Network Models for Malaria: Antigens, Dynamics, and Evolution Over Space and Time

Lauren M Childs, Department of Mathematics, Virginia Tech, Blacksburg, VA, United States
Daniel B Larremore, Department of Computer Science, University of Colorado Boulder, Boulder, CO, United States; BioFrontiers Institute, University of Colorado Boulder, Boulder, CO, United States

Introduction

The world is full of networks of various sorts. We live in social networks, travel on transportation networks, stay in touch via communication networks, and stay alive through gene regulatory networks. While these networks span completely different domains, they have a few things in common. First, they all represent relationships between pairs of objects—people, locations, genes, or devices. Network scientists call these objects *nodes* or *vertices*. Second, these pairwise relationships between nodes can be clearly defined for each network. Network scientists call these *links* or *edges*. Third, these example networks are all *complex*, meaning that their structure is arbitrary and not regular—they are not rings, simple lattices, or regular trees. Taken together, a complex network of nodes, connected pairwise by links, is an incredibly flexible and powerful way to manage and analyze complicated data.

The *Plasmodium* parasites that cause malaria provide no shortage of complicated processes and relationships to examine using networks, but in this Review we focus on three areas in particular. First, we examine the use of networks to understand the structure and evolution of highly polymorphic genes, including those in the *var* family. We then review networks used for analyses at a broader scale for applications in *var* family genomics, haplotype phylogeography, and epidemiology. Finally, we discuss the uses of networks for modeling dynamics, including antigenic variation, metabolism, and protein interaction. Together, these sections illustrate the vast diversity of applications that networks have found in the study of *Plasmodium* species.

The three application areas are also interesting because networks play a different role in each (Fig. 1). In the first, nodes are genes or constituent blocks within genes and links are shared sequence content, making networks a representation of *shared sequence structure*. In the second, nodes are entire parasite genomes, isolates, or spatial regions and links are overlapping genomic content, making networks a representation of *genomic evolution and spatial spread*. In the third, nodes represent expression and links are drawn when one gene or protein typically leads to, or is correlated with, another, making networks a representation of a *dynamical system*.

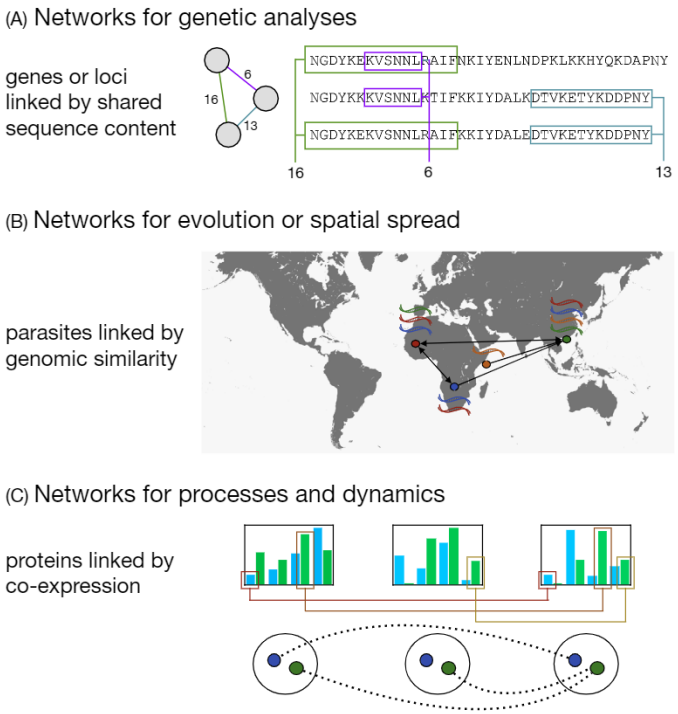

Fig. 1 Overview of three application areas with canonical examples. (A) In "var Genes and Repertoires: Shared Sequence Structure," nodes represent genes or constituent blocks and links are shared amino acid sequences or co-occurrence. Thus, patterns in networks represent the structural characteristics of genetic sequences. (B) In "Genomics and Phylogeography," nodes represent parasite genomes or isolates and links represent similarities or shared genes. Thus, networks represent genomic evolution or spatial spread. (C) In "Dynamics and Regulation," nodes represent gene or protein expression and links show correlations among them. Networks therefore represent dynamical processes. Images in (B) and (C) do not represent real data and are meant for explanatory purposes. Image in (A) reproduced with permission of D. Larremore: Larremore DB, Clauset A, and Buckee CO (2013) A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Computational Biology* 9(10): e1003268; underlying map in (B) reproduced from FreeVectorMaps.com.

The flexibility of networks to represent structure, spread, and dynamics (among many other things) under a unified vocabulary and set of analysis tools is a primary reason for networks' growing usage to better understand malaria.

Network-based methods are used widely across malaria research, spanning applications in genetics, epidemiology, model selection, ecology, and more. In this way, they are similar to other analysis methods, including phylogenetics and regression, and so as a consequence, the applications covered in the following sections are similarly broad. Nevertheless, a majority of our focus will be on analyses of the *var* gene family, which imported network analysis methods earlier and more extensively than other applications. Both the breadth of applications and the depth of *var* gene analyses featured here highlight the flexibility of networks, and their future use in analysis of malaria-related problems is guaranteed.

The rest of this Review is structured as follows. We begin with some background on key network concepts and vocabulary in "Network Basics." We then review the three primary applications described above, covering genetic structure and evolution in "var Genes and Repertoires: Shared Sequence Structure," genomics and phylogeography in "Genomics and Phylogeography," and dynamics and regulation in "Dynamics and Regulation." Finally, we highlight some additional areas in which networks have been used in "Other Applications of Networks" before conclusions and outlook in "Conclusion."

Network Basics

This section is written for the reader who is not yet deeply familiar with the vocabulary and concepts of complex networks. While this introduction is by no means meant to be exhaustive, it does cover many of the key concepts that are used in the later descriptions of network models of malaria's antigens, dynamics, and evolution. Readers who are looking to skip directly to malaria applications should turn to "var Genes and Repertoires: Shared Sequence Structure."

A network is a representation of the set of pairwise relationships among a set of objects. For instance, a social network might represent the friendships between pairs of people in a social group, an air transportation network might represent the direct flights available between pairs of airports, and a gene regulatory network might represent the influence that genes have on each other's expression. Two people are connected in the social network if they are friends; two cities are connected in the air transportation network when there is a direct flight between them; two genes are connected in the regulatory network when the expression of one

directly regulates the expression of the other. Depending on the context, the connections in a network may be called *links*, *edges*, or *ties*, and the things they connect may be called *nodes*, *vertices*, or *actors*.

Links in a network can represent different types of relationships, even when the nodes are the same. For instance, among a set of genes, we might draw a link between two genes when they are expressed together in one network, while in another network we might draw a link when those genes are located on the same chromosome. In any network analysis, it is therefore critical to be clear about precisely what the nodes are, and what the links between them mean.

Categorizing and Representing Networks

There are two broad ways of categorizing links across all network types which illustrate the diversity of relationships that can be represented. First, some links are *directed*, meaning that a link from node x to node y is distinct from a link from node y to node x, while others are *undirected*, simply indicating that nodes x and y are connected. Using our previous examples, friendships in social networks are typically represented using undirected links, while flights from one airport to another are typically represented using directed links. Second, some links are *weighted*, meaning that the link itself has a number attached that represents something quantitative about the relationship. Others are *unweighted*, and represent the mere existence of a relationship. The air travel network could be cast as either a weighted network in which each link represents the number of passengers per day who fly from one airport directly to another, or as an unweighted network in which each link represents the existence of a direct flight between airports. When a network has links of a certain type, we apply the same descriptors to the whole network, saying, for instance when we track how many passengers transit between cities, that the air transportation network is a directed and weighted network.

Networks are most commonly represented using an *adjacency matrix*, a square matrix A in which each row/column corresponds to a node, so that the (i, j) entry of the matrix stores both the presence and weight of any links between node i and node j. For an unweighted network, $A_{ij} = 1$ when i and j are linked, and $A_{ij} = 0$ if they are not. For a weighted network, A_{ij} is equal to the value of the weight from i to j. Note that for undirected networks, $A_{ij} = A_{ji}$, making its adjacency matrix symmetric. Directed networks typically have asymmetric adjacency matrices. Formulating a network as a matrix unlocks all the tools of linear algebra, which is one reason that networks have been able to become so widely used in analyses across fields.

Centrality

One of the most common questions that one can ask of a network is, "Which nodes are most important?" Generally, quantitative measures of which nodes are important or central to the network are called *centrality measures*, and there are many types. Each centrality measure is based on an intuitive notion of what it means to be central, and while many are correlated, they do not all identify the same nodes as important.

The most basic centrality measure is called *degree centrality*. The *degree* of a node is simply the number of edges attached to that node (Fig. 2), so degree centrality reflects a belief that more important nodes have more connections. In the case of directed networks, where a link from node i to node j does not guarantee the reverse, nodes will have an in-degree centrality and an out-degree centrality. While a node's degree is the number of links it has, a node's weighted degree is the sum of the weights of its links, and is sometimes called a node's *strength*.

Another straightforward centrality measure is called *closeness centrality*, which measures how close, on average, that node is to all other nodes of the network (Fig. 2). It is calculated by finding the length of the shortest path from each node to all other nodes, and

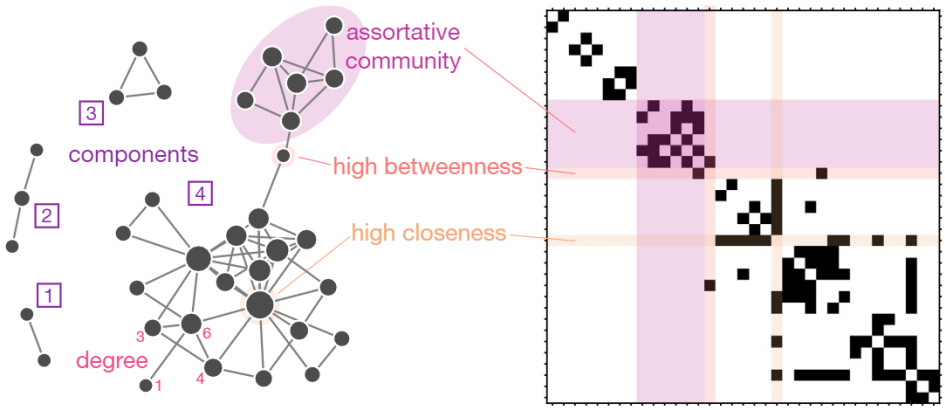

Fig. 2 Illustration of a force-directed network layout (*left*) and its corresponding adjacency matrix (*right*; see text for detailed definition), with some of the basic network concepts labeled, for illustrative purposes. The four components are labeled with boxed numbers; the degrees of four nodes are labeled with unboxed numbers; nodes with high betweenness centrality and high closeness centrality are annotated; an assortative community is also annotated. As shown, this network is undirected and unweighted. The sizes of the nodes in the force-directed layout have been scaled by degree. Force-directed layout made using *webweb* (Wapman and Larremore, 2019).

then computing the average of that set of path lengths. Building on the idea of the set of shortest paths between the pairs of nodes in a network, the *betweenness centrality* of a node counts the number of shortest paths, among all the possible shortest paths between pairs of nodes in the entire network, that travel through that particular node. Naturally, nodes with a smaller closeness centrality have better access to the rest of the network via shortest paths, while nodes with a high betweenness centrality are often bridge or gatekeeper nodes between different parts of a network (Fig. 2). There is a wide variety of other centrality measures, and choosing the appropriate centrality measure for any particular analysis requires some consideration about what, precisely, its intended use or interpretation.

