closely related to Toxoplasma and Neospora. *The Journal of Eukaryotic Microbiology*, 45(2):184–188.

[Casemore and Jackson, 1984] Casemore, D. P. and Jackson, F. B. (1984). Hypothesis: Cryptosporidiosis in Human Beings is not Primarily a Zoonosis. *Journal of Infection*, 9:153–156.

[Casemore et al., 1985] Casemore, D. P., Sands, R. L., and Curry, a. (1985). Cryptosporidium species a "new" human pathogen. *Journal of Clinical Pathology*, 38(12):1321–1336.

[Cavalier-Smith, 2014] Cavalier-Smith, T. (2014). Gregarine site-heterogeneous 18S rDNA trees, revision of gregarine higher classification, and the evolutionary diversification of Sporozoa. *European Journal of Protistology*, 50(5):472–495.

[Chalmers and Davies, 2010] Chalmers, R. M. and Davies, A. P. (2010). Minireview: Clinical cryptosporidiosis. *Experimental Parasitology*, 124(1):138–146.

[Chalmers et al., 2017] Chalmers, R. M., Robinson, G., Hotchkiss, E., Alexander, C., May, S., Gilray, J., Connelly, L., and Hadfield, S. J. (2017). Suitability of loci for multiple-locus variable-number of tandem-repeats analysis of Cryptosporidium parvum for inter-laboratory surveillance and outbreak investigations. *Parasitology*, 144(1):37–47.

[Char et al., 1996] Char, S., Kelly, P., Naeem, A., and Farthing, M. J. (1996). Codon usage in Cryptosporidium parvum differs from that in other Eimeriorina. *Parasitology*, 112 (Pt 4:357–362.

[Clode et al., 2015] Clode, P. L., Koh, W. H., and Thompson, R. C. A. (2015). Life without a Host Cell: What is Cryptosporidium?

[Compeau et al., 2011] Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991.

[Csuros et al., 2007] Csuros, M., Noe, L., and Kucherov, G. (2007). Reconsidering the significance of genomic word frequencies. *Trends in Genetics*, 23(11):543–546.

[Daly et al., 2015] Daly, G. M., Leggett, R. M., Rowe, W., Stubbs, S., Wilkinson, M., Ramirez-Gonzalez, R. H., Caccamo, M., Bernal, W., and Heeney, J. L. (2015). Host subtraction, filtering and assembly validations for novel viral discovery using next generation sequencing data. *PLoS ONE*, 10(6):1–28.

[De Roeck et al., 2018] De Roeck, A., Duchateau, L., Van Dongen, J., Cacace, R., Bjerke, M., Van den Bossche, T., Cras, P., Vandenberghe, R., De Deyn, P. P., Engelborghs, S., Van Broeckhoven, C., Sleegers, K., Goeman, J., Crols, R., Nuytten, D., Mercelis, R., Vandenbulcke, M., Sieben, A., De Bleecker, J. L., Santens, P., Versijpt, J., Michotte, A., Deryck, O., Vanopdenbosch, L., Bergmans, B., Willems, C., De Klippel, N., Delbeck, J., Ivanoiu, A., and Salmon, E. (2018). An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer's disease. *Acta Neuropathologica*, 135(6):827–837.

[Delehelle et al., 2018] Delehelle, F., Cussat-Blanc, S., Alliot, J. M., Luga, H., and Balaresque, P. (2018). ASGART: Fast and parallel genome scale segmental duplications mapping. *Bioinformatics*, 34(16):2708–2714.

[Denoeud and Vergnaud, 2004] Denoeud, F. and Vergnaud, G. (2004). Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. *BMC bioinformatics*, 5:4.

[Dohm et al., 2008] Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16).

[Edwards et al., 2012] Edwards, R. A., Olson, R., Disz, T., Pusch, G. D., Vonstein, V., Stevens, R., and Overbeek, R. (2012). Real Time Metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics*, 28(24):3316–3317.

[Eisenberg et al., 2005] Eisenberg, J. N. S., Lei, X., Hubbard, A. H., Brookhart, M. A., and Colford, J. M. (2005). The role of disease transmission and conferred immunity in outbreaks: Analysis of the 1993 Cryptosporidium outbreak in Milwaukee, Wisconsin. *American Journal of Epidemiology*, 161(1):62–72.

[Fayer, 2010] Fayer, R. (2010). Taxonomy and species delimitation in Cryptosporidium. *Experimental Parasitology*, 124(1):90–97.

[Gardner et al., 2002] Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Davis, R. W., Fraser, C. M., and Barrell, B. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511.

[Gardy et al., 2018] Gardy, J. L., Lee, R. S., Cowley, L. A., Hanage, W. P., and Martin, M. A. (2018). Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. *Microbial Genomics*, 4(10):1–8.

[Gelfand et al., 2014] Gelfand, Y., Hernandez, Y., Loving, J., and Benson, G. (2014). VNTRseek - a computational tool to detect tandem repeat variants in high-throughput sequencing data. *Nucleic Acids Research*, 42(14):8884–8894.

[Gilchrist et al., 2018] Gilchrist, C. A., Cotton, J. A., Burkey, C., Arju, T., Gilmartin, A., Lin, Y., Ahmed, E., Steiner, K., Alam, M., Ahmed, S., Robinson, G., Zaman, S. U., Kabir, M., Sanders, M., Chalmers, R. M., Ahmed, T., Ma, J. Z., Haque, R., Faruque, A. S., Berriman, M., and Petri, W. A. (2018). Genetic diversity of cryptosporidium hominis in a bangladeshi community as revealed by whole-genome sequencing. *Journal of Infectious Diseases*, 218(2):259–264.

[Graham, 2008] Graham, A. L. (2008). Ecological rules governing helminth microparasite coinfection. *Pnas*, 105(2).

[Grinberg et al., 2013] Grinberg, A., Biggs, P. J., Dukkipati, V. S., and George, T. T. (2013). Extensive intra-host genetic diversity uncovered in Cryptosporidium parvum using Next Generation Sequencing. *Infection, Genetics and Evolution*, 15:18–24.

[Grinberg and Widmer, 2016] Grinberg, A. and Widmer, G. (2016). Cryptosporidium within-host genetic diversity: systematic bibliographical search and narrative overview.

[Guo et al., 2015a] Guo, Y., Li, N., Lysén, C., Frace, M., Tang, K., Sammons, S., Roellig, D. M., Feng, Y., and Xiao, L. (2015a). Isolation and enrichment of cryptosporidium DNA and verification of DNA purity for whole-genome sequencing. *Journal of Clinical Microbiology*, 53(2):641–647.

[Guo et al., 2015b] Guo, Y., Tang, K., Rowe, L. A., Li, N., Roellig, D. M., Knipe, K., Frace, M., Yang, C., Feng, Y., and Xiao, L. (2015b). Comparative genomic analysis reveals occurrence of genetic recombination in virulent Cryptosporidium hominis subtypes and telomeric gene duplications in Cryptosporidium parvum. *BMC genomics*, 16:320.

[Gupta et al., 2016] Gupta, A., Jordan, I. K., and Rishishwar, L. (2016). stringMLST: a fast k-mer based tool for multi locus sequence typing. *Bioinformatics (Oxford, England)*, page btw586.

[Gymrek et al., 2012] Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, 22(6):1154–1162.

[Hadfield et al., 2015] Hadfield, S. J., Pachebat, J. A., Swain, M. T., Robinson, G., Cameron, S. J., Alexander, J., Hegarty, M. J., Elwin, K., and Chalmers, R. M. (2015). Generation of whole genome sequences of new Cryptosporidium hominis and Cryptosporidium parvum isolates directly from stool samples. *BMC genomics*, 16:650.

[Hawash, 2014] Hawash, Y. (2014). DNA extraction from protozoan oocysts/cysts in feces for diagnostic PCR. *Korean Journal of Parasitology*, 52(3):263–271.

[Henning et al., 2004] Henning, L., Schellenberg, D., Smith, T., Henning, D., Alonso, P., Tanner, M., Mshinda, H., Beck, H. P., and Felger, I. (2004). A prospective study of Plasmodium falciparum multiplicity of infection and morbidity in Tanzanian children. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 98(12):687–694.

[Highnam et al., 2013] Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., and Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research*, 41(1):1–7.

[Hijjawi et al., 2001] Hijjawi, N. S., Meloni, B. P., Morgan, U. M., and Thompson, R. C. A. (2001). Complete development and long-term maintenance of Cryptosporidium parvum human and cattle genotypes in cell culture. *International Journal for Parasitology*, 31(10):1048–1055.

[Hijjawi et al., 2004] Hijjawi, N. S., Meloni, B. P., Ng'anzo, M., Ryan, U. M., Olson, M. E., Cox, P. T., Monis, P. T., and Thompson, R. C. A. (2004). Complete development of Cryptosporidium parvum in host cell-free culture. *International Journal for Parasitology*, 34(7):769–777.

[Hosono et al., 2003] Hosono, S., Faruqi, A. F., Dean, F. B., Du, Y., Sun, Z., Wu, X., Du, J., Kingsmore, S. F., Egholm, M., and Lasken, R. S. (2003). Unbiased whole-genome amplification directly from clinical samples. *Genome Research*, 13(5):954–964.

[Hunter et al., 2007] Hunter, P. R., Hadfield, S. J., Wilkinson, D., Lake, I. R., Harrison, F. C., and Chalmers, R. M. (2007). Correlation between Subtypes of Cryptosporidium parvum in Humans and Risk. *Emerging Infectious Diseases*, 13(1):82–88.

[Hunter and Nichols, 2002] Hunter, P. R. and Nichols, G. (2002). Epidemiology and

clinical features of Cryptosporidium infection in immunocompromised patients. *Clin.Microbiol.Rev.*, 15(0893-8512 (Print)):145–154.

[Ifeonu et al., 2016] Ifeonu, O. O., Chibucos, M. C., Orvis, J., Su, Q., Elwin, K., Guo, F., Zhang, H., Xiao, L., Sun, M., Chalmers, R. M., Fraser, C. M., Zhu, G., Kissinger, J. C., Widmer, G., and Silva, J. C. (2016). Annotated draft genome sequences of three species of Cryptosporidium: Cryptosporidium meleagridis isolate UKMEL1, C. baileyi isolate TAMU-09Q1 and C. hominis isolates TU502 2012 and UKH1. *Pathogens and Disease*.

[Isaza et al., 2015] Isaza, J. P., Galván, A. L., Polanco, V., Huang, B., Matveyev, A. V., Serrano, M. G., Manque, P., Buck, G. A., and Alzate, J. F. (2015). Revisiting the reference genomes of human pathogenic Cryptosporidium species: reannotation of C. parvum Iowa and a new C. hominis reference. *Scientific Reports*, 5(October):16324.

[Jaskiewicz et al., 2018] Jaskiewicz, J. J., Sandlin, R. D., Swei, A. A., Widmer, G., Toner, M., and Tzipori, S. (2018). Cryopreservation of infectious Cryptosporidium parvum oocysts. *Nature Communications*, 9(1):1–8.

[Jones et al., 2001] Jones, E., Oliphant, T., Peterson, P., and Al, E. (2001). SciPy: Open source scientific tools for Python.

[Kaplinski et al., 2015] Kaplinski, L., Lepamets, M., and Remm, M. (2015). GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, 4:58.

[Kaupke et al., 2017] Kaupke, A., Gawor, J., Rzeżutka, A., and Gromadka, R. (2017). Identification of pig-specific Cryptosporidium species in mixed infections using Illumina sequencing technology. *Experimental Parasitology*.

[Koh et al., 2013] Koh, W., Clode, P. L., Monis, P., and Thompson, R. C. A. (2013). Multiplication of the waterborne pathogen Cryptosporidium parvum in an aquatic biofilm system. *Parasites & vectors*, 6:270.

[Koh et al., 2014] Koh, W., Thompson, A., Edwards, H., Monis, P., and Clode, P. L. (2014). Extracellular excystation and development of Cryptosporidium: tracing the fate of oocysts within Pseudomonas aquatic biofilm systems. *BMC microbiology*, 14:281.

[Kojima et al., 2016] Kojima, K., Kawai, Y., Misawa, K., Mimori, T., and Nagasaki, M. (2016). STR-realigner: A realignment method for short tandem repeat regions. *BMC Genomics*.

[Kristmundsdóttir et al., 2016] Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B., and Halldórsson, B. V. (2016). popSTR: population-scale detection of STR variants. *Bioinformatics*, page btw568.

[Krzywinski et al., 2009] Krzywinski, M., Schein, J., Birol, n., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos. *Genome Research*, 19(9):1639–1645.

[Kurtz et al., 2008] Kurtz, S., Narechania, A., Stein, J. C., and Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics*, 9:517.

[Kurtz et al., 2004] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

[Lasken and Egholm, 2003] Lasken, R. S. and Egholm, M. (2003). Whole genome amplification: Abundant supplies of DNA from precious samples or clinical specimens. *Trends in Biotechnology*, 21(12):531–535.

[Leander et al., 2003] Leander, B. S., Clopton, R. E., and Keeling, P. J. (2003). Phylogeny of grenarines (Apicomplexa) as inferred from a small-subunit rDNA and beta-tubulin. *International Journal of Systematic and Evolutionary Microbiology*, 53(1):345–354.

[Leggett et al., 2013] Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., and Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in Genetics*, 4(DEC):1–5.

[Leibler and Kullback, 1951] Leibler, R. A. and Kullback, S. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li et al., 2010] Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. pages 265–272.

[Lord et al., 1999] Lord, C. C., Trenholme, K., Day, K., Hargrove, J. W., Woolhouse, M. E. J., McNamara, J. J., Paul, R. E. L., and Barnard, B. (1999). Aggregation and distribution of strains in microparasites. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 354(1384):799–807.

[Manske and Kwiatkowski, 2009] Manske, H. M. and Kwiatkowski, D. P. (2009). SNP-o-matic. *Bioinformatics*, 25(18):2434–2435.

[Marçais and Kingsford, 2011] Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

[Marchet et al., 2017] Marchet, C., Lecompte, L., Da Silva, C., Cruaud, C., Aury, J. M., Nicolas, J., and Peterlongo, P. (2017). De novo Clustering Nanopore Long Reads of Transcriptomics Data by Gene. *Doi.Org*, page 170035.

[Mardis, 2010] Mardis, E. R. (2010). The $1,000 genome, the $100,000 analysis? *Genome Medicine*, 2(11):7–9.

[Marinier et al., 2019] Marinier, E., Enns, E., Tran, C., Fogel, M., Peters, C., Kidwai, A., Ji, H., and Van Domselaar, G. (2019). Quasitools: A Collection of Tools for Viral Quasispecies Analysis. *bioRxiv*, pages 1–4.

[McIver et al., 2011] McIver, L. J., Fondon, J. W., Skinner, M. A., and Garner, H. R. (2011). Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics*, 97(4):193–199.

[McIver et al., 2013] McIver, L. J., McCormick, J. F., Martin, A., Fondon, J. W., and Garner, H. R. (2013). Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene*, 516(2):328–334.

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*.

[Melsted and Halldorsson, 2014] Melsted, P. and Halldorsson, B. V. (2014). KmerStream: Streaming algorithms for k-mer abundance estimation. *Bioinformatics*, 30(24):3541–3547.

[Miller et al., 2018] Miller, C. N., Jossé, L., Brown, I., Blakeman, B., Povey, J., Yiangou, L., Price, M., Cinatl, J., Xue, W. F., Michaelis, M., and Tsaousis, A. D. (2018). A cell culture platform for Cryptosporidium that enables long-term cultivation and new tools for the systematic investigation of its biology. *International Journal for Parasitology*, 48(3-4):197–201.

[Mohammadi et al., 2005] Mohammadi, T., Reesink, H. W., Vandenbroucke-Grauls, C. M., and Savelkoul, P. H. (2005). Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *Journal of Microbiological Methods*, 61(2):285–288.

[Monfort, 2008] Monfort, P. (2008). Convergence of EU regions - Measures and evolution. *European Union*, Europa.(6):1–32.

[Morgan-Ryan et al., 2002] Morgan-Ryan, M., Fall, A., Ward, L. A., Hijjawi, N., Sulaiman, I., Fayer, R., Thompson, R. C. A., Olson, M., Lal, A., and Xiao, L. (2002). Cryptosporidium hominis n. sp. (Apicomplexa: Cryptosporidiidae) from Homo sapiens. *The Journal of Eukaryotic Microbiology*, 49(6):433–440.

[Morris et al., 2019a] Morris, A., Pachebat, J., Tyson, G., Robinson, G., Chalmers, R., and Swain, M. (2019a). Generating Reliable Genome Assemblies of Intestinal Protozoans from Clinical Samples for the Purpose of Biomarker Discovery. *Communications in Computer and Information Science*, In Press:1–25.

[Morris et al., 2019b] Morris, A. V., Pachebat, J., Robinson, G., Chalmers, R., and Swain, M. (2019b). Identifying and Resolving Genome Misassembly Issues Important for Biomarker Discovery in the Protozoan Parasite , Cryptosporidium. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS*, volume 3, pages 90–100. SciTePress.

[Morris et al., 2019c] Morris, A. V., Robinson, G., Swain, M. T., and Chalmers, R. M. (2019c). Direct sequencing of Cryptosporidium in stool samples for public health. *Frontiers in Pathogen Genomics*, In Press.

[Morrison and Ellis, 1997] Morrison, D. a. and Ellis, J. T. (1997). Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Molecular biology and evolution*, 14(4):428–441.

[Nader et al., 2019] Nader, J. L., Mathers, T. C., Ward, B. J., Pachebat, J. A., Swain, M. T., Robinson, G., Chalmers, R. M., Hunter, P. R., van Oosterhout, C., and Tyler, K. M. (2019). Evolutionary genomics of anthroponosis in Cryptosporidium. *Nature Microbiology*, 4(5):826–836.

[Ning et al., 2001] Ning, Z., Cox, A. J., Mullikin, J. C., Ning, Z., Cox, A. J., and Mullikin, J. C.

(2001). SSAHA : A Fast Search Method for Large DNA Databases SSAHA : A Fast Search Method for Large DNA Databases. pages 1725–1729.

[Oliphant, 2006] Oliphant, T. (2006). *A Guide to NumPy*. USA: Trelgol Publishing.

[O'Rawe et al., 2013] O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K., and Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Medicine*, 5(3):28.

[Otto et al., 2011] Otto, T. D., Dillon, G. P., Degrave, W. S., and Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 39(9):1–7.

[Otto et al., 2010] Otto, T. D., Sanders, M., Berriman, M., and Newbold, C. (2010). Iterative correction of reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, 26(14):1704–1707.

[Pajuste et al., 2017] Pajuste, F.-D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., and Remm, M. (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific Reports*, 7(1):2537.

[Patro et al., 2014] Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462–464.

[Peng et al., 2012] Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

[Perez-Cordon et al., 2016] Perez-Cordon, G., Robinson, G., Nader, J., and Chalmers, R. M. (2016). Discovery of new variable number tandem repeat loci in multiple Cryptosporidium parvum genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. *Experimental Parasitology*, 169(August):119–128.

[Puiu et al., 2004] Puiu, D., Enomoto, S., Buck, G. A., Abrahamsen, M. S., and Kissinger, J. C. (2004). CryptoDB: the Cryptosporidium genome resource. *Nucleic Acids Research*, 32(90001):329D–331.

[Ramirez-Gonzalez, 2014] Ramirez-Gonzalez, R. H. (2014). Kontaminant: kmer based screening and filtering of next generation reads.

[Read and Taylor, 2001] Read, A. F. and Taylor, L. H. (2001). The ecology of genetically diverse infections.

[Rice et al., 2000] Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(1):276–277.

[Rigoutsos et al., 2006] Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., and Platt, D. (2006). Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proc Natl Acad Sci U S A*, 103(17):6605–6610.

[Rizk et al., 2013] Rizk, G., Lavenier, D., and Chikhi, R. (2013). DSK: K-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653.

[Robinson and Chalmers, 2012] Robinson, G. and Chalmers, R. M. (2012). Assessment of polymorphic genetic markers for multi-locus typing of Cryptosporidium parvum and Cryptosporidium hominis. *Experimental Parasitology*, 132(2):200–215.

[Rosales et al., 2005] Rosales, M. J., Peréz Cordón, G., Sánchez Moreno, M., Marín Sánchez, C., and Mascaró, C. (2005). Extracellular like-gregarine stages of Cryptosporidium parvum. *Acta Tropica*, 95(1):74–78.

[Ross et al., 2013] Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*.

[Rustagi et al., 2016] Rustagi, N., Hampton, O. A., Li, J., Xi, L., Gibbs, R. A., Plon, S. E., Kimmel, M., and Wheeler, D. A. (2016). ITD assembler: an algorithm for internal tandem duplication discovery from short-read sequencing data. *BMC Bioinformatics*, 17(1):188.

[Ryan and Hijjawi, 2015] Ryan, U. and Hijjawi, N. (2015). New developments in Cryptosporidium research. *International Journal for Parasitology*, 45(6):367–373.

[Ryan et al., 2016] Ryan, U., Paparini, A., Monis, P., and Hijjawi, N. (2016). It's official - Cryptosporidium is a gregarine: What are the implications for the water industry? *Water Research*, 105:305–313.

[Salter et al., 2014] Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. (2014). Reagent and

laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12(1):1–12.

[Seppälä et al., 2012] Seppälä, O., Karvonen, A., Louhi, K.-R., Jokela, J., and Rellstab, C. (2012). Reciprocal Interaction Matrix Reveals Complex Genetic and Dose-Dependent Specificity among Coinfecting Parasites. *The American Naturalist*, 180(3):306–315.

[Seppälä et al., 2009] Seppälä, O., Karvonen, A., Valtonen, E. T., and Jokela, J. (2009). Interactions among co-infecting parasite species: A mechanism maintaining genetic variation in parasites? *Proceedings of the Royal Society B: Biological Sciences*.

[Simpson et al., 2009] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123.

[Sivarajah et al., 2013] Sivarajah, V., Ramamurthy, N. K., Rowe, S., and Devalia, K. (2013). Atypical distribution of pneumatosis intestinalis in a patient with AIDS. *BMJ Case Reports*.

[Spano and Crisanti, 2000] Spano, F. and Crisanti, A. (2000). Cryptosporidium parvum: the many secrets of a small genome. *International Journal for Parasitology*, 30(4):553–565.

[Steinbiss et al., 2016] Steinbiss, S., Silva-Franco, F., Brunk, B., Foth, B., Hertz-Fowler, C., Berriman, M., and Otto, T. D. (2016). Companion: a web server for annotation and analysis of parasite genomes. *Nucleic acids research*, 44(W1):W29–W34.

[Swain et al., 2012] Swain, M. T., Tsai, I. J., Assefa, S. a., Newbold, C., Berriman, M., and Otto, T. D. (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nature protocols*, 7(7):1260–84.

[Tan et al., 2010] Tan, J. C., Tan, A., Checkley, L., Honsa, C. M., and Ferdig, M. T. (2010). Variable numbers of tandem repeats in plasmodium falciparum genes. *Journal of Molecular Evolution*.

[Tang and Nzabarushimana, 2017] Tang, H. and Nzabarushimana, E. (2017). STRScan: Targeted profiling of short tandem repeats in whole-genome sequencing data. *BMC Bioinformatics*.

[Templeton et al., 2010] Templeton, T. J., Enomoto, S., Chen, W. J., Huang, C. G., Lancto, C. A., Abrahamsen, M. S., and Zhu, G. (2010). A genome-sequence survey for ascogregarina taiwanensis supports evolutionary affiliation but metabolic diversity between a Gregarine and Cryptosporidium. *Molecular Biology and Evolution*, 27(2):235–248.

[Thorvaldsdóttir et al., 2013] Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192.

[Troell et al., 2016] Troell, K., Hallström, B., Divne, A. M., Alsmark, C., Arrighi, R., Huss, M., Beser, J., and Bertilsson, S. (2016). Cryptosporidium as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics*, 17(1):1–12.

[Tsai et al., 2010] Tsai, I. J., Otto, T. D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, 11(4).

[Tyzzer, 1907] Tyzzer, E. (1907). A sporozoan found in the peptic glands of the common mouse. *Proc. Soc. Exp. Biol. Med.*, 5:12–13.

[Valigurová et al., 2007] Valigurová, A., Hofmannová, L., Koudela, B., and Vávra, J. (2007). An ultrastructural comparison of the attachment sites between Gregarina steini and Cryptosporidium muris. *Journal of Eukaryotic Microbiology*, 54(6):495–510.

[Valigurová et al., 2008] Valigurová, A., JirkÅ¯, M., Koudela, B., Gelnar, M., Modrý, D., and Šlapeta, J. (2008). Cryptosporidia: Epicellular parasites embraced by the host cell membrane. *International Journal for Parasitology*, 38(8-9):913–922.

[Van Der Horst et al., 2013] Van Der Horst, J., Buijs, M. J., Laine, M. L., Wismeijer, D., Loos, B. G., Crielaard, W., and Zaura, E. (2013). Sterile paper points as a bacterial DNA-contamination source in microbiome profiles of clinical samples. *Journal of Dentistry*, 41(12):1297–1301.

[Van Der Walt et al., 2011] Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30.

[Vinga and Almeida, 2003] Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison - A review. *Bioinformatics*, 19(4):513–523.

[Wetzel et al., 2005] Wetzel, D. M., Schmidt, J., Kuhlenschmidt, M. S., Dubey, J. P., and Sibley, L. D. (2005). Gliding motility leads

to active cellular invasion by Cryptosporidium parvum sporozoites. *Infection and Immunity*, 73(9):5379–5387.

[Widmer and Cacciò, 2015] Widmer, G. and Cacciò, S. M. (2015). A comparison of sequence and length polymorphism for genotyping Cryptosporidium isolates. *Parasitology*, (2006):1–6.

[Widmer and Lee, 2010] Widmer, G. and Lee, Y. (2010). Comparison of single- and multilocus genetic diversity in the protozoan parasites cryptosporidium parvum and C. hominis. *Applied and Environmental Microbiology*, 76(19):6639–6644.

[Widmer et al., 2012] Widmer, G., Lee, Y., Hunt, P., Martinelli, A., Tolkoff, M., and Bodi, K. (2012). Comparative genome analysis of two Cryptosporidium parvum isolates with different host range. *Infection, Genetics and Evolution*, 12(6):1213–1221.

[Widmer et al., 2002] Widmer, G., Lin, L., Kapur, V., Feng, X., and Abrahamsen, M. S. (2002). Genomics and genetics of Cryptosporidium parvum: The key to understanding cryptosporidiosis. *Microbes and Infection*, 4(10):1081–1090.

[Widmer and Sullivan, 2012] Widmer, G. and Sullivan, S. (2012). Genomics and population biology of Cryptosporidium species. *Parasite Immunology*, 34(2-3):61–71.

[Wood and Salzberg, 2014] Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.

[Xu et al., 2004] Xu, P., Widmer, G., Wang, Y., Ozald, L., Alves, J., Serrano, M. G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A., Pearson, W., Dear, P. H., Bankier, A. T., Peterson, D., Abrahamsen, M. S., Kapur, V., Tzipori, S., and Buck, G. A. (2004). The Genome of Cryptosporidium hominis. *Letters to Nature*, 431(October).

[Zahedi et al., 2017] Zahedi, A., Gofton, A. W., Jian, F., Paparini, A., Oskam, C., Ball, A., Robertson, I., and Ryan, U. (2017). Next Generation Sequencing uncovers within-host differences in the genetic diversity of Cryptosporidium gp60 subtypes. *International Journal for Parasitology*, 47(10-11):601–607.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829.

[Zhang et al., 2006] Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. (2006). Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences*, 89(13):5847–5851.

[Zhang et al., 2014] Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C., and Brown, C. T. (2014). These are not the K-mers you are looking for: Efficient online K-mer counting using a probabilistic data structure. *PLoS ONE*, 9(7).

[Zhong et al., 2018] Zhong, D., Koepfli, C., Cui, L., and Yan, G. (2018). Molecular approaches to determine the multiplicity of Plasmodium infections. *Malaria Journal*, 17(1):1–9.

[Zhou et al., 2008] Zhou, F., Olman, V., and Xu, Y. (2008). Barcodes for genomes and applications. *BMC Bioinformatics*, 9(1):546.

[Zielezinski et al., 2017] Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186.

[Ziemann, 2016] Ziemann, M. (2016). Accuracy, speed and error tolerance of short DNA sequence aligners. *bioRxiv*, page 053686.

[Zimin et al., 2013] Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21):2669–2677.

Appendix

Isolate	Post-PAGIT assembly length (Mb)	N50, Mean contig size, Max contig size	N-spaces	Genes Transferred by RATT
non-WGA				
UKP10	8.73	1.07 1.28 0.19	4366	2931
UKP11	7.41	0.9 1.17 0.24	4942	2284
UKP12	9.32	1.14 1.37 0.55	5627	2890
UKP13	9.02	1.1 1.33 0.26	6261	2612
UKP14	9.43	1.13 1.39 0.25	3156	3449
UKP15	9.4	1.14 1.38 0.11	1031	3723
UKP16	9.31	1.15 1.36 0.19	808	3740
WGA				
UKP90	9.15	1.11 1.34 1.14	71	3724
UKP94	9.15	1.11 1.34 1.14	78	3763
UKP95	9.14	1.11 1.34 1.14	93	3696
UKP97	9.16	1.11 1.34 1.15	146	3794
UKP98	9.14	1.11 1.35 1.14	513	3516
UKP99	9.15	1.11 1.34 0.87	50	3775
UKP102	9.11	1.11 1.34 0.88	67	3742
UKP103	9.16	1.11 1.34 0.88	100	3734
UKP104	9.11	1.11 1.34 0.88	70	3738
UKP106	9.15	1.11 1.34 1.14	74	3744
UKP107	9.18	1.11 1.35 1.15	199	3671
UKP118	9.16	1.11 1.34 1.14	77	3777
UKP119	0.48	0.07 0.1 0.06	317	173
UKP120	9.15	1.11 1.34 1.14	83	3757
UKP121	9.15	1.11 1.34 1.14	72	3769
UKP122	9.14	1.11 1.34 1.14	104	3618
UKP123	5.44	0.69 0.86 0.68	1994	2140
UKP124	9.01	1.08 1.34 1.13	892	3571
UKP125	9.15	1.12 1.34 1.14	112	3782
UKP126	6.66	0.79 1.03 0.83	2226	2576
UKP127	9.13	1.11 1.35 1.14	264	3705
UKP128	9.14	1.11 1.34 1.14	89	3781

Continued on next page

Table 1 – continued from previous page

Isolate	Post-PAGIT assembly length (Mb)	N50, Mean contig size, Max contig size	N-spaces	Genes Transferred by RATT
UKP129	9.15	1.11 1.34 1.14	177	3678
UKP130	9.12	1.11 1.34 1.14	89	3726
UKP131	9.14	1.11 1.34 1.14	64	3731
UKP132	3.01	0.41 0.54 0.38	1386	1197
UKP133	9.03	1.1 1.34 1.13	716	3637
UKP134	9.14	1.11 1.34 1.14	70	3777
UKP135	7.36	0.91 1.1 0.92	1368	2910
UKH51	9.08	1.12 1.33 1.14	50	3743
UKH55	9.09	1.11 1.33 1.14	52	3734
UKH56	9.08	1.11 1.33 1.14	55	3755
UKH57	9.09	1.11 1.33 1.14	82	3753
UKH58	8.75	1.06 1.28 1.09	1061	3550
UKH59	9.08	1.11 1.32 1.13	53	3755
UKH60	9.09	1.11 1.33 1.14	61	3748
UKH61	9.07	1.11 1.33 1.13	49	3749
UKH62	9.12	1.11 1.34 1.14	316	3771
UKH63	9.45	1.17 1.38 1.18	1299	3628
UKH64	9.45	1.17 1.38 1.18	1299	3628
UKH65	9.15	1.14 1.34 1.14	442	3744
UKH66	9.14	1.11 1.34 1.14	333	3762
UKH67	9.31	1.13 1.4 1.16	534	3704
UKH68	9.14	1.12 1.34 1.14	366	3757
UKH69	9.12	1.11 1.33 1.14	332	3772
UKH70	9.13	1.11 1.34 1.14	343	3764
UKH71	9.12	1.11 1.33 1.14	335	3773
UKH72	9.08	1.11 1.32 1.14	67	3752

Table 1: Extended assembly stats for IDBA-UD whole genome assemblies improved using PAGIT.

