Detecting sequence variants in clinically important protozoan parasites

Larissa Calarco[a]*, Joel Barratt[a], and John Ellis[a]

[a] School of Life Sciences, University of Technology Sydney,
PO Box 123, Broadway,
NSW 2007
Australia

***Corresponding author**: Larissa Calarco
School of Life Sciences, University of Technology Sydney,
PO Box 123, Broadway,
NSW 2007
Australia
E: Larissa.M.Calarco@student.uts.edu.au
P: +61 2 9514 4161
F: +61 2 9514 8206

ABSTRACT

Second and third generation sequencing methods are crucial for population genetic studies, and variant detection is a popular approach for exploiting this sequence data. While mini- and microsatellites are historically useful markers for studying important protozoa such as *Toxoplasma* and *Plasmodium* sp., detecting non-repetitive variants such as those found in genes, can be fundamental to investigating a pathogens' biology. These variants, namely single nucleotide polymorphisms (SNPs) and insertions and deletions (indels), can help elucidate the genetic basis of an organism's pathogenicity, identify selective pressures, and resolve phylogenetic relationships. They also have the added benefit of possessing a comparatively low mutation rate, which contributes to their stability. However, there are a plethora of variant analysis tools with nuanced pipelines and conflicting recommendations for best practice, which can be confounding. This lack of standardisation means that variant analysis requires careful parameter optimisation, an understanding of its limitations, and the availability of high-quality data. This review explores the value of variant detection when applied to non-model organisms, such as clinically important protozoan pathogens. The limitations of current methods are discussed, including special considerations that require the end-users' attention to ensure that the results generated are reproducible, and the biological conclusions drawn are valid.

Keywords: variant analysis, population genetics, evolutionary selection, NGS data, non-model organism, structural variants, SNPs

1. Introduction

The process of mutation gives rise to genetic variation in non-coding and coding genomic regions resulting in genetic differences that accumulate over time, slowly culminating in the molecular divergence of populations. These genetic differences can be exploited to distinguish divergent populations at the molecular level. Repetitive sequences that occur in coding and non-coding DNA, known as mini- and microsatellites, have become popular genotyping markers for use in population diversity studies involving many protozoan pathogens (Al-Qassab et al., 2010; Anderson et al., 2000; Basso et al., 2009; Das et al., 2016). However, it is the study of variants present in genes that advances our understanding of whether a mutation has altered a genes' function and consequently, the organism's phenotype. Genetic variants considered pertinent to the study of gene function are single nucleotide polymorphisms (SNPs), and insertions & deletions (indels). Larger sequence variants are also biologically and functionally important polymorphisms, and include structural variants (SVs) such as transversions, copy number variation (CNV), inversions, and duplications. Each of these polymorphisms can be identified through *in-silico* variant analysis methods applied to next- and third generation sequencing (TGS) data, which have superseded Sanger sequencing for many applications. Ultimately, demand for the development and improvement of *in-silico* tools to perform variant analyses stems from the continued advancement and increasing availability of high-throughput sequencing technologies.

This review provides an update on the state of the art regarding sequencing technologies and variant calling pipelines, and their applications in parasitology. These topics are first discussed from a generalist perspective by highlighting the importance, applications, and challenges associated with sequencing and variant detection. Next, the discussion focuses on how these technologies apply to the study of clinically important protozoan pathogens, and considerations that require attention when studying unique, complex, and peculiar parasite

genomes. Advances arising from the application of these technologies to several protozoan pathogens are highlighted, including population genetic studies of *Toxoplasma gondii*, *Leishmania* sp., and *Trypanosoma cruzi*, and the detection of drug resistance variants in *Plasmodium* species. Finally, we provide recommendations on best practice with regards to variant detection, including its application to non-model organisms, in which case, robust genomic resources are often unavailable.

2. Current sequencing technologies and their challenges

2.1. Overview of current technologies

The advancement of first and second generation sequencing (SGS) technologies over the past 20 years has revolutionised genetic research, facilitating several major scientific advancements. The evolution of next generation sequencing (NGS) technologies has seen the development of platforms that boast high speed, massive throughput, enormous data generation, and affordability (Ambardar et al., 2016). Such technology has also been applied in fields of diagnostics and sequencing of organellar genomes (Flaherty et al., 2018; Jex et al., 2010; Roeber et al., 2013). However, the ongoing development of new technologies continues, each offering a range of advantages and limitations that vary between platforms. Consequently, the choice of sequencing platform must be considered depending on the desired outcome or specific research question, as well as the nature of the organism under investigation.

Second generation sequencing technologies first emerged with the commercialisation of Roche's 454 pyrosequencing platform in 2005, which accommodates a wide scope of applications including RNA- and DNA-sequencing, metagenomics, and targeted amplicon sequencing (Ambardar et al., 2016; Tripathi et al., 2016). Following the release of the Genome Analyzer platform in 2007 however, commercially available Illumina platforms have become the standard for high-throughput, massively paralleled sequencing, and are the only platforms

capable of paired-end sequencing (Ambardar et al., 2016; Quail et al., 2012). This facilitates the production of high-quality data, leads to higher read coverage, and aids in the discovery of structural variants and repetitive sequence elements (Ambardar et al., 2016). Other notable SGS technologies include ABI/Life Technologies' SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform and Life Technologies' Ion Torrent PGM (Personal Genomic Machine). The SOLiD platform executes multiple sequencing rounds resulting in an overall base calling accuracy of >99.85% (Kchouk et al., 2017; Mardis, 2013), whereas Ion Torrent sequencers produce comparatively longer reads between 35-400 bp (average 200 bp), with higher throughput and faster run times compared to other SGS platforms (Eid et al., 2009; Rothberg et al., 2011).

Alternatively, rather than relying on PCR to enrich DNA template prior to sequencing, developing TGS technologies target single DNA molecules directly through single molecule real time (SMRT) sequencing technology (Braslavsky et al., 2003; Eid et al., 2009; Harris et al., 2008). This results in longer reads being generated, a faster sequencing time, and eliminates some sequencing biases introduced by PCR amplification (Lu et al., 2016; Schadt et al., 2010). Pacific Biosciences' (PacBio; www.pacb.com) commercialisation of SMRT sequencing in 2011 paved the way for this technology, and is currently the most widely used TGS technology commercially available (Eid et al., 2009). Its advantages include the production of reads substantially longer than those generated by any SGS technology, averaging up to 30,000 bp, and comparatively rapid sample preparation (Chin et al., 2016; Liu et al., 2012). In 2014 a new TGS platform known as the MinION device was released by Oxford Nanopore Technologies (ONT) (Lu et al., 2016), where its main appeal is portability and affordability, making it conducive to real-time applications. Furthermore, its ability to generate long reads (generally >10 kilobases (kb)) lends itself to the detection of structural genomic variants and repeat

sequences, which is especially relevant for the complex genomes of many Protozoa (Laehnemann et al., 2016; Lu et al., 2016; Mikheyev and Tin, 2014).

2.2. Limitations and challenges of sequencing technologies

While the advent of SGS technologies saw unprecedented large and affordable throughput, these platforms were not without their limitations. The short sequencing reads produced by SGS platforms for example are not conducive to *de novo* genome analyses, and can result in the generation of highly fragmented assemblies, which is particularly problematic for the large repetitive genomes of eukaryotic pathogens (Ambardar et al., 2016; Korhonen et al., 2016; Schatz et al., 2010). With the development of TGS technologies however, arguably the greatest concern is that TGS platforms currently introduce sequencing errors at rates approximately 10-15% higher than SGS platforms (Mardis, 2013; Nagarajan and Pop, 2013). Furthermore, ONT base calling errors are currently higher than PacBio, with correct base calling rates of approximately 65-88% (Ashton et al., 2015; Ip et al., 2015; Laver et al., 2015).

When selecting a specific SGS or TGS technology for a sequencing task, the main trade-off between platforms is data volume versus read length. The economical production of copious amounts of sequence data using the highly paralleled sequencing chemistries offered by SGS technologies is available at the expense of read length and PCR bias. This is compared to the rapid, real-time production of long reads from TGS technologies that improve *de novo* assembly quality, where these advantages exist at the expense of higher error rates and lower throughput. In an attempt to overcome these challenges, it has become common practice to perform hybrid *de novo* assemblies, whereby reads generated by SGS and TGS platforms are combined in the same assembly, where the lower error rates introduced by SGS platforms offset the high error rates of TGS platforms, while the longer TGS reads help to close genomes. This approach in particular has gained momentum for various protozoans, where previously

published genomes suffer from the disadvantages of using data generated from either sequencing technology alone (Batra et al., 2019; Bruske et al., 2018; Diaz-Viraque et al., 2019; Gonzalez-de la Fuente et al., 2018; Gonzalez-de la Fuente et al., 2017).

2.3. Sequencing considerations for clinically significant Protozoa

It is well accepted that the more complex and repetitive the genome, the lower the quality of the assembled genome sequence. This is especially true for pathogenic Protozoa, where the unique nature of their genomes poses many challenges pertinent to sequencing and subsequent analysis of generated reads. For example, short reads generated by SGS platforms make it nearly impossible to assemble repetitive regions that are characteristic of many parasite genomes. However, long reads generated using newer TGS platforms provide an attractive alternative for addressing such challenges. Subsequently, many studies have emerged recently that take advantage of new TGS technologies, to improve the genome quality, completeness, and annotation of clinically significant parasites (Berna et al., 2018; Chien et al., 2016; Otto et al., 2018; Vembar et al., 2016). However, the advantages and appeal of NGS platforms are impeded by their costs, sample preparation, and availability in remote hospitals and field settings. This is a concern in developing countries where many pathogenic Protozoa are endemic. New platforms such as Nanopore's MinION sequencer aim to address this, where laborious sample preparation and skilled technicians are not required (Lu et al., 2016).

Limitations in the genome assemblies of protozoans hinder precise comparative genomics and transcriptomics, gene expression studies, and gene content analysis, which are crucial for understanding the nature and progression of these diseases (Berna et al., 2018; Chien et al., 2016). Furthermore, gaps or absent regions within genome assemblies due to the limitations of available technologies, impede the detection and analysis of genetic variation such as indels, SVs, CNVs, chromosomal rearrangements, and hypervariable multi-gene

families (Kwiatkowski, 2015). Due to the inability of many NGS technologies to address the peculiarities of parasite genomes such as *P. falciparum*, sequencing studies tend to focus on small variants such as SNPs and indels, and neglect or underestimate the presence and importance of large structural variants in highly repetitive and hypervariable regions.

While *Plasmodium falciparum* has a comparatively small eukaryotic genome at ~23 Mb in length, it has a high repeat content of 51.8%, and is AT-rich with an AT content of 80.6% (Gardner et al., 2002; Girgis, 2015). Furthermore, the genomes of many *Plasmodium* species have polymorphic, repetitive subtelomeric regions encoding multi-gene virulence families (Chien et al., 2016; Su et al., 1995; Vembar et al., 2016). As a result, NGS capabilities have fallen short when employed to accurately sequence *P. falciparum* and other *Plasmodium* species (Oyola et al., 2014; Oyola et al., 2012; Quail et al., 2012), and the use of available genome sequences as references for clinical isolates has been called into question when studying genetic diversity (Kwiatkowski, 2015). Routine use of PCR-based whole genome amplification (WGA) has previously contributed to short read sequencing of clinical and laboratory-derived *P. falciparum* strains (Ariey et al., 2014; Kamau et al., 2015; Manske et al., 2012), where multiplexed Illumina libraries can also be generated using very low genomic DNA quantities for this species (Oyola et al., 2014). This is especially relevant in the context of processing clinical samples either in the field, or other resource limited settings. However, it has been suggested that PCR-induced bias or sequencing errors can result in overestimating SNP numbers (Oyola et al., 2014; Vembar et al., 2016).

Newer technologies such as PacBio's SMRT sequencing have subsequently been exploited to overcome these limitations. Vembar *et al.* (2016) performed amplification-free long-read sequencing of *P. falciparum* genomic DNA, where the produced reads were used to generate a complete telomere-to-telomere *de novo* assembly. This method also resolved AT-rich centromeres and repetitive subtelomeric regions, and identified large insertions,

duplications, and expansions, where the improved genome was in turn used to estimate *P. falciparum* genetic diversity.

Comparatively, while the publication of genomes from *Leishmania. major* (Ivens et al., 2005), *Trypanosoma brucei* (Berriman et al., 2005), and *T. cruzi* (El-Sayed et al., 2005) in 2005 represented important milestones in trypanosomatid research, each of these genomes were at varying degrees of completion, and were plagued by fragmentation. The sequencing of additional *T. cruzi* strains was subsequently performed using newer NGS technologies such as Roche 454 and Illumina, however issues of fragmentation and the collapsing of repetitive sequences still persisted (Franzen et al., 2012; Grisard et al., 2014). Another known characteristic of *T. cruzi* and other trypanosome genomes is that gene content is greatly expanded, mainly as a result of multi-gene families (Acosta-Serrano et al., 2001; Buscaglia et al., 2006; Frasch, 2000; Pita et al., 2019). The failure of NGS technologies to account for these novel genomic features therefore resulted in miscalculation of protein-coding genes, pseudogenes, copy number estimates, and tandem repeats in these species (Arner et al., 2007; El-Sayed et al., 2005). The advent of TGS technologies has subsequently allowed accurate estimations of gene copy number, tandem and dispersed repetitive sequences, and the correct assembly of homologous chromosomes to retrieve haplotypes (Berna et al., 2018; Callejas-Hernandez et al., 2018; Diaz-Viraque et al., 2019; Jayaraman et al., 2019).

Quail *et al*. (2012) compared the performance of three popular NGS platforms with respect to coverage, GC distribution, variant calling, and accuracy. This study reported several differences between the quality of the data produced by each platform, though the Ion Torrent, PacBio, and MiSeq platforms each displayed almost perfect coverage performance for GC-rich, AT-rich, and neutral genomes. However, approximately 30% of the AT-rich *Plasmodium falciparum* genome had no coverage on Ion Torrent's PGM platform. Additionally, while more true variants could be called from data produced by the PGM platform compared to that of the

MiSeq, the trade-off was a higher false positive rate. Similarly, while variants could also be identified using PacBio data, sequencing depth was lacking.

Ultimately, second and TGS data analysis is complex, requires powerful computational resources, usually involves multi-step workflows, and requires numerous algorithms and software for processing depending on the research question (Pabinger et al., 2014). Furthermore, with the rapid evolution of newer sequencing technologies, selecting the most suitable technology for a specific application, such as variant detection, is becoming increasingly difficult, especially in the context of unique protozoan genomes.

3. Applications of sequencing technologies: variant analysis workflows

An important use of SGS and TGS data is the identification of sequence variants within and between samples. After selecting the most suitable sequencing strategy for an intended application, the choice of analysis tools employed, the associated parameters, and the nature of the organism under investigation, can have a drastic impact on the success of a variant analysis workflow. A variant analysis workflow (Figure 1) first involves the generation of SGS and/or TGS data. Read quality control ensues and includes processes such as read trimming to remove adapter sequence and low quality bases at the ends of reads, and filtering to remove short and poor quality reads. Groomed reads are then typically mapped to a reference genome or transcriptome, to allow for subsequent variant calling and annotation.