Large-Scale Structure in Networks

In the previous subsections, we introduced vocabulary to describe networks at the level of individual nodes and links. In a sense, these are the fundamental building blocks of any network. However, networks exhibit large-scale structure as well, and it is arguably the ability to define and identify structures at that scale that has allowed networks to become particularly useful and interesting.

The easiest large-scale structure to identify in a network is called a *component*. A component is the complete set of nodes that are reachable from each other by traversing the network's links (Fig. 2). If no path exists between two nodes in a network, those nodes belong to separate components. A network with only one large component, such that one can reach any node from any other node, is said to be *connected*. For instance, in a network of roads between the world's cities, the cities of Australia would form their own component. The concept of components can be generalized to directed networks as well; see Newman (2018).

Often a network will feature groups of nodes that, while not entirely separate from the rest of the network, are nevertheless far more densely connected to nodes within their group than to nodes in other groups. These sets of nodes are called *assortative communities* or modules (Figs. 2 and 3). The process of automatically finding communities in a network is called community detection, and it is perhaps one of the most well studied topics in network science. In non-network data, this type of modular structure might commonly be called clustering, but in the study of networks, clustering refers to a separate, much smaller-scale phenomenon (Newman, 2018).

While assortative communities are groups with statistically high internal link density, the opposite is also possible: *disassortative* communities are groups of nodes with statistically low internal link density. For instance, in a network of genetic co-occurrence, a disassortative group of gene motifs with very low co-occurrence could indicate that those motifs are mutually exclusive or functionally equivalent (Rorick et al., 2018).

Some networks, called *bipartite* networks, consist of two types of nodes and have perfectly disassortative structure, meaning that links can only form between nodes of different types. Bipartite networks are often described in a way that makes clear what the two "parts" of the network are: actor-movie networks, plant-pollinator networks, or gene-motif networks (Larremore et al., 2015). Bipartite networks are typically recognizable because links only make sense between the two types of nodes, but never between nodes of the same type.

In the analysis of bipartite networks, it is common to eliminate one side of the network entirely, and connect two nodes of type *a* if they share a common neighbor of type *b*. For instance, an actor-movie bipartite network might be reduced to a co-starring network; an author-paper network might be reduced to a co-authorship network. Indeed, any *co-something* network is likely formed from a bipartite network! This unipartite version of a bipartite network is called a *projection*. As shown in Fig. 4, one can project a

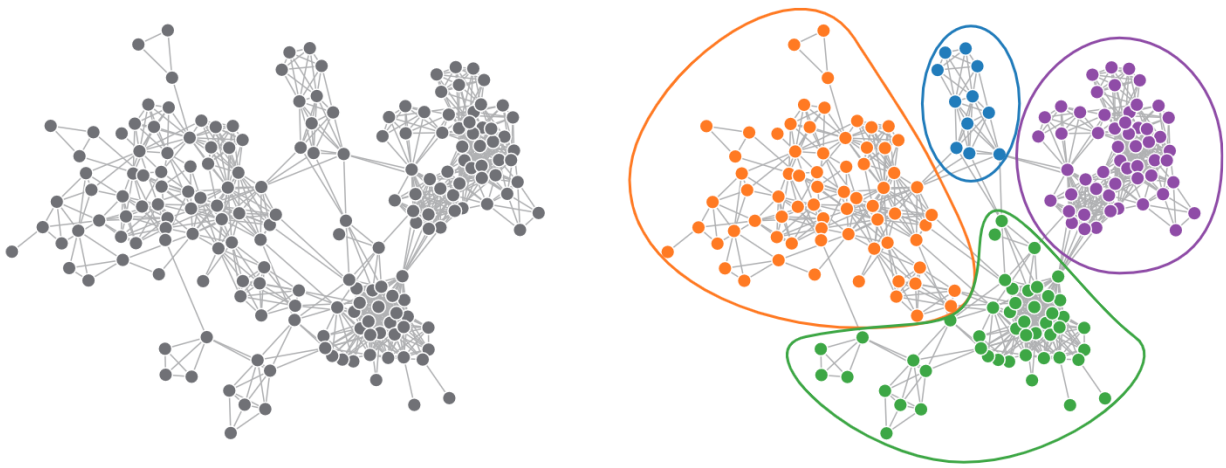

Fig. 3 Force-directed network layout and the result of community detection (*right*). The communities here are assortative, meaning that nodes within each group tend to connect with other nodes in the same group. Force-directed layout made using *webweb* (Wapman and Larremore, 2019). Modified from Larremore DB, Clauset A, and Buckee CO (2013) A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Computational Biology* 9(10): e1003268; with permission of D. Larremore.

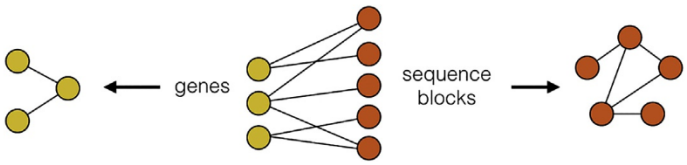

Fig. 4 Network representation of how genes and constituent sequence blocks (e.g., domains or homology blocks) can be differently represented. In the center, genes (*yellow circles*) are connected to sequence blocks (*orange circles*) found within them. Alternatively, this can be represented by a network of genes that share common sequence blocks (*left*) or a network of sequence blocks that are found in the same genes (*right*).

bipartite network in either direction, and thus the same dataset may be analyzed directly as a bipartite network or as either of its projections (Fig. 4).

Large-scale structure can often be revealed by network visualization. Typically, a network is visualized using either a heatmap of the adjacency matrix or a force-directed layout, and each has its advantages and disadvantages. If the rows/columns of an adjacency matrix are sorted, a heatmap can reveal assortative and disassortative groups of nodes that, more broadly, tend to have similar patterns of connection within a group and to other groups. Notice that plotting a heatmap of an adjacency matrix, particularly when it is unsorted, does not easily reveal the number of components of a network.

The second common network visualization is called a *force-directed layout.* Nodes are represented by circles, and links by lines connecting the circles (Fig. 2). Each node is imbued with a repulsive force, so that nodes tend to spread out. But each link is imbued with a force that attracts the nodes at its endpoints, so that linked nodes are drawn to each other. The balance of these attractive and repulsive forces tends to place together assortative communities with many internal links, and clearly spreads apart separate components. However, only networks with a moderate link density are easily visualized in a force-directed layout; networks with a high link density appear as an unresolvable "hairball." Note also that groups of nodes that are characterized by their lack of connections to each other (disassortative structure) will fail to be plotted together in a force-directed layout. *In general, it is advisable to remember that network visualizations of any form are merely pictures of a network, and that a network, in turn, is merely an abstraction of a set of relationships.*

var Genes and Repertoires: Shared Sequence Structure

Structural relationships between genes are typically represented using phylogenetic trees. However, because tree-fitting models assume there is no recombination, they are ill suited for analyses of highly recombinant genes, including *Plasmodium falciparum's var* gene family (Claessens et al., 2014). The highly diverse and multicopy *var* genes encode a protein called *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP-1), which the parasite uses to evade host immunity as described below. Networks analyses, in contrast to phylogenetic tree-fitting algorithms, can accommodate recombination. In this section, we briefly review the biology of *var* genes themselves, before delving into the many recent applications of networks to analyze them.

Biological Background of *var* Genes

Antigenic variation, the mechanism by which *P. falciparum* alters surface proteins to evade detection by the immune system, is mediated by the highly diverse multi-copy *var* gene family (Gardner et al., 2002; Scherf et al., 2008; Deitsch and Dzikowski, 2017). Each parasite contains approximately 60 of these genes, which we will refer to as its *var* repertoire. These genes encode various forms of a parasite protein trafficked to the surface of infected red blood cells, aptly known as *P. falciparum* Erythrocyte Membrane Protein 1 (*Pf*EMP1). The role of *Pf*EMP1 is twofold—binding to host receptors and generating a strong antibody response. These *Pf*EMP1 proteins presented on the surface of infected red blood cells bind to host receptors, such as CD36 (cluster determinant 36), ICAM-1 (intercellular adhesion molecule 1), and CSA (chondroitin sulfate A) (Deitsch and Dzikowski, 2017). Such binding is necessary as the parasite multiplies within the infected red blood cells filling more and more space and making the infected red blood cell stiffer. These inflexible cells are easily removed during filtration in the spleen; by remaining bound inside capillaries and among other host tissues during this portion of the parasite life cycle, the parasite ensures survival until its progeny have been released. In addition to binding the walls of host capillaries, certain *Pf*EMP1 also bind uninfected red blood cells, facilitating bursting progeny to easily find a new home, or bind immune cells, slowing their killing capabilities.

Genes of the multi-copy *var* gene family, although extremely diverse in sequence, share a number of similar features (Kyes et al., 2001). Their internal gene organization is consistent with two exons separated by a transmembrane domain. The larger 5′ exon contains the majority of the sequence variation and is composed of two or more domains (Fig. 5), separated by cytosine rich sequences (Kyes et al., 2001; Scherf et al., 1998). Two common domain types are Duffy binding-like (DBL) domains and cysteine-rich interdomain region (CIDR) domains; in fact, nearly all *var* genes across multiple species contain at least one DBL domain and one CIDR domain (Kyes et al., 2001). The domains themselves contain a similar pattern with areas of high conservation and high diversity; these can be further subdivided into short sequence regions of high homology, known as homology blocks, separated by regions of low homology (Fig. 6) (Rask et al., 2010; Larremore et al., 2013). Originally only described for DBL domains (Smith

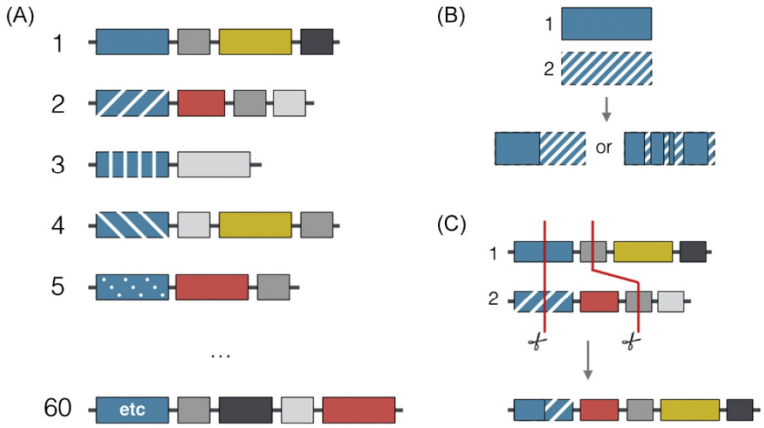

Fig. 5 Schematic of *var* gene structure and recombination. (A) Approximately 60 *var* genes are found in each genome spread across multiple chromosomes, primarily near the center or tails (location not shown). Each gene is composed of domains (represented by differently colored blocks) that show high diversity (represented by patterns within blue blocks). (B) Recombination between domains of the same type may occur through swapping of large regions (*left*) or of smaller homology blocks (*right*). (C) Recombination between genes may result in swapping of pieces of domains or whole domains.

