Detection of simple and complex *de novo* mutations without, with, or with multiple reference sequences

Kiran V Garimella[1,2,3]*, Zamin Iqbal[2,4], Michael A. Krause[2,5,6], Susana Campino[5], Mihir Kekre[5], Eleanor Drury[5], Dominic Kwiatkowski[3,5], Juliana M. Sa[6], Thomas E. Wellems[6], and Gil McVean[2,3]

[1]Data Sciences Platform, Broad Institute of MIT and Harvard
[2]Wellcome Trust Centre for Human Genetics, University of Oxford
[3]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford
[4]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus
[5]The Wellcome Trust Sanger Institute, Wellcome Genome Campus
[6]Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Abstract

The characterization of *de novo* mutations in regions of high sequence and structural diversity from whole genome sequencing data remains highly challenging. Complex structural variants tend to arise in regions of high repetitiveness and low complexity, challenging both *de novo* assembly, where short-reads do not capture the long-range context required for resolution, and mapping approaches, where improper alignment of reads to a reference genome that is highly diverged from that of the sample can lead to false or partial calls. Long-read technologies can potentially solve such problems but are currently unfeasible to use at scale. Here we present Corticall, a graph-based method that combines the advantages of multiple technologies and prior data sources to detect arbitrary classes of genetic variant. We construct multi-sample, coloured de Bruijn graphs from short-read data for all samples, align long-read-derived haplotypes and multiple reference data sources to restore graph connectivity information, and call variants using graph path-finding algorithms and a model for simultaneous alignment and recombination. We validate and evaluate the approach using extensive simulations and use it to characterize the rate and spectrum of *de novo* mutation events in 119 progeny from four *Plasmodium falciparum* experimental crosses, using long-read data on the parents to inform reconstructions of the progeny and to detect several known and novel non-allelic homologous recombination events.

1 Introduction

High genomic diversity within a population can confound variant and particularly *de novo* mutation (DNM) discovery efforts. As a single reference genome cannot capture the range of possible haplotypes, short read aligners assume that new haplotypes are small perturbations to a known canonical reference sequence. Divergent or absent loci violate this assumption, hence reads

---

*Correspondence and request for materials should be addressed to K.V.G. (email: kiran@broadinstitute.org).

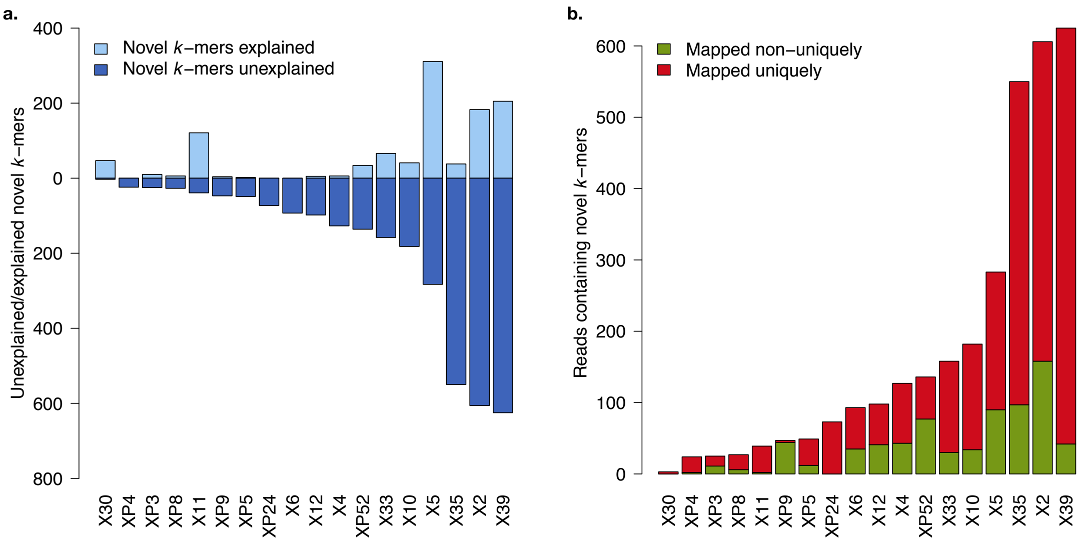


Figure 1: Extent of reference-characterized and uncharacterised novelty among 18 progeny from an experimental cross between 3D7 and HB3 *P. falciparum* isolates (76-bp reads, ~100*x* coverage). (**a**) Novel *k*-mers observed in the reference-based analysis ("explained", bars above zero-line) versus novel *k*-mers remaining from the reference-free analysis ("unexplained", bars below zero-line). (**b**) Reads that map uniquely to the reference genome ($MQ > 0$) versus mapping multiple times or not mapping at all ($MQ = 0$), conditioned on the read containing a novel *k*-mer. See Supplemental Material for details.

sampled from them may align incorrectly or not at all(19). This results in many false positives and false negatives in such regions, the combination of which can be sometimes be erroneously interpreted as complex forms of variation. Maps of "genome accessibility" can restrict variant calling to less diverse regions of the genome and reduce such errors(41; 25; 47; 30), but may lead to substantial undiscovered variation. For example, we examined 20 high-coverage *P. falciparum* pedigree samples where one parent is expected to be substantially divergent from the reference. We compared novel sequences present in short-read *de novo* assemblies of progeny versus haplotypes combinatorically produced from multiple reference-based callsets on the same data. Even with strong filtering on the *de novo* assembly data and no filtering on the reference-based callset data, 28% ± 22% (min=0%, max=94%) of novel sequences in the assemblies did not correspond to a reference-based variant call. Most of these unexplained novel "*k*-mers" (length *k* substrings from reads) were found in reads that failed to map or mapped non-uniquely to the reference sequence. This comparison is shown in Figure 1 (see Supplemental Material section S1 for further analysis details).

Of particular concern are *de novo* structural variants (SVs) driven by mutational mechanisms mediated by microhomology and repeat structure(3). Many SVs are predisposed to occur within repetitive loci around the genome. For example, non-allelic homologous recombination (NAHR) can occur between two low copy-number repeats (LCRs), repetitive sequences ranging from several to hundreds of kilobases in length and having > 95% sequence identity between them(23). Non-allelic copies will occasionally be aligned in meiosis and mitosis, with subsequent crossover employing them as the substrate for homologous recombination. Resolution of the misaligned sequences can yield successive insertions, deletions, duplications, inversions, and translocations(29). NAHR in humans has been associated with several genomic disorders (e.g. Charcot-Marie-Tooth 1A, hereditary neuropathy with liability to pressure palsies(24)), and cancer (e.g. hereditary breast/ovarian cancer(46)).

While short read *de novo* assembly may provide a means for overcoming reference bias, the repetitive nature of the SVs precludes the straightforward application of existing tools(1; 38). A typical assembly graph stores genomic subsequences k-mers as vertices and sequence overlaps (read-to-read alignments or $k-1$ substring matches) as edges(9). Repeats longer than the vertex length collapse into a single copy. Differing sequence contexts manifest as multiple edges, which is problematic for assembly as extracting unambiguous contiguous sequence from a graph requires runs of vertices with in-degree and out-degree of 1 ("unitigs").

For small sample sizes, *de novo* assembly using long read data from third generation sequencing is a viable strategy for overcoming reference bias and assembling through highly repetitive loci(31; 16). However, the high molecular weight gDNA input requirement relative to second-generation sequencing ($\sim 10,000$ ng versus ~ 1 ng) is difficult to satisfy with some samples. Many pathogens grow slowly in culture, requiring several months or even years to expand to sufficient amounts for long-read sequencing. Stromal contamination and high heterogeneity in cancer samples compromises the ability to acquire pure samples of such high mass, and amplification risks PCR replication artifacts masquerading as true *de novo* mutations.

Instead, it may be possible to sequence a small number of samples with long reads to augment a larger, short read data set. For a typical assembly from short read data (e.g. 76 bp reads, $> 20x$ coverage), sequencing is expected to recover nearly* every k-mer in the genome(20), even if the reads do not provide sufficient genomic context to navigate through repetitive regions. That context can be provided by aligning long haplotypes to the short read graph, annotating edge choices, and following these choices when traversing the graph(40). Importantly, these long haplotypes need not be from the sample itself; recent common ancestry among samples leads to extensive sharing of variation that can be used to guide assembly in related samples. By demanding that the short read genome graph is immutable (after initial construction and removal or correction of likely sequencing errors), the process of long haplotype alignment cannot add any new vertices, only provide connectivity information through existing vertices. This naturally constrains the alignments to informing connectivity in regions of high (but not necessarily perfect) homology between the long read and short read samples. Finally, by aligning multiple data sets to the graph (many long-read datasets, paired-end reads from the sample itself, etc.), we can assemble through recombination breakpoints by transitioning between annotation sets. In essence, rather than using existing tools to improve accuracy of long-read assemblies with short reads(17; 42; 34; 13), we improve the connectivity of short-read assemblies with long reads.