Isolate; Species; *gp60* subtype; BioProject number	Total No. base pairs sequenced (Mb)	Total No. base pairs mapped to reference genome (Mb)	Proportion of *Cryptosporidium* DNA	Fraction of reference genome covered	Average coverage of reference sequence	AT fraction of mapped reads
Pilot phase						
UKP2* *C.hominis* IbA10G2 PRJNA253834	521.277984	426.692177‡	0.8186	1	46.84	0.6785
UKP2† *C.parvum* IIaA19G1R2 PRJNA253836	510.081295	471.881392‡	0.9251	1	51.8	0.6792
UKH3† *C.hominis* IbA10G2 PRJNA253834	337.791948	305.024423§	0.903	0.98	34.71	0.6749
Main phase						
UKH4 *C.hominis* IaA14R3 PRJNA253838	2164.426378	1828.866488§	0.845	0.96	209.17	0.6272
UKH5 *C.hominis* IbA10G2 PRJNA253839	2182.317271	1765.458438§	0.809	0.96	201.92	0.6303
UKP3 *C.parvum* IIaA18G2R1 PRJNA253840	1703.132267	1514.828932‡	0.8894	0.99	166.42	0.6355
UKP4 *C.parvum* IIaA15G2R1 PRJNA253843	1967.147533	1751.97903‡	0.8906	0.99	192.48	0.636
UKP5 *C.parvum* IIaA15G2R1 PRJNA253845	288.922509	244.528063‡	0.8463	0.99	26.86	0.6767
UKP6 *C.parvum* IIaA15G2R1 PRJNA253846	1169.379989	954.176437‡	0.816	0.99	104.83	0.6773
UKP7 *C.parvum* IIaA17G1R1 PRJNA253847	795.715168	708.613859‡	0.8905	0.99	77.85	0.6364
UKP8 *C.parvum* IIdA22G1 PRJNA253848	1896.616473	1587.380412‡	0.837	0.98	174.39	0.632

Table 2: Bowtie2 mapping statistics of Dataset 1.2 for *C. parvum* and *C. hominis* taken from Hadfield *et al.*. *CsCl purified, †IMS purified, ‡mapped to *C. parvum* IowaII, §mapped to *C.hominis* TU502.

.1 Code Availability

Source code for a toolkit to calculate Gini and Gini-Granularity curves from coverage files can be found at https://github.com/ArthurVM/Read-Distribution-Toolkit.

Source code for VaNTA is available from https://github.com/ArthurVM/VaNTA. VaNTA was developed to run on laptops and personal computers, and therefore has very low computational requirements. This iteration of VaNTA is developed to run on UNIX systems via the command line. It will also be made available on the Galaxy platform.

Source code for BlooMine can be found at https://github.com/ArthurVM/BlooMine. BlooMine was developed and tested on a system running Ubuntu v18.04, with an Intel core i7 9th Gen and 32Gb DDR3 RAM. BlooMine was developed to run on laptops and personal computers, and therefore has very low computational requirements. In practical (but not theoretical) terms, there is no lower bound memory requirement. Multithreading is supported, and it is assumed that the user has access to at least 2 processing threads. Therefore, the minimum processor requirement can be taken as a modern dual core processor. BlooMine is developed to run on UNIX systems via the command line. It will also be made available on the Galaxy platform.

.2 BlooMine Benchmarking

Benchmarking statistics carried out against conventional pipelines used for investigating MOI are not provided due to the fact that BlooMine functions in a fundamentally different manner to standard read alignment software, such as Bowtie2 and BWA, which drive such approaches. Due to the often fairly strict "acceptable nucleotide-identity rates" utilised by such read aligners, using the simulated dataset which was used here would have been an improper method of comparison. Consequently, the text focuses on the fidelity of BlooMine in identifying target sequences with high levels of variance, rather than comparisons against existing tools and pipelines which are poor analogues, and the result of which would be too dependant on the type of dataset provided.

.3 Data Availability

The genome assemblies and annotations generated in Chapter 2 will be made available for public access on GenBank. The list of all Tandem Repeats reported by VaNTA in Chapter 3 on both the *P. falciparum* and *C. parvum* datasets can be found at https://github.com/ArthurVM/VaN

.4 Genome Sequencing

DNA extraction, purification and whole genome amplification (by MDA) of dataset 2 was carried out by myself at the Public Health Wales Cryptosporidium Reference unit in Singleton Hospital - Swansea (CRU). Sequencing of dataset 2 was organised by the CRU. One batch of 7 *C. hominis* genomes failed during sequencing, the reason for this is unknown. All genomes were derived from samples taken during UK outbreaks of Cryptosporidiosis between 2015-2018.

frontiers in Public Health

REVIEW
published: 11 December 2019
doi: 10.3389/fpubh.2019.00360

Direct Sequencing of *Cryptosporidium* in Stool Samples for Public Health

Arthur Morris[1†], *Guy Robinson*[2,3†], *Martin T. Swain*[1] *and Rachel M. Chalmers*[2,3*]

[1] *Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Aberystwyth, United Kingdom,*
[2] *Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea, United Kingdom,*
[3] *Swansea University Medical School, Swansea, United Kingdom*

The protozoan parasite *Cryptosporidium* is an important cause of diarrheal disease (cryptosporidiosis) in humans and animals, with significant morbidity and mortality especially in severely immunocompromised people and in young children in low-resource settings. Due to the sexual life cycle of the parasite, transmission is complex. There are no restrictions on sexual recombination between sub-populations, meaning that large-scale genetic recombination may occur within a host, potentially confounding epidemiological analysis. To clarify the relationships between infections in different hosts, it is first necessary to correctly identify species and genotypes, but these differentiations are not made by standard diagnostic tests and more sophisticated molecular methods have been developed. For instance, multilocus genotyping has been utilized to differentiate isolates within the major human pathogens, *Cryptosporidium parvum* and *Cryptosporidium hominis*. This has allowed mixed populations with multiple alleles to be identified: recombination events are considered to be the driving force of increased variation and the emergence of new subtypes. As yet, whole genome sequencing (WGS) is having limited impact on public health investigations, due in part to insufficient numbers of oocysts and purity of DNA derived from clinical samples. Moreover, because public health agencies have not prioritized parasites, validation has not been performed on user-friendly data analysis pipelines suitable for public health practitioners. Nonetheless, since the first whole genome assembly in 2004 there are now numerous genomes of human and animal-derived cryptosporidia publically available, spanning nine species. It has also been demonstrated that WGS from very low numbers of oocysts is possible, through the use of amplification procedures. These data and approaches are providing new insights into host-adapted infectivity, the presence and frequency of multiple sub-populations of *Cryptosporidium* spp. within single clinical samples, and transmission of infection. Analyses show that although whole genome sequences do indeed contain many alleles, they are invariably dominated by a single highly abundant allele. These insights are helping to better understand population structures within hosts, which will be important to develop novel prevention strategies in the fight against cryptosporidiosis.

Keywords: cryptosporidium, public health, genotyping, genome, sequencing, multiplicity of infection

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control, Sweden

Reviewed by:
Michael Arrowood, Centers for Disease Control and Prevention, United States
Lihua Xiao, South China Agricultural University, China
Michelle Power, Macquarie University, Australia

***Correspondence:**
Rachel M. Chalmers
Rachel.Chalmers@wales.nhs.uk

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *31 July 2019*
Accepted: *13 November 2019*
Published: *11 December 2019*

Citation:
Morris A, Robinson G, Swain MT and Chalmers RM (2019) Direct Sequencing of Cryptosporidium in Stool Samples for Public Health. Front. Public Health 7:360. doi: 10.3389/fpubh.2019.00360

INTRODUCTION

The parasite *Cryptosporidium* is a protozoan that occurs worldwide, and can cause the diarrheal disease cryptosporidiosis in humans and animals (**Figure 1**). The life cycle of *Cryptosporidium* (**Figure 2a**) (1) is completed within a single host. Both the asexual phase, and the production of thin-walled oocysts that enable autoinfection, mean the numbers of parasites are increased from possibly single figures in the initial infection, to result in clinically significant infections and the shedding of vast numbers of oocysts in feces (2). These shed oocysts have thick walls, conferring protection for the four infective sporozoites contained within, and enabling long-term survival, environmental transmission, and resistance to commonly used disinfectants including chlorine (3, 4). This means that, in addition to the variety of hosts that act as direct sources of infection (**Figure 1**; **Table 1**), contaminated food, water, or environmental vehicles are involved in transmission and need to be considered and investigated for effective disease control and prevention of outbreaks of cryptosporidiosis (5).

Human cryptosporidiosis is usually a gastrointestinal disease, although there is some evidence for respiratory cryptosporidiosis in some populations (6). Symptoms ranging from mild to severe depending upon a number of factors, including the host's age, immune status, nutrition, genetics, and the site of infection, as well as the infecting species and variant of *Cryptosporidium* (7–9). Clinical symptoms include diarrhea, abdominal pain, vomiting, nausea, and low-grade fever, which, although prolonged (2 weeks is not unusual) are generally self-limiting in immune competent hosts. However, infection can be more problematic and even life-threatening in some severely immunocompromised individuals, and in malnourished young children (10). There are few options for treatment or prevention. Recent studies have shown that in some low-resource countries, where access to safe drinking water, sanitation, hygiene, and healthcare is often poor, *Cryptosporidium* is one of the most important causes of moderate-to-severe diarrheal disease and death in young children (11, 12). Furthermore, long-term effects of infection such as malnutrition, growth, and cognitive deficits have been described, highlighting the socio-economic impact on the adverse outcomes of infection (10). A vicious cycle of malnutrition and diarrhea can become established with detrimental effects on these societies (13). For these reasons, *Cryptosporidium* was included in the World Health Organization's Neglected Diseases Initiative in 2004 (14), which served to raise awareness of the need for international and national investments in prevention and control.

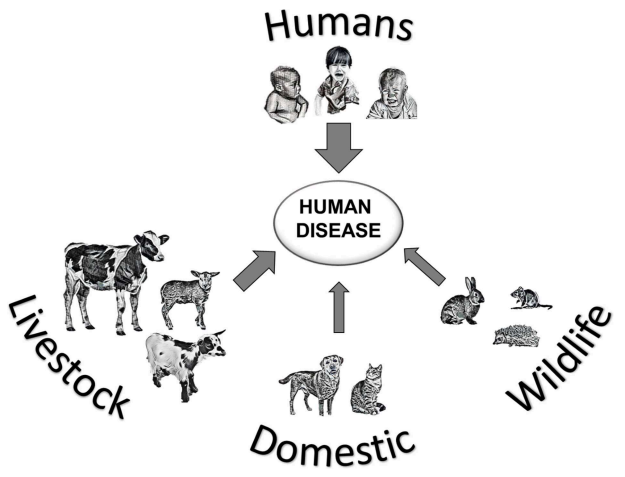


FIGURE 1 | Transmission of *Cryptosporidium* spp. leading to human cryptosporidiosis, arrow thickness represents likely global importance of source hosts.

Thirty-nine species of *Cryptosporidium* have been described at the time of writing (**Table 1**), but not all cause human disease. The vast majority of human cryptosporidiosis is caused by the zoonotic species *Cryptosporidium parvum* or anthroponotic *Cryptosporidium hominis,* with multiple variants that can cause varying severity of symptoms. The diagnostic target of laboratory tests, and those used to detect *Cryptosporidium* in water, is the oocyst, using stained microscopy or immunologically-based assays, or the sporozoite DNA. Routinely applied tests are not able to differentiate species, and molecular methods are needed to investigate true relationships between infections and contaminants and thus elucidate the complex transmission of *Cryptosporidium*. A range of samples need to be investigated, from feces (e.g., stools, diapers, livestock dung, manure, slurry, runoff, and wild life droppings), to contaminated water and food, but these present challenges to detection and genotyping. At present, amplification by culture is not an option in this context, and finding oocyst targets, which may be in low concentration in the sample matrix, can be a hit-and-miss affair. Recent advances in molecular methods generally, and particularly in genomics, have increased the amount of data available particularly on the major pathogenic *Cryptosporidium* species (**Table 1**). Continued generation and accessibility of genomic data will potentially improve the public health response to cryptosporidiosis by identifying new targets for incorporation into diagnostic and genotyping assays (15). Putative virulence and host adaption factors have been proposed (16), and potential chemotherapeutic targets and vaccine candidates are being sought (10, 17) and identified [e.g., (18)].

INTRODUCTION TO *CRYPTOSPORIDIUM* GENOTYPING

To identify *Cryptosporidium* species, genotyping was undertaken initially using conventional PCR combined with either restriction fragment length polymorphism (RFLP) or Sanger sequence analysis, most commonly of the 18S rRNA gene (19). The 18S rRNA gene includes conserved regions interspersed with highly polymorphic regions and is currently considered to provide the definitive sequences for discriminating *Cryptosporidium* species. It is present in multiple copies (5 per sporozoite; 20 per oocyst) facilitating the development of sensitive assays, which is especially important for testing samples such as water where small (but potentially significant) numbers of oocysts

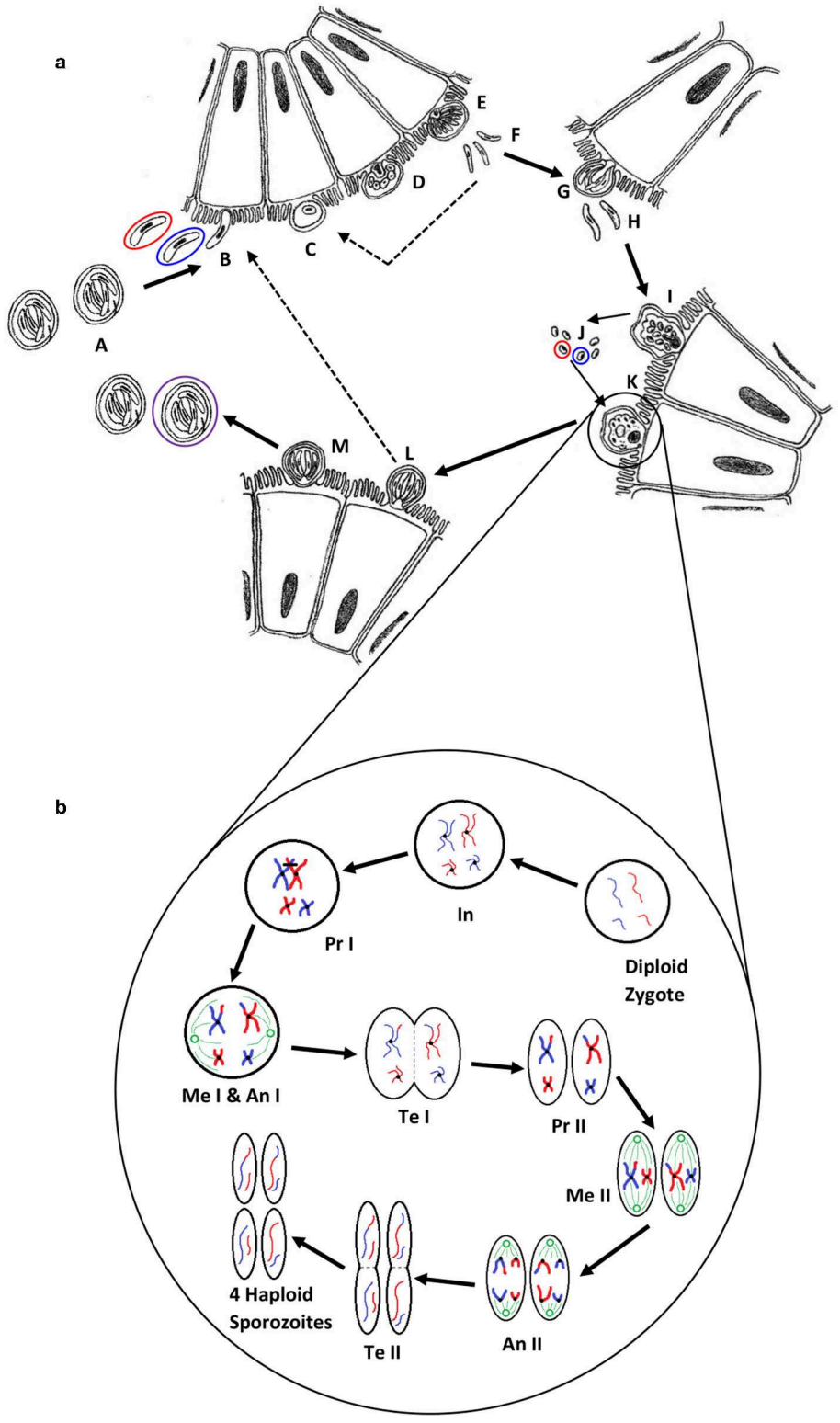


FIGURE 2 | (a) The life cycle of *Cryptosporidium* (1). Oocysts (A) are ingested by the host, most likely as a mixed population of different genotypes; haploid sporozoites (B) (variants are represented by red and blue) excyst and invade the brush border of epithelial cells; each sporozoite develops into a haploid trophozoite with a prominent nucleus (C); the trophozoite undergoes merogony by mitosis to form a type I meront (D,E); up to eight haploid merozoites (F) are released, invade

(Continued)

FIGURE 2 | another cell and undergo merogony again to form either further type I meronts (dotted line) or type II meronts (G), which release four haploid merozoites (H) and form either microgamonts (I) that become multinucleate and mature to form multiple haploid microgametes (J) by mitosis, or a haploid macrogamont (K). Microgamonts are released and potentially each fertilize a macrogamont to form a diploid zygote which undergoes sporogony by meiosis to produce either thin-walled oocysts (L) containing four haploid sporozoites that can autoinfect the host (dotted line), or thick-walled oocysts (M) that are shed in the feces ready to transmit four haploid sporozoites to a new host (the purple circle represents an oocyst that is the product of fertilization between the red and blue genotypes). **(b)** A simplified schematic of genetic recombination in *Cryptosporidium*, potentially generating variation between sporozoites within oocysts. In a mixed infection population, different fertilization scenarios potentially occur—between the same genotypes (resulting in identical daughter sporozoites) or between different genotypes, as in the example shown, that result in a variety of outcomes depending on the random genetic exchange, or lack of, that occurs during meiosis. For simplicity only two example chromosomes are shown with DNA from different genotypes represented by blue and red. The diploid zygote contains duplicate pairs of chromosomes, one set from each parent cell; during interphase (In) the DNA in each chromosome is replicated to produce two identical sister chromatids held together with a centromere; in prophase I (Pr I) the chromosomes start to condense and pair up with the homologous chromosome from the other parent cell, and cross-over can occur resulting in an genetic exchange; during metaphase I (Me I) the paired chromosomes line up along the center of the cell and microtubules connect the centromeres to the centrosomes (shown in green); during anaphase I (An I) each complete set of chromosomes (still paired as sister chromatids) are pulled toward each centrosome—the chromosomes from either parent are randomly combined at this phase introducing a further opportunity for recombination (a blue and a red chromosome are drawn to each centrosome in this example); in telophase I (Te I) the chromosomes start to unravel and cytokinesis starts to split the cell into two, resulting in two haploid cells; in prophase II (Pr II) the chromosomes condense again; during metaphase II (Me II) the chromosomes line up along the center of the cells and microtubules connect the centromeres to the centrosomes; this time during anaphase II (An II) the sister chromatids are separated and pulled apart toward the centrosomes, creating new daughter chromosomes; finally in telophase II (Te II) the chromosomes unravel and cytokinesis starts to split the cells, which in the case of this example due to the crossover event in prophase I, results in four genetically different haploid sporozoites. Depending upon whether random genetic exchanges take place between chromosomes from different genotype parents (either in prophase I or anaphase I) the resulting haploid sporozoites can either be all different, two pairs of identical sporozoites that are different from each parent, or two pairs of identical sporozoites that are the same as the two parents.

may be present. Species-level genotyping has provided improved understanding of human epidemiology in some countries, streamlined by the use of real-time PCR (see below). DNA extraction methods from stool and gene targets have been reviewed in detail by Khan et al. (17).

Beyond the species-level, Sanger sequencing part of the *gp60* gene is most commonly used for further discriminating some *Cryptosporidium* species, including *C. parvum* and *C. hominis* (19–21). The *gp60* gene is hypervariable both between and within *Cryptosporidium* species, and the presence of a highly variable serine repeat region in most species enables further discrimination (19). For nomenclature of *gp60* subtypes, the reader is referred to a review of molecular epidemiologic tools by Xiao and Feng (19). The use of this locus as a subtyping marker has been questioned as it is associated with host cell invasion, and therefore can be considered a virulence factor under selective pressure. Nevertheless, as shown below, it may still be an appropriate target for interrogation as a phenotype determining biomarker. Another issue arises from the use of a single locus; this may not be appropriate due to the genetic recombination that occurs within *Cryptosporidium* populations during the sexual stage of the life-cycle (**Figure 2b**). Whilst not likely or expected between different species, this may occur in populations of mixed subtypes of the same species (22–25). This necessitates the investigation of multiple loci to reveal a more accurate estimate of diversity and population structure (19, 26), and would confer greater discrimination for characterization of isolates (26, 27).

The reality is that genotyping tools are not currently widespread in their application for public health purposes and in most countries *Cryptosporidium* is under-diagnosed and isolates are not characterized (28). In low-resource countries where surveillance data are lacking, research studies have found that *C. hominis* or human-adapted *C. parvum* subtypes predominate (29, 30). *C. parvum* can also be the main species detected in some urban settings with no animals close to residences, further suggesting anthroponotic rather than zoonotic transmission (29). These findings indicate that measures to improve sanitation and hygiene would have greatest impact in these settings. Not only is there a high prevalence of *Cryptosporidium* in these populations, but there is also greater diversity within these species, especially noticeable in *C. hominis*, than is seen in industrialized countries (17, 31).

Genotyping in *Cryptosporidium* Surveillance and Outbreaks

The aim of genotyping in the public health context is to understand transmission and to improve the detection resolution, investigation, and interpretation of waterborne, zoonotic, person-to-person, and foodborne outbreaks. The potential impact lies in:

- Identifying the *Cryptosporidium* species and subtypes that most commonly cause human cryptosporidiosis, and their demographic and temporal-spatial distribution
- Monitoring for the emergence of new species and subtypes in human infection
- Improving detection, investigation, and interpretation of outbreaks
- Increasing the sensitivity of epidemiological investigations to identify links and risk factors, and identify the source of outbreaks and contamination.

In most countries, routine surveillance captures *Cryptosporidium* as an organism, but not species. Where genotyping is used to inform public health, it is mainly in industrialized countries but the framework varies. For example, in England and Wales, clinical diagnostic laboratories have been sending *Cryptosporidium*-positive stools for genotyping for many years, both for molecular surveillance and for outbreak investigations, and most diagnostic stools are genotyped (5, 32). In France, testing for *Cryptosporidium* is not part of routine diagnostic parasitological testing, but a national network of sentinel laboratories was established to test for and genotype new and outbreak cases of cryptosporidiosis (ANOFEL Cryptosporidium

TABLE 1 | *Cryptosporidium* species, their major hosts, oocyst dimensions, reported human infectivity and availability of genome data.

Cryptosporidium species	Mean oocyst dimensions (μm)	Major host(s)	Infections reported in humans	Genomes available (accession number)
C. alticolis	5.4 × 4.9	Voles	No	No
C. apodemi	4.2 × 4.0	Mice	No	No
C. andersoni	7.4 × 5.5	Cattle	Yes (rarely)	PRJNA354069
C. avium	6.3 × 4.9	Birds	No	No
C. baileyi	6.2 × 4.6	Birds	No	PRJNA222835
C. bovis	4.9 × 4.6	Cattle	Yes (rarely)	No
C. canis	5.0 × 4.7	Canids	Yes (occasionally)	No
C. cuniculus	5.6 × 5.4	Lagomorphs, Humans	Yes (occasionally)	PRJNA315496
C. ditrichi	4.7 × 4.2	Mice	Yes (rarely)	No
C. ducismarci	5.0 × 4.8	Tortoises	No	No
C. erinacei	4.9 × 4.4	Hedgehogs	Yes (rarely)	No
C. fayeri	4.9 × 4.3	Marsupials	Yes (rarely)	No
C. felis	4.6 × 4.0	Felids	Yes (occasionally)	No
C. fragile	6.2 × 5.5	Toads	No	No
C. galli	8.3 × 6.3	Birds	No	No
C. homai	Data not available	Guinea Pigs	No	No
C. hominis	4.9 × 5.2	Humans	Yes (commonly)	PRJEB10000 PRJNA13200 PRJNA252787 PRJNA222836 PRJNA222837 PRJNA307563 PRJNA253838 PRJNA253839 PRJNA253834
C. huwi	4.6 × 4.4	Fish	No	No
C. macropodum	5.4 × 4.9	Marsupials	No	No
C. meleagridis	5.2 × 4.6	Birds, mammals	Yes (occasionally)	PRJNA222838 PRJNA315503 PRJNA315502
C. microti	4.3 × 4.1	Voles	No	No
C. molnari	4.7 × 4.5	Fish	No	No
C. muris	7.0 × 5.0	Rodents	Yes (rarely)	PRJNA32283 PRJNA19553
C. occultus	5.2 × 4.9	Rodents	Yes (rarely)	No
C. parvum	5.0 × 4.5	Mammals	Yes (commonly)	PRJNA144 PRJNA320419 PRJNA439211 PRJNA253848 PRJNA253843 PRJNA253845 PRJNA253836 PRJNA253840 PRJNA253846 PRJNA253847 PRJNA320419 PRJNA315506 PRJNA437480 PRJNA315504 PRJNA315508 PRJNA315507 PRJNA315505 PRJNA13873
C. proliferans	7.7 × 5.3	Rodents, maybe Equids	No	No
C. proventriculi	7.4 × 5.7	Birds	No	No
C. rubeyi	4.7 × 4.3	Squirrels	No	No
C. ryanae	3.7 × 3.2	Cattle	No	No
C. scrofarum	5.2 × 4.8	Pigs	Yes (rarely)	No
C. serpentis	6.2 × 5.3	Reptiles	No	No
C. suis	4.6 × 4.2	Pigs	Yes (rarely)	No
C. testudinis	6.4 × 5.9	Tortoises	No	No
C. tyzzeri	4.6 × 4.2	Rodents	Yes (rarely)	No
C. ubiquitum	5.0 × 4.7	Mammals	Yes (occasionally)	PRJNA534291 PRJNA315509 PRJNA315510
C. varanii	4.8 × 4.7	Reptiles	No	No
C. viatorum	5.4 × 4.7	Humans, Rodents	Yes (occasionally)	PRJNA492837
C. wrairi	5.4 × 4.6	Guinea Pigs	No	No
C. xiaoi	3.9 × 3.4	Sheep, Goats	No	No

National Network, 2010). The Netherlands, Sweden and Scotland also use sentinel laboratories to provide sporadic and outbreak samples for genotyping in reference laboratories (28). In the USA, the Centers for Disease Control and Prevention is developing CryptoNet, a molecular-based surveillance system aimed at the systematic collection and molecular characterization of isolates using 18S rDNA PCR-RFLP and *gp60* sequencing (https://www.cdc.gov/parasites/crypto/cryptonet.html). In Germany, Norway, Spain, Ireland, Northern Ireland, Australia, and New Zealand, *Cryptosporidium* genotyping has been used in epidemiological research projects and/or for supporting outbreak investigations (28, 33, 34), while the focus in Asia, Africa, and South American countries has been on molecular epidemiological research (29, 30, 35).

Molecular surveillance data in the United Kingdom (UK) for example has shown that >95% of cases are caused by *C. hominis* or *C. parvum*. Two seasonal peaks in cases occur, with *C. parvum* consistently causing the majority of cases in spring and *C. hominis* predominating in the autumn peak, with much higher rates of foreign travel also reported during this second period (32, 36–38). A similar temporal pattern has been reported in New Zealand (39), but contrasts with the epidemiology in Ireland, where there is no autumn peak and *C. parvum* predominates all year (33, 40). This is likely due to the highly rural socio-geography of Ireland and the greater potential of zoonotic transmission, a feature also seen in rural regions of Great Britain (36, 38). In the UK, the highest incidence of cryptosporidiosis is in children under 5 years, with a second smaller peak in adults in their 20s and 30s; in England and Wales in the period 2000 to 2003, *C. hominis* predominated in infants and the 30–39 year age group (32), and in children <10 years and adults in the period 2004 to 2006 (37), suggesting transmission between children and caregivers. In Ireland, where *C. parvum* predominates, the adult peak does not appear but this may be a testing bias (33, 40).

Although the sentinel surveillance in France is not wholly representative of the French population due to the structure of the network resulting in the inclusion of a higher proportion of hospitalized cases (70%), particularly over-representing the proportion of HIV-infected patients, certain trends are noticeable (ANOFEL Cryptosporidium National Network, 2010). There appears to only be a late summer/autumn peak each year, but the case numbers per month were too low to determine any species-related seasonality. However, *C. parvum* was more prevalent each year compared to *C. hominis* (54.2 vs. 36.5%) and with the remaining 9.4% representing other species (particularly *C. felis*). The seemingly high number of unusual species were mainly found in the over-represented immunocompromised patients (82.8%), which may explain their higher prevalence than in the UK for example.

In the Netherlands, only an autumn peak in case numbers is present in surveillance data, and the predominant species infecting people does not seem to be stable between years. One study undertaken between 2003 and 2005 reported a higher prevalence of *C. hominis* (70.3%) than *C. parvum* (18.7%), with 9.9% cases having both species, and a single case of *C. felis* (41). The infecting species was significantly associated with patient age, with children (aged 0–9 years) more frequently infected with *C. hominis* and adults (over 25 years old) more frequently with *C. parvum* (41). However, over a 3-year study from April 2013, *C. parvum* was most prevalent in years one and two, but in year three (April 2015 to March 2016) *C. hominis* predominated and cases did not decline toward the winter as they had done in previous years (42). Whether these apparent shifts were a function of fluctuating participation in the sentinel scheme or another reason is not known. In England and Wales apparent shifts have also been seen; from 2000 to 2003 the ratio of *C. parvum*:*C. hominis* nationally was close to 1, but in the period 2004–2006 it was 1:1.5, most noticeable in 2005 when it was 1:2.3 and major *C. hominis* outbreaks may have influenced the distribution (37). The UK and the Netherlands both reported an excess in cases of *C. hominis* with similar epidemiology in the latter part of 2015, and despite *gp60* sequencing identifying subtype IbA10G2 and enhanced surveillance, no explanation was found. This was the second time an international *C. hominis* excess had been reported; in the late summer of 2012 the Netherlands, UK, and Germany reported similarly unexplained increases (43).