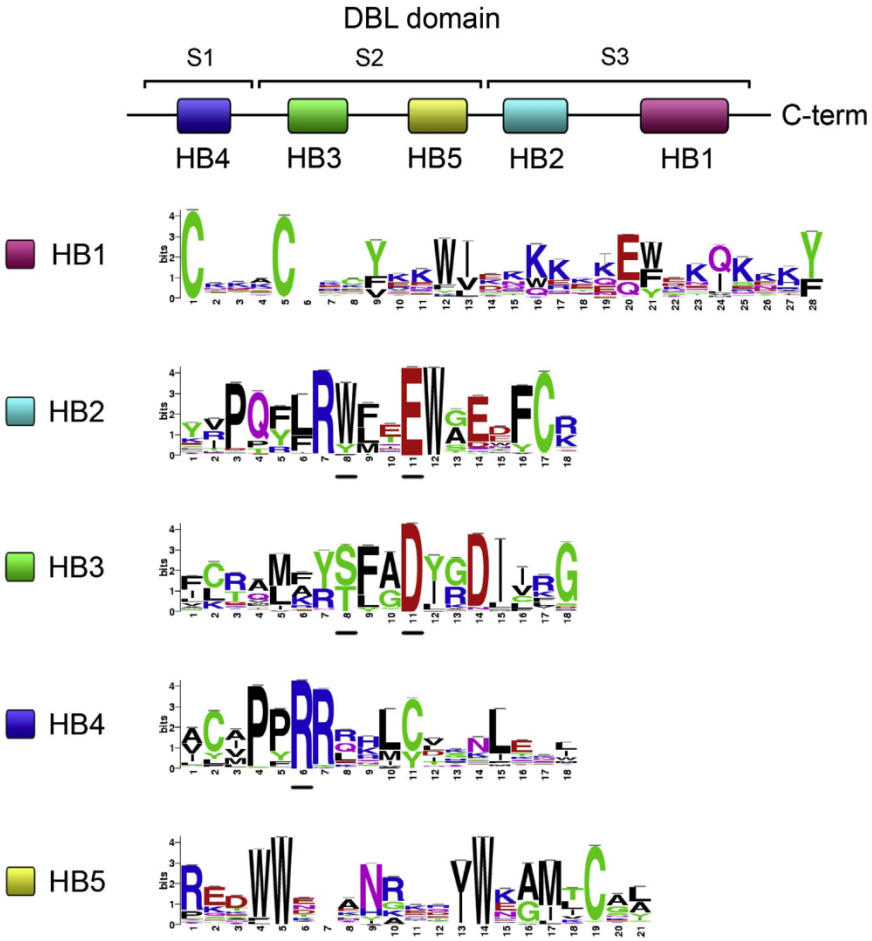

Fig. 6 Schematic of homology blocks within a DBL domain. Each homology block (HB) is labeled with a number, and the amino acid sequence conservation across each block is depicted by logos of sequence. Reproduced from Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, and Lavstsen T (2010) Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer. *PLoS Computational Biology* 6(9): e1000933; with permission of T. Lavstsen.

et al., 2000), they were extended to entire *var* genes with an hidden Markov model-driven homology block detection pipeline in Rask et al. (2010). Through their analyses, the authors uncovered an evolutionary signature suggesting a recombination break point within DBL domains as well as likely functional reasons for conserved regions, such as phosphorylation sites. They describe these conserved structures within and between domains as potential functional units, a building block for ensuing work on functional "strain" classification of sequences.

var Gene Classification

The significant diversity found among *var* genes coupled with the repeated appearance of certain genes and subsequences has led to extensive study of what structure can be determined within the population and how that diversity is developed and maintained. In attempts to understand this diversity, a myriad of classifications were introduced.

Initial classifications involved descriptive characteristics, such as chromosomal location, orientation, and gene length. While predominantly found near the ends of chromosomes, or telomeres, a fraction are found near the center of chromosomes, or centromeres. Recent evidence suggests epigenetic control of *var* gene transcription accessibility via histone modification (Deitsch and Dzikowski, 2017); their predominant location near centromeres and telomeres may mediate this regulation.

Consistent with the above patterns in chromosomal location, orientation and gene length, are particular upstream promotor sequences, known as *ups*, that led to categorization of the genes by these regions. Longer, more conserved genes pointing toward the telomeres, predominantly fall into the *upsA* designation while shorter, centrally-located ones are classified as *upsC* (Lavstsen et al., 2003). An intermediate category, *upsB*, contains *var* genes in both locations but with lengths more similar to those designated *upsC*. Importantly, several genes do not fit these categorizations (Lavstsen et al., 2003), leading to *upsD* and even an *upsE* categorization. The proliferation of these categories, like prior classification schemes, struggled to fully disentangle the high conservation and high diversity.

Internal sequence features also played a role in categorization. *var* genes can be grouped by the number of conserved cysteine (cys) residues they contain in their DBLα domain (Kyriacou et al., 2006). Most of the genes are found with 2 (cys2) or 4 (cys4), while a few contain an odd number or none at all. Conveniently, most of the cys2, although not all, fall into the *upsA* category; other *ups* groups predominantly are cys4. Coupled with the cysteine dichotomy, labeling based on a set of recurring sequence motifs at particular locations, known as positions of limited variability (PoLV), helped distinguish sequences (Bull et al., 2007). Such cys/PoLV categorizations provided a rapid way to categorize *var* genes (Bull et al., 2007).

Despite the existence of multiple classification schemes, such as chromosomal location, upstream region, length, and structural similarity, it rapidly became clear that none would fully describe the mix of high sequence diversity coupled with repeated appearance of the same genes or subsequences within the population. Furthermore, with every new parasite sequenced additional *var* genes are discovered. As the set of known *var* genes was continually expanding, it became apparent just how much diversity may exist. Novel methods beyond straightforward descriptive classification were needed to uncover relationships between the *var* genes and parasites containing them. Network analysis would provide such a tool.

Structure and Maintenance of *var* Diversity

Initial efforts to make sense of the diverse *var* gene family focused on structural similarity through the development of phylogenetic trees (Lavstsen et al., 2003; Kraemer and Smith, 2003). These analyses demonstrated that all the previously introduced classification systems, such as upstream promotor region, were insufficient to describe relationships among the *var* genes. This was abundantly apparent in work by Trimnell et al. (2006) who present their phylogeny with a tree of *var* gene relationships from 22 world-wide isolates. Their resulting star-like phylogeny, where genes are almost all on equal length branches from a central node, highlights both the vast diversity and high level of conservation of these genes. Zilversmit et al. (2013) expanded on this work by including genes found in *P. reichenowi,* a similar malaria species to *P. falciparum,* and qualitatively found a nearly identical tree structure (Fig. 7). The extension to include *P. reichenowi* demonstrates the ancient nature of this family. Opposing evolutionary forces are acting on the gene family, which itself is a primary agent of malaria antigenic variation.

A major reason for the inability of tree methods to discern structure among the *var* gene pool is the highly recombinant nature of these sequences (Freitas-Junior et al., 2000), which tree-building algorithms struggle to satisfactorily incorporate. Network analyses, however, are more robust to such complex biological relationships and have become an important tool to understand and examine *var* gene relationships. In one of the first studies to use network theory to analyze *var* gene relationships, Bull et al. (2008) determined which *var* genes share contiguous blocks of high sequence similarity and used this information to build relationships between *var* genes from a pool of patient isolates. These relationships are represented as a network of sequence blocks through presence of shared subsequences among different isolates. Importantly, they used their network based on sequence blocks, specifically position specific polymorphic blocks (PSPB), to evaluate previous types of categorizations such as cys/PoLV, and found that the network approach offered different information. For example, the PSPB network structure provided insight into phenotypic response, such as recombining communities, revealing functional as well as structural relationships. While the studies reviewed above have sought to discover the effects of recombination on *var* structure and function from isolates, the laboratory studies of Claessens et al. (2014) have turned toward the mechanistic constraints of the recombination process itself. Using whole-genome sequencing of many generations of clones, they observed a large number of mitotic recombination events during asexual replication. They built relationships around these mitotic recombinations, describing an unusually high level of in-frame mitotic