Here, we demonstrate that a small panel of long-read samples could be used to significantly improve *de novo* mutation calling efforts in short read samples. We show that by threading long-read-derived haplotypes through short-read graphs, contig lengths spanning putative DNM sites increases dramatically — often by an order of magnitude. We further elucidate a model of simultaneous alignment and "contig painting" wherein progeny contigs are probabilistically aligned to multiple candidate parental haplotypes and recombination between any haplotype and any site is permitted. This model provides a uniform approach to typing simple (e.g. single nucleotide variants — "SNVs", small insertions/deletions — "indels", multi-nucleotide variants — "MNVs") and complex (e.g. large indels or rearrangements manifesting as multiple successive breakends — "BND") variants.

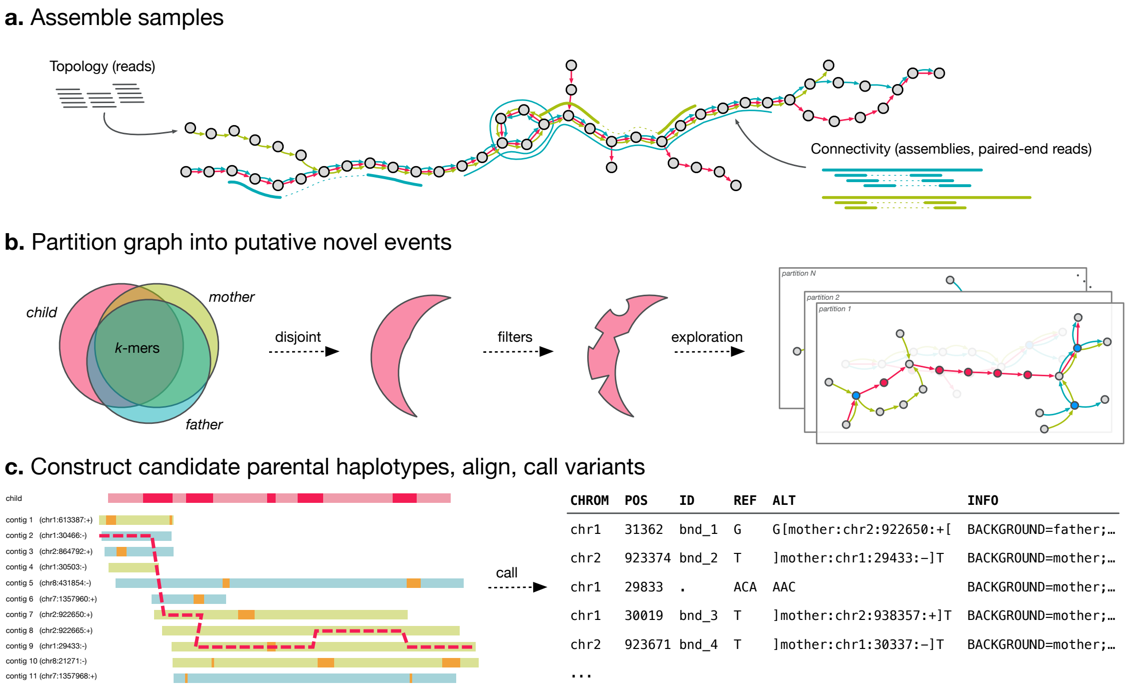
---

*barring systematic sequencing errors and ultra low-complexity sequences that fail to amplify.

After testing in simulated datasets, we apply our software to progeny from four experimental crosses of the etiological agent of malaria, *Plasmodium falciparum*. We specifically chose to examine the *P. falciparum* data as the parasite's genome is of intermediate size ($\sim$23Mb, small enough that PacBio sequencing of all parental genomes was feasible, and large enough to force several software design choices that would enable scaling to larger genomes and datasets), notoriously repetitive (AT content $\sim$ 80%), haploid (allowing phase between proximate structural variant breakends to be assumed and avoiding the complexity of diploid assembly), and known to possess pathogenically relevant NAHR activity with validated examples. We discover DNMs of all mutational classes in core genomic regions at rates consistent with previous literature, and observe statistically significant enrichment of variants in accessory regions of the genome. Furthermore, we rediscover a known NAHR event and redefine the boundaries beyond those previously established. We also discover new NAHR events, nearly all of which occur within or proximate to antigenic genes. Finally, we examine the transformation of our model from a *de novo* mutation calling framework to a general variant calling framework, showing that the iterative addition of draft reference sequence data yields stepwise improvements in variant discovery. These tools are implemented in a novel graph-based, connectivity-aware variant caller, Corticall.

2 RESULTS

Our DNM discovery approach consists of three steps. First, *de novo* assembly, based on multi-color linked de Bruijn graphs, or LdBGs, is used to store and link adjacent k-mers for each sample. These assemblies are error-cleaned - that is, low frequency k-mers likely to be the result of sequencing errors are removed from the graphs. Unlike error-correction, error-cleaning does not add new (and potentially unobserved) sequence to the graph. Second, trusted "novel" k-mers are identified; sequences unique to the individual progeny, indicative of DNMs, that are unlikely to arise from error or contamination. Finally, novel k-mer-spanning contigs are aligned to reconstructed sequences in the parents, identifying the nature of the event that generated the DNM. Figure 2 depicts these steps, detailed in the Supplemental Material and summarized below.



a. Assemble samples

b. Partition graph into putative novel events

c. Construct candidate parental haplotypes, align, call variants

Topology (reads)
Connectivity (assemblies, paired-end reads)
child
mother
k-mers
father
disjoint
filters
exploration
partition N
partition 2
partition 1
child
contig 1 (chr1:613387:+)
contig 2 (chr1:30466:-)
contig 3 (chr2:864792:+)
contig 4 (chr1:30503:-)
contig 5 (chr8:431854:-)
contig 6 (chr7:1357960:+)
contig 7 (chr2:922650:+)
contig 8 (chr2:922665:+)
contig 9 (chr1:29433:-)
contig 10 (chr8:21271:-)
contig 11 (chr7:1357968:+)
call

```
CHROM  POS    ID     REF  ALT                       INFO
chr1   31362  bnd_1  G    G[mother:chr2:922650:+[   BACKGROUND=father;…
chr2   923374 bnd_2  T    ]mother:chr1:29433:-]T    BACKGROUND=mother;…
chr1   29833  .      ACA  AAC                       BACKGROUND=mother;…
chr1   30019  bnd_3  T    ]mother:chr2:938357:+]T   BACKGROUND=mother;…
chr2   923671 bnd_4  T    ]mother:chr1:30337:-]T    BACKGROUND=mother;…
...
```

Figure 2: Overview of the Corticall algorithm. (**a**) Samples are assembled into a multi-color linked de Bruijn graph (LdBG). Short, accurate reads are used to determine graph topology. Longer sequences derived from paired-end reads or from draft/finished assemblies are thread through the graph, providing information on connectivity to overcome repeats, but not adding novel *k*-mers. (**b**) Novel *k*-mers - sequences present in the progeny and absent in the parents - are filtered and then used to signal the presence of putative *de novo* mutations. Subgraphs around such events are extracted, forming a set of variant candidates. (**c**) Regions flanking novel *k*-mers are assembled to reveal candidate parental haplotypes. The progeny's contig is probabilistically aligned to the set of candidate parental contigs, allowing for mismatches, indels, and (potentially non-allelic) recombination. The resulting alignment thus specifies parental background and (if reference sequences are available) coordinate information. Variants (SNVs/MNVs/indels/translocation breakends/etc.) within the novel *k*-mer regions are returned as likely *de novo* mutations.

2.1 Connectivity preserved in multi-color linked de Bruijn graphs (LdBGs)

We have previously reported on multi-sample and multi-color de Bruijn graphs (dBG) for straightforward reference-free genome comparison between multiple samples(15) and linked de Bruijn graphs (LdBG) for improved assembly via read-to-graph and reference-to-graph alignment annotations(40). Briefly, an LdBG is a multigraph(49) representation of multiple genomes that preserves "stackability" (easy comparison of multiple samples via inner joins by k-mer of per-sample coverage and edge information) and connectivity information inherent in reads and/or long input haplotypes. As illustrated in Figure 2a, input reads are decomposed into k-mers and stored as graph vertices. Each sample is assigned a unique identifier (or "color"). Colored edges are placed between vertices representing $k-1$ overlaps with another k-mer in the same sample. Reads and/or haplotype data (e.g. alternate reference assemblies) are then aligned to the graph (once per color) by trivial lookups of shared k-mers. Discrepancies between the sequence and the graph manifest as missing k-mers, correctable by traversing the graph between the gap boundaries, or truncating the alignment if the correction attempt fails. At junctions (vertices with in-degree or out-degree greater than 1), the edge consistent with the aligned sequence is recorded in an auxiliary file. All junctions spanned by an alignment are annotated with relevant link information, ensuring traversal can begin anywhere in the graph and still have access to complete navigation data. During traversal, we collect links in the order they are encountered, assigning each link an "age" reflecting the number of vertices traversed since collected, and using the oldest link to specify junction choices. If a conflict arises between multiple oldest links, we halt traversal.