In the United States (US) national cryptosporidiosis surveillance through CryptoNet is in its infancy, but there seems to be a high diversity of *Cryptosporidium* species and subtypes causing human cryptosporidiosis compared to other industrialized nations (19). While *C. hominis* and *C. parvum* cause the majority of cases, unusual species such as *C. ubiquitum* and the chipmunk genotype are also seen, particularly in rural areas and may suggest an important role of wildlife in transmission, either directly or through drinking untreated water (19). While general surveillance of *Cryptosporidium* species and genotypes in the US is still fairly new, outbreak surveillance has been carried out for many years through the National Outbreak Reporting System (NORS). Analysis of 444 outbreaks of cryptosporidiosis between 2009 and 2017 demonstrated most were in the autumn and caused mainly by waterborne and person-person transmission (44). Molecular data are available for some of the outbreaks on the NORS website https://wwwn.cdc.gov/norsdashboard/. Genotyping data for 131/178 (74%) outbreaks in the same time period in England and Wales showed 69 were caused by *C. parvum* (which caused all animal and environmental contact and food-borne outbreaks, and a minority of recreational water outbreaks), 60 were caused by *C. hominis* (most of the recreational water and all person-to-person spread outbreaks) and in two outbreaks both species were identified (5). Both *C. parvum* and *C. hominis* caused drinking waterborne outbreaks. *Gp60* sequencing established linkage between cases and suspected sources in nine animal contact, three swimming pool, and one drinking water outbreaks (5). Thus, the public health benefits of identifying infecting species and subtypes lie in the ability to identify and strengthen epidemiologic links between cases, and in indicating possible exposures and sources to inform outbreak management (5). However, the ability to differentiate zoonotic and anthroponotic *C. parvum* routinely in all cases would be useful.

Identification by sequencing has established that unusual species of *Cryptosporidium*, previously considered without zoonotic potential, can infect people. Enhanced surveillance has provided some understanding of the transmission of these infections. In the UK, cases with unusual species often reported zoonotic exposures; contact with unwell pets was a significant association, and in particular, contact with cats was reported by significantly more cases with *C. felis* (45). Genotyping *C. ubiquitum* from patients in the US revealed mainly the rodent-adapted subtype families (XIIb-XIId) in contrast to the UK where infections were mainly the ruminant-adapted XIIa subtype family (19, 46).

The potential for outbreaks is not limited to *C. parvum* and *C. hominis*. In 2007 *Cryptosporidium cuniculus* (previously rabbit genotype) was first identified in a patient during routine molecular surveillance in the UK (47). The following year an investigation into a drinking water quality incident in England established that oocysts detected in treated water were *C. cuniculus*. Soon afterwards, primary and secondary *C. cuniculus* cases appeared in the supplied local population, with the same *gp60* subtype, VaA18 (48). Importantly, matching the *Cryptosporidium* isolated from the drinking water, the remains of a rabbit discovered in a chlorine contact tank, and the case samples provided strong evidence for waterborne transmission. This was the first outbreak reported to have caused cryptosporidiosis where the etiological agent was a species other than *C. parvum* or *C. hominis*, and established *C. cuniculus* as a human pathogen. It re-enforced the importance of protecting water supplies not only from livestock and sewage contamination, but also from wildlife.

Sequencing of the *gp60* gene has identified changes in the circulation of predominant subtypes, and the emergence of virulent subtypes. *C. hominis* IbA10G2 continues to predominate in northern Europe, but in the US in 2007, 40 of 57 sporadic cases from four states were a rare subtype, IaA28R4, with IbA10G2 accounting for just eight cases (49). Since 2013, IaA28R4 has been displaced by IfA12G1R5 as the predominant *C. hominis*

genotype in the US associated with both sporadic and outbreak cases (19). In Africa and Asia there is greater variation in *C. hominis* subtypes. For example, in Bangladesh where *C. hominis* is the most common species (>95% of cases) and the seasonality demonstrates a summer peak corresponding to the monsoon, *gp60* analysis revealed 13 different subtypes over a 2 year period (31). Some, for example IaA18R3 and IbA9G3 were present year on year, but other subtypes predominated in some years and disappeared in subsequent years (e.g., IdA15G1 was very common in 2015, but not in 2016 when IaA19R3 and IeA11G3T3 were dominant), indicating a dynamic and frequent transmission (31).

In Europe there is more variation among *C. parvum* than *C. hominis*, although IIaA15G2R1 and IIaA17G1R1 are often (but not always) the most common (5, 19, 50). Genotyping has increased our capacity to detect, investigate and interpret outbreaks. For example, in 2012, *C. parvum* IIaA15G2R1 was used as part of the case definition in an analytical study to investigate a large outbreak (>300 cases) across England and Scotland. A statistically significant association was identified with consumption of pre-cut, bagged mixed salad leaves from a specific national retailer (51). Also in 2012, an outbreak in schoolchildren was associated with a visit to a holiday farm in Norway (52). Genotyping of isolates from cases and potential animal sources on the farm revealed the same rare subtype of *C. parvum*, IIaA19G1R1, in the cases, lambs and goat kids (52). The same holiday farm was also involved in a previous outbreak in 2009 and the same subtype was identified retrospectively, suggesting that in the absence of newly introduced subtypes, existing subtypes can be stable and circulate on the farms for many years (52).

Although *gp60* sequencing has played an important role in refining epidemiological investigations, it is somewhat surprising that there is no standardized multilocus genotyping scheme for *Cryptosporidium* surveillance and outbreaks. Additionally, the lack of suitable markers has hampered our understanding of the main transmission pathway (zoonotic or anthroponotic) of *Cryptosporidium* species and subtypes. As discussed in this paper, genomics has an important role to play in the identification of new markers and the development of a MLG scheme, and the aspiration is that application would eventually become nationally systematic.

Multilocus Genotyping

Currently multilocus genotyping (MLG) is mainly applied to study the population structure of *Cryptosporidium* spp. with few reports describing its utility in surveillance or outbreaks. One example is an investigation into a Swedish swimming pool outbreak in 2002, where multilocus genotyping revealed two concurrent *C. parvum* outbreaks, with different subtypes linked to the use of either the indoor or outdoor pool, indicating multiple contamination events (53). In England, the epidemiological association of *C. parvum* cases with a drinking water supply was strengthened by MLG (54). However, more often investigations have explored the population structure and biology of *Cryptosporidium*.

In 2015, Widmer and Caccio investigated the relationship between sequence and length polymorphism within a set of biomarkers in the *Cryptosporidium* genome. They compared genetic distances of sequence and length polymorphism, finding that there was a weak correlation between the two distance measures. Their results also indicated that the resolution of *Cryptosporidium* population structure was dependent on the genotyping method used (55). Differences in varying extents of host-associated (56, 57) and geographical segregation (24, 58–60), and the extent of panmixia vs. clonality, depending on the population studied (21), have been reported. For example, in Spain, *C. parvum* in cattle herds was reported to show a panmitic population structure contrasting with sheep where *C. parvum* populations appeared more clonal (19, 61, 62). This may have been a function of the predominance of *C. parvum gp60* subtype family (IId) in sheep in the study region of Northeastern Spain (63) as IId has been reported to be clonal in other regions/countries (64).

Pamixia in *Cryptosporidium* spp. may reflect the increased potential for genetic recombination between more diverse isolates than is available in these supposed clonal populations of parasites. The presence of mixed populations with multiple alleles is the driving force of increased variation and the emergence of new subtypes due to recombination events (65–67). In some studies, for example in Scotland *C. hominis* populations have shown clonality (58), but in a cohort of children in Peru, genetic recombination was detected in some *C. hominis* IbA10G2 samples using MLST of 32 polymorphic loci, despite the overall clonality of the *C. hominis* population (65).

However, with the vast majority of *C. hominis* isolates in many areas, including northern Europe and Australia, demonstrating the dominant IbA10G2 (21) the potential for recombination with other more diverse subtypes may be reduced through lack of exposure in those regions. In contrast, the wide variety of different *C. parvum* subtypes usually present in local geographic areas make mixed populations more likely. This has been suggested in a study of the global population structures of both *Cryptosporidium* species, where samples from Uganda showed similar panmitic population structures, contrasting with *C. hominis* samples from the United Kingdom and *C. parvum* from New Zealand which showed much more clonal population structures (68). The authors suggest that both *C. parvum* and *C. hominis* population structures appear to be shaped by local or host-related factors rather than being species-specific (68). This was borne out by a study in Sweden that applied a nine-locus SNP-based method to differentiate *C. hominis* IbA10G2 and grouped 44 isolates, from 12 countries (including 7 non-European), into 10 MLSTs with known epidemiologically-linked samples clustering together; geographical clustering was not obvious, however the numbers of isolates from each country were small (69). In the USA, the emergence and spread of *C. hominis* IaA28R4 was investigated by sequencing eight loci (67). Of 95 *C. hominis* samples (62 IaA28R4 samples) from four states, the sequence diversity identified two clear sub-populations separated geographically between Ohio and three southwestern states, and suggested that the Ohio subpopulation was a descendant of the subpopulation in the southwestern states. Furthermore,

genetic recombination was seen to occur in IaA28R4 isolates and was likely an important factor in its emergence (67), a finding supported by a comparative study of the genome along with the previously dominant IbA10G2 subtype (70).

For disease surveillance and outbreak investigations, there is a need to establish a common multilocus genotyping scheme to track the sources and spread of infection. In a review published in 2012, Robinson and Chalmers reported that different combinations of loci and methods of analysis had been used, with very few groups using comparable loci (27). For public health purposes it is desirable to have consensus to enable cross-boundary comparisons and investigations and track international spread. An initiative funded by EU COST Action FA1408 "A European Network for Foodborne Parasites: Euro-FBP" (http://www.euro-fbp.org) enabled a workshop to be held between 23 scientists and experts in public and animal health from 12 European countries and the USA on *Cryptosporidium* genotyping (71). The participants discussed the need for, and potential directions of, a standardized typing scheme specifically for surveillance and outbreak investigations. There was general agreement that a robust multilocus genotyping scheme should be developed through collaborative laboratory studies, to standardize a method for meaningful interpretation of genotype occurrence and distribution trends, and where possible incorporate into national surveillance programs (71). To achieve this multiple markers spread, sufficiently across the genome, are required. The recent generation of genome data facilitates the identification of markers that show potential to be combined for MLG investigations specifically for surveillance and outbreak investigations (15).

WHOLE GENOME SEQUENCING

While we aspire to using WGS routinely in public health investigations of *Cryptosporidium* cases in the way it is applied to some bacterial pathogens (72–74), the reality is that this is still a way off. Direct sequencing would provide timely investigation of public health incidents, but it poses a challenge for this parasite; it is difficult to culture and bioinformatics pipelines have not been validated for public health purposes as *Cryptosporidium* has suffered from lack of prioritization in genomics programs.

The first technical problem is the amount of DNA that is required. Although this varies depending on the technology used, for example, the Nextera XT DNA kits that have been used in several publications require 1 ng of DNA, and as each oocyst contains 40 fg of DNA it means that 2.5×10^4 oocysts are required without losses and in a practical volume (75). To generate sufficient DNA, oocysts may be propagated through animals, but *Cryptosporidium* populations have been shown to change through natural host-based preferential selection of individual subtypes or further recombination into new subtypes. For example, the "isolate" that provided the first reference *C. hominis* genome in 2004 (TU502) was subsequently serially propagated in gnotobiotic pigs over many years resulting in a different subtype in 2012, which was likely due to the original population being overgrown by another contaminating isolate (76). Additionally, the availability of host animals appropriate to the *Cryptosporidium* species in question (**Table 1**), and the ethics, time and cost resources that are associated with propagation are prohibitive. As propagating oocysts is not a practical solution, obtaining enough clinical sample is the next hurdle, as the volume of stools often submitted is very small. Purity is also a challenge because feces is the starting point, so *Cryptosporidium* DNA is overshadowed by non-target DNA from the biome and host. Lack of purity has been overcome by the combination of several techniques including harvesting by flotation, further purifying by immunomagnetic separation and using the natural chlorine resistance of *Cryptosporidium* oocysts to surface-sterilize them with bleach (75, 77).

The sufficiency of available *Cryptosporidium* DNA has also been addressed through the use of whole genome amplification (WGA) techniques, which now mean that very small amounts of DNA, even from single oocysts, can be used for genome sequencing (77, 78). Guo et al. used WGA to enrich *Cryptosporidium* DNA from six discrete species/genotypes extracted from 24 human and animal fecal samples (77). The results were encouraging, showing that *Cryptosporidium* DNA was significantly enriched, allowing for coverage of $>$ 94% of the genome (77). This ability to whole genome sequence from very low numbers of oocysts is a development that may help when investigating environmental samples and other transmission pathways. Additionally, it may also alleviate problems encountered when whole genome sequencing a mixed population of oocysts. The concern that WGA could result in higher numbers of errors introduced into the genome sequence due to the fidelity of the enzymes used is also unfounded. The presence of four sporozoite genomes in a single oocyst helps, as any errors introduced in the first cycle are unlikely to occur at exactly the same place in more than one genome, so subsequent copies from the other genomes (containing the correct sequence) should overshadow any errors. Although WGS technology has developed and some of the technical hurdles have been overcome to enable direct sequencing (75, 77, 78), we are still not at a point where it can be used to inform in real-time for meaningful surveillance or during outbreak investigations. Aside from technical and resource issues, the lack of user-friendly, validated pipelines specifically designed to generate data in a form that is useful to public health practitioners during the management of incidents, make direct whole genome sequencing currently impractical. Nevertheless, genomic data are being used for biomarker discovery and to understand genetic diversity in parasite populations in different settings. These developments are described below, and arise from the progression of *Cryptosporidium* whole genome sequencing and assembly over the last two decades.

Progression of Whole Genome Sequencing and Assembly

Attempts to sequence the genome of *Cryptosporidium* began in the early 2000s. Initial attempts involved cloning sheared fragments into plasmid vectors and Sanger sequencing. This approach resulted in $>$ 9x coverage of the genome and yielded a fragmented assembly of 221 contigs of length $>$ 5 kbp (79). A more advanced sequencing project was undertaken to resolve gaps, using large *C. parvum* fragments contained within lambda DASH II libraries, and sequence missing DNA using a primer

walk strategy (79). The completed genome of *C. parvum* (Iowa II) along with a preliminary annotation was first published in 2004 by Abrahamsen et al. (80) who passaged oocysts through an animal donor to produce enough parasitic material for the extraction and purification of sufficient amounts of DNA. A random shotgun sequencing approach was used, which yielded a complete genome with coverage of 13x over 18 large contigs (80) and was shortly followed by the publication of the first draft genome of *C. hominis* (TU502) in late 2004. However, this *C. hominis* genome proved to be much more fragmented than that of *C. parvum*, resulting in a sequence consisting of 1,422 contigs (81).

In 2015, the *C. parvum* (Iowa II) reference genome was reassembled and reannotated, and a new *C. hominis* reference genome (UdeA01) published (82). The updated assembly resolved all eight chromosomes from the 18 scaffolds in the previous genome, representing the first chromosome level assembly of *C. parvum*. The reannotation effort increased the number of putative genes from 3807 to 3865 for *C. parvum* Iowa II, and predicted the presence of 3819 genes in *C. hominis* UdeA01 (82). In 2016, Ifeonu et al. reassembled and reannotated the *C. hominis* TU502 genome, along with producing new draft genomes of human isolated *C. hominis* (UKH1) and *C. meleagridis* (UKMEL1) along with the avian species *Cryptosporidium baileyi* (TAMU-09Q1) (83). The *C. hominis* TU502 genome proved to be a considerable improvement on the previous 2004 version, being much more complete, and reducing the number of contigs down to 119. Annotation was facilitated by the RNAseq data generated from the oocyst stage of both *C. hominis* and *C. baileyi*, predicting the presence of 3745 protein coding genes in *C. hominis* TU502 and 3765 in *C. hominis* UKH1 (83).

As can be seen in **Table 2**, there is little difference between the genomes of *C. parvum* and *C. hominis*. They exhibit 95–97% DNA sequence identity; with 11 protein-coding sequences identified only in *C. hominis* and 5 in *C. parvum*, and no large indels or rearrangements apparent (84). The high conservation in the *C. hominis* genomes generated from European samples compared to the much more polymorphic *C. parvum* does not appear to be expressed in general observations on structure and base representation as illustrated in **Table 2**, suggesting that phenotypic differences are potentially due to more subtle sequence divergence (SNPs and Indels) and gene expression. This further illustrates the importance of large-scale sequence comparison of *Cryptosporidium* species to elucidate potentially exploitable variation. Widmer et al. identified a number of highly divergent genes by comparison of the genomes of *C. parvum gp60* subtype IIc and the Iowa II reference (85). Further investigation reveals that genomic evolution was largely reductive, resulting in *Cryptosporidium* depending mainly on host cells for basic nutrients (86).

As more genomes are becoming available at an ever-increasing rate, researchers are able to explore further the biology and evolution of *Cryptosporidium*. Recently, Nader et al. (87) used 21 whole genome sequences to show the existence of two subspecies lineages of *C. parvum* (*C. parvum parvum* and *C. parvum anthroponosum*) with different host-adapted infectivity. Additionally, they identified some of the historic genetic exchanges that have occurred between these lineages and *C. hominis* during the evolution of these different species and subspecies, even suggesting rough time-lines for when these events occurred (87, 88).

In an important epidemiological development, Gilchrist et al. (31) used the methods described by Hadfield et al. (75), to study the genetic diversity of *C. hominis* in slum dwelling infants in Dhaka, Bangladesh, over a 2-year period. As mentioned above, they found that *C. hominis* was more abundant during the monsoon periods and showed high levels of diversity at *gp60* locus. Furthermore, WGS revealed extensive SNP diversity, and very high levels of variation at seven distinct loci. They also detected high levels of recombination within the *C. hominis* populations, evidenced by linkage disequilibrium decay. The genetic diversity of *C. hominis* encountered in the Bangladesh study was found to be far greater than that seen in northern Europe, where the predominant *C. hominis* IbA10G2 subtype is highly conserved at the genome level (50, 71). This study reveals the importance of high-throughput, wide scale genomic sequencing and analysis in elucidating the complex population structure of the parasite worldwide (31).

In another study, WGS was also used for a comparative genomic analysis between two subtypes of *C. hominis* that have been dominant in the US at various times, IbA10G2 and IaA28R4, and *C. parvum* (70). Their genome comparison revealed evidence of genetic recombination in the two *C. hominis* subtypes, and also some unique genetic differences between *C. hominis* and *C. parvum*, and multigene families that may contribute to the host variation between these two species (70).

Genome Availability

The advent of the new techniques to facilitate the DNA extraction, enrichment, sequencing, and assembly of high quality *Cryptosporidium* genomes from clinical samples, provides

TABLE 2 | The progression of *C. hominis* and *C. parvum* whole genome assembly from initial attempts in 2004 to the completed genomes in 2015 and 2016 (80–83).

Feature	*C. parvum* Iowa II (2004)	*C. hominis* TU502 (2004)	*C. hominis* UdeA01 (2015)	*C. parvum* Iowa II (2015)	*C. hominis* TU502 (2016)
Genome length	9.10 Mbp	9.16 Mbp	9.05 Mbp	9.10 Mbp	9.10 Mbp
Coding genes (% genome)	3807 (75.3%)	3994 (69%)	3819 (75.4%)	3865 (75.7%)	3745 (77.8%)
GC content	0.3	0.32	0.32	0.32	0.3
Introns	0.05	0.05-0.20	0.109	0.108	not reported
Fragments	18	1422	8	8	119

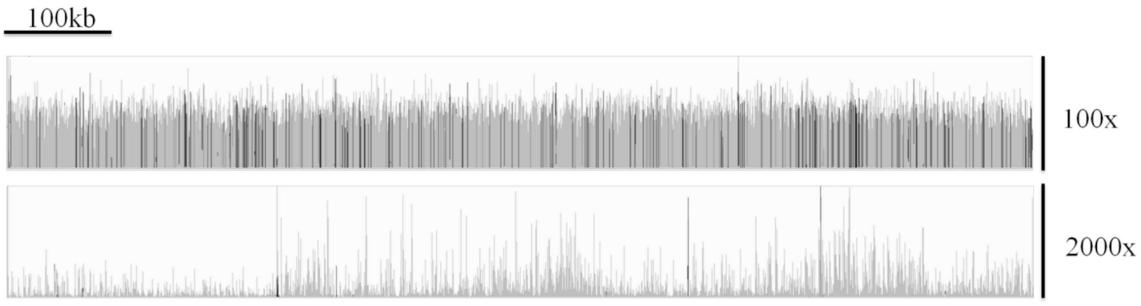


FIGURE 3 | A comparison of the coverage over chromosome 1 *of C. parvum* Iowa II (top track) and the clinical isolate UKP3 (bottom track), showing the highly uneven coverage typically exhibited from many clinical isolates. Reads were mapped using Bowtie v2.3.3.1 (93) and visualized using Integrative Genomics Viewer v2.4.16 (94).

an opportunity to greatly expand the number of genomes available. An EU funded collaboration (Aquavalens project, www.aquavalens.org) between several institutions generated 27 assemblies of *C. parvum*, *C. hominis*, *Cryptosporidium viatorum*, *C. ubiquitum*, *C. cuniculus*, and *C. meleagridis* directly from clinical isolates using the DNA extraction and purification protocol described by Hadfield et al. (75) and Nader et al. (87). Under another EU funded project, COMPARE (https://www.compare-europe.eu/), 31 new *C. parvum* and 19 new *C. hominis* genome assemblies were generated from clinical isolates, using the DNA extraction and purification protocol described by Hadfield et al. (75), and the DNA enrichment protocol described by Guo et al. (77). A further 14 *C. hominis* genomes, representing 9 different *gp60* subtypes, have also been published (89) and are available as a Bioproject (PRJNA307563) on the National Center for Biotechnology Information (NCBI) online databases. Currently, whole genome assemblies of isolates from human and animal derived *Cryptosporidium* spanning 9 species, are available as Bioprojects on NCBI databases (see **Table 1**), but this number is rapidly increasing as methods and technology become more available. The *Cryptosporidium* genomics resource CryptoDB (http://cryptodb.org/), provides access to species including *C. hominis*, *C. parvum*, other zoonotic species including *C. meleagridis*, and host-adapted species rarely found in humans (*Cryptosporidium muris*, *Cryptosporidium andersoni*, *C. baileyi*, and *Cryptosporidium tyzzeri*) and provides analytical tools to mine and compare the genomes sequences and their functionality (90, 91). A number of unassembled, unprocessed raw read sequences are also publically available via online repositories such as GenBank and the Welcome Trust Sanger Institute FTP servers.

Sequencing Using Long-Read Technology

Recently, there have been attempts to generate *Cryptosporidium* sequences using long-read technology, such as MinION by Oxford Nanopore, and Pacific Biosciences. There exist a few draft genomes from long reads generated by PacBio, but most are yet unpublished. However, a *C. parvum* PacBio sequence is available on the Welcome Trust Sanger Institute FTP servers (ftp://ftp.sanger.ac.uk/project/pathogens/Cryptosporidium) that was generated to map shorter Illumina reads to during the study in Dhaka that explored the genetic diversity of *C. hominis* (31). Currently, there have been no successful attempts at sequencing the genome using the MinION platform published. This is likely due to the large amount of DNA required to generate such reads using this particular technology, which is a known difficulty associated with *Cryptosporidium* genomic sequencing.

Pitfalls in Genome Assembly

Morris et al. have outlined difficulties associated with generating reliable and accurate genome assemblies from clinical isolates of *Cryptosporidium* (92). They demonstrated that the issues surrounding extracting sufficient DNA from clinical isolates resulted in highly uneven depth of coverage across the genome (for an example, see **Figure 3**) which can be seen in sequences generated from clinical isolates by a number of research teams. This, in tandem with the large number of low complexity regions within the *Cryptosporidium* genome, results in widespread genome misassembly when using the Spades assembler (95). Peng et al. further proposed an approach to generating reliable draft assemblies from clinical samples, and demonstrated how accurate resolution of low complexity regions are essential for biomarker discovery using the Iterative De-Bruijn Assembler (IDBA) (96).

Assembly of *C. parvum* and *C. hominis* is facilitated by high quality reference sequences (*C. parvum* IowaII and *C. hominis* UdeA01) which allow for reference-guided assembly. This, however, is not the case for other species of *Cryptosporidium*. It is therefore important to consider whether a reference guided assembly should be attempted, and what reference genome to use. The application of an inappropriate reference sequence may result genome assembly errors.

APPLICATIONS, FUTURE ISSUES, AND RESEARCH DIRECTIONS

With the recent expansion in the number of available raw read archives and genome assemblies generated from clinical samples, further *in silico* investigation can be carried out in an attempt to resolve a number of biological questions, such as:

- Can biomarkers differentiate genetic lineages of *Cryptosporidium* spp. virulence or pathogenicity, and therefore act as targets for diagnostic interrogation or novel therapeutics?

- How much variability exists within intergenic regions in species of *Cryptosporidium*?
- To what extent do multiple sub-populations of *Cryptosporidium* spp. exist within an infected host and in single clinical samples and impact of these during onward transmission and even the evolution of the parasite?

Biomarker Discovery and Analysis

The state of *Cryptosporidium* genotyping is far from resolved, and there is still a large amount of work to be done regarding the discovery, assessment, and selection of suitable biomarkers and genotyping conventions. Subsequent to the increasing availability of genomes is a bottle-neck in the analysis of these data, and there is a need to develop time-efficient, computationally inexpensive and high-throughput (automated) methods of genome analysis. "In house" pipelines have been used for biomarker detection and analysis. A typical example was reported by Perez-Cordon et al. (15), who used Tandem Repeats Finder (TRF) (97) to detect Variable Number Tandem Repeat (VNTR) regions within the genome of *Cryptosporidium parvum* Iowa II isolate and aligned them to homologues within a dataset of genomes generated by Hadfield et al. (75). This pipeline consisted of three primary steps:

1. Tandem Repeat (TR) identification in a reference genome.
2. Discovery of the TR's around the genome of a dataset of assembled genomes.
3. Assessment of these TR's for variation and subsequent viability as Biomarkers.

Using this pipeline, bioinformatic analysis of the Hadfield dataset alone has yielded a large number of novel VNTR regions (15), some of which compare favorably to the commonly used *gp60* marker in their ability to resolve discrete subtypes of *C. parvum*. Automating pipelines, can utilize the increasing amounts of whole genome sequence data available for *Cryptosporidium* allowing for the discovery of novel VNTRs in a high-throughput manner.

In addition to novel VNTR markers, genome analysis of other *Cryptosporidium* species and genotypes can allow for the redescription of known markers in these for the development of new subtyping tools. One example, is with the zoonotic species *Cryptosporidium ubiquitum*, where the homolog of *gp60* was diverse from those of *C. hominis* and *C. parvum* so could not be used to differentiate isolates (46). Li et al. used whole genome sequence data to identify and develop a *gp60* subtyping tool that allowed the differentiation and showed apparent host-adaptation (46). Another example, described the development from whole genome sequencing data of a two marker subtyping tool (*gp60* and a mucin protein gene) for the zoonotic chipmunk genotype I (98).

When developing genotyping assays, it is important that biomarkers are selected so as not to influence the outcome of the analysis. For example, markers must be distant enough from each other on the same chromosome or spread over the eight chromosomes to ensure genetic linkage does not occur, and markers must give high enough discrimination when combined to be appropriate for the application in question, such as demonstrating epidemiological relationships (27, 84).

Multiplicity of Infection in *Cryptosporidium*

It is both biologically plausible (due to unrestricted sexual recombination between sub-populations), and there is strong evidence (described below) that infections can arise from, and give rise to, multiple sub-populations of *Cryptosporidium* spp. which will be present in individual hosts (termed here multiplicity of infection—MOI) and thus clinical samples. This is driven by meiotic division in the zygote resulting in potential re-assortment of chromosomes (**Figure 2b**). As a result, the genomes of the haploid sporozoites within an oocyst may differ from each other and the parent sporozoites. Grinberg and Widmer demonstrated the common occurrence of MOI and provided evidence that the degree of MOI may depend on prevailing transmission patterns within geographical regions (25). The current approaches of Sanger sequencing results in the resolution of a single allele at each locus for the population, which, if MOI is present, would in effect simply represent the most populous sequence variant at each locus within the assembly. Grinberg and Widmer illustrated this from three hypothetical infections (25), but the potential extent for MOI is theoretically even greater (**Figure 2b**). This may confound epidemiological analysis, which generally relies on the assumption that large-scale genetic recombination does not occur within a host, and that a single host exhibits a single, clonal population. Furthermore, it has been suggested that MOI is a driving force behind the evolution of virulence, and has a complex relationship with both the virulence experienced by the host, and transmission (99, 100). It is therefore essential that MOI is well-understood and accounted for in order to develop novel prevention strategies in the fight against cryptosporidiosis and other parasitic diseases. The investigation into the impact of MOI relies on the accurate and reliable detection and discrimination of discrete populations of parasites, not readily achieved by current genotyping approaches. There are a few major alternatives to achieve this:

- Cloning and sequencing key loci to detect variation.
- Isolating and sequencing single oocysts from clinical samples.
- Comparing length polymorphism at multiple loci.
- Investigating sequence variation among reads within short read archives generated by Next Generation Sequencing (NGS).

These approaches investigate MOI from very different angles: variable locus cloning and single cell sequences from an experimental angle, and length polymorphism and sequence variation within reads from an *in silico* angle. This lends them unique challenges to overcome. By cloning PCR amplicons of selected loci (*gp60* and hsp70) and utilizing Next Generation Sequencing (NGS), Grinberg et al. reported the presence of numerous sub-populations within single isolates of *C. parvum*. They demonstrated the presence of two *hsp70* and 10 *gp60* alleles within their two isolate dataset. Furthermore, they reported that in both isolates there was a dominant allele, which represented the majority of the amplicons sequenced (101). Single oocysts were isolated and sequenced by Troell et al. (78) with a

view to elucidate these putative intra-isolate sub-populations. Sequencing 10 oocysts individually resulted in assemblies of 49.4–91.8% of the size of the *C. parvum* Iowa II reference genome. By pooling the reads from all 10 oocysts, they generated a 94.4% complete genome. Variation at multiple loci was detected between the assembled genomes, verifying the presence of discrete populations within the "isolate" (78). Analysis of fragment length polymorphism can highlight MOI, however, due to PCR-based amplification of the fragments, minority variants are largely undetectable (25). To compare the results obtained from Sanger sequencing and NGS, Zahedi et al. investigated *gp60* amplicons from 11 *C. hominis*, 22 *C. parvum*, and 8 *C. cuniculus* animal samples from Australia and China (102). They demonstrated that NGS is more effective at resolving the presence of multiple populations of *Cryptosporidium* within a sample, and the extent of MOI. There was concordance between the subtypes identified by both platforms, but additional subtypes were identified using NGS on *C. parvum* and *C. cuniculus gp60* amplicons, but not *C. hominis*.

The major issue with the experimental approaches detailed above is that they are expensive, extremely labor intensive and time consuming, leading to poor scalability. This leads to a major problem in generating sufficient data with which to begin to unravel the role of these parasite sub-populations, and to understand their overall impact on global public health. It is expected that they will have roles in affecting transmission by reducing host-fitness (virulence), and in generating novel subtypes via sexual recombination. There is therefore a great need to develop strategies which allow us to carry out investigations in a high-throughput manner, utilizing the wealth of raw genomic data is available for *Cryptosporidium* and other related parasites. Using biomarkers discovered from the analysis of the increasing number of high quality genomes, the opportunity arises to start to investigate MOI using *in silico* techniques, by mining raw read sets sequenced from clinical samples for information, which may have been previously unattainable. This approach involves three stages:

1. Identification of target regions for read interrogation. It is essential to select target regions, which are likely to show variation in-host, and it is therefore wise to select loci which show large amounts of variation between hosts.
2. Identification of reads within a single-host read set which have captured the target region.
3. Assessment of variation of the target sequence amongst reads which were identified in step 2.