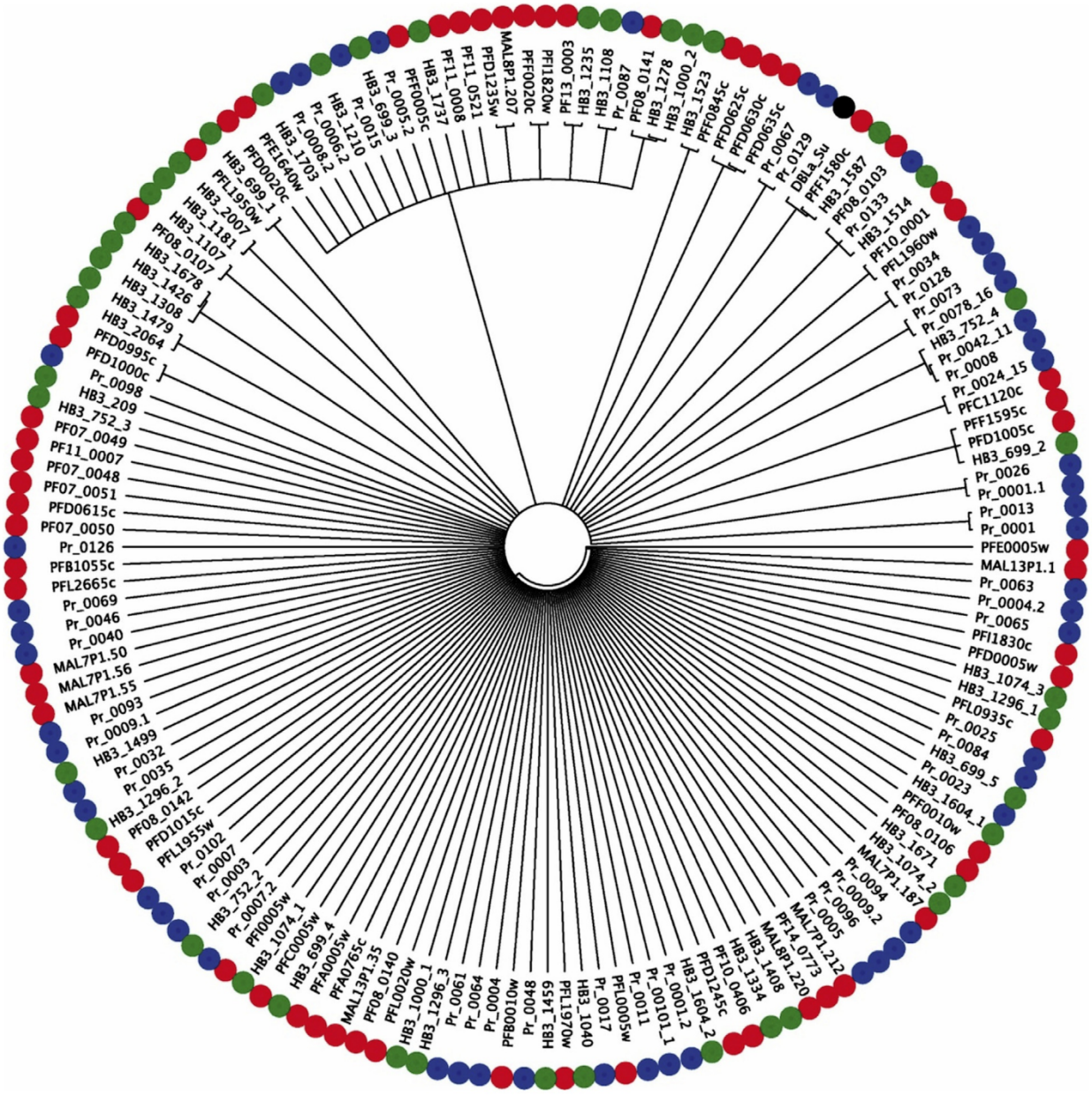

Fig. 7 Tree-based representation of relationships of amino acid sequences of DBLα sequences from *var* genes. A proportional maximum-likelihood phylogeny of sequences in *P. falciparum* isolates 3D7 (*red circles*) and HB3 (*green circles*) and *P. reichenowi* (*blue circles*) and reference (*black circle*). Reproduced from Zilversmit MM, Chase EK, Chen DS, Awadalla P, Day KP, and McVean G (2013) Hypervariable antigen genes in malaria have ancient roots. *BMC Evolutionary Biology* 13(1): 110; with permission of M. Zilversmit.

recombination which differs in form from standard homologous recombination. Their work helped establish some quantitative descriptions of *var* gene diversity generation, and showed ways in which previous classifications (Rask et al., 2010) may not be fully predictive of possible recombination events.

Origins of *Pf* Multigene Families in Ape-Infecting *Plasmodium Laverania* Parasites

Examining sequence similarity of *var* genes, subdomains, or sequence blocks is not only useful to examine relationships of *var* genes within a particular population but across space, time or even species. Networks of sequence similarity between genes can be used to uncover relationships and build phylogenies, disentangling sequence overlap arising from identity by descent (IBD)—inheriting the identical sequence from a common ancestor—as compared to recombination—acquiring an identical sequence from another currently present parasite.

The conservation of similar *var* gene fragments, in particular within the DBL and CIDR domains, speaks to the long history of *var* genes, and their relationships across the globe and through evolutionary time. Some of the work mentioned previously, such as Zilversmit et al., (2013), demonstrated the interspecies relationships between *P. falciparum* and a chimp-infecting malaria species, *P. reichenowi*. Larremore et al. (2015) expanded on the historical relationships by examining DBLα sequences amplified from chimp and gorilla fecal samples. Restricting to samples containing only single-species infections from *Plasmodium Laverania* spp. including *P. falciparum*, they used novel network methods developed in Larremore et al. (2013) to examine relationships between the species. For the DBLα domains, Larremore et al. (2015) demonstrated the history of the *var* genes dated back to the origin of the *Laverania* parasites. A recent paper by Otto et al. (2018) confirmed and extended this result including all known *Laverania* species and additional multi-copy gene families including *rifin* and *stevor*. They estimated the beginning of speciation for *P. falciparum* as well as other evolutionary constraints, such as a bottleneck. With their analyses, they open the door for determining why only a single *Laverania* species is able to infect humans.

Haplotypes and Vaccine Development

More generally, haplotype networks, representing similarities between different haploid genotypes, typically consider genes other than *var* genes. Many of the genes analyzed in depth via haplotype networks have possible therapeutic uses, such as potential vaccine candidates. We briefly introduce two.

Recent work by Chowdhury et al. (2018) compares two erythrocyte binding antigens (EBA-175 and EBA-140) to understand the role of selection for these two vaccine candidates. These EBA proteins are highly polymorphic and used for the parasite to bind and enter red blood cells during the blood-stage of parasite development. The authors build a haplotype network for each antigen using the median-joining method described in Forster et al. (2001) and show that the networks have distinct structure. The differences in the resulting networks are assumed to be an indication of differing evolution. Haplotypes for EBA-175 evolved from 3D7 through accumulation of variants while EBA-140 showed a population expansion.

Despite this review's focus on polymorphic genes, in particular *var* genes, haplotype networks are being used to examine other proteins. For example, the genetic diversity and sequence evolution of CelTOS (cell-traversal protein for ookinetes and sporozoites), another potential vaccine candidate, has also been examined through generation of a haplotype network (Pirahmadi et al., 2018). Using samples from around the world, the authors show that there is clear regional genetic structure with no shared haplotype across all regions. The minimal antigenic diversity in CelTOS, which they report, may indicate a critically conserved function. Unlike the high sequence diversity found in *var* genes, this network-based analysis further supports the potential for use of CelTOS as a vaccine candidate.

Outlook

The ability of network analyses to include recombination, known to occur frequently among polymorphic genes, such as *var* genes, has advanced our ability to understand both their population structure and their evolutionary history. Going forward, network analyses are likely to remain a primary method for examining, categorizing, and understanding these diverse groups. In that process, novel network-based methods will be needed to determine which observed evolutionary relationships are driven by recombination, selection, or other forces.

Analysis of genetic similarity, whether via networks or other methods, ignores the object assumed to be under selection: functionality. Furthermore, whether the unit under selection is a gene, a domain, a homology block or an epitope, remains unknown. Therefore, in the next section, we zoom out and instead of considering networks at the level of gene subsequences and individual genes, we look at applications of networks in the study of population genetics and related questions from ecology.

Genomics and Phylogeography

The previous section showed how networks have been used to understand the evolution of genes and gene families, and accordingly, the nodes in those networks were typically genes or their constituent sequence blocks. We now zoom out from genetic networks to networks of a broader scope, in which nodes are entire genomes, isolates, or even geographic regions.

Networks in this section are generally characterized by weighted links, due to the fact that a comparison between two complex objects, e.g., two genomes or regions, is more complicated than a presence-absence measurement. In the first set of analyses, which focus on the evolution of parasites and populations over time, links are often undirected, with weights that quantify similarity. In the second set of analyses, which focus on the spread and evolution of parasites over geographic space, links are typically directed, with weights that quantify flows of parasite genomes between cities, countries or regions.

Population Genetics Over Time: Immunity and Repertoire Selection

One of the fundamental puzzles in the study of *P. falciparum* is that those individuals in endemic regions who survive infection during childhood are likely to be repeatedly infected throughout their lives. On the surface, this is due to immune evasion by parasites, yet human immunity itself is not static, either for individual hosts whose immune systems adapt, or for populations of

hosts, whose immune systems collectively drive the evolution of locally circulating parasites. Unsurprisingly then, both parasite genomes and human immunity have left identifiable imprints on each other, and both have been investigated using networks.

How does human immunity change after an infection, and what are the implications for the antigenic diversity of parasites? This question was addressed by Buckee and Bull through the analysis of changes in networks of immune recognition over the course of recovery from infections in Kenya (Buckee et al., 2009). In these serological networks, each node represents a patient and their parasite isolate, and a directed edge is drawn from one node to another when the antibodies in the former's serum show a significant agglutination response to the latter's parasites.

Buckee and Bull constructed two serological recognition networks to compare change across time. The first used patient sera drawn at the time of acute clinical infection, and the second used patient sera drawn 3 weeks later, after convalescence. In such a construction, each link in the "acute" network accounts for recent history of infection, while each link in the convalescent network reflects that additional recognition has been acquired through an adaptive response, following initial presentation. The changes accumulated from one network to another are therefore reflective of both adaptation by individuals' antibodies and antigenic similarity between the parasites that infect them. Buckee and Bull use a suite of network metrics including reciprocity (the fraction of directed links that are reciprocated), transitivity (the extent to which edges from i to j and j to k predict an edge from j to k), and node degree in their analysis. Using these measures, they compare the changes in data-derived network structure with those from simulated data, and show that the *var* gene population structure arising in the patient data is most likely the result of encountering a mixture of common *var* genes and a large pool of high diversity *var* genes. In other words, this network analysis provided evidence for the hypothesis that the population of *var* antigens can be partitioned into a small and easily recognized set, and a large and highly diverse set. This hypothesis has since been substantiated by analyses of *var* gene block-sharing networks (Bull et al., 2007; Buckee et al., 2009; Larremore et al., 2013) and mathematical models (Buckee and Recker, 2012).

More recent work has continued to combine network analyses and immune assays, taken at different points in time. In Senegal, Bei et al. (2015) used plasma samples collected across different years to test for both the presence of antibodies that could inhibit parasite growth in culture and those recognizing the surfaces of infected erythrocytes (Bei et al., 2015). By combining this information with networks of *var* gene-sharing between parasites, they showed that parasites with similar genomic *var* repertoires tended to elicit similar immune responses. By comparing to controls over time, they showed that immune recognition at the human population level was stimulated by exposure to the locally prevalent parasites in circulation the previous year. In this way, parasite groups with similar genomic *var* repertoires have been linked to the acquisition of new immune responses in local populations. Taken together, combinations of serological and genomic networks (Buckee et al., 2009; Bei et al., 2015) show the potential for future studies of the feedback mechanisms between immunity and immune evasion, more broadly.

var Repertoires, Immune Selection, and Strain Theory

A substantial amount has been learned about the evolution of genomic *var* repertoires under immune selection by directly studying the structure of *var repertoire overlap* networks, even when immune assay data are unavailable. In a repertoire overlap network, each node corresponds to a single parasite, and an edge between two nodes represents the weighted similarity of those parasites' *var* gene repertoires. When both parasites' *var* genes are fully sequenced, similarity can be directly measured by the fraction of *var* types that are common to both parasites, divided by the total number of *var* genes sequenced from both parasites, a measure called *pairwise type sharing* or the Sorenson-Dice coefficient (Barry et al., 2007). When *var* repertoires are not fully sequenced, pairwise type sharing can lead to links with downwardly biased weights, which has motivated the recent development of unbiased estimates of *var* type sharing (Larremore, 2019). This Bayesian repertoire overlap method, however, has not yet been extended to the analysis of parasite isolates from infections with high multiplicity of infection, making their use for analysis of isolate networks less useful. More broadly, one can think of repertoire overlap networks as networks of what ecologists call β-diversity.

