Cell Host & Microbe

Co-transmission of related parasite lineages shapes within-host parasite diversity

--Manuscript Draft--

Manuscript Number:	CELL-HOST-MICROBE-D-19-00301R3
Full Title:	Co-transmission of related parasite lineages shapes within-host parasite diversity
Article Type:	Research Article
Keywords:	Malaria; Transmission; Genomics; Genetics; Single Cell Sequencing
Corresponding Author:	Ian H. Cheeseman Texas Biomedical Research Institute San Antonio, UNITED STATES
First Author:	Standwell C Nkhoma
Order of Authors:	Standwell C Nkhoma
	Simon G Trevino
	Karla M Gorena
	Shalini Nair
	Stanley Khoswe
	Catherine Jett
	Roy Garcia
	Benjamin Daniel
	Aliou Dia
	Dianne J Terlouw
	Stephan A Ward
	Timothy JC Anderson
	Ian H. Cheeseman
Abstract:	Malaria patients frequently carry one or more clonal lineage of the parasite, Plasmodium falciparum. In regions of high transmission, we might expect component parasites within complex infections to be unrelated as a result of parasite inoculations from different mosquitos. This project was designed to directly test this prediction. We generated 485 near-complete single-cell genome sequences isolated from fifteen P. falciparum patients from Chikhwawa, Malawi, an area of intense malaria transmission. Matched single-cell and bulk genomic analyses revealed that patients harbored up to seventeen unique lineages. Current statistical approaches were unable to accurately reconstruct infection composition from bulk sequence data. Surprisingly, our analysis demonstrated that parasite lineages within infections tend to be related, suggesting that superinfection by repeated mosquito bites is rarer than co-transmission of parasites from a single mosquito. Our single-cell analysis indicates strong barriers to establishment of secondary infections, providing new insights into the biology and transmission of malaria.
Suggested Reviewers:	Daniel Hartl, Ph.D Professor, Harvard dhartl@oeb.harvard.edu Prof. Hartl is a population geneticist who has worked extensively on malaria parasite genetics. He has previously published on intrahost relatedness in malaria infections.
	Dyann Wirth, Ph.D. Professor, Harvard dfwirth@hsph.harvard.edu Prof. Wirth is a malaria parasite geneticist who has worked extensive on malaria parasite population structure, drug resistance and natural selection in malaria parasite populations.

	Christopher Newbold, Ph.D. Professir, University of Oxford chris.newbold@imm.ox.ac.uk Prof. Newbold has worked extensively on parasite genomics and has spearheaded projects on natural variation in malaria parasite populations.
	Photini Sinnis, MD Professor, Johns Hopkins School of Public Health psinnis1@jhu.edu Prof. Sinnis has worked extensively on bottlenecks in transmission throughout the malaria parasite lifecycle. She has used mathematical modelling, cell biology, genetics and field studies to achieve this.
	Sarah Tishkoff, Ph.D. Professor, University of Pennsylvania tishkoff@pennmedicine.upenn.edu Prof. Tishkoff is a human geneticist who has often worked on understanding the impact of malaria infection on the human genome.
	Jane Carlton, Ph.D. Professor, New York Univeristy jane.carlton@nyu.edu Prof. Carlton has pioneered population genomics in the malaria parasite Plasmodium vivax. She has been the driving force behind the generation of genome references for multiple malaria parasite species.
Opposed Reviewers:	

TEXAS BIOMEDICAL RESEARCH INSTITUTE

Department of Genetics

Ian H. Cheeseman, Ph.D.

October 10th, 2019

Tel: 210 258 9834
Email: ianc@txbiomedgenetics.org

To the Editors:

Research Article submission: "Co-transmission of related parasite lineages shapes within-host parasite diversity"

We submit for your interest a revised version of our manuscript. We feel we have comprehensively addressed all comments from the reviewers and have considerably strengthened our manuscript. Following comments from Reviewer Two we have also altered the title from "Single cell genomic dissection of complex African malaria infections indicate strong barriers to reinfection" to "Co-transmission of related parasite lineages shapes within-host parasite diversity".

Malaria infections often contain multiple, genetically distinct parasites. Genetically distinct parasites co-infect an individual through two routes, superinfection, where an individual is bitten by two infected mosquitoes bearing distinct parasites, or co-transmission, where a single mosquito inoculates an individual with multiple parasite genotypes. As sexual recombination between parasite genotypes occurs in the mosquito midgut these two processes give rise to drastically different patterns of within-host variation, and influence the evolution of several biomedically important traits including parasite virulence, antimalarial drug resistance and malaria transmission. Genome sequencing of parasites from malaria infections is commonly used to understand the complexity of infection, but is unable to infer the numbers of clones present or determine the relatedness among parasites within a single host severely limiting the interpretations which can be made.

In the submitted manuscript we use single cell sequencing to understand superinfection and co-transmission in malaria infections. By implementing a highly optimized protocol (Trevino et al, Genome Biology and Evolution, 2017) we are able to generate near-complete capture of single parasite genomes. We use this approach to generate the most comprehensive portrait of within-host genetic variation to date, producing ~450 complete single cell genome sequences from across 15 infections from a region of intense malaria transmission.

Surprisingly, we find that superinfection to be rare compared to co-transmission, with most complex infections composed of multiple recently related lineages. These lineages persist across multiple mosquito transmission cycles, with substantial genetic variation maintained during each mosquito bottleneck. Our results strongly suggest that there are barriers that prevent superinfection of malaria infected patients.

These data provide a gold standard for the development of statistical tools for robust deconvolution of malaria infections using bulk sequence data. As we push for malaria elimination it will be crucial for us to understand how genetic variation is maintained in the

face of aggressive control and elimination efforts. Our study adds critical depth to our understanding of how within host genetic variation is generated and maintained in areas of intense malaria transmission.

Sincerely,

Ian Cheeseman, Ph.D.
Milton S. & Geraldine M. Goldstein Young Scientist
Assistant Scientist
Program Lead Host Pathogen Interactions

Dear Dr Lim,

We thank you and the two reviewers for the careful review of our manuscript. The comments were indeed clear and constructive, and have improved the manuscript.

We are mortified that a substandard version of the supplementary material was uploaded for the prior submission. As we described in our previous response there had been a concerted effort to improve the readability of this supplement, and we apologize for wasting time.

In addition, there were errors caught by the reviewers in the submission which we cannot attribute to this oversight. We have corrected these in a substantially revised manuscript. We feel that the comments raised have improved our manuscript and we hope that the revised version is suitable for publication in Cell Host and Microbe. We have responded below to each comment and underlined our response.

Dear Ian,

I am enclosing the comments that the reviewers made on your paper. Unfortunately, the consensus recommendation is against publication in Cell Host & Microbe in its current form. As discussed previously, we have asked both one of the original reviewers as well as a new reviewer with expertise in parasite genomics to evaluate your revised manuscript. The reviewers continue to refer to the potential interest of the topic and the approach, but cite that significant further work would be necessary for the conclusions to be definitive. Given their comments, we feel that it would be premature to proceed further with the manuscript based on the current presentation. Although we would not necessarily rule out resubmission of a revised manuscript that addresses the reviewers' concerns, it seems that substantial further work, detailed clarification and presentation would be required to achieve this aim. I should also mention that you will not be able to submit a revised version of this manuscript in EM without contacting me first to reactivate the file.

I am sorry that the outcome for this manuscript is not more positive. I hope, however, that the reviewers' comments are clear and constructive.

If you decide not to revise the paper for *Cell Host & Microbe*, you may want to consider transferring the paper to another Cell Press journal. If you are interested in having your paper considered at another Cell Press journal on the basis of these reviews, please contact the editor of the relevant journal directly and I would be glad to assist in transferring the files/information related to your paper. In particular, you may want to consider *Cell Reports* the newest sister journal at Cell Press (http://www.cell.com/cell-reports/home).

Best wishes,

Caeul Lim, Ph.D.
Scientific Editor, Cell Host & Microbe

Reviewers' Comments:

Reviewer #1: The author provide a revised manuscript in which they describe single cell sequencing of clones from a variety of different P. falciparum infections. The revised manuscript is more accessible to a broader audience and the authors have addressed a number of the reviewers' criticisms. However, some of the concerns about the supplementary data have not been addressed—in particular the concern that supplemental figures are illegible remains true. In addition, with this revision, the authors have uploaded a supplement which is apparently full of errors (e.g. missing figures) and which prevents a thorough review. Thus, the comments below are not a full assessment of the manuscript, which cannot be provided given the numerous mislabeled and missing figures.

We are disappointed to have uploaded incorrect supplementary files. This was done in error, and we deeply apologize. We have taken this opportunity to comprehensively review our supplement, figures and legends. We have pushed for greater legibility across the document. These changes include:

- **Detail added to the legend of Supplementary Figures 1,2,4-12**
- **Increased size and legibility of Supplementary Figures 1,5,6,9 & 10**
- **Inclusion of Supplementary Figures 1, 3, 5, 6 & 11 in the main supplement, as well as Supplementary Figure 12 (originally mistakenly described as Figure S8 in the text)**
- **Expansion of Supplementary Figure 8 to include bulk sequencing data**
- **Replacement of Supplementary Figure 7 with two new figures to support the identification of tetrads**

The authors write, "ERSA estimates relatedness between individuals from distribution of IBD tract lengths (Figure 5A, B) using individuals assumed to be unrelated from the same population as a reference. We see a spectrum of relationships within each infection (Figure 5C, Figure S6). It is not clear if the ERSA relationship is in Figure 5 as the legend to Figure 5 does not mention this.

We have corrected this is the revised manuscript. We had updated Figure 5 in the previously uploaded text, but not the stand-alone figure. Both are now in agreement.

Figure S5 and S6 appear to be missing.

Our updated supplement has been revised to include these figures.

The authors write "As seen in previous studies on P. falciparum genetic crosses (Jiang et al., 2011) chromosome length and crossover count were correlated (r2344 =0.8, p=0.0005, Pearsons correlation, Figure S8)." However, Figure S8 appears to show an IGV screenshot.

We have corrected this error, the figure is now included in the supplement.

Figure S8—two reads does not provide much confidence that the mitochondrial reads are real. Would need to see more data. I cannot assess the quality based on what is provided for review.

We agree with this statement. In our original manuscript we had sought to capture our uncertainty about this site. Line 370-371 "These may have arisen as *de novo* mutations during the current infection, or be artifacts of the WGA and genome sequencing pipeline". We have further clarified this Line 317-322 "However, as these may have arisen as *de novo* mutations during the current infection, or be artifacts of the WGA and genome sequencing pipeline we excluded this site from further analysis" and added additional detail to the supplementary figure supporting this statement. We do not believe the site supported by 2 reads in Cell 4 is the *de novo* mutation. Cell 4 shares complete identity to the dominant haplotype in the infection. We have included the bulk sample to the IGV plot supporting this in supplementary figures 7, and show complete mitochondrial genotypes in supplementary figure 8. We believe the *de novo* mutation we refer to is present in cells 9 and 22 and supported by 31 and 28 reads respectively. We thank the reviewer for the chance to clarify this point.

Overall the discussion of mitochondrial inheritance is poorly written and it is difficult to follow. Sparse figure legends do not help. The authors reference Figure 5D, which does not apparently exist.

We thank the review for these comments, we have rewritten much of this section to provide clarity and provided an updated version of Figure 5. Briefly, we show that parasites with identical genotypes in their nuclear genome also have identical genotypes in their mitochondrial genome. We have moved this section to improve the flow and substantially rewritten it changing to:

"Mitochondrial inheritance within individual infections

Mitochondria are inherited maternally during malaria parasite meiosis and therefore allow the identification of parasites which share a maternal lineage within infections. We found 93 SNPs in the mitochondrial DNA which varied within our dataset. Across all 498 parasites we were able to capture the genotype of 79.1% (36,645/46,314) of these 93 sites. We expect parasites which are identical across their nuclear chromosomes to also be identical across their mitochondrial genome. We find this to be the case. Within individual hosts, parasites which shared 100% of their genome IBD also shared 100% of their mitochondrial genome sequence. Across the 498 genome sequences for which we had reliable estimates of genome-wide IBD and mitochondrial genotypes there were only 2 of a possible 36,645 sites (0.0055%) where the mitochondrial genotype varied within a group of parasites which were 100% IBD. These arise at position 1,692 in cells 9 and 22 of infection MAL24. Visual inspection of these sites (Figure S7) and the presence of the reads supporting both genotypes in the bulk genome sequence supports these being genuine. However, as these may have arisen as *de novo* mutations during the current infection, or be artifacts of the WGA and genome sequencing pipeline we excluded this site from further analysis.

We identified a total of 20 unique mitochondrial haplotypes across the entire dataset (Figure S8 and Figure S9). We counted the number of mitochondrial haplotypes present in each infection (Figure 5D). This showed 9 infections contained a single mitochondrial haplotype, 3 infections contained 2 haplotypes, 1 infection contained 3 haplotypes, and 2 infections contained 4 haplotypes. All monoclonal infections contained a single mitochondrial haplotype, as did 4/10 (40%) polyclonal infections. Within polyclonal infections distinct mitochondrial haplotypes were observed between very close relatives (sharing >60% of their genomes IBD) supporting the retention of diversity we observe in the face of recurrent inbreeding. As mitochondria are uniparentally inherited the presence of multiple mitochondrial haplotypes across related parasites from within the same infection demonstrates multiple oocysts within a mosquito are responsible for initiating an infection. We tested if a higher level of shared pair-wise IBD is a predictor of sharing a mitochondrial haplotypes between two parasites. After excluding clonally identical parasites we estimate a 10% increase in IBD shared between parasites from the same infections increasing the odds of sharing a mitochondrial genotype by a factor of 10.082 (p=5.25x10^{-8}, logistic regression)."