A high level of variation within a single-host read set indicates the presence of multiple populations. Preliminary analysis of the Hadfield et al. dataset (75) indicated extensive variation at multiple tandem repeat loci around the *Cryptosporidium* genome, indicating highly complex in-host population structure. Results for variance mining at the *gp60* locus can be seen in **Figure 4**, which shows high levels of fragment length variation. However, there is invariably a single allele which appears to be

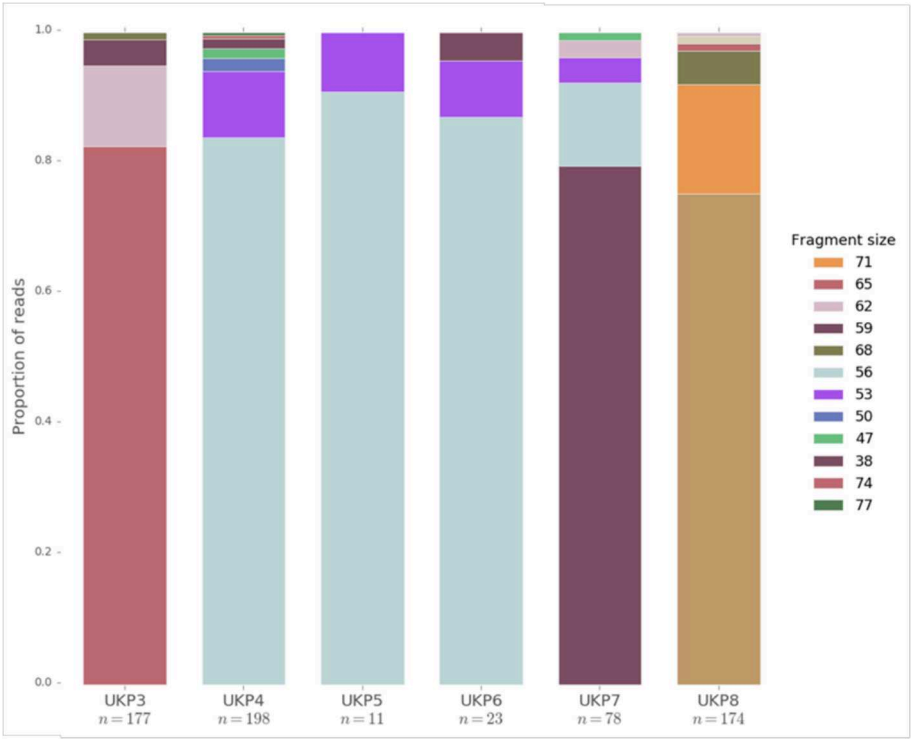


FIGURE 4 | The distribution of fragment lengths at the *gp60* locus mined from raw read sets generated from human clinical samples of UK isolated *C. parvum* by Hadfield et al. (75). Fragment lengths are given in the legend. *n* refers to the number of reads which fully captured the *gp60* region, and are therefore presented in the data.

most frequently exhibited within reads, and therefore considered dominant. This is in agreement with the findings reported by others, which show similar population structure (78, 101). There is, however, a disparity in the extent of MOI in *Cryptosporidium* between laboratory evidence by fragment sizing of key loci, and by mining NGS data. This is potentially due to the limited sensitivity of such approaches to identify multiple alleles of similar fragment size. Furthermore, PCR may preferentially amplify more abundant alleles, resulting in the less abundant alleles being obscured, as shown by Grinberg et al. who initially only identified the predominant alleles in their samples by PCR and Sanger sequencing (101). It may also be the case that such studies were not designed to detect multiple alleles within a single sample, and therefore underestimate the incidence of MOI. Consequently, care should be taken when interpreting entirely *in silico* results in the absence of experimental data. Due to MOI being a new area of investigation in *Cryptosporidium* research, the reliability of *in silico* approaches to elucidate in-host population diversity is still unclear, particularly in the light of studies indicating extensive contamination of samples (77). Preliminary results, however, appear to make predictions which are in accordance with experimental and epidemiological evidence, giving confidence in such data.

Natural transmission studies from analyzing secondary infections and those in farm settings has shown that dominant subtypes can be stable for many years or they can vary from year to year. For example, the outbreaks among visiting children on a holiday farm in Norway showed the same *gp60* subtype, IIaA19G1R1, was still circulating over several years and an investigation into secondary transmission within households after the children returned home also found the same subtype (103). While there was no evidence at the *gp60* gene of mixed populations in this example, in farm settings it is common for multiple subtypes to be present (104, 105). During a study of household transmission in a rural and urban setting in Bangladesh, a wide variety of *gp60* subtypes were found, particularly in the urban setting, but often there were concurrent infections with the same subtype within households and therefore it was mostly impossible to know the directionality of transmission (106). Where there were different subtypes within households it is unclear whether these stemmed from external sources rather than secondary transmission within the household (106). However, despite these studies there is a lack of data from mixed natural infections and the changes or dominance of subtypes that may occur during onward transmission, something that warrants further investigation using multilocus tools or whole genome data. Cama et al. used MLST to characterize differences in Iowa reference *C. parvum* isolates that had been maintained in different laboratories and described differences that were likely the results of passages through calves infected with exogenous *C. parvum* (107). This genetic drift in reference isolates was also seen with the TU502 reference *C. hominis* isolate between 2005 and 2012 following multiple animal passages (76). Therefore, the implications of MOI for surveillance and outbreak investigations are uncertain. As drift may happen in the longer term but not necessarily in the short term, detecting an outbreak "type" is reasonable, but equally it could be that two cases with apparently different subtypes are still actually linked if there is bias in the detection of dominant alleles.

CONCLUSIONS

WGS holds tantalizing promise for better understanding the transmission of cryptosporidosis, but there are still good reasons as to why it is not used routinely for diagnostics in a clinical setting. These include issues with extracting high quality pure DNA from clinical samples and issues with uneven depth of read coverage that leads to gaps in the assembled genome sequence. This later issue has important implications for cost: reducing costs by sequencing at a low depth of coverage is problematic, because it will increase the size and frequency of gaps in the assembled genome sequences. Nonetheless, while WGS is not yet on the horizon as method for routine clinical genotyping, it is indirectly having an important influence on clinical diagnostics. For instance, WGS is being used to guide and inform the development of MLST schemes, such as those based on VNTRs and fragment sizing. It is providing key insights into the evolutionary development of *Cryptosporidium*, including the discovery of new subspecies. Perhaps most important in terms of understanding the transmission of the disease, WGS is providing key insights into MOI. While evidence for MOI is occasionally found using fragment sizing, preliminary WGS analysis shows that is it is much more prevalent than the evidence from fragment sizing might suggest. WGS shows that although clinical samples do indeed contain multiple alleles, a single highly abundant allele usually dominates the data sets. It is highly likely that only the dominant allele that is detected via fragment sizing, with the other alleles remaining undetected. Resolution of these multiple populations is a stepping-stone to understanding the driving factors behind the evolution of virulence, and how new subtypes and genotypes arise in *Cryptosporidium*.

AUTHOR CONTRIBUTIONS

RC devised and revised the manuscript. GR, AM, and MS drafted the manuscript. All authors approved the final manuscript.

FUNDING

This work was funded by the Knowledge Economy Skills Scholarships (KESS 2), a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.

ACKNOWLEDGMENTS

We would like to thank Gregorio Perez Cordon for preparing **Figure 1** and for helpful comments on the manuscript, and Kevin Tyler for helpful discussions and assistance with **Figure 2b**.

REFERENCES

1. Current WL, Garcia LS. Cryptosporidiosis. *Clin Microbiol Rev*. (1991) 4:325–58. doi: 10.1128/cmr.4.3.325
2. Okhuysen PC, Chappell CL, Crabb JH, Sterling CR, DuPont HL. Virulence of three distinct *Cryptosporidium parvum* isolates for healthy adults. *J Infect Dis*. (1999) 180:1275–81. doi: 10.1086/315033
3. King BJ, Monis PT. Critical processes affecting *Cryptosporidium* oocyst survival in the environment. *Parasitology*. (2007) 134:309–23. doi: 10.1017/S0031182006001491
4. Jenkins MB, Eaglesham BS, Anthony LC, Kachlany SC, Bowman DD, Ghiorse WC. Significance of wall structure, macromolecular composition, and surface polymers to the survival and transport of *Cryptosporidium parvum* oocysts. *Appl Environ Microbiol*. (2010) 76:1926–34. doi: 10.1128/AEM.02295-09
5. Chalmers RM, Robinson G, Elwin K, Elson R. Analysis of the *Cryptosporidium* spp. and *gp60* subtypes linked to human outbreaks of cryptosporidiosis in England and Wales, 2009 to 2017. *Parasit Vectors*. (2019) 12:95. doi: 10.1186/s13071-019-3354-6
6. Sponseller JK, Griffiths JK, Tzipori S. The evolution of respiratory cryptosporidiosis: evidence for transmission by inhalation. *Clin Microbiol Rev*. (2014) 27: 575–86. doi: 10.1128/CMR.00115-13
7. Flores J, Okhuysen PC. Genetics of susceptibility to infection with enteric pathogens. *Curr Opin Infect Dis*. (2009) 22:471–6. doi: 10.1097/QCO.0b013e3283304eb6
8. Borad A, Ward H. Human immune responses in cryptosporidiosis. *Future Microbiol*. (2010) 5:507–19. doi: 10.2217/fmb.09.128
9. Chalmers RM, Davies AP. Minireview: clinical cryptosporidiosis. *Exp Parasitol*. (2010) 124:138–46. doi: 10.1016/j.exppara.2009.02.003
10. Checkley W, White AC Jr, Jaganath D, Arrowood MJ, Chalmers RM, Chen XM, et al. A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *Cryptosporidium*. *Lancet Infect Dis*. (2015) 15:85–94. doi: 10.1016/S1473-3099(14)70772-8
11. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*. (2013) 382:209–22. doi: 10.1016/S0140-6736(13)60844-2
12. GBD Diarrhoeal Diseases Collaborators. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Infect Dis*. (2017) 17:909–48. doi: 10.1016/S1473-3099(17)30276-1
13. Bartelt LA, Lima AA, Kosek M, Peñataro Yori P, Lee G, Guerrant RL. "Barriers" to child development and human potential: the case for including the "neglected enteric protozoa" (NEP) and other enteropathy-associated pathogens in the NTDs. *PLoS Negl Trop Dis*. (2013) 7:e2125. doi: 10.1371/journal.pntd.0002125
14. Savioli L, Smith H, Thompson A. *Giardia* and *Cryptosporidium* join the 'Neglected Diseases Initiative'. *Trends Parasitol*. (2006) 22:203–8. doi: 10.1016/j.pt.2006.02.015
15. Pérez-Cordón G, Robinson G, Nader J, Chalmers RM. Discovery of new variable number tandem repeat loci in multiple *Cryptosporidium parvum* genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. *Exp Parasitol*. (2016) 169:119–28. doi: 10.1016/j.exppara.2016.08.003
16. Bouzid M, Hunter PR, Chalmers RM, Tyler KM. *Cryptosporidium* pathogenicity and virulence. *Clin Microbiol Rev*. (2013) 26:115–34. doi: 10.1128/CMR.00076-12
17. Khan A, Shaik JS, Grigg ME. Genomics and molecular epidemiology of *Cryptosporidium* species. *Acta Trop*. (2018) 184:1–14. doi: 10.1016/j.actatropica.2017.10.023
18. Baragaña B, Forte B, Choi R, Nakazawa Hewitt S, Bueren-Calabuig JA, Pisco JP, et al. Lysyl-tRNA synthetase as a drug target in malaria and cryptosporidiosis. *Proc Natl Acad Sci USA*. (2019) 116:7015–20. doi: 10.1073/pnas.1814685116
19. Xiao L, Feng Y. Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and *Giardia duodenalis*. *Food Waterborne Parasitol*. (2017) 8–9:14–32. doi: 10.1016/j.fawpar.2017.09.002
20. Xiao L. Molecular epidemiology of cryptosporidiosis: an update. *Exp Parasitol*. (2010) 124:80–9. doi: 10.1016/j.exppara.2009.03.018
21. Feng Y, Ryan UM, Xiao L. Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol*. (2018) 34:997–1011. doi: 10.1016/j.pt.2018.07.009
22. Feng X, Rich SM, Tzipori S, Widmer G. Experimental evidence for genetic recombination in the opportunistic pathogen *Cryptosporidium parvum*. *Mol Biochem Parasitol*. (2002) 119:55–62. doi: 10.1016/S0166-6851(01)00393-0
23. Tanriverdi S, Blain JC, Deng B, Ferdig MT, Widmer G. Genetic crosses in the apicomplexan parasite *Cryptosporidium parvum* define recombination parameters. *Mol Microbiol*. (2007) 63:1432–9. doi: 10.1111/j.1365-2958.2007.05594.x
24. Mallon M, MacLeod A, Wastling J, Smith H, Reilly B, Tait A. Population structures and the role of genetic exchange in the zoonotic pathogen *Cryptosporidium parvum*. *J Mol Evol*. (2003) 56:407–17. doi: 10.1007/s00239-002-2412-3
25. Grinberg A, Widmer G. *Cryptosporidium* within-host genetic diversity: systematic bibliographical search and narrative overview. *Int J Parasitol*. (2016) 46:465–71. doi: 10.1016/j.ijpara.2016.03.002
26. Widmer G, Lee Y. Comparison of single- and multilocus genetic diversity in the protozoan parasites *Cryptosporidium parvum* and *C. hominis*. *Appl Environ Microbiol*. (2010) 76:6639–44. doi: 10.1128/AEM.01268-10
27. Robinson G, Chalmers RM. Assessment of polymorphic genetic markers for multi-locus typing of *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Exp Parasitol*. (2012) 132:200–15. doi: 10.1016/j.exppara.2012.06.016
28. Chalmers RM, Pérez-Cordón G, Cacció SM, Klotz C, Robertson LJ, on behalf of the participants of the *Cryptosporidium* genotyping workshop (EURO-FBP). *Cryptosporidium* genotyping in Europe: the current status and processes for a harmonised multi-locus genotyping scheme. *Exp Parasitol*. (2018) 191:25–30. doi: 10.1016/j.exppara.2018.06.004
29. Aldeyarbi HM, Abu El-Ezz NM, Karanis P. *Cryptosporidium* and cryptosporidiosis: the African perspective. *Environ Sci Pollut Res Int*. (2016) 23:13811–21. doi: 10.1007/s11356-016-6746-6
30. Mahmoudi MR, Ongerth JE, Karanis P. *Cryptosporidium* and cryptosporidiosis: the Asian perspective. *Int J Hyg Environ Health*. (2017) 220:1098–109. doi: 10.1016/j.ijheh.2017.07.005
31. Gilchrist CA, Cotton JA, Burkey C, Arju T, Gilmartin A, Lin Y, et al. Genetic diversity of *Cryptosporidium hominis* in a Bangladeshi community as revealed by whole-genome sequencing. *J Infect Dis*. (2018) 218:259–64. doi: 10.1093/infdis/jiy121
32. Chalmers RM, Elwin K, Thomas AL, Guy EC, Mason B. Long-term *Cryptosporidium* typing reveals the aetiology and species-specific epidemiology of human cryptosporidiosis in England and Wales, 2000 to 2003. *Euro Surveill*. (2009) 14:19086. doi: 10.2807/ese.14.02.19086-en
33. Zintl A, Proctor AF, Read C, Dewaal T, Shanaghy N, Fanning S, et al. The prevalence of *Cryptosporidium* species and subtypes in human faecal samples in Ireland. *Epidemiol Infect*. (2009) 137:270–7. doi: 10.1017/S0950268808000769
34. Waldron LS, Ferrari BC, Cheung-Kwok-Sang C, Beggs PJ, Stephens N, Power ML. Molecular epidemiology and spatial distribution of a waterborne cryptosporidiosis outbreak in Australia. *Appl Environ Microbiol*. (2011) 77:7766–71. doi: 10.1128/AEM.00616-11
35. Garcia-R JC, French N, Pita A, Velathanthiri N, Shrestha R, Hayman D. Local and global genetic diversity of protozoan parasites: spatial distribution of *Cryptosporidium* and *Giardia* genotypes. *PLoS Negl Trop Dis*. (2017) 11:e0005736. doi: 10.1371/journal.pntd.0005736
36. McLauchlin J, Amar C, Pedraza-Díaz S, Nichols GL. Molecular epidemiological analysis of *Cryptosporidium* spp. in the United Kingdom: results of genotyping *Cryptosporidium* spp. in 1,705 fecal samples from humans and 105 fecal samples from livestock animals. *J Clin Microbiol*. (2000) 38:3984–90.
37. Chalmers RM, Smith R, Elwin K, Clifton-Hadley FA, Giles M. Epidemiology of anthroponotic and zoonotic human cryptosporidiosis in England and Wales, 2004-2006. *Epidemiol Infect*. (2011) 139:700–12. doi: 10.1017/S0950268810001688
38. Pollock KG, Ternent HE, Mellor DJ, Chalmers RM, Smith HV, Ramsay CN, et al. Spatial and temporal epidemiology of sporadic human

cryptosporidiosis in Scotland. *Zoonoses Public Health*. (2010) 57:487–92. doi: 10.1111/j.1863-2378.2009.01247.x
39. Learmonth J, Ionas G, Pita A, Cowie R. Seasonal shift in *Cryptosporidium parvum* transmission cycles in New Zealand. *J Eukaryot Microbiol*. (2001) 48:34S—5. doi: 10.1111/j.1550-7408.2001.tb00444.x
40. Garvey P, McKeown P. Epidemiology of human cryptosporidiosis in Ireland, 2004–2006: analysis of national notification data. *Euro Surveill*. (2009) 14:19128. doi: 10.2807/ese.14.08.19128-en
41. Wielinga PR, de Vries A, van der Goot TH, Mank T, Mars MH, Kortbeek LM, et al. Molecular epidemiology of Cryptosporidium in humans and cattle in The Netherlands. *Int J Parasitol*. (2008) 38:809–17. doi: 10.1016/j.ijpara.2007.10.014
42. Nic Lochlainn LM, Sane J, Schimmer B, Mooij S, Roelfsema J, van Pelt W, et al. Risk factors for sporadic cryptosporidiosis in the netherlands: analysis of a 3-year population based case-control study coupled with genotyping, 2013-2016. *J Infect Dis*. (2019) 219:1121–9. doi: 10.1093/infdis/jiy634
43. Roelfsema JH, Sprong H, Cacciò SM, Takumi K, Kroes M, van Pelt W, et al. Molecular characterization of human *Cryptosporidium* spp. isolates after an unusual increase in late summer 2012. *Parasit Vectors*. (2016) 9:138. doi: 10.1186/s13071-016-1397-5
44. Gharpure R, Perez A, Miller AD, Wikswo ME, Silver R, Hlavsa MC. Cryptosporidiosis Outbreaks - United States, 2009-2017. *MMWR Morb Mortal Wkly Rep*. (2019) 68:568–72. doi: 10.15585/mmwr.mm6825a3
45. Elwin K, Hadfield SJ, Robinson G, Chalmers RM. The epidemiology of sporadic human infections with unusual cryptosporidia detected during routine typing in England and Wales, 2000-2008. *Epidemiol Infect*. (2012) 140:673–83. doi: 10.1017/S0950268811000860
46. Li N, Xiao L, Alderisio K, Elwin K, Cebelinski E, Chalmers R, et al. Subtyping *Cryptosporidium ubiquitum*, a zoonotic pathogen emerging in humans. *Emerg Infect Dis*. (2014) 20:217–24. doi: 10.3201/eid2002.121797
47. Robinson G, Elwin K, Chalmers RM. Unusual *Cryptosporidium* genotypes in human cases of diarrhea. *Emerg Infect Dis*. (2008) 14:1800–2. doi: 10.3201/eid1411.080239
48. Chalmers RM, Robinson G, Elwin K, Hadfield SJ, Xiao L, Ryan U, et al. *Cryptosporidium* sp. rabbit genotype, a newly identified human pathogen. *Emerg Infect Dis*. (2009) 15:829–30. doi: 10.3201/eid1505.081419
49. Xiao L, Hlavsa MC, Yoder J, Ewers C, Dearen T, Yang W, et al. Subtype analysis of *Cryptosporidium* specimens from sporadic cases in Colorado, Idaho, New Mexico, and Iowa in 2007: widespread occurrence of one *Cryptosporidium hominis* subtype and case history of an infection with the *Cryptosporidium* horse genotype. *J Clin Microbiol*. (2009) 47:3017–20. doi: 10.1128/JCM.00226-09
50. Cacciò SM, Chalmers RM. Human cryptosporidiosis in Europe. *Clin Microbiol Infect*. (2016) 22:471–80. doi: 10.1016/j.cmi.2016.04.021
51. McKerr C, Adak GK, Nichols G, Gorton R, Chalmers RM, Kafatos G, et al. An outbreak of *Cryptosporidium parvum* across England & Scotland associated with consumption of fresh pre-cut salad leaves, May 2012. *PLoS ONE*. (2015) 10:e0125955. doi: 10.1371/journal.pone.0125955
52. Lange H, Johansen OH, Vold L, Robertson LJ, Anthonisen IL, Nygard K. Second outbreak of infection with a rare *Cryptosporidium parvum* genotype in schoolchildren associated with contact with lambs/goat kids at a holiday farm in Norway. *Epidemiol Infect*. (2014) 142:2105–13. doi: 10.1017/S0950268813003002
53. Mattsson JG, Insulander M, Lebbad M, Björkman C, Svenungsson B. Molecular typing of *Cryptosporidium parvum* associated with a diarrhoea outbreak identifies two sources of exposure. *Epidemiol Infect*. (2008) 136:1147–52. doi: 10.1017/S0950268807009673
54. Hunter PR, Wilkinson DC, Lake IR, Harrison FC, Syed Q, Hadfield SJ, et al. Microsatellite typing of *Cryptosporidium parvum* in isolates from a waterborne outbreak. *J Clin Microbiol*. (2008) 46:3866–7. doi: 10.1128/JCM.01636-08
55. Widmer G, Cacciò SM. A comparison of sequence and length polymorphism for genotyping *Cryptosporidium* isolates. *Parasitology*. (2015) 142:1080–5. doi: 10.1017/S0031182015000396
56. Drumo R, Widmer G, Morrison LJ, Tait A, Grelloni V, D'Avino N, et al. Evidence of host-associated populations of *Cryptosporidium parvum* in Italy. *Appl Environ Microbiol*. (2012) 78:3523–9. doi: 10.1128/AEM.07686-11
57. Quílez J, Vergara-Castiblanco C, Monteagudo L, del Cacho E, Sánchez-Acedo C. Host association of *Cryptosporidium parvum* populations infecting domestic ruminants in Spain. *Appl Environ Microbiol*. (2013) 79:5363–71. doi: 10.1128/AEM.01168-13
58. Mallon ME, MacLeod A, Wastling JM, Smith H, Tait A. Multilocus genotyping of *Cryptosporidium parvum* Type 2: population genetics and sub-structuring. *Infect Genet Evol*. (2003) 3:207–18. doi: 10.1016/S1567-1348(03)00089-3
59. Gatei W, Hart CA, Gilman RH, Das P, Cama V, Xiao L. Development of a multilocus sequence typing tool for *Cryptosporidium hominis*. *J Eukaryot Microbiol*. (2006) 53:S43–8. doi: 10.1111/j.1550-7408.2006.00169.x
60. Cacciò SM, de Waele V, Widmer G. Geographical segregation of *Cryptosporidium parvum* multilocus genotypes in Europe. *Infect Genet Evol*. (2015) 31:245–9. doi: 10.1016/j.meegid.2015.02.008
61. Ramo A, Quílez J, Monteagudo L, Del Cacho E, Sánchez-Acedo C. Intra-species diversity and panmictic structure of *Cryptosporidium parvum* populations in cattle farms in Northern Spain. *PLoS ONE*. (2016) 11:e0148811. doi: 10.1371/journal.pone.0148811
62. Ramo A, Monteagudo LV, Del Cacho E, Sánchez-Acedo C, Quílez J. Intra-species genetic diversity and clonal structure of *Cryptosporidium parvum* in sheep farms in a confined geographical area in Northeastern Spain. *PLoS ONE*. (2016) 11:e0155336. doi: 10.1371/journal.pone.0155336
63. Quílez J, Torres E, Chalmers RM, Hadfield SJ, Del Cacho E, Sánchez-Acedo C. *Cryptosporidium* genotypes and subtypes in lambs and goat kids in Spain. *Appl Environ Microbiol*. (2008) 74:6026–31. doi: 10.1128/AEM.00606-08
64. Wang R, Zhang L, Axén C, Bjorkman C, Jian F, Amer S, et al. *Cryptosporidium parvum* IId family: clonal population and dispersal from Western Asia to other geographical regions. *Sci Rep*. (2014) 4:4208. doi: 10.1038/srep04208
65. Li N, Xiao L, Cama VA, Ortega Y, Gilman RH, Guo M, et al. Genetic recombination and *Cryptosporidium hominis* virulent subtype IbA10G2. *Emerg Infect Dis*. (2013) 19:1573–82. doi: 10.3201/eid1910.121361
66. Feng Y, Torres E, Li N, Wang L, Bowman D, Xiao L. Population genetic characterisation of dominant *Cryptosporidium parvum* subtype IIaA15G2R1. *Int J Parasitol*. (2013) 43:1141–7. doi: 10.1016/j.ijpara.2013.09.002
67. Feng Y, Tiao N, Li N, Hlavsa M, Xiao L. Multilocus sequence typing of an emerging *Cryptosporidium hominis* subtype in the United States. *J Clin Microbiol*. (2014) 52:524–30. doi: 10.1128/JCM.02973-13
68. Tanriverdi S, Grinberg A, Chalmers RM, Hunter PR, Petrovic Z, Akiyoshi DE, et al. Inferences about the global population structures of *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Appl Environ Microbiol*. (2008) 74:7227–34. doi: 10.1128/AEM.01576-08
69. Beser J, Hallström BM, Advani A, Andersson S, Östlund G, Winiecka-Krusnell J, et al. Improving the genotyping resolution of *Cryptosporidium hominis* subtype IbA10G2 using one step PCR-based amplicon sequencing. *Infect Genet Evol*. (2017) 55:297–304. doi: 10.1016/j.meegid.2017.08.035
70. Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, et al. Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. *BMC Genomics*. (2015) 16:320. doi: 10.1186/s12864-015-1517-1
71. Chalmers RM, Caccio S. Towards a consensus on genotyping schemes for surveillance and outbreak investigations of *Cryptosporidium*, Berlin, June 2016. *Euro Surveill*. (2016) 21:30338. doi: 10.2807/1560-7917.ES.2016.21.37.30338
72. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol*. (2012) 13:R118. doi: 10.1186/gb-2012-13-12-r118
73. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis*. (2013) 13:137–46. doi: 10.1016/S1473-3099(12)70277-3
74. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis*. (2015) 61:305–12. doi: 10.1093/cid/civ318

75. Hadfield SJ, Pachebat JA, Swain MT, Robinson G, Cameron SJ, Alexander J, et al. Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics*. (2015) 16:650. doi: 10.1186/s12864-015-1805-9
76. Widmer G, Ras R, Chalmers RM, Elwin K, Desoky E, Badawy A. Population structure of natural and propagated isolates of *Cryptosporidium parvum*, *C. hominis* and *C. meleagridis*. *Environ Microbiol*. (2015) 17:984–93. doi: 10.1111/1462-2920.12447
77. Guo Y, Li N, Lysén C, Frace M, Tang K, Sammons S, et al. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. *J Clin Microbiol*. (2015) 53:641–7. doi: 10.1128/JCM.02962-14
78. Troell K, Hallström B, Divne AM, Alsmark C, Arrighi R, Huss M, et al. *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics*. (2016) 17:471. doi: 10.1186/s12864-016-2815-y
79. Widmer G, Lin L, Kapur V, Feng X, Abrahamsen MS. Genomics and genetics of *Cryptosporidium parvum*: the key to understanding cryptosporidiosis. *Microbes Infect*. (2002) 4:1081–90. doi: 10.1016/S1286-4579(02)01632-5
80. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, et al. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*. (2004) 304:441–5. doi: 10.1126/science.1094786
81. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, et al. The genome of *Cryptosporidium hominis*. *Nature*. (2004) 431:1107–12. doi: 10.1038/nature02977
82. Isaza JP, Galván AL, Polanco V, Huang B, Matveyev AV, Serrano MG, et al. Revisiting the reference genomes of human pathogenic *Cryptosporidium* species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference. *Sci Rep*. (2015) 5:16324. doi: 10.1038/srep16324
83. Ifeonu OO, Chibucos MC, Orvis J, Su Q, Elwin K, Guo F, et al. Annotated draft genome sequences of three species of *Cryptosporidium*: *Cryptosporidium meleagridis* isolate UKMEL1, *C. baileyi* isolate TAMU-09Q1 and *C. hominis* isolates TU502 2012 and UKH1. *Pathog Dis*. (2016) 74:ftw080. doi: 10.1093/femspd/ftw080
84. Widmer G, Sullivan S. Genomics and population biology of *Cryptosporidium* species. *Parasite Immunol*. (2012) 34:61–71. doi: 10.1111/j.1365-3024.2011.01301.x
85. Widmer G, Lee Y, Hunt P, Martinelli A, Tolkoff M, Bodi K. Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range. *Infect Genet Evol*. (2012) 12:1213–21. doi: 10.1016/j.meegid.2012.03.027
86. Keeling PJ. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. *Dev Cell*. (2004) 6:614–6. doi: 10.1016/S1534-5807(04)00135-2
87. Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, et al. Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol*. (2019) 4:826–36. doi: 10.1038/s41564-019-0377-x
88. Kissinger JC. Evolution of *Cryptosporidium*. *Nat Microbiol*. (2019) 4:730–1. doi: 10.1038/s41564-019-0438-1
89. Sikora P, Andersson S, Winiecka-Krusnell J, Hallström B, Alsmark C, Troell K, et al. Genomic variation in IbA10G2 and other patient-derived *Cryptosporidium hominis* subtypes. *J Clin Microbiol*. (2017) 55:844–58. doi: 10.1128/JCM.01798-16
90. Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC. CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res*. (2004) 32:D329–31. doi: 10.1093/nar/gkh050
91. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, et al. CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Res*. (2006) 34:D419–22. doi: 10.1093/nar/gkj078
92. Morris A, Pachebat J, Robinson G, Chalmers R, Swain M. Identifying and resolving genome misassembly issues important for biomarker discovery in the protozoan parasite, *Cryptosporidium*. In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. *Bioinformatics*. (2019) 3:90–100. doi: 10.5220/0007397200900100
93. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. (2009) 10:R25. doi: 10.1186/gb-2009-10-3-r25
94. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. (2013) 14:178–92. doi: 10.1093/bib/bbs017
95. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. (2012) 19:455–77. doi: 10.1089/cmb.2012.0021
96. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. (2012) 28:1420–8. doi: 10.1093/bioinformatics/bts174
97. Benson G. Tandem Repeats Finder: a program to analyse DNA sequences. *Nucleic Acids Res*. (1999) 27:573–8. doi: 10.1093/nar/27.2.573
98. Guo Y, Cebelinski E, Matusevich C, Alderisio KA, Lebbad M, McEvoy J, et al. Subtyping novel zoonotic pathogen *Cryptosporidium* chipmunk genotype I. *J Clin Microbiol*. (2015) 53:1648–54. doi: 10.1128/JCM.03436-14
99. Alizon S, de Roode JC, Michalakis Y. Multiple infections and the evolution of virulence. *Ecol Lett*. (2013) 16:556–67. doi: 10.1111/ele.12076
100. Sondo P, Derra K, Lefevre T, Diallo-Nakanabo S, Tarnagda Z, Zampa O, et al. Genetically diverse *Plasmodium falciparum* infections, within-host competition and symptomatic malaria in humans. *Sci Rep*. (2019) 9:1–9. doi: 10.1038/s41598-018-36493-y
101. Grinberg A, Biggs PJ, Dukkipati VS, George TT. Extensive intra-host genetic diversity uncovered in *Cryptosporidium parvum* using Next Generation Sequencing. *Infect Genet Evol*. (2013) 15:18–24. doi: 10.1016/j.meegid.2012.08.017
102. Zahedi A, Gofton AW, Jian F, Paparini A, Oskam C, Ball A, et al. Next Generation Sequencing uncovers within-host differences in the genetic diversity of *Cryptosporidium gp60* subtypes. *Int J Parasitol*. (2017) 47:601–7. doi: 10.1016/j.ijpara.2017.03.003
103. Johansen ØH, Hanevik K, Thrana F, Carlson A, Stachurska-Hagen T, Skaare D, et al. Symptomatic and asymptomatic secondary transmission of *Cryptosporidium parvum* following two related outbreaks in schoolchildren. *Epidemiol Infect*. (2015) 143:1702–9. doi: 10.1017/S095026881400243X
104. Wells B, Shaw H, Hotchkiss E, Gilray J, Ayton R, Green J, et al. Prevalence, species identification and genotyping Cryptosporidium from livestock and deer in a catchment in the Cairngorms with a history of a contaminated public water supply. *Parasit Vectors*. (2015) 8:66. doi: 10.1186/s13071-015-0684-x
105. Thomson S, Innes EA, Jonsson NN, Katzer F. Shedding of *Cryptosporidium* in calves and dams: evidence of re-infection and shedding of different gp60 subtypes. *Parasitology*. (2019) 146:1404–13. doi: 10.1017/S0031182019000829
106. Korpe PS, Gilchrist C, Burkey C, Taniuchi M, Ahmed E, Madan V, et al. Case-control study of *Cryptosporidium* transmission in bangladeshi households. *Clin Infect Dis*. (2019) 68:1073–9. doi: 10.1093/cid/ciy593
107. Cama VA, Arrowood MJ, Ortega YR, Xiao L. Molecular characterization of the *Cryptosporidium parvum* IOWA isolate kept in different laboratories. *J Eukaryot Microbiol*. (2006) 53:S40–2. doi: 10.1111/j.1550-7408.2006.00168.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generating Reliable Genome Assemblies of Intestinal Protozoans from Clinical Samples for the Purpose of Biomarker Discovery

Arthur Morris[1(✉)], Justin Pachebat[1], Graeme Tyson[1], Guy Robinson[2], Rachel Chalmers[2], and Martin Swain[1]

[1] IBERS, Aberystwyth University, Aberystwyth, Wales, UK
arm21@aber.ac.uk
[2] Cryptosporidium Reference Unit, Public Health Wales, Swansea, Wales, UK

Abstract. Protozoan parasites that cause diarrhoeal diseases in humans take a massive toll on global public health annually, with over 200,000 deaths in children of less than two years old in Asia and Sub-Saharan Africa being attributed to Cryptosporidium alone. They can, in particular, be a serious health risk for immuno-incompetent individuals. Genomics can be a valuable asset in helping combat these parasites, but there are still problems associated with performing whole genome sequencing from human stool samples. In particular there are issues associated with highly uneven sequence coverage of these parasite genomes, which may result in critical errors in the genome assemblies produced using a number of popular assemblers. We have developed an approach using the Gini statistic to better characterise depth of sequencing coverage. Furthermore, we have explored the sequencing biases resulting from Whole Genome Amplification approaches, and have attempted to relate those to the Gini statistic. We discuss these issues in two parasite genera: Cryptosporidium and Cyclospora, and perform an analysis of the sequencing coverage depth over these genomes. Finally we present our strategy to generate reliable genome assemblies of sufficient quality to facilitate discovery of new Variable Number Tandem Repeat (VNTR) biomarkers.