231

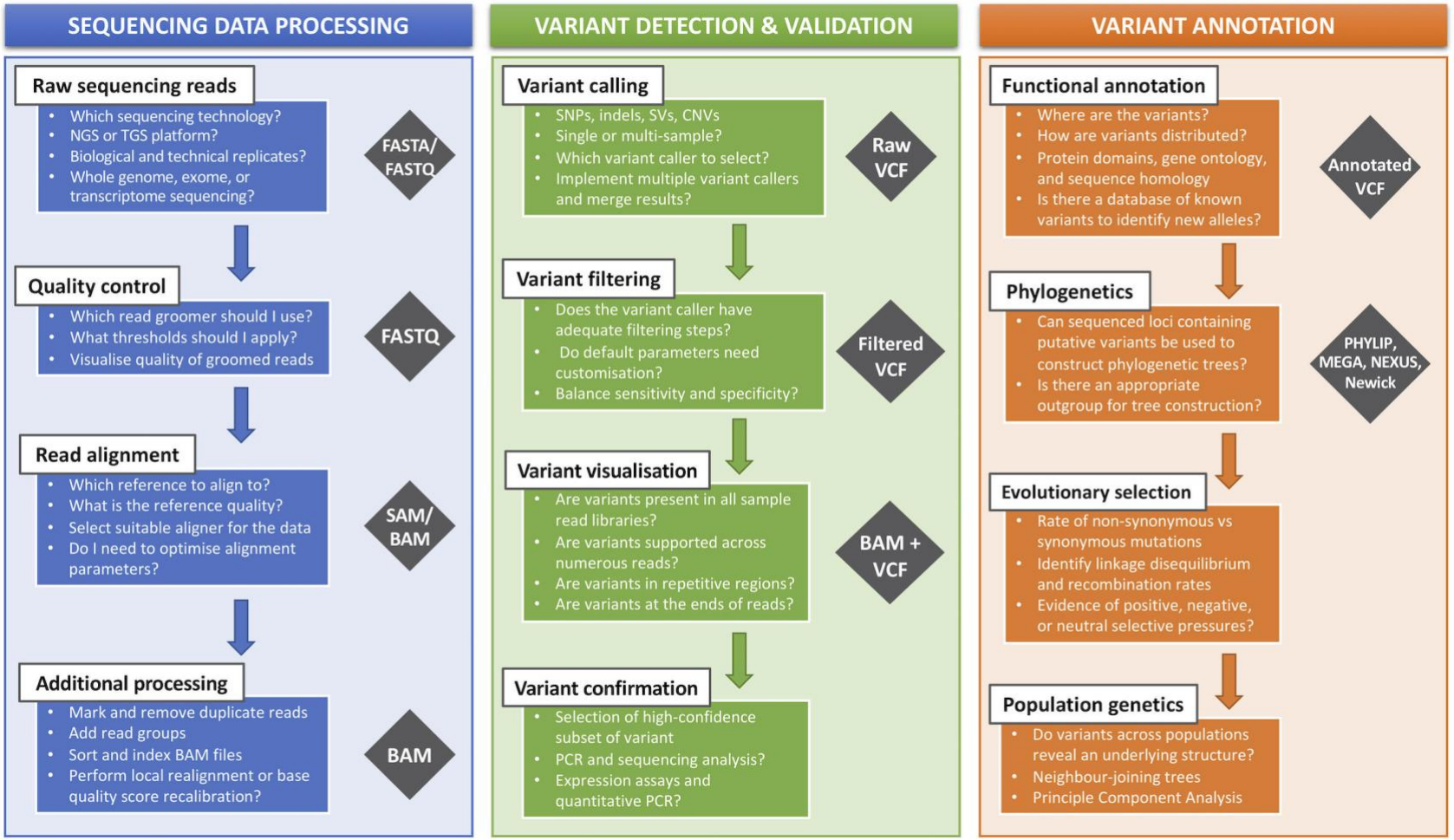


Figure 1. Summary of a typical variant detection workflow for NGS and TGS data.

Following selection of an appropriate platform and generation of sequencing data, analysis begins with *in-silico* data pre-processing. Initially, sequenced reads are quality controlled using an appropriate tool, and subsequently visualised using an interface such as FastQC. The groomed

reads are then aligned to a reference genome using a read mapper such as BWA or Bowtie2, where the user needs to consider alignment parameters, and the quality of the reference. Additional processing steps can also be executed, such as those recommended by the Genome Analysis Toolkit (GATK), to generate an analysis read Binary Alignment Map (BAM) file. Variants such as single nucleotide polymorphisms (SNPs), insertions and deletions (indels), structural variants (SVs), and copy number variants (CNVs) are detected based on sequence differences between the reference and sample reads under investigation, with respect to read mapping, quality, and coverage. The putative variants in a Variant Call File (VCF) are subsequently filtered and visualised to produce a high-quality callset. Next, a set of high-confidence variants identified *in-silico*, is subject to laboratory validation, which can be problematic depending on the type of variant under investigation. This can involve designing primers flanking the predicted variant, and subsequent PCR and sequencing analysis, where returned sequencing data can be evaluated for the presence of the putative variant. Lastly, *in-silico* variants are annotated to elucidate their biological relevance, phylogenetic relationships, population genetics, genotypes, and gene/protein function. Common file types used and generated for each stage of the variant analysis workflow are show in the grey diamonds.

3.1. Read quality control and mapping

Implementation of appropriate quality control measures for raw SGS and TGS reads prior to data analysis is essential to remove reads containing obvious base calling errors, poor quality sequence, small indels, and adaptor sequences (Dai et al., 2010). There are several tools available for performing each of these tasks, many of which perform overlapping functions. These include the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), which is a command-line tool designed to pre-process NGS reads, with the capacity to trim and quality filter reads, in addition to converting file formats and providing summary statistics. Additionally, Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) offers automation in a Perl wrapper script to trim adapter sequences, filter base quality scores, and remove short reads. Subsequently, tools such as FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) can assess the quality of sequence data in an easy to use interface, and report the results in summary graphs and tables. Read grooming results in retention of only high-quality reads which facilitate an accurate alignment, mitigating the risk of calling false variants during downstream analyses (Nielsen et al., 2011).

Accurate alignment of groomed reads from a RNA-seq or DNA-seq experiment is critical for variant calling accuracy, as correct mapping will avoid the erroneous interpretation of misaligned reads as true variants (Piskol et al., 2013). Hence, when selecting one of the several alignment tools available (Table 1), consideration of its tolerance for imperfect matches is critical (Bao et al., 2014). For example, Bowtie2 offers accurate read alignment for reads of varying lengths, as generated from a range of sequencing technologies (Langmead and Salzberg, 2012). This tool has evolved with advancing sequencing chemistries, accommodating increasing throughputs and read lengths. The algorithm also achieves sensitive gapped-read alignments, where gaps can be an error source associated with new single-molecule sequencing technologies. By comparison, the Burrows-Wheeler Aligner's (BWA)

‘BWA-MEM alignment algorithm has been reported to perform well for longer reads, and be more accurate than Bowtie2 (Li, 2013), whereas Bowtie2 is faster and more accurate when handling indels (Langmead and Salzberg, 2012). Furthermore, tools including TopHat2 (Kim et al., 2013), STAR (Spliced Transcripts Alignment to a Reference) (Dobin et al., 2013), and RUM (RNA-seq Unified Mapper) (Grant et al., 2011) are specifically designed for aligning RNA-seq reads, whilst addressing associated challenges such as alternative splicing, indels, gene fusions, and introns.

265 **Table 1. List of available sequence alignment tools for NGS analysis.**

Sequence alignment tool	Features	Reference
TopHat and TopHat2	Spliced aligner for RNA-seq data which can also identify novel splice sites, and produce accurate alignments for highly repetitive genomes in the presence of indels and gene fusions	(Kim et al., 2013; Trapnell et al., 2009)
Bowtie and Bowtie2	Offers accurate alignments for reads of varying lengths produced from a range of sequencing technologies. Bowtie2 achieves sensitive gapped-read alignments, where gaps can be an error source associated with single-molecule sequencing technologies	(Langmead et al., 2009)
Burrows-Wheeler Aligner (BWA)	Efficient and accurate alignment of short and long sequencing reads against large reference sequences. Allows for mismatches and gaps originating from sequencing when performing alignments	(Li and Durbin, 2009, 2010)
SOAP and SOAP2	An ultrafast and memory efficient short read aligner that supports multiple input and output file formats. SOAP and SOAP2 are compatible with both	(Li et al., 2008; Li et al., 2009b)

	single- and paired-end reads and are capable of gapped and ungapped alignments	
Spliced Transcripts Alignment to a Reference (STAR)	Flexible, ultrafast RNA-seq alignment tool compatible with a range of second and third generation sequencing platforms. Able to align high-throughput short and long RNA-seq reads	(Dobin et al., 2013)
MapSplice	Splice detection algorithm that can align both short and long RNA-seq reads. Can be applied to data from model organisms, as well as those with limited transcript annotations	(Wang et al., 2010b)
RNA-seq Unified Mapper (RUM)	RNA-seq alignment algorithm that addresses the main challenges associated with RNA alignment such as alternative splicing, indels, base substitutions, base calling errors, and introns	(Grant et al., 2011)

266

A common optimisation strategy for not only selecting a suitable alignment tool, but also an appropriate set of parameters and their thresholds, is to execute multiple rounds of read mapping within and between aligners (Calarco et al., 2018). The suitability of both the aligner and its specific parameters can subsequently be assessed using the read statistics produced by tools such as Bowtie2 and TopHat2. For example, the overall percentage of reads mapped to the reference can be compared, in addition to the number of concordant or discordant pairs mapped, in the case of paired-end reads. Through this approach, the user's own data is being employed to optimise and tailor the read mapping process, as opposed to using the pre-defined default thresholds of the respective tools, which are usually trained on and designed for data generated from model organisms such as humans.

Additional processing of sequencing data and aligned reads is also routinely required by downstream tools. Manipulation of such data in file formats such as SAM, BAM, and VCF can involve marking duplicate reads, performing realignment around potential indel sites, sorting and indexing alignment files, collecting metrics, and converting files. Picard tools (Broad Institute, 2019) and SAMtools (Li et al., 2009a) are extremely valuable toolkits that can perform such commands amongst a plethora of others, and are incorporated into many "gold standard" or "best practice" workflows, such as the Genome Analysis Toolkit (GATK) (McKenna et al., 2010) and VarScan (Koboldt et al., 2009). For example, identifying and removing duplicate sequenced reads is an important processing step, where such reads can occur as a result of library preparation during PCR enrichment. If a PCR duplicate is sequenced multiple times and contains an amplification-derived error, this can introduce bias in downstream variant analysis, where a variant caller may incorrectly identify this error as a true variant, or miscalculate the frequency in which the allele is represented (Ebbert et al., 2016). Adding read group information using these toolkits is also extremely useful, and even required by many variant callers such as VarScan, when attempting to identify variants across multiple

samples or populations. This is especially relevant for Protozoa, where sequencing is frequently performed on multiple isolates or passages, and on clinical samples that may be pooled from multiple patients. The toolkits discussed are user-friendly and are accompanied by extensive documentation and usage recommendations, making them ideal for streamlining analysis pipelines, and also for inexperienced users.

3.2. Variant calling and visualisation

Following alignment, the mapped reads are then subject to a variant detection workflow for identification of SNPs and indels (Pabinger et al., 2014). As some variants may result from sequencing or mapping errors, a balance between sensitivity to minimise false negatives, and specificity to minimise false positives, is essential. Consequently, the variant calling step is generally designed to maximise sensitivity, while downstream filtering offers specificity. Manual visualisation of at least a subset of alignments can be a crucial step in a variant identification workflow, as this can aid in interpreting results and determining the confidence of variant calls. This is also useful as an additional validation step prior to confirmation of certain variants by downstream PCR and Sanger sequencing.

Several visualisation tools are available, which possess useful capabilities, including the visualisation of mapped reads in the context of the reference genome, displaying read mapping quality, and highlighting variants. Visualising read alignment files in the context of a reference can also assist in assessing the suitability of software and pipelines employed. This can include visualising the adequacy of read coverage across specific loci, and for *de novo* assemblies, can be used to aid in the selection and optimisation of assembly and alignment tools and their parameters. Additionally, manually visualising alignments can be used to assess the potential existence of mixed infections or multiple populations present in a sample, based on the proportion of reads containing SNPs and indels. Popular user-friendly tools include the

Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., 2013), Artemis (Carver et al., 2012), and Savant (Fiume et al., 2010).

3.3. Annotating variants and identifying functionally significant mutations

The variants are then annotated to elucidate their functional and biological relevance (Pabinger et al., 2014). Annotation of variants (i.e., assigning relevant biological information to these sites) can include identifying genes effected by the variant, determining whether it falls in a non-coding or coding region, introduces (or removes) a stop codon, or whether it is a silent, missense or nonsense mutation (McCarthy et al., 2014). Annotation tools typically assign general attributes to each putative variant, which helps investigators assess their potential impact on the organism. For example, tools such as ANNOVAR (Wang et al., 2010a) and SnpEff (Cingolani et al., 2012) can provide information on the impact a sequence variant may have on a genes' function, compare results to existing variant databases, predict the coding effects of SNPs and indels, and identify mutation effects such as non-synonymous and synonymous substitutions, frame shifts, stop codon insertions, and mutated start codons.

If a mutation is identified within a protein-coding gene, its functional impact can be predicted using orthology, where similar sequences of known function might be present in databases such as UniProt (The UniProt Consortium, 2017), and identified by BLAST analysis (Conesa et al., 2016). InterPro is a valuable resource that provides functional analysis of protein sequences by integrating several different databases (Finn et al., 2017). The collective database assigns protein sequences to families, predicts domains and other important motifs, provides residue-level annotation, and intrinsic protein disorder predictions. A limitation of some annotation tools however, is their inability to support the submission of a large unified set of variants, rendering their use for only the manual analysis of select variants.

Alternatively, identifying sequence variants within non-coding genomic regions has gained momentum over the last decade, with a focus on how they affect regulatory elements, such as promoters, enhancers, and transcription factor binding sites (TFBSs), and consequently gene expression and disease (Narlikar and Ovcharenko, 2009). Tools such as the VariantAnnotation Bioconductor package (Obenchain et al., 2014) available in R (www.R-project.org), contain useful commands to allocate SNPs to either coding, intron, intergenic, 3′ untranslated, 5′ untranslated, promoter, or splice site regions within a gene. Tools such as these generally require a Generic Feature Format (GFF) file to be available for the organism under investigation, that contains gene sequence annotations. The region-based annotation function component of the ANNOVAR package is also able to identify variants that disrupt enhancers, repressors, and promoters, and those that are located within TFBSs. However, such tools generally only accommodate well-studied model organisms, for which required datasets are available.

The final step in a variant calling workflow involves prioritising the variants down to a reportable or experimentally confirmable set that can be validated in the laboratory by PCR and Sanger sequencing if deemed necessary (Pabinger et al., 2014). One limitation with popular variant calling pipelines (Table 2), is the absence of well-defined filtering strategies and thresholds to apply to individual callsets, where there is currently a scarcity of any consensus or direction in the literature. Nonetheless, Table 3 summarises popular techniques to select for high-quality, functionally significant variants.