Figure S9 and S10 are still illegible. What is UPGMA? What are the axes? "Height" is upside down.

We have increased the font size, provided a more appropriate label for the axis and greater detail in the figure legends for these two plots. In addition, they are now correctly labelled as Figures S10 and S11.

Reviewer #2: This is an intriguing and important manuscript which uses single cell genomics to carefully define the complexity of infection in Plasmodium falciparum clinical isolates. It is well established that individuals in high transmission regions are often infected with multiple different P. falciparum haplotypes, but exactly how many, and how they are related to each other, has been difficult to determine. The approach used, single cell genome sequencing, is cutting edge, because while scRNAseq is becoming increasingly applied to Plasmodium research, single cell genome sequencing is more challenging to generate because there is only a single copy of the genome in each cell, rather than the multiple copies of most RNA species. This is therefore an important question using novel technologies. Quality control of the single cell data is clearly key, but appears to be rigorous - there is little evidence of contaminating DNA from other sources, and samples with mixed calls, which could indicate the presence of two cells in the source material instead of one, are eliminated. The central findings are really two fold - one technical, that single cell sequencing can be used to deconvolute mixed infections, and one biological, and that individuals are more often infected with related parasite genotypes than might be expected. The former seems clearly proven, with some technical clarification required as noted below. The latter is also clear and important, although there is some confusion in the way the methodologies are applied and inferences derived, and the title rather overstates the generalisability of the findings. Suggested areas for clarification are enumerated below:

Major issues

1. Amplification bias. A key question is obviously one of bias in the single cell sequence data. Whole Genome Amplification is used to amplify the material, but why WGA and not selective

WGA which might give more even coverage across the genome? No details about the WGA method are provided in line 548, and depending on the method, the worry is that it will result in bias towards some (presumably slightly less AT-rich) regions of the genome. Is there evidence of this? One would guess so, based on the fact that only 60,002 SNPs were scored, and 10,997 used for downstream analysis, which is a relatively low number for P. falciparum genomic analysis, but no details are given. Given that the same SNPs are used in all samples, it doesn't necessarily invalidate the interpretation, but much more clarity about the single cell sequencing methodology and limitiations thereof is required.

We thanks the reviewer for these comments and the opportunity to provide greater detail to our approach. We have published two previous papers, Nair et al, Genome Research, 2012 and Trevino et al, Genome Biology and Evolution, 2017 where we have included extensive details about the development of the WGA protocol we implement here. We appreciate that these were not appropriately cited in the manuscript. This has been corrected, and we have added some additional detail to this section at lines 530-532, and 579-586 of the revised manuscript.

In short, we have opted for WGA as it outperforms sWGA. In a recent paper (Oyola, Ariani et al, Malaria Journal 2016) using an input of 40 parasites less than 50% of the genome was captured. Selective WGA, even with an excess of parasite templates, appears to plateau at 85% of the genome covered. When limiting samples to the same criteria as that paper (libraries with at least 20M reads) our approach captures 95% of the genome with 5 or more reads, similar to our published estimates (Trevino et al, Genome Biology and Evolution, 2017). We have previously shown (Nair et al, Genome Research, 2012) that any bias from this approach is stochastic, rather than deterministic and not correlated with AT content. As we have published multiple papers describing the methodology applied here we opt to refer readers to these more comprehensive analysis rather than include further details here.

There was an additional concern that we had scored a low number of mutations for a malaria genomic survey. We agree there was insufficient detail included to appropriately judge how comparable our data is to other malaria genome sequencing efforts. We have added further details on our SNP calling pipeline, and variant filtering to highlight that the numbers here are not low, but are conservatively called. Notably, we include a larger subset of variant sites (~175K SNPs, Line 143) used during initial quality control of the data. To investigate this further we compared our results to SNP calls from the Pf3K project a large, well curated dataset limiting our comparison to chromosome 1 and the 317 samples in Chikhwawa, Malawi present in the Pf3K data. Our genotyping pipeline and filtering was based upon the Pf3K methods and correspondingly we find very similar numbers of common sites (for MAF>0.05: 1,021 in our data and 1,102 in Pf3K and for MAF>0.01: 2,666 in our data and 2,791 in Pf3K). Due to the larger sampling Pf3K has a larger amount of variants when no minor allele frequency is applied (2,666 in our data and 8,652 in Pf3K). In short, we believe the SNP calling in our data is comparable with contemporary approaches.

2. Correlation between number of haplotypes called by bulk sequencing cf single cell data. Figure 2C suggests that there is considerable disconnect between the calls of complexity made from bulk and single cell sequence data, with several infections on the left hand side of the x axis having 1-3 haplotypes identified from single cell rarefaction, despite having lower Fws scores. Purely visually, this doesn't seem in keeping with Figure 3A, where the correlation seems much higher. Is this because the "effective number of haplotypes" estimated from single cell data used in Figure 3 is different from the number of haplotypes estimated from the same data in Figure 1C? If these are indeed different methods (which is what I guess is happening), it does not come across clearly at all in the manuscript. Clearer explanation and pros/cons of each method would seem useful to include. On a related note, in Figure 3A, the outliers all lie under the trend line suggest that either Fws is overestimating complexity, or single cell is underestimating. That is the opposite of what I would expect, given the difficulty of detecting strains present in minor abundance. Is there an explanation for this? This comment speaks to a general issue of the paper - multiple different methods and computational tools are used somewhat interchangeably, without clear rationale or explanation in some places. For the non-expert reader, it will still be difficult to follow, and more clarity is required.

We thank the reviewer for the ability to clarify this important point in our manuscript. There are multiple distinct estimates of the number of haplotypes present in an infection presented here. The first, FWS, captures unfixed sites and is highly effective at capturing clonal vs. non-clonal. Beyond this there is no expectation that a specific FWS value will capture "complexity" in a discrete way. There are two estimates of the actual number of haplotypes, one from DEploid and inferred from bulk sequencing, the other by single cell sequencing. These measure the same thing, though do not agree perfectly in this analysis. In an effort to compare across all measures the effective number of haplotypes (introduced as part of DEploid by Zhu et al) normalizes the number of haplotypes by their abundance. These measures are concordant as they all account for abundance. We have added additional detail on each approach in lines 225-229 to address the lack of clarity.

3. In Figure 4, are all the bulk sequencing squares coloured using the same code that the single cell sequencing squares are? A few clusters (the colour code is hard to follow, see below) seems to lack a bulk sample - how is this possible? And is it useful to include all the bulk samples in this figure, given that most didn't have single cell data called, and so simply sit at the upper right of the figure?

We have included an updated version we hope will be clearer for readers. All infections with single cell data also have a bulk sequence. For the figure there are occasions where the bulk and single cell are identical, and this has obscured the bulk dot. We have made these more prominent in the revised figure. We feel it is important to include the data which does not fall within the network. The alternative would suggest far greater connectivity is present in the general population than we detect. We have wrestled with the color schemes to attempt to improve clarity. This is not trivial with a large number of infections.

4. Possible causes of higher than expected level of relatedness observed in complex infections. The data in Figure 5 is compelling, and novel. A key question is how generalisable it is - whether there is something unusual or unique about this sample collection/study site that could explain the finding. While expanding to more locations is clearly out of scope of this manuscript, some consideration of this issue in the Discussion might be useful. For example, if the infections were from a region where transmission was high by highly fragmented, with pockets of high transfection surrounded by arid, low transmission regions, then this kind of pattern might be more expected. If transmission is more uniform in the area, then the findings are more surprising and significant. Before making a general case that complex infections are frequently formed from related parasites, it is critical to state the epidemiological context and possible confounders.

We agree that the sample size possible in this work does limit how generalizable the findings are. We have included a more detailed discussion of the context in lines 417-425.

"In this work we have generated a detailed picture of within-host genetic variation across fifteen malaria patients. While this is a modest number of infections, these findings are directly informative about the limits of malaria complexity. We have surveyed a high transmission setting and focused on polyclonal infections. This sampling enriches for infections with the highest likelihood of genetic diversity arising from superinfection rather than co-transmission. In spite of this we find co-transmission to be widespread, and likely underappreciated as a mechanism generating and maintaining genetic diversity in natural malaria populations. There is a pressing need to extend these observations across the range of malaria endemicity to fully capture the transmission network of malaria infections."

5. Title. I'm not convinced that the manuscript actually shows that there are strong barriers to reinfection. What has been shown is a higher than expected rate of inbreeding. There could be many explanations for this, some epidemiological as noted above, and until repeated in different contexts, I think I would be very hesitant to say anything as definitive as the title implies….

We have now modified the title to "Co-transmission of related parasite lineages shapes within-host parasite diversity" to better convey the key observation from this work

Minor issues

1. Figure 4. Given the colour scheme has multiple relatively similar colours, it's quite hard to translate from the key to individual infections. Given this figure is used to highlight examples of superinfection and cotransmission, alternative labelling is needed - either by indicating more clearly (for example with circles, or arrows) on the figure the examples that are referred to in lines 258-260 (MAL5, 15, 23 and 24), or coming up with a different scheme to label all the infections which doesn't have the same issue of similar colouring…

We thank the reviewer for this suggestion. We agree the suggested scheme improves the ability to understand the figure and have updated the figure.

Typographical errors
1. Line 52 - analyse not analysis?
2. Line 79 - single parasite genotype present presumably refers only to Figure 1A?
3. Line 258 - missing "that"?

Thank you, we have corrected these errors

Co-transmission of related parasite lineages shapes within-host parasite diversity

Standwell C. Nkhoma[1,2,3,4,*,†], Simon G. Trevino[4], Karla M. Gorena[5], Shalini Nair[4], Stanley Khoswe[1], Catherine Jett[4], Roy Garcia[4], Benjamin Daniel[5], Aliou Dia[4], Dianne J. Terlouw[1,2], Stephen A. Ward[2], Timothy J.C. Anderson[4], Ian H. Cheeseman[4,6,*]

[1]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, University of Malawi College of Medicine, Blantyre, Malawi.
[2]Liverpool School of Tropical Medicine, Liverpool, United Kingdom.
[3]Wellcome Trust Liverpool Glasgow Centre for Global Health Research, Liverpool, United Kingdom.
[4]Texas Biomedical Research Institute, San Antonio, Texas, United States of America.
[5]University of Texas Health Science Center San Antonio, San Antonio, Texas, United States of America.
[6]Lead Contact
*Correspondence: ianc@txbiomed.org, snkhoma@atcc.org

† Present Address**: Malaria Research and Reference Reagent Resource Center (MR4), BEI Resources, ATCC, Manassas, VA, USA**

SUMMARY

In high transmission regions, we expect parasite lineages within complex malaria infections to be unrelated due to parasite inoculations from different mosquitoes. This project was designed to test this prediction. We generated 485 single-cell genome sequences from fifteen *P. falciparum* malaria patients from Chikhwawa, Malawi, an area of intense transmission. Patients harbored up to seventeen unique parasite lineages. Surprisingly, parasite lineages within infections tend to be closely related, suggesting that superinfection by repeated mosquito bites is rarer than co-transmission of parasites from a single mosquito. Both closely and distantly related parasites comprise an infection, suggesting sequential transmission of complex infections between multiple hosts. We identified tetrads and reconstructed parental haplotypes which revealed the inbred ancestry of infections and non-Mendelian inheritance. Our analysis suggests strong barriers to secondary infection and outbreeding amongst malaria parasites from a high transmission setting, providing unexpected insights into the biology and transmission of malaria.

INTRODUCTION

Malaria remains a major global health problem, with ~400,000 malaria-related deaths in 2015, and over 200 million clinical cases (WHO, 2016) cases The intensity of malaria transmission is correlated with the complexity of infection (COI), the number of genetically distinct parasites observed within a single infection. Genetically distinct malaria parasites can infect an individual through two routes (Figure 1). A single individual may be bitten by two (or more) infected mosquitoes, each bearing a unique parasite genotype (Figure 1B), or an individual may be bitten by a single mosquito bearing more than one parasite genotype (Figure 1C). Throughout, we refer to these two processes as superinfection and co-transmission respectively. Superinfection of an individual by multiple infectious mosquito bites is often used to explain the high complexity of infection found in high transmission regions(Volkman et al., 2012). Following a bloodmeal, gametocyte stage parasites fuse in the mosquito midgut, and an obligate round of sexual recombination occurs. If only a single parasite genotype is present, all offspring will be identical (Figure 1A, B). When multiple parasite genotypes are present recombinant progeny

may arise (Conway et al., 1991, Mu et al., 2005, Nkhoma et al., 2012, Wong et al., 2017, Wong et al., 2018) (Figure 1C). At a population level recombination of parasite genotypes shapes the local decay of linkage disequilibrium and haplotype variation (Mu et al., 2005, Mu et al., 2007, Neafsey et al., 2008).