Keywords: Cryptosporidium · Genome assembly · Biomarker discovery

1 Introduction

Gastrointestinal parasitic protozoans have a significant impact on global public and veterinary health. In recent years, *Cyclospora cayetanensis* has been responsible for a number of significant outbreaks in the United Kingdom, Canada, and the United States [15,23]. The incidence of this parasite appears to be increasing [15]. In the developing world, Cryptosporidium infection alone is one of

A. Roque et al. (Eds.): BIOSTEC 2019, CCIS 1211, pp. 216–241, 2020.
https://doi.org/10.1007/978-3-030-46970-2_11

the main causes of childhood morbidity. A recent large-scale study identified it as contributing to approximately 202,000 deaths per year in children less than 24 months old [24]. In the UK, *C. parvum* and *C. hominis* cause most cases of cryptosporidiosis. While self-limiting after prolonged duration of symptoms (2–3 weeks) in immunocompetent hosts, severely immunocompromised patients suffer severe, sometimes life threatening disease.

The sequencing and assembly of whole or partial genomes has become an essential tool in modern science, facilitating research in every area of biology. A primary concern for parasites such as Cyclospora and Cryptosporidium is extracting from clinical samples sufficient amounts of high quality, low contaminant DNA for sequencing. Without this, sequencing may result in low coverage sequence, variable sequencing depth and poor quality genome assemblies. The impact of genomics has been limited by the fact that Cyclospora and Cryptospridium are currently unculturable *in vitro*. In 2015 this problem was overcome through an approach that now allows genomic Cryptosporidium DNA suitable for whole genome sequencing to be prepared directly from human stool samples [7]. Hadfield *et al.* [7] applied their method to the whole genome sequencing of eight *C. parvum* and *C. hominis* isolates. This method is being applied to Cyclospora, however purification still remains an issue, with no Immuno-Magnetic Seperation (IMS) kits available for this parasite. Presently, the Cryptosporidium genomics resource, CryptoDB [22], currently gives access to 13 complete genomes, with a total of 10 available from the NCBI, including a high-quality *C.parvum* reference genome [1] exhibiting a highly compact 9.1Mb genome, bearing 3,865 genes. However, there exists no such reference genome for Cyclospora, with the best quality assembly being 44.6 Mb over 865 contigs (PRJNA292682).

Currently improved understanding of Cryptosporidium epidemiology relies on conventional genotyping tests, however, such typing tools are limited for use in Cyclospora molecular investigations. The availability of whole genome sequences provides much higher resolution information for genotyping. In addition, the genomes can be used to study a wide array of aspects of pathogen biology, such as identity, taxonomy in relation to other pathogens, sensitivity or resistance to drugs, development of novel therapeutic agents, and virulence. Our interest is to develop novel genotyping approaches by identifying and evaluating variable regions around the genome of these parasites. This will allow sources of infection and routes of transmission to be characterized and compared in a cost- and time-efficient manner [6,21]. Here variable-number of tandem-repeats (VNTR) are used, with recent investigations concluding that additional loci need to be identified and validated [6]. Our work is building on that of Perez-Cordon *et al.* (2016), who used Tandem Repeats Finder [5] to identify polymorphic VNTRs around the genome of *C. parvum*, and analysed them for variation across the eight genomes sequenced by Hadfield *et al.* [7]. We aim to use whole genome sequencing of additional isolates and species to help achieve this goal, but this work is hampered by the quality of available genome sequences [21].

This paper is presented as an extension of the paper titled "Identifying and Resolving Genome Misassembly Issues Important for Biomarker Discovery in the Protozoan Parasite, Cryptosporidium" [17]. Here, we argue that the problems associated with the generation of genomes from clinical samples is seen in other gastrointestinal Apicomplexans, presenting genome assemblies of clinically isolated *Cyclospora cayetanensis*, and subjecting them to similar analysis. Furthermore we present a novel method of investigating and characterising the distribution of reads across a genome, termed Gini-granularity curves, which resolves issues associated with data granularity when calculating the Gini coefficient [16]. Finally we investigate the effect of using DNA enrichment by Whole Genome Amplification (WGA) to resolve the low DNA yields typically extracted from clinical samples of these parasites.

This paper is structured as follows. First, we explain the quality issues associated with genome sequences extracted from clinical stool samples. Then we describe our methods, including the data sets used, the novel utilisation of Gini and Gini-granularity curves to measure the distribution of read depth in a set of sequenced reads, the process of assembly with the identification of misassemblies, and the effect WGA has on coverage distribution. In the results and discussion sections, we summarise properties of the sequenced reads, show how they can lead to misassemblies, and give evidence of the types of misassembly we encounter. We also describe how using the Gini coefficient, and analysis of Gini-granularity curves can explain some of these assembly errors and characterise the distribution of reads across a genome. We then give a brief outline of the strategy we use to generate genome assemblies of sufficient quality to use for the discovery of novel VNTRs in Cryptosporidium, and extend this approach to Cyclospora, less the genome improvement (due to the lack of a reference genome). Finally we briefly discuss the value of WGA in generating high quality assemblies from low DNA yield intestinal protozoan clinical samples.

2 Sequencing and Assembly Issues in Gastrointestinal Parasitic Protozoans

Although it is possible to derive high quality DNA by culturing some parasites in donor animals [1], this is expensive, time consuming, and raises ethical concerns. It is not appropriate for clinical samples, where maintaining sequence identity is essential. Furthermore, no animal model has yet been identified for *Cyclospora cayetanensis*. Sequencing intestinal protozoan genomes directly from clinical samples suffers from three major problems:

1. The yield of oocysts from clinical samples is low.
2. The oocysts are extracted directly from faeces, necessitating extensive cleaning and purification before DNA extraction.
3. The DNA yield per oocyst is low.

These three problems commonly result in sequenced data sets with very uneven depth of coverage, see Fig. 1 for examples. The reasons for uneven depth

of coverage are unclear; in this paper we have attempted to elucidate some of the issues. Uneven sequencing depth has been identified in datasets obtained from published and unpublished paired end read libraries generated by different groups, and which were prepared using the standard Nextera XT DNA sample preparation kit. Moreover, many groups use Whole Genome Amplification (WGA) to increase the quantity of extracted DNA. This may have additional impact on the depth of coverage. WGA has been touted as a potential solution to samples which yield low levels of DNA or for which little DNA exists [8,13,31]. There has, however, been limited rigorous research into coverage bias introduced by such DNA enrichment techniques. Uneven sequencing depth may lead to genome misassembly, and we have identified this an issue with a number of popular *de novo* assemblers. Poor quality genome assemblies can find their way into public repositories of genome sequence and this can confound the development of novel prevention strategies, therapeutics, and diagnostic approaches.

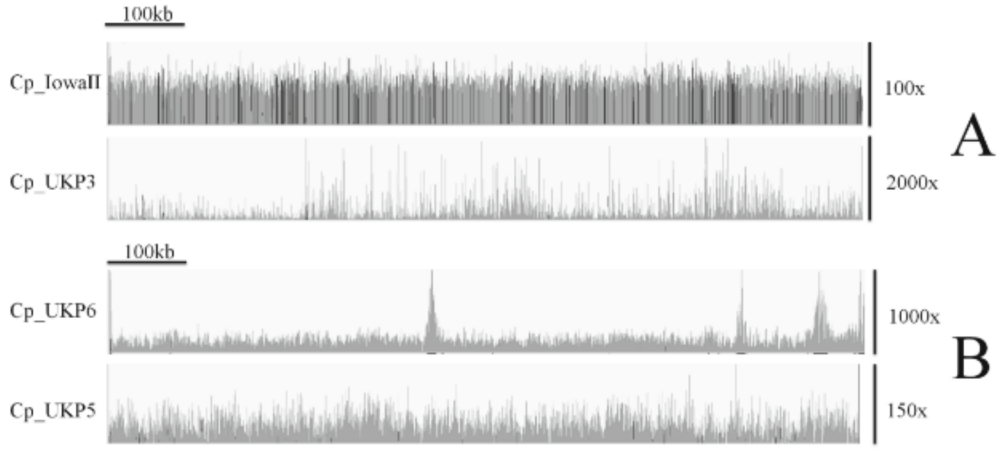


Fig. 1. A: Coverage across chromosome 7 of the *C.parvum* IowaII reference genome (top track) and *C.parvum* UKP3 (bottom track) genomes to illustrate the extreme coverage inequality of the UKP3 isolate genome (UKP3 $Gini = 0.5489$, IowaII $Gini = 0.112$). Image produced using IGV. Note that the IowaII DNA sequences were derived from an animal model, and have low or "normal" read depth variation, whereas UKP3 is more typical of DNA sequences extracted from clinical samples. **B:** The coverage over chromosome 1 for 2 genomes: Cp_UKP6 ($G \mid W_1 = 0.255$, $nAUC = 0.921$) and Ch_UKP5 ($G \mid W_1 = 0.278$, $nAUC = 0.884$) with $G \mid W_1$ across chromosome 1 alone of 0.262 and 0.264 respectively. See Sect. 4.2 for an explanation of the annotation.

3 Whole Genome Sequence Datasets Available for Analysis

As a dataset, we utilised 12 isolates of *Cyclospora cayetanensis* and 10 isolates of *Cryptosporidium spp.* There exists no high-quality Cyclospora reference genome, so reference guided scaffolding and other post-assembly processing using the PAGIT software was not undertaken for these isolates. The Cryptosporidium

dataset consisted of 7 UK isolates of *C. parvum* and 3 UK isolates of *C. hominis*, presented by Hadfield *et al.* [7]. The *C. parvum* IowaII reference genome [1] was used to guide assembly and annotation of the *C. parvum* assemblies, and the *C. hominis* TU502 [9] reference genome to guide assembly and annotation of *C. hominis*. The DNA within this dataset was un-enriched by Whole Genome Amplification (WGA).

For the purpose of identifying a correlation between genes transferred to chimeric regions and Gini, and to investigate the effect of WGA on genome sequencing, unpublished isolates consisting of 29 UK *C. parvum* and 19 UK *C. hominis* isolates where also used, giving a combined total of 48 genomes. These isolates were subjected to DNA enrichment pre-sequencing, using ϕ29 Multiple Displacement Amplification (MDA) WGA.

4 Sequencing and Read Analysis Methodology

4.1 Raw Read Analysis

Raw reads were mapped to a reference genome using Burrows Wheeler Aligner (BWA) v0.7.16. [14]. Cyclospora reads were mapped back to the assemblies they were used to generate, and Cryptosporidium reads were mapped to species specific reference genomes (*C. parvum* IowaII for *C. parvum* and *C. hominis* TU502 [29] for *C. hominis*). Coverage analysis was then performed using Samtools v1.5 [14].

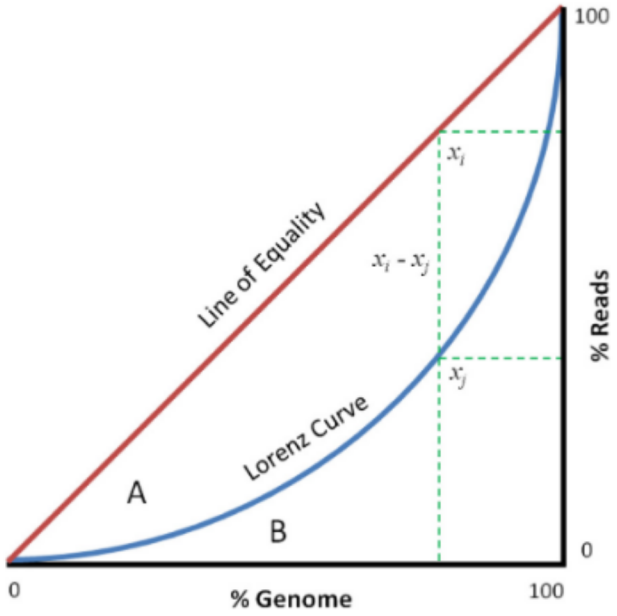


Fig. 2. Graphical representation of the Gini coefficient. In this graph, the Gini coefficient can be calculated as $A/(A+B)$, which represented area under the Lorenz curve (blue) inversely proportional to the line of equality (red). The green dotted lines denote the percentage of reads which cover 80% of a genome used to generate the Lorenz curve (unequal coverage depth) as compared to a perfectly equal distribution of reads. Taken from Morris *et al.* [17]. (Color figure online)

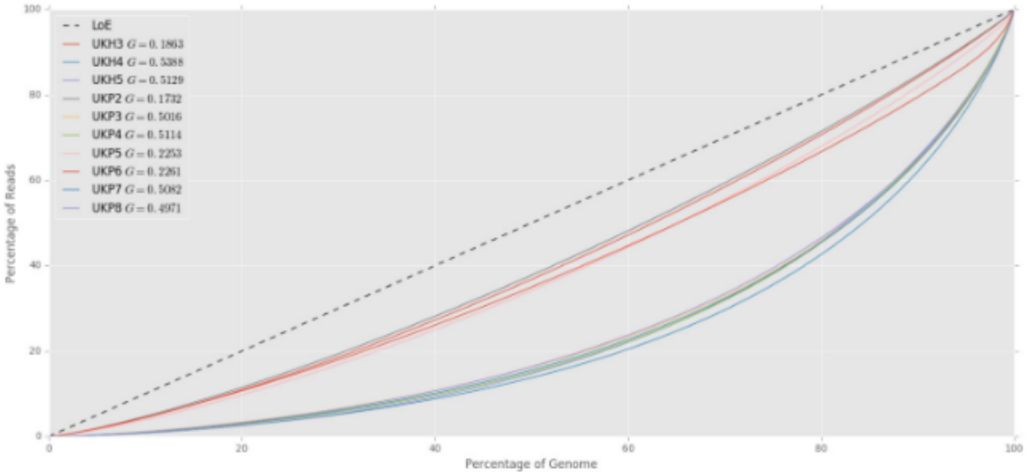


Fig. 3. Gini curves for the Hadfield *et al.* Cryptosporidium dataset. LoE refers to the Line of Equality, wherein theoretic perfect equality of the dataset it represented, achieving a Gini of 0. These Gini curves were generated using a window size of 500.

Read depth was calculated using the 'depth' tool within the samtools package [14]. The Gini coefficient is a metric used to measure the inequality within a dataset. It is commonly used in economics to measure the distribution of income within a population, where it is represented by a value between 0 and 1, with 0 representing perfectly even distribution, and higher values representing higher inequality of distribution. Here we have applied this coefficient to measure inequality of depth of coverage across a genome. For each of the genomes, we calculated the Gini coefficient of read depth. The Gini coefficient is calculated as:

$$G = A/(A + B)$$

where A is the area under the line of equality, and B the area under the Lorenz curve, on the graph of distribution inequality (see Fig. 2). The green dotted lines (marked at 80% on the x axis) in Fig. 2 gives an example of how, in the dataset used to generate the Lorenz curve, 80% of the genome is covered by only 40% of reads (the value at the position of collision of the green dotted line on the y axis), whereas in a perfect distribution it would be covered by 80% of reads.

The algorithm for calculating a genome's Gini coefficient of read depth coverage involves first calculating the mean depth of coverage of 1bp windows ($W = 1$) over the genome [17]. These windows are ordered according to their depth of coverage values, and these values rescaled between 0 and 100. This ordered set of read depth values is used to generate the Lorenz curve, L, where the value at every position i on the curve represents the sum of all values at positions $\leq i$. A line of equality, E, was generated to represent perfectly even distribution of reads across a genome. The difference between the values at each position on E and L is then calculated and the summed inverse proportional difference (the Gini coefficient) of these values calculated. This was performed using the following equation:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i} \tag{1}$$

where n refers to the number of windows (read depth values) across the genome, x_i is a depth of coverage value at position i on the line of equality E, and x_j is the value at position j on the Lorenz curve L.

The Gini coefficient for each genome represents the unevenness of read depth across the genome sequence (see Fig. 1A for an example of uneven coverage across chromosome 7 of UKP3 as compared to Iowa II).

4.2 Gini-Granularity Curves

To further investigate the read distribution throughout each assembly, the Gini coefficient was calculated across window sizes of 1–10,000 nucleotides. Furthermore, these curves were normalised such that the Gini at maximum granularity (which is obtained by calculating G at window size of 1) is adjusted to 1 for the purpose of simplified comparison. For each of these normalised and unnormalised curves, the area under the curve (AUC) was calculated. These curves are hereafter referred to as Gini-granularity curves.

More formally, consider a sequence s where the array of depth of coverage for each position in s is c. A partitioned array of c is generated using window size w, forming N partitions where $N = \frac{|c|}{w}$:

$$P_w^c = \{c_{[j,j+w]} \mid j = iw,\ 0 \leq i < N\} \tag{2}$$

The mean coverage over each partition is consequently:

$$C_w^c = \{\bar{n} \mid n \in P_w^c\} \tag{3}$$

where $\bar{n}$ is the mean depth of coverage of partition n. Taking G_w^c as the Gini of C_w^c (see Eq. 1 for the calculation of the Gini coefficient), an array of Gini values over a range of partition sizes is, $r = [i, j]$ where $r \subset \mathbb{N}$ is:

$$G_r^c = \{G_w^c \mid w \in r\} \tag{4}$$

A_r^c is the area under the curve generated by G_r^c. The normalised area under the curve, nA_r^c is calculated as the area under nG_r^c where:

$$nG_r^c = \{G_i^c + (1 - G_1^c) \mid 0 < i \leq |G_r^c|,\ i \in \mathbb{N}\} \tag{5}$$

4.3 DNA Enrichment Using Whole Genome Amplification

Due to the low DNA yield of Cryptosporidium, WGA was utilised to enrich the DNA for sequencing. The protocol was followed as documented in the protocol:

'Amplification of Purified Genomic DNA using the REPLI-g Mini Kit' by Qiagen. This was carried out as follows: 5 µl Buffer D1 was added to 5 µl template DNA, vortexed to mix, and briefly centrifuged. These were then incubated at room temperature for 3 min. During this time a master mix was prepared using 29 µl REPLI-g Mini Reaction Buffer and 1 µl REPLI-g Mini DNA Polymerase, to a total of 30 µl. 10μl Buffer N1 was added to the samples and mixed by vortexing, and centrifuged briefly. The master mix was then added to 20 µl of this denatured DNA, and incubated at 30 °C for 16 h. After this incubation period, the REPLI-g Mini DNA Polymerase was inactivated by heating the sample at 65 °C for 3 min.

4.4 Sequencing Bias Analysis in WGA Datasets

To investigate bias which may exist in how DNA is enriched using WGA, or sequenced, we used the `depth` tool within the Samtools package. Bespoke python scripts were written to analyse the relationship between coverage and GC content across windows of various sizes. Kernel density estimation was carried out on these datasets to investigate the relationship between coverage and genomic content using the SciPy package in Python [10].

5 Assembly and Post-assembly Improvement Methodology

De novo assembly was carried out in the same manner for both Cyclospora and Cryptosporidium. However, due to the lack of a reliable reference genome for Cyclospora, genome improvement and annotation using the PAGIT toolkit [25] was not possible. Consequently any statistics provided for Cyclospora assemblies are derived from the *de novo* assembly alone.

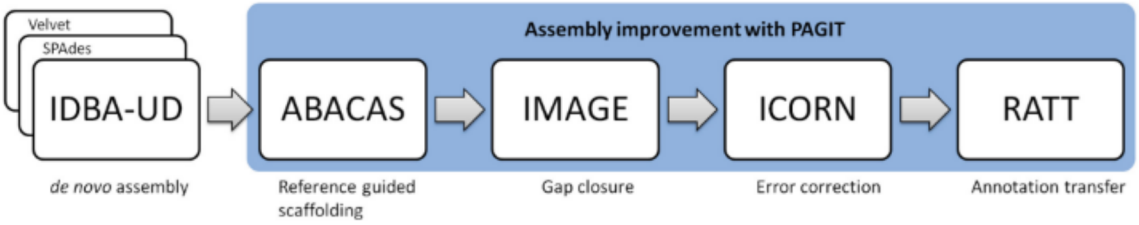


Fig. 4. The workflow for Cryptosporidium assembly, adapted from that used by Hadfield *et al.* for the assembly of genomes with high coverage depth inequality. Cyclospora genomes were assembled using only the IDBA-UD assembler. Adapted from Morris *et al.* [17].

5.1 *De novo* Assembly

First *de novo* assembly was undertaken using SPAdes v3.7.1 [3] *de novo* assembler to construct scaffolds from paired end read files. Kmer sizes of 23, 33, 55, 65, 77 & 89 were used in the assembly, with 1 iteration used for error correction, repeat resolution was enabled and the coverage cut-off set to 'off'. Various kmer sizes, coverage cut-offs, repeat masking, and a reference guided assembly approach were used in an attempt to improve assembly quality.

A second *de novo* assembly was undertaken using velvet v1.2.10 *de novo* assembler [30] on paired end read files using a maximum kmer length of 31, coverage cut-off set to auto, coverage mask set to 2, and the '-short' parameter enabled.

A third assembly was undertaken using IDBA-UD [20], to resolve low coverage regions whilst attempting to prevent generation of chimeric fragments during assembly and scaffolding.

5.2 Post Assembly Improvement

The Cryptosporidium assemblies were improved using the Post Assembly Genome Improvement toolkit (PAGIT) [25]: a pipeline consisting of four standalone tools with the aim of improving the quality of genome assemblies. The tools are, in suggested order of execution: ABACAS [2], IMAGE [28], ICORN [19], & RATT [18]. The workflow of this assembly pipeline can be found in Fig. 4. Cyclospora assemblies were not improved due to the lack of a high quality reference genome.

ABACAS: Algorithm Based Automatic Contiguation of Assembled Sequences. ABACAS is a contig-ordering and orientation tool which is driven by alignment of the draft genome against a suitable reference. Suitability of the reference is defined by amino acid similarity of at least 40%. Alignment is performed by NUCmer or PROmer from the MUMmer package [12]: a tool designed for large scale genome alignment. Contigs from the draft assembly are positioned according to alignment to the reference genome, with spaces between the contigs being filled with 'N's, generating a scaffold of the draft assembly. ABACAS was executed using the updated (All 8 chromosomes resolved) *C.parvum* IowaII [1] reference genome with default parameters.

IMAGE: Iterative Mapping and Assembly for Gap Extension. IMAGE uses Illumina paired end reads to extend contigs by closing gaps within the scaffolds of the draft genome assembly. IMAGE uses read pairs where one read aligns to the end of a contig and the other read overhangs beyond the end of the contig into the gap. This gap can then be partially closed using the overhanging sequence and by extending the contig. IMAGE was run in groups of three iterations at kmer sizes of 91, 81, 71, 61, 51, 41, & 31, totalling 21 iterations. Scaffolding was then performed with a minimum contig size of 500, joining contigs with gaps of 300 N's.

ICORN: Iterative Correction of Reference Nucleotides. ICORN was developed to identify small errors in the nucleotide sequence of the draft genome, such as those which may occur due to low base quality scores. It was designed to correct small erroneous indels, and is not suitable for, or capable of, correcting larger indels or misassemblies. ICORN was run using 8 iterations and a fragment size of 300.

RATT: Rapid Annotation Transfer Tool. RATT is an annotation transfer too used to infer orthology/homology between a reference genome and a draft assembly. This is achieved by utilising NUCmer from the MUMmer package to identify shared synteny between annotated features within the reference genome, and sequence within the draft assembly. Annotation files (EMBL format) are produced which contain regions which are inferred to be common features. The regions are filtered and transferred dependant on whether the transfer is between strains (Strain, similarity rate of 50–94%), species (Species, similarity rate of 95–99%), or different assemblies (Assembly, similarity rate of ≥99%). RATT was run using *C. parvum* IowaII annotations in EMBL format, downloaded from CryptoDB, as a reference. The Strain parameter was used to transfer feature annotations to the draft assembly.

5.3 Analysis of Draft Genomes

VNTR's around the reference and draft Cryptosporidium genomes were identified for the purpose of VNTR comparison and polymorphism analysis. Tandem Repeats Finder v4.09 [5] was used to identify VNTR's around the *C. parvum* IowaII reference genome using a matching weight of 2, mismatch and indel penalties of 5, match and indel probabilities of 80 and 10 respectively, minimum score of 50 and maximum period size of 15. The number of VNTRs per gene is included as a heat map in Fig. 8.

5.4 Identification of Misassembly

The Cryptosporidium draft genomes were analysed in two ways (1) by transferring gene annotations from the reference genome to the drafts using RATT, and (2) by aligning the contigs (from IDBA-UD) or scaffolds (from SPAdes/Velvet) from the draft assemblies to the IowaII reference genome. RATT was used to identify the number of genes which were transferred between genomes: it provided a convenient way of identifying putative chimeric regions i.e. regions on a draft chromosome that contained genes from 2 or more reference chromosomes. NUCmer was then used to investigate these putative chimeric regions by performing whole genome alignments. NUCmer (from the MUMmer package [12]) was used with a minimum length of match set to 100, preventing the report of small regions of similarity, a maximum gap of 90, and a minimum cluster length of 65.

5.5 Quality Assessment with Gini

The Gini coefficient for each isolate was calculated and plotted against the number of genes transferred to chimeric regions within the Cryptosporidium genome assemblies (detailed in Sect. 5.4). The coefficient of determination (R^2) was used to calculate the amount of variance in the number of genes transferred to chimeric regions explained by the Gini coefficient. Gini values at window size 1 were calculated and plotted against nAUC to investigate read distribution across each genome.

5.6 Data Visualisation

The *C. parvum* assemblies (UKP2-8) and VNTR annotations were visualised alongside the *C.parvum* IowaII reference genome using the Circos package v0.69 [11]. Mapped reads were visualised using Integrative Genomics Viewer v2.4.16 [26].

6 Results and Discussion for Sequencing and Read Analysis

Table 1 indicates high depth of coverage inequality throughout the genomes, represented by relatively high Gini coefficient values in comparison to that exhibited by the *C. parvum* Iowa II reference genome (0.112), which the mean depth and breadth of coverage (fraction of the reference covered) will not indicate. This appears to be a common issue when sequencing intestinal protozoans from human clinical samples. Paired end read libraries accessed from GenBank, sequenced by the Welcome Trust Sanger Institute (Bioproject PRJEB3213), and those published by Troell *et al.* (Bioproject PRJNA308172), who was attempting to generate whole genome sequences from single cells using whole genome amplification [27], also suffered from very high Gini coefficients, indicating that this problem is not restricted to a single research team. See Fig. 1 for an example of how the Gini value corresponds to actual read depth variation.

Gini granularity curves are presented here as a more complete indication of the coverage over a genome. The premise behind this is based on two shortcomings of using the Gini coefficient alone as a measure of depth of coverage inequality:

- Two genomes with identical ordered coverage arrays will produce identical Lorentz curves, and therefore an identical Gini. This does not take into account the distribution of depth of coverage across the genome.
- The Gini coefficient is known to be confounded by data granularity [16].

Curves generated from Gini granularity analysis are found in Fig. 6. These curves show a similar set of characteristics, which can be defined by two phases:

1. Decline phase: The Gini value (G) decreases quickly as the window size (W) increases.

Table 1. BWA mapping statistics for each assembly. The Gini coefficient was calculated using window size of 1. Cyclospora reads were mapped to the assemblies they were used to generate. Cryptosporidium reads were mapped to appropriate reference genomes for each species: **C. parvum* IowaII, †*C. hominis* TU502. Cyclospora reads were mapped back to the assembly they were used to generate. Included is the the area under the normalised gini granularity curves as an indication of read distribution (see Sect. 4.2).

