RESEARCH Open Access

Investigating within-host population diversity of *Cryptosporidium parvum* using BlooMine

Arthur V. Morris[1*], Thomas Connor[1], Justin Pachebat[3] and Martin Swain[2,3]

Abstract
Investigating multiplicity-of-infection (MOI) in pathogen populations is central to understanding within-host evolutionary dynamics. In *Cryptosporidium parvum*, MOI may play a pivotal role in diversification due to the parasite's sexually recombinogenic life cycle, which occurs entirely within a single host. Subtyping of *C. parvum* typically relies on variation at short tandem repeat (STR) loci - such as that found in the 60kDa surface glycoprotein gene, Gp60 - but accurate profiling of these regions from next-generation sequencing (NGS) data remains technically challenging. Here, we apply BlooMine, an alignment-free, Bloom filter based tool that isolates STR containing reads via a novel pseudo-alignment strategy robust to structural and sequence variation. Using this approach, we analyse the full publicly available *C. parvum* Illumina dataset to quantify polyclonality at the Gp60 locus. After stringent artefact suppression, we detect strong evidence of in-host diversity across multiple continents and hosts. Cattle samples exhibited a 2.3-fold higher odds of harbouring multiple Gp60 subtypes compared to human samples, a difference that remained significant after adjustment for geography, sequencing depth, and subtype family. Allele co-occurrence analysis revealed closely related pairs likely arising from replication slippage, as well as mutually exclusive combinations suggestive of within-host competition or transmission structuring. Several subtypes traditionally considered host-specific were detected in unexpected host contexts, indicating greater host plasticity than previously assumed. Our results support a model in which STR-driven microevolution, recombination, and polyclonal infection jointly shape *C. parvum* population structure. We propose that MOI may act as a genetic crucible, facilitating subtype diversification within individual hosts. This study represents the largest genomic survey of Gp60 polyclonality to date and provides key insights into the evolutionary and epidemiological dynamics of *C. parvum*.

Keywords Cryptosporidium, Genomic, Zoonosis, Disease, Bioinformatics

Introduction

Cryptosporidium is a single celled eukaryotic parasite which causes substantial mortality and morbidity worldwide, with it being implicated in over 200,000 deaths annually [1]. Extracting sufficient DNA from clinical samples of *Cryptosporidium* for the purpose of whole genome analysis is a challenge due to the often low faecal mass provided in clinical samples, and low oocyst count of these samples [2, 3]. Despite recent advances being made both experimentally and computationally, there is currently no workflow which allows for routine generation and analysis of whole genome data from clinical *Cryptosporidium* samples [4–6].

Multiplicity of infection (MOI) is a well documented observation within the natural world, and has been reported widely in parasitic protozoan infection. In 1999, Lord et al. reported that the majority of adults infected with

*Correspondence:
Arthur V. Morris
morrisa28@cardiff.ac.uk
[1] School of Biological Sciences, Cardiff University, Cardiff, United Kingdom
[2] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, United Kingdom
[3] Department of Life Sciences, Aberystwyth University, Aberystwyth, United Kingdom

Plasmodium falciparum were host to more than five distinct strains [7]. MOI may arise by two principal processes: infection by multiple strains; and in-host divergence into genetically distinct sub-populations. The presence of multiple clonal populations of a parasite in a single host (referred to as 'polyclonality') has a significant impact on their incidence, transmission, and virulence, factors of enormous importance to public health. An incomplete understanding of parasite polyclonality leads to potentially inaccurate assumptions about the epidemiology, clinical presentation and outlook of the infection. The virulence experienced by a host which exhibits polyclonality is a product of the interactions between the different clonal populations, resulting in an overall virulence of anywhere from greater than the most virulent to less than the least virulent, depending on various biological factors [8, 9]. Efforts have been made to identify general patterns when ecological and evolutionary theory are applied to within-host dynamics, leading to the generalisation that 'basic ecological rules govern the outcome of co-infection across a broad spectrum of parasite taxa' [10]. The investigation into the impact of MOI on a host relies on the accurate and reliable detection and discrimination of the clonal component populations of the mixture. With modern methods of DNA isolation and purification, and the fall in price of using high-throughput Next-Generation sequencing (NGS) technologies to perform whole genome sequencing (WGS), investigating MOI by mining WGS reads has become a more feasible approach [2, 3]. 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 conventional molecular wet-lab techniques. It is both biologically plausible (due to unrestricted sexual recombination between sub-populations), and there is strong evidence, that mixed-infections can give rise to novel strains of *Cryptosporidium* spp., which may be genotypically and phenotypically distinct from their parents, representing a serious challenge to public health control services which rely on epidemiological surveillance [11–14]. It is therefore essential that polyclonality can be characterised in order to improve epidemiological surveillance, which is essential in outbreak investigation and subsequent interruption of transmission - a critical component in the fight against cryptosporidiosis and other parasitic diseases.

Alignment-free sequence analysis has the capacity to resolve the significant bottleneck 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 [15]. 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. There are a great many tools which employ alignment-free methods to identify nucleotide and protein sequence similarity, developed to resolve problems in many disciplines within Bioinformatics. However, the majority of the tools developed to resolve these from pathogen clinical data utilise single nucleotide polymorphism (SNP) data, for which variant callers are highly optimised. Variable Number Tandem Repeat (VNTR) loci have been regions subject to heavy interrogation, and have been used to define subtypes and variants in many pathogen species [16]. Despite great advances in the use of NGS data for assembly-free genomic analysis, low concordance of variant calling pipelines has been reported, and such pipelines are often incapable of resolving VNTR repeat subunit copy number even across small VNTRs which can be captured in a single read, which remains a troubling issue [17].

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, and a novel pseudo-alignment algorithm to perform set membership queries and infer sequence homology [18]. We use BlooMine to investigate MOI in clinical samples of *Cryptosporidium parvum*, demonstrating that it can be used to perform both *in silico* subtyping from NGS reads, and local genomic heterogeneity analysis, facilitating both informed epidemiological investigations in public health, and robust hypothesis testing critical to the development of novel approaches of transmission interruption.

Materials and methods

Implementation

BlooMine is written using C++ and Python 3. It consists of two modules:

1. BlooMine_screen: Screen a FASTQ dataset for reads which contain a target sequence.
2. BlooMine_demix: Analyse reads for target sequence heterogeneity.

The BlooMine_screen module consists of 3 steps:

- BlooMine_gen: Generate a Bloom filter from the target sequence.
- BlooMine_FPscreen: Carry out a first pass screen of reads using the target Bloom filter.
- BlooMine_SPaln: Carry out a second pass target sequence pseudo-alignment step.

BlooMine_gen

A target sequence, T is used to generate a *k*mer array, K_T^k. This is then passed through hash functions to generate a bit signature for T, B_t. The number of hash functions can be adjusted to meet user type-I and type-II error rate requirements.

BlooMine_FPscreen

Reads within the FASTQ file are screened against the Bloom filter, B_t, generated during the BlooMine_gen process to detect the presence of the target sequence. This step constitutes the first pass screening. A *k*mer array is generated from each read within the FASTQ, which is then hashed using the Bloom filter hash functions. The subsequent bit signature generated by this hashing process is screened against B_t to infer set membership. The detection of a hit increments the score by 1. If the overall score for the read exceeds the *k*mer intersection threshold, sequence similarity is inferred and the read returned for further processing. Trivially (without a description of Bloom filter functionality) the set of all reads which exceed the *k*mer intersection threshold is therefore

$$R' = \{r \in R \mid |K_T^k \cap K_r^k| \geq \theta\}$$

where K_r^k is a *k*mer array of the read sequence, T is the target sequence, and θ is *k*mer intersection threshold.

