Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

The impact of genomics on population genetics of parasitic diseases

Daniel N Hupalo, Martina Bradic and Jane M Carlton

Parasites, defined as eukaryotic microbes and parasitic worms that cause global diseases of human and veterinary importance, span many lineages in the eukaryotic Tree of Life. Historically challenging to study due to their complicated life-cycles and association with impoverished settings, their inherent complexities are now being elucidated by genome sequencing. Over the course of the last decade, projects in large sequencing centers, and increasingly frequently in individual research labs, have sequenced dozens of parasite reference genomes and field isolates from patient populations. This 'tsunami' of genomic data is answering questions about parasite genetic diversity, signatures of evolution orchestrated through anti-parasitic drug and host immune pressure, and the characteristics of populations. This brief review focuses on the state of the art of parasitic protist genomics, how the peculiar genomes of parasites are driving creative methods for their sequencing, and the impact that next-generation sequencing is having on our understanding of parasite population genomics and control of the diseases they cause.

Addresses
Center for Genomics and Systems Biology, Department of Biology, New York University, 12 Waverly Place, New York, NY 10003, United States

Corresponding author: Carlton, Jane M (jane.carlton@nyu.edu)

Current Opinion in Microbiology 2015, **23**:49–54

This review comes from a themed issue on **Genomics**

Edited by **Neil Hall** and **Jay Hinton**

http://dx.doi.org/10.1016/j.mib.2014.11.001

The state of parasite whole genome sequencing

As of August 2014, sixty-five reference genomes of parasitic protists and their close relatives have been deposited in GenBank (Figure 1). The majority of these genomes fall within two phyla: first, the Kinetoplastidae, which contains parasites responsible for diseases ranging from African sleeping sickness (*Trypanosoma brucei*) and Chagas disease (*Trypanosoma cruzi*) to visceral leishmaniasis (*Leishmania* spp.); and second, the Apicomplexa, including the genus *Plasmodium*, whose transmission by the *Anopheles* mosquito causes malaria in more than 100 countries. Accordingly, research utilizing 'comparative genomics' of parasite genomes has been focused within the *Plasmodium* [1], *Leishmania* [2] and *Trypanosoma* [3] clades. The majority of these genomes were sequenced using first-generation Sanger technology, and some have taken many years to complete assembly, gene finding and annotation [4]. More recently, the advent of cheaper, faster and more accurate next-generation sequencing (NGS) platforms such as those provided by Illumina (e.g. HiSeq series), Roche 454 (e.g. GS Junior), and Life Technologies (e.g. Ion Torrent Personal Genome Machine), has enabled whole genome sequencing of multiple field isolates from patients. These unassembled genomes are deposited in GenBank's Sequence Read Archive or the European Bioinformatics Institute European Nucleotide Archive (http://www.ncbi.nlm.nih.gov/sra and http://www.ebi.ac.uk/ena, respectively) as well as in organism-specific databases such as those hosted by EuPathDB [5]. This new wave of parasite genome sequence data is revolutionizing the study of population genetics of parasites in two major ways: first, by generating preliminary descriptions of the population genetics of commonly used lab strains, and second, by enabling 'real-time' population genetics of patient field isolates. Examples of these are given below.

The challenges to genomics posed by parasite biology

Parasite genomes have highly diverse architectures. They vary in properties such as nucleotide bias, for example the extremely AT-rich *Plasmodium falciparum* genome [6], or the 'isochore' structure of *Plasmodium vivax* chromosomes that have GC-rich cores but AT-rich subtelomeric regions [7]. Many genomes are highly repetitive or replete with transposable elements, for example the amoebic dysentery-causing parasite *Entamoeba histolytica* [8]. Genome sizes of parasites also vary widely. The first eukaryotic parasite genome to be published, from the microsporidian *Encephalitozoon cuniculi*, was found to be 2.3 Mb [9], whereas the sexually transmitted parasite *Trichomonas vaginalis* has a ~160 Mb genome that has undergone recent genome expansion [10].

Such diversity poses unique challenges to whole genome sequencing, including attaining adequate genome coverage, identifying polymorphisms, and obtaining reliable estimates of population genetic parameters. These challenges have fostered new sequencing strategies for sampling patient isolates, such as the 'reduced representation' methods [11] that are being used to develop

Glossary

Admixture: interbreeding of individuals issued from two or more distinct populations or species

Coalescent theory: a theory describing the genealogy of chromosomes or genes. The genealogy is constructed backwards-in-time, starting with the present-day sample. Lineages coalesce until the most recent common ancestor (MRCA) of the sample is reached

F_{ST}: the mean fixation index is a measure of population differentiation due to genetic structure

Linkage disequilibrium: when a genotype present at one locus is dependent on the genotype at a second locus. LD decays each generation at a rate determined by the degree of recombination.