361 **Table 2. Features and considerations of popular variant calling software.**

Variant Caller	Features	Considerations	Reference
GATK[I]	Superior processing steps including realignment and base quality score recalibration	Requires a database of known variant sites to perform base recalibration, which is not suitable for non-model organisms or those lacking such resources	(McKenna et al., 2010)
SAMtools	Contains tools for sorting, indexing and formatting input reads, to subsequently subject to variant detection using BCFtools	Support for and extent of filtering options available for variant calls is limited	(Li et al., 2009a)
Geneious	Provides an all-inclusive interface for analysing NGS data, including read pre-processing, alignment, variant calling, visualisation, and annotation	Not available as a free, open-access software package	www.geneious.com
Atlas2	Separately calls and identifies SNPs[II] and indels[III]	Only supports single sample variant calling and is specifically designed for exome sequencing analysis	(Challis et al., 2012)
VarScan	Separately calls SNPs and indels with a short run-time and compatibility with several short aligners	Using default parameters can result in a high false-positive variant discovery rate, and therefore requires optimisation by the user	(Koboldt et al., 2009)

Snippy	Calls SNPs and indels between a haploid reference genome and NGS reads. The results can be used to generate a core SNP alignment, and phylogenomic trees	Originally designed for short reads generated from bacterial genomes, and uses an internal variant caller (FreeBayes)	(Seeman, 2015)
appreci8	An automated variant calling pipeline to detect SNPs and indels, by integrating eight variant calling tools	Tool-specific filtration steps are not taken into consideration, where instead default options for variant calling are applied	(Sandmann et al., 2018)
FreeBayes	Models multi-allelic loci in sets of individuals with non-uniform copy number	The pipeline is based on a minimal pre-processing approach, and does not support additional recalibration steps	(Garrison and Marth, 2012)
DeepVariant	Detects variants with a greater accuracy than conventional methods, using deep neural networks	Run time is considerably slower compared to other gold standard variant callers	(Poplin et al., 2018)

[I] Genome Analysis Toolkit
[II] single nucleotide polymorphisms
[III] insertions and deletions

371 **Table 3. List of recommended filtering strategies to obtain high confidence variant**
372 **callsets, following *in-silico* detection.**

Filtering strategy	Purpose	Reference
Sequence coverage/depth	To ensure the existence of a variant can be substantiated across multiple reads, during the visualisation step	(Reumers et al., 2012)
Reported base quality	Filter based on quality scores assigned to each base during sequencing, that represent the confidence of each base called	(Park et al., 2014)
Strand bias	Filter based on a reported metric that uses the Fisher's Exact Test to detect strand bias in the reads	(Park et al., 2014)
Variants within homopolymer runs	Error source associated with DNA sequencing	(Reumers et al., 2012)
Annotation	Annotate variants to select for those located within functionally significant genomic regions	(Pabinger et al., 2014)
Consensus variants	Final callset should be comprised of consensus variants called by multiple variant calling pipelines	(Bao et al., 2014; O'Rawe et al., 2013; Pabinger et al., 2014)

373

Next Generation Sequencing data analysis can be daunting due to the wealth of tools available, the optimisation and tailoring of pipelines required, and the need to be familiar with implementing algorithms and scripts via command-line. As a result, many easy to use and publicly accessible interfaces and software platforms have been developed to help streamline and automate NGS analysis, including variant detection pipelines with recommendations on best practices. For example, Geneious (www.geneious.com) is a sequence analysis software platform that provides a user-friendly interface of bioinformatics tools and workflows. Importing raw sequencing data in a variety of formats is a simple 'drag and drop' process, where such data can subsequently be pre-processed with integrated tools for trimming, filtering, adaptor removal, and normalisation. The Geneious package accommodates for the analysis of reads of any length generated by Illumina, PacBio, Roche 454, Nanopore, and Ion Torrent platforms, including *de novo* assembly, read alignment to a reference, variant detection, genome visualisation and annotation, and gene expression. Furthermore, the platform offers a range of tutorials and application support for researchers planning on implementing various types of analysis pipelines.

The Broad Institute's GATK offers a range of tools for variant identification and genotyping using high-throughput sequencing data (McKenna et al., 2010). GATK offers an industry standard, best practice pipeline for germline and somatic short variant and structural variant discovery using DNA and RNA-seq data (https://software.broadinstitute.org/gatk/best-practices). While originally designed for the processing of whole genomes or exomes produced by Illumina platforms, this toolkit can be adapted to accommodate for other sequencing technologies and any organism, not just for studying human genetics. It can also perform additional tasks pertinent to pre-processing of high-throughput sequencing data, and offers extensive tutorials and support.

4. The importance of variant detection in molecular research

In biological and medical fields, the association between genotype and phenotype is an essential line of research (Consortium et al., 2010). The advent of NGS technologies has delivered large volumes of DNA sequence data paving the way for an improved understanding of disease processes, gene expression, and population genetics (Nielsen et al., 2011). The increasing availability of SGS and TGS technologies has led to a shift from simply performing genome sequencing for the sake of generating new genomes, towards analysing sequence data to discover novel sequence variants between genomes.

Detection of SNPs and indels offers several advantages over the use of alternative markers such as mini- and microsatellites for research applications in population diversity and genotyping. By nature, SNPs are extremely stable, exhibit low mutation rates, and are present throughout the entire genome (Picoult-Newberg et al., 1999). They are the most common genetic marker (Sachidanandam et al., 2001; Sherry et al., 1999), and are consequently very informative, providing a genome-wide representation of natural variation in populations (Vera et al., 2013). Additionally, despite microsatellites commonly exhibiting greater allelic diversity per locus, SNPs reportedly exhibit strong segregation among populations (Karlsson et al., 2011; Vera et al., 2013), making them an ideal target for identifying loci that may be subject to neutral variation or undergoing selection (Helyar et al., 2011).

4.1. Detecting evolutionary selection

There are a myriad of methods available to support the downstream analysis of confirmed sequence variants that complement the burgeoning field of SNP detection. Generally, these tools attempt to predict the type of selection that may be acting on a protein-coding gene, and to what effect, which provides information on their biological significance (Jeffares et al., 2015). A mutation that surfaces in a population can be classified as advantageous, deleterious,

or neutral (Thiltgen et al., 2017), and elucidating the mechanisms that either result in the maintenance or loss of these sequence polymorphisms is an important question in population genetics (Escalante et al., 1998). Estimating the ratio of non-synonymous to synonymous mutations (d_N/d_S) can reveal whether positive diversifying, negative purifying, or neutral selection is acting on a gene (Jeffares et al., 2015). There are several tools widely available for determining rates of mutation and calculating these statistics.

The PAML software package uses maximum likelihood (ML) for phylogenetic analyses of DNA and protein sequences (Yang, 1997, 2007). Various PAML programs can estimate non-synonymous and synonymous substitution rates in protein-coding sequences from several species within a population and can detect positive Darwinian selection. DnaSP, offers numerous tools for the analysis and visualisation of sequence variation both within and between populations (Rozas et al., 2017). In addition to the commonly exploited loci selection tests centred around synonymous and non-synonymous substitution rates, DnaSP also includes tests that estimate linkage disequilibrium, identify recombination, and test for neutrality (i.e., Tajima's *D* (Tajima, 1989) and Fu and Li's *D* and *F* statistic (Fu and Li, 1993)).

PopGenome exploits the full range of capabilities of the R statistical and graphical environment for population genetics research (Pfeifer et al., 2014). This R package reads DNA alignments and SNP data in a range of formats (FASTA, MEGA, PHYLIP, and VCF to name a few), as well as annotation files in GFF (general feature file) format, and links this data to functionally significant annotations. A key advantage of this software is its support for analysing genome-scale data, and its ability to produce an array of population genetics statistics such as linkage disequilibrium, neutrality, and recombination. In addition to these commonly used statistics, PopGenome offers tests of non-neutral evolution, including the McDonald-Kreitmann test (McDonald and Kreitman, 1991), and calculates a range of fixation indices (i.e., F_{ST}).

Goodswen *et al*. (2018) implemented a pipeline optimised for eukaryotic pathogens that predicts positive selection sites through comparison of synonymous and non-synonymous mutation rates within protein coding genes. When tested on *T. gondii*, the pipeline provided a set of proteins representing potential vaccine candidates, as they were predicted to contain residues exposed to the immune system that are under positive selection. As part of this workflow, specific proteins were predicted to be naturally exposed to the immune system following submission of a set of protein or nucleotide sequences to *Vacceed* (Goodswen et al., 2014), which is an automated, *in-silico* pipeline based on reverse vaccinology, that assigns protein candidates a score between one and zero, where one represents the highest confidence that a given protein is a suitable vaccine candidate. This pipeline incorporates various tools to identify secreted and/or membrane-associated proteins, based on predicted subcellular location, transmembrane topology, signal peptides, and peptide binding to MHC class I and II molecules. Specifically, Goodswen *et al.* (2018) identified surface antigens, and dense granule, microneme, and rhoptry proteins as potential vaccine candidates, as well as two rhoptry proteins (ROP5 and ROP18), that are known determinants of *T. gondii* virulence (Lei et al., 2014; Ma et al., 2017). Similarly, the high rate of polymorphisms detected in genes encoding *Plasmodium* sp. surface proteins, led to the hypothesis that these proteins were experiencing positive selection as a consequence of the pressure exerted by the host's immune system (Hughes and Hughes, 1995). The high rate of non-synonymous compared to synonymous mutations in these genes was indicative of diversifying Darwinian selection. Understanding the selective processes experienced by specific genes can be invaluable for understanding a protein's function, processes of adaptation and gene-level natural selection, gene conservation, and the evolutionary dynamics of genes (Thiltgen et al., 2017). However, the foundation of these analyses is the accurate detection of SNPs and indels.

4.2. Population structure and genetics

Variants detected *in-silico* can subsequently be exploited to discern a populations' genetic structure, where genome-wide SNP studies have the potential to provide a framework for understanding a species' population genetics. Principle component analysis (PCA) is routinely used to analyse SNP data to reveal geographical segregation and genetic diversity within and between populations (Abraham and Inouye, 2014; Aydemir et al., 2018; Iantorno et al., 2017; Su et al., 2012). The construction of neighbour-joining (NJ) trees is another popular method for investigating a populations' genetic structure (Saitou and Nei, 1987). This can be performed using various tools such as the 'nj' function in R's 'ape' package (https://cran.r-project.org/web/packages/ape/).

Revealing the population-level genetic structure of a species is crucial for understanding the distribution of its phenotypic features, epidemiology, and molecular evolution. For protozoan parasites, this might include drug susceptibility patterns or virulence markers that exist between geographically dispersed populations. For example, examining genetic differences between *T. gondii* strains globally led to the discovery of four clonal lineages responsible for most human infections in the Northern hemisphere (Khan et al., 2011a). The observation that little to no sequence variation exists in chromosome Ia between *T. gondii* lineages, also resulted in this entire chromosome being deemed relatively homogenous between the predominate lineages on different continents (Khan et al., 2006; Khan et al., 2011b). This led to the conclusion that chromosome Ia experienced a genetic sweep approximately 10,000 years ago, where the genetic variants on chromosome Ia afforded a significant Darwinian advantage resulting in their rapid geographical spread (Boyle et al., 2006; Khan et al., 2011b).

5. Limitations and challenges of variant analysis

5.1. Considerations within and between variant callers

As the process of SNP and indel detection is based on relatively new technologies, they are not without their limitations. Furthermore, the large number of variant analysis tools available means that the challenge of standardisation and accuracy persists (Hanlee, 2012). It would be erroneous to presume all variant calling tools employ similar approaches to variant detection, and indeed, some tools possess markedly different sensitivities and specificities (O'Rawe et al., 2013). These differences result from inconsistencies in data collection, read alignment methods, the alignment parameters selected, post-alignment processing and variant analysis algorithms. While relatively accurate alignment tools are available for mapping reads to reference sequences, difficulties still exist in determining whether a variant is real or the result of error (Hanlee, 2012). Unfortunately, variant calling remains highly variable depending on the tools and methods used, highlighting the need for improved standardisation.

Several studies have evaluated and compared variant calling pipelines with respect to data type, computational considerations, choice of tools, and interpretation of the results (Altmann et al., 2012; Oliver et al., 2015; Vyas et al., 2016; Xu, 2018). O'Rawe *et al.* (2013) analysed raw sequence data with five available variant calling pipelines, under near-default software parameters and identified a significant number of discrepancies between the tools, including the omission of true functional variants by some of them. It was therefore recommended that the variants called by multiple pipelines be considered for downstream analysis to decrease the possibility of false positives and negatives (Bao et al., 2014; Pabinger et al., 2014). Ideally, several aligners and variant callers should be employed in a consensus approach to identify variants of high confidence (Bao et al., 2014; O'Rawe et al., 2013; Pabinger et al., 2014). Importantly, the calling of variants on multiple replicate samples should

522 be incorporated into a workflow to mitigate the influence of random sequencing errors on false
523 positive variant identification (Bao et al., 2014) (Figure 2).

524

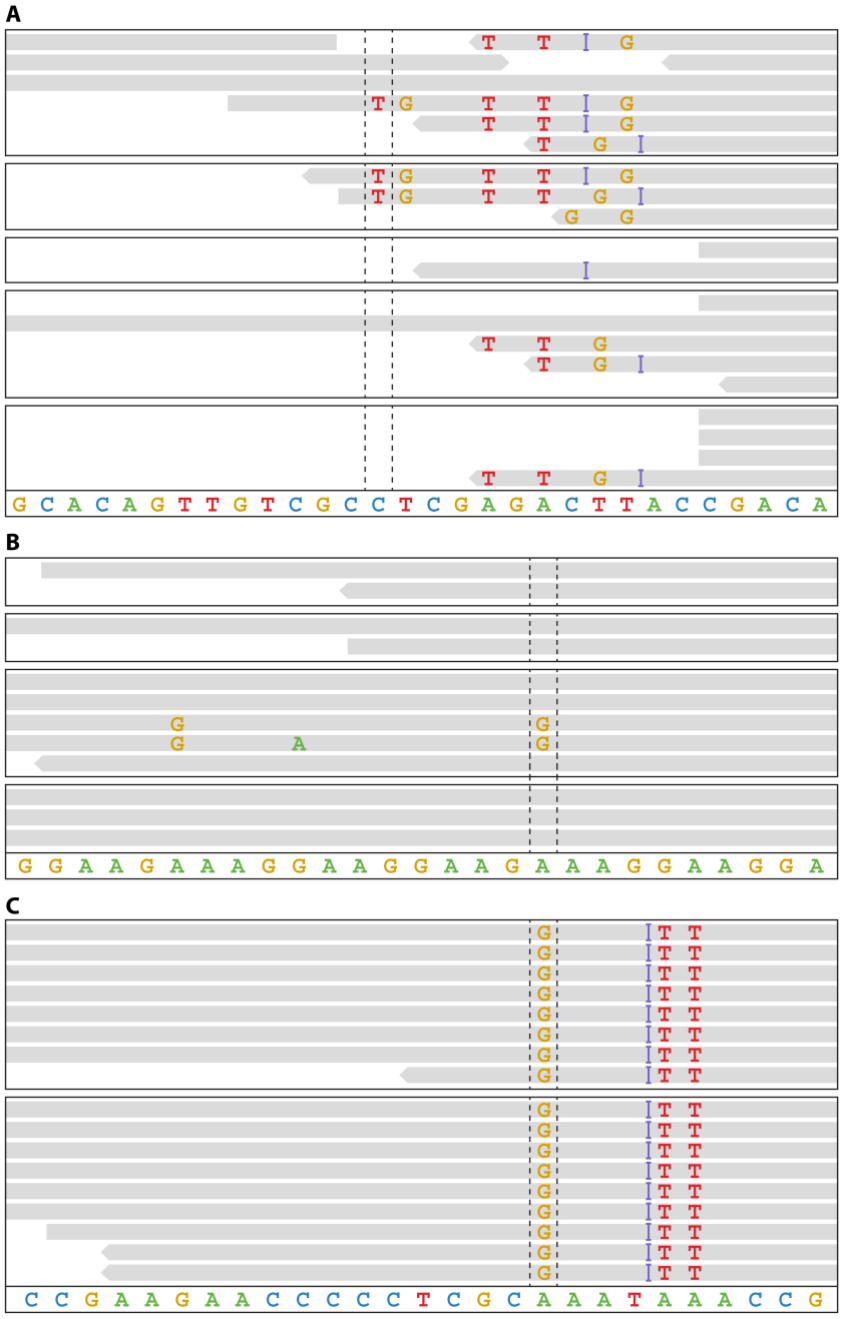


Figure 2. Example of false-positive and true variants identified by a variant analysis and visualised in IGV.

As displayed in IGV (Thorvaldsdottir et al., 2013), the reference sequence is shown along the bottom of each panel and the horizontal grey bars represent the individual Illumina reads that successfully aligned to the reference sequence. Vertical dashed lines highlight the location of a putative variant that was selected for viewing. The bases in each read that differ to the

reference are also shown on the read, and a purple I reflects the presence of an inserted base. Panel (a) shows false-positive variant calls due to the incorrect alignment of reads to the reference sequence likely due to a large insertion in the reference sequence. Panel (b) shows some reads that have been misaligned due to the presence of a repetitive region in the reference sequence, resulting in the calling a false positive variant (a "G" base). Panel (c) shows a set of variants identified that were confirmed and validated by PCR amplification followed by Sanger sequencing. The consistent mapping of reads to this region and the fact that the variants occur towards the middle of the reads (as opposed to the ends) and are present in all reads demonstrate the appearance of true variants in an alignment. Ultimately, this figure highlights the importance of an accurate alignment for calling variants.