BlooMine_SPaln

BlooMine_SPaln filters reads using a novel soft-alignment algorithm (Algorithm 1) based on positional *k*-mer mapping. For each read in the filtered set R', the positions of all matching *k*mers are identified and used to construct a sparse positional vector aligned to the read.

```
Input: $R'$: set of hit reads from first-pass screening;
       $T$: the target sequence;
       $\omega$: gap threshold (see Equation 1);
       $k$: k-mer size;
       $MST$: minimum score to trigger a hit.
Output: $R''$: set of hit reads.
$R'' \leftarrow \emptyset$;
$K_t \leftarrow \texttt{Kmerise}(T, k)$;
foreach $read \in R'$ do
    $K_r \leftarrow \texttt{Kmerise}(read, k)$;
    Initialise array $N$ of 0's of length $|read|$;
    foreach $kmer \in K_r$ do
        if $kmer \in K_t$ then
            $i \leftarrow$ index position $kmer$ maps to in $K_r$;
            $N[i : i + k] \leftarrow kmer$;
    $gap\_count \leftarrow 0$;
    foreach $p \in N$ do
        if $p$ is a gap then
            $gap\_count \leftarrow gap\_count + 1$;
        else
            if $gap\_count \geq \omega$ then
                split $N$ at $p - 1$;
            $gap\_count \leftarrow 0$;
    $s \leftarrow \texttt{score\_alignments}(N)$;
    if $s \geq MST$ then
        add $read$ to $R''$;
return $R''$;
```

Algorithm 1 BlooMine SPaln Second-pass screen: Alignment-free alignment inference

From this vector, gapped alignment regions are inferred by grouping together adjacent *k*mer hits that are separated by no more than a threshold number of positions. These are defined as subalignment chunks. Each chunk is scored using a gap-sensitive function to determine the optimal alignment region.

The gap threshold ω is defined as:

$$\omega = \left\lceil \frac{hk - g}{n} \right\rceil \tag{1}$$

This value is chosen such that a single matching *k*mer followed by a gap of size ω results in a total alignment score near zero, thereby discouraging long, sparse alignments.

Each chunk c is then scored using the following scoring algorithm. Let chunk $c = \{c_0, c_1, ..., c_{|c|-1}\}$ be a sequence of aligned positions (indexed by position in the read), we define the score of the chunk s_c as:

$$s_c = \sum_{i=0}^{|c|-1} \delta(c_i, c_{i-1}) \tag{2}$$

Where $\delta(c_i, c_{i-1})$ is defined as:

$$\delta(c_i, c_{i-1}) = \begin{cases} +h & \text{if } c_i \in K_T \\ -g & \text{if } c_i \notin K_T \wedge c_{i-1} \in K_T \\ -n & \text{if } c_i \notin K_T \wedge c_{i-1} \notin K_T \end{cases} \tag{3}$$

Where:

k = the size of each *k*mer (default=7).

h = the value the score is incremented by when a hit is triggered (10.0).

g = the value the score is penalised by when a gap is opened (15).

n = the value the score is penalised by when a gap is extended (7.0).

c_i = position *i* within the aligned chunk *c*.

K_T = the *k*mer array of target sequence *T*.

Alignment regions are split at any gap of size $\geq \omega$. All contiguous sub-sequences of alignment chunks (i.e. $\mathcal{O}(n(n+1)/2)$ combinations for n chunks) are evaluated using the scoring function. The highest-scoring alignment is selected. If this score exceeds the minimum score threshold *MST* (see Minimum score threshold section), the read is deemed to contain the target sequence. Figure 1 presents a graphical representation of this process using a full target region consisting of two flanking regions and a polymorphic STR locus (see Fig. 2 for an example model of an STR target region).

Minimum score threshold

Due to the polymorphic nature of certain genomic regions (e.g. VNTRs), traditional alignment algorithms such as Smith-Waterman and Needleman–Wunsch often fail to produce reliable results due to their assumption of globally consistent match/mismatch patterns [19, 20]. These methods are not robust to high densities of SNPs, insertion/deletion events (INDELs), or rearrangements.

To address this, we propose an approximate alignment scoring scheme based on *k*-mer matches across sliding windows. The key insight is that mismatches often cluster, particularly in low-complexity and repetitive regions, rather than present as uniformly distributed across a sequence.

Let s be a quest sequence of length $|s|$, and let κ be the expected mismatch spacing (i.e. the average number of bases between mismatches). The mismatch rate is then defined as $\epsilon = \frac{1}{\kappa}$, and the expected number of mismatches in *s* is $|s| \cdot \epsilon$.

We define a sliding window of size $k + \kappa - 1$. The expected number of mismatches per window, *e* is therefore

$$e = \frac{k + \kappa - 1}{\kappa} \tag{4}$$

For example, with $k = 5$ and $\kappa = 10$, a window spans 14 bases and is expected to contain 1.4 mismatches.

Let $w = \frac{|K_t|}{k+\kappa-1}$ be the number of such windows over the target sequence. The total expected mismatch count, *M*, is

$$M = e(w - 1) \tag{5}$$

The minimum alignment score is then calculated as

$$MST = h \cdot |K_t| - wg + nM \tag{6}$$

This novel alignment algorithm uses concepts from both alignment and alignment-free sequence analysis approaches to resolve alignments from sequences which have undergone rearrangement or INDEL events, since regions flanking highly polymorphic VNTR regions may be poorly conserved. Furthermore, alignments are carried out on small sequences (single reads) potentially many thousands of times, which may render conventional hard-alignment approaches less efficient.

BlooMine_MOI

BlooMine functionality can be leveraged to investigate the presence of multiple alleles within a single FASTQ file, which may indicate MOI in clinical pathogen samples. Reads are filtered in two steps using probe sequences flanking the target region as proxy-targets. First, reads containing flank 1 are identified within the full dataset. Second, the resulting subset is further

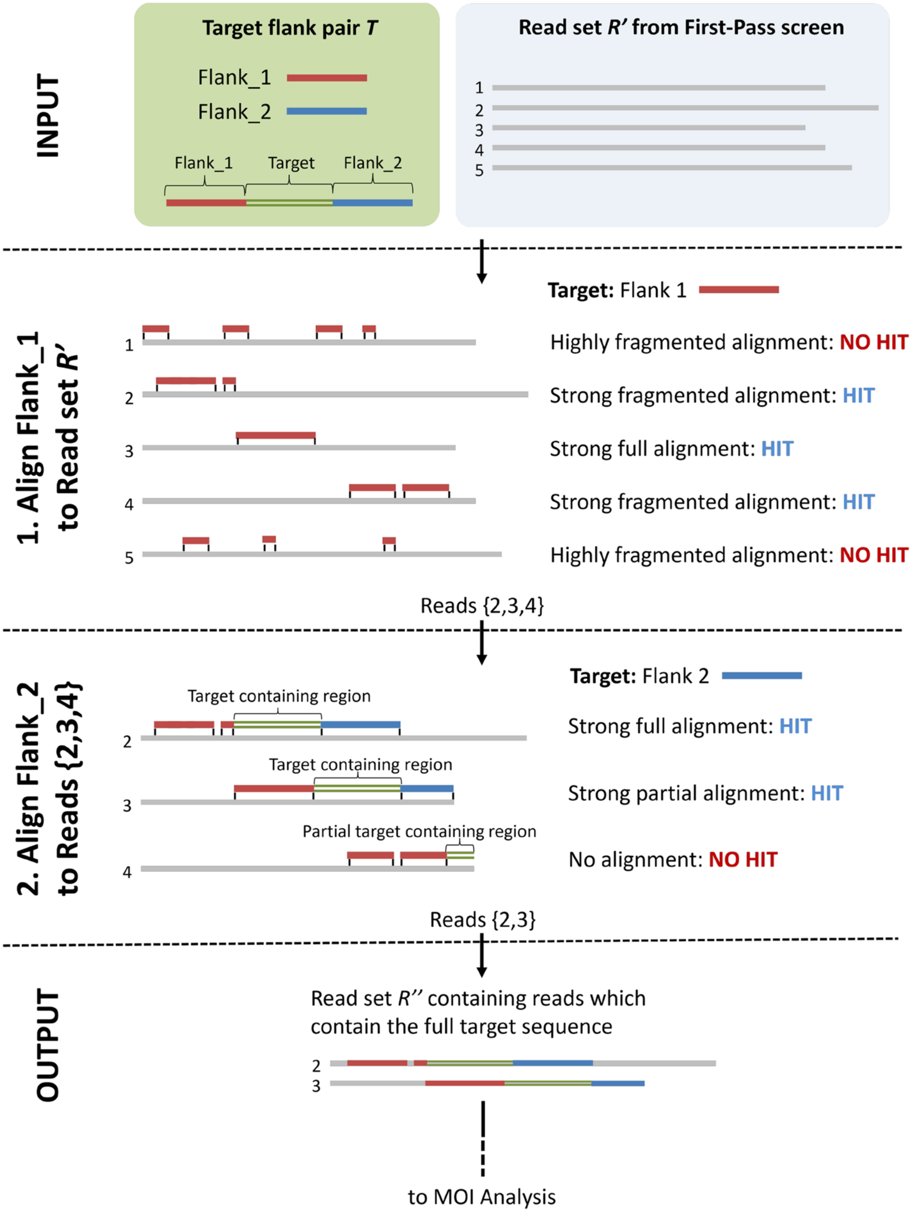
INPUT
Target flank pair *T*
Flank_1
Flank_2
Flank_1
Target
Flank_2
Read set *R'* from First-Pass screen
1
2
3
4
5
1. Align Flank_1 to Read set *R'*
Target: Flank 1
Highly fragmented alignment: NO HIT
Strong fragmented alignment: HIT
Strong full alignment: HIT
Strong fragmented alignment: HIT
Highly fragmented alignment: NO HIT
Reads {2,3,4}
2. Align Flank_2 to Reads {2,3,4}
Target: Flank 2
Target containing region
Strong full alignment: HIT
Target containing region
Strong partial alignment: HIT
Partial target containing region
No alignment: NO HIT
Reads {2,3}
OUTPUT
Read set *R''* containing reads which contain the full target sequence
2
3
to MOI Analysis

