PopNet: A Markov Clustering Approach to Study Population Genetic Structure

Javi Zhang, Asis Khan, Andrea Kennard, Michael E. Grigg, and John Parkinson

1Department of Biochemistry, University of Toronto, Toronto, ON, Canada
2Program in Molecular Medicine, Hospital for Sick Children, Toronto, ON, Canada
3Molecular Parasitology Section, Laboratory of Parasitic Diseases, NIAID, National Institutes of Health, Bethesda, MD
4Departments of Computer Science and Molecular Genetics, University of Toronto, Toronto, ON, Canada

*Corresponding authors: E-mails: john.parkinson@utoronto.ca; griggm@niaid.nih.gov.
Associate Editor: Hideki Innan

Abstract

With the advent of low cost, high-throughput genome sequencing technology, population genomic data sets are being generated for hundreds of species of pathogenic, industrial, and agricultural importance. The challenge is how best to analyze and visualize these complex data sets to yield intuitive representations capable of capturing complex evolutionary relationships. Here we present PopNet, a novel computational method that identifies regions of shared ancestry in the chromosomes of related strains through clustering patterns of genetic variation. These relationships are subsequently visualized within a network by a novel implementation of chromosome painting. We apply PopNet to three diverse populations that feature differential rates of recombination and demonstrate its ability to capture evolutionary relationships as well as associate traits to specific loci. Compared with existing tools, PopNet provides substantial advances by both removing the need to predefine a single reference genome that can bias interpretation of population structure, as well as its ability to visualize multiple evolutionary relationships, such as recombination events and shared ancestry, across hundreds of strains.

Key words: population genomics, single nucleotide polymorphisms, network visualization, software, chromosome painting.

Introduction

Population genetic structure, represented by patterns of inheritance, is dynamic and depends upon multiple forces including mode of reproduction and selection pressures. For example, recombination during sexual reproduction results in admixed populations in which different sections of chromosomes exhibit alternate patterns of inheritance. Traditionally such patterns have relied on analysis of individual or limited sets of marker genes to infer genetic hybridization (Grigg et al. 2001). The advent of chip-based technologies, exploiting common sequence variants such as single nucleotide polymorphisms (SNPs) resulted in significant gains in resolution and throughput (Evans et al. 2015). More recently, with decreasing costs and the ability to provide information at single base resolution, next generation sequencing (NGS) is finding wider application in population based studies focused on areas such as cancer, drug resistance, and agriculture (Singh and Singh 2008; Minot et al. 2012; Miotto et al. 2013; Fu et al. 2016; Johann et al. 2016; Pang et al. 2016).

As new technologies have emerged, parallel efforts to develop the algorithms and tools to best exploit these data sets are required. For example, data generated from SNP chips have resulted in the development of methods devoted to the identification of haplotype blocks (Haploblocks—sets of SNPs exhibiting strong linkage disequilibrium), in order to better infer patterns of recombination within populations (International HapMap 2003; Minot et al. 2012). Haploblocks can be seen as genetic islands or blocks of shared ancestry within the whole genome context, where each haploblock serves as a marker that can be used to define ancestry as well as demarcate genes associated with distinct phenotypes (Minot et al. 2012; Guan et al. 2013). Moreover, in admixed individuals whose genomes are inherited from members of different subpopulations, haploblocks may be used to identify regions that have been exposed to different selective pressures and evolution histories (Khanyile et al. 2015). Since haploblocks are traditionally defined on the basis of linkage disequilibrium, they typically capture only a limited fraction of the genome (Lawson et al. 2012).

Recently, chromosome painting has emerged as an effective technique to examine patterns of inheritance across an entire genome (Lawson et al. 2012; Minot et al. 2012). For example, in a recent study of 62 strains of Toxoplasma gondii, chromosomes are divided into discrete sections containing a fixed number (e.g., 1,000) of SNPs on which local admixture analyses are performed (Lorenzi et al. 2016). Tools such as STRUCTURE and fineSTRUCTURE (Pritchard et al. 2000; Lawson et al. 2012) can then be used to group sections sharing similar ancestry to visualize the fraction of the genome associated with distinct patterns of inheritance. Subsequent
refinement of these tools allows such patterns to be visualized in the context of the chromosomal region in which they occur, allowing the identification of defined regions of coinheritance (Yahara et al. 2013; Lorenzi et al. 2016). However, such visualizations provide limited insights into the origins of these shared patterns and the population structure as a whole and rely on imprecise methods to estimate the number of ancestries present. Other methods, such as network based approaches are useful in inferring recombination events within the context of a global population structure (Huson and Bryant 2006). For example, Neighbor-Net represents all individuals within a split network, where shared ancestry is represented by the reticulation of edges providing insight into recombination events that occurred between their ancestors (Bryant and Moulton 2004).

Compared with the traditional rooted phylogenetic tree, the additional dimension provided by Neighbor-Net allows for a much more accurate depiction of the relationship between sub populations. Yet, while the network approach is effective at defining global structure, it provides only limited insights into how alleles associated with individual genes or local chromosome regions are distributed within a population. Furthermore, it yields no positional information concerning recombination events that result in admixed chromosomes where different lineages may exhibit regions of shared ancestry.

In an attempt to build on graph based approaches such as Neighbor-Net to infer population structure, we present a novel computational platform, termed PopNet, which integrates chromosome painting within a network structure to generate intuitive representations of genome-scale population data. Where previous implementations of chromosome painting have required an independent analysis, such as cross validation (Lorenzi et al. 2016), to define ancestral populations, PopNet incorporates an agnostic clustering approach to define the set of possible ancestries as clades. The results are visualized positionally as a network with the degree and delineation of the shared ancestry directly incorporated into the graph. We demonstrate the effectiveness of PopNet through its application to three diverse populations: *Saccharomyces cerevisiae*, *Toxoplasma gondii*, and *Plasmodium falciparum*, showing that *S. cerevisiae* lineages form around habitat or function; North American *T. gondii* families possess admixed genomes that share multiple regions of similarity, and Asian *P. falciparum* populations show a higher degree of genetic hybridization compared with their African counterparts.