Building and analyzing networks of the *var* repertoire overlap can help make sense of parasite population structure, and in particular, to refine and extend so-called "strain" theories of malaria. In the 1990s, Gupta et al. showed that when antigens consist of multiple loci which can be recombined, those loci may persist in linkage disequilibrium to form what appear to be discrete and non-overlapping strains, in spite of ongoing recombination (Gupta and Day, 1994; Gupta et al., 1996). In other words, immunity itself can divide a population of antigens into non-overlapping groups, overcoming the recombination processes that would otherwise mix those groups. To examine this theory in the context of *var* repertoires, Artzy-Randrup et al. (2012) used an agent-based simulation and found that *var* repertoires are indeed partitioned at the population level, based on the accumulation of specific immunity. However, going beyond the previous strain theory (Gupta and Day, 1994; Gupta et al., 1996), they observed that while the population is partitioned into co-existing dominant *var* repertoires, the overlap between those repertoires is dependent on the intensity of transmission.

The model by Artzy-Randrup et al. (2012) focuses its simulations on populations of moderate diversity, and operates in a closed system without the entry of new variants. In contrast, real populations may have far higher diversity, with turnover in the gene pool. In this context, Day et al. (2017) examined data from Bakoumba, Gabon, to analyze the structure of empirical *var* repertoires and their overlap. They constructed networks of pairwise type sharing between isolates and then used a randomized network null model to measure the extent to which the observed data are statistically unexpected. In this form of random network, nodes retain their degrees but the edges are otherwise randomized, also known as the configuration model (Fosdick et al., 2018) in the broader network science literature. Day et al. (2017) used evidence from these analyses to argue for the maintenance of weakly overlapping repertoire structure in the population, finding that repertoires are less similar to each other than expected by chance. Thus, deviating

from classical strain theory which predicts complete segregation of antigenic variants (Gog and Grenfell, 2002; Gupta and Day, 1994; Gupta et al., 1996, 1998), Artzy-Randrup et al. (2012) showed via simulation that *var* repertoires under immune selection, particularly at higher transmission settings, will retain some level of overlap (Artzy-Randrup et al., 2012), and Day et al. (2017) subsequently found partial overlap between repertoires in field data via in-depth sequencing (Day et al., 2017). Importantly, the work of Day and coauthors found significant deviations from expectation under a randomized network null model, but whether these deviations could be attributed to immune selection remained remained an open question. This motivated further analyses which could differentiate between deviations from statistical null models and evolutionarily neutral models.

To further understand real-world *var* population structure in the presence of high rates of recombination, importation and turnover, and transmission, He et al. (2018) used static snapshots of *var* population structure, as represented by a network at various points in time. This work introduced a new way of defining the network edges between repertoires, letting a directed edge from repertoire *i* to repertoire *j* represent the number of shared *var* genes between *i* and *j* divided by the number of unique *var* genes found in *i*. Note that not only is this measurement different from pairwise type sharing, it is also asymmetric, so that the direction of the stronger edge indicates whether one repertoire can outcompete the other. Using this novel definition, a model incorporating immune selection was compared to two neutral models (one employing immunity based on the number of past infections, and one lacking such a history) through simulation. The authors then used a suite of network properties, including centrality and particular motifs, to distinguish between networks resulting from the contrasting models. They found that network patterns from Ghana were better explained by the immune selection model than either neutral model, deepening the connection between *var* repertoire evolution and immunity, by bringing to bear "repertoire competition" networks.

Most recently, Pilosof et al. (2019) extended the previous network analyses to explicitly include temporal relationships in *var* population networks by making use of *multilayer* networks. In a *multilayer* network that represents temporal relationships, nodes, which represent parasites from a set of different time periods, are separated into a corresponding set of different layers. This means that although individual networks from particular time periods could be analyzed separately, the layers of the network can instead be linked together by edges that connect pairs of parasites that are similar from 1 year to the next. Thus, by using multilayer networks, Pilosof et al. were able to analyze *var* repertoire relationships while accounting for transmission dynamics between seasons (Fig. 8). This multilayer network representation, when combined with community detection methods for multilayer networks, allowed the authors to differentiate between the patterns one would expect from immune selection and antigenic drift. From a population genetics perspective, they therefore showed that *var* populations can be temporally organized into groups of repertoires that are more similar to each other than to the repertoires in other groups, and that those community structures persist for longer under models incorporating immune selection than the neutral models of He et al. (2018). By finding similar long-lasting modules in real data, these multilayer network analyses support the notion of functionally defined "strain" structure in *var* gene populations with epidemiological consequences.

The analyses described above involve repertoire similarity networks, which are fundamentally based on the idea that the scale at which repertoires ought to be compared is at the level of genes. In other words, to construct such networks, the individual *var* genes of various parasites are compared, and two repertoires are generally more similar when they share more of the same genes. However, this premise, while reasonable, also overlooks the possibility that there exist *var* genes which are functionally equivalent, yet genetically different. To explore this possibility, Rorick et al. (2018) examined a network of co-occurrence of homology blocks (HBs) in *var* genes by drawing an edge between two HBs when they were found in the same *var* gene. (This approach reverses the more typical networks where nodes are *var* genes which are connected when they share sequence blocks (Bull et al., 2007; Larremore et al., 2013).) It is helpful to note that both approaches can be thought of as projections of a larger bipartite network in which genes are linked to the sequence content they contain. Thus, typical methods project away the sequence content to obtain a gene-to-gene network, while Rorick et al. (2018) projected away the genes to create an HB-to-HB network. Using this novel perspective, along

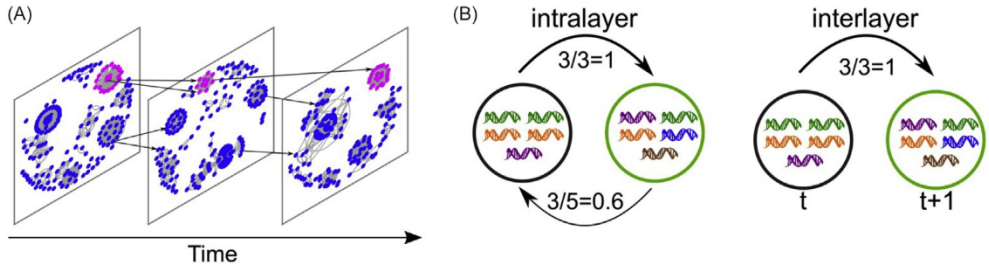

Fig. 8 A temporal multilayer network of repertoire genetic similarity and its associated modular structure. (A) A multilayer network where each layer's nodes represent genomes collected at particular time points. Edges indicate the genetic similarity between pairs of *var* repertoires, and in particular, edges between layers show how genetic similarity exists across adjacent years, with one particular multi-year module highlighted in pink. (B) A schematic of the construction of directed edges within and between layers, based on the contents of *var* repertoires. Here, the black- and green-circled repertoires share three alleles, depicted as colored DNA icons, but the black repertoire has three unique alleles while the green has five, meaning that the latter can outcompete the former. For clarity, only a few interlayer edges are represented in (A). Reproduced from Pilosof S, He Q, Tiedje KE, Ruybal-Pesantez S, Day KP, and Pascual M (2019) Competition for hosts modulates vast antigenic diversity to generate persistent strain structure in Plasmodium falciparum. *PLoS Biology* 17(6): e3000336; with permission of S. Pilosof.

with flexible community detection methods which are capable of finding disassortative patterns—that is, groups of HBs which very rarely co-occur, and are therefore likely to be substitutes for each other—they reduced the total diversity from over 10,000 *var* types to only 48 key *var* "functional types." Examining individual parasites, they showed that functional types are quite common, indicating a shared parasite strategy on a functional level, which does not necessarily extend to the genetic level, providing a mechanism for the conserved functionality but high diversity exhibited by this gene family.

Spatial Population Genetics and Phylogeography

Just as in phylogenetic studies, network analyses can be used to understand patterns of selection in space, as well as time. And, just as in phylogenetics, analyses can be performed with or without ever forming an explicit tree or network. For instance, in geographic studies, repertoire overlap networks are commonly analyzed without explicitly describing the pairwise relationships as a network; here, repertoire overlaps are calculated for every possible pair of parasites in a sample spanning multiple geographies, and then, the distribution of repertoire overlaps among all the parasites within a region are compared to the distribution of repertoire overlaps between two regions. Such an approach has been used in analyses of *P. falciparum* in Papua New Guinea (Barry et al., 2007; Tessema et al., 2015), Brazil (Albrecht et al., 2010), and across multiple locations (Albrecht et al., 2010; Larremore, 2019) (Fig. 9). Were such analyses explicitly cast in the framework of networks, these statistical tests would be comparing link weight distributions within and between communities.

Naturally, phylogeographic analyses using networks are not restricted to *var* repertoire overlap. Taylor et al. (2017) argue that existing approaches, based on common population genetic measurements of the fixation index (F_{ST}), are sufficient for large spatial scales but not small spatial scales. In these networks, nodes correspond to locations, and edges are related to the proportion of parasite pairs that are related between locations. In particular, they show that at small spatial scales, spatial genomics and genetic epidemiology are better served by identity-by-descent (IBD) measurements of relatedness than F_{ST}. Henden et al. (2018) used IBD to build relatedness networks to determine how the population structure has changed across continents. Note that from the network perspective, the difference between using summaries of IBD on the parasite population level and F_{ST} amounts to changing the definition of the network's links, but not the network's nodes, and as with any other measurement, it is not surprising that some measurements are better suited for particular scales of analysis than others.