Fig. 1 Workflow for identifying target regions in short reads

filtered for reads containing flank 2. The output is the set of reads which contain both flanks, and therefore the target region. These reads are then used to calculate the level of heterogeneity in the target region.

Probe sets for three major *C. parvum* Gp60 families (IIa, IId, IIc) were developed using sequences flanking the STR region. These probes were determined to cover a large portion of the *C. parvum* subtype space. To account for probe cross-matching between Gp60 families, the major Gp60 family was heuristically determined using probe set which yielded the largest number of read hits. Instances of multiple Gp60 family hits were handled using bespoke scripts which determined the best matching probe pair for each read.

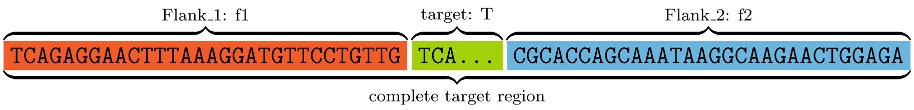
Flank_1: f1
target: T
Flank_2: f2
TCAGAGGAACTTTAAAGGATGTTCCTGTTG
TCA...
CGCACCAGCAAATAAGGCAAGAACTGGAGA
complete target region

Fig. 2 A model of a target sequence, showing the target region consisting of a TCA repeat region, and flanking regions. Where... refers to an arbitrary number of the preceding repeat motif

MOI investigation of *C. parvum*

A dataset of all publicly available *Cryptosporidium parvum* (n=667) NGS samples was utilised in this study. Samples which were a product of artificial infection, or lab strains, were filtered out (n=38). The remaining samples (n=629) were screened using the highly polymorphic Gp60 microsatellite locus as a target. These samples were taken from Europe, North America, Africa, and Asia, and from 7 hosts: Cattle (*Bos taurus*), Human (*Homo sapiens*), Sheep (*Ovis aries*), Goat (*Capra hircus*), Water Buffalo (*Bubalus bubalis*), Bactrian camel (*Camelus bactrianus*) and Mouse (*Mus musculus*). The BlooMine MOI pipeline was run on each of these samples using sequences taken from conserved regions flanking the Gp60 locus.

BlooMine was executed using a false-positive rate of 0.0001, a FP-screen threshold of 35.0 and a SP-screen threshold (κ) of 4.0. All data were analysed on a laptop running Ubuntu v22.04.5, with 64Gb memory and using 12 threads.

Gp60 subtyping of Cryptosporidium

Subtyping of *Cryptosporidium* involves interrogation of a highly polymorphic VNTR (serine repeat) region within the Gp60 surface glycoprotein gene on chromosome 6 [21]. Due to its highly polymorphic nature in *C. parvum*, it has been used to perform epidemiological investigations in this species for many decades. Subtypes are elucidated by determining the fragment length and sequence content of the Gp60 allele (a variant of a target locus).

Here, allele naming follows a convention analogous to that used when reporting the Gp60 subtype of *Cryptosporidium*, which was thoroughly described by Robinson *et al* [22, 23]. A, G, and R are followed by integers representing the numbers of TCA, TCG and terminal ACATCA counts respectively. This yields an allele code like 'A15G2R1'. Gp60 family is determined by the sequence content of the Gp60 gene, and takes the form of a latinised alphanumeric code (e.g. IIa) which indicates both the species and Gp60 family of the sample, whereby the two major species, *C. hominis* and *C. parvum* are designated I and II respectively. This yields a full Gp60 subtype code which indicates the family and the allele present in any given sample (e.g. IIaA15G2R1). However, since the full Gp60 region was not utilised in determining the Gp60 family in this study (due to limitations of read length), we present the probe which exhibited the highest fidelity with each allele after the allele name. For example A15G2R1-IIa would refer to the A15G2R1 allele extracted using probes designed against sequences from Gp60 family IIa. This is done to distinguish it from the conventional nomenclature and ensure that the Gp60 family designations are not taken as definitive.

Allele filtering and PCR stutter artefact correction

To accurately characterise the allelic composition of STR loci from sequencing data, particularly the *Gp60* locus, we employ a two-stage filtering process designed to exclude sequencing noise and PCR stutter artefacts.

Allele Filtering

If we consider the set of all reads which have been determined to contain a target region by BlooMine as R'', then we presume that all variants of the target sequence ('alleles') are embedded within this read set. An allele count vector is computed for each allele within R'' as

$$\mathcal{C}_{R''} = [c_1, c_2, \ldots, c_n] \tag{7}$$

where c_i is the number of reads in R'' assigned to allele a_i (the allele count of a_i), and *n* is the total number of distinct alleles identified. From this, we define a normalised frequency vector:

$$\mathcal{F}_{R''} = [f_1, f_2, \ldots, f_n] = \left[\frac{c_1}{\sum_{j=1}^{n} c_j}, \frac{c_2}{\sum_{j=1}^{n} c_j}, \ldots, \frac{c_n}{\sum_{j=1}^{n} c_j}\right] \tag{8}$$

where f_i is the relative frequency of allele a_i across all reads in R'' (the allele frequency of a_i). The major allele is defined as the allele with the highest frequency in $\mathcal{F}_{R''}$, and minor alleles are all other alleles.

For example, suppose we have 40 reads which fully cover the *Gp60* region, including its flanking regions. This ensures that the STR locus is not clipped and allows for confident allele calling. The total allele count at this locus is therefore $N = 40$, with observed counts:

$$\mathcal{C} = \begin{bmatrix} a_1 : 20 \\ a_2 : 10 \\ a_3 : 6 \\ a_4 : 4 \end{bmatrix}$$

The corresponding frequency vector is:

$$\mathcal{F} = \begin{bmatrix} a_1 : 0.5 \\ a_2 : 0.25 \\ a_3 : 0.15 \\ a_4 : 0.10 \end{bmatrix}$$

The major allele is the most frequent, in this case a_1, and all others are considered minor alleles. We then apply a naive hard-filter using a minimum count threshold T_c and frequency threshold T_f. Alleles that fall below either threshold are removed. If we assign the naive hard-filtering parameters in this examples as $T_c = 3$ and $T_f = 0.15, a_4$ is filtered out due to its low frequency even though its count exceeds the minimum. This dual-threshold approach is functions to remove noise in both low and high depth scenarios: in high-depth datasets (e.g., $N = 1000$), noise may exhibit high counts but low frequency; in low-depth datasets, even a single erroneous read can represent a substantial proportion.