Sample	Proportion of reads mapped to reference	Fraction of reference genome covered	Average coverage of reference sequence	G | W_1	nAUC
Ch_UKH3	0.903	0.98	34.71	0.237	0.888
Ch_UKH4	0.845	0.96	209.17	0.602	0.786
Ch_UKH5	0.809	0.96	201.92	0.585	0.777
Cp_UKP2	0.9251	1.00	51.8	0.224	0.892
Cp_UKP3	0.8894	0.99	166.42	0.556	0.815
Cp_UKP4	0.8906	0.99	192.48	0.566	0.806
Cp_UKP5	0.8463	0.99	26.86	0.278	0.884
Cp_UKP6	0.816	0.99	104.83	0.255	0.921
Cp_UKP7	0.8905	0.99	77.85	0.556	0.804
Cp_UKP8	0.837	0.98	174.39	0.566	0.796
Cc_7064046	N/A	N/A	51.4	0.556	0.896
Cc_7064047	N/A	N/A	82.87	0.466	0.735
Cc_7064048	N/A	N/A	69.54	0.413	0.820
Cc_7064049	N/A	N/A	49.82	0.256	0.886
Cc_7064050	N/A	N/A	94.07	0.125	0.957
Cc_7064051	N/A	N/A	49.81	0.296	0.892
Cc_7064052	N/A	N/A	18.37	0.713	0.780
Cc_7064053	N/A	N/A	61.23	0.531	0.795
Cc_7064054	N/A	N/A	73.26	0.328	0.857
Cc_7064055	N/A	N/A	55.64	0.379	0.867
Cc_7064056	N/A	N/A	77.92	0.293	0.891
Cc_21_S4	N/A	N/A	66.36	0.455	0.874
Cc_22_S5	N/A	N/A	58.69	0.514	0.904

2. Perturbation phase: The Gini value plateaus, and perturbation increases, as window size increases.

Figure 7 ($G \mid W_1$ plotted against $nAUC$) indicates that there is great variability in the read distribution throughout each sample. Samples of interest are coloured to show how placement within this plot relates to curve characteristics in Fig. 6, and therefore read distribution.

The two phases of the Gini granularity curves (Fig. 6) may be indicative of characteristics of each dataset, and the method by which they were generated. The magnitude of the drop exhibited during the decline phase appears to vary considerably, with some isolates presenting a large decrease in G over smaller window size increases (e.g. Cc_7064052), and others presenting with very little drop. The initial drop in G may be as a result of the window size being less than the insert size of fragments used to generate these reads. Variation in the rate of

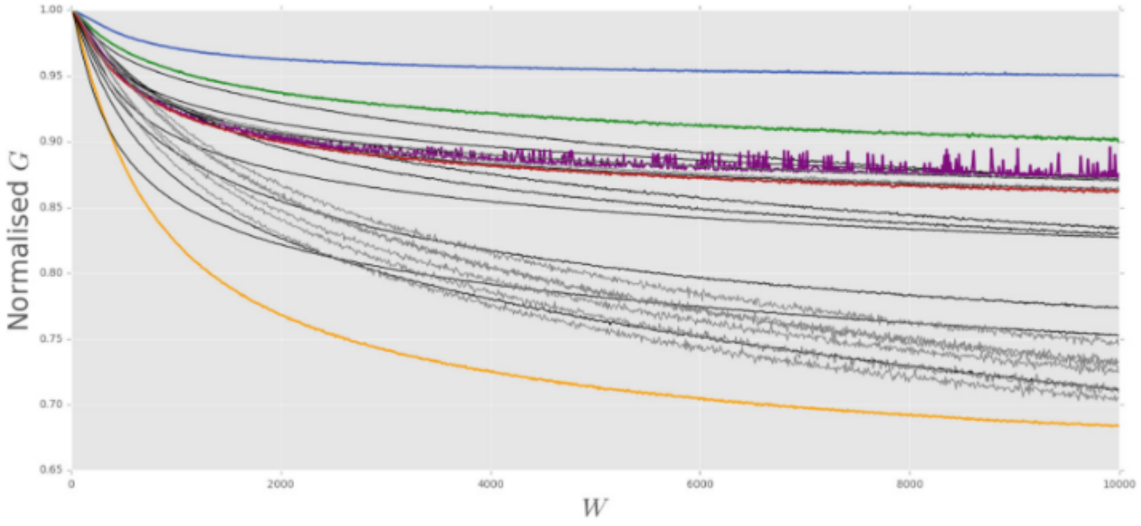


Fig. 5. Normalised Gini granularity curves generated from normalised Gini values (G) calculated using different window sizes (W) for each genome within the Cyclospora and Cryptosporidium dataset. Black samples are Cyclospora, grey samples are Cryptosporidium *spp.* Samples of interest are colourised as: blue = Cc_7064050, red = Cp_UKP5, green = Cp_UKP6, purple = Cc_7064051, and yellow = Cc_7064047. (Color figure online)

drop during the decline phase may indicate higher coverage spike granularity[1]. There appears to be some amount of variation in the characteristics of the curves generated for each species. For example, the Cryptosporidium data (seen in grey) differentiates into two discrete groups in Fig. 7, when using $G \mid W_1$ (the Gini calculated using a window size of 1, and therefore at maximum granularity). This is also seen in Fig. 3:

1. Ch_UKH4, Cp_UKP3, Cp_UKP4, Cp_UKP7, & Cp_UKP8 exhibiting $G = 0.55 - 0.60$
2. Ch_UKH3, Ch_UKH5, Cp_UKP2, Cp_UKP5, & Cp_UKP6 exhibiting $G = 0.22 - 0.28$

The magnitude and length of the decline phase appears to vary depending on the Gini value calculated over single base windows ($W = 1$ or W_1). Greater $G \mid W_1$ values appear to exhibit an extended decline phase (see Table 1). Likewise, the perturbation phase differs between the two groups, where group 1 (higher G) levels off at a much slower rate, and shows large levels of G perturbation, and group 2 (lower G) levels off at a quicker rate and shows lower levels of G perturbation. However, the Cyclospora dataset (seen in black in Fig. 7) does not appear to separate into groups. Furthermore, levels of perturbation does not appear to increase in relation to $G \mid W_1$. The variation in the characteristics of the perturbation phase may be as a result of a number of factors, such as

[1] Spike granularity is used here to describe the density of peaks and troughs in coverage across a sequence, wherein high spike granularity refers to a larger number of peaks and troughs within a sequence.

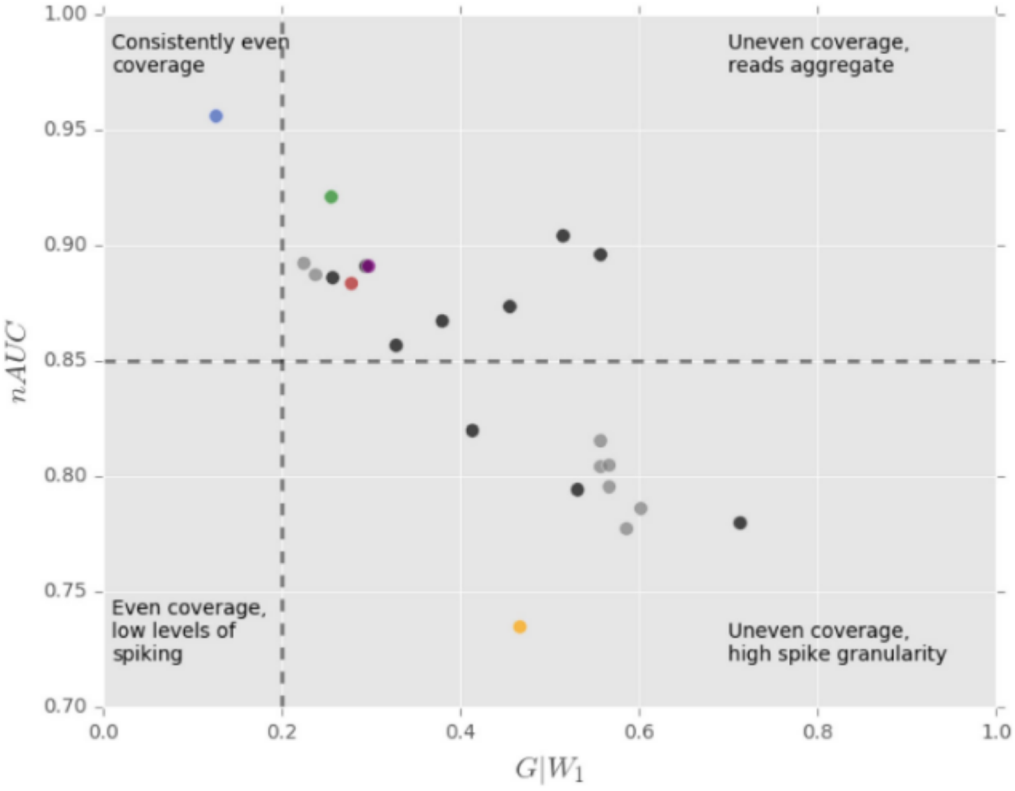


Fig. 6. $G \mid W_1$ plotted against $nAUC$. Quadrants are shown as an indication of how this graph can be interpreted. Black samples are Cyclospora, grey samples are Cryptosporidium *spp.* Samples of interest are colourised as: blue = Cc_7064050, red = Cp_UKP5, green = Cp_UKP6, purple = Cc_7064051, and yellow = Cc_7064047. (Color figure online)

the level of noise within the dataset, genome incompleteness, and the number of contigs within the final assembly. These results indicate that the analysis of these Gini curves hints at coverage features which are lost by considering a single Gini value alone.

Within high Gini isolates, a lower area under the normalised curve indicates uniformly high spike granularity, whereas a higher area under the normalised curve indicates aggregation of reads throughout the genome (see Fig. 6). Analysis of these curves allows for a more comprehensive analysis of problematic genomes with high depth of coverage inequality, wherein high spike granularity indicates a general problem with sequencing, and high read aggregation indicates problems sequencing particular regions. Colourised are curves of particular interest, such as green (Cp_UKP6) and red (Cp_UKP5) which represent the differences which may be exhibited by two genomes with similar $G \mid W_1$. These curves suggest two different distribution types, due to Cp_UKP5 bearing a more pronounced decline phase than Cp_UKP6, and therefore bearing a lower area under the normalised Gini-granularity curve, suggesting greater spike granularity. Figure 1B shows the coverage over chromosome 1 for both of these genomes to be very different in character, despite there being only a 0.002 different in Gini

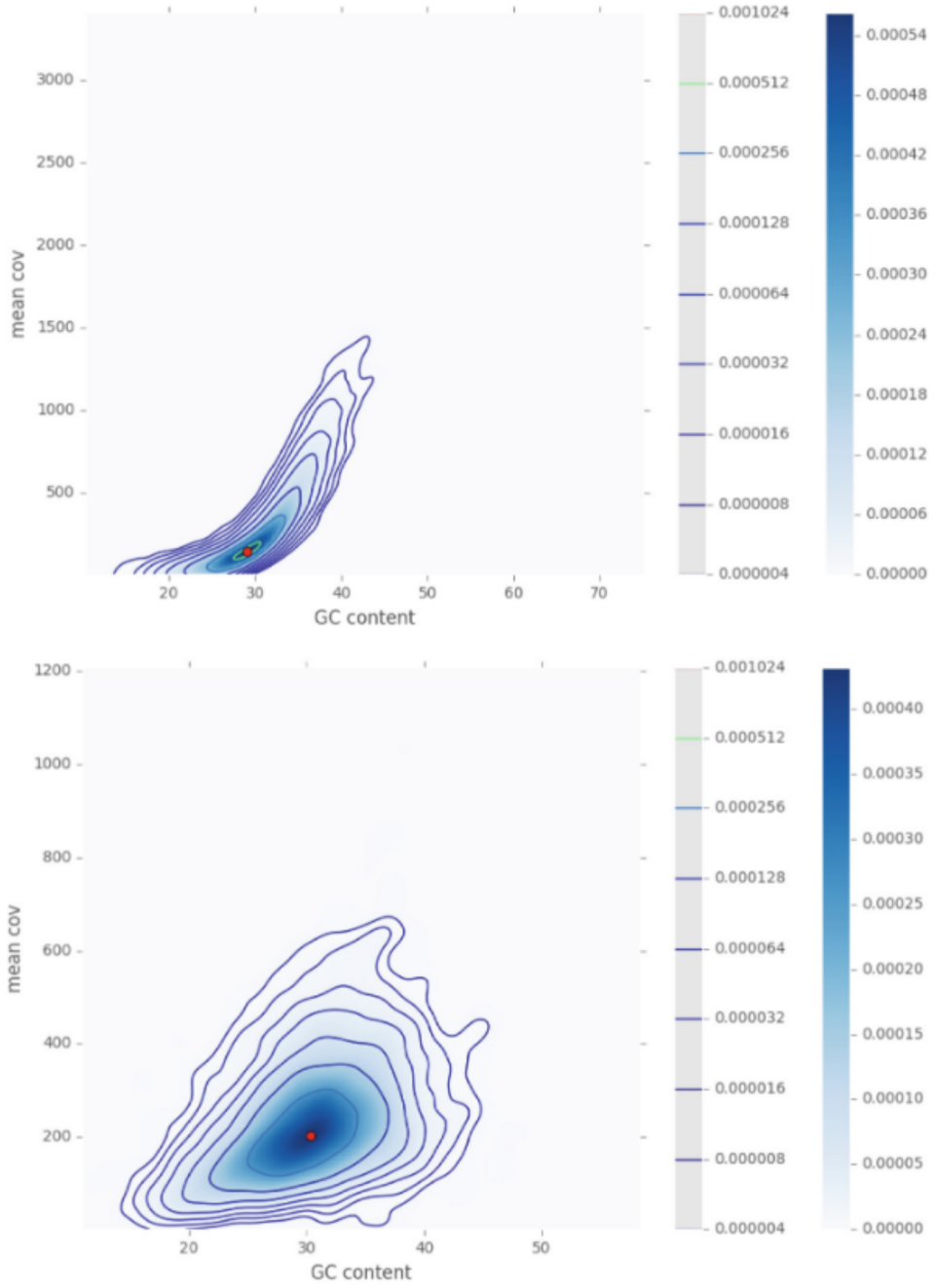


Fig. 7. Above: UKP4 $G = 0.4693$. Below: UKP94 $G = 0.2539$. Coverage vs GC contents plotted within 1000bp windows for 2 UK isolated *C. parvum* genomes. DNA of UKP4 was not subjected to enrichment by a Whole Genome Amplification (WGA) process. DNA of UKP94 was enriched by a WGA process (phi29) prior to sequencing. The plots were generated using kernel density estimation, overlaid with contour lines. The red dot marks the estimated centre of mass of the graph object. (Color figure online)

at absolute granularity. Coverage over Cp_UKP6 appears to present as relatively even, but with localised spikes of coverage reaching and exceeding 1000x (low spike-granularity). In contrast Cp_UKP5 presents with non-localised homogeneous pronounced 'spiking', with little width (high spike-granularity), and reach

similar depth. This should serve as a clear example of how the difference in the nAUC of genomes which bear similar $G \mid W_1$ related to the distribution of read coverage across a genome.

In Fig. 6), the curve generated using Cc_7064050 (blue) demonstrates that this genome is very evenly covered with little inequality of read distribution or coverage, quantified by $G \mid W_1$ and $nAUC$ (0.125 and 0.957). Furthermore, Fig. 7 places this sample in the upper left quadrant, indicating high read coverage equality. Table 2 shows that the spades assembly for this sample is very good, with large mean contig size and n50 (245.5 Kb and 31.6 Kb respectively). However, as can be seen in Table 3, the IDBA assembly is highly fragmented, consisting of 32309 contigs. This indicates that IDBA performs poorly in comparison to SPAdes in instances where the coverage over a genome is even. Gini may therefore be used to indicate which assembler is most appropriate given a particular dataset. Care should be taken when attempting to predict the quality of an assembly from Gini-derived metrics, as they are a measure of the distribution of reads throughout a genome only. In particular, care should be taken as to the window size used in calculating the Gini of coverage over a genome, since the mean contig size for the Cc_7064050 IDBA assembly is 1.433 Kb, which indicates that results may be unreliable when calculating the Gini of depth of coverage over window of a greater size than this.

The curve generated from Cc_7064051 is highlighted in purple as it exhibits some unusual characteristics during the perturbation phase. The level of perturbation during this phase is unusually high, in comparison to the rest of the dataset. There appears to be a trend in the perturbation of $G \mid W_{5000<}$, wherein significant levels of perturbation are interrupted by regions of low perturbation. Furthermore the incidence of these regions of low perturbation reduces as W increases. This is very likely to be an artefact generated by behaviour influenced by the window size and the contig size distribution, since high levels of Gini-granularity curve perturbation is seen in a number of the assemblies which bear fewer contigs and a larger mean contig size.

Considering the effects of Whole Genome Amplification, Fig. 5 shows kernel density estimation plots from 2 isolates of *C. parvum*, where GC content is plotted against mean coverage over windows of 1000 nucleotides. The isolates within this dataset can be split into two cohorts:

1. Genomes which were sequenced from DNA extracted and purified from clinical isolates without any further processing to enrich DNA prior to sequencing.
2. Genomes which were sequenced from DNA extracted and purified from clinical isolates and then enriched using WGA.

The results illustrate much more dispersion in the graphs generated from enriched genomes than those generated from un-enriched genomes. This indicates that the distribution, and subsequent shape of the graph, is at least partially associated with the Gini score. There appear to be two major distinct types of distribution:

Type I. A very tight distribution with a greater than linear distribution, exhibiting a defined increase in coverage at 30% GC content. An example of this type of distribution can be seen in Cp_UKP4 in Fig. 5.

Type II. A radially dispersed distribution with no clear trend, and a centre of mass at 30% GC content. An example of this type of distribution can be seen in Cp_UKP94 in Fig. 5.

These types account for the majority of all distributions seen within the extended Cryptosporidium dataset of 45 genomes. However, there are a small number which do not clearly conform to these two distribution types.

The results detailed in Fig. 5 illustrate that the implementation of WGA as a means to enrich DNA prior to sequencing significantly alters the coverage over the genome. Rather than increasing the mean depth of coverage over the genome uniformly, however, more our analysis highlights that it selectively amplifies certain sequences. Due to the existing GC bias which has been reported within Illumina sequencing data [4] (illustrated by Type I distribution, as seen in Cp_UKP4 in Fig. 5), this has the effect of obscuring this bias, resulting in a much more radially dispersed distribution with a far less clear positive correlation between coverage and GC content. These results highlight the value of using DNA enrichment by WGA for generating high quality, reliable genome assemblies from clinically isolated samples of gastrointestinal Apicomplexan parasites. However, it also hints at a complex relationship between depth of coverage and sequence content across enriched genomes, which warrants further investigation.

7 Results and Discussion for Assembly and Post-assembly Processing

After *de novo* assembly, both the Cryptosporidium *de novo* assemblies were run through the PAGIT pipeline to make the improvements described in the methods section, including gap closing and the transfer of gene annotations. The results can be found in Tables 2 and 3. The results from assembly with Velvet were comparable to that of SPAdes, and therefore are not shown here. The SPAdes assemblies required fewer gaps to be closed by IMAGE. The mean percentage of genes transferred by RATT to the improved SPAdes assemblies is >99%. The mean percentage of genes transferred to chimeric regions is 10.6%.

Table 3 shows the results of assembly using IDBA-UD, and subsequent improvement and annotation using PAGIT. These genomes benefited greatly from gap closure by IMAGE over those produced by SPAdes (see Tables 2 and 3), since gaps in intragenic repetitive regions were much more common, potentially confounding VNTR analysis. The mean percentage of genes transferred by RATT to the improved IDBA-UD assemblies is 98%. The mean percentage of genes transferred to chimeric regions is 0.2%. In the IDBA-UD assemblies, the *C. hominis* genomes performed slightly worse, with 0, 44, and 32 genes transferred to chimeric regions respectively across UKH3, UKH4, and UKH5. Cyclospora post-PAGIT statistics are not available due to the lack of a reference genome, preventing reference guided scaffolding and improvement using PAGIT.

Table 2. The assembly statistics (SPAdes and post-PAGIT) include the number of scaffolds (No.), scaffold N50 metric, scaffold mean length (Av.), and the total size of the final assembly. The assembly size for Cryptosporidium is after improvement using PAGIT, and for Cyclospora is of the *de novo* assembly without improvement. Gene annotations were transferred to Cryptosporidium assemblies by RATT out of a total of 3805 gene annotations in the reference assembly. Genes erroneously transferred refers to genes transferred to regions which have been identified as chimeric (and therefore misassemblies). Within *C. hominis*, the erroneous transfers are putative, due to differences between *C. parvum* and *C. hominis*. Due to Cyclospora being devoid of a suitable reference genome, IMAGE and RATT were not utilised on these assemblies.

Isolate	Total length before PAGIT: No. N50 Av. (kb)	Assembly size (kb)	Gaps closed by IMAGE	Genes transferred: all (erroneously)
Ch_UKH3	168 149.9 54.0	9293	12	3792 (401)
Ch_UKH4	522 57.4 17.5	9594	95	3791 (467)
Ch_UKH5	463 54.6 19.6	9357	92	3787 (496)
Cp_UKP2	157 216.0 58.2	9254	23	3720 (356)
Cp_UKP3	270 109.8 33.7	9336	23	3688 (453)
Cp_UKP4	235 175.2 38.7	9226	22	3770 (349)
Cp_UKP5	447 70.7 20.3	9271	51	3800 (430)
Cp_UKP6	689 332.6 14.1	9826	13	3731 (96)
Cp_UKP7	521 62.6 17.3	9257	19	3797 (475)
Cp_UKP8	369 93.0 24.7	9473	26	3803 (518)
Cc_7064046	41279 4.2 2.1	86061	N/A	N/A
Cc_7064047	34386 8.0 2.2	75256	N/A	N/A
Cc_7064048	8526 10.8 5.2	44656	N/A	N/A
Cc_7064049	1846 111.1 24.3	44771	N/A	N/A
Cc_7064050	1429 245.5 31.6	45117	N/A	N/A
Cc_7064051	2753 167.4 16.6	45628	N/A	N/A
Cc_7064052	48274 1.6 1.3	64415	N/A	N/A
Cc_7064053	51427 1.6 1.3	67729	N/A	N/A
Cc_7064054	3019 34.2 14.8	44689	N/A	N/A
Cc_7064055	12577 35.8 4.2	52497	N/A	N/A
Cc_7064056	1507 94.2 29.5	44459	N/A	N/A
Cc_21_S4	17470 3.5 2.1	37339	N/A	N/A
Cc_22_S5	22122 3.5 2.0	44314	N/A	N/A

The dramatic decrease in the number of genes transferred to chimeric regions indicates significantly fewer misassemblies in improved genomes generated by IDBA-UD than in those of SPAdes, marking a significant improvement. This indicates the effectiveness of using ABACAS to identify gaps within the IDBA-UD assemblies, and IMAGE to close them, which SPAdes would resolve during assembly.

NUCmer, from the MUMMER package was used to identify misassembly, as detailed in Sect. 5.3. Figure 8 shows the extent of misassembly in the isolate genomes, denoted by coloured bars corresponding to which chromosomes regions

Table 3. Statistics for draft genomes assembled using IDBA-UD as per Table 2.

Isolate	IDBA-UD assembly statistics: No. N50 Av. (kb)	Assembly size (kb)	Gaps closed by IMAGE	Genes transferred: all (erroneously)
Ch_UKH3	419 52.9 21.5	9102.3	104	3757 (0)
Ch_UKH4	627 39.7 14.3	9212.5	229	3688 (44)
Ch_UKH5	619 38.7 14.5	9197.0	247	3699 (32)
Cp_UKP2	360 63.9 25.2	9143.7	241	3776 (0)
Cp_UKP3	563 47.8 16.0	9168.6	312	3767 (1)
Cp_UKP4	509 53.7 17.7	9154.9	292	3772 (0)
Cp_UKP5	1830 11.2 4.8	9273.8	1791	3552 (1)
Cp_UKP6	768 51.4 12.1	9135.7	105	3702 (2)
Cp_UKP7	829 32.0 10.7	9184.0	288	3775 (6)
Cp_UKP8	614 40.7 14.7	9177.8	293	3756 (0)
Cc_7064046	38758 4.0 2.1	83253	N/A	N/A
Cc_7064047	37957 6.0 2.1	79422	N/A	N/A
Cc_7064048	3233 37.0 13.7	44314	N/A	N/A
Cc_7064049	1839 111.1 24.3	44768	N/A	N/A
Cc_7064050	32309 1.9 1.4	46217	N/A	N/A
Cc_7064051	4021 52.5 11.3	45363	N/A	N/A
Cc_7064052	62780 1.2 1.2	73955	N/A	N/A
Cc_7064053	62428 1.6 1.3	80922	N/A	N/A
Cc_7064054	3859 26.9 11.5	44244	N/A	N/A
Cc_7064055	9557 26.0 5.2	49426	N/A	N/A
Cc_7064056	2757 41.5 16.0	44119	N/A	N/A
Cc_21_S4	18760 2.8 1.9	35554	N/A	N/A
Cc_22_S5	23440 2.8 1.8	42504	N/A	N/A

Table 4. The number of VNTR regions missing within the IDBA-UD assemblies pre and post gap closing with IMAGE. Taken from Morris *et al.* [17].

Isolate	VNTR regions missing before IMAGE	VNTR regions missing post-IMAGE
Cp_UKP2	48	7
Cp_UKP3	56	12
Cp_UKP4	63	10
Cp_UKP5	209	33
Cp_UKP6	62	13
Cp_UKP7	62	8
Cp_UKP8	67	13

belong to according to NUCmer. Extensive misassembly was identified in all of the genomes, to varying degrees. The most consistently misassembled chromosome is chromosome 7, with a consistent chromosome 8 misassembly. The most

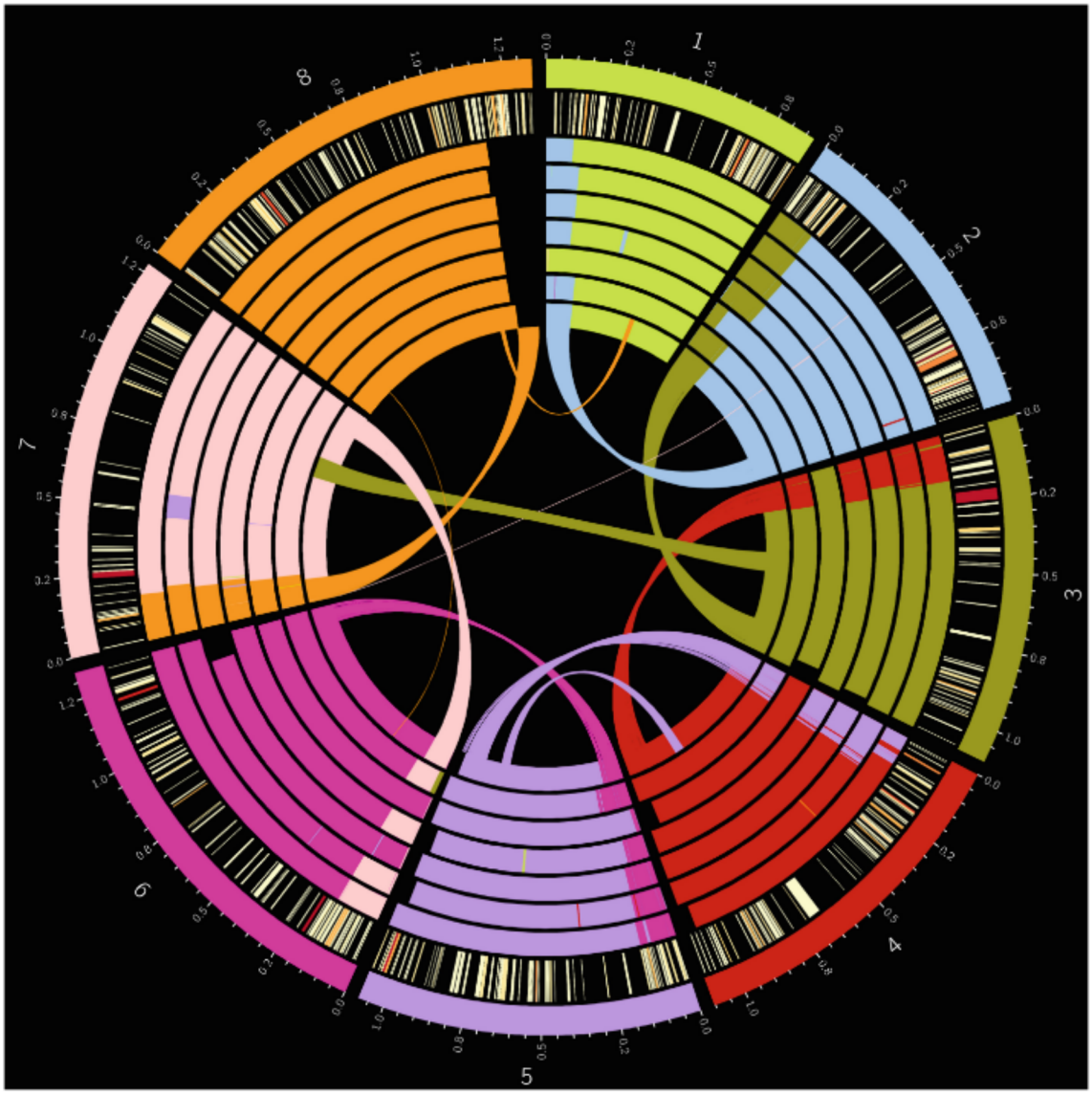


Fig. 8. Misassembled regions on each SPAdes assembled Hadfield *et al. C.parvum* genome. Regions are colour coordinated by which chromosome of the *C.parvum* IowaII reference genome (represented by the outer track) they map to. From outermost to innermost, the inner tracks represent the genomes of each isolate from UKP2-8. The innermost track (UKP8) also includes a linkage map showing precisely where the regions map to in the IowaII reference genome. The second from outer track shows a heatmap of genes bearing Tandem Repeats (TRs), from light yellow denoting a single VNTR within the gene to dark red indicating many TRs within the gene. TRs were identified using Tandem Repeats Finder (see Sect. 5.3). Taken from Morris *et al.* [17]. (Color figure online)

misassembled isolates where UKP3 and UKP8, with 8 misassemblies of larger than 10kb. These two isolates have very high Gini scores (see Table 1), of 0.550 and 0.556 respectively.