The clinical impact of complex infections has been studied in mouse malaria models. Here, interactions between genetically distinct malaria parasites are known to influence the evolution of parasite virulence, antimalarial drug resistance, immunity, gametocyte sex ratios, and malaria transmission (Bell et al., 2006, de Roode et al., 2005, Wargo et al., 2007a, Wargo et al., 2007b, Reece et al., 2008). However, translating these findings to human malaria has been a major challenge. This is due to the paucity of appropriate tools for resolving infection complexity on a large-scale at the level of single parasitized cells: we cannot directly infer the composition of malaria infections by bulk sequencing of infected blood samples. Complex infections confound most traditional genetic analysis, preventing the accurate inference of allele frequencies and even simple genotype-phenotype associations (Nair et al., 2014, Wong et al., 2018). However, powerful new approaches to analyze individual malaria infections are emerging, aided by recent advances in targeted capture of singly-infected erythrocytes from complex mixtures and improved methods for single-cell sequencing (Nair et al., 2014, Trevino et al., 2017), and computational approaches for interpreting infection complexity (Chang et al., 2017, Zhu et al., 2018).

Using single-cell sequencing and cloning parasites by limiting dilution from a single individual, we previously saw a range of inferred relationships amongst co-infecting parasite haplotypes, including identical clonal lineages, siblings and unrelated individuals (Nair et al., 2014, Nkhoma et al., 2012, Trevino et al., 2017). However, it is unknown to what extent these findings can be generalized across a population. This project was designed to understand the degree to which the genetic diversity of individual infections is driven by superinfection of unrelated strains, or co-transmission of related ones. To do this, we performed whole genome sequencing of bulk infections and single-cell sequencing of parasite-infected cells isolated from malaria patients in Chikhwawa, a high transmission region in Malawi.

RESULTS

Infection complexity in bulk sequenced samples

To resolve the within-host population structure of malaria infections, we performed a cross-sectional survey of individuals infected with uncomplicated *P. falciparum* malaria in Chikhwawa, Malawi, an area of high malaria transmission (entomological inoculation rate 183 infectious bites per person per year (Mzilahowa et al., 2012). In this setting, we might expect that patients will contain a mixture of unrelated parasites, resulting from independent mosquito inoculations (i.e. that superinfections will predominate). We performed bulk parasite genome sequencing of 49 infections to a median read depth of 31 (interquartile range 20.93-48.37). We estimated the complexity of infection from bulk sequence data using 10,997 unfixed SNP positions with a minor allele frequency (MAF) >0.05 using the F_{WS} statistic (Auburn et al., 2012, Manske et al., 2012) and DEploid (Zhu et al., 2018) (Figure 2a,b, Table S1). F_{WS} grades infections on a continuous scale of complexity where infections with an F_{WS}>0.95 are considered clonal and DEploid estimates the number of haplotypes (K) present in sequence data by jointly estimating haplotypes and their abundances. In close agreement with contemporary estimates of within-host diversity (Early et al., 2018) from the same location, 22 of 49 infections (44.9%) were considered clonal by F_{WS}. The within-host allele frequency (WHAF) captured from deep sequencing can be used to infer the presence of related parasites (Pearson et al., 2016). The patterns of unfixed mutations in the remaining 27 infections suggest a simple model of

superinfection is insufficient to universally capture all patterns of within-host relatedness (Data S1). Across the genome the WHAF in superinfected patients cluster around three values: fixed to the reference allele; fixed to the alternative allele; or at an intermediate frequency determined by the proportion of the two strains. Analysis of bulk sequencing data could not definitively resolve superinfection and co-transmission, so we selected 15 infections across the range of F_{WS} and inferred K (the number of haplotypes present) for single-cell sequencing, using a recently optimized method capable of near-complete genome capture (Trevino et al., 2017). The malaria parasite undergoes 4-5 rounds of DNA replication within a single-cell producing segmented schizont stage parasites with an average of 16 genome copies (Reilly et al., 2007). We isolate individual schizonts by Fluorescence Activated Cell Sorting (FACS), followed by whole genome amplification (WGA) under highly sterile conditions before sequencing the amplified product.

Single-cell sequencing of malaria parasites

In total we sequenced the genomes of 485 single-cells subjected to WGA (437 unique to this study), 49 bulk infections and 24 clones isolated from a single patient by limiting dilution (Nkhoma et al., 2012, Rosario, 1981). Prior to genotype filtering we scored 175,543 biallelic SNPs with a VQSLOD>0 across the 558 genome sequences. The highly repetitive and AT-rich *P. falciparum* genome (Gardner et al., 2002) presents unique challenges with generating an accurate picture of the variation present in a single-cell. We were particularly concerned with capture of DNA from more than one genetic background during the single-cell sequencing protocol and implemented stringent quality checks. Using sequencing data from the 24 clones we estimated the threshold for identifying single-cell sequences where there was potential contamination from exogenous DNA at 1% of mixed base calls. The sequences from the cloned lines were integrated into the single-cell dataset for downstream analysis. After excluding low coverage libraries (<75,000 calls, n=23) and sequences with >1% mixed base calls (n=38) 424 single-cell sequences remained. After including 23 of the sequences from *ex vivo* expanded clones there were 13-45 sequences per infection (mean 29.9 sequences; Figure S1). The number of haplotypes per sample attempted was estimated by rarefaction analysis (described below).

After quality control we retained 60,002 SNPs scored in at least 90% of the 496 sequences, 10,997 of which had a MAF>0.05 across the 49 bulk sequenced infections. As the 10,997 SNPs were ascertained from population data which did not undergo whole genome amplification this category of SNPs are unlikely to be artifacts from the MDA reaction, nor mutations arising during the course of an infection. Across our dataset parasites classified as clonally identical were identical at 99.97% of the original 60,002 sites, equivalent to a genome-wide error-rate of 1.2×10^{-6} mutations per basepair per cell, suggesting our approach is comparable (or superior) to leading single-cell sequencing methods (Leung et al., 2015, Hou et al., 2013, Lu et al., 2012). As an initial characterization of our data we estimated the genetic diversity in each infection from the number of unfixed sites from read pileups in bulk sequencing or across called genotypes in single-cell sequencing. For paired bulk/single-cell data from the same infection a mean of 1.6 fold (range 0.7-9.1 fold) more polymorphic sites were discovered by single-cell sequencing than by bulk sequencing (Figure S2). This is likely due to the limits in discovery of very low frequency SNPs by bulk sequencing. By subsampling our single-cell data we saw diminishing returns from sequencing additional cells, with 90% of the observed polymorphic sites captured by sampling a mean of 21.6 cells (range 7-43, Figure S2).

Haplotypic diversity of malaria infections

An important goal in malaria genomics is estimating the number of unique haplotypes (or complexity of infection) within an infection (Volkman et al., 2012). We estimated the number of

unique haplotypes directly from the single-cell data. To exclude potential confounding of *de novo* mutation and sequencing error, we restricted analysis to 10,997 conservatively called sites with a MAF >0.05 in the 49 bulk sequenced infections. We estimated the number of unique haplotypes per infection by collapsing haplotypes from the same infection that were different at <1% of sites. For each infection, we applied individual-based rarefaction to the haplotype abundances and sequenced additional single genomes until a plateau in the rarefaction curve was reached (Figure S2). Using this approach, between 1 and 17 haplotypes were observed in each infection (Figure 1C, Table S1). Rarefaction of haplotype abundance suggested we had captured all haplotypes present in 14/15 infections. For these 14 samples the number of haplotypes captured was within the 95% confidence interval of the Chao I estimator. One infection (MAL15) showed exceptionally high diversity with 17 of an estimated 30.21 (95% CI=19.7-81.7) haplotypes detected. Two infections (MAL37 and MAL33) show a single haplotype from single-cell sequencing, although F_{WS} scores <0.95 and patterns of segregating sites suggest we have incompletely captured all haplotypes (Data S1). Sequencing more cells did not capture additional haplotypes.

Inference of infection composition from bulk sequences

There has been a concerted effort to develop statistical methods for inferring the parasite haplotypes within complex infections using information from bulk sequence data (Chang et al., 2017, Zhu et al., 2018). Such methods aim to phase haplotype data from complex mixtures, extending the methods used to infer haplotypes from diploid genotypes. Our single-cell resolution data from natural infections provides "gold standard" data for comparison with inferences. We found a strong correlation between the effective number of haplotypes (Zhu et al., 2018) observed by single-cell sequencing and the effective K from DEploid (Pearson's r^2=0.61) and F_{WS} (Pearson's r^2=-0.51, Figure 3A,B). The effective K and effective number of haplotypes normalizes the absolute haplotype number by the abundance of each haplotype and are distinct from the number of haplotypes presented in Figure 2. As within-host abundance is accounted for in these measures (and in F_{WS}) it is not surprising these agree more closely than when comparing absolute haplotype number or Inferred K**.** However, while DEploid performed well in determining the predominant haplotype present within infections, it was unable to accurately determine the additional haplotypes present within infections (Figure 3C). The poor performance most likely stems from the assumption that haplotypes found within infections are unrelated. We suggest that incorporating relatedness into these models may improve performance of these inference methods. Importantly, our data provide a suitable dataset for optimizing and improving these statistical models.

Recent ancestry of individual infections

The size of chromosomal blocks that are shared identical-by-descent (IBD) between infections provides a metric for assessing parasite relatedness: recent relatives share large blocks, while in distant relatives these blocks are smaller because they have been broken up by recombination events. As recombination occurs in the mosquito (Figure 1) related parasites are likely to have arisen from a single mosquito bite and have been co-transmitted. To better characterize levels of relatedness within infections we identified blocks of chromosomes shared IBD between all paired sequences using a hidden Markov model (Schaffner et al., 2018). IBD sharing between clonal bulk sequenced infections was rare, with a mean of 0.73 blocks shared between infections (range 0-5), encompassing a mean of 88.5kb (range 3.8-342.7kb) of each genome, with a mean block length of 50.8kb (range 3.8-142.4kb). In contrast, within infections parasites shared a mean of 13.0 (range 0-30) IBD blocks between parasite genomes, encompassing a mean of 16,334.2kb (range 3.1-20,577.0kb) of each genome, with a mean shared block length of 1,143.6kb (range 3.1-1469.8kb). As we limit inference of IBD to the 'core' genome (Miles et al., 2016) identical parasites share 20,577kb of their genomes IBD in 14

blocks (one per chromosome). The presence of IBD sharing between individuals supports recent shared ancestry. For 10/15 of the infections there was at least one block of IBD shared in all pairwise comparisons. As our filtering of IBD blocks was limited to >2.5cM we are limited to inference of relatedness over the last 25 generations (~6 years (Henden et al., 2018)).

Recent studies have highlighted the power of IBD networks to capture the structure of a parasite population (Henden et al., 2018). We built a network of pairwise shared IBD, creating links between parasites with >15% of their genomes shared IBD (Figure 4, Data S2). This revealed close connectivity between parasites from the same infection, with much sparser connectivity between parasites from different infections. We observed subdivision within individual infections. For instance, MAL5 and MAL24 form two clusters of parasites that were connected by the sequence derived from bulk sequencing to one another in agreement with expectations of superinfection. MAL15 and MAL23 show either direct connections or indirect connections (passing through another genotype) between all parasites in agreement with the expectations of co-transmission. Varying the minimum IBD required to connect genomes allowed us to visualize how relatedness subdivides individual infections across a range of IBD sharing (Data S2).

The distribution of total pairwise shared IBD and the average shared block lengths can be used to infer the relationships between individual genomes (Huff et al., 2011, Li et al., 2014). We inferred the degree of relatedness from our data using the Estimation of Recent Shared Ancestry (ERSA) algorithm. Under this scheme 0 denotes identical clones, 1 denotes parent-sibling, 2 denotes full/half siblings with higher numbers denoting increasingly distant relationships. ERSA estimates relatedness between individuals from distribution of IBD tract lengths (Figure 5A, B, Figure S3) using individuals assumed to be unrelated from the same population as a reference. We see a spectrum of relationships within each infection (Figure 5C). In MAL5 this confirmed the lack of relatedness between the two clusters of parasites, suggesting this infection was the result of a genuine superinfection. However, no other infections can be classified so simply, commonly showing relationships as distant as 4th degree (equivalent to ‘first cousins’). Within our data this suggests that it is not uncommon for parasites to be transmitted through two generations (human-mosquito-human-mosquito-human), with up to four generations of co-transmission seen in our data in infection MAL24 and MAL17. In our data we see only a single unambiguous instance of superinfection of two unrelated parasites with no concurrent co-transmission (MAL5). Across the analysis, the genetic diversity of three infections (MAL17, MAL24 and MAL48) appears to be driven by both superinfection of unrelated parasites and co-transmission of related parasites (in addition to MAL5 where only superinfection is suspected).

Mitochondrial inheritance within individual infections

Mitochondria are inherited maternally during malaria parasite meiosis and therefore allow the identification of parasites which share a maternal lineage within infections. We found 93 SNPs in the mitochondrial DNA which varied within our dataset. Across all 498 parasites we were able to capture the genotype of 79.1% (36,645/46,314) of these 93 sites. We expect parasites which are identical across their nuclear chromosomes to also be identical across their mitochondrial genome. We find this to be the case. Within individual hosts, parasites which shared 100% of their genome IBD also shared 100% of their mitochondrial genome sequence. Across the 498 genome sequences for which we had reliable estimates of genome-wide IBD and mitochondrial genotypes there were only 2 of a possible 36,645 sites (0.0055%) where the mitochondrial genotype varied within a group of parasites which were 100% IBD. These arise at position 1,692 in cells 9 and 22 of infection MAL24. Visual inspection of these sites (Figure S4) and the presence of the reads supporting both genotypes in the bulk genome sequence supports these being genuine. However, as these may have arisen as *de novo* mutations during the current

infection, or be artifacts of the WGA and genome sequencing pipeline we excluded this site from further analysis.