PCR Stutter Filtering

To further reduce artefacts arising from PCR amplification, we apply an additional stutter filter based on known dynamics of replication slippage in STR regions [24]. The most common stutter artefact is a loss of a single repeat unit (back-stutter; $L1$), followed less frequently by a loss of two repeat units ($L2$). Forward stutter and larger contractions are rare [25]. These stutter products are not sequencing errors, but true sequences generated from mis-amplified templates. We account for this by comparing the frequencies of minor alleles to that of the presumed template (major allele). For alleles a_j which are one or two repeat units shorter (thus potential $L1$ or $L2$ artefacts) than the major allele, a_M, we compute a stutter ratio:

$$s_j = \frac{f_j}{f_M} \tag{9}$$

where f_j and f_M are the frequency of a_j and a_M respectively. We then assign a stutter ratio threshold for $L1$ artefacts (R_1), and an $L2$ stutter ratio threshold of $R_2 = R_1/2$. $L1$ and $L2$ candidates are considered artefacts if $s_j < R_1$ or $s_j < R_2$ respectively, and thus filtered out [26, 27]. For all other alleles (not plausible stutter products), only the naive filtering thresholds apply.

Continuing the example, suppose the filtered allele set is:

$$\begin{bmatrix} a_1 : 20 \\ a_2 : 10 \\ a_3 : 6 \end{bmatrix}$$

with frequencies:

$$\begin{bmatrix} a_1 : 0.5 \\ a_2 : 0.25 \\ a_3 : 0.15 \end{bmatrix}$$

Assume the following STR sequences:

- a_1: TCATCATCATCATCGTCG
- a_2: TCATCATCATCA (2 repeat unit loss)
- a_3: TCATCATCATCATCG (1 repeat unit loss)

Table 1 The host and continent stratified dataset used to conduct the CMH test. Values are given as: monoclonal count (polyclonal count)

	Human	Cattle	Continent total
Europe	62 (3)	77 (11)	139 (14)
North America	163 (13)	44 (8)	207 (21)
Total	225 (16)	121 (19)	346 (35)

Allele a_2 is a candidate $L2$ stutter, a_3 is a candidate $L1$ stutter. Normalising the frequencies using Eq. 9 gives us:

$$s_{a_2} = \frac{0.25}{0.5} = 0.5, \quad s_{a_3} = \frac{0.15}{0.5} = 0.3$$

If we assign aggressive stutter threshold parameters of $R_1 = 0.34$ (thus $R_2 = 0.16$), a_3 is filtered $(0.3 < 0.34)$, but a_2 is retained $(0.5 > 0.16)$, yielding the final filtered set:

$$\mathcal{C}' = \begin{bmatrix} a_1 : 20 \\ a_2 : 10 \end{bmatrix}, \quad \mathcal{F}' = \begin{bmatrix} a_1 : 0.667 \\ a_2 : 0.333 \end{bmatrix}$$

Polyclonality rate in cattle and humans

Allele calls from the Gp60 locus were filtered to suppress PCR-stutter and low-support artefacts using a pre-specified parameter set: $L1 = 0.15, L2 = 0.075$ (i.e., $L2 = L1/2$), minimum allele count $= 3$, and minimum allele frequency $= 0.05$. Samples from hosts other than cattle (*Bos taurus*) or human (*Homo sapiens*) were excluded. A sample was classified polyclonal if ≥ 2 alleles passed all filters; otherwise monoclonal.

To control for geographic confounding and mitigate Simpson's paradox, analysis was stratified by continent. Continents with fewer than two samples in either host were excluded *a priori*. The final analysis set comprised $n = 381$ cattle and human samples from Europe ($n = 153$) and North America ($n = 228$), of which 346 were monoclonal and 35 were polyclonal (see Table 1).

Within each retained continent, a 2×2 table (host × poly/monoclonal sample count) was constructed and per-continent odds ratios (ORs) for polyclonality (cattle vs. human) computed. This stratification controls for uneven geographic distribution, and mitigates Simpson's paradox. Mantel-Haenszel pooled OR (OR_{MH}) with 95% confidence intervals were obtained across continents, and the null hypothesis $H_0 : OR_{MH} = 1$ tested using the Cochran-Mantel-Haenszel (CMH) test (two-sided, $\alpha = 0.05$).

Consistency of stratum specific *OR*s across continents was evaluated with the Breslow-Day (BD) test for homogeneity of stratum-specific ORs, and results interpreted under the criterion that BD $p < 0.10$ indicates heterogeneity. Pooled results were interpreted with caution when heterogeneity was detected.

To assess the robustness of hypothesis testing under different filtering parameters, an exploratory threshold sweep over minimum allele count $= 1 - 30$ and $L1 = 0.01 - 0.30$ (with $L2$ fixed at $\frac{L1}{2}$, and minimum allele frequency fixed at 0.05) was carried out. For each grid point, the continent-stratified CMH analysis was repeated and summarised by constructing a heatmap from OR_{MH} with CMH p-values overlaid. This sweep is descriptive, with all inferences relying on pre-specified thresholds of $L1 = 0.15$ and minimum allele count = 3, and as such no multiplicity adjustments were carried out.

To probe additional predictors, a multivariable logistic regression was fit with polyclonality as the outcome and covariates: host, continent, sequencing depth (summed Gp60 allele counts across each sample), major-allele length, and Gp60 family (IIa vs IId, IIc excluded due to zero polyclonal samples).

Allele co-occurrence

To investigate allele co-occurrence and mutual exclusivity within polyclonal datasets, samples were stratified by continent of origin, and allele pairs with geographical contiguity—defined as co-presence within at least one continent—were tested using the Cochran–Mantel–Haenszel (CMH) procedure. This approach controls for confounding due to uneven geographic distribution, yielding pooled odds ratios and p-values across strata, with inclusion restricted to allele pairs exhibiting geographic contiguity (i.e. allele pairs occurring within at least one shared continent) and meeting a minimum per-continent occurrence threshold of three samples per allele. The Breslow-Day test was used to detect heterogeneity in odds-ratio across different continents. Fisher's exact test was not applied due to its sensitivity to highly abundant alleles, which can lead to inflated estimates of both exclusivity and co-occurrence signal strength in globally unstratified datasets.

Results

Dataset summary

Of the 629 *C. parvum* genomic NGS datasets used, 592 fully covered the Gp60 locus with at least 3 reads and a minimum allele frequency of 0.05. Under the first pass filter only (without stutter filtering), 398 were monoclonal and 194 were polyclonal, exhibiting at least 2 alleles at the Gp60 locus.

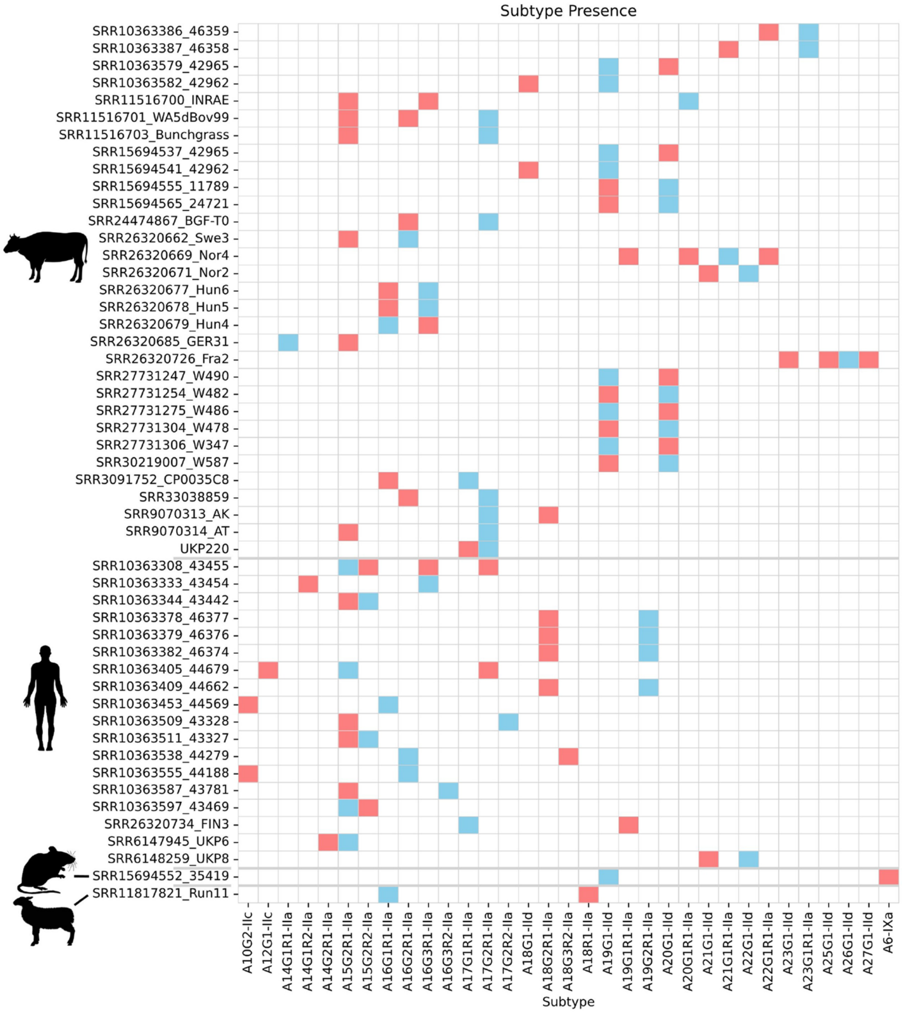

Fig. 3 Major/minor allele presence across all 51 samples which were resolved as polyclonal after applying a 0.15 stutter filter. Blue = major allele; Red = minor allele