2.2 Novel k-mers are signposts for DNMs

We build upon this genome comparison framework by first identifying regions of the joint pedigree graph (an LdBG containing sequence data for parents, progeny, and optional reference sequences) to explore for potential DNMs. As such mutations are by definition present in the progeny and absent in the parents, k-mers spanning these events would also be expected to be exclusive to the progeny.

An accurate list of novel k-mers serves both as an indicator of DNM presence around the graph and a measure of how many mutational events are available for discovery. However, iteration over the graph and selection of putative novel k-mers (those with 0 coverage in the parents and > 0 coverage in the progeny) will yield a set enriched for sequencing errors and other artifacts that obscure the small fraction of k-mers arising from genuine DNMs. We apply multiple filters to remove such artifacts (specifically, contamination; graph tips; low-complexity sequence; "orphans", sequence found in the progeny but with no edges to parental sequences; low-coverage k-mers; "unanchored" k-mers, k-mers in branches that have no unique alignment in any provided genome; and k-mers shared by other progeny, see Supplemental Material section S2.4 for details). We verified these filters by examining novelty in simulated *P. falciparum* crosses and a real trio for which we obtained PacBio sequencing on both parents and progeny.

2.3 Contigs spanning novel *k*-mers contain putative *de novo* events

Next, we "partition" the graph into subgraphs, grouping novel *k*-mers into separate bins based on their proximity to one another within the graph. This is illustrated in Figure 2b. Each partition may harbor one or more DNMs, but DNMs are not split across multiple partitions. At each novel *k*-mer, we walk along the progeny's color in the pedigree graph, exploring outwards and constructing the longest possible contigs. To maximize contig length (and thus increase our sensitivity to complex variation), we employ two strategies. First, links derived from haplotype alignments (e.g. draft references, paired-end reads, etc.) are used to disambiguate junction choices. Second, as DNMs will typically yield a succession of novel *k*-mers in a graph, and as the previous filtering step will have removed most artifacts, we walk past junctions when one (and only one) of the outgoing edges at a novel *k*-mer connects to another novel *k*-mer. This procedure, which we have termed "novel *k*-mer aggregation", ensures that proximate novel *k*-mers are considered together, useful for large structural variants that may manifest as a series of nearby, but non-adjacent, runs of novel *k*-mers.

2.4 Assembling adjacent parental contigs for event decoding

We then construct parental sequences that constitute the candidate haplotypic background(s) for a DNM. At each parentally-shared *k*-mer in a partition, we initiate a contig assembly in the parents. The presence of novel *k*-mers in the partition may lead to gaps in the parental contigs not automatically filled by this assembly step. We close these gaps via depth-first searches (DFS) between bordering *k*-mers. To prevent a combinatoric explosion of considered paths, we limit our explorations to depths of $1,000$ bp by default. For gaps we fail to close in this manner, we assemble flanking boundaries up to a maximum of 500 bp.

Each contig is given a label specifying the parental background from which it was reconstructed and a unique index. If draft/finished reference sequence data is available, we additionally attach coordinate information by aligning each parental contig to the associated draft reference sequence via a built-in version of bwa mem(21)[†].

2.5 "Mosaic" alignment reveals simple and complex mutations

To identify mutations, determine parental background and assign genomic coordinates, we apply a pair-HMM to simultaneously align and phase progeny contigs over candidate parental haplotypes. This model, originally used to study evolutionary relationships in a set of highly diverse antigenic genes from the *P. falciparum var* gene family(48), combines the probabilistic models for sequence alignment(8) and the detection of recombination events(22). Recast in a structural variant framework, it enables simultaneous discovery of both simple/complex mutations in a panel of sequences that are not pre-aligned to one another. As our model permits recombination between any site and any candidate parental haplotype, it also enables the detection of non-allelic events, such as NAHR.

[†] bwa mem Java bindings developed by Pierre Lindenbaum, https://github.com/lindenb/jbwa

Briefly, the method is as follows. Consider a query sequence (the contig in the progeny) and a set of N source sequences (contigs in both parents, partially or completely spanning the target sequence). Our goal is to describe the target sequence as a set of match/mismatch, insertion, deletion, and recombination operations on the source sequences. We choose the starting point in the source sequence uniformly across all sites in the source sequences, beginning in the match or insert states with some probability. At each position, there exists the probability of jumping to any target sequence and any position via recombination. The maximum likelihood alignment (and trajectory through the target panel sequence space) is obtained using the Viterbi algorithm. Variant calls are obtained by examining the traceback path and identifying differences with respect to the query sequence. This process is depicted in Figure 2c. Further mathematical specification is available in the Supplemental Material section S2.7.

A simple set of post-processing filters are applied to keep false discovery rates low. For all mutational types, we reject events containing less than 5 novel k-mers. We additionally require NAHR events to satisfy one of two conditions: (1) multiple breakends are detected within a single contig; (2) single breakends are detected within $2,000$ bp of breakends satisfying (1).

2.6 Simulation: novel *k*-mer detection and increased contig lengths

To evaluate our ability to correctly detect DNMs in assembly data, we generated an *in silico* pedigree, simulating full-length (23 Mb) genomes for $1,000$ *P. falciparum* progeny and incorporating a wide range of *de novo* events for later evaluation. Annotated draft reference sequences constructed for two *P. falciparum* isolates (HB3 and DD2, see Supplemental Material section S3) were used as parental genomes. We computed k-mer-based homology maps per sister chromatid and modeled crossovers per chromosome based on empirical rates(26), keeping track of the relocated members of each parent's *var* gene repertoire. We then added simple and complex DNMs, simulating small $(1-100$ bp), intermediate $(101-500)$, and large $(501-1,000$ bp) events, and placing them randomly throughout the genome. In addition to simulating SNVs, MNVs/indels with random sequence, and inversions, special care was taken to simulate variants arising from repeat expansion and contraction by searching for existing repetitive regions in the genome and adding or subtracting repeat units. NAHR events were simulated by recombining members of the progeny's *var* gene repertoire after meiotic recombination. Assuming a low *de novo* mutation rate, 3 random events were simulated per progeny.

For each progeny's genome, we simulated 76 bp paired-end reads with an insert size distribution of 250 ± 50 bp, stochastic coverage of $100X$, and a sequencing error rate of 0.5% ($\sim Q23$). These values were comparable to existing data on the HB3xDD2 cross(44; 45; 25). We constructed joint pedigree graphs using existing Illumina data for the HB3xDD2 parents along with our simulated reads for the progeny, applying the assembly procedure detailed in S4.5 (initial assembly at $k=47$, error cleaning, and paired-end read and draft reference threading),

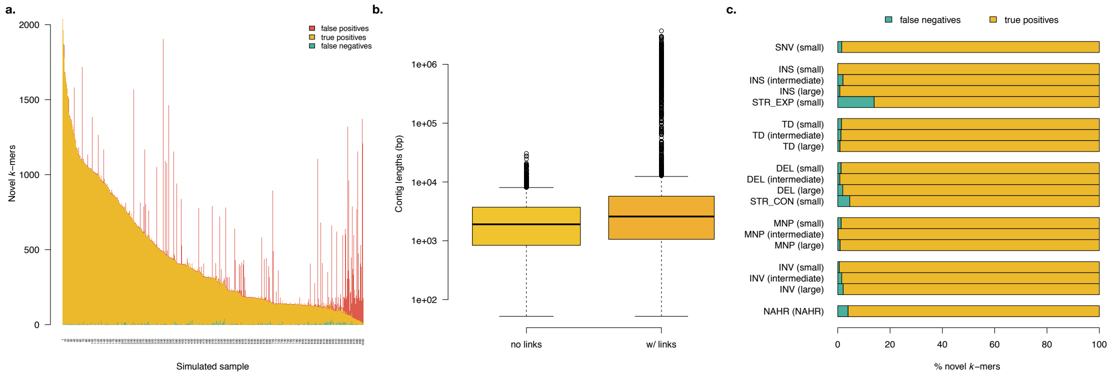


Figure 3: Simulation-based evaluation of novel *k*-mer detection and subsequent reassembly quality for contigs spanning novel *k*-mers in error-containing short read data. (**a**) Number of *k*-mers in the progeny correctly identified as novel (true positives), undetected (false negatives), and misidentified as novel (false positives). (**b**) Contig lengths resulting from reconstruction strategies without and with link information. (**c**) For all simulated alleles, the fraction assembled completely (i.e. wholly contained within a single contig) and incompletely (i.e. only partially reconstructed).

and extracted novel *k*-mers according to the procedure in S2.4.

We first evaluated our novel *k*-mer detection procedure on these simulated datasets. Figure 3a summarizes our detection of true and false novel *k*-mers. We were able to detect $90.0\% \pm 22.7\%$ of expected novel *k*-mers per sample. Novel *k*-mers that we failed to detect were typically low-complexity or repetitive sequences (generated *de novo* by the mutation process but also occurring elsewhere in the genome). For these events, a *k*-mer size of 47 bp was insufficient to resolve the sequences as novel.