We identified a total of 20 unique mitochondrial haplotypes across the entire dataset (Figure S5 and Figure S6). We counted the number of mitochondrial haplotypes present in each infection (Figure 5D). This showed 9 infections contained a single mitochondrial haplotype, 3 infections contained 2 haplotypes, 1 infection contained 3 haplotypes, and 2 infections contained 4 haplotypes. All monoclonal infections contained a single mitochondrial haplotype, as did 4/10 (40%) polyclonal infections. Within polyclonal infections distinct mitochondrial haplotypes were observed between very close relatives (sharing >60% of their genomes IBD) supporting the retention of diversity we observe in the face of recurrent inbreeding. As mitochondria are uniparentally inherited the presence of multiple mitochondrial haplotypes across related parasites from within the same infection demonstrates multiple oocysts within a mosquito are responsible for initiating an infection. We tested if a higher level of shared pair-wise IBD is a predictor of sharing a mitochondrial haplotypes between two parasites. After excluding clonally identical parasites we estimate a 10% increase in IBD shared between parasites from the same infections increasing the odds of sharing a mitochondrial genotype by a factor of 10.082 (p=5.25x10^{-8}, logistic regression).

Identification of meiotic siblings and reconstruction of parental haplotypes

During meiosis a tetrad of four recombinant progeny arise from a single paired set of chromosomes. Genetic characterization of tetrads has enabled precise capture of tracts of gene conversion, cross-overs and non-cross-overs in yeast, mammalian and plant genetics (Cole et al., 2014, Hou et al., 2013, Mancera et al., 2011, Li et al., 2015). Identifying progeny which came from the same tetrad, rather than simply sharing the same parents, allows us to estimate the number of ookinetes giving rise to an infection. As meiotic siblings share reciprocal recombination breakpoints, it also enables a fine-scale measurement of gene conversion and non-Mendelian inheritance in natural infections. Chance identification of potential meiotic siblings in low transmission regions (Dharia et al., 2010) and *P. vivax* (Bright et al., 2014) has suggested that tetrad analysis may be feasible. We identified potential members of tetrads as those sharing the same mitochondrial genotype with a coefficient of relatedness of 1, which would contain half-sibs, full-siblings, and meiotic siblings. For each group we inferred the parental haplotypes and recombination breakpoints using Hapi (Li et al., 2018). Following this we filtered out parasites which did not share reciprocal breakpoints consistent with shared meiosis.

Across the dataset we were able to capture one complete tetrad, from infection MAL15 (Figure 6). In the absence of complete tetrads it is difficult to definitively conclude parasites came from the same meiosis, or simply share identical parents. However, we found 10 groups of parasites sharing reciprocal breakpoints between 2-4 genotypes (inclusive of the tetrad from MAL15) suggesting they arose from the same meiosis (Data S3). Based on these potential mating's we suggest the minimum number of ookinetes required in an infected mosquito for each infection as between 1 and 5 (Table S1). Importantly, despite exhaustive capture of genetic diversity in nearly all infections, most members of a tetrad do not reach observable frequencies in the blood stream. This suggests there is a considerable attrition in the representation of progeny between the mosquito midgut and the human blood stage. We estimated crossover locations across the dataset using a total of 570 unique crossovers. As seen in previous studies on *P. falciparum* genetic crosses (Jiang et al., 2011) chromosome length and crossover count were correlated (r^2=0.9, p=8.43x10^{-6}, Pearson's correlation, Figure S7).

The capture of a complete tetrad allows us to observe non-Mendelian inheritance in a natural malaria genetic cross for the first time. We identified 33 recombination events across the genome, with no cross-overs detected on chromosomes 1,7 and 9. We found evidence for 2 major tracts of skewed inheritance resulting from gene conversion. Adjacent to the right-most telomere of chromosome 2 a tract of 872 bp covering 18 markers, and adjacent to the left-most telomere of chromosome 4 a tract of 180,970bp. Within the chromosome 4 region there is a block of 19,432bp from one parent, and 161,538bp from the other. In addition to regions of skewed inheritance we saw 2 large regions (770,168bp on chromosome 8 and 51,426bp on chromosome 9) where variation between the parents had been eliminated by inbreeding. Other regions with no variation distinguishing parents surrounded known hypervariable genes (Miles et al., 2016), and were not included in our analysis.

DISCUSSION

There has been a concerted effort to understand the complexity of malaria infections from either deep sequencing data (Assefa et al., 2014, O'Brien et al., 2016, Zhu et al., 2018), or from genotyping a limited number of markers (Chang et al., 2017, Galinsky et al., 2015). We show here there is considerable depth to complex infections which may be challenging to infer from bulk analysis alone. Through a combination of deep sequencing of bulk infections and single-cell sequencing, we have generated the most comprehensive picture of the within-host diversity of malaria infections to date. This provides a much-needed standard for developing novel tools for probing the complexity of infections from deep sequencing data. By using multiple estimates of relatedness targeting distinct features of the data, we argue that most complex infections result from parasites co-transmitted from single mosquito bites in our dataset. Strikingly, our analysis supports only a single infection where simple superinfection of two unrelated strains has occurred (MAL5), and a further three infections where both superinfection and co-transmission have concurrently contributed to diversity (MAL17, MAL24, MAL48). The remaining infections were either monomorphic or showed strong support for co-transmission of related strains only. Notably, the most diverse infection we studied (MAL15) was explained entirely by co-transmission of related parasites. In the two infections where we were unable to capture the minor strains (MAL33 and MAL37), patterns of unfixed SNPs within the infection suggest the uncaptured strain was related to the captured strain (Data S1). It may be that the minor isolate failed to develop to DNA-rich late stages or was present at a fraction lower than we were able to sample within the constraints of this work.

In this work we have generated a detailed picture of within-host genetic variation across fifteen malaria patients. While this is a modest number of infections, these findings are directly informative about the limits of malaria complexity. We have surveyed a high transmission setting and focused on polyclonal infections. This sampling enriches for infections with the highest likelihood of genetic diversity arising from superinfection rather than co-transmission. In spite of this we find co-transmission to be widespread, and likely underappreciated as a mechanism generating and maintaining genetic diversity in natural malaria populations. There is a pressing need to extend these observations across the range of malaria endemicity to fully capture the transmission network of malaria infections.

These results could not have been obtained by statistical inference of bulk sequence data alone. There have been impressive advances in imputation of individual parasite haplotypes from bulk sequence data with the development of DEploid (Zhu et al., 2018). We identify two to seventeen haplotypes in the infections we dissect here, with 50% of our infections bearing six or more unique haplotypes. As DEploid typically limits inference to five haplotypes, reliance on bulk inference would have discounted information on the additional haplotypes. When we examine the haplotypes inferred by DEploid we see the accuracy of imputation is not equivalent

across all haplotypes in an infection (Figure 3C). Majority haplotypes inferred by DEploid share high similarity (median of 97.8%) to single-cell haplotypes, suggesting DEploid is a highly effective at capturing common haplotypes. However, minority haplotypes (5-50% abundance) were captured poorly (median 84.3% similarity). Encouragingly, we see particularly good performance for DEploid in inferring the composition of a single infection with two unrelated haplotypes (MAL5). The relatively poor performance of inference in other infections suggests that incorporating inbreeding and complex relatedness structures may lead to algorithmic improvements. In general, the data we present here provides an ideal training set for improvements in the implementation of statistical tools for understanding polyclonal infections and in generating guidelines for the interpretation of haplotypes inferred from bulk sequence data.

Only parasites which transmit gametes to the same mosquito can produce recombinant offspring. Patterns of parasite diversity and relatedness within individual mosquitoes (Annan et al., 2007) (albeit in a distinct population) are in general agreement with our results – most mating is between related parasites. The mechanisms underlying why inbreeding is common, even in high transmission settings, is less clear. Malaria transmission is intense in Chikhwawa (Mzilahowa et al., 2012) and we expected superinfection to be more prevalent than we observed. A mechanism controlling the outcome of superinfection, perhaps by hepcidin based inhibition of liver development in superinfecting sporozoites (Portugal et al., 2011), could explain why we do not see more superinfection. Alternatively, the low numbers of superinfecting parasites emerging from the liver relative to those present in established infections (which may contain $10^{11\text{-}12}$ blood stage parasites) may limit establishment of superinfections. This would represent an infectious disease example of the “priority” effects (De Meester et al., 2016) that are important in determining assemblies of ecological communities or microbiomes. Immune-mediated selection of parasite variants sharing alleles at the major antigenic loci (Farnert et al., 2008) could also generate a strong relatedness structure within infections. However, in this study, the effects of host immunity are likely to be limited because only infections from children who have little or no pre-existing malaria immunity were studied. In analyzing why superinfection is less common, it is also worth noting that analysis of parasite diversity is generally limited to single blood draws due to the need to treat symptomatic patients expediently. As this sampling strategy may overlook sub-populations circulating at lower frequencies, there may be additional genetic variation which escapes routine analysis. Most importantly, the strong relatedness structure we observe will limit the amount of outbreeding in malaria parasites, even in regions of intense transmission. Restrained outbreeding amongst malaria parasites could profoundly shape the evolution of parasite virulence, drug resistance and malaria transmission dynamics as shown in mice (Wargo et al., 2007a, Wargo et al., 2007b, Huijben et al., 2010, Huijben et al., 2011, Alizon, 2013).

The depletion of genetic variation during repeated rounds of co-transmission has been previously modelled (Wong et al., 2018), suggesting a substantial decline in the number of clonal lineages and an increase in average relatedness can arise through a single transmission cycle. We directly observe this by reconstructing tetrads. In all but one case we do not observe complete tetrads in our data. Some members of the tetrads were either lost during the infection or were present at too low a frequency to be sampled in this study. Either of these scenarios suggest that genetic variation will be depleted over successive transmission cycles, a prospect directly explored in Figure 5. Our data suggest complex infections comprise parasites which have been co-transmitted longer than two transmission cycles. Due to the lengthy developmental cycle of the parasite, and the propensity for mosquitoes to disperse the potential for these patterns to be driven by local population structure is minimal (Conway and McBride, 1991, Prugnolle et al., 2008). We observe that substantial genetic variation is maintained

despite the bottleneck of mosquito transmission with up to 17 unique haplotypes likely inoculated by a single mosquito. Understanding how patterns of transmission and within host dynamics contribute to the diversity and relatedness structure within malaria infections will help inform ongoing elimination and control efforts.

ACKNOWLEDGEMENTS

We thank all children and their families who participated in this study in Chikhwawa, Malawi. We also thank Andrew Mtande, Ruth Daiman and Miriam Phiri for their help with participant recruitment and clinical management. We are also grateful to Clement Masesa and Lumbani Makhaza for designing our data collection tool and for managing the study database. This study was supported by a Wellcome Trust Intermediate Fellowship in Tropical Medicine and Public Health (Grant # 099992/Z/12/Z to SCN) and an NIH grant (NIAID AI110941-01A1 to IHC). FACS data were generated in the Flow Cytometry Shared Resource Facility (supported by UTHSCSA, NIH-NCI P30 CA054174, UL1RR025767 (CTSA)).

AUTHOR CONTRIBUTIONS

S.C.N., S.G.T., S.A.W., D.J.T, T.J.C.A and I.H.C designed the study. S.G.T. and I.H.C. developed tools. S.C.N., S.G.T., K.G., S.N., A.D., C.J., R.G., B.D., and I.H.C. performed experiments. S.C.N., S.K., and D.J.T. collected samples. S.C.N, S.G.T., T.J.C.A and I.H.C. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests. Correspondence and requests for materials should be addressed to ianc@txbiomed.org.

Figure Legends

Figure 1. The within host genetic diversity of malaria parasites is shaped by mosquito transmission. (A) A simple monoinfection is generated when an uninfected individual is bitten by a mosquito bearing a single parasite genotype. (B) A superinfection occurs when an individual is bitten by two mosquitoes, each bearing a single parasite genotype. (C) Co-transmission of parasites occurs when a single mosquito bearing multiple genetically distinct parasites bites an uninfected individual. As genetic recombination is an obligate stage of mosquito transmission multiple related parasites may infect an individual through this route.

Figure 2. Complexity of infection inferred from bulk and single-cell sequencing. (A) F_{WS} scores for 49 bulk sequenced infections. Infections above the dashed line (F_{WS}=0.95) are assumed to be clonal. (B) Inferred number of haplotypes (K) inferred by DEploid, infections are ordered by the F_{WS} score. Black dots in (A) and (B) denote infections also deconvoluted by single-cell sequencing. (C) Number of unique haplotypes inferred by single-cell sequencing. This data is included in Table S1.

Figure 3. Inferring the composition of malaria infections from bulk sequencing data. Correlation between the effective number of haplotypes inferred from single-cell sequencing and either F_{WS} (A) or Effective K inferred from DEploid (B). (C) The maximum similarity between each haplotype inferred by DEploid and single-cell sequences from the same infection. The abundance of the haplotype (as estimated by DEploid) for each haplotype is shown on the y-axis. Majority haplotypes were inferred with high accuracy (73.6-99.9% similarity), with reduced accuracy for minority haplotypes. After exclusion of low abundance (<10%) haplotypes and haplotypes from clonal infections there was a significant relationship between abundance and maximum similarity (adjusted r^2=0.32, p=4.6x10^{-5}, linear model). Clonal infections (MAL29,

MAL31, MAL38) and simple superinfections (MAL5) perform uniformly well with DEploid, though other infections show variable success.