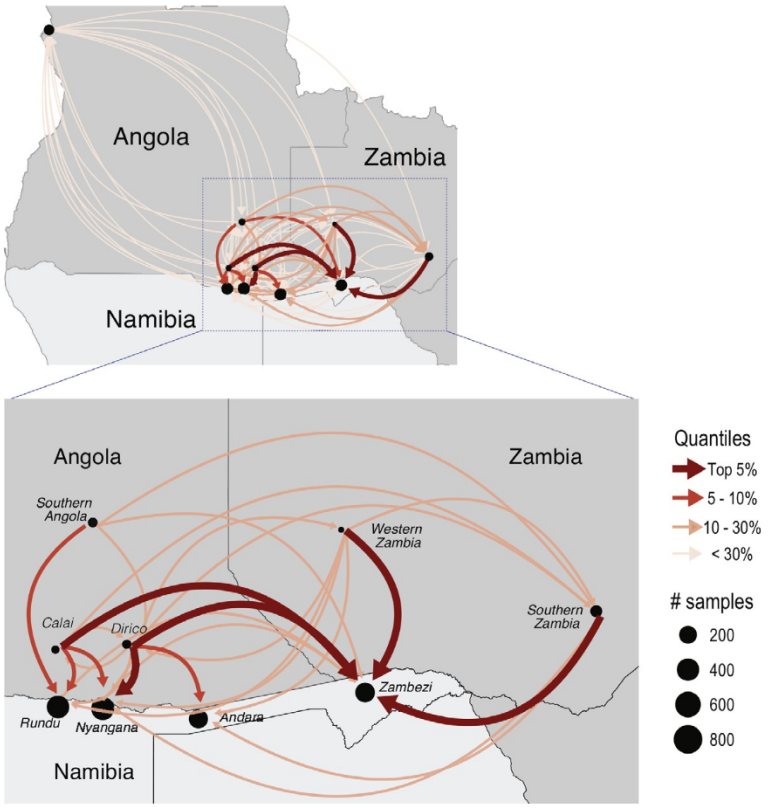

Fig. 9 Genetic and malaria incidence data are used to estimate importation of malaria and directionality of parasite flow at the border of Angola, Zambia and Namibia. Direction and thickness of arrow indicate malaria importation shown by quantiles. Details of the methodology can be found in (Tessema et al., 2015). Reproduced from Tessema SK, Monk SL, Schultz MB, Tavul L, Reeder JC, Siba PM, Mueller I, and Barry AE (2015) Phylogeography of var gene repertoires reveals fine-scale geospatial clustering of Plasmodium falciparum populations in a highly endemic area. *Molecular Ecology* 24(2): 484–497; with the permission of B. Greenhouse.

Finally, networks have been used extensively to model the directional flows and movements of people, who may carry parasite genomes from one location to another (Tatem and Smith, 2010). By combining migration data with the rate of infection among 2–10 year olds, they estimate the *P. falciparum* migration metric for each pair of countries, a value that represents the directed rates of parasite importation into one country from another. In other words, the authors combined multiple data sources to create a directed network that estimates parasite exchange between countries, which they then coarse-grain by using various community detection approaches.

Rather than using migration estimates from census data, others have used the usage of mobile phones to estimate the movement of humans, and therefore parasites. Early work using human mobility and transmission models estimated the impact of mobility networks on the spatial spread of malaria across Kenya, finding source and sink locations at a national scale (Wesolowski et al., 2012).

More recent work using mobile phone data from Bangladesh combines mobility with genetics, learning the relationship between pairwise relatedness using SNPs and geographical distance (Chang et al., 2019). In this work, Chang et al. ask over what distance genetic relatedness is informative, further refining the relationship between genes, mobility, and the estimation of spatial spread. Of note, the presence of mobile phone infrastructure tends to be anti-correlated with extensive public health infrastructure, so mobile phone studies often use travel surveys to corroborate and strengthen their findings (Chang et al., 2019). This work then amounts to using directed mobility networks and genetic measurements to estimate parasite importation networks, with clear implications for targeting interventions and network locations, which will have larger cascading impacts.

Outlook

Networks are likely to remain powerful tools for understanding the complicated patterns of genetic similarities between isolates or parasites, over space and/or time. This applies equally to analyses of *var* gene repertoires, which this review covers in detail, as well as patterns of SNPs (single nucleotide polymorphisms), which we spend relatively less time on. We believe that multilayer networks are likely to become a primary tool for the analysis of longitudinal datasets, in particular because of recent methodological advances. Multilayer networks may also become relevant for spatial epidemiology, particularly as researchers are able to join datasets from human mobility, genetic similarity, geography and climate. However, in parallel, there is an urgent need to better understand the impacts of measurement error, modeling assumptions, and sampling biases, in the context of these powerful new network methods.

Dynamics and Regulation

Beyond the extensive examination of *var* structure at a population level—whether spatial, temporal, or lineage—networks have also been used to study antigenic variation within individual infections. The dynamics of *var* gene expression across the total set of parasites present within a single infection has a rich potential for network representation and analyses.

var Switching

The high sequence variation within the *var* gene family couples with the repeated presence of *var* genes within a parasite's genome: all parasite isolates contain approximately 60 *var* genes. Although these genes may be similar in some ways—chromosomal location, upstream regions, sequence homology blocks—little exact repetition is found of even single genes across parasite *var* repertoires. A notable exception to this pattern is the gene, whose ensuing *Pf*EMP1 protein binds the host receptor chondroitin sulphate A (CSA) found within the placenta (Viebig et al., 2005; Scherf et al., 2008). Notably, this gene, known as *var2CSA*, can only be successfully employed by the parasite during infection of a pregnant woman due the presence of the binding substrate—the placenta. Despite the rarity of its niche, this gene is found in nearly every *var* repertoire sequenced (Rask et al., 2010).

*Pf*EMP1 is a primary protein responsible for the binding between infected red blood cells and host receptors on tissues such as capillary walls (Deitsch and Dzikowski, 2017; Scherf et al., 2008). However, its prominent location sticking out from the surface of red blood cells, also makes it highly immunogenic—leading to the development of an antibody-mediated response. Thus, in order to avoid the development of a strong immune response, the parasite evolved the strategy of mutually exclusive expression—the expression of a single copy of a gene from its large, multi-copy gene family while simultaneously silencing the other gene copies (Dzikowski et al., 2006). In other words, a single parasite only transcribes and expresses a single *var* gene at one time. Thus, that parasite is able to bind only a single host receptor, but it is also exposing only a single *Pf*EMP1 protein to the immune system. The dual function—binding and immunogenicity—of *Pf*EMP1 and its encoding *var* gene family must balance to ensure long periods of survival of the parasite in the human host, allowing ample time for onward transmission.

The mechanisms that regulate mutually exclusive *var* gene expression are complex and under continued investigation (Deitsch and Dzikowski, 2017). The process requires the simultaneous upregulation of one *var* gene coupled with the downregulation of the currently expressed one, coordinated extracellularly across the parasite population. Several layers of the regulation have been described, although the coordination of expression of genes among populations of parasites remains poorly understood (Deitsch and Dzikowski, 2017). It is likely that a combination of activation, silencing, epigenetic regulation, and perhaps even response to immune pressure impact these complex expression patterns (Deitsch and Dzikowski, 2017). Previous work has focused on the

potential roles of *var* gene activation, silencing and the control of switching—the combination of which determines the patterns of *var* gene expression in the absence of selection pressures (Fig. 10). Given the complexity of the process and the inherent connectedness of transcriptional regulation pathways, networks have provided an underlying data structure from which to understand *var* gene switching.

Dynamics of var *switching*

With each individual parasite, expressing only a single *var* gene at a time through the complicated process of mutually exclusive expression, whose mechanism remains under investigation (Deitsch and Dzikowski, 2017), the timing and order of the appearance of individual *var* genes within an infection is of interest. This dynamics is especially important to understand when and how chronic infections occur, and knowledge of the switching patterns and regulation have the potential to lead to new therapeutic options. Due to the sequential nature of *var* gene expression by a single parasite through mutually exclusive expression, a natural representation of *var* gene expression profiles is a network of activation order as appears in (Buckee and Recker, 2012; Noble et al., 2013; Childs and Buckee, 2015). While a network representation is natural here, and essentially required for efficient simulation of this process, rarely are network analyses directly employed to understand features of the structure of the *var* switching network.

With all the complexities in the process of antigenic variation, a parsimonious understanding of underlying antigenic switching networks is challenging. A combination of experimental and theoretical techniques provided insight into the process. Controlled in vitro experiments provide biological information on the order and rates of switching of *var* gene transcription and expression in the absence of selection pressure by the immune system (Noble and Recker, 2012; Frank et al., 2007). Due to constraints on the experimental setup, however, information is lacking. The data is temporarily sparse and complicated by measurement error making it difficult to draw conclusions, particularly on rarely expressed *var* genes (Noble and Recker, 2012). Initial studies showed that *var* gene control was gene specific, with non-random activation and deactivation occurring at highly dissimilar rates across the *var* family (Horrocks et al., 2004). There was some evidence that gene location impacted these rates, with the genes located near the telomeres more likely to be silenced (Frank et al., 2007). At a more global level, in order to avoid exposure of the full *var* repertoire to the immune system, a highly structured switching pattern was predicted, at least early in an infection (Recker et al., 2011).

Noble and Recker (2012) developed a statistically rigorous method for determining the underlying switching network from gene transcription profiles. Honed on simulated gene transcription profiles, the method is able to distinguish the role of on and off rates of genes as opposed to the network structure of antigenic switching. This allows for hypothesis on the structure of switching pathways to be empirically tested, as done in (Noble et al., 2013). Here, the authors use their methods from (Noble and Recker, 2012) to determine on/off rates and the structure of switching networks for *P. falciparum* from in vitro data. They find that certain genes, in particular ones exhibiting high sequence diversity found near centromeres, are favored for activation. In contrast to previous suggestions that this increased level of centrally-located genes was due to low de-activation rates (Frank et al., 2007), they find that off rates contribute little to the higher levels of transcription of certain genes versus others. This method has the potential to be used for a more detailed examination of immune pressures during in vivo studies, although this appears not to have been done yet.