5.2. Sources of false positive variants

Correct alignment is essential for accurate variant calling. However, as is the case for many eukaryotic organisms, alignment accuracy is sometimes hampered by the inability of some algorithms to handle differential RNA splicing (Piskol et al., 2013). While some aligners can satisfactorily predict alternatively spliced RNAs from RNA-seq data, they still generate an objectionably high error rate. The use of paired-end sequencing can facilitate the accurate detection of RNA splice variants, and their use is strongly recommended for whole-exome sequencing (Pabinger et al., 2014). In addition to splice variants, short indels and repetitive sequences can be problematic for alignment algorithms, and accurate alignment is often sacrificed for speed (Bao et al., 2014; Piskol et al., 2013). This can result in erroneous alignments that can give rise to false variant calls. Short erroneous indels in sequencing reads can make it difficult for tools to achieve correct alignment and these represent a major source of false positive errors. Variant calling can be improved by performing a realignment step that focuses on areas with potential indels, which is a step recommended in the GATK's Best Practices Workflow (https://software.broadinstitute.org/gatk/best-practices). This step aids in producing clean reads with a consensus indel for subsequent variant identification approaches, for specific regions where misalignments resulting from indels is a possibility. Consequently, manual examination of variant calls is recommended wherever practical to ensure the selected alignment algorithm is performing correctly (Figure 2).

As SGS and TGS data are prone to errors that can lead to false positive variant calls, the tools, filters and parameters employed are crucial to mitigate this. Various studies have investigated the cause of false positive variants and the most effective strategies to improve the accuracy of variant calls (Park et al., 2014; Ribeiro et al., 2015). These studies have found that variant calling accuracy is dependent on several factors including the quality of the reference, the selection of alignment algorithm and variant calling software, the alignment stringency,

and sequencing depth. Considering these caveats, limitations and challenges during the experimental design process is imperative, as every potential variant called represents a hypothesis to be tested. Consequently, the identification of false positive or false negative variants can have significant consequences including the loss of time and resources (O'Rawe et al., 2013), and may also lead to the spread of misinformation which can potentially be more damaging in the long term.

6. Variant detection in non-model organisms such as parasites

A major inadequacy of available variant analysis tools is the lack of recommendations for adaptation to non-model organisms, including many clinically important protozoa such as *Plasmodium* species, *Toxoplasma gondii*, *Leishmania* species, and other trypanosomatids. This is problematic as the molecular biology of protozoan pathogens is drastically different to that of model organisms, which includes a limited number of Metazoa, bacteria, and fungi that have been extensively studied (e.g., *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Escherichia coli*, and *Saccharomyces cerevisiae*). Nonetheless, it is helpful to use studies conducted on well-researched model organisms as a general guide for the study of non-model species, whilst retaining a certain degree of caution. In the human genome for example, SNPs occur on average once every 300 bases (~10 million SNPs), comprising approximately 0.1% of the entire genome (International HapMap Consortium, 2005; Jorde and Wooding, 2004; Reich et al., 2003). When comparing SNP frequencies across a diverse collection of taxa, expected rates are estimated at one SNP every 200–500 bases in non-coding DNA, and one SNP in every 500–1000 bases for coding DNA (Brumfield et al., 2003).

The frequency of polymorphisms in *P. falciparum* is approximately 1 in every 400-800 bases (Jeffares et al., 2007; Mu et al., 2007; Volkman et al., 2007), and 1 in 100 bases for *T. gondii* (Khan et al., 2006). Furthermore, the *P. falciparum* genome is extremely AT-rich,

consists of numerous repetitive sequences, and low complexity regions in protein sequences (Gardner et al., 2002; Pizzi and Frontali, 2001). As a result, sequencing and subsequent variant identification in protozoal genomes can be challenging, as these repetitive and low complexity sequences can hinder accurate read alignment and assembly (Battistuzzi et al., 2016; Talavera-Lopez and Andersson, 2017). Ribeiro *et al*. (2015) explored potential sources of false positive variants, and deemed the quality of the reference used as having the largest effect on the rate of false positive calls. This is a cause for concern regarding the use of these variant calling algorithms for non-model organisms, as their genomes are sometimes highly fragmented or in the very early stages of analysis and annotation. These draft-quality genomes may have been misassembled, or may be the result of inadequate or incomplete sequencing, poor quality control, or insufficient validation.

Protozoan parasites and the diseases they cause are a significant public health burden, causing in excess of a million deaths annually (Lozano et al., 2012). The highest contributors to this figure include malaria, leishmaniasis, and trypanosomiasis. The decreasing effectiveness of treatment options or vaccines, in concert with the increasing threat of drug resistance are applicable concerns for each of these diseases. As a result, SGS and TGS based genome-wide studies involving these organisms are becoming increasingly important, particularly those aiming to identify the genetic mechanisms of drug resistance and to monitor the expansion of resistance alleles in populations.

6.1. Parasites, ploidy, and pooled samples

Important considerations pertaining to clinically significant Protozoa and NGS sequencing and variant analysis, include ploidy, pooled samples, and mixed infections. While the majority of organisms such as plants, animals, and humans, are dominated by the diploid lifecycle stage, many eukaryotes, including Protozoa, alternate between different ploidy phases (Nuismer and

Otto, 2004). It has been reported that ploidy phases are a result of evolutionary selection, where diploidy is more likely to be favoured in a host species, compared to haploidy in a parasite species, based on host-parasite interactions. Challenges associated with detecting sequence variants in non-diploid parasite lifecycle stages however, include the ability to distinguish between sequencing errors and true variants that exist at a low frequency. This therefore has implications for the selection and implementation of both alignment and variant calling algorithms.

How accurately genomic variation can be identified and assigned to sub-genomes within a sample or individual, is dependent on experimental design, software selection and implementation, and the biological history or context of a species, including the lifecycle stage (Clevenger et al., 2015). Furthermore, while the presence of a unique set of haplotypes within an infection can be a direct measure of diversity, resolving these haplotypes is hindered by sequencing errors and *de novo* mutations in individual haplotypes (Trevino et al., 2017). While traditional Sanger sequencing is conducive to identifying major resistance alleles (i.e. those with $\geq$50% frequency), it is not sensitive enough to accurately detect minor alleles and mixed genotype infections (Talundzic et al., 2018). Advances in NGS and bioinformatics pipelines have addressed these shortcomings by offering a cost-effective, high-throughput alternative that requires a significantly reduced amount of template DNA, and the multiplexing of hundreds of samples and markers in one run. Consequently, many tools and protocols have been designed to handle such data generated from different organism lifecycle stages, and therefore ploidy levels, as well as those that accommodate pooled samples.

Both FreeBayes (Garrison and Marth, 2012) and the GATK (McKenna et al., 2010) allow users to specify the ploidy of the organism under investigation, without restricting this option to diploid or haploid. FreeBayes uses a Bayesian framework to assist with detecting multi-allelic haplotypes, and can also operate as a frequency-based pooled variant caller, as

opposed to describing variants and haplotypes in terms of genotypes (Garrison and Marth, 2012). Similarly, the GATK's HaplotypeCaller is able to both deal with non-diploid organisms or lifecycle phases, whether they be haploid or polyploid, as well as pooled samples (McKenna et al., 2010). The user can either use the '-ploidy' argument to specify the ploidy, or allow the tool to correctly predict the ploidy of a given sample at a given site. While the HaplotypeCaller can only process one ploidy phase at a time, the results from additional runs can later be combined, allowing multiple samples to be individually genotyped. The tool subsequently calls SNPs and indels via local reassembly of haplotypes.

Commonly in parasitology studies, samples may represent mixed infections or require pooling, where there is a need to not only detect rare or novel variants at low frequencies, but to also estimate allele frequencies from such pooled samples (Brockman et al., 2008). Allele frequencies can be accurately estimated through deep sequencing protocols of pooled populations, representing a rapid and economical method (Boitard et al., 2012). Pooling sequences from malaria infections however presents complications, including sample contamination with human DNA (Venkatesan et al., 2012), and the multiple potential origins of drug resistance mutations that can lead to soft sweeps, which in turn are difficult to detect (Nair et al., 2008; Nair et al., 2007). Cheeseman *et al*. (2015) for example described a two-tier approach for rare variant association testing of malaria parasites acquired directly from infections, by incorporating pooled Illumina sequencing and subsequent resequencing of limited parasite haplotypes. This method was able to accurately and robustly identify a known causal drug-resistance marker.

Initially, many variant calling tools were limited to a specific sequencing platform, read alignment algorithm, and/or single sample variant analysis. However, tools such as VarScan (Koboldt et al., 2009) are designed to detect sequence variants from a number of short read alignment algorithms, with high specificity and sensitivity, and across both individual and

pooled samples. Variant calling with VarScan is compatible with sequencing data generated from both Roche/454 sequencing of single samples, as well as deep sequencing of pooled samples from Illumina platforms. Furthermore, VarScan's documentation provides recommendations for input parameters and thresholds that are specific to each compatible alignment tool, which is especially appealing for users new to NGS analysis. Compensating for pooled data is a matter of selecting appropriate input parameters and thresholds such as read coverage and variant frequency, where you can for example specify a high read coverage threshold and a lower variant allele frequency to detect rare or novel variants. For variant calling across multiple samples, the 'mpileup' command from the SAMtools package is first run simultaneously for all input BAM files, where the output can be piped straight to VarScan for SNP and indel calling. Tools such as VarScan offer a powerful method for large-scale genetic variation studies for both individual and pooled samples, in concert with the high-throughput and massively parallel sequencing technologies offered by current sequencing platforms.

Malaria infections in endemic regions often exhibit multiple-genotype infections, consisting of mixtures of diverse parasite lineages (Anderson et al., 2000; Conway et al., 1991; Conway and McBride, 1991; Nkhoma et al., 2012). Such infections are thought to influence drug resistance (Hastings, 2006; Huijben et al., 2011), virulence evolution (Bell et al., 2006), and recombination rates (Conway et al., 1999), however they are poorly understood and challenging to address through traditional PCR genotyping and deep sequencing approaches. As a result, single-cell-sequencing (SCS) methods have been developed and exploited to elucidate the impact of such malaria infections, and isolate individual malaria haplotypes. Using a combination of cell sorting and WGA, Nair *et al.* (2014) produced high-quality material from red blood cells infected with *P. falciparum* and *Plasmodium vivax*, for sequencing on the Illumina HiSeq 2000 platform, and subsequent genotyping. Such an

675 approach is also valuable with respect to sampling, low parasitaemia, and culturing malaria
676 parasites. While some malaria species such as *P. falciparum* are culturable long term, this is
677 not feasible for other species such as *P. vivax*, where such alternative approaches can therefore
678 be helpful (Noulin et al., 2013). The data revealed the presence of within-host variation and
679 drug resistance haplotypes, where this SCS technique resulted in the accurate resolution of
680 single-cell genotypes from complex infections, which can be used in the future to obtain
681 parasite genome sequences directly from clinical blood samples.
682
683 *6.2. Population genetic studies of Toxoplasma gondii*
684 The global population genetic structure of *T. gondii* has been of major interest for decades,
685 with studies on the topic confirming the existence of at least four major clonal lineages (Khan
686 et al., 2011a). The within-lineage variation for three of the four major lineages occurring in the
687 Northern Hemisphere is <0.01%, whereas the between lineage variation ranges from
688 approximately 1-3% (Boyle et al., 2006). Based on genome-wide SNP comparisons of various
689 clonal-lineage strains, the ancestor of *T. gondii* type II crossed with ancestral strains
690 approximately 10,000 years ago to produce lineages I and III. Another study identified $\geq 10^6$
691 SNPs between ten *T. gondii* strains from Europe, North America, and South America, that
692 could potentially reveal strain-specific phenotypes (Minot et al., 2012). This SNP data was
693 used to identify shared haplotype blocks across the strains, and generate a haplotype map for
694 the species. Based on extensive SNP identification across various populations of *T. gondii*,
695 even a limited number of mating events can drastically modify the population structure of a
696 sexually reproducing pathogen and facilitate the emergence of new clonal genotypes (Boyle et
697 al., 2006). Characterisation of the almost non-existent polymorphisms within clonal lineages
698 revealed a history of infrequent yet important sexual recombination events followed by strong
699 selective sweeps, causing rapid clonal expansion within the species.

The cyst-forming apicomplexan parasite *Neospora caninum* causes hind limb paralysis in canines and abortion or stillbirth in cattle, and is closely related to *T. gondii* (Dubey et al., 1988). Calarco *et al*. (2018) generated RNA-seq data using Illumina HiSeq2000 paired-end sequencing, for two *N. caninum* isolates with distinct differences in pathogenicity in murine models. The implementation of a variant analysis pipeline using the sequencing data produced enabled the identification of over 3000 SNPs differentiating the two isolates. Numerous non-synonymous SNPs were present within protein-coding genes, and 19 SNP-dense regions were identified and found to be unevenly distributed along the *N. caninum* genome.

6.3. Sequencing and population genetics of Trypanosomatids

The leishmaniases includes several neglected tropical diseases caused by species of the genus *Leishmania*, where over 350 million people live at risk of these diseases globally (Alvar et al., 2012). *Leishmania* sp. are endemic in 98 countries, with an estimated 0.7-1 million new cases, and 20,000-30,000 *Leishamania*-associated deaths reported per annum. In 2005, the genome of the first *Leishmania* species, *L. major*, was sequenced using classical shotgun Sanger sequencing technology (Ivens et al., 2005). The advancement of sequencing technologies in subsequent years however, saw draft genomes being generated for an increasing number of *Leishmania* species, most of which took advantage of popular Illumina NGS platforms, which boasted the highest throughput and lowest sequencing costs per base (Leprohon et al., 2015). However, using short-read sequencing approaches presented challenges when attempting to handle highly repetitive DNA sequences and tandemly arranged identical genes, which are characteristic of *Leishmania* genomes (Alonso et al., 2016; Batra et al., 2019; Requena, 2011; Ubeda et al., 2014). As a result, TGS technologies are now being exploited to improve and re-sequence the draft genomes available for a number *Leishmania* species and strains.

For example, Gonzalez-de la Fuente *et al.* (2017) re-sequenced the *L. infantum* genome using a combined sequencing approach, taking advantage of long reads generated by PacBio sequencing, and short paired-end reads produced by Illumina technology. This study demonstrated the value of including PacBio reads when assembling a quality *Leishmania* genome, and the relevance of Illumina reads when joining contigs and extending chromosome ends. This *de novo* assembly was suggested to replace previous draft genomes, based on the resulting increased genome size, the identification of incorrectly assembled regions, and the numerous newly annotated or corrected genes presented. Similarly, Lypaczewski *et al*. (2018) published a complete reference genome assembly for *L. donovani*, after exploiting sequencing data from both SGS Illumina and TGS PacBio technologies. Previously, the *L. donovani* genome assembly contained 2,154 contigs, consisting of 7,969 protein coding genes, and an N50 value of 45,436, representing a measure of contiguity (Downing et al., 2011). The new assembly published by Lypaczewski *et al.* (2018) however, contained 36 contigs, 8,633 protein coding genes, and a 22-fold increase in N50. This study therefore improved on the quality of the previously published assembly by closing an estimated 2000 gaps across the 36 chromosomes, presenting new and re-annotated protein-coding genes and non-coding RNA genes, and extending multiple chromosomes. This approach also resulted in the correct assembly of highly repetitive *L. donovani* virulence gene clusters, and the accurate identification of SNPs and indels between distinct strains of the species, highlighting how complete, high-quality reference genome assemblies are vital for functional genomic studies.

It is through advances in -omics technologies that determinants of disease phenotype and drug efficacies are being investigated, to improve our knowledge of the pathogenesis of leishmaniasis and the drug resistance mechanisms employed. In 2011, a high-quality reference genome was generated using the combined SGS technologies of 454 Life Sciences and Illumina platforms for *L. donovani*, which is a major cause of the fatal visceral form of

749 leishmaniasis (VL) (Downing et al., 2011). This approach allowed errors within homopolymer
750 stretches produced by pyrosequencing, to be corrected using reads from Illumina's Genome
751 Analyser, in addition to resolving gaps and read errors in the assembly. The resulting high-
752 quality genome was used to study intra-species genetic diversity across 16 Nepalese and Indian
753 clinical isolates of *L. donovani*, possessing diverse drug susceptibility profiles. Read alignment
754 to the new reference genome provided important information on mechanisms of drug resistance
755 utilised by *L. donovani*, which were not apparent using traditional multilocus typing
756 approaches. Furthermore, the SNP diversity of these isolates when compared with other
757 *Leishmania* species, provided evidence that selection was acting on various surface- and
758 transport-related genes in this population of *L. donovani*, including several genes associated
759 with drug resistance.