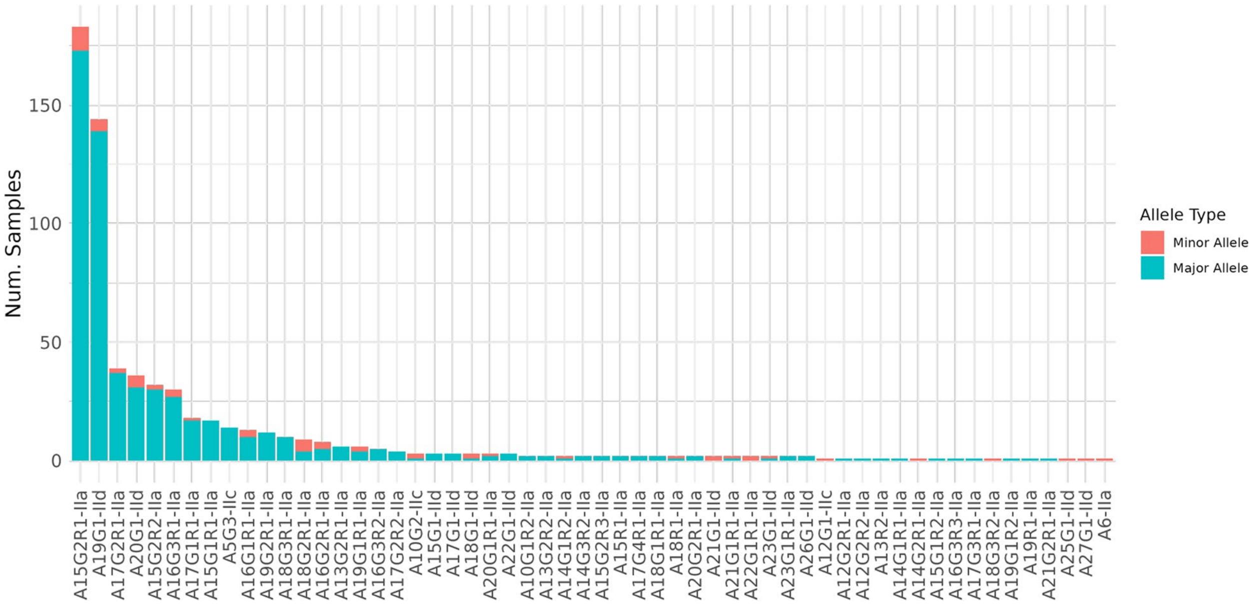

Fig. 4 The number of samples which exhibit each allele as major or minor allele

Polyclonality prevalence

An $L1$ stutter filter of 0.15 was used to investigate allele co-occurrence and host/geographical distributions. Using this stutter filter, 51/592 (8.6%) polyclonal samples were resolved: 18 human, 31 cattle, and one each from sheep and mouse (see Fig. 3). All polyclonal samples were from North America ($n = 23$), Europe ($n = 15$) and Asia ($n = 13$).

In total, 54 unique alleles were detected across clonal and polyclonal samples (Fig. 4); eight occurred only as minor alleles (A14G1R1-IIa, A17G2R1-IIa, A20G1R1-IIa, A22G1-IId, A23G1-IId, A23G1R1-IIa, A25G1-IId, A26G1-IId).

A15G2R1-IIa was the most common allele in the entire dataset ($n = 183$) both as a major allele ($n = 158$) and minor allele ($n = 11$). It was the most frequent major allele in human ($n = 109$), sheep ($n = 5$) and goat ($n = 5$) samples, and the second most common allele in cattle ($n = 64$). A15G2R1-IIa comprised 28% of all alleles detected across the dataset and 44% of human samples. The most common major allele in cattle was A19G1-IId ($n = 141$) was found in 45% of cattle samples. This allele was also found in 80% ($n = 143$) of all Asian samples, which were dominated by Chinese cattle samples, and a single North American sample.

Multiple host-specific alleles were found in humans ($n = 18$), cattle ($n = 10$), sheep ($n = 2$), and mouse ($n = 1$). Six of these host-specific alleles were only detected in polyclonal human (A10G2-IIc, A12G1-IIc*, A14G2R1-IIa*, A16G3R2-IIa, A17G2R2-IIa, and A18G3R2-IIa*) and seven in cattle (A14G1R1-IIa, A21G1R1-IIa, A22G1R1-IIa*, A23G1R1-IIa, A23G1-IId*, A25G1-IId*, A27G1-IId*) samples, of which six presented only as minor alleles (denoted by *).

Four samples showed cross-family polyclonality. Three were North American human infections with concurrent IIa and IIc alleles (including one IIa/IIa/IIc triple mixture). The fourth case was a mouse sample in which an atypical allele was recovered using the IId probe set. Subsequent alignment of this cryptic allele against a curated database of Gp60 sequences identified its closest match as *Cryptosporidium tyzzeri*, subtype family IXa.

Host differences in polyclonality

After filtering, within-continent ORs (cattle vs. humans) were 3.23 (95% CI $0.87 - 12.02$) in Europe and 1.8 ($0.72 - 4.49$) in North America, corresponding to polyclonality rates of 0.116 vs. 0.039 and 0.140 vs. 0.083, respectively. The Mantel–Haenszel pooled estimate across continents was $OR_{MH} = 2.26$ (95% CI $1.08 - 4.71$; CMH $p = 0.048$), with no evidence of heterogeneity (BD $p = 0.467$). Human IIa/IIc co-infections were observed only in North America and therefore account for the lower continent-specific OR observed there. Cattle had ~ 2.3-fold higher odds of harbouring a polyclonal *C. parvum* infection than humans after conditioning on continent.

In the multivariable logit model, the host effect remained significant after adjustment for continent, major-allele length, depth, and Gp60 family (adjusted OR for cattle vs. human ≈ 2.69; 95% CI $1.21 - 5.95$; $p = 0.015$). Sequencing depth ($p = 0.253$), continent (North America vs Europe $p = 0.266$) were not associated with polyclonality in this model. Major-allele length showed a positive association (per-unit $OR = e^{0.1421} \approx 1.15, 95\% \;\; \text{CI} \;\; 1.08 - 1.23; \;\; p < 0.001$); interpreted over the interquartile range, this corresponds to $OR_{IQR} \approx 2.43$. We consider plausible biological and technical explanations for this signal in the Discussion. Gp60 family ($p = 0.729$) was not associated with polyclonality, however, because only IIa and IId were observed across both continents (IIc was only detected in North American samples), the family contrast is sparse and should be interpreted cautiously.

Robustness to filtering choices was evaluated with a parameter sweep over minimum allele count $(1 - 30)$ and $L1$ threshold $(0.01 - 0.30$; $L2$ fixed at $L1/2$, min $AF = 0.05)$. Across a plateau surrounding the pre-specified thresholds (min count $\geq 3, L1 \geq 0.12$), OR_{MH} consistently exceeded 1, with many settings reaching CMH $p < 0.05$ (Fig. 5). As expected, very lenient thresholds (min count < 3 and/or $L1 < 0.05$) diluted the signal, consistent with increasing stutter noise rather than a change in the underlying association. Together, these analyses support a continent-adjusted excess of polyclonality in cattle relative to humans, and show that the conclusion is stable to reasonable choices of filtering parameters.