Figure 4. A network representation of pairwise IBD sharing across the genomes. Each node represents a single parasite colored by the infection of origin. Nodes are joined if >15% of the genomes are shared IBD. Each node is colored by the infection it was derived from, with bulk sequences denoted by a square and single-cell sequences by a circle. The parasite from a single infection are highlighted by dashed lines. MAL23 and MAL48 both contain multiple unlinked clusters indicative of superinfected parasites.

Figure 5. Recent ancestry inferred from IBD sharing. (A) Density plot of the total IBD shared between parasites from a single infection (labeled to the left of the plot). (B) Density plot of the mean IBD block length between parasites from a single infection. The dotted lines in A and B show the expected value for parasites separated by differing number of meiosis, e.g. clonally identical (~21Mb total IBD/~1.5Mb Mean IBD length) and separated by a single meiosis (~10Mb total IBD/~0.6Mb Mean IBD length). The most distant relationship shown is 5 meiosis. (C) Relative frequency of different degrees of relatedness inferred between parasites from the same infection using the ERSA algorithm (ns - no significant relatedness observed). (D) The number of mitochondrial haplotypes identified in this infection.

Figure 6. Patterns of recombination in a tetrad formed during meiosis. Two parental haplotypes (red and blue) were inferred using Hapi and the inheritance of these haplotypes was inferred in the four parasite genotypes shown here. Across the genome there is consistent 1:1 inheritance of parental genomes aside from the proximal end of chromosome 4 and the distal end of chromosome 2 (green boxes, also shown expanded). Inbreeding among the ancestors of the parental genotypes has eroded variation on chromosomes 8 and 9 (orange boxes).

STAR METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ian Cheeseman (ianc@txbiomed.org). Raw sequence data has been deposited at the sequence read archive (https://www.ncbi.nlm.nih.gov/sra*)* under study number SRP155167.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Malaria-infected blood samples (5 ml; thin smear parasitaemia: 0.2 to 21.8%) were obtained prior to treatment from children aged 19 to 116 months old presenting to Chikhwawa District Hospital in Malawi with uncomplicated *P. falciparum* malaria from February to June 2016. Blood samples were collected only from children whose parents or legal guardians provided consent. Ethical approval for this study was obtained from the University of Malawi College of Medicine Research and Ethics Committee (Protocol number P.02/13/1528) and the Liverpool School of Tropical Medicine Research Ethics Committee (Protocol number 14.035). Each research subject was assigned a unique ID number at the time of enrollment. Similarly, each blood sample was assigned a unique barcode at the time of collection. The sample barcode was used to identify and track each sample while being processed and analyzed in the laboratory. Detailed information about all research subjects and blood samples they donated are provided in Table S1.

METHOD DETAILS

Sample Collection

Venous blood (5ml) from each subject was collected directly into an Acid Citrate Dextrose tube (BD, UK). The sample was immediately placed in an ice-cold container and transported to our laboratory in Blantyre within six hours of collection. Half of each blood sample was washed using incomplete RPMI 1640 media (Sigma-Aldrich, UK). Three aliquots of the sample were cryopreserved in glycerolyte 57 solution (Fenwal, Lake Zurich, IL, USA) and stored in liquid nitrogen. Parasites used in fluorescence-activated single-cell sorting were cultured from one of these aliquots. The second half of the sample was filtered using CF11 columns to deplete human leucocytes (Venkatesan et al., 2012), and was stored at -80^{0}C until needed. Parasite DNA was extracted from this sample using a DNA Mini Kit (QIAGEN, USA) and directly sequenced on an Illumina HiSeq instrument.

Single-cell capture

We performed single-cell capture, whole genome amplification, sequencing analysis and followed guidelines for preventing contamination using the approaches described in (Nair et al., 2014, Trevino et al., 2017). We outline these approaches below.

Parasite culture

Approximately 1 mL of cryopreserved blood sample was thawed at 37^{0}C to revive intact cells (~200ul recovered pellet, ~1% parasitemia). The sample was washed twice by adding 10mL of complete media (filter-sterilized incomplete RPMI 1640 media to which 5% w/v of hypoxanthine and 8% w/v of albumax II were added). Following the final centrifugation step (425 x g for 5 minutes) cells were resuspended and grown in 8 mL complete media in a sealed T25 tissue culture flask flushed with 5% CO_2, 5% O_2 and 90% N_2 prior to being sealed. The culture flask was incubated at 37^{0}C for 40 hours to allow for parasite progression to late stages, which generates higher quality genomic data after MDA and library preparation (Trevino et al., 2017). To stain parasitized cells in readiness for FACS, ~8 µl of an infected red blood cell pellet (~10^8 cells) was resuspended in 10 mL of 1X PBS (Lonza, USA) which included 5 µl of Vibrant DyeCycle Green at 37^{0}C for 30 minutes with intermittent manual inversion of the tube approximately every 10 minutes. Cells were washed once in 1X PBS and resuspended in 5 - 8 mL of 1X PBS in a foil-covered tube to protect the dye from photobleaching in preparation for FACS sorting. Catalog numbers for the critical reagents used in parasite culture and capture of single parasitized cells are provided in the *Key Resources Table*.

FACS Sorting

Cells were sorted by MoFlo Astrios (Beckman Coulter) by gating the two brightest observed populations according to DNA fluorescence, the sort was run in single-cell sort mode with a drop envelope of 0.5. Individual cells were sorted into 0.2 mL PCR tubes containing 5 µl autoclaved sterile PBS (Lonza), which had been prepared under sterile conditions in a PCR hood. Each event required about 15 seconds to open the tube, place on the sorting rack, recover, and close the tube. Tubes were then immediately stored on dry ice and transferred to -80 °C longer-term storage within an hour.

Generation of single-cell DNA libraries

Library preparation for individually sorted late-stage parasites was carried out using the Qiagen Single-Cell FX DNA kit without library amplification according to manufacturer's instructions. Whole genome amplification preparation was carried out under a PCR hood and DNA was amplified on a dedicated PCR machine. Library products were analyzed by TapeStation and included off-target peaks typical of MDA DNA inputs. Adapter-ligated DNA products were quantified by KAPA Hyperplus Kits. All sequencing was performed on an Illumina HiSeq 2500. A detailed description of the development of the single-cell sequencing methods used here, and the protocols in place to control for contamination are available (Nair et al., 2014, Trevino et al.,

2017). Raw sequence data has been deposited at the sequence read archive (https://www.ncbi.nlm.nih.gov/sra) under study number SRP155167.

Sequence analysis

We aligned raw sequencing reads to v3 of the 3D7 genome reference (http://www.plasmodb.org) using BWA MEM v0.7.5a (Li, 2013). After removing PCR duplicates and reads mapping to the ends of chromosomes (Picard v1.56) we recalibrated base quality scores, realigned around indels and called genotypes using GATK v3.5 (DePristo et al., 2011) in the GenotypeGVCFs mode using QualByDepth, FisherStrand, StrandOddsRatio VariantType, GC Content and max_alterate_alleles set to 6. We recalibrated quality scores and calculated VQSLOD scores using SNP calls conforming to Mendelian inheritance, excluding sites where the VQSLOD score was <0. Median read depth of WGA single-cells was 28.3 (interquartile range (IQR) 12.5-46.4) with median of 90.5% (IQR 78.1-96.0%) of the genome covered by at least one read. In contrast the non-WGA samples had a median read depth of 31.11 (IQR 20.93-48.37) and a median of 95.8% (IQR 93.1-97.4%) of the genome covered by at least one read. A potential source of error in single-cell genomics is the inclusion of exogenous DNA amplified alongside the target genome in downstream analysis. As an initial indication of the potential of non-target DNA being introduced to our analysis we first examined the proportion of reads mapping to the *P. falciparum* genome (Gardner et al., 2002) in each sequence. We observed a median of 93.3% (IQR 87.0-95.4%) of reads map to the parasite genome for single-cell sequences, compared to 35.7% (IQR 19.7-48.5%) for bulk patient samples and 79.4% (IQR 74.5-86.9%) for clonally expanded samples suggesting our stringent handling protocols were effective at eliminating environmental DNA. For a more rigorous test we identified lines with potential cross contamination based on unfixed basecall frequency. As the parasite genome is haploid during blood stages all variants are expected to be fixed in genome sequencing data. The highly AT-rich and repetitive nature of the parasite genome makes alignment challenging, generating false positive unfixed variants in clonal lines. After excluding highly error-prone genomic regions (calls outside of the “core genome”(Miles et al., 2016) or within microsatellites) we measured the proportion of mixed base calls (>5% of reads at a locus mapping to the minority allele) at high confidence biallelic SNPs (>10 reads mapped, VQSLOD>0, GQ>70). Using the cloned lines and bulk population samples as a guide we estimated 1% as an appropriate threshold for excluding putatively mixed lines (Figure S1).

QUANTIFICATION AND STATISTICAL ANALYSIS

Estimating the complexity and diversity of bulk sequenced samples

F_{WS} was calculated in moimix (https://github.com/bahlolab/moimix) for all bulk patient samples. We estimated the number of unique haplotypes and their sequence from deep sequence of bulk infections using DEploid (Zhu et al., 2018) v0.5 (https://github.com/mcveanlab/DEploid). We used 10,997 HQ SNPs with a MAF >5%. For a reference panel we used 10 bulk Malawian samples presumed to be clonal (F_{WS}>0.95) and population level allele frequencies from across the complete bulk sequencing data. We inferred the most likely number of haplotypes (K) using the command:

./dEploid -ref sample_reference_allele_counts.txt -alt sample_alternative_allele_counts.txt -plaf population_allele_freq.txt -o sample_out -ibd -noPanel -exclude highly_variable_sites.txt -sigma 7 -seed 2

Estimating relatedness between sequences

SNP data were imported into R using SeqArray (Zheng et al., 2017). Between all samples passing quality control we calculated the proportion of shared alleles and using SNPs which

were at >5% MAF in the bulk sequenced samples. We used a distance matrix generated from this data (1-pairwise allele sharing) to estimate the number of unique haplotypes in each infection by collapsing together sequences which differed at <1% of sites. Rarefaction of haplotype abundance was performed using the rareNMtests package (Cayuela and Gotelli, 2014) in R. We called regions of IBD between all samples passing quality control using hmmIBD v2.0.0 (Schaffner et al., 2018) (https://github.com/glipsnort/hmmIBD). We performed maximum-likelihood estimation of recent shared ancestry using ERSA 2.0 (Huff et al., 2011, Li et al., 2014) (http://www.hufflab.org/software/ersa/) using the output from hmmIBD using the flags --min_cm=1.5 --adjust_pop_dist=true --number_of_chromosomes=14 --rec_per_meioses=19. We converted the basepair positions to a uniform genetic map using the scaling factor 1cM=9.6kb (Jiang et al., 2011) and excluded IBD chunks <1cM in length. As identical clones are not specifically modelled in ERSA we excluded these from analysis, though their abundance is shown in the '0' bar in Figure 5C. All other statistical analysis and visualization was performed in R v3.4.0 (Team, 2017). Inference of parental genotypes and recombination breakpoints was performed using Hapi.

DATA AND CODE AVAILABILITY

This study did not generate any novel code. The data generated during this study are available from the sequence read archive (https://www.ncbi.nlm.nih.gov/sra) under study number SRP155167.

REFERENCES

ALIZON, S. 2013. Parasite co-transmission and the evolutionary epidemiology of virulence. *Evolution,* 67**,** 921-33.

ANNAN, Z., DURAND, P., AYALA, F. J., ARNATHAU, C., AWONO-AMBENE, P., SIMARD, F., RAZAKANDRAINIBE, F. G., KOELLA, J. C., FONTENILLE, D. & RENAUD, F. 2007. Population genetic structure of Plasmodium falciparum in the two main African vectors, Anopheles gambiae and Anopheles funestus. *Proc Natl Acad Sci U S A,* 104**,** 7987-92.

ASSEFA, S. A., PRESTON, M. D., CAMPINO, S., OCHOLLA, H., SUTHERLAND, C. J. & CLARK, T. G. 2014. estMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics,* 30**,** 1292-4.

AUBURN, S., CAMPINO, S., MIOTTO, O., DJIMDE, A. A., ZONGO, I., MANSKE, M., MASLEN, G., MANGANO, V., ALCOCK, D., MACINNIS, B., ROCKETT, K. A., CLARK, T. G., DOUMBO, O. K., OUEDRAOGO, J. B. & KWIATKOWSKI, D. P. 2012. Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. *PLoS One,* 7**,** e32891.

BELL, A. S., DE ROODE, J. C., SIM, D. & READ, A. F. 2006. Within-host competition in genetically diverse malaria infections: parasite virulence and competitive success. *Evolution,* 60**,** 1358-71.

BRIGHT, A. T., MANARY, M. J., TEWHEY, R., ARANGO, E. M., WANG, T., SCHORK, N. J., YANOW, S. K. & WINZELER, E. A. 2014. A high resolution case study of a patient with recurrent Plasmodium vivax infections shows that relapses were caused by meiotic siblings. *PLoS Negl Trop Dis,* 8**,** e2882.

CAYUELA, L. & GOTELLI, N. J. 2014. rareNMtests: Ecological and biogeographical null model tests for comparing rarefaction curves.