760 The causative agent of Chagas disease (American trypanosomiasis) is *Trypanosoma*
761 *cruzi*, which affects over 8 million people per annum (Rassi et al., 2010). The first whole
762 genome sequence for *T. cruzi* was published in 2005, which was based on shotgun Sanger
763 sequencing technology (El-Sayed et al., 2005). While this draft genome was valuable at the
764 time, it was highly fragmented with a total of 4,098 contigs, most of which were less than 150
765 kb in length, and only 12 contigs exceeded 100 kb in size. Inherent complexities of
766 trypanosomatid genomes such as *Trypanosoma* and *Leishmania* species, include repetitive
767 sequences and tandemly arranged genes, which can now be tackled by exploiting the longer
768 reads generated by TGS technologies, to generate genome assemblies of higher quality than
769 their predecessors (Berna et al., 2018). As discussed in section 2.3, Berna *et al.* (2018)
770 assembled and annotated the genomes of two *T. cruzi* clones using PacBio sequencing
771 technology, improving on previous versions by resolving fragmented assemblies and repetitive
772 sequences. The final genome assemblies contained 1142 and 599 contigs, with improved N50
773 values of 265 and 318 kb. Using the assemblies obtained from PacBio SMRT sequencing

774 technology, novel repetitive sequences were revealed, and copy numbers of multi-gene
775 families and tandemly arrayed genes could be accurately calculated.

776 With respect to population genetics, a 2012 study used sequence data generated from
777 strains of *T. cruzi* belonging to various lineages, to facilitate the generation of a map of the
778 genetic diversity present within the species, and to highlight the polymorphic nature of the *T.*
779 *cruzi* genome (Ackermann et al., 2012). The study took advantage of the plethora of sequencing
780 data now available for the species to detect SNPs, including transcriptome data and genomes
781 generated using 454 Life Sciences' FLX Titanium platform. Focusing on protein coding
782 genomic regions, 97% of high-quality SNPs present across 47 loci were validated, where a set
783 of core, highly conserved genes were identified as being under purifying selection. There were
784 also a number of mutations that introduced or removed a stop codon, and tri-allelic and tetra-
785 allelic SNPs that could be utilised in strain typing assays.

786

787 *6.4. The importance of SNP detection in malaria causing Plasmodium falciparum*

788 The annual WHO World Malaria Report reported approximately 216 million malaria cases in
789 2017, and just under half a million deaths resulting from malaria. Consequently, malaria
790 research efforts generally focus on the mechanisms of *Plasmodium* sp. drug resistance,
791 potential vaccine targets, and vector control strategies. In 2002, the first draft genome for *P.*
792 *falciparum* was sequenced using Sanger shotgun sequencing technology (Gardner et al., 2002).
793 However, extensive efforts since then have been dedicated to resequencing *Plasmodium*
794 genomes using NGS approaches, to assist in tackling challenges associated with sequencing
795 the AT-rich genome of the malaria parasite, and to identify genes and loci associated with
796 clinical outcomes and drug resistance (Le Roch et al., 2012). Illumina sequencing technology
797 has been considered the most popular method for sequencing *Plasmodium* species, and a range

of techniques and combined approaches have been used to further improve and study these genomes (Bartfai et al., 2010; Kozarewa et al., 2009; Ponts et al., 2010).

While first and second generation sequencing technologies provide accuracy, massive parallelisation, and high-throughput, their availability and use in developing countries, especially in field hospitals, is not always feasible (Runtuwene et al., 2018). However, the development of TGS platforms has become increasingly attractive for sequencing *Plasmodium* genomes, especially for laboratory strains. Such samples can be useful *in-vitro* models for investigating parasite pathogenesis, and for clinically important species lacking available genetic information (Benavente et al., 2018; Bryant et al., 2018; Rutledge et al., 2017). For example, Runtuwene *et al.* (2018) applied ONT's portable MinION sequencing platform with PCR amplification, for genotyping laboratory adapted strains of *P. falciparum* and clinical samples containing the parasite. This study showed that the MinION device could generate long reads of acceptable quality, though at a sequencing accuracy of typically less than 90%. Since the average base-calling accuracy of the sequence was only 74.3%, it was suggested that a sequencing depth >50 greatly improved the accuracy of SNP calling.

A 2014 study (Preston et al., 2014) investigated the genetic variation in the mitochondria and apicoplast of 711 *P. falciparum* isolates from 14 countries. The study established a geographically informative, highly specific 23-SNP barcode, based on a high degree of linkage, where the linkage disequilibrium analyses inferred the co-transmission of each organellar genome and the non-recombining nature of the SNPs identified. There was also a higher proportion (77.8%) of non-synonymous mutations in SNPs within coding regions of the apicoplast compared to 61.8% on the nuclear genome and only 31.3% on the mitochondrial genome. This suggests that the organellar genomes are subject to different selective pressures, such that the conserved mitochondrial genes appear to be under purifying selection, whereas the apicoplast genes may instead be experiencing diversifying selection.

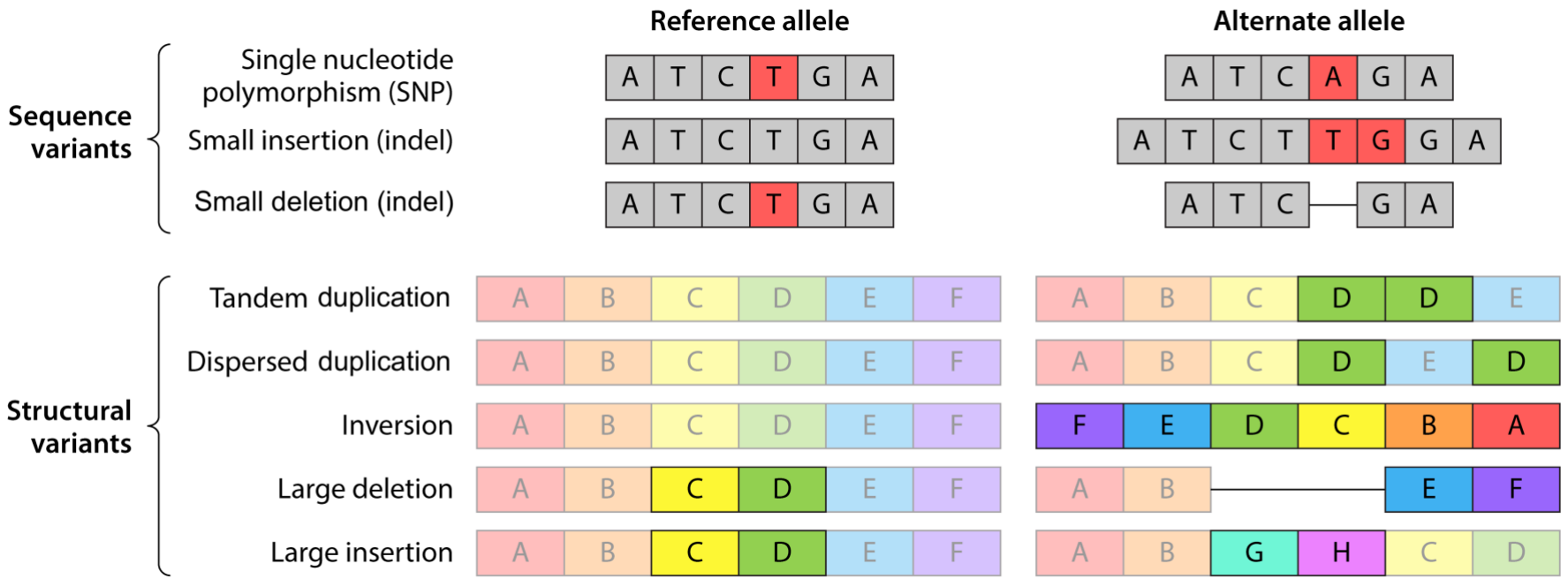
The emergence of chloroquine-resistant *P. falciparum* parasites is at least partially attributable to mutations in the molecular markers *pfmdr1*and *pfcrt* (Moers et al., 2015; Reed et al., 2000). To address the increased morbidity and mortality associated with malaria, as a result of selection of *pfmdr1*and *pfcrt* resistance alleles (Ashley et al., 2014; Nag et al., 2017), in the mid-1990s artemisinin-based combination therapies (ACTs) were introduced, and subsequently recommended by the WHO in 2005 as first-line treatments for *P. falciparum* malaria infections. However, progress made towards controlling and eradicating malaria worldwide by the availability and use of ACTs, is constantly under threat due to the geographical spread of artemisinin resistance. *Plasmodium falciparum* has experienced selective pressure due to the widespread and long term administration of numerous antimalarials including the abandoned drugs chloroquine and quinine, which are now obsolete for treating malaria (Nag et al., 2017). To prevent the recurrence of widespread resistance to ACTs, recent efforts have focused on identifying the *P. falciparum* genes, and specifically, the mutations in these genes that are indicators of resistance to ACTs and other antimalarials.

Mutations in the regions of the *pfcrt* gene encoding three transmembrane domains are responsible for chloroquine resistance (Cooper et al., 2007), while point mutations in both the *pfdhfr* and *pfdhps* genes are associated with resistance to sulfadoxine-pyrimethamine (Abdul-Ghani et al., 2013). Ariey *et al.* (2014) demonstrated an association between mutations in the Kelch 13 gene propeller domain and artemisinin resistance using whole-genome paired-end sequencing of clinical isolates from Cambodia, performed on an Illumina HiSeq platform. This study classified the polymorphic K13-propeller domain as a useful molecular marker for monitoring the emergence and expansion of artemisinin resistant *P. falciparum* across South East Asia. Numerous synonymous and non-synonymous mutations in the K13-propeller region related to the slow clearance of the parasites during treatment have now been identified (Ariey et al., 2014; Ashley et al., 2014; Takala-Harrison et al., 2015).

Key to monitoring the nature, development, and expansion of antimalarial drug resistance, is an understanding of the parasite's susceptibility to available drugs and the geographical origin and spread of resistance alleles. It is clear that the high-throughput capabilities, resolution, and scalability that SGS and TGS technologies offer, are conducive to developing tools that improve our knowledge on the mechanisms of drug resistance in malaria-causing parasites.

7. Identifying large structural variants using NGS data

In addition to the detection of polymorphisms located within functionally significant genes, larger variations such as SVs and copy number variation (CNVs) are routinely explored using SGS and TGS data (Figure 3). As expected, these large sequence variants affect phenotypic diversity within and between populations, and are implicated in a range of human diseases (Tattini et al., 2015). Structural variants are estimated to represent 1.2% of sequence variation in human genomes, compared to the existence of SNPs occupying only 0.1% of the genome (Pang et al., 2010). In protozoa, large SVs such as deletions and CNVs have been linked to clinically significant phenotypes, including drug resistance (Cowman et al., 1994; Downing et al., 2011; Papadopoulou et al., 1998), virulence (Khan et al., 2009), and changes in gene expression (Gonzales et al., 2008; Mackinnon et al., 2009). There are several *in-silico* tools available for detecting SVs using SGS and TGS data, each with their own unique inputs and output formats, underlying models, advantages and limitations. It has been suggested however, that it is not possible to identify the complete spectrum of SVs within genomes using a single tool, where a consensus approach is recommended (Lam et al., 2012; Mimori et al., 2013; Wong et al., 2010).



871

Figure 3. Visual summary of the types of sequence variants and large structural variants.

The sequence variants panel displays both single base changes (SNPs) and single insertions or deletions of bases (indels) spanning small nucleotide regions. These changes can either result in non-synonymous or synonymous mutations depending on whether an amino acid change or frame shift occurs in the corresponding protein coding sequence. The letters in the structural variants panel represent large spanning genomic segments or genes

In *T. gondii* differences in virulence observed between lineages I and III were attributed to expression differences in *ROP18* (Khan et al., 2009), a serine/threonine kinase secreted by rhoptries that phosphorylates host cell proteins (Taylor et al., 2006). This differential expression was traced to a large upstream DNA segment in the regulatory element of *ROP18* present only within the avirulent type III strain, which alters transcription of the gene. As this upstream region was also found to exist in the closely related parasite *N. caninum*, it was proposed that this segment of DNA was present in a common ancestor of all surviving *T. gondii* strains, though lost through a large DNA rearrangement in the more recently derived ancestor of the virulent lineages I and II. Additionally, strong evidence for positive selection was observed for *ROP18*, which possesses three atypically divergent alleles making it unusually polymorphic.

Previously, the detection of CNVs has exploited quantitative PCR (qPCR) methodologies, which are also imperfect (Beghain et al., 2016). However, the advancement of whole-genome sequencing technologies has facilitated more extensive analyses of such genomic variations, which subsequently requires the development of detection tools to respond to the availability of such data. Beghain *et al.* (2016) addressed the ability to detect CNVs from Cambodian *P. falciparum* isolates, using classical qPCR, compared to short paired-end reads from whole-genome sequencing, generated on the Illumina HiSeq platform. The algorithm PlasmoCNVScan was developed to better handle the unique nature of *Plasmodium* CNVs, which are not accommodated for by other available methods. Comparable results were observed between the two approaches taken in the study, demonstrating how such tools and sequencing technologies are conducive to studying the mechanisms of variations such as CNVs, to better understanding adapting parasite genomes.

Through the identification of SNPs and SVs using whole-genome sequencing of clinical *L. donovani* isolates generated from both Life Sciences and Illumina platforms,

Downing *et al.* (2011) detected genes with variable patterns of diversity in drug resistant samples, specifically associated with CNVs. Tests for selective pressures regarding the presence of SVs and SNPs, identified a set of protein-coding genes subject to adaptive evolution in this *L. donovani* population. While there was minimal SNP variation present, which is typically reflective of a homogenous genetic background, there were extensive SVs thought to be responsible for locus-specific changes in gene copy number, including whole chromosome CNVs and the generation of extrachromosomal fragments. Within the 17 strains studied, a pattern of ancient adaptive evolution was observable for six genes related to translation and RNA stability. This study also provided evidence of positive selection operating at loci encoding ribosomal components and RNA-binding proteins. This included SVs at two loci essential for translation, and thought to responsible for differences in gene expression between antimonial resistant and antimonial sensitive parasite lines.

8. Concluding Remarks

There is an increasing demand for robust tools that exploit SGS and TGS data. This includes tools that perform variant analysis, facilitating the identification of functionally significant sequence polymorphisms within and between populations. These polymorphisms include SNPs, indels, large SVs, and CNVs for which there are a plethora of *in-silico* tools available that are lacking in standardization, often varying drastically in their performance and outputs. While selecting the most appropriate SGS/TGS workflow and software settings for answering a specific research question may seem trivial, these decisions will often be crucial for accurate variant calling and any associated downstream investigations. Several additional challenges exist with respect to many protozoan pathogens of clinical significance, including the absence of high-quality reference genomes for many species, and the fact that much of the software developed to answer pertinent questions, has not been optimised on non-model organisms that

often possess drastically different molecular characteristics. Other considerations when generating and analysing sequencing data for pathogenic Protozoa include the unique, complex nature of their genomes, the presence of mixed infections, preparing and pooling samples, and ploidy phases. However, as the field continues to develop it is expected these challenges will be overcome, particularly as SGS and TGS technologies are becoming increasingly available, making it simpler to generate high-quality reference genomes. Until such a time as tools optimised for protozoan pathogens become available, parasitologists embarking on SGS and TGS related projects are encouraged to consider their choice of sequencing technology and analysis tools carefully. To this end, we hope this review assists others in preventing unnecessary downstream expenses by avoiding the generation of erroneous data, and the use procedures that may lead one towards inaccurate biological conclusions.