**Results**

**PopNet Reveals Population Structure and Identifies Recombination Events through Clustering Chromosome Segments**

The PopNet pipeline uses patterns of SNP distributions, generated for example by the Genome Analysis Toolkit (GATK; McKenna et al. 2010) or the Nucmer alignment algorithm (Kurtz et al. 2004), to define and visualize population structure based on a series of graph clustering steps (fig. 1a; supplementary methods, Supplementary Material online). First, the algorithm divides the genome into a series of non-overlapping segments of user-defined length (e.g., 10 kb—see below). A SNP similarity matrix is then generated for each segment based on the number of alleles shared between each pair of strains and the Markov Clustering algorithm (MCL) (van Dongen 2000a) applied to identify groups of strains sharing similar patterns of alleles. The results of this “primary” clustering are then used to construct a new “global” similarity matrix, which defines the number of segments in which each pair of individuals are placed in the same cluster. This global matrix is subsequently clustered, also with MCL. The resulting clusters generated in this step define the total number of discrete groups of related strains (clades) in the population, each of which is represented by a unique color within the network visualization.

Visualization of SNP inheritance patterns is performed using a chromosome painting approach (Miotto et al. 2013; Lorenzi et al. 2016) in which each segment of a strain’s genome is colored according to its closest matching clade. Specifically, if a strain shares a cluster with at least N% of the members of another clade, the chromosome segment associated with that strain is assigned the color of that clade (supplementary fig. 1, Supplementary Material online). N is a tuneable parameter that impacts the number of chromosome “features” (defined as chromosome segments that share the same ancestry) identified (supplementary fig. 2a and b, Supplementary Material online). Here, we set N to 30 to provide a balance between the identification of too many features (increasing noise) and too few features (reducing signal). Where a strain shares a cluster with multiple clades, based on the prior definition, its clade relationship is defined through a “chain-extension” algorithm: in an initial step, the algorithm compiles a list of putative clades associated with each segment. Next, starting at the beginning of each chromosome, the algorithm identifies the longest consecutive chain of segments sharing ancestry with the same clade. During this step, we allow “gaps” in this chain to allow for strain-specific divergence due to, for example, genetic drift or sequencing artifacts. Default settings allow for one gap for every eight segments up to a maximum of five consecutive gaps. Although, we note that the number of allowed consecutive gaps has little impact on results (supplementary fig. 2c and d, Supplementary Material online), increasing the number of segments allowed per gap results in an increase in regions of shared ancestry and a concomitant decrease in granularity (a term used to describe cluster distributions, with low granularity indicating fewer clusters with more members, supplementary fig. 2e and f, Supplementary Material online). Once assigned to a specific clade, the chain of consecutive segments is assigned the color of that clade and the algorithm continues on the following chain. Finally, if the segment does not form part of a chain sharing ancestry with other clades, the segment assumes the color of the strains associated clade. For each chromosome, this procedure results in a linear representation of SNP inheritance patterns. Chromosomes are subsequently concatenated and circularized to form a ring (annulus) for each individual represented in the population.
**Fig. 1.** PopNet algorithm and benchmarking. (a) Schematic of the PopNet algorithm. PopNet requires a matrix of SNPs as input. A primary clustering step defines ancestries for each segment based on the number of shared SNPs within that segment. A secondary clustering step, based on the frequency of coclustering from the primary clustering step, is used to define clades. Clade memberships are then used to define colors which are used to define regions of shared ancestry across a chromosome. Chromosomes are then concatenated within an annulus to yield a network view of population structure. (b–d) Relationship of cluster numbers, clustering efficiency, and distribution of chromosome features to segment length, as well as the inflation (λ) and preinflation (π) parameters used by MCL in the primary clustering step. (e) Process of generating a simulated ancestry featuring rounds of recombination between individuals. (f) PopNet representation of a simulated data set. Node labels indicate ancestors (A1–4) and offspring (O1–11). Regions of shared ancestry with other members of the population are represented through colored segments indicating the clade sharing that segment.
Annuli are then used as nodes in a network where edges between nodes represent the frequency of two individuals being assigned to the same cluster (as defined by the initial clustering step). The color of the outer edge of the annulus, indicates which clade that individual belongs to, which is predefined as the clade that the individual shares a “tunable” percent majority of sequence similarity (i.e., blocks of shared SNPs).

To define segment clusters PopNet relies on three further parameters: two inflation parameters that impact cluster granularity used by the MCL algorithm, and the length of the chromosome segment used in the primary clustering step. Based on an initial analysis of 67 yeast genomes (see below), for the primary clustering step, we find that using inflation and preinflation parameters of 8 and 19, respectively, in addition to a segment length of 10,000 results in a reasonable trade-off between maximizing cluster granularity (and thereby reducing false positive cluster assignments), clustering efficiency—a metric related to the extent that resultant clusters capture edge weights between nodes where higher values indicate that more nodes with stronger connections to other nodes (i.e., more shared SNPs) are placed in the same cluster (van Dongen 2000b) and the number of chromosome features (figs. 1b–d and supplementary fig. 3a and b, Supplementary Material online).

For the global clustering step, we find that setting the inflation and preinflation parameters for MCL to 4 and 1.5, respectively, results in reasonable trade-offs between cluster granularity, clustering efficiency, and inter and intracluster distances (supplementary fig. 3c–f, Supplementary Material online).