Allele distribution and co-occurrence

Within-sample allele associations were tested across 140 allele pairs using a continent-stratified CMH model (colour = signed $-log10$ p, bubble size = $|log2(OR)|$). Three pairs showed statistically significant positive associations (co-occurrence): A19G2R1-IIa with A18G2R1-IIa ($|log2(OR)| = 6.96$), A10G2-IIc with A16G2R1-IIa ($|log2(OR)| = 4.80$), and A16G2R1-IIa with A17G2R1-IIa ($|log2(OR)| = 2.87$). The latter reflects a moderate effect size. The cross-family association between A10G2-IIc and A16G2R1-IIa is likely a result of the low abundance of A10G2-IIc in combination with its tendency to present as a minor allele within this dataset. In contrast, several pairs exhibited significant negative associations (mutual exclusivity), notably A5G3-IIc with A15G1R1-IIa ($|log2(OR)| = -5.70$), A19G1-IId with A20G1-IId ($|log2(OR)| = -8.13$), and A15G1R1-IIa with A20G1-IId ($|log2(OR)| = -4.39$). The highly prevalent A15G2R1-IIa showed many statistically significant exclusions with other alleles; however, most of these had small effect sizes (small bubbles indicating low $|log2(OR)|$).

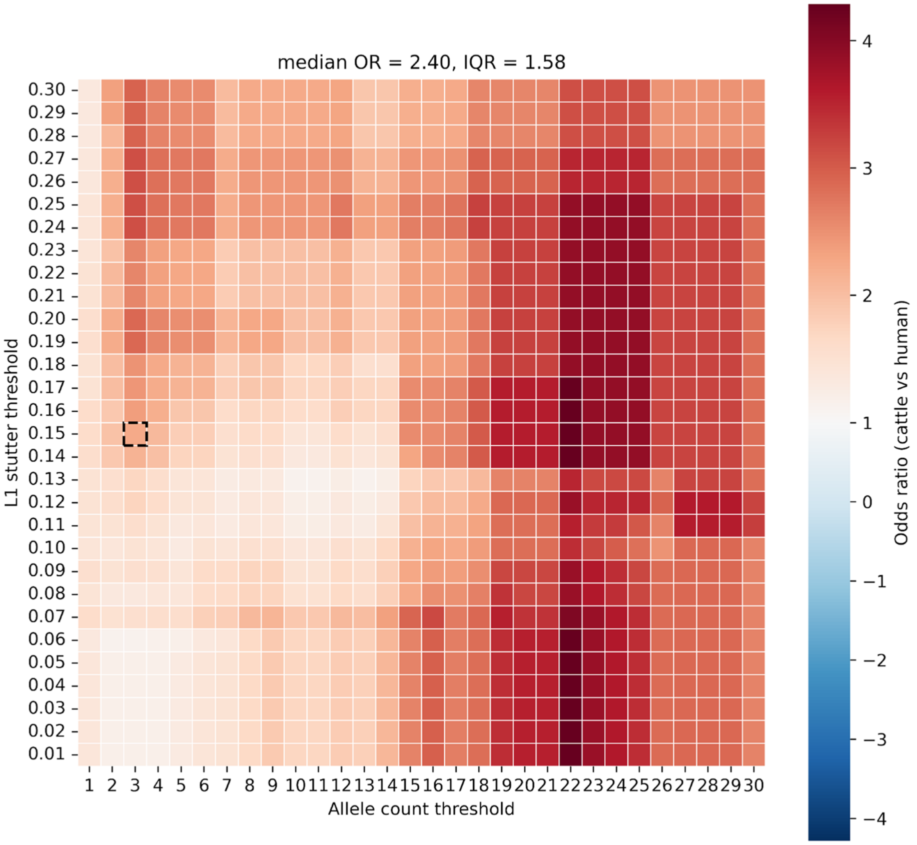

Fig. 5 Signed log odds ratios from a continent-stratified Cochran–Mantel–Haenszel test comparing the frequency of polyclonality in cattle versus human samples across a grid of allele count and stutter filter (L_1) thresholds. This parameter sweep was conducted to evaluate the robustness of the observed host effect. The black box denotes the pre-specified filtering regime used in all inferential analyses. *OR* = odds ratio; *IQR* = interquartile range

Discussion

Cattle are more likely to exhibit multiple Gp60 alleles than humans

The stratified analysis revealed that cattle are substantially more likely than humans to harbour multiple Gp60 alleles (approximately 2.3-fold higher odds), which is likely to reflect a greater Gp60 subtype polyclonality rate. Inference is strengthened by the persistence of the host effect after adjustment for continent, sequencing depth, and Gp60 family in a multivariable model, indicating that the contrast is not readily explained by measured confounding. An association was observed between polyclonality and major-allele length, with longer alleles exhibiting increased odds of within-sample diversity. This effect was independent of host: neither a per-continent Mann–Whitney U test nor an OLS model adjusting for host, continent, and their interaction detected any meaningful difference in major-allele length between humans and cattle (Europe: $p = 0.60$, North America: $p = 0.95$, OLS $p = 0.48$, adjusted $R^2 = -0.001$). This length effect is biologically plausible, as longer STRs are more prone to slipped-strand mispairing during replication. However, because the generation of PCR slippage artefacts share a similar mechanistic origin, distinguishing biological from technical sources of variation in longer alleles is non-trivial. Accordingly, the observed association should be interpreted with caution, and future studies leveraging high-fidelity amplification or long-read sequencing may help disentangle these effects.

Because estimates of multiplicity at STR loci are sensitive to artefactual alleles, a conservative, pre-specified filtering scheme was adopted to suppress PCR stutter. An $L1$ near 0.15 is widely used in forensic STR work [28, 29], although lower values are sometimes chosen [27, 30]. Importantly, those thresholds were developed for fragment sizing rather than NGS allele frequencies, so they should be treated as heuristic. To assess sensitivity, we performed an exploratory sweep over allele-count and $L1$ values. The cattle-human contrast remained directionally stable across a neighbourhood around the pre-specified settings, whereas very lenient filters attenuated the signal in a manner consistent with rising stutter noise rather than a biological reversal. We therefore base formal conclusions on the pre-specified threshold set and view the sweep as a robustness check rather than a multiple-comparison exercise.

The elevated polyclonality observed in cattle, combined with the association between major allele length and polyclonality, suggests two non-exclusive mechanisms underlying within-host diversity at the Gp60 locus:

1. **In-host diversification** via slipped-strand mispairing, which generates novel alleles during replication. This mechanism is supported by the observed positive association between major-allele length and polyclonality, consistent with the known propensity of longer STR tracts to undergo slippage.
2. **Superinfection** from multiple sources, resulting in the co-occurrence of genetically distinct Gp60 alleles within the same host. This is supported by the markedly higher polyclonality rates observed in cattle, which are likely driven by intense transmission pressure in agricultural settings, high environmental oocyst burdens, and repeated exposure via shared water or fomites.

Together, these findings point to both de novo generation of allelic variants and repeated exogenous introduction as contributors to Gp60 polyclonality. From a public-health perspective, frequent mixed infections in livestock may increase the risk of zoonotic spillover and facilitate genetic exchange between divergent lineages, underscoring the importance of targeted surveillance at the animal–human interface.

Gp60 subtyping interrogates a single hypervariable locus, and consequently mixed infections detected here may under-represent genome-wide diversity. Longitudinal sampling and whole-genome assays will be important to confirm the host contrast, quantify within-host dynamics, and determine whether polyclonality in cattle reflects repeated exposure, prolonged carriage, or enhanced within-host diversification.

Allelic associations, host plasticity, and geographic structure

Analysis of allele co-occurrence among polyclonal samples, conditioned on continent of origin, revealed both positive and negative associations (Fig. 6). Several positively associated pairs differed by a single repeat unit, consistent with recent divergence via replication slippage or co-transmission under high-exposure conditions. In contrast, multiple allele pairs exhibited statistically significant mutual exclusivity, including cross-family exclusions (e.g., A5G3-IIc with A15G1R1-IIa). These were detected within continents, suggesting that geographic segregation alone cannot explain the pattern. The most prevalent allele, A15G2R1-IIa, showed widespread exclusion of other alleles, likely reflecting a combination of high baseline frequency and genuine competitive incompatibility. These patterns could arise from within-host competition (e.g., differential fitness, immune evasion, or niche occupancy) or from biased introduction histories; distinguishing these mechanisms will require replication at additional loci, whole-genome metrics, and functional follow-up.