Next, we studied our novel *k*-mer-spanning contig reconstructions. Contigs spanning novel *k*-mers were assembled with three different procedures; resulting lengths were calculated and aggregated across all 1,000 simulated samples and are presented in Figure 3b. Without link information, median contig N50 lengths (minimum contig length required to cover 50% of the genome) were 4,186 bp. This increased by $\sim 71\%$ to a median of 7,161 bp with the addition of links from paired-end reads and parental assembly links (Figure 3b).

Finally, we sought to understand the relationship between missed novel *k*-mers and the type of variant event from which they arose, summarized in Figure 3c. Across all variant types, $97.8\% \pm 3.1\%$ of novel *k*-mers generated by mutational events are detected. The bottom three performers are short tandem repeat (STR) contractions, STR expansions, and NAHR events, where the percentage of novel *k*-mers detected are 86.1%, 95.4%, and 96.0% respectively. This is to be expected; all three mutational classes are manipulations of repetitive sequence, the expansion/contraction/recombination of which would be plausibly expected to generate *k*-mers already present in other repeats in the genome.

Table 1: Variant classification and F_1 score in $1,000$ simulated *P. falciparum* genomes.

			Without link information		With link information	
Type	Length (bp)	N	F1	F1 (w/ phase)	F1	F1 (w/ phase)
SNVs	0	627	0.85	0.85	0.86	0.86
Insertions						
random sequence	1-100	29	0.81	0.74	0.83	0.76
	101-500	111	0.99	0.91	0.99	0.91
	501-1000	145	0.98	0.90	0.99	0.91
STR expansions	*(variable)*	323	0.91	0.83	0.92	0.85
tandem duplications	*(variable)*	287	0.69	0.64	0.70	0.65
Deletions						
random sequence	1-100	29	0.75	0.70	0.81	0.73
	101-500	125	0.84	0.80	0.84	0.81
	501-1000	444	0.53	0.52	0.54	0.52
STR contractions	*(variable)*	301	0.91	0.87	0.92	0.90
MNVs						
random sequence	1-100	32	0.85	0.79	0.93	0.82
	101-500	106	0.73	0.72	0.76	0.75
	501-1000	135	0.77	0.76	0.79	0.78
inversions	1-100	29	0.86	0.84	0.91	0.82
	101-500	116	0.88	0.88	0.89	0.88
	501-1000	137	0.88	0.88	0.89	0.88
NAHRs						
single breakpoints	-	1142	0.63	0.63	0.70	0.70
all breakpoints	-	270	0.60	0.60	0.66	0.66

[a] For NAHR events, partial (complete) reconstruction indicates recovery of any (all) of the simulated breakends.

2.7 Simulation: mutation detection and evaluation

We applied our software to the simulated set of $1,000$ HB3xDD2 progeny, measuring performance using F_1 scores for variants. The aggregated results for all simulated samples are shown in Table 1.

Overall, we found that greater than 90% of detected novel *k*-mers are assignable to variant events, and more than 97% of simulated variants are identified (either partially or completely reconstructed). This changes very little with assembly mode as, aside from some light filtering, the absence or presence of link information does not alter the detection of novel *k*-mers. Instead, it simply alters the number of contigs into which a variant assembles. For complete reconstruction of each variant event, F_1 uniformly increases between the link-uninformed and link-informed reconstruction as link information provides a means to overcome repetitive regions of the assembly. This is particularly valuable for DNMs in repetitive elements: STR expansions, STR contractions, and tandem duplications. For these variant classes, link information conferred a $\sim 10\%$, $\sim 10\%$, and $\sim 14\%$ increase in F_1 respectively.

We measured calling performance on breakends, and further our ability to group proximate breakends into single NAHR events. While both reconstructions are generally able to detect the presence of a breakend, the LdBG reconstructions show marked improvement in event characterization. This permit multiple breakends to be observed on a single contig, enabling detection and assignment of all breakends within a simulated NAHR event to a single call.

2.8 Core, accessory, and repetitive region DNM detection in *P. falciparum*

To characterize the number and type of DNMs occurring in the genome of the malaria parasite *P. falciparum*, we applied our software to data from four *P. falciparum* experimental crosses: 3D7xHB3(43), HB3xDD2(44), 7G8xGB4(14), and 803xGB4(33). In addition to the 3D7 canonical reference genome(12), we generated high-quality draft reference assemblies for the remaining five parental genomes (HB3, DD2, 7G8, GB4, and 803) using PacBio RSII sequencing (Supplemental Table 3). We obtained Illumina data for all samples in the experimental crosses (Supplemental Table 4) and generated McCortex assemblies at $k = 47$. After contaminant and outlier removal, we called DNMs in 119 progeny. These calls are summarized in Figure 4.

Across samples, we assigned putative variant calls to $(89 \pm 11)\%$ of novel k-mers (Figure 4a). Their impact was greatest in the 803xGB4 cross, where the 803 and GB4 draft reference sequences have comparatively poorer assembly qualities of $Q28$ and $Q23$, respectively (compared to $Q28$ and $Q29$ for 3D7 and HB3). After filtering, we detected a total of 972 *de novo* mutations (163 SNVs, 348 insertions, 322 deletions, 19 MNVs, 7 NAHR events, and 113 incompletely assembled events).

The average per sample DNM count is low, with short indels (~ 5.47 per sample) outnumbering SNVs (~ 1.29 per sample). We applied the RepeatMasker(36) software to annotate repetitive genomic sequences and Spine(28) to annotate accessory genomic regions (sequences private to each parasite isolate, typically encompassing subtelomeric/hypervariable regions) across all parental genomes. We then inspected variant locations with respect to these annotations (Figure 4b-c). Aggregated across all samples and crosses, we found a 3-fold enrichment of mutations occurring in repetitive genomic regions, $\sim 90\%$ of which fell within tandem duplications. Mutations were enriched in the accessory (~ 2 Mb) versus core (~ 21 Mb) genomic compartments (SNVs: $p = 1.5 \times 10^{-7}$; INS/DEL/MNV: $p < 2.2 \times 10^{-16}$; BND: $p < 2.2 \times 10^{-16}$; based on chi-squared tests accounting for indel lengths and number of novel k-mers appearing in NAHR events, see Supplemental Material). We observed similar per-sample mutation distributions across samples.

We computed per-sample per-nucleotide mutation rates across all four crosses. Additionally, as DNMs can continue to accumulate in each parasite during the *in vitro* intraerythrocytic lifecycle, we computed mutational rates per nucleotide and generation. However, culture time and lifecycle time for cross progeny was not always known. Assuming a culture time of 52 days between initial cloning and sequencing (the average of the documented culture times for the 3D7xHB3 and HB3xDD2 cross progeny), and a mitotic generation time of 48 hours(39), Per-nucleotide mutational rates are presented in Table 2. These rates are broadly consistent across