To examine PopNet performance, we simulated a simple population of strains of a model haploid organism using defined rules of recombination and crossing (fig. 1e). Our model featured a founder population of four strains with genomes of size 21 Mb (a typical size for a single celled eukaryotic) divided into 14 1.5 Mb chromosomes. For each chromosome, we assigned one SNP for every 500 bp (3,000 SNPs/chromosome—consistent with previous population studies); alleles were assigned either as A, T, G, or C depending on the strain (i.e., the alleles for the first strain were all assigned as “A”). A simulated population was then generated through a series of “cross-over” events in which a strain is chosen at random and chromosome segments randomly replaced with equivalent segments from a second randomly selected strain at a rate of 1 every 5.2 Mb (equivalent to an average recombination rate of 52 kb/cM, a value half between those observed for populations of two related parasites—T. gondii and P. falciparum [Jiang et al. 2011]). To ensure chromosome features could be readily visualized, we used a simple model of segment replacement based on a Gaussian distribution, with a mean of 100 kb and sigma of 10 kb (as noted below, the average feature length for S. cerevisiae was found to be ~110 kb). Repeating this process 11 times generated a population of 15 strains, featuring a total of 54 recombination events. PopNet analysis of the model population revealed four clades, associated with the four progenitors, and correctly identified all 54 recombination events (fig. 1f). For example, strain O1 features five teal bands on a background of deep blue, representing contributions from strain A3 (teal) into A1 (deep blue). Each band corresponds to one recombination event between the two ancestors, except for chromosome XI, where two crossover events occurred within the same region resulting in a single band.

Although, this analysis demonstrates the ability of PopNet to capture admixture and gene flow between populations, it is important to note that for natural populations, parental strains may not be present. Hence, regions of common ancestry, as identified by PopNet should not be attributed to recombination between sampled individuals. In the following three sections, we apply PopNet to three diverse populations: Toxoplasma gondii, S. cerevisiae, and Plasmodium falciparum, to demonstrate how PopNet results in intuitive interpretations of population genomics data.

PopNet Reveals Population Structure and Extent of Admixture between Isolates of T. gondii from North American
The unicellular parasite, Toxoplasma gondii, is the leading cause of infectious retinitis globally and can be fatal if acquired during pregnancy or when immunosuppressed (Kaye 2011; Saadatnia and Golkar 2012). Genetic variation driven through the parasites sexual phase has resulted in a diverse set of strains, each featuring different virulence potentials, together with distinct host and tissue tropisms (Wendte et al. 2011; Pan et al. 2012). In North America and Europe, populations of T. gondii are thought to be largely clonal, dominated by three major types: Type I, Type II, and Type III (Lawson et al. 2012). A fourth family, largely infecting wild animals in North America, has been designated Type X or Haplogroup 12 (HG12) based on a limited set of genetic markers. This lineage is thought to represent the product of a cross between Type II and a unique lineage termed “v” (Miller et al. 2008; Khan et al. 2011). Recent analysis of 62 T. gondii genomes reveal mosaic chromosomes composed of conserved haploblocks shaped through multiple recombination events (Lorenzi et al. 2016). Clusters of genes associated with these haploblocks are known to influence transmission, host range, and pathogenesis. Here we investigated the utility of PopNet to reveal the genetic ancestry and population structure of 24 strains of T. gondii. In this analysis, we combined 17 previously sequenced strains with de novo genome data from two additional HG12 isolates (3142 and TGSKUNK), a new Type II isolate (TGOATUS21), three progeny from a cross between a Type II strain (ME49) and a type III strain (CTG), designated S22, S23 and S30 (Sibley et al. 1992). In addition, to ensure that our pipeline is consistent with previously generated genome data, we resequenced strain ARI (designated ARLMG).

We first examined the performance of standard population structure tools. As expected, by adopting whole genome sequence data instead of a limited set of genetic markers, Neighbor-Net (Bryant and Moulton 2004) does not support the HG designation (Lorenzi et al. 2016), with many isolates from the same haplogroup, appearing on separate branches (fig. 2a). However, while Neighbor-Net did distinguish bona
**Fig. 2.** Population structure of 20 strains of *Toxoplasma gondii*. (a) Neighbor-Net representation of *T. gondii* population indicating four previously defined major clades along with major haplogroups. (b) FineSTRUCTURE representation, revealing lack of resolution amongst Type II and haplogroup 12 strains, as well as the progeny of a cross between CTG and ME49. (c) (i) Chromosome painting representation of patterns of inheritance for ME49, compared with the type I strain GT1 and type II strain VEG. (ii) Chromosome painting representation of three progeny and their parents for three chromosomes. Cross-over events are indicated with black arrows (2 in strain S23 and 5 in strain S30—only 4 are shown here). Color indicates ancestral relationships of each chromosome segment (see inset). Black lines indicate absence of SNPs in that region. (d) PopNet
fide Type II and III strains from those that possess hybrid genomes, the evolutionary relationships of the progeny from the cross between ME49 and CTG is not clear. The phylogenetic tree and matrix generated by fineSTRUCTURE (Lawson et al. 2012) revealed clear groupings for many strains and further resolved the shared ancestry across groups. However, although fineSTRUCTURE can theoretically infer the number of populations (clades) and capture detailed patterns of inheritance at the chromosome level, such features are not provided through the standard visualization interface. As a result, evolutionary relationships between hybrids, such as the progeny of the cross between ME49 and CTG, are not readily resolved (fig. 2b). In contrast, by integrating chromosome painting within a network visualization framework, PopNet builds on these tools to yield a more intuitive understanding of patterns of inheritance.

We first applied PopNet to reveal the extent of shared ancestry between strain ME49 and the Type I and III strains, GT1 and VEG (fig. 2c(i)). Consistent with previous studies (Minot et al. 2012; Lorenzi et al. 2016), PopNet reveals a mosaic structure across chromosomes, as indicated, for example, by cyan blocks indicating shared ancestry between strains ME49 and VEG on chromosomes Ib, III, VI, VIIb, VIII, IX, XI, and XII. However, unlike previous studies, we identify many more blocks of shared ancestry. Applied to the progeny of CTG and ME49, PopNet reveals the ancestry of each chromosome (fig. 2c(ii)). Amongst these patterns we identified seven recombination events, including a double cross-over in chromosome X of S30, a rate consistent with previous observations (Khan et al. 2005).