Metabolic Networks

Metabolic interactions and pathways have extensively relied on networks as representation of their various mechanistic links. Here, links are built from known interactions, such as correlated increases or decreases between two components. Original compilations of all the metabolic connections required a careful combing of the literature to uncover all the known components and their interactions, followed by pain-staking work to fit all the pieces together. More recently, algorithms have been developed to computationally complete this task (Fatumo et al., 2011, 2013; Yeh et al., 2004). Network analysis were thus invoked to evaluate

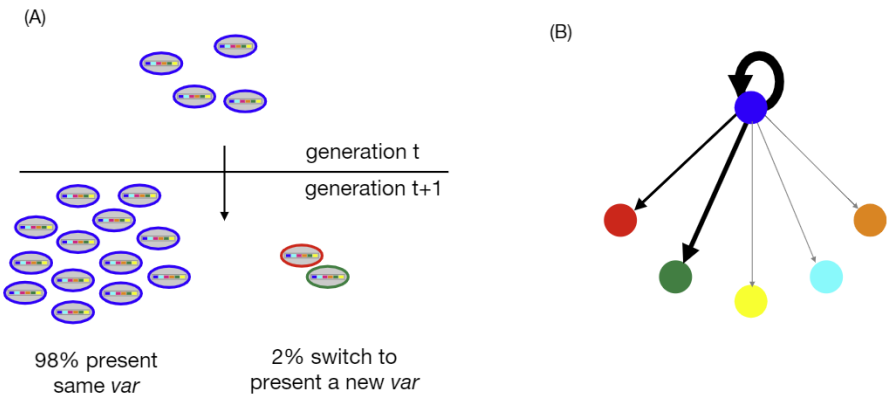

Fig. 10 Presentation of *var* genes at a population level changes with each generation. (A) Schematic *var* repertoire represented by string of colors, with the outer edge color representing the *Pf*EMP1 protein presented. (B) Network representation of switching of presented *var* genes.

and distinguish between computationally derived networks, especially in the absence of a gold standard of comparisons (Fatumo et al., 2011). A variety of databases, including, but not limited to, PlasmoCyc (Karp et al., 2005) Malaria Parasite Metabolic Pathways (MPMP) (Ginsburg, 2009), and MetaSHARK (Pinney et al., 2005), exist to facilitate these constructions. Some such as MPMP manually construct the network (Ginsburg, 2009) while others use bioinformatic algorithms (e.g., (Pinney et al., 2005)) to automatically generate the networks. A major use of these complicated metabolic networks is the detection of potential drug and intervention targets in silico. The development of metabolic networks in malaria is extensive, rapidly expanding with new computational tools, and not the focus of this review, so we encourage readers to explore the literature on metabolic networks of *P. falciparum* and *P. vivax*.

Protein-Protein Interaction Networks

Protein-protein interaction (PPI) networks, in contrast, have used network analysis extensively to identify the most essential proteins for the parasite. As their name suggests, PPI link proteins that are known to be correlated through some interaction. From the original description of the *P. falciparum* PPI in (LaCount et al., 2005), network analyses have shown malaria PPI to be quite divergent from other eukaryotes, similar to the divergence found at the sequence level (LaCount et al., 2005; Suthram et al., 2005; Hase et al., 2010). These large PPI networks and network metrics such as centrality are extensively used to find highly interactive proteins, which are central to network integrity (Bhattacharyya and Chakrabarti, 2015). Such important proteins are excellent candidates for additional screening as drug targets.

The PPI may vary in space and time and have significant dependencies on the presence and utilization of host proteins. Thus, extensions examine interactions between parasite proteins and host proteins to identify the most important proteins for parasite success and survival. A recent review by Soyemi et al. (2018) summarizes current computational techniques to predict host-parasite protein interactions (HPPI). Given the scale, complexity, and temporal variability of these sets of interactions, networks serve as more than a visualization technique and can truly drive discovery.

Outlook

Networks, both as a representation of an underlying structure and of a dynamical system, showcase their power in the study of *var* genes and other interaction networks. Due to the complexity of *var* gene switching, open network research questions remain, including the question of what determines timing and changes in gene activation, particularly in vivo with the added selection pressures of an immune response. Although little in vivo data is currently available for validation, the analysis of synthetically generated data may still lead to an understanding of which processes, i.e., activation or de-activation, dominate during an infection. This will have consequences for determining and ultimately intervening on the severity and length of infection. However, there is an urgent need for the development of techniques that sufficiently differentiate the underlying selective processes from limited real-world or synthetic data. In the context of metabolic and PPI networks, there is an opportunity to move beyond simple visual representations and use network analyses to decipher which proteins (or nodes) are most important.

Other Applications of Networks

Network analysis has also been used as a tool in other areas of malaria research focusing on within-host processes. In particular, structural networks themselves are an essential basis for connecting and visualizing these sets of interactions. A major goal of these endeavors is the identification and validation of therapeutic targets, often through in silico knockouts.

Biophysical Networks

A purely mechanistic application of networks also arise in biophysical models of infected red blood cell dynamics (Imai et al., 2010; Lai et al., 2015; Ye et al., 2013). Within the membrane, proteins are described by interacting masses attached by springs. As the parasites develop intracellularly in red blood cells, more proteins are trafficked to the surface and alter the deformability of these infected red blood cells. These models demonstrate the biophysical reasoning for the known stiffening of red blood cells as parasites develop from the ring stage to the trophozoite and finally to become schizonts. Incorporation of these biophsyical constraints into a multi-scale model of blood flow indicate a need to account for red blood cell shape and deformability, especially in smaller blood vessels (Pan et al., 2011).

Networks as a Logical Filter

Understanding within-host dynamics of malaria parasites can be further complicated by the fact that, longitudinally, repeated infections in the same individual may be caused by reinfection, recrudescence, or, in the case of *P. vivax*, relapse. Work by Taylor et al. (2018) employs networks as a filter, to distinguish between these scenarios. Under the assumptions that (i) reinfection with the same parasite is impossible (while allowing reinfection with parasites which are identical-by-state), and (ii) recrudescent parasites must be present in the immediately preceding infection, certain network motifs are compatible with different scenarios.

The motifs considered include only those that are viable in terms of shared relationships. For example, it is not possible for two parasites, *A* and *B*, to be identical clones, but have a third parasite *C* be an identical clone of *A* but only a sibling with some fractional overlap of *B*. Utilizing network motifs as filtering tool for understanding messy and incomplete data, such as is common with inhost parasite data, has the potential to reduce the space of possibilities.

Conclusion

In this review, we have focused on the past and recent use of networks in the study of the malaria-causing parasites in the genus *Plasmodium*. We discussed three main topics which highlight both the flexibility of the network science toolkit in two key ways. First, the networks discussed in this review are defined differently from application to application, including networks where nodes are genetic loci, whole genomes, geographical locations, and metabolic pathways. Second, the subsequent analyses of those networks have also varied widely, from large-scale structural inference problems like the detection of communities, to dynamics problems like the identification of the most epidemiologically important nodes. In sum, complex networks provide a flexible data structure in combination with a growing set of analysis techniques, explaining much of their wide application in the context of malaria.

Many of malaria's puzzles are particularly well suited for network analyses. The complexity of interactions and dynamics exhibited by *Plasmodium* spp., particularly in the context of antigenic diversity, is staggeringly difficult to disentangle without network techniques. In this review, we particularly highlighted three applications: (i) the sequence-based relationship of highly polymorphic *var* genes or domains, which are important for immune evasion; (ii) the evolution of parasite genomes, as described by *var* gene overlap, through space and time; and (iii) the dynamic changes of *var* gene expression across the course of parasite generations in individual infections. Looking ahead, we fully expect that the network-based innovations in these and other areas are likely to produce more network-based advances.

Networks are not, however, a panacea for general problems with study design, biases, or undersampling. Particular care will have to be taken to understand how sampling biases may creep into network analyses. Research must also work to understand the key differences between network null models, like the configuration model, and the more generic null models that could be appropriate in particular analyses, such as label randomization. Like any other statistical or analytical tool, flexibility in definition and analysis do not imply that networks are one-size-fits-all or that the biological problem is readily captured by networks.

On a promising note, the increasingly wide application of network analyses in studies of *Plasmodium* spp. is likely to be further driven by continuing advances in network science more broadly. A new generation of complex networks research provides more and more statistical guarantees and analyses, augmenting the previous generation of asymptotically correct results. This turn toward statistical rigor will make networks more useful and even more reliable than in the past. Multilayer networks, in particular, are becoming far more developed, following big sets of advances in multilayer ranking, community detection, and centralities. Nevertheless, even without significant mathematical complexity, the creative use of network representations can change our understanding of a system, and streamlined analysis and visualization packages make the use of complex networks in the analysis of malaria parasites easier than ever.

Acknowledgments

The authors would like to thank Aimee Taylor and Shai Pilosof for feedback on an early version of the manuscript.

References

Albrecht L, Castineiras C, Carvalho BO, Ladeia-Andrade S, Santos da Silva N, Hoffmann EH, dalla Martha RC, Costa FT, and Wunderlich G (2010) The South American *Plasmodium falciparum* var gene repertoire is limited, highly shared and possibly lacks several antigenic types. *Gene* 453(1–2): 37–44.

Artzy-Randrup Y, Rorick MM, Day K, Chen D, Dobson AP, and Pascual M (2012) Population structuring of multicopy, antigen-encoding genes in *Plasmodium falciparum*. *eLife* 1: e00093.

Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, McVean GAV, and Day KP (2007) Population genomics of the immune evasion (var) genes of *Plasmodium falciparum*. *PLoS Pathogens* 3(3): e34.

Bei AK, Diouf A, Miura K, Larremore DB, Ribacke U, Tullo G, Moss EL, Neafsey DE, Daniels RF, Zeituni AE, et al. (2015) Immune characterization of *Plasmodium falciparum* parasites with a shared genetic signature in a region of decreasing transmission. *Infection and Immunity* 83(1): 276–285.

Bhattacharyya M and Chakrabarti S (2015) Identification of important interacting proteins (IIPs) in *Plasmodium falciparum* using large-scale interaction network analysis and in-silico knock-out studies. *Malaria Journal* 14(1): 70.

Buckee CO and Recker M (2012) Evolution of the multi-domain structures of virulence genes in the human malaria parasite, *Plasmodium falciparum*. *PLoS Computational Biology* 8(4): e1002451.

Buckee CO, Bull PC, and Gupta S (2009) Inferring malaria parasite population structure from serological networks. *Proceedings of the Royal Society of London B: Biological Sciences* 276(1656): 477–485.

Bull PC, Kyes S, Buckee CO, Montgomery J, Kortok MM, Newbold CI, and Marsh K (2007) An approach to classifying sequence tags sampled from *Plasmodium falciparum* var genes. *Molecular and Biochemical Parasitology* 154(1): 98.

Bull PC, Buckee CO, Kyes S, Kortok MM, Thathy V, Guyah B, Stoute JA, Newbold CI, and Marsh K (2008) *Plasmodium falciparum* antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. *Molecular Microbiology* 68(6): 1519–1534.

Chang H-H (2019) Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. *Elife* 8: e43481.