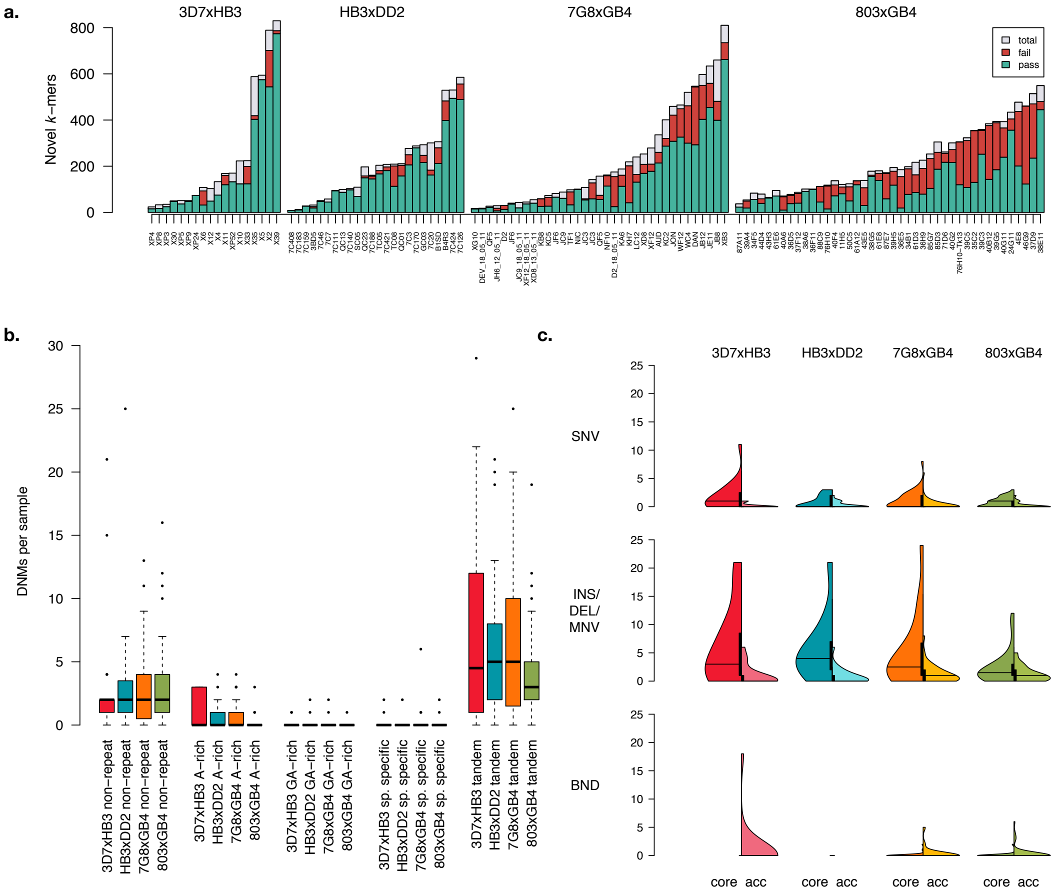


Figure 4: Per-sample DNM discovery metrics in 119 ***P. falciparum*** **progeny.** (**a**) Novel *k*-mers per cross and sample (grey bars). For those contained within successfully assembled variants, *k*-mers in variant passing filters are shown in green; the rest are shown in red. (**b**) Per cross DNM sample distributions for mutations appearing in repetitive regions of the respective parental genomes. (**c**) Violin plots showing DNM sample distributions per cross, split by those in core genomic regions (left) and accessory regions (right).

Table 2: Mutation rates per cross, sample, mutational class, and genome compartment

	3D7xHB3	HB3xDD2	7G8xGB4	803xGB4
Culture time (days)[a]	47	57	52[b]	52[b]
Lifecycle time (hours)[c]	48	48	48	48
Progeny	18	24	35	42
Genome length (bp)[d]				
core	20810915	21052828	21325706	21303692
(accessory)	(1860495)	(1603876)	(2368812)	(2389390)
Total variants[e]				
SNVs	32 (3)	21 (7)	43 (14)	33 (10)
Indels	114 (19)	141 (21)	188 (46)	106 (54)
NAHRs	0 (2)	0 (0)	0 (2)	0 (3)
Rate ($sample^{-1}\ bp^{-1}$)[e]				
SNVs	8.1e-09 (8.5e-08)	3.9e-08 (1.7e-07)	5.3e-08 (1.6e-07)	3.5e-08 (9.5e-08)
Indels	1.7e-06 (4.0e-06)	1.8e-06 (1.4e-06)	1.4e-06 (4.6e-06)	8.1e-07 (3.3e-06)
NAHRs	0 (6.3e-06)	0 (0)	0 (7.9e-07)	0 (9.4e-07)
Rate ($sample^{-1}\ bp^{-1}\ gen^{-1}$)[e]				
SNVs	3.4e-09 (3.6e-09)	1.4e-09 (6.1e-09)	2.0e-09 (5.9e-09)	1.4e-09 (3.7e-09)
Indels	1.2e-08 (2.3e-08)	9.4e-09 (1.8e-08)	8.9e-09 (1.9e-08)	4.3e-09 (1.9e-08)
NAHRs	0 (2.4e-09)	0 (0)	0 (8.5e-10)	0 (1.1e-09)

[a] Culture time estimates from Claessens *et al.* 2014.

[b] Assumed as mean of 3D7xHB3 and HB3xDD2 culture times.

[c] Assumed from Trampuz *et al.* 2003.

[d] Averaged lengths of core and accessory regions for each parent.

[e] Core (accessory) rates shown outside (inside) parentheses.

crosses and compartments, and with previous estimates based on parasite clone trees(2; 4).

2.9 Hypothesis-free discovery of NAHR events to basepair resolution

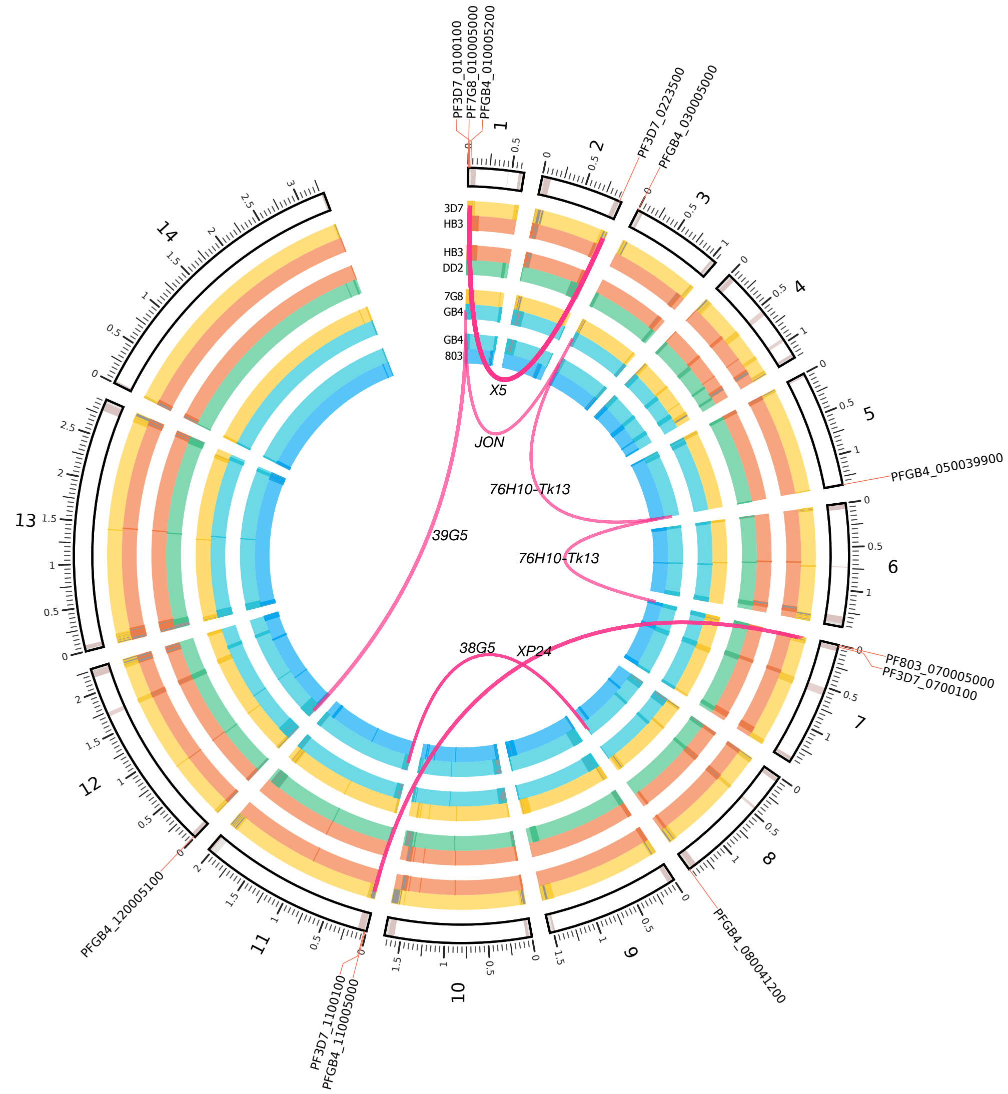


Figure 5: Circos(18) plot of NAHR events detected in all 119 **samples across four *P. falciparum* experimental crosses.** Parental genomes for each cross are depicted in the inner grouped circular tracks. Bézier curves depict each translocation event, with termini indicating the parent(s) of origin and a label at the apex of the curve identifying the sample in which it was found. Closest gene names annotated on outer circumference. Dark bands indicate accessory regions determined by the Spine(28) software, except in the outer ideogram, which is based on alignability maps for the canonical 3D7 reference genome(25).

To detect NAHR events, we grouped proximate breakend calls and applied three filtration criteria: (1) events must contain 20 or more novel *k*-mers, (2) consist of 3 or more breakends, and (3) at least one contig must link distal genomic loci within the same contig. We detected

Table 3: Nearest genes to NAHR breakends

Gene	3D7 ortholog	Encodes	Function
Pf3D7_0100100	-	PfEMP1	immune evasion
Pf3D7_0223500	-	PfEMP1	immune evasion
Pf3D7_0700100	-	PfEMP1	immune evasion
Pf3D7_1100100	-	PfEMP1	immune evasion
Pf7G8_010005000	unknown	hypothetical protein, conserved	unknown
Pf803_070005000	Pf3D7_0100100	PfEMP1	immune evasion
PfGB4_010005200	unknown	hypothetical protein, conserved	unknown
PfGB4_030005000	unknown	hypothetical protein, conserved	unknown
PfGB4_050039900	PF3D7_0700200	RIF	variant surface antigen
PfGB4_080041200	PF3D7_0100100	PfEMP1	immune evasion
PfGB4_110005000	PF3D7_0223500	PfEMP1	immune evasion
PfGB4_120005100	unknown	hypothetical protein, conserved	unknown

7 NAHR events in total after applying, depicted in Figure 5. All occurred in subtelomeric accessory regions of the genomes. To determine genes nearest to each NAHR breakend, we transferred existing 3D7 gene models and performed *ab initio* gene prediction on each parental genome via the Companion(37) annotation server. All but four of the genes closest to a breakend were related to antigenic gene families and immune evasion.