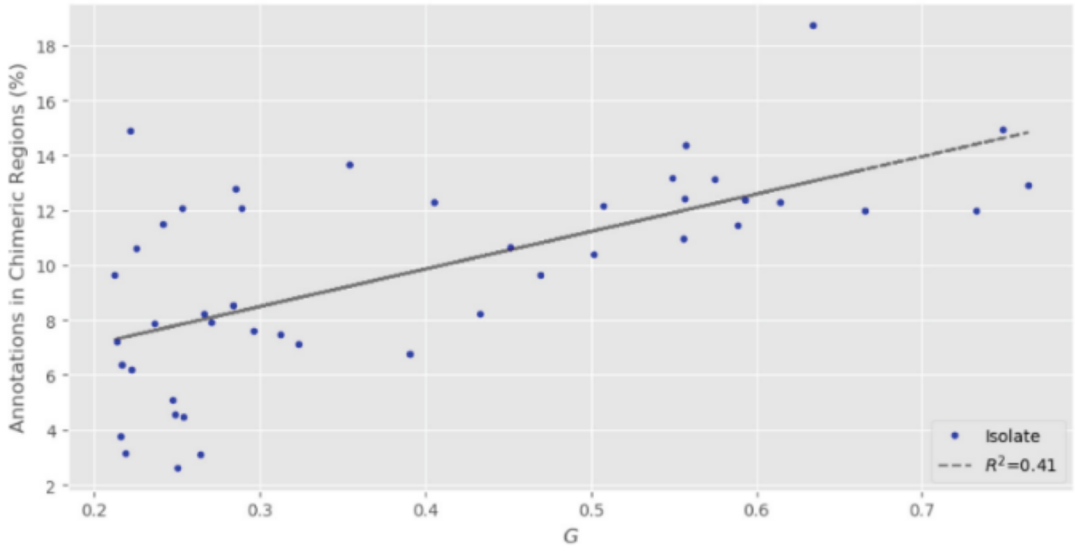


Fig. 9. The percentage of genes transferred to chimeric (misassembled) regions against Gini coefficient of coverage for 45 isolates of *C.parvum* and *C.hominis*. $R^2 = 0.41$

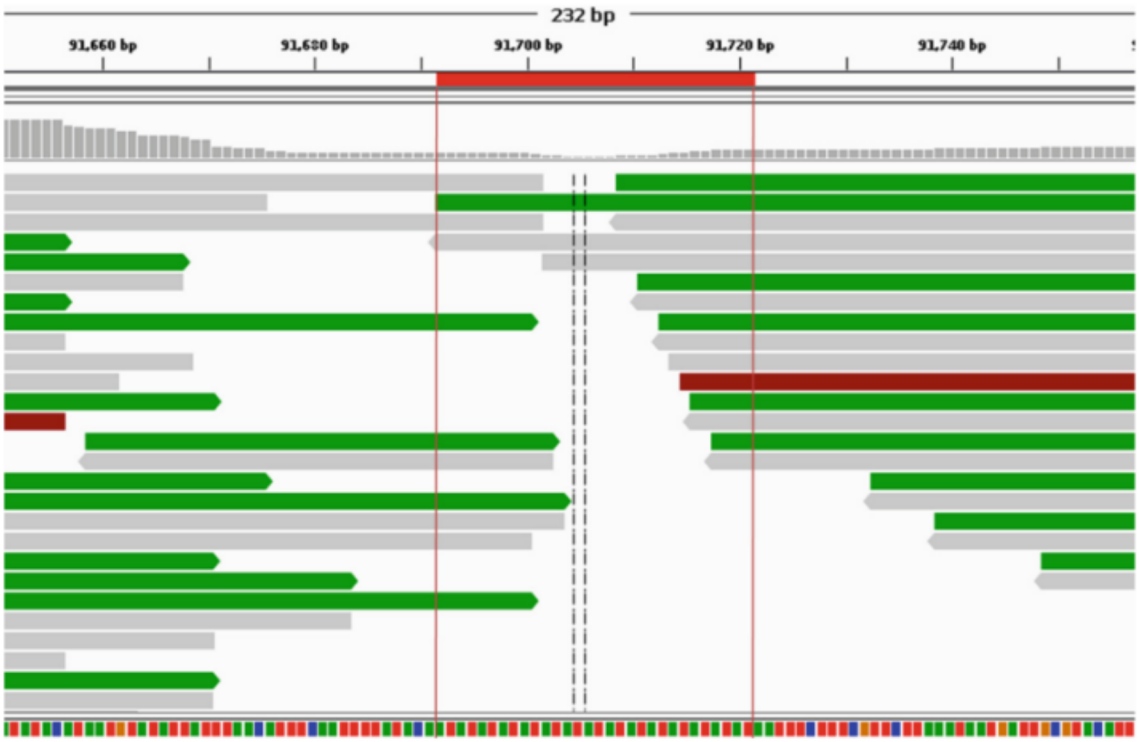


Fig. 10. The misassembly interface between fragments from chromosomes 8 and 7 on the chimeric chromosome 7 of UKP3. Single reads are shown, as is a colourised sequence track (A = Green, T = Red, C = Blue, G = Orange) at the bottom where the repeat region implicated in the formation of this chimeric contig can be seen. Image produced using IGV. Taken from Morris *et al.* [17]. (Color figure online)

Figure 9 illustrates a moderate correlation ($R^2 = 0.41$) between the Gini coefficient and number of misplaced genes within misassembled chromosomal regions across 45 isolates of the extended *C. parvum* and *C. hominis* dataset.

Table 4 shows the number of VNTR regions that were missing from the *C. parvum* IBDA-UD assemblies before and after gap closure with IMAGE. These results show that a large amount of VNTR regions were resolved using IMAGE, indicating the importance of post-assembly genome improvement in the generation of accurate and reliable genome assemblies.

Whole genome alignments were used to identify *in silico* translocation events (considered putative misassemblies), as detailed in Sect. 5.3. Figure 8 shows putatively misassembled regions (translocations) within the *C. parvum* UKP2-8 [7] PAGIT-improved SPAdes assemblies. A heatmap showing the number of VNTR's per coding sequence (CDS) is included. Every genome assembly within the dataset exhibits significant misassembly across all chromosomes, particularly at the terminal ends. Figure 8 illustrates that translocation occurred in a similar fashion throughout each of the assemblies, with the same areas being merged into similar chimeric genomes, as can be seen in chromosome 7, where the initial 120 kb region has merged into the end of chromosome 8 throughout all of the genomes. It is interesting to note that only on UKP3 was a 70 kb area from chromosome 5 seen starting at 500 kb on chromosome 7. Similarly only in UKP8 was a unique 70 kb translocated region seen in chromosome 7 from chromosome 3. These two genomes bear high Gini coefficients, as detailed in Table 1, which may contribute to this. A peculiarity of these misassemblies is the observed trend of chimeric chromosomes being a result of the native chromosome being flanked upstream by 80 kb of the downstream extreme portion of the subsequent chromosome. This is illustrated very clearly in Fig. 8.

Taxonomic evaluation carried out by Hadfield *et al.* utilising the gp60 marker show that there are five gp60 subtypes within the *C. parvum* dataset. This variation within the Hadfield *C. parvum* isolates is supported by Perez-Cordon *et al.* [21] which shows clear variation across 28 VNTR loci, suggesting a number of genetic lineages. The very low likelihood of similar translocation occurring across different populations of *C. parvum* indicates that these events are as a result of misassembly by SPAdes, rather than a biological observations.

Examination of one such chimeric contig (the chr8-chr7 chimeric region at 0-0.14Mb of UKP3 on Fig. 8) revealed that the region has very low depth of coverage, with no single read spanning the chromosomal fragments. Moreover, the sequences from different chromosomes are joined using a simple "AT" repetitive region with only three reads spanning the repeat region and no reads pairing across it (see Fig. 10). This was observed in a number of other chimeric interface regions. Due to the low complexity, high repeat rich nature of the Cryptosporidium genome, coupled with the difficulties associated with DNA extraction and sequencing of this parasite, there is insufficient evidence to suggest that this represents true biological variation. Instead, it may be attributed to a misassembly by the Spades software. This kind of assembly error was also typical of the assemblies produced by using Velvet *de novo* assembler.

Unlike SPAdes, the IDBA assembler leaves these sequence fragments unjoined, with the result that significantly less chimeric regions are seen in the IDBA assemblies. This is because IDBA is designed for the task of assembling

genomes of highly uneven depth of coverage. Although IDBA-UD did not create so many chimeric contigs, the low complexity regions were often left unassembled, with the result that CDS regions contained gaps. Unfortunately, these gaps often included the VNTRs that might be suitable for incorporation into a multi-locus genotyping scheme.

Both SPAdes and Velvet (data from Velvet not shown) produced full, ungapped CDS regions (see Table 2). Thus the IDBA assemblies were not suitable for VNTR analysis and further biomarker identification without significant improvement. PAGIT was used to improve the genomes from all assemblers (see Sect. 5.2), and this improved the resolution of low complexity regions within the IDBA-UD assemblies. Within PAGIT, ABACAS performs scaffolding on the genome assemblies and introduces gaps across the unassembled regions, the IMAGE tool then performs gap closure on these regions, resulting in high quality intragenic VNTR's for biomarker analysis. The number of gaps closed within the IDBA-UD assemblies was significantly higher than within the SPAdes assemblies. This difference in gaps closed was expected, as IDBA-UD was designed for the purpose of assembling genomes which suffer from poor depth of coverage equality, and is therefore more conservative in extending reads across regions with shallow coverage.

The *C. parvum* assemblies produced by IDBA-UD and PAGIT exhibited very few misassemblies compared to the SPAdes assemblies. However, the *C. hominis* genomes suffered from a greater amount of putative misassemblies within the IDBA-UD genomes, as measured by the number of genes being transferred between chromosomes. Note that, genes are transferred from the *C. parvum* IowaII reference genome, which is as different, albeit similar species, and so some biological changes may be expected. Further analysis is required to fully eliminate assembly error as a cause of these chromosomal translocations. Table 4 shows that IMAGE is essential within this workflow for the resolution of repetitive regions which are not resolved during assembly with IDBA-UD. The results show a five to six-fold decrease in the number of VNTR regions missing within the assemblies.

8 Conclusion

In this paper we have performed a detailed analysis of genome sequencing and assembly on 23 genomes from 2 genera of gastrointestinal Apicomplexans.

To investigate sequencing depth and breadth of coverage, we have developed a novel approach that uses the Gini coefficient to determine coverage inequality. We also present a novel technique which allows for further investigation of depth of coverage inequality by generating Gini-granularity curves. We demonstrate how these curves characterise the distribution of reads across a genome and relate this to the quality of subsequent genome assemblies.

We have demonstrated that the use of WGA to enrich DNA within clinical samples is a viable way of increasing read coverage. However, these results also suggest that there is a complex relationship between the selectivity of amplified

DNA during WGA, and its sequence content, which is not explained by GC content alone. Due to the protocol required to extract DNA from clinical samples, these genome sequences often have highly uneven sequencing depth even if the coverage across the genome sequence is relatively high.

We found the SPAdes and Velvet assemblies to be problematic on our datasets. This led to misassemblies across low coverage, low complexity regions, resulting in the creation of chimeric chromosomes: up to 15% of all genes were being placed within these chimeric chromosomes. Although the assemblies generated by IDBA-UD did not suffer from the problem of chimeric sequences, they were problematic due to a different assembly approach, leading to a large number of gaps, particularly in repetitive regions. This is a significant issue because these gaps often contained the VNTR sequences that are important to us for developing new clinical genotyping strategies. However, the IMAGE gap closing tool from the genome improvement pipeline, PAGIT, was able to resolve these missing low complexity regions. Using this strategy, of assembly with IDBA followed by gap closing with IMAGE, we will be able to perform more in depth VNTR analysis with the intention of identifying biomarkers that will facilitate the development of novel prevention strategies in the fight against the diseases caused by these organisms.

References

1. Abrahamsen, M.S., et al.: Complete genome sequence of the Apicomplexan, Cryptosporidium parvum. Science **304**(5669), 441–445 (2004). https://doi.org/10.1126/science.1094786. http://www.ncbi.nlm.nih.gov/pubmed/15044751
2. Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., Berriman, M.: ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics **25**(15), 1968–1969 (2009). https://doi.org/10.1093/bioinformatics/btp347
3. Bankevich, A., et al.: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**(5), 455–477 (2012). https://doi.org/10.1089/cmb.2012.0021
4. Benjamini, Y., Speed, T.P.: Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. **40**(10), 1–14 (2012). https://doi.org/10.1093/nar/gks001
5. Benson, G.: Tandem repeats finder: a program to analyse DNA sequences. Nucleic Acids Res. **27**(2), 573–578 (1999)
6. Chalmers, R.M., et al.: Suitability of loci for multiple-locus variable-number of tandem-repeats analysis of Cryptosporidium parvum for inter-laboratory surveillance and outbreak investigations. Parasitology **144**(1), 37–47 (2017). https://doi.org/10.1017/S0031182015001766
7. Hadfield, S.J., et al.: Generation of whole genome sequences of new Cryptosporidium hominis and Cryptosporidium parvum isolates directly from stool samples. BMC Genom. **16**, 650 (2015). https://doi.org/10.1186/s12864-015-1805-9. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1805-9
8. Hosono, S., et al.: Unbiased whole-genome amplification directly from clinical samples. Genome Res. **13**(5), 954–964 (2003). https://doi.org/10.1101/gr.816903

9. Ifeonu, O.O., et al.: Annotated draft genome sequences of three species of Cryptosporidium: Cryptosporidium meleagridis isolate UKMEL1, C. baileyi isolate TAMU-09Q1 and C. hominis isolates TU502 2012 and UKH1. Pathogens Dis. (2016). https://doi.org/10.1093/femspd/ftw080
10. Jones, E., Oliphant, T., Peterson, P., Al, E.: SciPy: open sourcescientific tools for Python (2001)
11. Krzywinski, M., et al.: Circos. Genome Res. **19**(9), 1639–1645 (2009). https://doi.org/10.1186/1471-2105-14-244. http://genome.cshlp.org/content/19/9/1639.short
12. Kurtz, S., et al.: Versatile and open software for comparing large genomes. Genome Biol. **5**(2), R12 (2004). https://doi.org/10.1186/gb-2004-5-2-r12. http://genomebiology.com/2004/5/2/R12
13. Lasken, R.S., Egholm, M.: Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. Trends Biotechnol. **21**(12), 531–535 (2003). https://doi.org/10.1016/j.tibtech.2003.09.010
14. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). https://doi.org/10.1093/bioinformatics/btp324
15. Marques, D.F., et al.: Cyclosporiasis in travellers returning to the United Kingdom from Mexico in summer 2017: lessons from the recent past to inform the future. Eurosurveillance (2017). https://doi.org/10.2807/1560-7917.ES.2017.22.32.30592
16. Monfort, P.: Convergence of EU regions - measures and evolution. Eur. Union **Europa**(6), 1–32 (2008)
17. Morris, A.V., Pachebat, J., Robinson, G., Chalmers, R., Swain, M.: Identifying and resolving genome misassembly issues important for biomarker discovery in the protozoan parasite, cryptosporidium. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS, vol. 3, pp. 90–100. SciTePress (2019). https://doi.org/10.5220/0007397200900100
18. Otto, T.D., Dillon, G.P., Degrave, W.S., Berriman, M.: RATT: rapid annotation transfer tool. Nucleic Acids Res. **39**(9), 1–7 (2011). https://doi.org/10.1093/nar/gkq1268
19. Otto, T.D., Sanders, M., Berriman, M., Newbold, C.: Iterative correction of reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics **26**(14), 1704–1707 (2010). https://doi.org/10.1093/bioinformatics/btq269
20. Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L.: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28**(11), 1420–1428 (2012). https://doi.org/10.1093/bioinformatics/bts174
21. Perez-Cordon, G., Robinson, G., Nader, J., Chalmers, R.M.: Discovery of new variable number tandem repeat loci in multiple Cryptosporidium parvum genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. Exp. Parasitol. **169**(August), 119–128 (2016). https://doi.org/10.1016/j.exppara.2016.08.003
22. Puiu, D., Enomoto, S., Buck, G.A., Abrahamsen, M.S., Kissinger, J.C.: CryptoDB: the Cryptosporidium genome resource. Nucleic Acids Res. **32**(90001), 329D–331 (2004). https://doi.org/10.1093/nar/gkh050. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh050
23. Qvarnstrom, Y., et al.: Draft genome sequences from Cyclospora cayetanensis oocysts purified from a human stool sample. Genome Announc. (2015). https://doi.org/10.1128/genomeA.01324-15

24. Sow, S.O., et al.: The Burden of Cryptosporidium diarrheal disease among children <24 months of age in moderate/high mortality regions of Sub-Saharan Africa and South Asia, utilizing data from the Global Enteric Multicenter Study (GEMS). PLoS Negl. Trop. Dis. **10**(5), 1–20 (2016). https://doi.org/10.1371/journal.pntd.0004729
25. Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M., Otto, T.D.: A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat. Protocols **7**(7), 1260–84 (2012). https://doi.org/10.1038/nprot.2012.068. http://www.nature.com/doifinder/10.1038/nprot.2012.068%5Cn
26. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinf. **14**(2), 178–192 (2013). https://doi.org/10.1093/bib/bbs017
27. Troell, K., et al.: Cryptosporidium as a testbed for single cell genome characterization of unicellular eukaryotes. BMC Genom. **17**(1), 1–12 (2016). https://doi.org/10.1186/s12864-016-2815-y. http://dx.doi.org/10.1186/s12864-016-2815-y
28. Tsai, I.J., Otto, T.D., Berriman, M.: Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. **11**(4), R41 (2010). https://doi.org/10.1186/gb-2010-11-4-r41
29. Xu, P., et al.: The Genome of Cryptosporidium hominis. Lett. Nat. **431**(October), 1107–1112 (2004). https://doi.org/10.1038/nature02990
30. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**(5), 821–829 (2008). https://doi.org/10.1101/gr.074492.107
31. Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., Arnheim, N.: Whole genome amplification from a single cell: implications for genetic analysis. Proc. Natl. Acad. Sci. **89**(13), 5847–5851 (2006). https://doi.org/10.1073/pnas.89.13.5847

Identifying and Resolving Genome Misassembly Issues Important for Biomarker Discovery in the Protozoan Parasite, Cryptosporidium

Arthur Morris[1], Justin Pachebat[1], Guy Robinson[2], Rachel Chalmers[2] and Martin Swain[1]
[1]*IBERS, Aberystwyth University, Aberystwyth, U.K.*
[2]*Cryptosporidium Reference Unit, Public Health Wales, Swansea, U.K.*

Keywords: Genomics, Cryptosporidium, Assembly, Biomarker Discovery, Gini, Clinical Microbiology, Pathogen Genomics.

Abstract: Cryptosporidium is a protozoan parasite that causes a diarrhoeal disease in humans, and which may be spread by swimming pools or infected municipal water supplies. It can be a serious health risk for individuals with weakened immune systems. Genomics has the potential to help control this pathogen, but until recently, it has not been possible to perform whole genome sequencing directly from human stool samples. This is no longer the case, and there are now at least a dozen high quality genomes available via resources like CryptoDB and NCBI, with other isolates being sequenced. The analysis of these genomes will improve current approaches for tracking sources of contamination and routes of transmission by allowing the identification of biomarkers, such as multiple-locus variable tandem repeat regions (VNTRs). However, problems remain due to highly uneven sequence coverage, which causes serious errors and artefacts in the genome assemblies produced by a number of popular assemblers. Here we discuss these assembly issues, and describe our strategy to generate genome assemblies of sufficient quality to enable the discovery of new VNTR biomarkers.

1 INTRODUCTION

Cryptosporidium is an Apicomplexan parasite causing gastrointestinal disease (Cryptosporidiosis) in humans and animals. In the developing world, Cryptosporidium is one of the main causes of childhood morbidity. A recent large-scale study has evaluated the aetiology, burden and clinical syndromes of moderate-to severe diarrhoea across seven sites in sub-Saharan Africa and South Asia. It identified Cryptosporidium as contributing to approximately 202,000 deaths per year in children less than 24 months old (Sow et al., 2016). In the UK, *C. parvum* and *C. hominis* cause most cases of Cryptosporidiosis. While self-limiting after prolonged duration of symptoms (2-3 weeks) in immunocompetent hosts, severely immunocompromised patients suffer severe, sometimes life threatening disease. *C.parvum* has a small, very compact genome, with the IowaII (Abrahamsen et al., 2004) reference exhibiting a 9.1Mb genome, bearing 3,865 genes, of which 89.1% are intronless.

The sequencing and assembly of whole or partial genomes has become an essential tool in modern science, facilitating research in every area of biology. A primary concern for Cryptosporidium is extracting from clinical samples sufficient amounts of high quality, low contaminant DNA for sequencing. Without this, sequencing may result in low coverage sequence, variable sequencing depth and poor quality genome assemblies. In the area of Cryptosporidiosis the impact of genomics has been limited by the need to propagate the parasite in animals to generate enough oocysts from which to extract DNA of sufficient quantity and purity for analysis (Abrahamsen et al., 2004). In 2015 this problem was overcome through an approach that now allows genomic Cryptosporidium DNA suitable for whole genome sequencing to be prepared directly from human stool samples (Hadfield et al., 2015). Hadfield *et al.* (2015) applied their method to the whole genome sequencing of eight *C. parvum* and *C. hominis* isolates. Presently, the Cryptosporidium genomics resource, CryptoDB (Puiu et al., 2004), currently gives access to 13 complete genomes, with a total of 10 available from the NCBI.

Currently clinical diagnosis of Cryptosporidium relies on conventional genotyping tests. The availability of whole Cryptosporidium genome sequences provides much higher resolution information for geno-

Morris A., Pachebat J., Robinson G., Chalmers R. and Swain M.
Identifying and Resolving Genome Misassembly Issues Important for Biomarker Discovery in the Protozoan Parasite, Cryptosporidium.
DOI: 10.5220/0007397200900100
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 10th International Conference on Bioinformatics Models, Methods and Algorithms)*, pages 90-100
ISBN: 978-989-758-353-7

typing. In addition, the genomes can be used to study a wide array of aspects of pathogen biology, such as identity, taxonomy in relation to other pathogens, sensitivity or resistance to drugs, development of novel therapeutic agents, virulence, and epidemiology. Our interest is to build on current genotyping tests by developing a standardised multi-locus typing scheme. This will allow sources of contamination and routes of transmission to be characterized and compared in a cost- and time-efficient manner (Perez-Cordon et al., 2016; Chalmers et al., 2017). Here variable-number of tandem-repeats (VNTR) are used, with recent investigations concluding that additional loci need to be identified and validated (Chalmers et al., 2017). Our work is building on that of Perez-Cordon *et al.* (2016), who used Tandem Repeats Finder (Benson, 1999) to identify polymorphic VNTR's around the genome of *C. parvum*, and analysed them for variation across the eight genomes sequenced by Hadfield *et al.* (2015). We aim to use whole genome sequencing of additional isolates and species to help achieve this goal, but this work is hapered by the quality of available genome sequences (Perez-Cordon et al., 2016).

This paper is structured as follows. First, we explain the quality issues associated with genome sequences extracted from clinical stool samples. Then we describe our methods, including the data sets used, a novel metric we use to measure the distribution of read depth in a set of sequenced reads, and the process of assembly with the identification of misassemblies. In the results and discussion sections, we summarise properties of the sequenced reads, show how they can lead to misassemblies, and give evidence of the types of misassembly we encounter. We also describe how our novel metric can explain some of these assembly errors. Finally, we conclude with a brief outline the strategy we use to generate genome assemblies of sufficient quality to use for the discovery of novel VNTRs.

2 THE PROBLEM

Although it is possible to derive high quality Cryptosporidium DNA by culturing the parasite in donor animals (Abrahamsen et al., 2004), this is expensive and time consuming, and is not appropriate for clinical samples, where maintaining sequence identity is essential. Sequencing Cryptosporidium from clinical samples suffers from three major problems:

- The yield of oocysts from clinical samples is low.
- The oocysts are extracted directly from faeces, necessitating extensive cleaning and purification before DNA extraction.
- The DNA yield per oocyst is low.

These three problems commonly result in sequenced data sets with very uneven depth of coverage, which makes assembly and analysis difficult. Uneven sequencing depth has been identified in datasets obtained from published and unpublished paired end read libraries generated by different groups, and which were prepared using the standard Nextera XT DNA sample preparation kit. Uneven sequencing depth may lead to genome misassembly, and we have identified this an issue with a number of popular *de novo* assemblers. Poor quality genome assemblies can find their way into public repositories of genome sequence and this can confound the development of novel prevention strategies, therapeutics, and diagnostic approaches.

3 METHOD

Our initial choice of assembly software was to use SPAdes (Bankevich et al., 2012), following the Hadfield *et al.* (2015) paper. However, after aligning the assembled genomes to the reference genome, and visualising genome features such as genes and VNTRs, a number of issues became apparent (see Figure 4) such as the transfer of large sequence fragments between chromosomes. We assumed this was a computational artefact, rather than a true biological signal, and therefore we have investigated the assembly process in the following manner.

3.1 Dataset

We used the dataset presented by Hadfield *et al.* (Hadfield et al., 2015), consisting of 7 UK isolates of *Cryptosporidium parvum* and 3 UK isolates of *Cryptosporidium hominis*: UKP2 to UKP8 & UKH3 to UKH5. An updated *C. parvum* IowaII reference assembly was utilised, which included all 8 chromosomes resolved, rather than the 18 fragment IowaII assembly (Abrahamsen et al., 2004) that was used by Hadfield *et al*. This dataset was used because they currently represent the largest collection of published Cryptosporidium draft genomes from clinical isolates.

For the purpose of identifying a correlation between genes transferred to chimeric regions and Gini, unpublished isolates consisting of 29 UK *C. parvum* and 19 UK *C. hominis* isolates where also used.

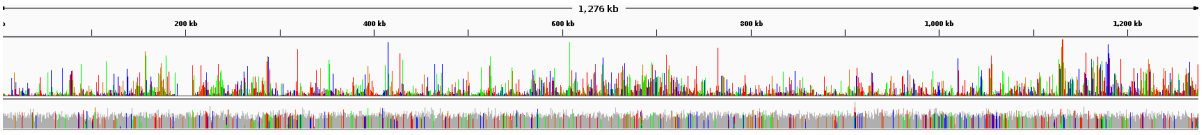


Figure 1: Coverage across chromosome 7 of the *C.parvum* UKP3 (top track) and IowaII reference (bottom track) genomes to illustrate the extreme coverage inequality of the UKP3 isolate genome (UKP3 $Gini = 0.5489$, IowaII $Gini = 0.112$). Image produced using IGV. Note that the IowaII DNA sequences were derived from an animal model, and have low or "normal" read depth variation, whereas UKP3 is more typical of DNA sequences extracted from clinical samples.

3.2 Sequenced Read Analysis

The reads were mapped to a reference genome (*C. parvum* IowaII for *C. parvum* and *C. hominis* TU502 (Xu et al., 2004) for *C. hominis*) using Bowtie2 v2.3.3.1. (Langmead et al., 2009) Coverage analysis was then performed using Samtools v1.5 (Li and Durbin, 2009).

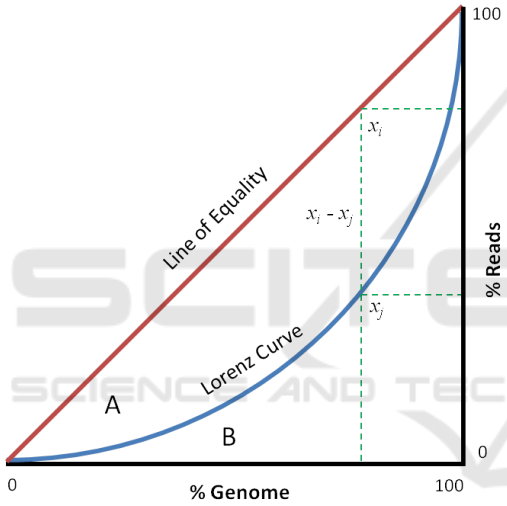


Figure 2: Graphical representation of the Gini coefficient. In this graph, the Gini coefficient can be calculated as $A/(A+B)$, which represented area under the Lorenz curve (blue) inversely proportional to the line of equality (red). The green dotted lines denote the percentage of reads which cover 80% of a genome used to generate the Lorenz curve (poor coverage depth equality) as compared to a perfect distribution of reads.

Read depth was calculated using the 'depth' tool within the samtools package. The Gini coefficient is a measure used to identify inequality in the distribution of a quantifiable metric. It is commonly used in economics to measure income inequality within a population, where it is represented by a value between 0 and 1, with 0 representing perfectly even distribution, and higher values representing higher inequality of distribution. Here we have applied this coefficient to measure inequality of depth of coverage across a genome. For each of the 10 Hadfield genomes, we calculated the Gini coefficient of read depth. The Gini coefficient is defined using the following equation:

$$G = A/(A+B)$$

where A is the area under the line of equality, and B the area under the Lorenz curve, on the graph of distribution inequality (see Figure 2). The green dotted lines (marked at 80% on the x axis) in Figure 2 gives an example of how, in the dataset used to generate the Lorenz curve, 80% of the genome is covered by only 40% of reads (the value at the position of collision of the green dotted line on the y axis), whereas in a perfect distribution it would be covered by 80% of reads.

The algorithm for calculating a genome's Gini coefficient of read depth coverage involves first calculating the mean depth of coverage of 1Kb windows over the genome. These windows are ordered according to their depth of coverage values, and these values rescaled between 0 and 100. This ordered set of read depth values is used to generate the Lorenz curve, L, where the value at every position i on the curve represents the sum of all values at positions $\leq i$. A line of equality, E, was generated to represent perfectly even distribution of reads across a genome. The difference between the values at each position on E and L is then calculated and the summed inverse proportional difference (The Gini coefficient) of these values calculated. This was performed using the following equation:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n \sum_{i=1}^{n} x_i}$$

where n refers to the number of windows (read depth values) across the genome, x_i is a depth of coverage value at position i on the line of equality E, and x_j is the value at position j on the Lorenz curve L.

The Gini coefficient for each genome represents the unevenness of read depth across the genome sequence (an example of uneven coverage across chromosome 7 of UKP3 as compared to Iowa II can be seen in Figure 1).

3.3 *De novo* Assembly

First *de novo* assembly was undertaken in the same manner as those reported by Hadfield *et al.* (2015). SPAdes v3.7.1 (Bankevich et al., 2012) *de novo* assembler was used to construct scaffolds from paired end read files. Kmer sizes of 23, 33, 55, 65, 77 & 89 were used in the assembly, with 1 iteration used for error correction, repeat resolution was enabled and the coverage cut off set to 'off'. Various kmer sizes, coverage cut-offs, repeat masking, and a reference guided assembly approach were used in an attempt to improve assembly quality.

A second *de novo* assembly was undertaken using velvet v1.2.10 *de novo* assembler (Zerbino and Birney, 2008) on paired end read files using a maximum kmer length of 31, coverage cut-off set to auto, coverage mask set to 2, and the '-short' parameter enabled.

A third assembly was undertaken using IDBA-UD (Peng et al., 2012), to resolve low coverage regions whilst attempting to prevent generation of chimeric fragments during assembly and scaffolding.

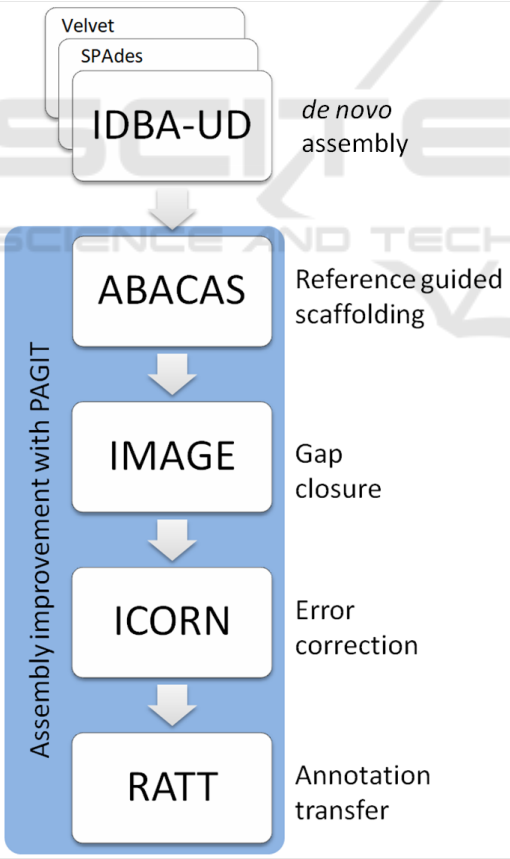


Figure 3: The workflow for assembly, adapted from that used by Hadfield *et al.* for the assembly of genomes with high coverage depth inequality.

3.4 Post Assembly Processing

The assemblies were improved using the Post Assembly Genome Improvement toolkit (PAGIT) (Swain et al., 2012): a pipeline consisting of four standalone tools with the aim of improving the quality of genome assemblies. The tools are, in suggested order of execution: ABACAS (Assefa et al., 2009), IMAGE (Tsai et al., 2010), ICORN (Otto et al., 2010), & RATT (Otto et al., 2011).

The workflow of this assembly pipeline can be found in Figure 3.

3.4.1 ABACAS: Algorithm based Automatic Contiguation of Assembled Sequences

ABACAS is a contig-ordering and orientation tool which is driven by alignment of the draft genome against a suitable reference. Suitability of the reference is defined by amino acid similarity of at least 40%. Alignment is performed by NUCmer or PROmer from the MUMmer package (Kurtz et al., 2004): a tool designed for large scale genome alignment. Contigs from the draft assembly are positioned according to alignment to the reference genome, with spaces between the contigs being filled with 'N's, generating a scaffold of the draft assembly.

ABACAS was executed using the updated (All 8 chromosomes resolved) *C.parvum* IowaII (Abrahamsen et al., 2004) reference genome with default parameters.