Expanding the analysis to our 24 strains, PopNet recapitulates the Type I, Type II, Type III, and HG12 lineages and reveals the influence of local admixture on population structure (fig. 2d). As with the linear representations, the network view clearly delineates chromosomal regions of shared ancestry. For example, the Type III strains, VEG and M7741 (blue) share many blocks of sequence with Type II strains (green) as exemplified by corresponding segments of the other groups color. In comparison, the Type I strains GT1 and RH88 appear more isolated, with limited regions of shared ancestry with Type II strains on chromosome IV and Type III strains on chromosomes VIII and XII. Furthermore, unlike previous studies, PopNet captures the divergent nature of chromosome la within the type I strains, GT1 and RH88, as well as CAST as indicated by multiple colored regions for this chromosome.

Analysis of the newly sequenced North American isolates, confirms TgSKUNK, ARI and 3142 as members of HG12, and TGGOATUS21 as related to Type II strains ME49 and PRU. We are reassured that PopNet generates identical patterns for ARI and ARLMG, indicating the robustness of our sequencing and analysis pipeline. Within HG12, TgSKUNK, and RAY share more ancestral blocks with Type II strains (28.7% and 29.1% of their chromosomes, respectively) than ARI (15.3%) and 3142 (16.8%). The positional inheritance of these blocks supports a model whereby ARI and 3142 represent related sister progeny that possess different Type II recombination blocks than TgSKUNK and RAY. Further sampling within HG12 strains will yield a clearer view of the evolutionary events shaping the genomes within this group of strains that circulate among wildlife in North America.

Within each group, strain-specific patterns are also observed consistent with multiple independent introgressions of sequence blocks bearing distinct ancestries. For example, the progeny of the strain II by III cross, each feature unique patterns of inheritance, whereas P89 shares more regions with Type I members than VEG, CTG, and M7741. Even relatively subtle contributions from other clades are captured. For example, M7741, which was previously considered a close relative of VEG and CTG, shows introgressions from a Type I strain in chromosomes Ib, VIIa, and X, suggesting that it represents an admixture cross between VEG/CTG and a Type I strain. A finding not captured using NeighborNet or fineSTRUCTURE. The ability to identify such regions is important as TGSHUS28 is known to contain both Type I and Type III markers, and exhibits Type I-like virulence in mice (Dubey et al. 2008). Here, we show that TGSHUS28 forms a clade with Type III strains but has inherited chromosomes II, VI, VIIa, X, and parts of chromosomes VIII, IX, XI, and XII from an ancestor of Type I strains.

Through the analysis of 24 strains of *T. gondii*, this analysis shows how PopNet yields a detailed view of the extent to which recombination events between the ancestors of the present day haplogroups have shaped population structure. Furthermore, through identifying such events, PopNet paves the way for future studies aimed at exploiting haploblock data sets to identify SNPs associated with altered patterns of virulence.

**PopNet Associates Phenotypic Traits with Distinct S. cerevisiae Subpopulations**

Large scale population studies of domesticated and wild yeasts reveal their segregation into clades based on geography and/or function (Liti et al. 2009; Liti 2015). However, while outcrossing is rare, detailed knowledge concerning the extent...
and functional impact of gene flow between strains at the population level is limited. Here we applied PopNet to analyze the population structure of 67 strains of S. cerevisiae, including clinical, laboratory, industrial, and environmental isolates, and exploit biochemical data to gain insights into genetic association of function (Cherry et al. 2012; Bergstrom et al. 2014).

PopNet defines eight clades that display reduced shared ancestry compared with the T. gondii population (fig. 3a). For example, most members of clades I, II, and III possess few regions of common ancestry with other groups, with an average of 93.5% of their chromosomes unique to each clade. Only two members, Y55 (clade I) and UWOPS87_2421 (clade II), reveal visible sections shared with clade IV. Conversely, clade V shares considerable ancestry with clades IV and VI (14% and 22% of their chromosomes, respectively), whereas members of clades IV, VII, and VIII display more mosaic chromosomes. Such findings suggest that clades I, II, and III are largely genetically isolated, whereas clades V and VI, form a distinct population with shared ancestry. This is consistent with a previous study, in which the Malaysian strains of clade III, were found to be reproductively isolated from other lineages (Cubillos et al. 2011). The remaining clades (IV, VII, and VIII) represent genetic admixtures resulting from multiple recombination events across clades. Consistent with previous studies, clades tend to be associated with both function and geographical source of strain (fig. 3b and c). For example, clades II and III consist of environmental isolates isolated from plants with members deriving from Hawaii and Bahamas (clade II) or Malaysia and Africa (clade III). Clades I, V, and VI largely comprise isolates used in fermentation, with those from clades V and VI originating from Asia and associated with the production of sake (with the exception of the bio-ethanol producing strain ZTW1), whereas those in clade I originate from Africa and France and are associated with wine, palm wine, and ginger beer. Members of clade IV are more diverse and consist of brewing, baking, environmental, and clinical isolates.