CHANG, H. H., WORBY, C. J., YEKA, A., NANKABIRWA, J., KAMYA, M. R., STAEDKE, S. G., DORSEY, G., MURPHY, M., NEAFSEY, D. E., JEFFREYS, A. E., HUBBART, C., ROCKETT, K. A., AMATO, R., KWIATKOWSKI, D. P., BUCKEE, C. O. & GREENHOUSE, B. 2017. THE REAL McCOIL: A method for the concurrent estimation

of the complexity of infection and SNP allele frequency for malaria parasites. *PLoS Comput Biol,* 13**,** e1005348.

COLE, F., BAUDAT, F., GREY, C., KEENEY, S., DE MASSY, B. & JASIN, M. 2014. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat Genet,* 46**,** 1072-80.

CONWAY, D. J., GREENWOOD, B. M. & MCBRIDE, J. S. 1991. The epidemiology of multiple-clone Plasmodium falciparum infections in Gambian patients. *Parasitology,* 103 Pt 1**,** 1-6.

CONWAY, D. J. & MCBRIDE, J. S. 1991. Genetic evidence for the importance of interrupted feeding by mosquitoes in the transmission of malaria. *Trans R Soc Trop Med Hyg,* 85**,** 454-6.

DE MEESTER, L., VANOVERBEKE, J., KILSDONK, L. J. & URBAN, M. C. 2016. Evolving Perspectives on Monopolization and Priority Effects. *Trends Ecol Evol,* 31**,** 136-146.

DE ROODE, J. C., PANSINI, R., CHEESMAN, S. J., HELINSKI, M. E., HUIJBEN, S., WARGO, A. R., BELL, A. S., CHAN, B. H., WALLIKER, D. & READ, A. F. 2005. Virulence and competitive ability in genetically diverse malaria infections. *Proc Natl Acad Sci U S A,* 102**,** 7624-8.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., MCKENNA, A., FENNELL, T. J., KERNYTSKY, A. M., SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43**,** 491-8.

DHARIA, N. V., PLOUFFE, D., BOPP, S. E., GONZALEZ-PAEZ, G. E., LUCAS, C., SALAS, C., SOBERON, V., BURSULAYA, B., KOCHEL, T. J., BACON, D. J. & WINZELER, E. A. 2010. Genome scanning of Amazonian Plasmodium falciparum shows subtelomeric instability and clindamycin-resistant parasites. *Genome Res,* 20**,** 1534-44.

EARLY, A. M., LIEVENS, M., MACINNIS, B. L., OCKENHOUSE, C. F., VOLKMAN, S. K., ADJEI, S., AGBENYEGA, T., ANSONG, D., GONDI, S., GREENWOOD, B., HAMEL, M., ODERO, C., OTIENO, K., OTIENO, W., OWUSU-AGYEI, S., ASANTE, K. P., SORGHO, H., TINA, L., TINTO, H., VALEA, I., WIRTH, D. F. & NEAFSEY, D. E. 2018. Host-mediated selection impacts the diversity of Plasmodium falciparum antigens within infections. *Nat Commun,* 9**,** 1381.

FARNERT, A., LEBBAD, M., FARAJA, L. & ROOTH, I. 2008. Extensive dynamics of Plasmodium falciparum densities, stages and genotyping profiles. *Malar J,* 7**,** 241.

GALINSKY, K., VALIM, C., SALMIER, A., DE THOISY, B., MUSSET, L., LEGRAND, E., FAUST, A., BANIECKI, M. L., NDIAYE, D., DANIELS, R. F., HARTL, D. L., SABETI, P. C., WIRTH, D. F., VOLKMAN, S. K. & NEAFSEY, D. E. 2015. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malar J,* 14**,** 4.

GARDNER, M. J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R. W., CARLTON, J. M., PAIN, A., NELSON, K. E., BOWMAN, S., PAULSEN, I. T., JAMES, K., EISEN, J. A., RUTHERFORD, K., SALZBERG, S. L., CRAIG, A., KYES, S., CHAN, M. S., NENE, V., SHALLOM, S. J., SUH, B., PETERSON, J., ANGIUOLI, S., PERTEA, M., ALLEN, J., SELENGUT, J., HAFT, D., MATHER, M. W., VAIDYA, A. B., MARTIN, D. M., FAIRLAMB, A. H., FRAUNHOLZ, M. J., ROOS, D. S., RALPH, S. A., MCFADDEN, G. I., CUMMINGS, L. M., SUBRAMANIAN, G. M., MUNGALL, C., VENTER, J. C., CARUCCI, D. J., HOFFMAN, S. L., NEWBOLD, C., DAVIS, R. W., FRASER, C. M. & BARRELL, B. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature,* 419**,** 498-511.

HENDEN, L., LEE, S., MUELLER, I., BARRY, A. & BAHLO, M. 2018. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS Genet,* 14**,** e1007279.

HOU, Y., FAN, W., YAN, L., LI, R., LIAN, Y., HUANG, J., LI, J., XU, L., TANG, F., XIE, X. S. & QIAO, J. 2013. Genome analyses of single human oocytes. *Cell,* 155**,** 1492-506.

HUFF, C. D., WITHERSPOON, D. J., SIMONSON, T. S., XING, J., WATKINS, W. S., ZHANG, Y., TUOHY, T. M., NEKLASON, D. W., BURT, R. W., GUTHERY, S. L., WOODWARD, S. R. & JORDE, L. B. 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res,* 21**,** 768-74.

HUIJBEN, S., NELSON, W. A., WARGO, A. R., SIM, D. G., DREW, D. R. & READ, A. F. 2010. Chemotherapy, within-host ecology and the fitness of drug-resistant malaria parasites. *Evolution,* 64**,** 2952-68.

HUIJBEN, S., SIM, D. G., NELSON, W. A. & READ, A. F. 2011. The fitness of drug-resistant malaria parasites in a rodent model: multiplicity of infection. *J Evol Biol,* 24**,** 2410-22.

JIANG, H., LI, N., GOPALAN, V., ZILVERSMIT, M. M., VARMA, S., NAGARAJAN, V., LI, J., MU, J., HAYTON, K., HENSCHEN, B., YI, M., STEPHENS, R., MCVEAN, G., AWADALLA, P., WELLEMS, T. E. & SU, X. Z. 2011. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. *Genome Biol,* 12**,** R33.

LEUNG, M. L., WANG, Y., WATERS, J. & NAVIN, N. E. 2015. SNES: single nucleus exome sequencing. *Genome Biol,* 16**,** 55.

LI, H., GLUSMAN, G., HU, H., SHANKARACHARYA, CABALLERO, J., HUBLEY, R., WITHERSPOON, D., GUTHERY, S. L., MAULDIN, D. E., JORDE, L. B., HOOD, L., ROACH, J. C. & HUFF, C. D. 2014. Relationship estimation from whole-genome sequence data. *PLoS Genet,* 10**,** e1004144.

LI, H. J. A. P. A. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

LI, R., QU, H., CHEN, J., WANG, S., CHATER, J. M., ZHANG, L., WEI, J., ZHANG, Y.-M., XU, C., ZHONG, W.-D., ZHU, J., LU, J., MA, R., FERRANTE, S. P., ROOSE, M. L. & JIA, Z. 2018. Inference of Chromosome-length Haplotypes using Genomic Data of Three to Five Single Gametes. *bioRxiv*.

LI, X., LI, L. & YAN, J. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat Commun,* 6**,** 6648.

LU, S., ZONG, C., FAN, W., YANG, M., LI, J., CHAPMAN, A. R., ZHU, P., HU, X., XU, L., YAN, L., BAI, F., QIAO, J., TANG, F., LI, R. & XIE, X. S. 2012. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science,* 338**,** 1627-30.

MANCERA, E., BOURGON, R., HUBER, W. & STEINMETZ, L. M. 2011. Genome-wide survey of post-meiotic segregation during yeast recombination. *Genome Biol,* 12**,** R36.

MANSKE, M., MIOTTO, O., CAMPINO, S., AUBURN, S., ALMAGRO-GARCIA, J., MASLEN, G., O'BRIEN, J., DJIMDE, A., DOUMBO, O., ZONGO, I., OUEDRAOGO, J. B., MICHON, P., MUELLER, I., SIBA, P., NZILA, A., BORRMANN, S., KIARA, S. M., MARSH, K., JIANG, H., SU, X. Z., AMARATUNGA, C., FAIRHURST, R., SOCHEAT, D., NOSTEN, F., IMWONG, M., WHITE, N. J., SANDERS, M., ANASTASI, E., ALCOCK, D., DRURY, E., OYOLA, S., QUAIL, M. A., TURNER, D. J., RUANO-RUBIO, V., JYOTHI, D., AMENGA-ETEGO, L., HUBBART, C., JEFFREYS, A., ROWLANDS, K., SUTHERLAND, C., ROPER, C., MANGANO, V., MODIANO, D., TAN, J. C., FERDIG, M. T., AMAMBUA-NGWA, A., CONWAY, D. J., TAKALA-HARRISON, S., PLOWE, C. V., RAYNER, J. C., ROCKETT, K. A., CLARK, T. G., NEWBOLD, C. I., BERRIMAN, M., MACINNIS, B. & KWIATKOWSKI, D. P. 2012. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature,* 487**,** 375-9.

MILES, A., IQBAL, Z., VAUTERIN, P., PEARSON, R., CAMPINO, S., THERON, M., GOULD, K., MEAD, D., DRURY, E., O'BRIEN, J., RUANO RUBIO, V., MACINNIS, B., MWANGI, J., SAMARAKOON, U., RANFORD-CARTWRIGHT, L., FERDIG, M., HAYTON, K., SU, X. Z., WELLEMS, T., RAYNER, J., MCVEAN, G. & KWIATKOWSKI, D. 2016. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Res,* 26**,** 1288-99.

MU, J., AWADALLA, P., DUAN, J., MCGEE, K. M., JOY, D. A., MCVEAN, G. A. & SU, X. Z. 2005. Recombination hotspots and population structure in Plasmodium falciparum. *PLoS Biol,* 3**,** e335.

MU, J., AWADALLA, P., DUAN, J., MCGEE, K. M., KEEBLER, J., SEYDEL, K., MCVEAN, G. A. & SU, X. Z. 2007. Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nat Genet,* 39**,** 126-30.

MZILAHOWA, T., HASTINGS, I. M., MOLYNEUX, M. E. & MCCALL, P. J. 2012. Entomological indices of malaria transmission in Chikhwawa district, Southern Malawi. *Malar J,* 11**,** 380.

NAIR, S., NKHOMA, S. C., SERRE, D., ZIMMERMAN, P. A., GORENA, K., DANIEL, B. J., NOSTEN, F., ANDERSON, T. J. & CHEESEMAN, I. H. 2014. Single-cell genomics for dissection of complex malaria infections. *Genome Res,* 24**,** 1028-38.

NEAFSEY, D. E., SCHAFFNER, S. F., VOLKMAN, S. K., PARK, D., MONTGOMERY, P., MILNER, D. A., JR., LUKENS, A., ROSEN, D., DANIELS, R., HOUDE, N., CORTESE, J. F., TYNDALL, E., GATES, C., STANGE-THOMANN, N., SARR, O., NDIAYE, D., NDIR, O., MBOUP, S., FERREIRA, M. U., MORAES SDO, L., DASH, A. P., CHITNIS, C. E., WIEGAND, R. C., HARTL, D. L., BIRREN, B. W., LANDER, E. S., SABETI, P. C. & WIRTH, D. F. 2008. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. *Genome Biol,* 9**,** R171.

NKHOMA, S. C., NAIR, S., CHEESEMAN, I. H., ROHR-ALLEGRINI, C., SINGLAM, S., NOSTEN, F. & ANDERSON, T. J. 2012. Close kinship within multiple-genotype malaria parasite infections. *Proc Biol Sci,* 279**,** 2589-98.

O'BRIEN, J. D., IQBAL, Z., WENDLER, J. & AMENGA-ETEGO, L. 2016. Inferring Strain Mixture within Clinical Plasmodium falciparum Isolates from Genomic Sequence Data. *PLoS Comput Biol,* 12**,** e1004824.

PEARSON, R. D., AMATO, R., AUBURN, S., MIOTTO, O., ALMAGRO-GARCIA, J., AMARATUNGA, C., SUON, S., MAO, S., NOVIYANTI, R., TRIMARSANTO, H., MARFURT, J., ANSTEY, N. M., WILLIAM, T., BONI, M. F., DOLECEK, C., HIEN, T. T., WHITE, N. J., MICHON, P., SIBA, P., TAVUL, L., HARRISON, G., BARRY, A., MUELLER, I., FERREIRA, M. U., KARUNAWEERA, N., RANDRIANARIVELOJOSIA, M., GAO, Q., HUBBART, C., HART, L., JEFFERY, B., DRURY, E., MEAD, D., KEKRE, M., CAMPINO, S., MANSKE, M., CORNELIUS, V. J., MACINNIS, B., ROCKETT, K. A., MILES, A., RAYNER, J. C., FAIRHURST, R. M., NOSTEN, F., PRICE, R. N. & KWIATKOWSKI, D. P. 2016. Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat Genet,* 48**,** 959-964.

PORTUGAL, S., CARRET, C., RECKER, M., ARMITAGE, A. E., GONCALVES, L. A., EPIPHANIO, S., SULLIVAN, D., ROY, C., NEWBOLD, C. I., DRAKESMITH, H. & MOTA, M. M. 2011. Host-mediated regulation of superinfection in malaria. *Nat Med,* 17**,** 732-7.