Childs LM and Buckee CO (2015) Dissecting the determinants of malaria chronicity: Why within-host models struggle to reproduce infection dynamics. *Journal of the Royal Society Interface* 12(104): 20141379.
Chowdhury P, Sen S, Kanjilal SD, and Sengupta S (2018) Genetic structure of two erythrocyte binding antigens of *Plasmodium falciparum* reveals a contrasting pattern of selection. *Infection, Genetics and Evolution* 57: 64–74.
Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, and Kwiatkowski D (2014) Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of var genes during mitosis. *PLoS Genetics* 10(12): e1004812.
Day KP, Artzy-Randrup Y, Tiedje KE, Rougeron V, Chen DS, Rask TS, Rorick MM, Migot-Nabias F, Deloron P, Luty AJF, et al. (2017) Evidence of strain structure in *Plasmodium falciparum* var gene repertoires in children from Gabon, West Africa. *Proceedings of the National Academy of Sciences* 114(20): E4103–E4111.
Deitsch KW and Dzikowski R (2017) Variant gene expression and antigenic variation by malaria parasites. *Annual Review of Microbiology* 71: 625–641.
Dzikowski R, Frank M, and Deitsch K (2006) Mutually exclusive expression of virulence genes by malaria parasites is regulated independently of antigen production. *PLoS Pathogens* 2(3): e22.
Fatumo S, Plaimas K, Adebiyi E, and König R (2011) Comparing metabolic network models based on genomic and automatically inferred enzyme information from plasmodium and its human host to define drug targets in silico. *Infection, Genetics and Evolution* 11(4): 708–715.
Fatumo S, Adebiyi M, and Adebiyi E (2013) In silico models for drug resistance. In: *In Silico Models for Drug Discovery*, pp. 39–65. Springer.
Forster P, Torroni A, Renfrew C, and Röhl A (2001) Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Molecular Biology and Evolution* 18(10): 1864–1881.
Fosdick BK, Larremore DB, Nishimura J, and Ugander J (2018) Configuring random graph models with fixed degree sequences. *SIAM Review* 60(2): 315–355.
Frank M, Dzikowski R, Amulic B, and Deitsch K (2007) Variable switching rates of malaria virulence genes are associated with chromosomal position. *Molecular Microbiology* 64(6): 1486–1498.
Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, and Scherf A (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of p. falciparum. *Nature* 407(6807): 1018.
Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906): 498.
Ginsburg H (2009) Caveat emptor: Limitations of the automated reconstruction of metabolic pathways in plasmodium. *Trends in Parasitology* 25(1): 37–43.
Gog JR and Grenfell BT (2002) Dynamics and selection of manystrain pathogens. *Proceedings of the National Academy of Sciences* 99(26): 17209–17214.
Gupta S and Day KP (1994) A strain theory of malaria transmission. *Parasitology Today* 10(12): 476–481.
Gupta S, Maiden MCJ, Feavers IM, Nee S, May RM, and Anderson RM (1996) The maintenance of strain structure in populations of recombining infectious agents. *Nature Medicine* 2(4): 437.
Gupta S, Ferguson N, and Anderson R (1998) Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* 280(5365): 912–915.
Hase T, Niimura Y, and Tanaka H (2010) Difference in gene duplicability may explain the difference in overall structure of protein-protein interaction networks among eukaryotes. *BMC Evolutionary Biology* 10(1): 358.
He Q, Pilosof S, Tiedje KE, Ruybal-Pesantez S, Artzy-Randrup Y, Baskerville EB, Day KP, and Pascual M (2018) Networks of genetic similarity reveal non-neutral processes shape strain structure in *Plasmodium falciparum*. *Nature Communications* 9(1): 1817.
Henden L, Lee S, Mueller I, Barry A, and Bahlo M (2018) Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genetics* 14(5): e1007279.
Horrocks P, Pinches R, Christodoulou Z, Kyes SA, and Newbold CI (2004) Variable var transition rates underlie antigenic variation in malaria. *Proceedings of the National Academy of Sciences* 101(30): 11129–11134.
Imai Y, Kondo H, Ishikawa T, Teck Lim C, and Yamaguchi T (2010) Modeling of hemodynamics arising from malaria infection. *Journal of Biomechanics* 43(7): 1386–1393.
Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahrén D, Tsoka S, Darzentas N, Kunin V, and López-Bigas N (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* 33(19): 6083–6089.
Kraemer SM and Smith JD (2003) Evidence for the importance of genetic structuring to the structural and functional specialization of the *Plasmodium falciparum* var gene family. *Molecular Microbiology* 50(5): 1527–1538.
Kyes S, Horrocks P, and Newbold C (2001) Antigenic variation at the infected red cell surface in malaria. *Annual Review of Microbiology* 55(1): 673–707.
Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, Thera MA, Koné AK, Doumbo OK, Plowe CV, and Rowe JA (2006) Differential var gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia. *Molecular and Biochemical Parasitology* 150(2): 211–218.
LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, et al. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438(7064): 103.
Lai L, Xu X, Lim CT, and Cao J (2015) Stiffening of red blood cells induced by cytoskeleton disorders: A joint theory-experiment study. *Biophysical Journal* 109(11): 2287–2294.
Larremore DB (2019) Bayes-optimal estimation of overlap between populations of fixed size. *PLoS Computational Biology* 15(3): e1006898.
Larremore DB, Clauset A, and Buckee CO (2013) A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Computational Biology* 9(10): e1003268.
Larremore DB, Sundararaman SA, Liu W, Proto WR, Clauset A, Loy DE, Speede S, Plenderleith LJ, Sharp PM, Hahn BH, et al. (2015) Ape parasite origins of human malaria virulence genes. *Nature Communications* 6: 8368.
Lavstsen T, Salanti A, Jensen ATR, Arnot DE, and Theander TG (2003) Sub-grouping of *Plasmodium falciparum* 3d7 var genes based on sequence analysis of coding and non-coding regions. *Malaria Journal* 2(1): 27.
Newman M (2018) *Networks*. Oxford University Press.
Noble R and Recker M (2012) A statistically rigorous method for determining antigenic switching networks. *PLoS One* 7(6): e39335.
Noble R, Christodoulou Z, Kyes S, Pinches R, Newbold CI, and Recker M (2013) The antigenic switching network of *Plasmodium falciparum* and its implications for the immuno-epidemiology of malaria. *eLife* 2: e01074.
Otto TD, Gilabert A, Crellen T, Böhme U, Arnathau C, Sanders M, Oyola SO, Okouga AP, Boundenga L, Willaume E, et al. (2018) Genomes of all known members of a plasmodium subgenus reveal paths to virulent human malaria. *Nature Microbiology* 3(6): 687.
Pan W, Fedosov DA, Caswell B, and Karniadakis GE (2011) Predicting dynamics and rheology of blood flow: A comparative study of multiscale and low-dimensional models of red blood cells. *Microvascular Research* 82(2): 163–170.
Pilosof S, He Q, Tiedje KE, Ruybal-Pesantez S, Day KP, and Pascual M (2019) Competition for hosts modulates vast antigenic diversity to generate persistent strain structure in *Plasmodium falciparum*. *PLoS Biology* 17(6): e3000336.
Pinney JW, Shirley MW, McConkey GA, and Westhead DR (2005) metaSHARK: Software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Research* 33(4): 1399–1409.
Pirahmadi S, Zakeri S, Mehrizi AA, and Djadid ND (2018) Analysis of genetic diversity and population structure of gene encoding cell-traversal protein for ookinetes and sporozoites (CelTOS) vaccine candidate antigen in global *Plasmodium falciparum* populations. *Infection, Genetics and Evolution* 59: 113–125.
Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, and Lavstsen T (2010) *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes–divide and conquer. *PLoS Computational Biology* 6(9): e1000933.
Recker M, Buckee CO, Serazin A, Kyes S, Pinches R, Christodoulou Z, Springer AL, Gupta S, and Newbold CI (2011) Antigenic variation in *Plasmodium falciparum* malaria involves a highly structured switching pattern. *PLoS Pathogens* 7(3): e1001306.
Rorick MM, Baskerville EB, Rask TS, Day KP, and Pascual M (2018) Identifying functional groups among the diverse, recombining antigenic var genes of the malaria parasite *Plasmodium falciparum* from a local community in Ghana. *PLoS Computational Biology* 14(6): e1006174.

Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, Pouvelle B, Gysin J, and Lanzer M (1998) Antigenic variation in malaria: In situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in *Plasmodium falciparum*. *The EMBO Journal* 17(18): 5418–5426.

Scherf A, Lopez-Rubio JJ, and Riviere L (2008) Antigenic variation in *Plasmodium falciparum*. *Annual Review of Microbiology* 62: 445–470.

Smith JD, Subramanian G, Gamain B, Baruch DI, and Miller LH (2000) Classification of adhesive domains in the *Plasmodium falciparum* erythrocyte membrane protein 1 family. *Molecular and Biochemical Parasitology* 110(2): 293–310.

Soyemi J, Isewon I, Oyelade J, and Adebiyi E (2018) Inter-species/host-parasite protein interaction predictions reviewed. *Current Bioinformatics* 13(4): 396–406.

Suthram S, Sittler T, and Ideker T (2005) The plasmodium protein network diverges from those of other eukaryotes. *Nature* 438(7064): 108.

Tatem AJ and Smith DL (2010) International population movements and regional *Plasmodium falciparum* malaria elimination strategies. *Proceedings of the National Academy of Sciences* 107(27): 12222–12227.

Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawat K, Pyae Phyo A, Nosten F, Neafsey DE, and Buckee CO (2017) Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. *PLoS Genetics* 13(10): e1007065.

Taylor AR, Watson JA, Chu CS, Puaprasert K, Duanguppama J, Day NPJ, Nosten F, Neafsey DE, Buckee CO, Imwong M, et al. (2018) Estimating the probable cause of recurrence in *Plasmodium vivax* malaria: Relapse, reinfection or recrudescence? *BioRxiv*. page 505594.

Tessema SK, Monk SL, Schultz MB, Tavul L, Reeder JC, Siba PM, Mueller I, and Barry AE (2015) Phylogeography of var gene repertoires reveals fine-scale geospatial clustering of *Plasmodium falciparum* populations in a highly endemic area. *Molecular Ecology* 24(2): 484–497.

Trimnell AR, Kraemer SM, Mukherjee S, Phippard DJ, Janes JH, Flamoe E, Su X-Z, Awadalla P, and Smith JD (2006) Global genetic diversity and evolution of var genes associated with placental and severe childhood malaria. *Molecular and Biochemical Parasitology* 148(2): 169–180.

Viebig NK, Gamain B, Scheidig C, Lépolard C, Przyborski J, Lanzer M, Gysin J, and Scherf A (2005) A single member of the *Plasmodium falciparum* var multigene family determines cytoadhesion to the placental receptor chondroitin sulphate A. *EMBO Reports* 6(8): 775–781.

Wapman K and Larremore D (2019) webweb: A tool for creating, displaying, and sharing interactive network visualizations on the web. *Journal of Open Source Software* 4(40): 1458.. ISSN 2475-9066. 8. URL https://doi.org/10.21105/joss.01458.

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, and Buckee CO (2012) Quantifying the impact of human mobility on malaria. *Science* 338(6104): 267–270.

Ye T, Phan-Thien N, Khoo BC, and Lim CT (2013) Stretching and relaxation of malaria-infected red blood cells. *Biophysical Journal* 105(5): 1103–1109.

Yeh I, Hanekamp T, Tsoka S, Karp PD, and Altman RB (2004) Computational analysis of *Plasmodium falciparum* metabolism: Organizing genomic information to facilitate drug discovery. *Genome Research* 14(5): 917–924.

Zilversmit MM, Chase EK, Chen DS, Awadalla P, Day KP, and McVean G (2013) Hypervariable antigen genes in malaria have ancient roots. *BMC Evolutionary Biology* 13(1): 110.