Acknowledgements

This review was completed by L. Calarco in partial fulfilment of the Ph.D. degree at UTS.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

References

Abdul-Ghani, R., Farag, H.F., Allam, A.F., 2013. Sulfadoxine-pyrimethamine resistance in Plasmodium falciparum: a zoomed image at the molecular level within a geographic context. Acta Trop 125, 163-190.
Abraham, G., Inouye, M., 2014. Fast principal component analysis of large-scale genome-wide data. PLoS One 9, e93766.
Ackermann, A.A., Panunzi, L.G., Cosentino, R.O., Sanchez, D.O., Aguero, F., 2012. A genomic scale map of genetic diversity in Trypanosoma cruzi. BMC Genomics 13, 736.
Acosta-Serrano, A., Almeida, I.C., Freitas-Junior, L.H., Yoshida, N., Schenkman, S., 2001. The mucin-like glycoprotein super-family of Trypanosoma cruzi: structure and biological roles. Mol Biochem Parasitol 114, 143-150.
Al-Qassab, S., Reichel, M.P., Ellis, J., 2010. A second generation multiplex PCR for typing strains of Neospora caninum using six DNA targets. Molecular and cellular probes 24, 20-26.
Alonso, G., Rastrojo, A., Lopez-Perez, S., Requena, J.M., Aguado, B., 2016. Resequencing and assembly of seven complex loci to improve the Leishmania major (Friedlin strain) reference genome. Parasit Vectors 9, 74.
Altmann, A., Weber, P., Bader, D., Preuss, M., Binder, E.B., Muller-Myhsok, B., 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131, 1541-1554.
Alvar, J., Velez, I.D., Bern, C., Herrero, M., Desjeux, P., Cano, J., Jannin, J., den Boer, M., Team, W.H.O.L.C., 2012. Leishmaniasis worldwide and global estimates of its incidence. PLoS One 7, e35671.
Ambardar, S., Gupta, R., Trakroo, D., Lal, R., Vakhlu, J., 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol 56, 394-404.
Anderson, T.J., Haubold, B., Williams, J.T., Estrada-Franco, J.G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., Whitworth, J., Velez, I.D., Brockman, A.H., Nosten, F., Ferreira, M.U., Day, K.P., 2000. Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum. Mol Biol Evol 17, 1467-1482.
Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., Lim, P., Leang, R., Duong, S., Sreng, S., Suon, S., Chuor, C.M., Bout, D.M., Menard, S., Rogers, W.O., Genton, B., Fandeur, T., Miotto, O., Ringwald, P., Le Bras, J., Berry, A., Barale, J.C., Fairhurst, R.M., Benoit-Vical, F., Mercereau-Puijalon, O., Menard, D., 2014. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature 505, 50-55.
Arner, E., Kindlund, E., Nilsson, D., Farzana, F., Ferella, M., Tammi, M.T., Andersson, B., 2007. Database of Trypanosoma cruzi repeated genes: 20,000 additional gene variants. BMC Genomics 8, 391.
Ashley, E.A., Dhorda, M., Fairhurst, R.M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J.M., Mao, S., Sam, B., Sopha, C., Chuor, C.M., Nguon, C., Sovannaroth, S., Pukrittayakamee, S., Jittamala, P., Chotivanich, K., Chutasmit, K., Suchatsoonthorn, C., Runcharoen, R., Hien, T.T., Thuy-Nhien, N.T., Thanh, N.V., Phu, N.H., Htut, Y., Han, K.T., Aye, K.H., Mokuolu, O.A., Olaosebikan, R.R., Folaranmi, O.O., Mayxay, M., Khanthavong, M., Hongvanthong, B., Newton, P.N., Onyamboko, M.A., Fanello, C.I., Tshefu, A.K., Mishra, N., Valecha, N., Phyo, A.P., Nosten, F., Yi, P., Tripura, R., Borrmann, S., Bashraheil, M., Peshu, J., Faiz, M.A., Ghose, A., Hossain, M.A., Samad, R., Rahman, M.R., Hasan, M.M., Islam, A., Miotto, O., Amato, R., MacInnis, B., Stalker, J., Kwiatkowski, D.P., Bozdech, Z., Jeeyapant, A., Cheah, P.Y., Sakulthaew, T., Chalk, J., Intharabut, B., Silamut,

K., Lee, S.J., Vihokhern, B., Kunasol, C., Imwong, M., Tarning, J., Taylor, W.J., Yeung, S., Woodrow, C.J., Flegg, J.A., Das, D., Smith, J., Venkatesan, M., Plowe, C.V., Stepniewska, K., Guerin, P.J., Dondorp, A.M., Day, N.P., White, N.J., Tracking Resistance to Artemisinin, C., 2014. Spread of artemisinin resistance in Plasmodium falciparum malaria. N Engl J Med 371, 411-423.
Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., O'Grady, J., 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 33, 296-300.
Aydemir, O., Janko, M., Hathaway, N.J., Verity, R., Mwandagalirwa, M.K., Tshefu, A.K., Tessema, S.K., Marsh, P.W., Tran, A., Reimonn, T., Ghani, A.C., Ghansah, A., Juliano, J.J., Greenhouse, B.R., Emch, M., Meshnick, S.R., Bailey, J.A., 2018. Drug-Resistance and Population Structure of Plasmodium falciparum Across the Democratic Republic of Congo Using High-Throughput Molecular Inversion Probes. J Infect Dis 218, 946-955.
Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W.A., Jiang, H., Feng, G., 2014. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer informatics 13, 67-82.
Bartfai, R., Hoeijmakers, W.A., Salcedo-Amaya, A.M., Smits, A.H., Janssen-Megens, E., Kaan, A., Treeck, M., Gilberger, T.W., Francoijs, K.J., Stunnenberg, H.G., 2010. H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. PLoS Pathog 6, e1001223.
Basso, W., Schares, S., Barwald, A., Herrmann, D.C., Conraths, F.J., Pantchev, N., Vrhovec, M.G., Schares, G., 2009. Molecular comparison of Neospora caninum oocyst isolates from naturally infected dogs with cell culture-derived tachyzoites of the same isolates using nested polymerase chain reaction to amplify microsatellite markers. Veterinary parasitology 160, 43-50.
Batra, D., Lin, W., Narayanan, V., Rowe, L.A., Sheth, M., Zheng, Y., Loparev, V., de Almeida, M., 2019. Draft Genome Sequences of Leishmania (Leishmania) amazonensis, Leishmania (Leishmania) mexicana, and Leishmania (Leishmania) aethiopica, Potential Etiological Agents of Diffuse Cutaneous Leishmaniasis. Microbiol Resour Announc 8.
Battistuzzi, F.U., Schneider, K.A., Spencer, M.K., Fisher, D., Chaudhry, S., Escalante, A.A., 2016. Profiles of low complexity regions in Apicomplexa. BMC Evol Biol 16, 47.
Beghain, J., Langlois, A.C., Legrand, E., Grange, L., Khim, N., Witkowski, B., Duru, V., Ma, L., Bouchier, C., Menard, D., Paul, R.E., Ariey, F., 2016. Plasmodium copy number variation scan: gene copy numbers evaluation in haploid genomes. Malar J 15, 206.
Bell, A.S., de Roode, J.C., Sim, D., Read, A.F., 2006. Within-host competition in genetically diverse malaria infections: parasite virulence and competitive success. Evolution 60, 1358-1371.
Benavente, E.D., de Sessions, P.F., Moon, R.W., Grainger, M., Holder, A.A., Blackman, M.J., Roper, C., Drakeley, C.J., Pain, A., Sutherland, C.J., Hibberd, M.L., Campino, S., Clark, T.G., 2018. A reference genome and methylome for the Plasmodium knowlesi A1-H.1 line. Int J Parasitol 48, 191-196.
Berna, L., Rodriguez, M., Chiribao, M.L., Parodi-Talice, A., Pita, S., Rijo, G., Alvarez-Valin, F., Robello, C., 2018. Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. Microb Genom 4.
Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B., Bohme, U., Hannick, L., Aslett, M.A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U.C., Arrowsmith, C., Atkin, R.J., Barron, A.J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T.J., Churcher, C., Clark, L.N., Corton, C.H., Cronin, A., Davies, R.M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M.C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B.R.,

Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A.X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P.J., Moule, S., Martin, D.M., Morgan, G.W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C.S., Peterson, J., Quail, M.A., Rabbinowitsch, E., Rajandream, M.A., Reitter, C., Salzberg, S.L., Sanders, M., Schobel, S., Sharp, S., Simmonds, M., Simpson, A.J., Tallon, L., Turner, C.M., Tait, A., Tivey, A.R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M.D., Embley, T.M., Gull, K., Ullu, E., Barry, J.D., Fairlamb, A.H., Opperdoes, F., Barrell, B.G., Donelson, J.E., Hall, N., Fraser, C.M., Melville, S.E., El-Sayed, N.M., 2005. The genome of the African trypanosome Trypanosoma brucei. Science 309, 416-422.

Boitard, S., Schlotterer, C., Nolte, V., Pandey, R.V., Futschik, A., 2012. Detecting selective sweeps from pooled next-generation sequencing samples. Mol Biol Evol 29, 2177-2186.

Boyle, J.P., Rajasekar, B., Saeij, J.P., Ajioka, J.W., Berriman, M., Paulsen, I., Roos, D.S., Sibley, L.D., White, M.W., Boothroyd, J.C., 2006. Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like Toxoplasma gondii. Proc Natl Acad Sci U S A 103, 10514-10519.

Braslavsky, I., Hebert, B., Kartalov, E., Quake, S.R., 2003. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A 100, 3960-3964.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., Jaffe, D.B., 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18, 763-770.

Brumfield, R.T., Beerli, P., Nickerson, D.A., Edwards, S.V., 2003. The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology & Evolution 18, 249-256.

Bruske, E., Otto, T.D., Frank, M., 2018. Whole genome sequencing and microsatellite analysis of the Plasmodium falciparum E5 NF54 strain show that the var, rifin and stevor gene families follow Mendelian inheritance. Malar J 17, 376.

Bryant, J.M., Baumgarten, S., Lorthiois, A., Scheidig-Benatar, C., Claes, A., Scherf, A., 2018. De Novo Genome Assembly of a Plasmodium falciparum NF54 Clone Using Single-Molecule Real-Time Sequencing. Genome Announc 6.

Buscaglia, C.A., Campo, V.A., Frasch, A.C., Di Noia, J.M., 2006. Trypanosoma cruzi surface mucins: host-dependent coat diversity. Nat Rev Microbiol 4, 229-236.

Calarco, L., Barratt, J., Ellis, J., 2018. Genome Wide Identification of Mutational Hotspots in the Apicomplexan Parasite Neospora caninum and the Implications for Virulence. Genome Biol Evol 10, 2417-2431.

Callejas-Hernandez, F., Rastrojo, A., Poveda, C., Girones, N., Fresno, M., 2018. Genomic assemblies of newly sequenced Trypanosoma cruzi strains reveal new genomic expansion and greater complexity. Sci Rep 8, 14631.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., McQuillan, J.A., 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28, 464-469.

Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., Yu, F., 2012. An integrative variant analysis suite for whole exome next-generation sequencing data. BMC bioinformatics 13, 8.

Cheeseman, I.H., McDew-White, M., Phyo, A.P., Sriprawat, K., Nosten, F., Anderson, T.J., 2015. Pooled sequencing and rare variant association tests for identifying the determinants of emerging drug resistance in malaria parasites. Mol Biol Evol 32, 1080-1090.

Chien, J.T., Pakala, S.B., Geraldo, J.A., Lapp, S.A., Humphrey, J.C., Barnwell, J.W., Kissinger, J.C., Galinski, M.R., 2016. High-Quality Genome Assembly and Annotation for

Plasmodium coatneyi, Generated Using Single-Molecule Real-Time PacBio Technology. Genome Announc 4.
Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G.R., Delledonne, M., Luo, C., Ecker, J.R., Cantu, D., Rank, D.R., Schatz, M.C., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13, 1050-1054.
Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92.
Clevenger, J., Chavarro, C., Pearl, S.A., Ozias-Akins, P., Jackson, S.A., 2015. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant 8, 831-846.
Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol 17, 13.
Consortium, G.P., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
Conway, D.J., Greenwood, B.M., McBride, J.S., 1991. The epidemiology of multiple-clone Plasmodium falciparum infections in Gambian patients. Parasitology 103 Pt 1, 1-6.
Conway, D.J., McBride, J.S., 1991. Population genetics of Plasmodium falciparum within a malaria hyperendemic area. Parasitology 103 Pt 1, 7-16.
Conway, D.J., Roper, C., Oduola, A.M., Arnot, D.E., Kremsner, P.G., Grobusch, M.P., Curtis, C.F., Greenwood, B.M., 1999. High recombination rate in natural populations of Plasmodium falciparum. Proc Natl Acad Sci U S A 96, 4506-4511.
Cooper, R.A., Lane, K.D., Deng, B., Mu, J., Patel, J.J., Wellems, T.E., Su, X., Ferdig, M.T., 2007. Mutations in transmembrane domains 1, 4 and 9 of the Plasmodium falciparum chloroquine resistance transporter alter susceptibility to chloroquine, quinine and quinidine. Mol Microbiol 63, 270-282.
Cowman, A.F., Galatis, D., Thompson, J.K., 1994. Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. Proc Natl Acad Sci U S A 91, 1143-1147.
Dai, M., Thompson, R.C., Maher, C., Contreras-Galindo, R., Kaplan, M.H., Markovitz, D.M., Omenn, G., Meng, F., 2010. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics 11 Suppl 4, S7.
Das, R., Dhiman, R.C., Savargaonkar, D., Anvikar, A.R., Valecha, N., 2016. Genotyping of Plasmodium vivax by minisatellite marker and its application in differentiating relapse and new infection. Malar J 15, 115.
Diaz-Viraque, F., Pita, S., Greif, G., de Souza, R.C.M., Iraola, G., Robello, C., 2019. Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite Trypanosoma cruzi. Genome Biol Evol 11, 1952-1957.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.
Downing, T., Imamura, H., Decuypere, S., Clark, T.G., Coombs, G.H., Cotton, J.A., Hilley, J.D., de Doncker, S., Maes, I., Mottram, J.C., Quail, M.A., Rijal, S., Sanders, M., Schonian, G., Stark, O., Sundar, S., Vanaerschot, M., Hertz-Fowler, C., Dujardin, J.C., Berriman, M., 2011. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides

insights into population structure and mechanisms of drug resistance. Genome Res 21, 2143-2156.
Dubey, J.P., Hattel, A.L., Lindsay, D.S., Topper, M.J., 1988. Neonatal Neospora caninum infection in dogs: isolation of the causative agent and experimental transmission. Journal of the American Veterinary Medical Association 193, 1259-1263.
Ebbert, M.T., Wadsworth, M.E., Staley, L.A., Hoyt, K.L., Pickett, B., Miller, J., Duce, J., Alzheimer's Disease Neuroimaging, I., Kauwe, J.S., Ridge, P.G., 2016. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC bioinformatics 17 Suppl 7, 239.
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133-138.
El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G., Westenberger, S.J., Caler, E., Cerqueira, G.C., Branche, C., Haas, B., Anupama, A., Arner, E., Aslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D.A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J.F., de Jong, P., Edwards, K., Englund, P.T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A.C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M.J., Lorenzi, H., Louie, T., Machado, C.R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J.C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J.L., Rinta, J., Robertson, L., Salzberg, S.L., Sanchez, D.O., Seyler, A., Sharma, R., Shetty, J., Simpson, A.J., Sisk, E., Tammi, M.T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P.N., Wickstead, B., Wortman, J., White, O., Fraser, C.M., Stuart, K.D., Andersson, B., 2005. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science 309, 409-415.
Escalante, A.A., Lal, A.A., Ayala, F.J., 1998. Genetic polymorphism and natural selection in the malaria parasite Plasmodium falciparum. Genetics 149, 189-202.
Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C., Wu, C.H., Xenarios, I., Yeh, L.S., Young, S.Y., Mitchell, A.L., 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic acids research 45, D190-D199.
Fiume, M., Williams, V., Brook, A., Brudno, M., 2010. Savant: genome browser for high-throughput sequencing data. Bioinformatics 26, 1938-1944.
Flaherty, B.R., Talundzic, E., Barratt, J., Kines, K.J., Olsen, C., Lane, M., Sheth, M., Bradbury, R.S., 2018. Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in blood via targeted amplicon deep sequencing. Microbiome 6, 164.
Franzen, O., Talavera-Lopez, C., Ochaya, S., Butler, C.E., Messenger, L.A., Lewis, M.D., Llewellyn, M.S., Marinkelle, C.J., Tyler, K.M., Miles, M.A., Andersson, B., 2012.

Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei. BMC Genomics 13, 531.
Frasch, A.C., 2000. Functional diversity in the trans-sialidase and mucin families in Trypanosoma cruzi. Parasitol Today 16, 282-286.
Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693-709.
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., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419, 498-511.
Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing, arXiv:1207.3907. 2012. 1207.3907.
Girgis, H.Z., 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC bioinformatics 16, 227.
Gonzales, J.M., Patel, J.J., Ponmee, N., Jiang, L., Tan, A., Maher, S.P., Wuchty, S., Rathod, P.K., Ferdig, M.T., 2008. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. PLoS Biol 6, e238.
Gonzalez-de la Fuente, S., Camacho, E., Peiro-Pastor, R., Rastrojo, A., Carrasco-Ramiro, F., Aguado, B., Requena, J.M., 2018. Complete and de novo assembly of the Leishmania braziliensis (M2904) genome. Mem Inst Oswaldo Cruz 114, e180438.
Gonzalez-de la Fuente, S., Peiro-Pastor, R., Rastrojo, A., Moreno, J., Carrasco-Ramiro, F., Requena, J.M., Aguado, B., 2017. Resequencing of the Leishmania infantum (strain JPCM5) genome and de novo assembly into 36 contigs. Sci Rep 7, 18050.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2014. Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. Bioinformatics 30, 2381-2383.
Goodswen, S.J., Kennedy, P.J., Ellis, J.T., 2018. A Gene-Based Positive Selection Detection Approach to Identify Vaccine Candidates Using Toxoplasma gondii as a Test Case Protozoan Pathogen. Front Genet 9, 332.
Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., Pierce, E.A., 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27, 2518-2528.
Grisard, E.C., Teixeira, S.M., de Almeida, L.G., Stoco, P.H., Gerber, A.L., Talavera-Lopez, C., Lima, O.C., Andersson, B., de Vasconcelos, A.T., 2014. Trypanosoma cruzi Clone Dm28c Draft Genome Sequence. Genome Announc 2.
Hanlee, J.P., 2012. Improving bioinformatic pipelines for exome variant calling. Genome medicine 4, 7.
Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., Xie, Z., 2008. Single-molecule DNA sequencing of a viral genome. Science 320, 106-109.
Hastings, I.M., 2006. Complex dynamics and stability of resistance to antimalarial drugs. Parasitology 132, 615-624.
Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M.I., Ogden, R., Limborg, M.T., Cariani, A., Maes, G.E., Diopere, E., Carvalho, G.R., Nielsen, E.E., 2011. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol Ecol Resour 11 Suppl 1, 123-136.

Hughes, M.K., Hughes, A.L., 1995. Natural selection on Plasmodium surface proteins. Mol Biochem Parasitol 71, 99-113.
Huijben, S., Sim, D.G., Nelson, W.A., Read, A.F., 2011. The fitness of drug-resistant malaria parasites in a rodent model: multiplicity of infection. J Evol Biol 24, 2410-2422.
Iantorno, S.A., Durrant, C., Khan, A., Sanders, M.J., Beverley, S.M., Warren, W.C., Berriman, M., Sacks, D.L., Cotton, J.A., Grigg, M.E., 2017. Gene Expression in Leishmania Is Regulated Predominantly by Gene Dosage. MBio 8.
Institute, B., 2019. Picard Toolkit. GitHub Repository. http://broadinstitute.github.io/picard/.
International HapMap, C., 2005. A haplotype map of the human genome. Nature 437, 1299-1320.
Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Leggett, R.M., Eccles, D.A., Zalunin, V., Urban, J.M., Piazza, P., Bowden, R.J., Paten, B., Mwaigwisya, S., Batty, E.M., Simpson, J.T., Snutch, T.P., Birney, E., Buck, D., Goodwin, S., Jansen, H.J., O'Grady, J., Olsen, H.E., Min, I.O.N.A., Reference, C., 2015. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Res 4, 1075.
Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S.M., Bianchettin, G., Borzym, K., Bothe, G., Bruschi, C.V., Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R.M., Cronin, A., Cruz, A.K., Davies, R.M., De Gaudenzi, J., Dobson, D.E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A.C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J.C., Muller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O'Neil, S., Pentony, M., Pohl, T.M., Price, C., Purnelle, B., Quail, M.A., Rabbinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J.C., Rutter, S., Saunders, D., Schafer, M., Schein, J., Schwartz, D.C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D.F., Blackwell, J.M., Stuart, K.D., Barrell, B., Myler, P.J., 2005. The genome of the kinetoplastid parasite, Leishmania major. Science 309, 436-442.
Jayaraman, S., Harris, C., Paxton, E., Donachie, A.M., Vaikkinen, H., McCulloch, R., Hall, J.P.J., Kenny, J., Lenzi, L., Hertz-Fowler, C., Cobbold, C., Reeve, R., Michoel, T., Morrison, L.J., 2019. Application of long read sequencing to determine expressed antigen diversity in Trypanosoma brucei infections. PLoS Negl Trop Dis 13, e0007262.
Jeffares, D.C., Pain, A., Berry, A., Cox, A.V., Stalker, J., Ingle, C.E., Thomas, A., Quail, M.A., Siebenthall, K., Uhlemann, A.C., Kyes, S., Krishna, S., Newbold, C., Dermitzakis, E.T., Berriman, M., 2007. Genome variation and evolution of the malaria parasite Plasmodium falciparum. Nature genetics 39, 120-125.
Jeffares, D.C., Tomiczek, B., Sojo, V., dos Reis, M., 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. Methods Mol Biol 1201, 65-90.
Jex, A.R., Littlewood, D.T., Gasser, R.B., 2010. Toward next-generation sequencing of mitochondrial genomes--focus on parasitic worms of animals and biotechnological implications. Biotechnol Adv 28, 151-159.
Jorde, L.B., Wooding, S.P., 2004. Genetic variation, classification and 'race'. Nature genetics 36, S28-33.
Kamau, E., Campino, S., Amenga-Etego, L., Drury, E., Ishengoma, D., Johnson, K., Mumba, D., Kekre, M., Yavo, W., Mead, D., Bouyou-Akotet, M., Apinjoh, T., Golassa, L., Randrianarivelojosia, M., Andagalu, B., Maiga-Ascofare, O., Amambua-Ngwa, A., Tindana,

P., Ghansah, A., MacInnis, B., Kwiatkowski, D., Djimde, A.A., 2015. K13-propeller polymorphisms in Plasmodium falciparum parasites from sub-Saharan Africa. J Infect Dis 211, 1352-1355.
Karlsson, S., Moen, T., Lien, S., Glover, K.A., Hindar, K., 2011. Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. Mol Ecol Resour 11 Suppl 1, 247-253.
Kchouk, M., Gilbrat, J., Elloumi, M., 2017. Generations of Sequencing Technologies: From First to Next Generation. Biology and Medicine 9.
Khan, A., Bohme, U., Kelly, K.A., Adlem, E., Brooks, K., Simmonds, M., Mungall, K., Quail, M.A., Arrowsmith, C., Chillingworth, T., Churcher, C., Harris, D., Collins, M., Fosker, N., Fraser, A., Hance, Z., Jagels, K., Moule, S., Murphy, L., O'Neil, S., Rajandream, M.A., Saunders, D., Seeger, K., Whitehead, S., Mayr, T., Xuan, X., Watanabe, J., Suzuki, Y., Wakaguri, H., Sugano, S., Sugimoto, C., Paulsen, I., Mackey, A.J., Roos, D.S., Hall, N., Berriman, M., Barrell, B., Sibley, L.D., Ajioka, J.W., 2006. Common inheritance of chromosome Ia associated with clonal expansion of Toxoplasma gondii. Genome Res 16, 1119-1125.
Khan, A., Dubey, J.P., Su, C., Ajioka, J.W., Rosenthal, B.M., Sibley, L.D., 2011a. Genetic analyses of atypical Toxoplasma gondii strains reveal a fourth clonal lineage in North America. Int J Parasitol 41, 645-655.
Khan, A., Miller, N., Roos, D.S., Dubey, J.P., Ajzenberg, D., Darde, M.L., Ajioka, J.W., Rosenthal, B., Sibley, L.D., 2011b. A monomorphic haplotype of chromosome Ia is associated with widespread success in clonal and nonclonal populations of Toxoplasma gondii. MBio 2, e00228-00211.
Khan, A., Taylor, S., Ajioka, J.W., Rosenthal, B.M., Sibley, L.D., 2009. Selection at a single locus leads to widespread expansion of Toxoplasma gondii lineages that are virulent in mice. PLoS Genet 5, e1000404.
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.
Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283-2285.
Korhonen, P.K., Young, N.D., Gasser, R.B., 2016. Making sense of genomes of parasitic worms: Tackling bioinformatic challenges. Biotechnol Adv 34, 663-686.
Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., Turner, D.J., 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6, 291-295.
Kwiatkowski, D., 2015. Malaria genomics: tracking a diverse and evolving parasite population. Int Health 7, 82-84.
Laehnemann, D., Borkhardt, A., McHardy, A.C., 2016. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform 17, 154-179.
Lam, H.Y., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M.B., Kidd, J.M., Bustamante, C.D., Snyder, M., 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol 30, 226-229.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.
Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., Studholme, D.J., 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 3, 1-8.
Le Roch, K.G., Chung, D.W., Ponts, N., 2012. Genomics and integrated systems biology in Plasmodium falciparum: a path to malaria control and eradication. Parasite Immunol 34, 50-60.
Lei, T., Wang, H., Liu, J., Nan, H., Liu, Q., 2014. ROP18 is a key factor responsible for virulence difference between Toxoplasma gondii and Neospora caninum. PLoS One 9, e99744.
Leprohon, P., Fernandez-Prada, C., Gazanion, E., Monte-Neto, R., Ouellette, M., 2015. Drug resistance analysis by next generation sequencing in Leishmania. Int J Parasitol Drugs Drug Resist 5, 26-35.
Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv. 1303. https://arxiv.org/abs/1303.3997.
Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760.
Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595.
Li, H., Handsake, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009a. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
Li, R., Li, Y., Kristiansen, K., Wang, J., 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713-714.
Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., Wang, J., 2009b. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.
Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012, 251364.
Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., Abraham, J., Adair, T., Aggarwal, R., Ahn, S.Y., Alvarado, M., Anderson, H.R., Anderson, L.M., Andrews, K.G., Atkinson, C., Baddour, L.M., Barker-Collo, S., Bartels, D.H., Bell, M.L., Benjamin, E.J., Bennett, D., Bhalla, K., Bikbov, B., Bin Abdulhak, A., Birbeck, G., Blyth, F., Bolliger, I., Boufous, S., Bucello, C., Burch, M., Burney, P., Carapetis, J., Chen, H., Chou, D., Chugh, S.S., Coffeng, L.E., Colan, S.D., Colquhoun, S., Colson, K.E., Condon, J., Connor, M.D., Cooper, L.T., Corriere, M., Cortinovis, M., de Vaccaro, K.C., Couser, W., Cowie, B.C., Criqui, M.H., Cross, M., Dabhadkar, K.C., Dahodwala, N., De Leo, D., Degenhardt, L., Delossantos, A., Denenberg, J., Des Jarlais, D.C., Dharmaratne, S.D., Dorsey, E.R., Driscoll, T., Duber, H., Ebel, B., Erwin, P.J., Espindola, P., Ezzati, M., Feigin, V., Flaxman, A.D., Forouzanfar, M.H., Fowkes, F.G., Franklin, R., Fransen, M., Freeman, M.K., Gabriel, S.E., Gakidou, E., Gaspari, F., Gillum, R.F., Gonzalez-Medina, D., Halasa, Y.A., Haring, D., Harrison, J.E., Havmoeller, R., Hay, R.J., Hoen, B., Hotez, P.J., Hoy, D., Jacobsen, K.H., James, S.L., Jasrasaria, R., Jayaraman, S., Johns, N., Karthikeyan, G., Kassebaum, N., Keren, A., Khoo, J.P., Knowlton, L.M., Kobusingye, O., Koranteng, A., Krishnamurthi, R., Lipnick, M., Lipshultz, S.E., Ohno, S.L., Mabweijano, J., MacIntyre, M.F., Mallinger, L., March, L., Marks, G.B., Marks, R., Matsumori, A., Matzopoulos, R., Mayosi, B.M., McAnulty, J.H., McDermott, M.M., McGrath, J., Mensah, G.A., Merriman, T.R., Michaud, C., Miller, M., Miller, T.R., Mock, C., Mocumbi, A.O., Mokdad, A.A., Moran, A., Mulholland, K., Nair, M.N., Naldi, L., Narayan, K.M., Nasseri, K., Norman, P., O'Donnell, M., Omer, S.B., Ortblad, K., Osborne, R., Ozgediz, D., Pahari, B., Pandian, J.D., Rivero, A.P., Padilla, R.P., Perez-Ruiz, F., Perico, N., Phillips, D., Pierce, K., Pope, C.A., 3rd, Porrini, E., Pourmalek, F., Raju, M., Ranganathan, D., Rehm, J.T., Rein, D.B., Remuzzi,

G., Rivara, F.P., Roberts, T., De Leon, F.R., Rosenfeld, L.C., Rushton, L., Sacco, R.L., Salomon, J.A., Sampson, U., Sanman, E., Schwebel, D.C., Segui-Gomez, M., Shepard, D.S., Singh, D., Singleton, J., Sliwa, K., Smith, E., Steer, A., Taylor, J.A., Thomas, B., Tleyjeh, I.M., Towbin, J.A., Truelsen, T., Undurraga, E.A., Venketasubramanian, N., Vijayakumar, L., Vos, T., Wagner, G.R., Wang, M., Wang, W., Watt, K., Weinstock, M.A., Weintraub, R., Wilkinson, J.D., Woolf, A.D., Wulf, S., Yeh, P.H., Yip, P., Zabetian, A., Zheng, Z.J., Lopez, A.D., Murray, C.J., AlMazroa, M.A., Memish, Z.A., 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380, 2095-2128.

Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14, 265-279.

Lypaczewski, P., Hoshizaki, J., Zhang, W.W., McCall, L.I., Torcivia-Rodriguez, J., Simonyan, V., Kaur, A., Dewar, K., Matlashewski, G., 2018. A complete Leishmania donovani reference genome identifies novel genetic variations associated with virulence. Sci Rep 8, 16549.

Ma, L., Liu, J., Li, M., Fu, Y., Zhang, X., Liu, Q., 2017. Rhoptry protein 5 (ROP5) Is a Key Virulence Factor in Neospora caninum. Front Microbiol 8, 370.

Mackinnon, M.J., Li, J., Mok, S., Kortok, M.M., Marsh, K., Preiser, P.R., Bozdech, Z., 2009. Comparative transcriptional and genomic analysis of Plasmodium falciparum field isolates. PLoS Pathog 5, e1000644.

Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O'Brien, J., Djimde, A., Doumbo, O., Zongo, I., Ouedraogo, J.B., Michon, P., Mueller, I., Siba, P., Nzila, A., Borrmann, S., Kiara, S.M., Marsh, K., Jiang, H., Su, X.Z., Amaratunga, C., Fairhurst, R., Socheat, D., Nosten, F., Imwong, M., White, N.J., Sanders, M., Anastasi, E., Alcock, D., Drury, E., Oyola, S., Quail, M.A., Turner, D.J., Ruano-Rubio, V., Jyothi, D., Amenga-Etego, L., Hubbart, C., Jeffreys, A., Rowlands, K., Sutherland, C., Roper, C., Mangano, V., Modiano, D., Tan, J.C., Ferdig, M.T., Amambua-Ngwa, A., Conway, D.J., Takala-Harrison, S., Plowe, C.V., Rayner, J.C., Rockett, K.A., Clark, T.G., Newbold, C.I., Berriman, M., MacInnis, B., Kwiatkowski, D.P., 2012. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature 487, 375-379.