Previous work on NAHR events in *P. falciparum* – based on observations of apparent translocations of *var* gene sequences, and limited by inadequate reference sequences for parasites other than 3D7 – have only reported NAHR events within the exon 1 of *var*-gene family members(2; 4; 6; 11; 10; 7; 35). As we enforce no *a priori* hypothesis on which loci are likely to harbor such recombinations, the discovered events in our dataset extend beyond *var* exon 1. While the events still occur in the subtelomeric regions of the genome (within which many other genes related to immune evasion reside), five of the twelve genes proximate to the NAHR breakends were not *var* genes. A single event occurred near a gene from the *rif* gene family.

Beyond identifying new NAHR events outside of the usual *var*-gene repertoire, we were also able to clarify the extent of previously observed events. Figure 6 depicts three of the detected NAHR events. In Figure 6a, our calls recapitulate previously reported rearrangements (breakends 5-9) within the long exons of Pf3D7_0100100 and Pf3D7_0223500[‡](35). Flanking these known breakends are a number of mutations that have not been previously reported, including an additional series of breakends upstream of each of the *var* genes (1 and 2), two MNVs (3 and 4), and an SNV within the coding region of the antigenic gene on chromosome 2. In panel b, a novel NAHR event is shown with a recombination path that weaves in and out of coding regions, touching upon the previously unexamined exon 2. The recombination path within the novel event in panel c (within a sample in the previously unpublished 803xGB4 cross), remains wholly within the coding sequence.

[‡] PFA0005w and PFB1055c in older nomenclature

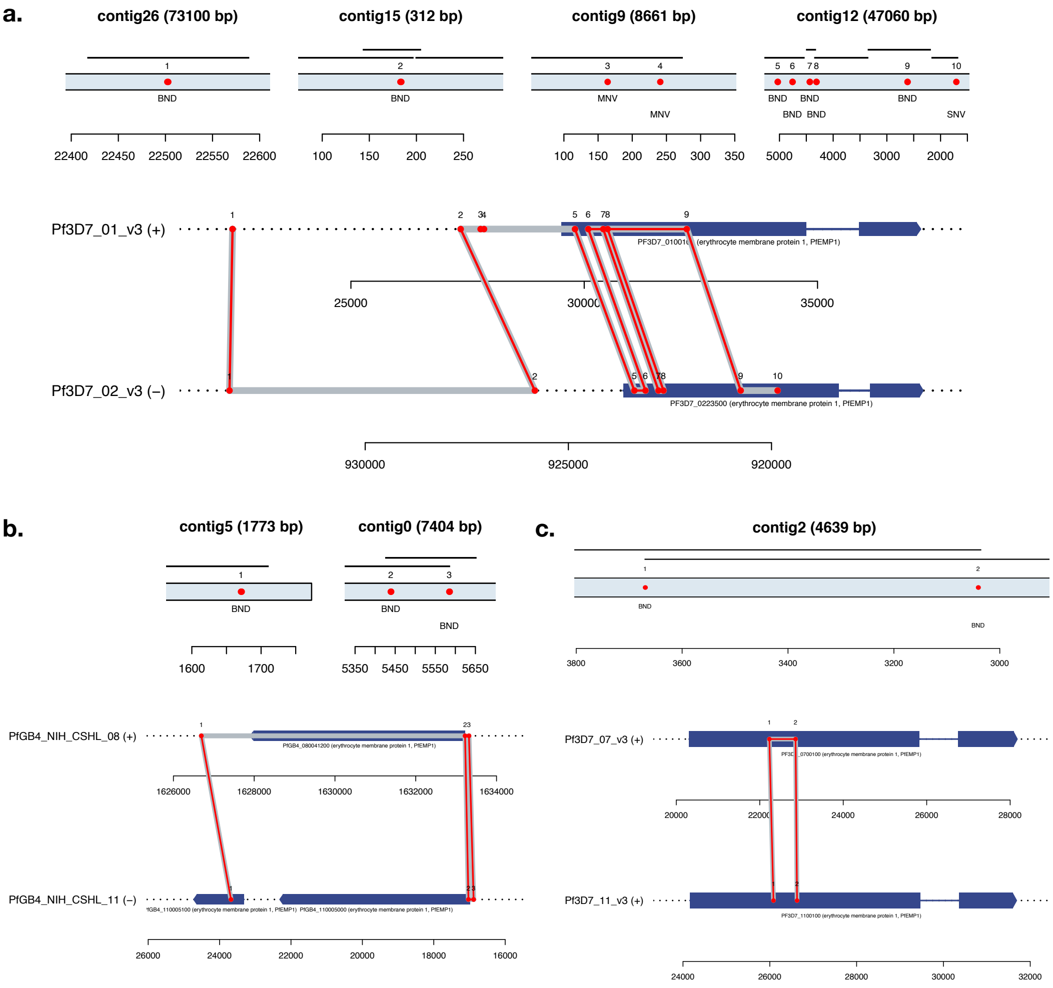

a.
contig26 (73100 bp)
contig15 (312 bp)
contig9 (8661 bp)
contig12 (47060 bp)
1
BND
2
BND
3
4
MNV
MNV
5
6
7 8
9
10
BND
BND
BND
BND
BND
SNV
22400
22450
22500
22550
22600
100
150
200
250
100
150
200
250
300
350
5000
4000
3000
2000
Pf3D7_01_v3 (+)
Pf3D7_02_v3 (–)
PF3D7_01001 (erythrocyte membrane protein 1, PfEMP1)
PF3D7_0223500 (erythrocyte membrane protein 1, PfEMP1)
25000
30000
35000
930000
925000
920000
b.
contig5 (1773 bp)
contig0 (7404 bp)
1
BND
2
3
BND
BND
1600
1700
5350
5450
5550
5650
PfGB4_NIH_CSHL_08 (+)
PfGB4_NIH_CSHL_11 (–)
PfGB4_080041200 (erythrocyte membrane protein 1, PfEMP1)
fGB4_110005100 (erythrocyte membrane protein 1, PfEMP1)
PfGB4_110005000 (erythrocyte membrane protein 1, PfEMP1)
1626000
1628000
1630000
1632000
1634000
26000
24000
22000
20000
18000
16000
c.
contig2 (4639 bp)
1
2
BND
BND
3800
3600
3400
3200
3000
Pf3D7_07_v3 (+)
Pf3D7_11_v3 (+)
PF3D _0700100 (erythrocyte membrane protein 1, PfEMP1)
PF3D7_1100100 (erythrocyte membrane protein 1, PfEMP1)
20000
22000
24000
26000
28000
24000
26000
28000
30000
32000

Figure 6: Three of the detected NAHR events in the *P. falciparum* crosses. (**a**) NAHR event involving two *var* genes in 3D7xHB3 progeny X5 (Pf3D7_0100100 on chr1, Pf3D7_0223500 on chr2). Top: LdBG contigs spanning mutation (dBG contig shown as thin black line for comparison). Called mutations shown along contig as red points. Bottom: mutations from LdBG contigs in genomic context shown in red. Gene models shown in dark blue (thick lines: exonic sequence; thin lines: intronic sequence). Inferred recombination path shown in gray. (**b**) NAHR event in 803xGB4 sample 38G5 (PfGB4_080041200 on chr8, PfGB4_11005100 and PfGB4_11005000 on chr11). (**c**) NAHR event in 3D7xHB3 sample XP24 (Pf3D7_0700100 on chr7 and Pf3D7_1100100 on chr11).

2.10 Variant calling with cumulatively expanding reference set

Exploring beyond comparisons of progeny-to-progenitor genomes, we hypothesized that genomic novelty present in a sample but not placeable on the background of an evolutionarily distant reference sequence would be better elucidated through the simultaneous use of multiple reference sequences. We obtained Illumina data and constructed a PacBio draft assembly of an 803xGB4 progeny (36F11). From the 36F11 data, we extracted *k*-mers that were novel with respect to the 3D7 reference genome, further filtering these *k*-mers based on presence in the counterpart 36F11 clone assembly, thus constructing a conservative *k*-mer list that flags true variation in the 36F11 parasite. We used this list to seed variant calls, increasing the number of reference sequences provided with each callset.

Figure 7 depicts the calling results on 36F11 with the cumulative addition of 3D7, HB3, 7G8, DD2, GB4, and 803 reference sequences. As novel *k*-mers are computed with respect to 3D7, calls at these *k*-mers can only be homozygous-variant. As additional reference sequences are added, variants are described against a new background sequence. However, many novel *k*-mers tagging variation against 3D7 are no longer considered novel w.r.t. another reference sequence, and their reconstructed sequence for the progeny is perfectly homologous to the additional reference. Thus, as more reference sequences are added, apparent variation against 3D7 is re-described as homozygous-variant (hashed bars) or homozygous-reference to a sequence other than 3D7 (solid bars). When using all six reference sequences, our ability to characterize apparent novelty to 3D7 grows from 40% to 95%.