PRUGNOLLE, F., DURAND, P., JACOB, K., RAZAKANDRAINIBE, F., ARNATHAU, C., VILLARREAL, D., ROUSSET, F., DE MEEUS, T. & RENAUD, F. 2008. A comparison of Anopheles gambiae and Plasmodium falciparum genetic structure over space and time. *Microbes Infect,* 10**,** 269-75.

REECE, S. E., DREW, D. R. & GARDNER, A. 2008. Sex ratio adjustment and kin discrimination in malaria parasites. *Nature,* 453**,** 609-14.

REILLY, H. B., WANG, H., STEUTER, J. A., MARX, A. M. & FERDIG, M. T. 2007. Quantitative dissection of clone-specific growth rates in cultured malaria parasites. *Int J Parasitol,* 37**,** 1599-607.

ROSARIO, V. 1981. Cloning of naturally occurring mixed infections of malaria parasites. *Science,* 212**,** 1037-8.

SCHAFFNER, S. F., TAYLOR, A. R., WONG, W., WIRTH, D. F. & NEAFSEY, D. E. 2018. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malar J,* 17**,** 196.

TEAM, R. C. 2017. R: A Language and Environment for Statistical Computing.

TREVINO, S. G., NKHOMA, S. C., NAIR, S., DANIEL, B. J., MONCADA, K., KHOSWE, S., BANDA, R. L., NOSTEN, F. & CHEESEMAN, I. H. 2017. High-Resolution Single-Cell Sequencing of Malaria Parasites. *Genome Biol Evol,* 9**,** 3373-3383.

VOLKMAN, S. K., NEAFSEY, D. E., SCHAFFNER, S. F., PARK, D. J. & WIRTH, D. F. 2012. Harnessing genomics and genome biology to understand malaria biology. *Nat Rev Genet,* 13**,** 315-28.

WARGO, A. R., DE ROODE, J. C., HUIJBEN, S., DREW, D. R. & READ, A. F. 2007a. Transmission stage investment of malaria parasites in response to in-host competition. *Proc Biol Sci,* 274**,** 2629-38.

WARGO, A. R., HUIJBEN, S., DE ROODE, J. C., SHEPHERD, J. & READ, A. F. 2007b. Competitive release and facilitation of drug-resistant parasites after therapeutic chemotherapy in a rodent malaria model. *Proc Natl Acad Sci U S A,* 104**,** 19914-9.

WHO 2016. *World malaria report 2015*, World Health Organization.

WONG, W., GRIGGS, A. D., DANIELS, R. F., SCHAFFNER, S. F., NDIAYE, D., BEI, A. K., DEME, A. B., MACINNIS, B., VOLKMAN, S. K., HARTL, D. L., NEAFSEY, D. E. & WIRTH, D. F. 2017. Genetic relatedness analysis reveals the cotransmission of genetically related Plasmodium falciparum parasites in Thies, Senegal. *Genome Med,* 9**,** 5.

WONG, W., WENGER, E. A., HARTL, D. L. & WIRTH, D. F. 2018. Modeling the genetic relatedness of Plasmodium falciparum parasites following meiotic recombination and cotransmission. *PLoS Comput Biol,* 14**,** e1005923.

ZHENG, X., GOGARTEN, S. M., LAWRENCE, M., STILP, A., CONOMOS, M. P., WEIR, B. S., LAURIE, C. & LEVINE, D. 2017. SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics,* 33**,** 2251-2257.

ZHU, S. J., ALMAGRO-GARCIA, J. & MCVEAN, G. 2018. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. *Bioinformatics,* 34**,** 9-15.

Figure 1

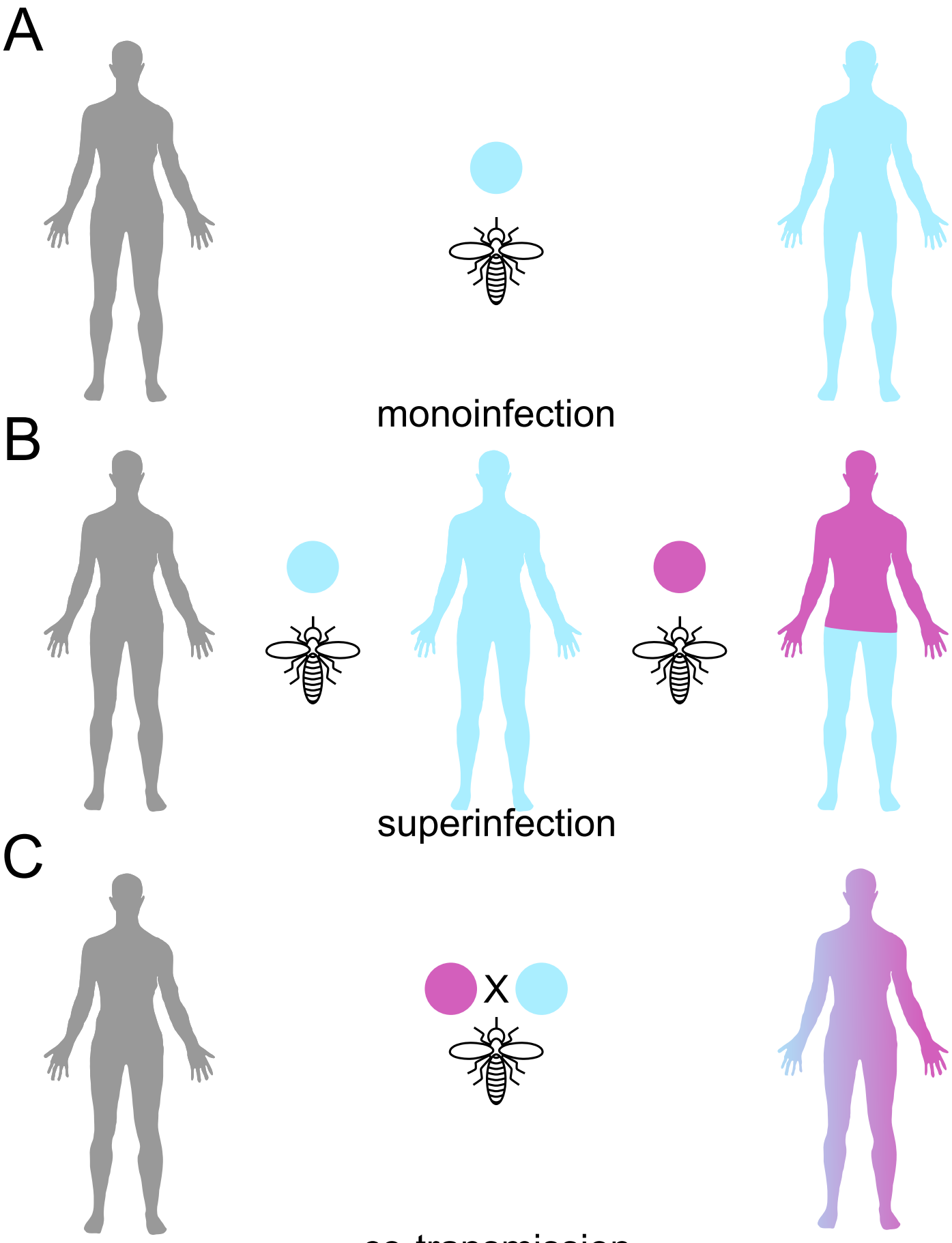


Figure 2

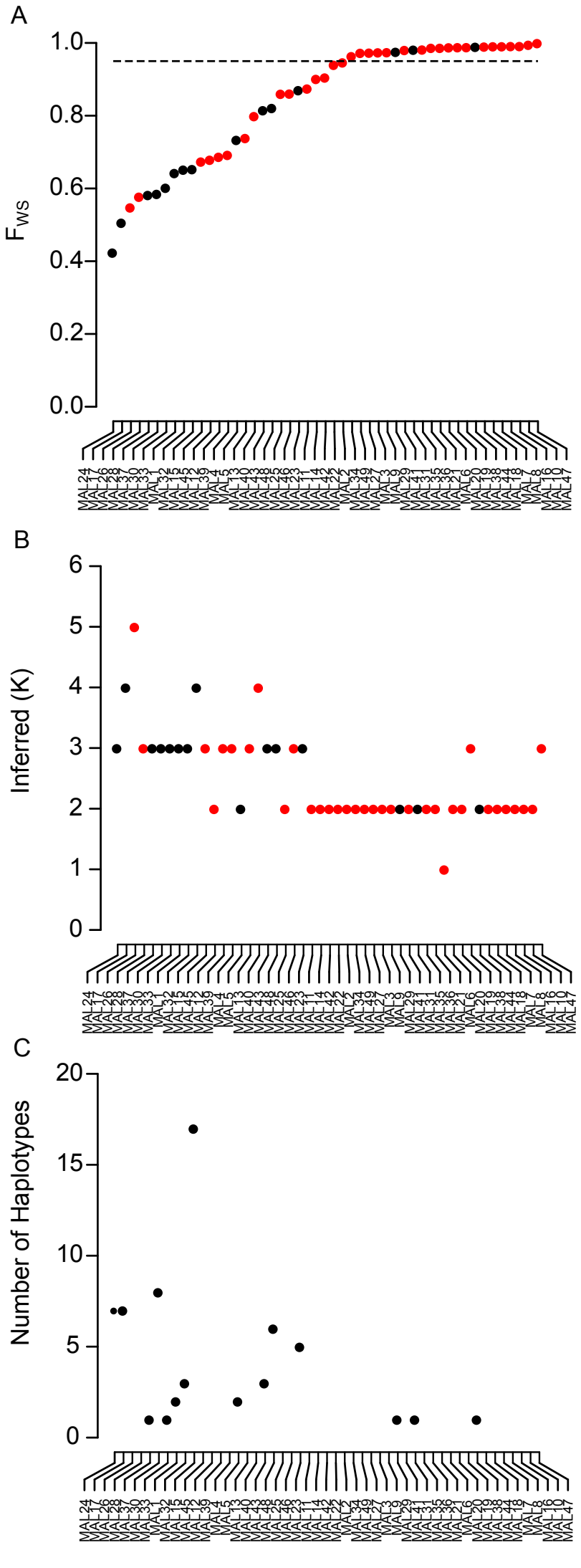

A
1.0
0.8
0.6
0.4
0.2
0.0
F_{ws}
B
6
5
4
3
2
1
0
Inferred (K)
C
20
15
10
5
0
Number of Haplotypes
MAL24 MAL17 MAL26 MAL28 MAL37 MAL30 MAL33 MAL1 MAL32 MAL15 MAL45 MAL12 MAL39 MAL4 MAL5 MAL13 MAL40 MAL43 MAL48 MAL25 MAL46 MAL23 MAL11 MAL14 MAL42 MAL22 MAL34 MAL49 MAL27 MAL3 MAL9 MAL29 MAL41 MAL31 MAL35 MAL36 MAL21 MAL6 MAL20 MAL19 MAL38 MAL44 MAL18 MAL7 MAL8 MAL16 MAL10 MAL47

Figure 3

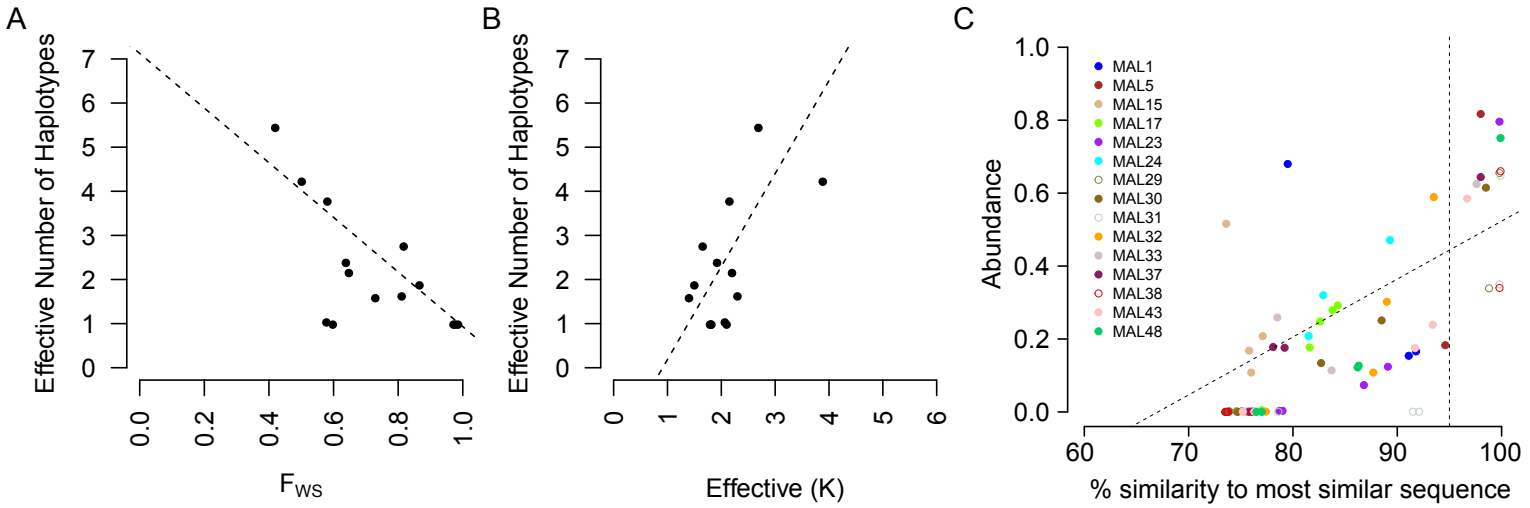

A
Effective Number of Haplotypes
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8
1.0
F_{WS}
B
Effective Number of Haplotypes
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
Effective (K)
C
Abundance
0.0
0.2
0.4
0.6
0.8
1.0
60
70
80
90
100
% similarity to most similar sequence
MAL1
MAL5
MAL15
MAL17
MAL23
MAL24
MAL29
MAL30
MAL31
MAL32
MAL33
MAL37
MAL38
MAL43
MAL48