Given overlap in function between members of several clades, we explored overlap in specific phenotypic traits previously generated across a range of experimental conditions (Liti et al. 2009) for a subset of 34 strains (fig. 3d). Consistent with their reproductive isolation, members of clade III displayed greatest phenotypic similarity between themselves than with other strains (mean intraclide Pearson correlation coefficient (PCC) = 0.970 ± 0.005 v mean interclade PCC = 0.488 ± 0.105, P = 2.3 × 10−63 Students t-test), whereas clades I, II, and IV also showed significantly distinct similarity in phenotypes (clade I: mean intraclide PCC = 0.551 ± 0.079 v mean interclade PCC = 0.472 ± 0.117, P = 0.022; clade II: mean intraclide PCC = 0.709 ± 0.133 v mean interclade PCC = 0.456 ± 0.103, P = 1 × 10−5; clade IV: mean intraclide PCC = 0.594 ± 0.159 v mean interclade PCC = 0.477 ± 0.119, P = 1 × 10−14). Clade V did not display significant phenotypic similarity, reflected by a lack of coclustering (fig. 3d). For clade II, we note that UWOPS87_2421 displays a markedly different profile to other members of its clade (fig. 3d). For example, in the presence of 450 mM aminotriazole, UWOPS87_2421 displays a lower relative growth rate (logarithmic strain coefficient (LSC) = −0.004), compared with other clade II members (LSC = 0.834 ± 0.037), that was similar to members of clade IV, which display a mean LSC of −0.012 ± 0.626. PopNet analysis identified 14 genomic regions in which strain UWOPS87_2421 displayed an anomalous clustering pattern to other clade II strains, resembling instead, members of clade IV with similar LSC scores (i.e., DBVPG6040, BC187, and L_1374). Among the 65 genes associated with these regions was the ATR1 locus, which encodes a drug efflux pump providing resistance against aminotriazole (Kanazawa et al. 1998). UWOPS87_2421 clustering with clade IV members at this locus suggests that the resistance allele may have been lost during a recombination event with an ancestor of this clade. In addition to confirming the segregation of domestic and wild type yeast populations into clades based on habitat and function, respectively (Liti et al. 2009; Wang et al. 2012), this analysis demonstrates the potential of PopNet to predict novel genotype–phenotype relationships; anomalous patterns of clustering of genomic regions that coincide with shared phenotypic characteristics, can be used to prioritize genes for detailed functional characterization.

PopNet Reveals Geographical Boundaries have a Major Impact on Genetic Hybridization and the Spread of Artemisinin Resistance in P. falciparum

The parasite Plasmodium falciparum is the causative agent of malaria affecting an estimated 214 million people worldwide (World Health Organization 2015). Studies of P. falciparum population structure are currently being used to guide the development and implementation of control and elimination programs (Mita and Tanabe 2012; Mioto et al. 2013; Yin et al. 2013). Unlike T. gondii which can bypass its sexual cycle for transmission, P. falciparum has an obligate sexual cycle for its transmission between human hosts. Consequently, the recombination rate of P. falciparum (~10 kb/cM) is much higher than the estimated rate for T. gondii (~100 kb/cM) (Jiang et al. 2011). What is debated is the extent to which self-mating versus outcrossing impacts the population genetic structure. Here, we analyzed 177 isolates of P. falciparum to examine the performance of PopNet on a frequently recombining population.

Applying PopNet, the 177 isolates were grouped into 18 clades, with two major clades (I and II) consisting of 89 and 38 members, respectively, and 16 minor groups consisting of between two and five members (fig. 4a). Integration of the geographical source of each strain revealed considerable overlap between clade membership and geographical origin (fig. 4b). For example, clade I contains 61 of 69 African isolates, 14 out of 20 isolates from Northern Cambodian province of Ratanakiri, and ten isolates from other Asian regions. Clade II is composed of Asian isolates from Cambodia and Thailand, whereas members of the smaller clades were largely associated with a single region.

Unlike T. gondii strains, chromosome representation of isolates from both major groups reveal fewer large contiguous
FIG. 3. PopNet representation and analysis of 67 strains of Saccharomyces cerevisiae. (a) Network visualization of population structure. Nodes, representing individual strains, are colored according to defined clades. Edges between nodes indicate frequency of coclustering during primary clustering step. (b) Relationship between strain function and clade membership. (c) Relationship between geographical source of strain and clade membership. (d) Phenotype comparisons across 34 selected strains. Growth phenotypes were obtained from a previous population genomics study and feature: growth lag (adaptation time, h), growth rate (doubling time, h) and growth efficiency (change in cell density) under a variety of conditions (Liti et al. 2009). For each condition, values (x) were normalized (nx) to the minimum value (y) for each condition: \(nx = (x - y)/(X - y)\). Hierarchical clustering (absolute correlation/average linkage) was performed using Cluster 3.0. Color indicates ratio of phenotype of the indicated strain relative to reference strain BY4741. (e) Clade relationships for the 35 strains depicted in (d) for the chromosome region around the ATR1 locus. Note how strain UWOPS87_2421, which differs in growth rate in the presence of 450 mM 1,2,4-Aminotriazole relative to other clade II strains also differs in clade assignment for the region containing the ATR1 locus.
**Fig. 4.** PopNet representation and analysis of 177 strains of *Plasmodium falciparum*. (a) Global network view—nodes, representing individual strains, are colored according to defined clades. Edges between nodes indicate frequency of coclustering during primary clustering step. (b) Geographical source of strain isolates. Nodes, placed as in (A) are colored according to region of origin (see inset). Node borders are colored according to clade membership defined in (a). (c) Subset of strains from Pursat and Ratanakiri regions illustrating differences in artemisinin resistance. Node color indicates half-life clearance times (see inset).
blocks of shared ancestry, featuring instead many smaller regions reflecting the frequent recombination events that occur in natural \( P. falciparum \) populations (Conway et al. 1999). Consequently, new genomic elements are rapidly disseminated through the population, resulting in clade members possessing a common genomic background introgressed with short (< 10 kb) sections obtained through crosses with members of other clades. Comparing between continents, the African population appears largely panmictic, with isolates from different countries belonging to the same population. However, also present are three relatively small genetically distinct subpopulations (clades XVI–XVIII). Members of each of these clades share only limited regions (~12% of their chromosomes) with other clades. On the other hand, whereas further small clades are identified in Asia, they exhibit frequent interactions with other clade members, sharing from 26% to 51% of their chromosomes with other clades. Although the characteristics of the Asian population has previously been reported, and mainly attributed to lower transmission rates as well as population expansion following the development of drug resistance (Miotto et al. 2013), we believe this is the first report of genetically isolated African subpopulations. Given that eradication events have not been recently reported in Africa, the existence of highly homogeneous populations might be explained by the rapid expansion of a highly successful strain. For example, previous studies have shown that the widespread use of chloroquine resulted in a selective sweep for drug-resistant genotypes (Wootton et al. 2002). Intriguingly, we also note the presence of isolates in Asia that share clade membership with the majority of African isolates. Either indicative of modern African strains being introduced to Asia or the progeny of ancestral strains common to both regions.