3 DISCUSSION

We have presented a graph-based *de novo* mutation calling method, available through our software Corticall, that is capable of discovering simple and complex variants in pedigrees and experimental crosses without bias towards a reference sequence. Our approach leverages long-haplotype data derived from any source (existing finished genomes, draft assemblies from third-generation sequencing, targeted sequencing of specific loci, etc.) to improve the assemblies of other short-read datasets. These long-haplotype samples need not be from the same sample. Short-read data is used to establish graph topology, while long-haplotype data is aligned to the graph but constrained to specify connectivity information only. Sequencing errors (and possibly mutations) are always adjudicated in favor of the existing graph, thus no new sequence is added, only navigation information. This approach opens the opportunity for multiple long-read datasets to be used to improve the connectivity of many more short-read assemblies.

Corticall can leverage many finished or draft reference-quality datasets, seamlessly transitioning between connectivity information sets during assembly. This affords a powerful approach to the hypothesis-free study of *de novo* mutations. As many of these events occur in repetitive or genetically diverse regions of the genome, the use of multiple reference sequences during assembly helps to provide access to so-called genomic “dark matter” - loci underserved

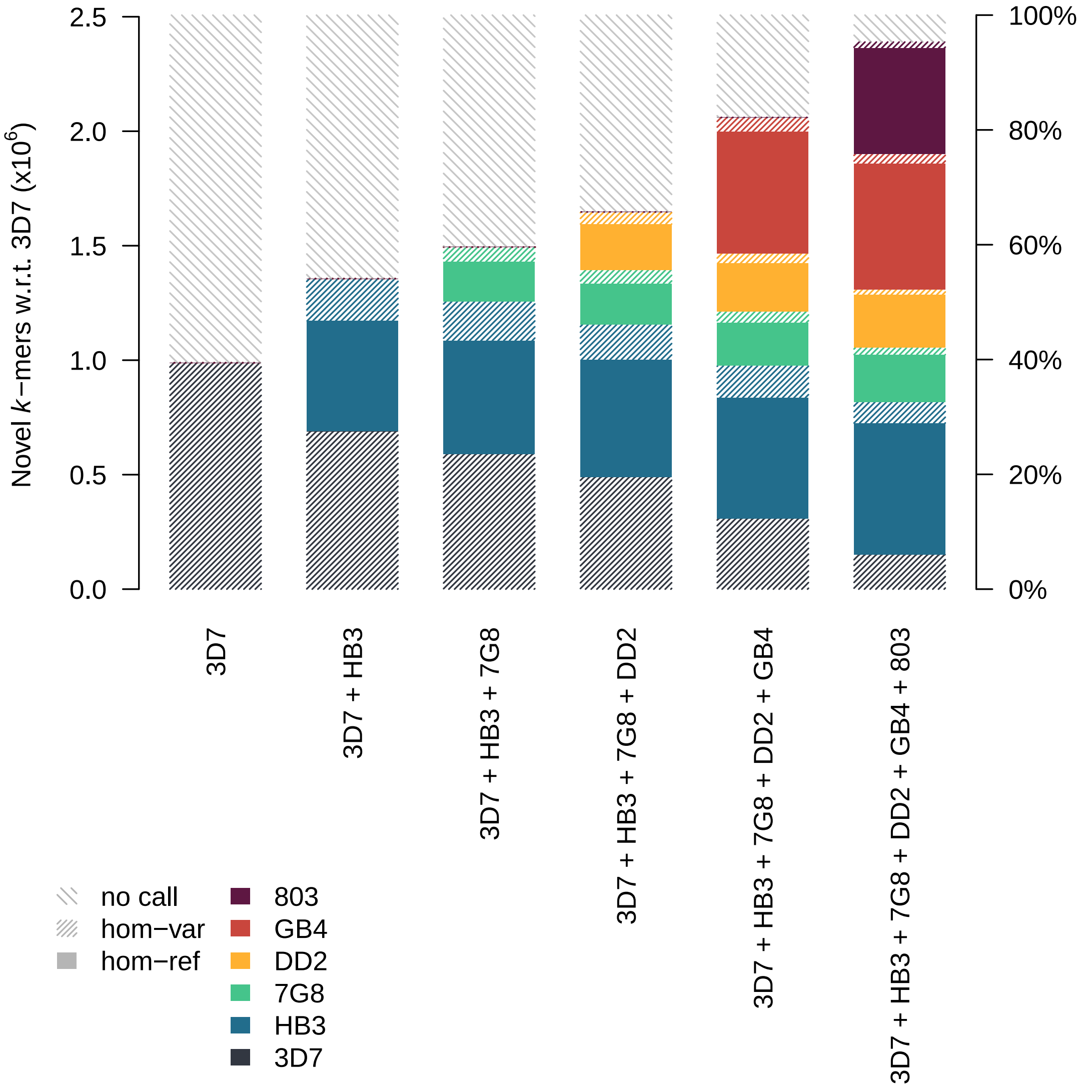


Figure 7: Calls tagged by 36F11 *k*-mers novel with respect to 3D7, redescribed against combinations of other reference sequences. Stacked bars represent fraction of novel *k*-mers linked to homozygous-reference (hom-ref) and homozygous-variant (hom-var) calls, or *k*-mers where no call could be made. Colors represent the specific haplotypic background the call was placed on (if a call can be equally described on multiple backgrounds, one is chosen at random).

by pure short-read *de novo* assembly or a single canonical reference.

Corticall assembles variants, not genomes, and keeps false discovery rates low by only inspecting regions of the genome harboring novel *k*-mers. By combining local, multi-sample assembly with a simultaneous alignment/recombination model, we are able to detect a wide variety of mutational types with a single, consistent framework. Additionally, tracking the number of novel *k*-mers explained by each variant call provides a useful metric for determining the completeness of the final callset.

In the *P. falciparum* crosses, we detected SNVs at rates broadly consistent with previous work, and indels at more than four times the SNV rate. We detected new NAHR events, all in subtelomeric regions of the genome that are not represented in the canonical reference. For previously discovered NAHR events, we are able to find additional breakends in nearby non-coding regions, establishing a more complete picture of non-allelic recombination behavior in these pathogens. Much of the *de novo* mutational spectrum appears in accessory regions. These compartments are diverse in the population precisely because they typically harbor clinically relevant genes underlying drug resistance or immune escape functionality. The mapping-free, reference-agnostic approach espoused by Corticall thus enables the detection of this clinically-relevant variation, and removes the requirement for determining the appropriate genome reference for mapping and analysis.

The fixed record size structure of Cortex graphs used with Corticall enables storage in an ordered, randomly accessible manner, thus keeping memory requirements low as the entire graph need not be loaded into memory in order to be inspected. Pre-determining the novel *k*-mers to inspect, along with intelligent caching to prevent redundant lookups when assembling multiple samples over shared *k*-mers, reduces disk accesses. As a result, Corticall is able to scale to genomes of any size. This may provide a valuable approach to the study of Mendelian disease in large pedigrees or tumor/normal pairs (wherein the normal can be considered as the parent of the tumor samples).

Corticall has several limitations, addressable by future work. While Corticall need not load an entire graph into memory to perform variant calling, the genome assembly software upon which it relies *does* require that the entire graph be stored in RAM as it is being constructed. Thus, even though the variant calling step on human data can be done in as little as 1 Gb of RAM, the initial *de novo* assembly step still requires hundreds of gigabytes of memory to execute. Recent approaches to streaming graph construction(32) and/or succinct de Bruijn graphs(5; 27) may well address this limitation.

Additionally, our use of long haplotype data is restricted to sequences that have been substantially error-corrected. Typically *k*-mer sizes used in de Bruijn graph-based short read assemblies (e.g. $k = 31 - 96$) are still too high for the long, error-prone reads generated by third-generation sequencers. However, lowering the *k*-mer size of the short read assemblies to a length more likely to result in a perfect match on the long read data (e.g. $k = 11$) would result in too many junctions from homologous sequences in the graph. Our current approach to error-correcting long reads against the graph requires that the path through the existing graph contain no junctions, and would thus be impaired by setting the *k*-mer size to low. A more computationally expensive read-to-graph alignment procedure could remedy this limitation.

As third-generation sequencing continues to mature, the construction of additional draft reference genomes will become more accessible. The utility of this data extends beyond pure *de novo* assembly for constructing new reference sequences, or for elucidating structural variation in single samples. Strategic choices as to which samples to sequence with long reads can enable simple and complex variant discovery in a much larger cohort while simultaneously keeping costs low, provided that variant calling methods are capable of leveraging such information. Corticall is a step forward in this direction, presenting a uniform approach to variant discovery and typing that combines assembly, alignment, recombination models, and third-generation reference sequence panels. Such approaches will assist in overcoming bias to a single canonical reference sequence and enable a more complete description of variation in diverse populations.