Figure 4

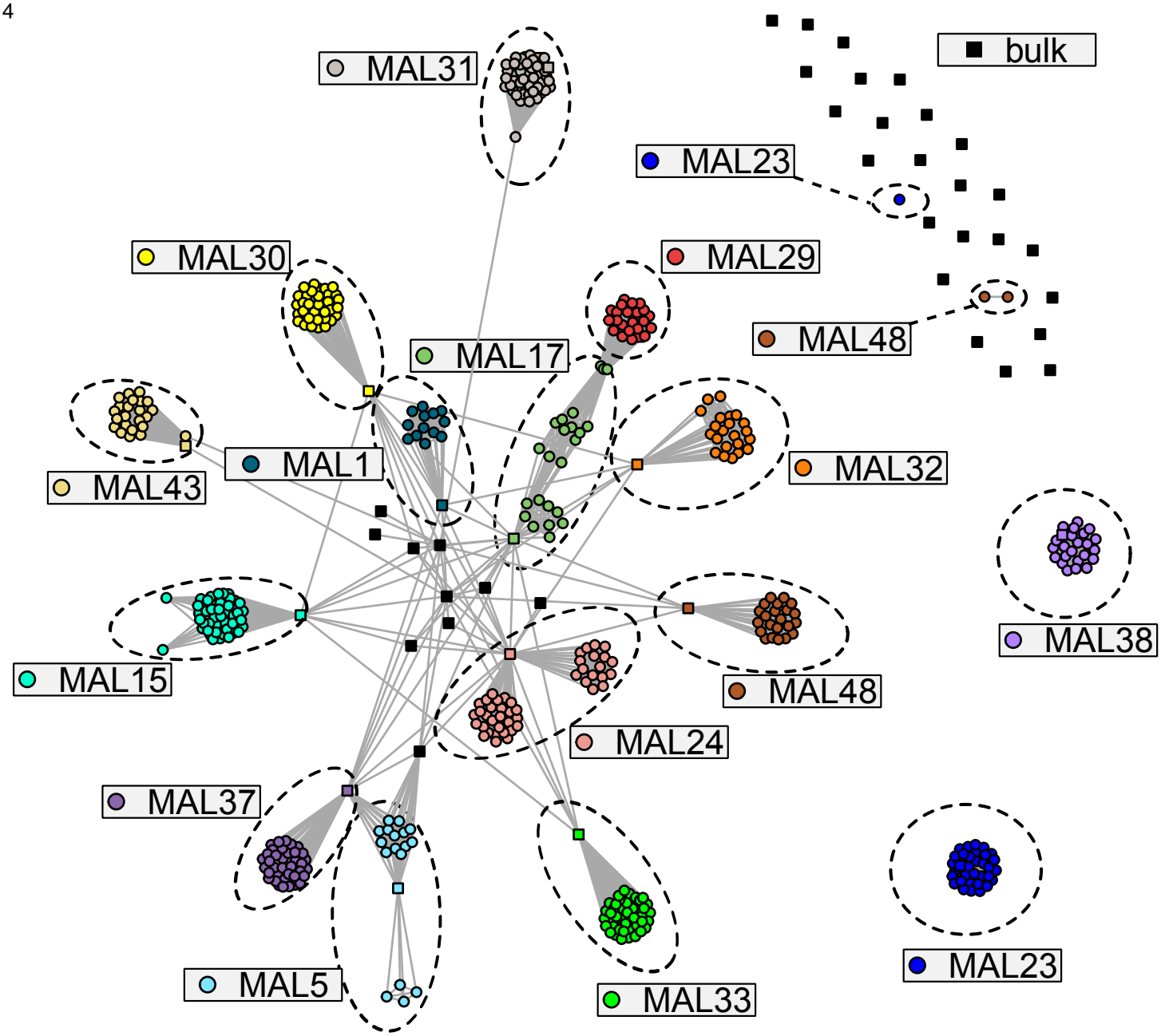

MAL31
bulk
MAL23
MAL30
MAL29
MAL48
MAL17
MAL43
MAL1
MAL32
MAL38
MAL15
MAL48
MAL24
MAL37
MAL23
MAL5
MAL33

Figure 5

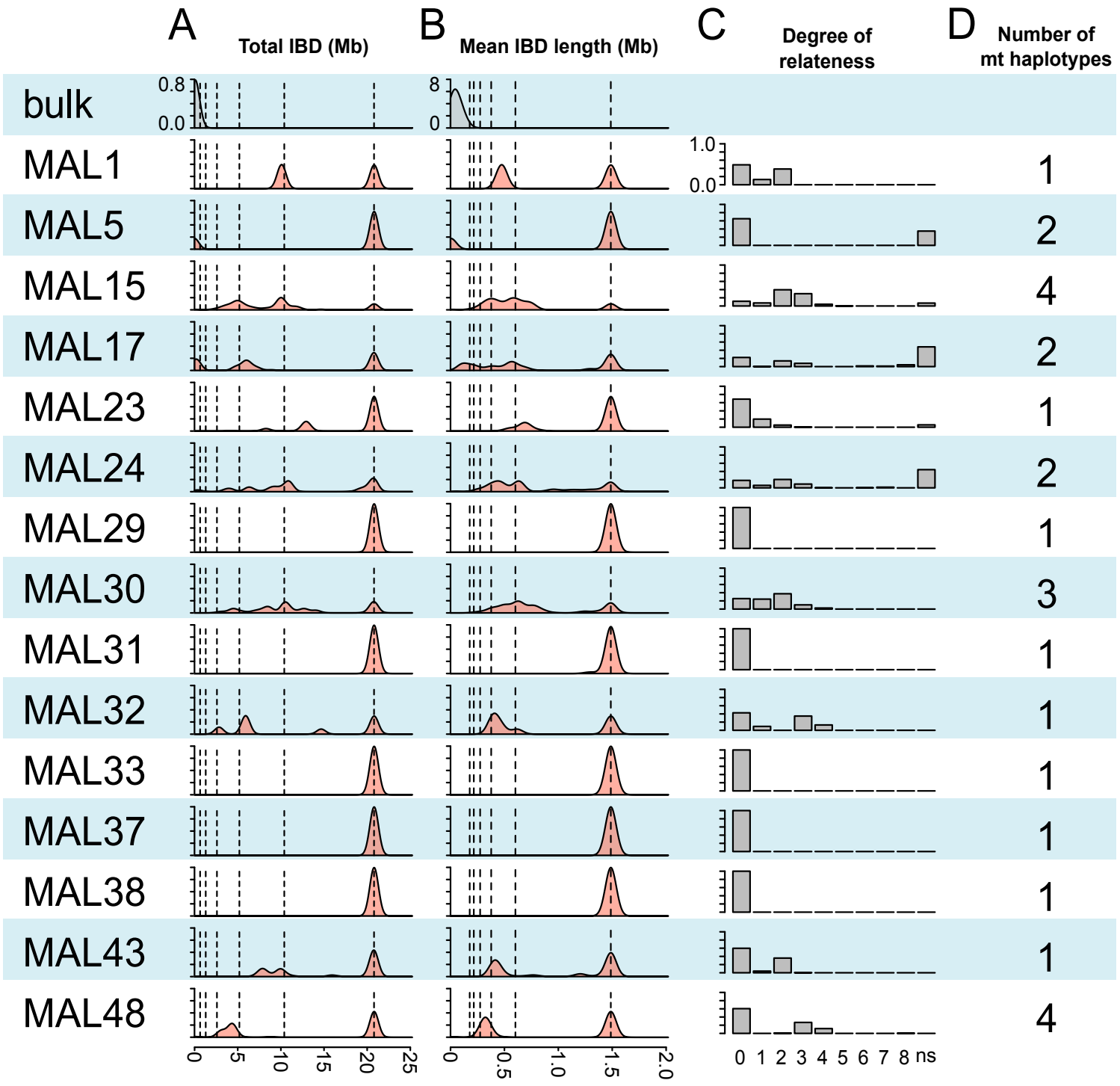
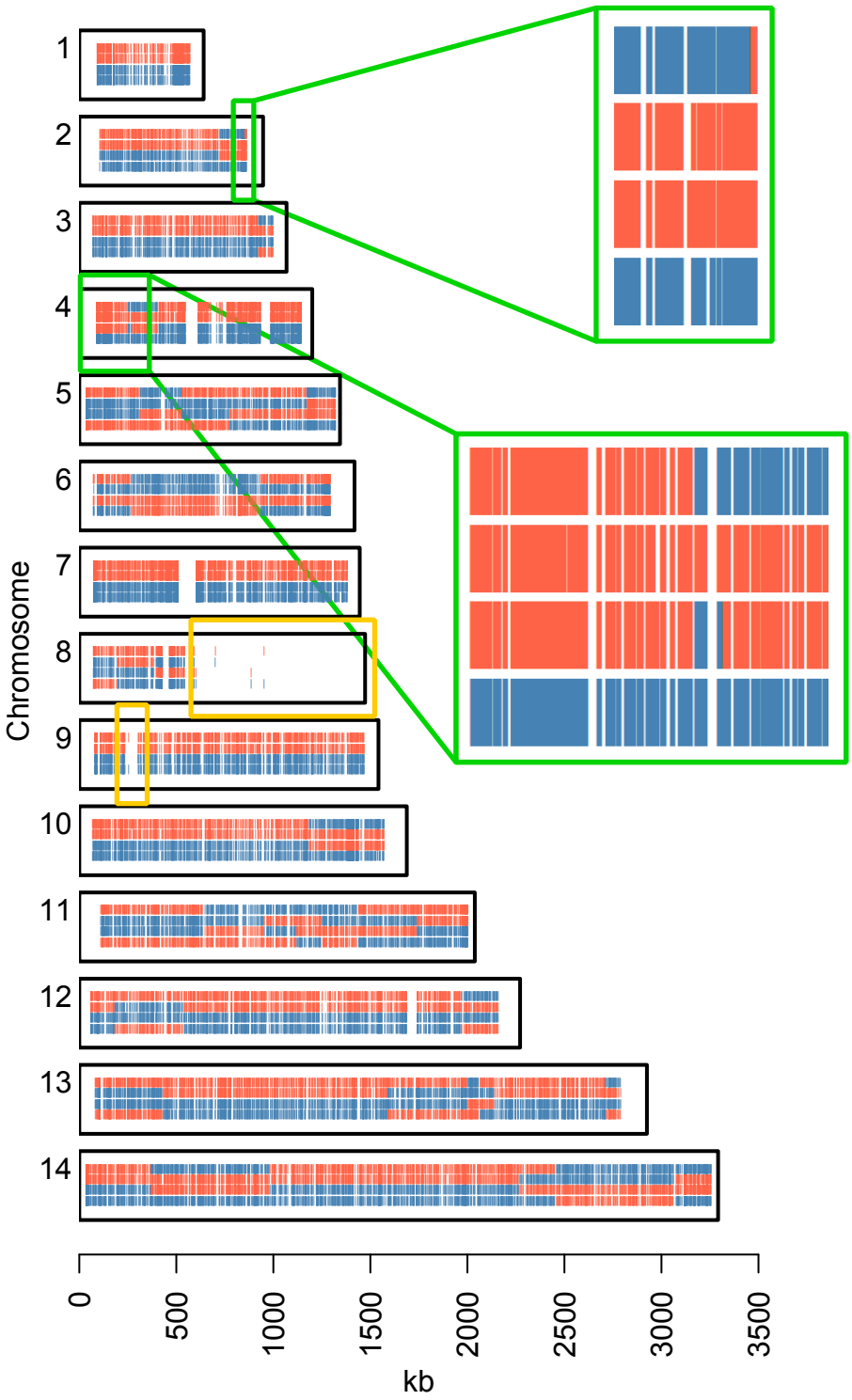

A
Total IBD (Mb)
B
Mean IBD length (Mb)
C
Degree of relateness
D
Number of mt haplotypes
bulk
MAL1
MAL5
MAL15
MAL17
MAL23
MAL24
MAL29
MAL30
MAL31
MAL32
MAL33
MAL37
MAL38
MAL43
MAL48
0.8
0.0
8
0
1.0
0.0
1
2
4
2
1
2
1
3
1
1
1
1
1
4
0 5 10 15 20 25
0 0.5 1.0 1.5 2.0
0 1 2 3 4 5 6 7 8 ns

Figure 6


Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
0
500
1000
1500
2000
2500
3000
3500
kb

Supplemental Table 1

Click here to access/download

Supplemental Videos and Spreadsheets

Supplementary Table 1.xlsx

Click here to access/download

Supplemental Videos and Spreadsheets

Supplementary Data 1.pptx

Click here to access/download

Supplemental Videos and Spreadsheets

Supplementary Data 2.pptx

Click here to access/download

Supplemental Videos and Spreadsheets

Supplementary Data 3.pptx