Population genetics: the study of the interrelated patterns of phenotypic, genotypic and allelic frequencies within populations, and how these frequencies change due to influences like selection and chance.

Selective sweep: the fixation of an advantageous mutation that reduces levels of linked silent polymorphism. The size of the chromosomal region impacted by a selective sweep is determined by the level of local recombination.

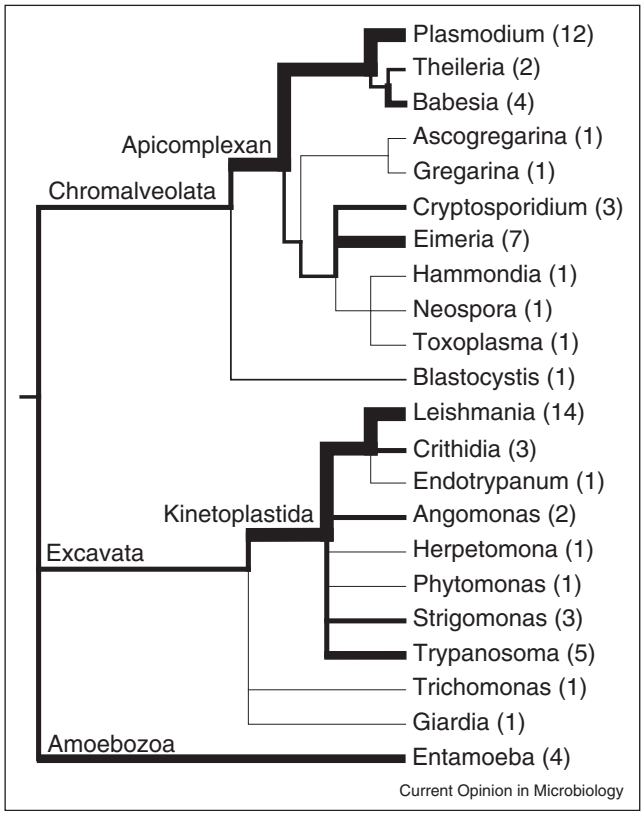
Tajima's *D*: a test that distinguishes between an allele evolving neutrally, that is, randomly, and one evolving under a non-random process such as positive selection, balancing selection, or selective sweep.

genetic markers for population genomic surveys. One such method, 'restriction-site associated DNA sequencing' uses either one (RAD) or a pair (ddRAD) of restriction enzymes in combination with partial sequencing [12,13•], and has been employed to resequence ~180 *T. vaginalis* genomes ([14]; M Bradic, New York University, unpublished data). A second new technology adopted by the parasite genomics community is 'hybrid selection', which uses biotinylated RNA baits designed from a parasite reference genome sequence to capture parasite DNA from a host-parasite DNA mixture [15]. Starting from exceedingly small quantities of patient material, *Plasmodium* DNA has been purified and enriched up to 40-fold [16,17•] — a key achievement in our ability to undertake population genetic surveys of parasites that cannot be grown in culture or are grossly contaminated with host genetic material.

Using genome sequence data to investigate parasite population genetics

Prior to the era of fast and cheap NGS, population studies of parasites were limited to a few genetic loci because of the lack of parasite genome sequence data. These initial studies using small numbers of microsatellite (MS) markers across chromosomes and single nucleotide polymorphisms (SNPs) in single-copy genes provided a preliminary glimpse of the genetic diversity, local and global population structure, and gene flow within and between populations of several different parasite species (see for example [18–20]), and identified loci suitable for classifying patient isolates [21,22]. Such genotyping has also been invaluable in epidemiology studies and disease classification [23,24]. Single-locus studies have also been used to identify mutations associated with parasite phenotypes such as virulence and drug resistance (reviewed for species of the malaria parasite in [25]). More recently,

Figure 1



A cartoon phylogenetic tree of parasite genera (and some closely-related free-living relatives) with reference genomes, with branch widths weighted according to the number of genomes (indicated in parentheses) in GenBank as of August 2014. Monophyletic supergroup and phylum name are labeled above their branches.

whole genome sequence data have enabled a genome-wide approach to population studies of commonly used parasite lab-adapted strains, and also of natural isolates taken directly from patients. In many instances, this has improved initial estimates of important population genetic parameters (see **Glossary**). We concentrate here on those parasites for which NGS data are now available that illustrate some of the impact that NGS data are having on population genetic studies of these parasites.

Recombination is an important population genetic parameter to consider in parasites, since the ability of a species to undergo sexual recombination directly impacts the spread of important genes through populations, such as those involved in virulence or drug resistance. Population genetic studies of several parasite species have indicated that genetic exchange is likely to occur or has taken place evolutionarily recently in the species (see for example in *T. vaginalis* [18], *Giardia* [26] and other parasite species reviewed in [27]). In the enteric pathogen *E. histolytica*, analysis of the first reference genome revealed a complement of genes necessary for