Previously, it has been shown that different Cambodian subpopulations have different levels of resistance to artemisinin, likely associated with the transmission of polygenic traits (Miotto et al. 2013, 2015). Here, we examined the performance of PopNet to identify such traits. Focusing on a subset of 33 Cambodian isolates for which artemisinin resistance data was available, we found that isolates from the Pursat province, clustered into a single clade, were more resistant (half-life clearance (hlc) = 6.4 h) relative to isolates from Ratanakiri province, that clustered into a different clade (hlc = 2.9 h, \( P < 0.01 \), two tailed Students t-test) (fig. 4c). Hybrid strains associated with other clades that shared edges with both Pursat and Ratanakiri isolates, possessed intermediate half-life clearance rates. These findings, consistent with previous results traits (Miotto et al. 2013, 2015), again illustrate the capacity of PopNet to associate population structure with phenotypic traits.

**Discussion**

Here, we introduce a new pipeline, PopNet, for the analysis of recombination within and between populations. Although tools such as Neighbor-Net (Bryant and Moulton 2004) and STRUCTURE (Pritchard et al. 2000) provide some insight into population structure, they offer only limited resolution of shared ancestry. To overcome these limitations, chromosome painting approaches have been used to identify regions of chromosomes with shared ancestry. Applied to organisms which are thought to be largely reliant on asexual reproduction of a haploid stage for population growth, such studies are beginning to reveal the extent to which meiotic expansion drives genetic diversity within natural populations (Minoret et al. 2012; Miotto et al. 2013; Lorenzi et al. 2016). PopNet represents the first software package that automates the process of integrating shared ancestry estimates with chromosome painting. Like fineSTRUCTURE (Lawson et al. 2012), beyond the initial task of defining SNPs, PopNet does not rely on defining a single reference genome with which to relate population structure, which can otherwise obscure patterns of shared ancestry. For example, in a recent chromosome painting analysis of 62 \( T. gondii \) strains which predefined six clades, designated A–F, many haploblocks in strains assigned to clades B–F were colored according to common ancestry to clade A members, however, clade A members did not display reciprocal relationships (Lorenzi et al. 2016). PopNet, by using an all-against-all approach, overcomes this limitation by coloring segments according to the clade with which is shares common ancestry. Furthermore, the agnostic clustering process employed by PopNet (secondary clustering step), enables the software to empirically determine a statistically supported number of clades within a population rather than having to predefine such a number (whereas fineSTRUCTURE explores a range of values for the number of clades to determine a value that best partitions the population). Such features allow PopNet to avoid potential biases that might arise from founder-like effects during interpretation of results. In addition, through adopting a network approach, clade relationships are readily visualized, whereas maintaining the ability to identify strains that exhibit unique patterns of ancestry. During analysis of the \( T. gondii \) population, we further demonstrated the robustness of the pipeline by: (1) including two independent sequencing runs, using different sequencing technologies for a single strain (ARI); and (2) including genomes of variable depth of sequence coverage: ARI, 3041 and TgSkunk sequenced to \(~50–60\times\) coverage and TgGOAT sequenced to \(8\times\) coverage. In this study, we first demonstrated the ability of PopNet to resolve recombination events using a simulated population. We then illustrated its application to three diverse populations spanning a spectrum of admixture, including strains of \( S. cerevisiae \) that appear genetically isolated, strains of \( T. gondii \) that appear to be largely reliant on clonal propagation and \( P. falciparum \) isolates that appear to have undergone frequent recombination. Consistent with these histories, although \( T. gondii \) and to a lesser extent, \( S. cerevisiae \) strains exhibited distinct blocks of shared ancestry across strains, \( P. falciparum \) chromosomal segments tend to feature more heterogeneous patterns of shared ancestry with many other isolates. Nonetheless, for all three populations, PopNet recapitulated known clades and also proved capable of revealing novel subgroupings, as well as genotype–phenotype associations in \( Plasmodium \) and yeast populations. For example, applied to \( Plasmodium \) population data, PopNet reveals Asian strains to exhibit frequent
recombination events between clades, whereas African strains include three genetically distinct subpopulations that appear to be isolated from the main population. Applied to the yeast population, we found that genomic regions sharing anomalous patterns of inheritance could be used to identify a gene associated with a specific phenotypic trait. Our current aim is to develop a statistical framework, analogous to that used in genome wide association studies, to further refine PopNet’s ability to reveal genotype–phenotype associations.

Together these results showcase PopNet’s ability to analyze large population genomics data sets to reveal novel genetic relationships. Future work will investigate additional refinements to expand the scope of the current framework. These include: developing a statistical framework to associate genetic loci with phenotypic traits; expanding analysis to polyploid populations; examining the influence of lateral gene transfer events, particularly in the context of prokaryotic populations; and introducing an overlapping sliding window to better define boundaries of shared ancestry. PopNet is made freely available under the Creative Commons Attribution 4.0 International license (http://creativecommons.org/licenses/by/4.0/, last accessed April 3, 2017).

**Materials and Methods**

**Data Sets**

Genome sequence data for 17 strains of *T. gondii* (Lorenzi et al. 2016), representing the four previously defined North American lineages were obtained from EuPathDB (Aurrecoechea et al. 2013). In addition, we generated genome sequence data for seven additional strains: TGGOATUS21 (isolated from a goat, Maryland), 3142 (isolated from a sea otter, California, U.S.A.), TGSKUNK (isolated from a skunk, British Columbia, Canada), ARLMG and three progeny obtained from a previous genetic cross between strain CTG and ME49 (Sibley et al. 1992). *T. gondii* strains were harvested from monolayers of human foreskin fibroblast cells (HFF), maintained in complete Dulbecco’s Modified Eagle Medium (DMEM, Invitrogen, U.S.A.) with 10% fetal bovine serum, 2 mM glutamine (Invitrogen, U.S.A.), and 10 µg/ml gentamicin (Invitrogen, U.S.A.) at 37°C with 5% CO2 as described previously (Howe and Sibley 1997). Parasites were filtered through 3 µm polycarbonate filter (Fisher Scientific, UK) and washed with 1× phosphate buffered saline. For whole genome sequencing total genomic DNA was prepared using Illumina HiSeq 2500 at Genomics Unit, Research Technologies Branch, National Institute of Allergy and Infectious Diseases, NIH. Libraries were prepared using the TruSeq Nano library protocol (Illumina, U.S.A.) and 2 × 150 bp paired end reads generated using Illumina HiSeq 2000 at the Markov Clustering (van Dongen 2000a).