3.4.2 IMAGE: Iterative Mapping and Assembly for Gap Extension

IMAGE uses Illumina paired end reads to extend contigs by closing gaps within the scaffolds of the draft genome assembly. IMAGE uses read pairs where one read aligns to the end of a contig and the other read overhangs beyond the end of the contig into the gap. This gap can then be partially closed using the overhanging sequence and by extending the contig.

IMAGE was run in groups of three iterations at kmer sizes of 91, 81, 71, 61, 51, 41, & 31, totalling 21 iterations. Scaffolding was then performed with a minimum contig size of 500, joining contigs with gaps of 300 N's.

3.4.3 ICORN: Iterative Correction of Reference Nucleotides

ICORN was developed to identify small errors in the nucleotide sequence of the draft genome, such as those which may occur due to low base quality scores. It was designed to correct small erroneous indels, and

is not suitable for, or capable of, correcting larger indels or misassemblies.

ICORN was run using 8 iterations and a fragment size of 300.

3.4.4 RATT: Rapid Annotation Transfer Tool

RATT is an annotation transfer too used to infer orthology/homology between a reference genome and a draft assembly. This is achieved by utilising NUCmer from the MUMmer package to identify shared synteny between annotated features within the reference genome, and sequence within the draft assembly. Annotation files (EMBL format) are produced which contain regions which are inferred to be common features. The regions are filtered and transferred dependant on whether the transfer is between strains (Strain, similarity rate of 50-94%), species (Species, similarity rate of 95-99%), or different assemblies (Assembly, similarity rate of >=99%).

RATT was run using IowaII annotations in EMBL format, downloaded from CryptoDB, as a reference. The Strain parameter was used to transfer feature annotations to the draft assembly.

3.5 Analysis of Draft Genomes

VNTR's around the reference and draft genomes were identified for the purpose of VNTR comparison and polymorphism analysis. Tandem Repeats Finder v4.09 (Benson, 1999) was used to identify VNTR's around the *C. parvum* IowaII reference genome using a matching weight of 2, mismatch and indel penalties of 5, match and indel probabilities of 80 and 10 respectively, minimum score of 50 and maximum period size of 15. The number of VNTR's per gene is included as a heat map in Figure 4.

3.6 Identification of Misassembly

The draft genomes were analysed in two ways (1) by transferring gene annotations from the reference genome to the drafts using RATT, and (2) by aligning the contigs (from IDBA-UD) or scaffolds (from SPAdes/Velvet) from the draft assemblies to the IowaII reference genome. RATT was used to identify the number of genes which were transferred between genomes: it provided a convenient way of identifying putative chimeric regions i.e. regions on a draft chromosome that contained genes from 2 or more reference chromosomes. NUCmer was then used to investigate these putative chimeric regions by performing whole genome alignments. NUCmer (from the MUMmer package (Kurtz et al., 2004)) was used with a minimum length of match set to 100, preventing the report of small regions of similarity, a maximum gap of 90, and a minimum cluster length of 65.

3.7 Quality Assessment with Gini

The Gini coefficient for each isolate was calculated and plotted against the number of genes transferred to chimeric regions (detailed in section 3.6). The coefficient of determination (R^2) was used to calculate the amount of variance in the number of genes transferred to chimeric regions explained by the Gini coefficient.

3.8 Data Visualisation

The *C.parvum* assemblies (UKP2-8) were visualised alongside the *C.parvum* IowaII reference genome using the Circos package v0.69 (Krzywinski et al., 2009). Mapped reads were visualised using Integrative Genomics Viewer v2.4.16 (Thorvaldsdóttir et al., 2013).

4 RESULTS

Statistics from the sequencing of the Hadfield *et al.* genomes can be found in Table 1. The Gini coefficient values are high (>0.25) in five of the ten paired end read libraries. See Figure 1 for an example of how the Gini value corresponds to actual read depth variation within UKP3 and IowaII. Apart from the variation in read depth, the sets of sequences generally appear to be of good quality, with high genome coverage, and little sign of contamination.

Table 2 shows the results of assembly using SPAdes. The results from assembly with Velvet were comparable to that of SPAdes, and therefore are not shown here. Table 3 shows the results of assembly using IDBA-UD. The results shown in these tables indicate that SPAdes produced assemblies with longer and fewer contigs than IDBA-UD, highlighting the differences between the assembly approaches adopted by the assemblers.

Both the assemblies were then run through the PAGIT pipeline to make the improvements described in the methods section, including gap closing and the transfer of gene annotations. The results can be found in Tables 2 and 3. The SPAdes assemblies required fewer gaps to be closed by IMAGE. The mean percentage of genes transferred by RATT to the improved SPAdes assemblies is >99%. The mean percentage of genes transferred to chimeric regions is 10.6%.

Table 1: Bowtie2 mapping statistics for *C. parvum* and *C. hominis* reads generated by Hadfield *et al.*. The Gini coefficient is included in this table as an indication of uneven depth of coverage (IowaII=0.112). **C. parvum* IowaII, †*C. hominis* TU502.

Isolate	Total base pairs sequenced (Mb)	Proportion overall read alignment	Fraction of ref. covered	Average cov. of ref. seq.	Gini coefficient
UKH3†	305.02	0.903	0.98	34.71	0.1634
UKH4†	1828.87	0.845	0.96	209.17	0.4935
UKH5†	1765.46	0.809	0.96	201.92	0.2895
UKP2*	426.69	0.819	1.00	46.84	0.2121
UKP3*	1514.83	0.889	0.99	166.42	0.5489
UKP4*	1751.98	0.891	0.99	192.48	0.4693
UKP5*	244.53	0.846	0.99	26.86	0.2895
UKP6*	954.18	0.816	0.99	104.83	0.2106
UKP7*	708.61	0.891	0.99	77.85	0.5494
UKP8*	1587.38	0.837	0.98	174.39	0.5570

Table 2: The assembly statistics (SPAdes and post-PAGIT) include the number of scaffolds (No.), scaffold N50 metric, scaffold mean length (Av.), and the total size of the final assembly. Gene annotations were transferred by RATT out of a total of 3805 gene annotations in the reference assembly. Genes erroneously transferred refers to genes transferred to regions which have been identified as chimeric (and therefore misassemblies). Within *C. hominis*, the erroneous transfers are putative, due to differences between *C. parvum* and *C. hominis*.

Isolate	Total length before PAGIT: No.	N50	Av. (kb)	Assembly size post-PAGIT (kb)	Gaps closed by IMAGE	Genes transferred: all (erroneously)
UKH3	168	149.9	54.0	9293	12	3792 (401)
UKH4	522	57.4	17.5	9594	95	3791 (467)
UKH5	463	54.6	19.6	9357	92	3787 (496)
UKP2	157	216.0	58.2	9254	23	3720 (356)
UKP3	270	109.8	33.7	9336	23	3688 (453)
UKP4	235	175.2	38.7	9226	22	3770 (349)
UKP5	447	70.7	20.3	9271	51	3800 (430)
UKP6	689	332.6	14.1	9826	13	3731 (96)
UKP7	521	62.6	17.3	9257	19	3797 (475)
UKP8	369	93.0	24.7	9473	26	3803 (518)

Table 3 shows the results of assembly using IDBA-UD, and subsequent improvement and annotation using PAGIT. These genomes benefited greatly from gap closure by IMAGE over those produced by SPAdes (see Tables 2 and 3), since gaps in intragenic repetitive regions were much more common, potentially confounding VNTR analysis. The mean percentage of genes transferred by RATT to the improved IDBA-UD assemblies is 98%. The mean percentage of genes transferred to chimeric regions is

Table 3: Statistics for draft genomes assembled using IDBA-UD as per Table 2.

Isolate	IDBA-UD assembly statistics: No.	N50	Av. (kb)	Assembly size post-PAGIT (kb)	Gaps closed by IMAGE	Genes transferred: all (erroneously)
UKH3	419	52.9	21.5	9102	104	3757 (0)
UKH4	627	39.7	14.3	9212	229	3688 (44)
UKH5	619	38.7	14.5	9197	247	3699 (32)
UKP2	360	63.9	25.2	9143	241	3776 (0)
UKP3	563	47.8	16.0	9168	312	3767 (1)
UKP4	509	53.7	17.7	9154	292	3772 (0)
UKP5	1830	11.2	4.8	9273	1791	3552 (1)
UKP6	768	51.4	12.1	9135	105	3702 (2)
UKP7	829	32.0	10.7	9184	288	3775 (6)
UKP8	614	40.7	14.7	9177	293	3756 (0)

0.2%. In the IDBA-UD assemblies, the *C. hominis* genomes performed slightly worse, with 0, 44, and 32 genes transferred to chimeric regions respectively across UKH3, UKH4, and UKH5.

The dramatic decrease in the number of genes transferred to chimeric regions indicates significantly fewer misassemblies in improved genomes generated by IDBA-UD than in those of SPAdes, marking a significant improvement. This indicates the effectiveness of using ABACAS to identify gaps within the IDBA-UD assemblies, and IMAGE to close them, which SPAdes would resolve during assembly.

NUCmer, from the MUMMER package was used to identify misassembly, as detailed in section 3.5. Figure 4 shows the extent of misassembly in the isolate genomes, denoted by coloured bars corresponding to which chromosomes regions belong to according to NUCmer. Extensive misassembly was identified in all of the genomes, to varying degrees. The most consistently misassembled chromosome is chromosome 7, with a consistent chromosome 8 misassembly. The most misassembled isolates where UKP3 and UKP8, with 8 misassemblies of larger than 10kb. These two isolates have very high Gini scores (see Table 1), of 0.5489 and 0.5570 respectively.

Figure 5 illustrates a moderate correlation ($R^2 = 0.41$) between the Gini coefficient and number of misplaced genes within misassembled chromosomal regions across 45 isolates of *C. parvum* and *C. hominis*.

Table 4 shows the number of VNTR regions that were missing from the IBDA-UD assemblies before and after gap closure with IMAGE. These results show that a large amount of VNTR regions were resolved using IMAGE, indicating the importance of post-assembly genome improvement in the generation of accurate and reliable genome assemblies.

Table 4: The number of VNTR regions missing within the IDBA-UD assemblies pre and post gap closing with IMAGE.

Isolate	VNTR regions missing before IMAGE	VNTR regions missing post-IMAGE
UKP2	48	7
UKP3	56	12
UKP4	63	10
UKP5	209	33
UKP6	62	13
UKP7	62	8
UKP8	67	13

Figure 4 shows putatively misassembled regions (translocations) within the *C.parvum* UKP2-8 (Hadfield et al., 2015) PAGIT-improved SPAdes assemblies. A heatmap showing the number of VNTR's per coding sequence (CDS) is included. Every genome assembly within the dataset exhibits significant misassembly across all chromosomes, particularly at the terminal end.

5 DISCUSSION

Table 1 indicates high depth of coverage inequality throughout the genomes, represented by relatively high Gini coefficient values in comparison to that exhibited by Iowa II (0.112), which the mean depth and breadth of coverage (fraction of the reference covered) will not indicate. This appears to be a common issue when sequencing Cryptosporidium from human clinical samples. Paired end read libraries accessed from GenBank, sequenced by the Welcome Trust Sanger Institute (Bioproject PRJEB3213), and those published by Troell *et al.* (2016) (Bioproject PRJNA308172), who was attempting to generate whole genome sequences from single cells using whole genome amplification (Troell et al., 2016), also suffered from very high Gini coefficients, indicating that this problem is not restricted to a single research team. Figure 5 indicates that there is some correlation between the Gini coefficient and the amount of misassembly within genomes assembled by SPAdes. Although this correlation is weak ($R^2 = 0.41$).

Whole genome alignments were used to identify *in silico* translocation events (considered putative misassemblies), as detailed in section 3.5. Figure 4 illustrates that translocation occurred in a similar fashion throughout each of the assemblies, with the same areas being merged into similar chimeric genomes, as can be seen in chromosome 7, where the initial 120kb region has merged into the end of chromosome 8 throughout all of the genomes. It is interesting to note that only on UKP3 was a 70kb area from chromosome 5 seen starting at 500kb on chromosome 7. Similarly only in UKP8 was a unique 70kb translocated region seen in chromosome 7 from chromosome 3. These two genomes bear high Gini coefficients, as detailed in Table 1, which may contribute to this. A peculiarity of these misassemblies is the observed trend of chimeric chromosomes being a result of the native chromosome being flanked upstream by 80kb of the downstream extreme portion of the subsequent chromosome. This is illustrated very clearly in Figure 4.

Taxonomic evaluation carried out by Hadfield *et al.* utilising the gp60 marker show that there are five gp60 subtypes within the *C. parvum* dataset. This variation within the Hadfield *C. parvum* isolates is supported by Perez-Cordon *et al.* (Perez-Cordon et al., 2016) which shows clear variation across 28 VNTR loci, suggesting a number of genetic lineages. The very low likelihood of similar translocation occurring across different populations of *C. parvum* indicates that these events are as a result of misassembly by SPAdes, rather than a biological observations.

Examination of one such chimeric contig (the chr8-chr7 chimeric region at 0-0.14Mb of UKP3 on Figure 4) revealed that the region has very low depth of coverage, with no single read spanning the chromosomal fragments. Moreover, the sequences from different chromosomes are joined using a simple "AT" repetitive region with only three reads spanning the repeat region and no reads pairing across it (see Figure 6). This was observed in a number of other chimeric interface regions. Due to the low complexity, high repeat rich nature of the Cryptosporidium genome, coupled with the difficulties associated with DNA extraction and sequencing of this parasite, there is insufficient evidence to suggest that this represents true biological variation. Instead, it may be attributed to a misassembly by the Spades software. This kind of assembly error was also typical of the assemblies produced by using Velvet *de novo* assembler.

Unlike SPAdes, the IDBA assembler leaves these sequence fragments unjoined, with the result that significantly less chimeric regions are seen in the IDBA assemblies. This is because IDBA is designed for the task of assembling genomes of highly uneven depth of coverage. Although IDBA-UD did not create so many chimeric contigs, the low complexity regions were often left unassembled, with the result that CDS regions contained gaps. Unfortunately, these gaps often included the VNTRs that we require for our multi-locus subtyping scheme.

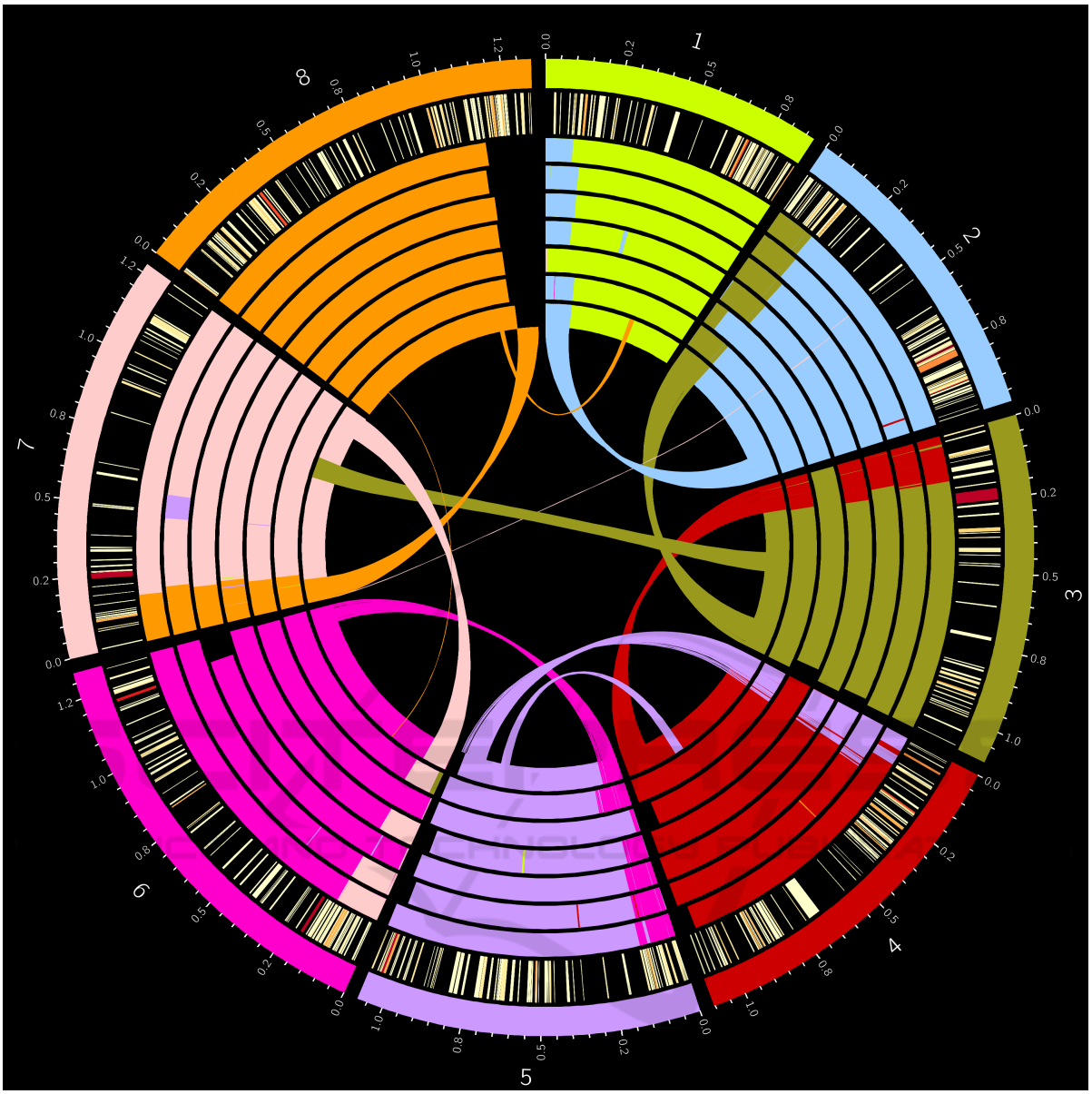


Figure 4: Misassembled regions on each SPAdes assembled Hadfield *et al. C.parvum* genome. Regions are colour coordinated by which chromosome of the *C.parvum* IowaII reference genome (represented by the outer track) they map to. From outermost to innermost, the inner tracks represent the genomes of each isolate from UKP2-8. The innermost track (UKP8) also includes a linkage map showing precisely where the regions map to in the IowaII reference genome. The second from outer track shows a heatmap of genes bearing Tandem Repeats (TRs), from light yellow denoting a single VNTR within the gene to dark red indicating many TRs within the gene. TRs were identified using Tandem Repeats Finder (see section 3.5).

Both SPAdes and Velvet (data from Velvet not shown) produced full, ungapped CDS regions (see Table 2). Thus the IDBA assemblies were not suitable for VNTR analysis and further biomarker identification without significant improvement. PAGIT was used to improve the genomes from all assemblers (see section 3.4), and this improved the resolution of low complexity regions within the IDBA-UD assemblies. Within PAGIT, ABACAS performs scaffolding on the genome assemblies and introduces gaps across the unassembled regions, the IMAGE tool then performs gap closure on these regions, resulting in high quality intragenic VNTR's for biomarker analysis.

An example of a region resolved by IMAGE can be seen in Figure 7, which shows a multiple alignment of the cgd5_350 gene from each of the Had-

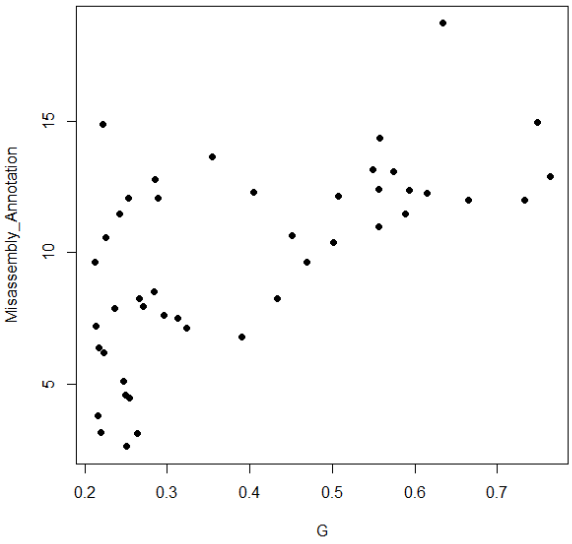


Figure 5: The percentage of genes transferred to chimeric (misassembled) regions against Gini coefficient of coverage for 45 isolates of *C.parvum* and *C.hominis*. $R^2 = 0.41$

field *C. parvum* assemblies. This region exhibits 4 distinct alleles, and can therefore be used to define specific genotypes: an essential tool of clinical diagnostics. The number of gaps closed within the IDBA-UD assemblies was significantly higher than within the SPAdes assemblies. This difference in gaps closed was expected, as IDBA-UD was designed for the purpose of assembling genomes which suffer from poor depth of coverage equality, and is therefore more conservative in extending reads across regions with shallow coverage.

The *C.parvum* assemblies produced by IDBA-UD and PAGIT exhibited very few misassemblies compared to the SPAdes assemblies. However, the *C.hominis* genomes suffered from a greater amount of putative misassemblies within the IDBA-UD genomes, as measured by the number of genes being transferred between chromosomes. Note that, genes are transferred from the *C.parvum* IowaII reference genome, which is as different, albeit similar species, and so some biological changes may be expected. Further analysis is required to fully eliminate assembly error as a cause of these chromosomal translocations. Table 4 shows that IMAGE is essential within this workflow for the resolution of repetitive regions which are not resolved during assembly with IDBA-UD. The results show a five to six-fold decrease in the number of VNTR regions missing within the assemblies.

6 CONCLUSION

In this paper we have performed a detailed analysis of 10 Cryptosporidium genomes assembled with 3 popular assemblers. In summary, the results indicate that assembly with IDBA-UD followed by improvement with PAGIT (with particular emphasis on IMAGE) is an effective and reliable way of assembling high quality draft genomes generated using the protocol detailed by Hadfield *et al.* (2015). Due to the protocol required to extract DNA from clinical samples, these genome sequences often have highly uneven sequencing depth even if the coverage across the genome sequence is relatively high. To investigate sequencing depth, we have developed a novel approach that uses the Gini coefficient to determine coverage inequality. We found the SPAdes and Velvet assemblies to be problematic, leading to misassemblies across low coverage, low complexity regions leading to the creation of chimeric chromosomes: up to 15% of all genes were being placed within these chimeric chromosomes. Although the assemblies generated by IDBA-UD did not suffer from the problem of chimeric sequences, they were problematic due to a different assembly approach, leading to a large number of gaps, particularly in repetitive regions. This is a significant issue because these gaps often contained the VNTR sequences that are important to us for developing new clinical genotyping strategies. However, the IMAGE gap closing tool from the genome improvement pipeline, PAGIT, was able to resolve these missing low complexity regions. Using this strategy, of assembly with IDBA followed by gap closing with IMAGE, we will be able to perform more in depth VNTR analysis with the intention of identifying biomarkers that will facilitate the development of novel prevention strategies in the fight against this important disease.

ACKNOWLEDGMENTS

We would like to thank Grigorio Perez-Cordon for his helpful discussion and support in the early stages of this work. This work was funded by the Knowledge Economy Skills Scholarships (KESS 2), a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.

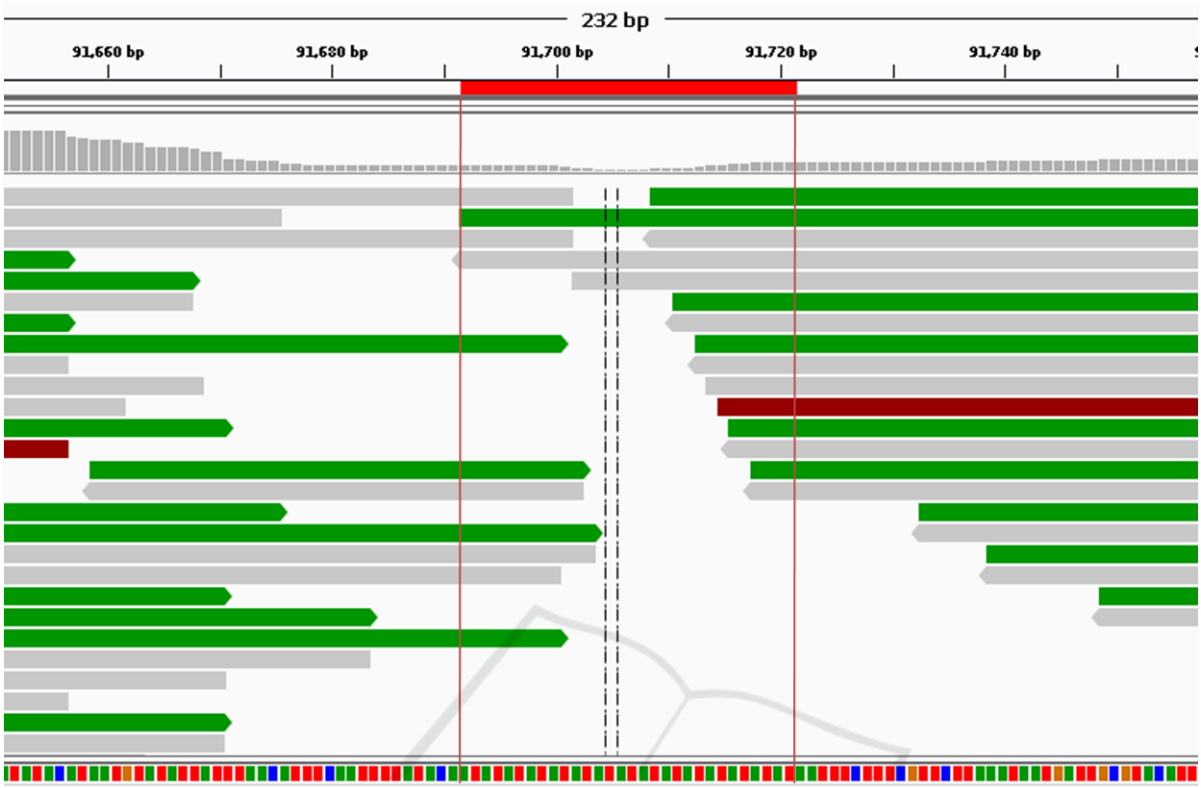


Figure 6: The misassembly interface between fragments from chromosomes 8 and 7 on the chimeric chromosome 7 of UKP3. Single reads are shown, as is a colourised sequence track (A = Green, T = Red, C = Blue, G = Orange) at the bottom where the repeat region implicated in the formation of this chimeric contig can be seen. Image produced using IGV.

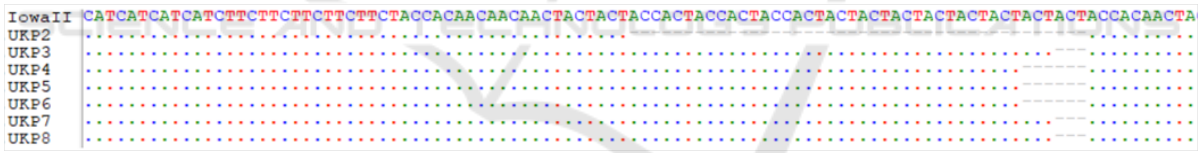


Figure 7: A multiple alignment of a VNTR region within cgd5_350 in the *C. parvum* dataset utilised in this paper that was resolved within IDBA-UD assemblies using IMAGE. Four alleles are seen within this alignment, defined by variation in the number of 'ACC' and 'ACT' codons present within the region. Differences such as this within a VNTR region can be used to define distinct genotypes, used for diagnostic evaluation.

REFERENCES

Abrahamsen, M. S., Lancto, C. A., Deng, M., Liu, C., Bankier, A. T., Dear, P. H., Konfortov, B. A., Spriggs, H. F., Iyer, L., Anantharaman, V., Aravind, L., Kapur, V., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Widmer, G., Tzipori, S., Buck, G. a., and Xu, P. (2004). Complete Genome Sequence of the Apicomplexan, Cryptosporidium parvum. *Science*, 304(5669):441–445.

Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., and Berriman, M. (2009). ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15):1968–1969.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a., and Pevzner, P. a. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477.

Benson, G. (1999). Tandem Repeats Finder: a program to analyse DNA sequences. *Nucleic Acids Res.*, 27(2):573–578.

Chalmers, R. M., Robinson, G., Hotchkiss, E., Alexander, C., May, S., Gilray, J., Connelly, L., and Hadfield, S. J. (2017). Suitability of loci for multiple-locus variable-number of tandem-repeats analysis of Cryptosporidium parvum for inter-laboratory surveillance and outbreak investigations. *Parasitology*, 144(1):37–47.

Hadfield, S. J., Pachebat, J. A., Swain, M. T., Robinson, G., Cameron, S. J., Alexander, J., Hegarty, M. J., Elwin, K., and Chalmers, R. M. (2015). Generation of whole genome sequences of new Cryptosporidium hominis and Cryptosporidium parvum isolates directly from stool samples. *BMC genomics*, 16:650.

Krzywinski, M., Schein, J., Birol, n., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos. *Genome Research*, 19(9):1639–1645.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Otto, T. D., Dillon, G. P., Degrave, W. S., and Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research*, 39(9):1–7.

Otto, T. D., Sanders, M., Berriman, M., and Newbold, C. (2010). Iterative correction of reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, 26(14):1704–1707.

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

Perez-Cordon, G., Robinson, G., Nader, J., and Chalmers, R. M. (2016). Discovery of new variable number tandem repeat loci in multiple Cryptosporidium parvum genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. *Experimental Parasitology*, 169(August):119–128.

Puiu, D., Enomoto, S., Buck, G. A., Abrahamsen, M. S., and Kissinger, J. C. (2004). CryptoDB: the Cryptosporidium genome resource. *Nucleic Acids Research*, 32(90001):329D–331.

Sow, S. O., Muhsen, K., Nasrin, D., Blackwelder, W. C., Wu, Y., Farag, T. H., Panchalingam, S., Sur, D., Zaidi, A. K., Faruque, A. S., Saha, D., Adegbola, R., Alonso, P. L., Breiman, R. F., Bassat, Q., Tamboura, B., Sanogo, D., Onwuchekwa, U., Manna, B., Ramamurthy, T., Kanungo, S., Ahmed, S., Qureshi, S., Quadri, F., Hossain, A., Das, S. K., Antonio, M., Hossain, M. J., Mandomando, I., Nhampossa, T., Acácio, S., Omore, R., Oundo, J. O., Ochieng, J. B., Mintz, E. D., O'Reilly, C. E., Berkeley, L. Y., Livio, S., Tennant, S. M., Sommerfelt, H., Nataro, J. P., Ziv-Baran, T., Robins-Browne, R. M., Mishcherkin, V., Zhang, J., Liu, J., Houpt, E. R., Kotloff, K. L., and Levine, M. M. (2016). The Burden of Cryptosporidium Diarrheal Disease among Children < 24 Months of Age in Moderate/High Mortality Regions of Sub-Saharan Africa and South Asia, Utilizing Data from the Global Enteric Multicenter Study (GEMS). *PLoS Neglected Tropical Diseases*, 10(5):1–20.

Swain, M. T., Tsai, I. J., Assefa, S. a., Newbold, C., Berriman, M., and Otto, T. D. (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nature protocols*, 7(7):1260–84.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192.

Troell, K., Hallström, B., Divne, A. M., Alsmark, C., Arrighi, R., Huss, M., Beser, J., and Bertilsson, S. (2016). Cryptosporidium as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics*, 17(1):1–12.

Tsai, I. J., Otto, T. D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, 11(4).

Xu, P., Widmer, G., Wang, Y., Ozald, L., Alves, J., Serrano, M. G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A., Pearson, W., Dear, P. H., Bankier, A. T., Peterson, D., Abrahamsen, M. S., Kapur, V., Tzipori, S., and Buck, G. A. (2004). The Genome of Cryptosporidium hominis. *Letters to Nature*, 431(October).

Zerbino, D. R. and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829.