While artefactual contributions from PCR stutter cannot be fully excluded, filtering parameters and robustness checks across a broad threshold sweep support the qualitative validity of these findings. Nonetheless, because PCR stutter and slipped-strand mispairing can produce indistinguishable one-repeat variants at the Gp60 locus,

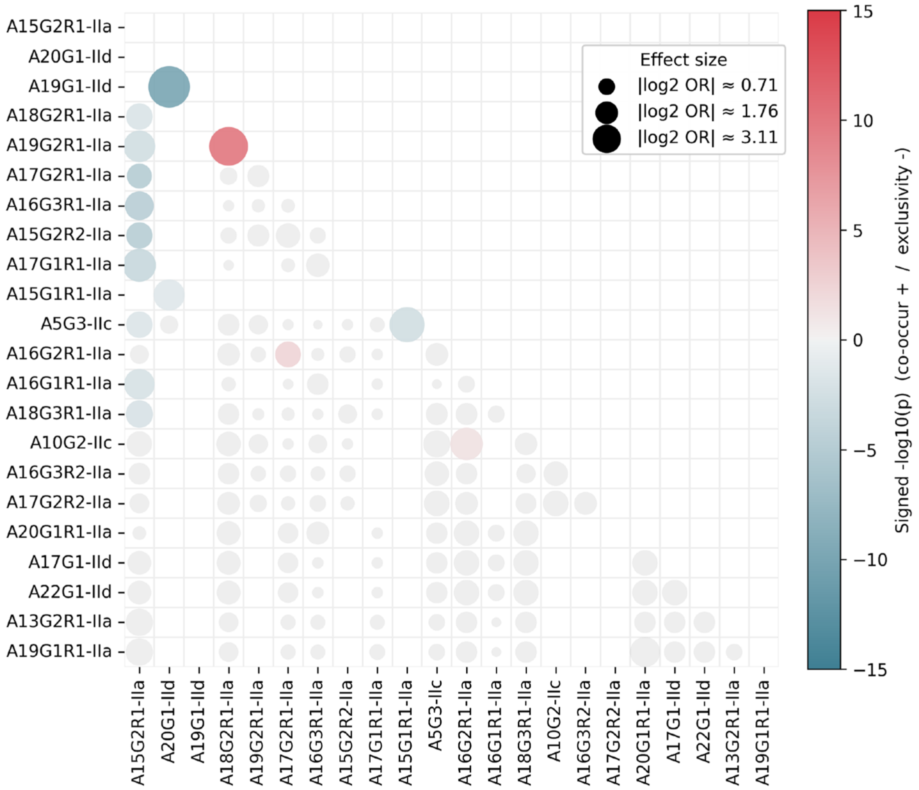

Fig. 6 Allele associations conditioned on continent using the Cochran–Mantel–Haenszel (CMH) test. Colour = signed $-log10$(one-sided p); red indicates positive association (co-occurrence), blue negative (mutual exclusivity). Bubble area is proportional to $|log2(OR)|$ where *OR* is the CMH pooled odds ratio. Grey bubbles mark non-significant pairs ($p \geq 0.05$)

complete removal of stutter-derived artefacts cannot be guaranteed. Consequently, caution is warranted when interpreting associations involving near-neighbour repeat variants due to the possibility of residual stutter artefacts.

Subtype distribution patterns were broadly consistent with established host associations and geographic trends. A15G2R1-IIa was the most frequently detected allele across the dataset, predominating in human and small ruminant samples, and aligning with its previously reported global ubiquity and non-host specificity [31–33]. In contrast, A19G1-IId dominated cattle samples from Asia, although this signal likely reflects sampling bias, as all Asian sequences originated from Chinese cattle. North American and European samples exhibited considerable overlap in subtype composition but also included region-specific alleles such as A15G2R2-IIa in North America and A17G1R1-IIa in Europe. A subset of alleles traditionally regarded as host-restricted displayed unexpected distributions. Notably, A19G2R1-IIa, previously described as cattle-associated [34], was recovered from 12 human samples in this dataset. These findings indicate that certain Gp60 subtypes may exhibit broader host plasticity than previously recognised, and underscore the value of continued genomic surveillance to track potential shifts in host adaptation.

MOI may act as a 'genetic crucible' for diversification

Across continents, A15G2R1-IIa was widespread in humans and small ruminants, whereas A19G1-IId dominated cattle samples taken in China. The detection of IIa/IIc co-infections in North American humans and IId co-existing with a *C. tyzzeri* IXa-like allele in a mouse highlights the potential for subtypes, subspecies, and even entirely different species with different host affinities to undergo recombination through sexual reproduction. We hypothesise that such mixed infections, a necessary prerequisite for recombination, may act as a *genetic crucible*, facilitating diversification through within-host recombination. Such recombinant populations of *Cryptosporidium spp.* are likely to present substantial challenges in the control of this parasite worldwide. These results warrant an investigation into the association between MOI and recombination on larger scale, multi-continent datasets, and using multiple markers or genomics to robustly test this hypothesis.

Technical limitations of BlooMine

The primary limitation of BlooMine is its reliance on the complete query (target sequence + flanking regions) being captured by enough reads to carry out MOI analysis. Larger query regions will require larger read sizes which are not produced by illumina sequencing platforms. However, the advent and employment of long-read technologies will facilitate the analysis of longer and more complex repeat rich regions, with which polyclonality may be characterised. Future work on BlooMine will support the analysis of formats commonly utilised by these long-read sequencing platforms.

Conclusion

Polyclonality in *Cryptosporidium parvum* shapes transmission, host adaptation, and diversification. Continent-stratified analyses indicate cattle have higher odds of multi-subtype Gp60 infections, consistent with the hypothesis that intensive agriculture elevates transmission and, in turn, polyclonality. Patterns of allele co-occurrence may suggest rapid STR-mediated microevolution, while mutual exclusivity between certain alleles points to possible competitive exclusion. Some subtypes exhibited unexpected host distributions, challenging assumptions about host restriction. These findings raise the hypothesis that elevated polyclonality may act as a genetic crucible, facilitating within-host recombination and subtype diversification. This hypothesis is consistent with the observed allele diversity and co-occurrence patterns we report here, and should be rigorously explored in future multi-locus or whole-genome studies. By enabling high-resolution, alignment-free detection of STR variation from NGS data, BlooMine provides a scalable tool for investigating within-host diversity and provides a framework for future MOI surveillance and evolutionary studies.

Abbreviations

AF	Allele frequency
BD	Breslow–Day test for homogeneity of stratum-specific odds ratios
CI	Confidence interval
CMH	Cochran–Mantel–Haenszel test / pooled analysis across strata
DNA	Deoxyribonucleic acid
FASTQ	Text-based format for sequencing reads (per-base calls and quality)
INDEL	Insertion/deletion variant
IQR	Interquartile range
k-mer	Substring of length *k* used in sequence analysis
MH OR (a.k.a. OR_MH, ORMH)	Mantel–Haenszel pooled odds ratio across strata
MOI	Multiplicity of infection
NGS	Next-generation sequencing
OLS	Ordinary least squares (regression)
OR	Odds ratio
PCR	Polymerase chain reaction
SNP	Single nucleotide polymorphism
SPaln	Second-pass pseudo-alignment step in BlooMine
STR	Short tandem repeat (microsatellite)
VNTR	Variable-number tandem repeat
WGS	Whole-genome sequencing

Gene/locus shorthand

Gp60	60 kDa surface glycoprotein gene used for Cryptosporidium subtyping; alleles denoted as patterns such as `A15G2R1` (counts of TCA, TCG, and terminal ACATCA motifs in *C. parvum*), with family label (e.g., `A15G2R1-IIa` in this paper; `IIaA15G2R1` conventionally when molecular typing is carried out)

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-12206-4.

Supplementary Material 1

Acknowledgements

We thank Rachel Chalmers and Guy Robinson at the UK Cryptosporidium Reference Unit for their support over the term of the KESS-II funded project.