Whole genome sequencing for 3142, ARLMG, and TgSKUNK was performed using the Illumina HiSeq 2000 at the Center for Advanced Genomics, Hospital for Sick Children (http://www.tcag.ca, last accessed April 3, 2017). Libraries were prepared following Illumina’s genomic DNA library preparation protocol (Illumina, U.S.A.) and 2 × 101 paired end reads generated (3142–42 million reads/56× coverage; ARLMG 45 million reads/55× coverage; TgSKUNK—59 million reads/62× coverage). Whole genome sequencing of TGGOATUS, S22, S23, and S30 was conducted using Illumina HiSeq 2000 at Genomics Unit, Research Technologies Branch, National Institute of Allergy and Infectious Diseases, NIH. Libraries were prepared using the TruSeq Nano library protocol (Illumina, U.S.A.) and 2 × 150 bp paired end reads generated using Illumina HiSeq 2000 at the Markov Clustering (van Dongen 2000a).

Running PopNet

Taking aligned SNP data as input, PopNet first performs a series of “primary” clustering to identify regions of shared ancestry between strains. In brief, SNPs for all genomes within a single population are divided into segments representing fixed-length sections of the genome. For each segment, a matrix is generated from the number of SNPs shared between every pair of strains; only SNPs shared by two or more strains are included in this analysis. The matrix is then clustered using the Markov clustering algorithm (van Dongen 2000a).
Next, a “secondary” clustering step is performed to define major ancestries within the population. For this a global similarity matrix is generated that defines, for each pair of strains, the frequency with which they cocluster during the primary clustering step. The matrix is again clustered using MCL with inflation and preinflation values being tuneable parameters that impact cluster resolution and size of chromosome features—see main text for a description of how these are selected. Resulting clusters from this secondary clustering step define clades and are assigned unique colors during visualization. Networks reflecting strain relationships are visualized using Cytoscape version 3.2.1 using the Enhanced Graphics plugin (Shannon et al. 2003; Morris et al. 2014). Nodes depicted as annuli, represent the concatenated chromosomes of individual strains with edges representing the frequency that two strains cocluster during the primary clustering step. The color of the node border indicates its clade membership from the secondary clustering step; sectors within the annular depiction are colored with reference to the groups defined during the primary clustering step—see main text. Genomic positions with missing data may be denoted as “-” in the input SNP table; such sections will subsequently be colored grey. The PopNet pipeline is performed using a series of inhouse scripts using the Python programming language.

Supplementary Material
Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments
This study was financially supported by the Canadian Institute for Health Research (CIHR-MOP 84556 to J.P. and M.E.G.), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-06664 to J.P.) and the Intramural Research Program of the NIH and NIAID (M.E.G.). M.E.G. is a scholar of the Canadian Institute for Advanced Research (CIFAR) Integrated Microbial Biodiversity Program. Computing resources were provided by the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund—Research Excellence; and the University of Toronto. Artemisinin resistance data was kindly provided by Rick Fairhurst (NIH, Bethesda, U.S.A.). We thank Xinzhuang Su (NIH/NIAID) for critically reading the manuscript and the editor and anonymous reviewers for their helpful comments to improve the clarity of the manuscript and directions for future refinements.

References
Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, et al. 2013. EuPathDB: the eukaryotic pathogen database. Nucleic Acids Res. 41:D684–D691.
Bergstrom A, Simpson JT, Salinas F, Barre B, Parts L, Zia A, Nguyen Ba AN, Moses AM, Louis EJ, Mustonen V, et al. 2014. A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol. 31:872–888.
Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.
Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40:D700–D705.
Conway DJ, Roper C, Oduola AM, Arnot DE, Kremsner PG, Grobusch MP, Curtis CF, Greenwood BM. 1999. High recombination rate in natural populations of Plasmodium falciparum. Proc Natl Acad Sci U S A. 96:4506–4511.
Cubillos FA, Billi E, ZÖRGÖ E, Parts L, Fargier P, Omholt S, Blumberg A, Warringer J, Louis EJ, Liti G. 2011. Assessing the complex architecture of polygenic traits in diverged yeast populations. Mol Ecol. 20:1401–1413.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DeFristo MA, Handsaker RE, Lunter G, Marsh GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:1566–1566.
Dubey JP, Sundar N, Hill D, Velmurugan GV, Bandini LA, Kwok OC, Majumdar D, Su C. 2008. High prevalence and abundant atypical genotypes of Toxoplasma gondii isolated from lambs destined for human consumption in the USA. Int J Parasitol. 38:999–1006.
Evans BR, Gloria-Soria A, Hou L, McBride C, Bonizzoni M, Zhao H, Powell JR. 2015. A multipurpose, high-throughput single-nucleotide polymorphism chip for the dengue and yellow fever mosquito, Aedes aegypti. G3 (Bethesda) 5:711–718.
Fu CY, Liu WG, Liu DL, Li JH, Zhu MS, Liao YL, Liu ZR, Zeng XQ, Wang F. 2016. Genome-wide DNA polymorphism in the indica rice varieties RGD-75 and Taifeng B as revealed by whole genome re-sequencing. Genome 59:197–207.
Grigg ME, Bonnefoy S, Hehl AB, Suzuki Y, Boothroyd JC. 2001. Success and virulence in Toxoplasma as the result of sexual recombination between two distinct ancestries. Science 294:161–165.
Guo F, Wei S, Zhang C, Zhang H, Zhang B, Xing B, Feng Z, Gao C, Liu X, Li S. 2013. A population-based association study of 2q32.3 and 8q21.3 loci with schizophrenia in Han Chinese. J Psychiatr Res. 47:712–717.
Howe DK, Sibley LD. 1997. Development of molecular genetics for Neospora caninum: a complementary system to Toxoplasma gondii. Methods 13:123–133.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.
International HapMap C. 2003. The International HapMap Project. Nature 426:789–796.
Jiang H, Li N, Gopalan V, Zilversmit MM, Varma S, Nagarajan V, Li J, Mu J, Hayton K, Henschen B, et al. (Jiang2011 co-authors). 2011. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. Genome Biol. 12:1–15.
Johann PD, Erkek S, Zapatka M, Kerl K, Buchhalter I, Hovestadt V, Jones DT, Sturm D, Hermann C, Segura Wang M, et al. 2016. Atypical teratoid/rhabdoid tumors are comprised of three epigenetic subgroups with distinct enhancer landscapes. Cancer Cell 29:379–393.
Kanazawa S, Driscoll M, Struhl K. 1988. ATR1, a gene encoding a transmembrane protein required for aminotriazole resistance. Genes to Cells 3:711–718.
Khanyile KS, Dzomba EF, Muchadeyi FC. 2015. Haplo-block structure of Southern African village chicken populations inferred using genome-wide SNP data. Genet Mol Res. 14:12276–12287.
Kurtz S, Phillippy A, Delcher A, Smoot M, Shumway M, Antonescu C, Salzberg S. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Lawson DJ, Helenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8:e1002453.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome Project Data Processing S. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Liti G. 2015. The fascinating and secret wild life of the budding yeast S. cerevisiae. eLife 4.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458:337–341.