Mardis, E.R., 2013. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) 6, 287-303.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.B., Donnelly, P., 2014. Choice of transcripts and software has a large effect on variant annotation. Genome medicine 6, 26.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652-654.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.

Mikheyev, A.S., Tin, M.M., 2014. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour 14, 1097-1102.

Mimori, T., Nariai, N., Kojima, K., Takahashi, M., Ono, A., Sato, Y., Yamaguchi-Kabata, Y., Nagasaki, M., 2013. iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. BMC Syst Biol 7 Suppl 6, S8.

Minot, S., Melo, M.B., Li, F., Lu, D., Niedelman, W., Levine, S.S., Saeij, J.P., 2012. Admixture and recombination among Toxoplasma gondii lineages explain global genome diversity. Proc Natl Acad Sci U S A 109, 13458-13463.

Moers, A.P., Hallett, R.L., Burrow, R., Schallig, H.D., Sutherland, C.J., van Amerongen, A., 2015. Detection of single-nucleotide polymorphisms in Plasmodium falciparum by PCR primer extension and lateral flow immunoassay. Antimicrob Agents Chemother 59, 365-371.
Mu, J., Awadalla, P., Duan, J., McGee, K.M., Keebler, J., Seydel, K., McVean, G.A., Su, X.Z., 2007. Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. Nature genetics 39, 126-130.
Nag, S., Dalgaard, M.D., Kofoed, P.E., Ursing, J., Crespo, M., Andersen, L.O., Aarestrup, F.M., Lund, O., Alifrangis, M., 2017. High throughput resistance profiling of Plasmodium falciparum infections based on custom dual indexing and Illumina next generation sequencing-technology. Sci Rep 7, 2398.
Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. Nat Rev Genet 14, 157-167.
Nair, S., Miller, B., Barends, M., Jaidee, A., Patel, J., Mayxay, M., Newton, P., Nosten, F., Ferdig, M.T., Anderson, T.J., 2008. Adaptive copy number evolution in malaria parasites. PLoS Genet 4, e1000243.
Nair, S., Nash, D., Sudimack, D., Jaidee, A., Barends, M., Uhlemann, A.C., Krishna, S., Nosten, F., Anderson, T.J., 2007. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol Biol Evol 24, 562-573.
Nair, S., Nkhoma, S.C., Serre, D., Zimmerman, P.A., Gorena, K., Daniel, B.J., Nosten, F., Anderson, T.J., Cheeseman, I.H., 2014. Single-cell genomics for dissection of complex malaria infections. Genome Res 24, 1028-1038.
Narlikar, L., Ovcharenko, I., 2009. Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic 8, 215-230.
Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12, 443-451.
Nkhoma, S.C., Nair, S., Cheeseman, I.H., Rohr-Allegrini, C., Singlam, S., Nosten, F., Anderson, T.J., 2012. Close kinship within multiple-genotype malaria parasite infections. Proc Biol Sci 279, 2589-2598.
Noulin, F., Borlon, C., Van Den Abbeele, J., D'Alessandro, U., Erhart, A., 2013. 1912-2012: a century of research on Plasmodium vivax in vitro culture. Trends Parasitol 29, 286-294.
Nuismer, S.L., Otto, S.P., 2004. Host-parasite interactions and the evolution of ploidy. Proc Natl Acad Sci U S A 101, 11036-11039.
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., Lyon, G.J., 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome medicine 5, 28.
Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., Morgan, M., 2014. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics 30, 2076-2078.
Oliver, G.R., Hart, S.N., Klee, E.W., 2015. Bioinformatics for clinical next generation sequencing. Clin Chem 61, 124-135.
Otto, T.D., Bohme, U., Sanders, M., Reid, A., Bruske, E.I., Duffy, C.W., Bull, P.C., Pearson, R.D., Abdi, A., Dimonte, S., Stewart, L.B., Campino, S., Kekre, M., Hamilton, W.L., Claessens, A., Volkman, S.K., Ndiaye, D., Amambua-Ngwa, A., Diakite, M., Fairhurst, R.M., Conway, D.J., Franck, M., Newbold, C.I., Berriman, M., 2018. Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. Wellcome Open Res 3, 52.
Oyola, S.O., Manske, M., Campino, S., Claessens, A., Hamilton, W.L., Kekre, M., Drury, E., Mead, D., Gu, Y., Miles, A., MacInnis, B., Newbold, C., Berriman, M., Kwiatkowski, D.P., 2014. Optimized whole-genome amplification strategy for extremely AT-biased template. DNA Res 21, 661-671.

Oyola, S.O., Otto, T.D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D.J., Macinnis, B., Kwiatkowski, D.P., Swerdlow, H.P., Quail, M.A., 2012. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics 13, 1.
Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z., 2014. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15, 256-278.
Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., Kirkness, E.F., Levy, S., Feuk, L., Scherer, S.W., 2010. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11, R52.
Papadopoulou, B., Kundig, C., Singh, A., Ouellette, M., 1998. Drug resistance in Leishmania: similarities and differences to other organisms. Drug Resist Updat 1, 266-278.
Park, M.H., Rhee, H., Park, J.H., Woo, H.M., Choi, B.O., Kim, B.Y., Chung, K.W., Cho, Y.B., Kim, H.J., Jung, J.W., Koo, S.K., 2014. Comprehensive analysis to improve the validation rate for single nucleotide variants detected by next-generation sequencing. PLoS One 9, e86664.
Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S.E., Lercher, M.J., 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31, 1929-1936.
Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M., 1999. Mining SNPs from EST databases. Genome Res 9, 167-174.
Piskol, R., Ramaswami, G., Li, J.B., 2013. Reliable identification of genomic variants from RNA-seq data. American journal of human genetics 93, 641-651.
Pita, S., Diaz-Viraque, F., Iraola, G., Robello, C., 2019. The Tritryps Comparative Repeatome: Insights on Repetitive Element Evolution in Trypanosomatid Pathogens. Genome Biol Evol 11, 546-551.
Pizzi, E., Frontali, C., 2001. Low-complexity regions in Plasmodium falciparum proteins. Genome Res 11, 218-229.
Ponts, N., Harris, E.Y., Prudhomme, J., Wick, I., Eckhardt-Ludka, C., Hicks, G.R., Hardiman, G., Lonardi, S., Le Roch, K.G., 2010. Nucleosome landscape and control of transcription in the human malaria parasite. Genome Res 20, 228-238.
Poplin, R., Chang, P.C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., Gross, S.S., Dorfman, L., McLean, C.Y., DePristo, M.A., 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983-987.
Preston, M.D., Campino, S., Assefa, S.A., Echeverry, D.F., Ocholla, H., Amambua-Ngwa, A., Stewart, L.B., Conway, D.J., Borrmann, S., Michon, P., Zongo, I., Ouedraogo, J.B., Djimde, A.A., Doumbo, O.K., Nosten, F., Pain, A., Bousema, T., Drakeley, C.J., Fairhurst, R.M., Sutherland, C.J., Roper, C., Clark, T.G., 2014. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun 5, 4052.
Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341.
Rassi, A., Jr., Rassi, A., Marin-Neto, J.A., 2010. Chagas disease. Lancet 375, 1388-1402.
Reed, M.B., Saliba, K.J., Caruana, S.R., Kirk, K., Cowman, A.F., 2000. Pgh1 modulates sensitivity and resistance to multiple antimalarials in Plasmodium falciparum. Nature 403, 906-909.

Reich, D.E., Gabriel, S.B., Altshuler, D., 2003. Quality and completeness of SNP databases. Nature genetics 33, 457-458.
Requena, J.M., 2011. Lights and shadows on gene organization and regulation of gene expression in Leishmania. Front Biosci (Landmark Ed) 16, 2069-2085.
Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B., Despierre, E., Vergote, I., Hilbush, B., Lambrechts, D., Del-Favero, J., 2012. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol 30, 61-68.
Ribeiro, A., Golicz, A., Hackett, C.A., Milne, I., Stephen, G., Marshall, D., Flavell, A.J., Bayer, M., 2015. An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. BMC bioinformatics 16, 382.
Roeber, F., Jex, A.R., Gasser, R.B., 2013. Advances in the diagnosis of key gastrointestinal nematode infections of livestock, with an emphasis on small ruminants. Biotechnol Adv 31, 1135-1152.
Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T., Bustillo, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348-352.
Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sanchez-Gracia, A., 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol 34, 3299-3302.
Runtuwene, L.R., Tuda, J.S.B., Mongan, A.E., Makalowski, W., Frith, M.C., Imwong, M., Srisutham, S., Nguyen Thi, L.A., Tuan, N.N., Eshita, Y., Maeda, R., Yamagishi, J., Suzuki, Y., 2018. Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum. Sci Rep 8, 8286.
Rutledge, G.G., Bohme, U., Sanders, M., Reid, A.J., Cotton, J.A., Maiga-Ascofare, O., Djimde, A.A., Apinjoh, T.O., Amenga-Etego, L., Manske, M., Barnwell, J.W., Renaud, F., Ollomo, B., Prugnolle, F., Anstey, N.M., Auburn, S., Price, R.N., McCarthy, J.S., Kwiatkowski, D.P., Newbold, C.I., Berriman, M., Otto, T.D., 2017. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature 542, 101-104.
Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S., Altshuler, D., International, S.N.P.M.W.G., 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928-933.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-425.
Sandmann, S., Karimi, M., de Graaf, A.O., Rohde, C., Gollner, S., Varghese, J., Ernsting, J., Walldin, G., van der Reijden, B.A., Muller-Tidow, C., Malcovati, L., Hellstrom-Lindberg, E., Jansen, J.H., Dugas, M., 2018. appreci8: a pipeline for precise variant calling integrating 8 tools. Bioinformatics 34, 4205-4212.

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum Mol Genet 19, R227-240.
Schatz, M.C., Delcher, A.L., Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. Genome Res 20, 1165-1173.
Seeman, T., 2015. Snippy: Fast Bacterial Variant Calling from NGS reads.
Sherry, S.T., Ward, M., Sirotkin, K., 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9, 677-679.
Su, C., Khan, A., Zhou, P., Majumdar, D., Ajzenberg, D., Darde, M.L., Zhu, X.Q., Ajioka, J.W., Rosenthal, B.M., Dubey, J.P., Sibley, L.D., 2012. Globally diverse Toxoplasma gondii isolates comprise six major clades originating from a small number of distinct ancestral lineages. Proc Natl Acad Sci U S A 109, 5844-5849.
Su, X.Z., Heatwole, V.M., Wertheimer, S.P., Guinet, F., Herrfeldt, J.A., Peterson, D.S., Ravetch, J.A., Wellems, T.E., 1995. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell 82, 89-100.
Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-595.
Takala-Harrison, S., Jacob, C.G., Arze, C., Cummings, M.P., Silva, J.C., Dondorp, A.M., Fukuda, M.M., Hien, T.T., Mayxay, M., Noedl, H., Nosten, F., Kyaw, M.P., Nhien, N.T., Imwong, M., Bethell, D., Se, Y., Lon, C., Tyner, S.D., Saunders, D.L., Ariey, F., Mercereau-Puijalon, O., Menard, D., Newton, P.N., Khanthavong, M., Hongvanthong, B., Starzengruber, P., Fuehrer, H.P., Swoboda, P., Khan, W.A., Phyo, A.P., Nyunt, M.M., Nyunt, M.H., Brown, T.S., Adams, M., Pepin, C.S., Bailey, J., Tan, J.C., Ferdig, M.T., Clark, T.G., Miotto, O., MacInnis, B., Kwiatkowski, D.P., White, N.J., Ringwald, P., Plowe, C.V., 2015. Independent emergence of artemisinin resistance mutations among Plasmodium falciparum in Southeast Asia. J Infect Dis 211, 670-679.
Talavera-Lopez, C., Andersson, B., 2017. Parasite genomics-Time to think bigger. PLoS Negl Trop Dis 11, e0005463.
Talundzic, E., Ravishankar, S., Kelley, J., Patel, D., Plucinski, M., Schmedes, S., Ljolje, D., Clemons, B., Madison-Antenucci, S., Arguin, P.M., Lucchi, N.W., Vannberg, F., Udhayakumar, V., 2018. Next-Generation Sequencing and Bioinformatics Protocol for Malaria Drug Resistance Marker Surveillance. Antimicrob Agents Chemother 62.
Tattini, L., D'Aurizio, R., Magi, A., 2015. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol 3, 92.
Taylor, S., Barragan, A., Su, C., Fux, B., Fentress, S.J., Tang, K., Beatty, W.L., Hajj, H.E., Jerome, M., Behnke, M.S., White, M., Wootton, J.C., Sibley, L.D., 2006. A secreted serine-threonine kinase determines virulence in the eukaryotic pathogen Toxoplasma gondii. Science 314, 1776-1780.
The UniProt, C., 2017. UniProt: the universal protein knowledgebase. Nucleic acids research 45, D158-D169.
Thiltgen, G., Dos Reis, M., Goldstein, R.A., 2017. Finding Direction in the Search for Selection. J Mol Evol 84, 39-50.
Thorvaldsdottir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14, 178-192.
Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111.
Trevino, S.G., Nkhoma, S.C., Nair, S., Daniel, B.J., Moncada, K., Khoswe, S., Banda, R.L., Nosten, F., Cheeseman, I.H., 2017. High-Resolution Single-Cell Sequencing of Malaria Parasites. Genome Biol Evol 9, 3373-3383.

Tripathi, R., Sharma, P., Chakraborty, P., Varadwaj, P.K., 2016. Next-generation sequencing revolution through big data analytics. Frontiers in Life Science 9, 119-149.
Ubeda, J.M., Raymond, F., Mukherjee, A., Plourde, M., Gingras, H., Roy, G., Lapointe, A., Leprohon, P., Papadopoulou, B., Corbeil, J., Ouellette, M., 2014. Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite Leishmania. PLoS Biol 12, e1001868.
Vembar, S.S., Seetin, M., Lambert, C., Nattestad, M., Schatz, M.C., Baybayan, P., Scherf, A., Smith, M.L., 2016. Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. DNA Res 23, 339-351.
Venkatesan, M., Amaratunga, C., Campino, S., Auburn, S., Koch, O., Lim, P., Uk, S., Socheat, D., Kwiatkowski, D.P., Fairhurst, R.M., Plowe, C.V., 2012. Using CF11 cellulose columns to inexpensively and effectively remove human DNA from Plasmodium falciparum-infected whole blood samples. Malar J 11, 41.
Vera, M., Alvarez-Dios, J.A., Fernandez, C., Bouza, C., Vilas, R., Martinez, P., 2013. Development and Validation of Single Nucleotide Polymorphisms (SNPs) Markers from Two Transcriptome 454-Runs of Turbot (Scophthalmus maximus) Using High-Throughput Genotyping. Int J Mol Sci 14, 5694-5711.
Volkman, S.K., Sabeti, P.C., DeCaprio, D., Neafsey, D.E., Schaffner, S.F., Milner, D.A., Jr., Daily, J.P., Sarr, O., Ndiaye, D., Ndir, O., Mboup, S., Duraisingh, M.T., Lukens, A., Derr, A., Stange-Thomann, N., Waggoner, S., Onofrio, R., Ziaugra, L., Mauceli, E., Gnerre, S., Jaffe, D.B., Zainoun, J., Wiegand, R.C., Birren, B.W., Hartl, D.L., Galagan, J.E., Lander, E.S., Wirth, D.F., 2007. A genome-wide map of diversity in Plasmodium falciparum. Nature genetics 39, 113-119.
Vyas, G., Tiwari, T., Mehta, A., Patel, M., Gupta, H., Ghosh, A., Surendra, K.C., 2016. Evaluation of next generation sequencing platforms for whole exome variant analysis. Clin Med Biochemistry 2.
Wang, K., Li, M., Hakonarson, H., 2010a. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164.
Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., MacLeod, J.N., Chiang, D.Y., Prins, J.F., Liu, J., 2010b. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research 38, e178.
Wong, K., Keane, T.M., Stalker, J., Adams, D.J., 2010. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol 11, R128.
Xu, C., 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 16, 15-24.
Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13, 555-556.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24, 1586-1591.