Authors' contributions

A.M. developed and wrote BlooMine, carried out all testing, and wrote the first draft of the manuscript. M.S. and J.P. wrote the proposal which lead to the development of BlooMine, and aided in its development. T.C. advised on statistical analysis of BlooMine and interpretation of PCR stutter. All authors contributed to the finalisation of the manuscript.

Funding

This work was funded by the Wellcome Trust (grant ID 215800/Z/19/Z), Health and Care Research Wales (grant ID AF-24-10), and the Knowledge Economy Skills Scholarship (KESS-II).

Data availability

The source code for BlooMine is available open source at https://github.com/ArthurVM/BlooMine. Supplementary data, metadata for samples included in this analysis, BlooMine reports, filtering scripts, and python notebooks containing code used in analysis can be found at https://github.com/ArthurVM/BlooMine_paper_supplementary.

Declarations

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 22 May 2025 / Accepted: 9 October 2025
Published online: 20 November 2025

References

1. Sow SO, Muhsen K, Nasrin D, Blackwelder WC, Wu Y, Farag TH, 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. 2016;10(5):e0004729. https://doi.org/10.1371/journal.pntd.0004729.
2. Morris AV, Robinson G, Swain MT, Chalmers RM. Direct sequencing of *Cryptosporidium* in stool samples for public health. Front Public Health. 2019;7:360. https://doi.org/10.3389/fpubh.2019.00360.
3. 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. https://doi.org/10.1186/s12864-015-1805-9.
4. Morris AV, Robinson G, Chalmers R, Cacciò S, Connor T. Parapipe: a pipeline for parasite next-generation sequencing data analysis applied to

Cryptosporidium. Access Microbiol. 2025;7(8):000993.v3. https://doi.org/10.1099/acmi.0.000993.v3. PMID: 40988887; PMCID: PMC12451305.

5. Morris AV, 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 ofthe 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 2019. pp. 90–100. https://doi.org/10.5220/0007397200900100.
6. Morris AV, Pachebat J, Tyson G, Robinson G, Chalmers R, Swain M. Generating Reliable Genome Assemblies of Intestinal Protozoans from Clinical Samples for the Purpose of Biomarker Discovery. In: Biomedical Engineering Systems and Technologies. Communications in Computer and Information Science. 2020. pp. 216–241. https://doi.org/10.1007/978-3-030-46970-2_11.
7. Lord CC, Trenholme K, Day K, Hargrove JW, Woolhouse MEJ, McNamara JJ, et al. Aggregation and distribution of strains in microparasites. Philos Trans R Soc Lond B Biol Sci. 1999;354:799–807. https://doi.org/10.1098/rstb.1999.0432.
8. Seppälä O, Karvonen A, Valtonen ET, Jokela J. Interactions among co-infecting parasite species: a mechanism maintaining genetic variation in parasites? Proc Biol Sci. 2009;276(1657):691–7. https://doi.org/10.1098/rspb.2008.1229.
9. Seppälä O, Karvonen A, Rellstab C, Louhi KR, Jokela J. Reciprocal interaction matrix reveals complex genetic and dose-dependent specificity among coinfecting parasites. Am Nat. 2012;180:306–15. https://doi.org/10.1086/666985.
10. Graham AL. Ecological rules governing helminth microparasite coinfection. Proc Natl Acad Sci U S A. 2008. https://doi.org/10.1073/pnas.0707221105.
11. 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. https://doi.org/10.1016/j.meegid.2012.08.017.
12. 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. https://doi.org/10.1016/j.ijpara.2017.03.003.
13. Kaupke A, Gawor J, Rzeżutka A, Gromadka R. Identification of pig-specific cryptosporidium species in mixed infections using illumina sequencing technology. Exp Parasitol. 2017. https://doi.org/10.1016/j.exppara.2017.09.020.
14. 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;17:1–12. https://doi.org/10.1186/s12864-016-2815-y.
15. Swain MT, Vickers M. Interpreting alignment-free sequence comparison: what makes a score a good score? NAR Genomics Bioinforma. 2022 09;4(3):lqac062. https://doi.org/10.1093/nargab/lqac062.
16. Abe N, Matsubayashi M, Kimata I, Iseki M. Subgenotype analysis of *cryptosporidium parvum* isolates from humans and animals in Japan using the 60-kda glycoprotein gene sequences. Parasitol Res. 2006;99:303–5. https://doi.org/10.1007/s00436-006-0140-0.
17. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013;5(3):28. https://doi.org/10.1186/gm432.
18. Bloom BH. Space/time trade-offs in hash coding with allowable errors. Commun ACM. 1970;13(7):422–6. https://doi.org/10.1145/362686.362692.
19. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7. https://doi.org/10.1016/0022-2836(81)90087-5.
20. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53. https://doi.org/10.1016/0022-2836(70)90057-4.
21. Strong WB, Gut J, Nelson RG. Cloning and sequence analysis of a highly polymorphic *Cryptosporidium parvum* gene encoding a 60-kilodalton glycoprotein and characterization of its 15- and 45-kilodalton zoite surface antigen products. Infect Immun. 2000;68(7):4117–34. https://doi.org/10.1128/iai.68.7.4117-4134.2000.
22. Robinson G, Chalmers RM, Elwin K, Guy RA, Bessonov K, Troell K, et al. Deciphering a cryptic minefield: a guide to cryptosporidium gp60 subtyping. Curr Res Parasitol Vector-Borne Dis. 2025;7:100257. https://doi.org/10.1016/j.crpvbd.2025.100257.
23. Xiao L, Feng Y. Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and *Giardia duodenalis*. Food Waterborne Parasitol. 2017;8–9:14–32. https://doi.org/10.1016/j.fawpar.2017.09.002.
24. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992;20(2):211–5. https://doi.org/10.1093/nar/20.2.211.
25. Riman S, Iyer H, Borsuk LA, Vallone PM. Understanding the behavior of stutter through the sequencing of STR alleles. Forensic Sci Int Genet Suppl Ser. 2019;7(1):115–6. https://doi.org/10.1016/j.fsigss.2019.09.045.
26. Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, et al. DNA commission of the international society of forensic genetics: recommendations on the interpretation of mixtures. Forensic Sci Int. 2006;160(2):90–101. https://doi.org/10.1016/j.forsciint.2006.04.009.
27. Chen DP, Tseng CP, Tsai SH, Wu TL, Chang PY, Sun CF. Systematic analysis of stutters to enhance the accuracy of chimerism testing. Ann Clin Lab Sci. 2008;38(3):264–72.
28. Seo SB, Ge J, King JL, Budowle B. Reduction of stutter ratios in short tandem repeat loci typing of low copy number DNA samples. Forensic Sci Int Genet. 2014;8(1):213–8. https://doi.org/10.1016/j.fsigen.2013.10.004.
29. Leclair B, Frégeau CJ, Bowen KL, Fourney RM. Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous str loci for forensic casework and database samples. J Forensic Sci. 2004;49(5):968–80.
30. Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM, Smerick JB, Budowle B. Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. J Forensic Sci. 2001;46(3):647–60.
31. Bacchetti R, Connelly L, Browning L, Alexander CL. Changing molecular profiles of human cryptosporidiosis cases in Scotland as a result of the coronavirus disease, COVID-19 pandemic. Br J Biomed Sci. 2023. https://doi.org/10.3389/bjbs.2023.11462.
32. Bellinzona G, Nardi T, Castelli M, Biffignandi GB, Adjou K, Betson M, et al. Comparative genomics of *Cryptosporidium parvum* reveals the emergence of an outbreak-associated population in Europe and its spread to the United States. Genome Res. 2024. https://doi.org/10.1101/gr.278830.123.
33. Avendaño C, Ramo A, Vergara-Castiblanco C, Sánchez-Acedo C, Quílez J. Genetic uniqueness of *Cryptosporidium parvum* from dairy calves in Colombia. Parasitol Res. 2018. https://doi.org/10.1007/s00436-018-5818-6.
34. Lichtmannsperger K, Harl J, Freudenthaler K, Hinney B, Wittek T, Joachim A. Cryptosporidium parvum, Cryptosporidium ryanae, and Cryptosporidium bovis in samples from calves in Austria. Protozoology. 2020:4291–5. https://doi.org/10.1007/s00436-020-06928-5/Published.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.