Lorenzi H, Khan A, Behnke MS, Namavivayam S, Swapna LS, Hadjithomas M, Karomycheva S, Pinney D, Brunk BP, Ajikia JW, et al. 2016. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. Nat Commun. 7:10147.

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

Miller MA, Miller WA, Conrad PA, James ER, Melli AC, Leutenegger CM, Dabritz HA, Packham AE, Paradies D, Harris M, et al. 2008. Type X Toxoplasma gondii in a wild mussel and terrestrial carnivores from coastal California: new linkages between terrestrial mammals, runoff and toxoplasmosis of sea otters. Int J Parasitol. 38:1319–1328.

Minot S, Melo MB, Li F, Lu D, Niedelman W, Levine SS, Saeij JP. 2012. Admixture and recombination among Toxoplasma gondii lineages explain global genome diversity. Proc Natl Acad Sci U S A. 109:13458–13463.

Miotto O, Almagro-Garcia J, Manske M, Macinnis B, Campino S, Rockett KA, Amarantunga C, Lim P, Suon S, Sreng S, et al. 2013. Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. Nat Genet. 45:648–655.

Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, Amarantunga C, Lim P, Mead D, Oyoa SO, Dhorda M, et al. 2015. Genetic architecture of artemisinin-resistant Plasmodium falciparum. Nat Genet. 47:226–234.

Mita T, Tanabe K. 2012. Evolution of Plasmodium falciparum drug resistance: implications for the development and containment of artemisinin resistance. Jpn J Infect Di. 65:465–475.

Morris JH, Kuchinsky A, Ferrin TE, Pico AR. 2014. enhancedGraphics: a Cytoscape app for enhanced node graphics. F1000Res 3:147.

Pan S, Thompson RC, Grigg ME, Sundar N, Smith A, Lymbery AJ. 2012. Western Australian marsupials are multiply infected with genetically diverse strains of Toxoplasma gondii. PLoS One 7:e45147.

Pang B, Du P, Zhou Z, Diao B, Cui Z, Zhou H, Kan B. 2016. The transmission and antibiotic resistance variation in a multiple drug resistance clade of Vibrio cholerae circulating in multiple countries in Asia. PLoS One 11:e0149742.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Saadatnia G, Golkar M. 2012. A review on human toxoplasmosis. Scand J Infect Dis. 44:805–814.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498–2504.

Sibley LD, LeBlanc AJ, Pfefferkorn ER, Boothroyd JC. 1992. Generation of a restriction fragment length polymorphism linkage map for Toxoplasma gondii. Genetics 132:1003–1015.

Singh P, Singh BN. 2008. Population genetics of Drosophila ananassae. Genet Res. (Camb) 90:409–419.

van Dongen S. 2000a. Graph clustering by flow simulation. PhD thesis. University of Utrecht.

van Dongen S. 2000b. Performance criteria for graph clustering and Markov cluster algorithms. Technical Report. Amsterdam, The Netherlands: CWI (Centre for Mathematics and Computer Science).

Woo P, Dhanda S, Wang SA, Bai FY. 2012. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. Mol Ecol. 21:5404–5417.

Wendte JM, Gibson AK, Grigg ME. 2011. Population genetics of Toxoplasma gondii: new perspectives from parasite genotypes in wildlife. Vet Parasitol. 182:96–111.

Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, Magill AJ, Su X-z. 2002. Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature 418:320–323.

World Health Organization. 2015. World Malaria Report 2015. Geneva, Switzerland: World Health Organization. pp. 280.

Yahara K, Furuta Y, Oshima K, Yoshida M, Azuma T, Hattori M, Uchiyama I, Kobayashi I. 2013. Chromosome painting in silico in a bacterial species reveals fine population structure. Mol Biol Evol. 30:1454–1464.

Yin JH, Yang MN, Zhou SS, Wang Y, Feng J, Xia ZG. 2013. Changing malaria transmission and implications in China towards National Malaria Elimination Programme between 2010 and 2012. PLoS One 8:e74228.