URLs. Corticall, part of the CortexJDK package: https://github.com/mcveanlab/CortexJDK .

4 METHODS

Methods and associated references are available in the Supplemental Material for this manuscript.

Accession codes. PacBio and Illumina sequence data have been deposited at the European Nucleotide Archive (ENA) (see Supplemental Material for accession numbers).

ACKNOWLEDGEMENTS
The authors thank Eric Antoniou, Sara Goodwin, Michael Schatz, and the CSHL PacBio sequencing service; Winni Kretzschmar and Karl Johan Westrin for helpful code improvements reducing the memory usage of the Tesserae model; and Isaac Turner, Patrick Albers, Jerome Kelleher, and Marcus Tutert for helpful technical discussions and manuscript review. This work was supported by grants from the Wellcome Trust (numbers 090532/Z/09/Z and 100956/Z/13/Z) and the Li Ka Shing Foundation (to G.M.). K.V.G. was supported by Wellcome Trust Research Studentship award (097310/Z/11/Z). Z.I. was funded by a Wellcome Trust/Royal Society Sir Henry Dale Fellowship (102541/Z/13/Z).

AUTHOR CONTRIBUTIONS
K.V.G. developed algorithms and pipelines for identifying mutations in sequence assembly graphs. M.A.K., S.C., M.K., E.D., and J.S. oversaw the culturing of malaria parasites and subsequent DNA extraction for PacBio sequencing. D.K. and T.W. provided lab support for the parasite culturing and access to Illumina data on all four *P. falciparum* crosses. Z.I. and G.M. provided access to critical resources. K.V.G., Z.I., and G.M. wrote the manuscript.

COMPETING FINANCIAL INTERESTS
G.M. is a founder and director of Genomics PLC and a partner in Peptide Groove LLP.

References

[1] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, Mar. 2011.

[2] S. E. R. Bopp, M. J. Manary, A. T. Bright, G. L. Johnston, N. V. Dharia, F. L. Luna, S. McCormack, D. Plouffe, C. W. McNamara, J. R. Walker, D. A. Fidock, E. L. Denchi, and E. A. Winzeler. Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families. *PLoS genetics*, 9(2):e1003293, Feb. 2013.

[3] C. M. B. Carvalho and J. R. Lupski. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet*, 17(4):224–238, Apr. 2016.

[4] A. Claessens, W. L. Hamilton, M. Kekre, T. D. Otto, A. Faizullabhoy, J. C. Rayner, and D. Kwiatkowski. Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLoS genetics*, 10(12):e1004812, Dec. 2014.

[5] T. C. Conway and A. J. Bromage. Succinct data structures for assembling large genomes. 27(4):479–486, Feb. 2011.

[6] K. W. Deitsch, A. del Pinal, and T. E. Wellems. Intra-cluster recombination and var transcription switches in the antigenic variation of Plasmodium falciparum. *Molecular and biochemical parasitology*, 101(1-2):107–116, June 1999.

[7] M. F. Duffy, T. J. Byrne, C. Carret, A. Ivens, and G. V. Brown. Ectopic recombination of a malaria var gene during mitosis associated with an altered var switch rate. *Journal of molecular biology*, 389(3):453–469, June 2009.

[8] R. Durbin. *Biological Sequence Analysis*. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Apr. 1998.

[9] P. Flicek and E. Birney. Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6(11s):S6–S12, Nov. 2009.

[10] M. Frank, L. Kirkman, D. Costantini, S. Sanyal, C. Lavazec, T. J. Templeton, and K. W. Deitsch. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of Plasmodium falciparum. *International Journal for Parasitology*, 38(10):1099–1109, Aug. 2008.

[11] L. H. Freitas-Junior, E. Bottius, L. A. Pirrit, K. W. Deitsch, C. Scheidig, F. Guinet, U. Nehrbass, T. E. Wellems, and A. Scherf. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature*, 407(6807):1018–1022, Oct. 2000.

[12] M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M.-S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. A. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, Oct. 2002.

[13] S. Goodwin, J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. C. Schatz, and W. R. McCombie. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research*, 25(11):1750–1756, Nov. 2015.

[14] K. Hayton, D. Gaur, A. Liu, J. Takahashi, B. Henschen, S. Singh, L. Lambert, T. Furuya, R. Bouttenot, M. Doll, F. Nawaz, J. Mu, L. Jiang, L. H. Miller, and T. E. Wellems. Erythrocyte Binding Protein PfRH5 Polymorphisms Determine Species-Specific Pathways of Plasmodium falciparum Invasion. *Cell Host & Microbe*, 4(1):40–51, Jan. 2008.

[15] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44(2):226–232, Feb. 2012.

[16] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 14:S18, Jan. 2018.

[17] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and A. M. Phillippy. Hybrid error correction and ¡i¿de novo¡/i¿ assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700, July 2012.

[18] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. 19(9):1639–1645, June 2009.

[19] G. Landan and D. Graur. Characterization of pairwise and multiple sequence alignment errors. *Gene*, 441(1-2):141–147, July 2009.

[20] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239, Apr. 1988.

[21] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org*, page arXiv:1303.3997, Mar. 2013.

[22] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, Dec. 2003.

[23] J. R. Lupski. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. *Genome Biology*, 5(10):242, Sept. 2004.

[24] J. R. Lupski. Genomic disorders ten years on. *Genome Medicine*, 1(4):42, Apr. 2009.

[25] A. Miles, Z. Iqbal, P. Vauterin, R. Pearson, S. Campino, M. Theron, K. Gould, D. Mead, E. Drury, J. O'Brien, V. Ruano Rubio, B. MacInnis, J. Mwangi, U. Samarakoon, L. Ranford-Cartwright, M. Ferdig, K. Hayton, X.-Z. Su, T. Wellems, J. Rayner, G. McVean, and D. Kwiatkowski. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Research*, 26(9):1288–1299, Sept. 2016.

[26] A. Miles, Z. Iqbal, P. Vauterin, R. Pearson, S. Campino, M. Theron, K. Gould, D. Mead, E. Drury, J. O'Brien, V. R. Rubio, B. MacInnis, J. Mwangi, U. Samarakoon, L. Ranford-Cartwright, M. Ferdig, K. Hayton, X. Su, T. Wellems, J. Rayner, G. McVean, and D. Kwiatkowski. Genome variation and meiotic recombination in Plasmodium falciparum: insights from deep sequencing of genetic crosses. *bioRxiv*, page 024182, Aug. 2015.

[27] M. D. Muggli, A. Bowe, N. R. Noyes, P. S. Morley, K. E. Belk, R. Raymond, T. Gagie, S. J. Puglisi, and C. Boucher. Succinct colored de Bruijn graphs. 33(20):3181–3187, Oct. 2017.

[28] E. A. Ozer, J. P. Allen, and A. R. Hauser. Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic tools Spine and AGEnt. *BMC genomics*, 15(1):737, Dec. 2014.

[29] M. M. Parks, C. E. Lawrence, and B. J. Raphael. Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biology*, 16(1):704, 2015.

[30] S. N. Redmond, B. M. MacInnis, S. Bopp, A. K. Bei, D. Ndiaye, D. L. Hartl, D. F. Wirth, S. K. Volkman, and D. E. Neafsey. De Novo Mutations Resolve Disease Transmission Pathways in Clonal Malaria. *Molecular biology and evolution*, 11(6):e0164397, May 2018.

[31] A. Rhoads and K. F. Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5):278–289, Oct. 2015.

[32] R. Rozov, G. Goldshlager, E. Halperin, and R. Shamir. Faucet: streaming de novo assembly graph construction. 34(1):147–154, Jan. 2018.

[33] J. M. Sá, S. R. Kaslow, M. A. Krause, V. A. Melendez-Muniz, R. E. Salzman, W. A. Kite, M. Zhang, R. R. M. Barros, J. Mu, P. K. Han, J. P. Mershon, C. E. Figan, R. L. Caleon, R. S. Rahman, T. J. Gibson, C. Amaratunga, E. P. Nishiguchi, K. F. Breglio, T. M. Engels, S. Velmurugan, S. Ricklefs, J. Straimer, N. F. Gnadig, B. Deng, A. Liu, A. Diouf, K. Miura, G. S. Tullo, R. T. Eastman, S. Chakravarty, E. R. James, K. Udenze, S. Li, D. E. Sturdevant, R. W. Gwadz, S. F. Porcella, C. A. Long, D. A. Fidock, M. L. Thomas, M. P. Fay, B. K. L. Sim, S. L. Hoffman, J. H. Adams, R. M. Fairhurst, X.-z. Su, and T. E. Wellems. Artemisinin resistance phenotypes and K13 inheritance in a Plasmodium falciparum cross and Aotus model. *Proceedings of the National Academy of Sciences*, pages 1–6, Nov. 2018.

[34] L. Salmela and E. Rivals. LoRDEC: accurate and efficient long read error correction. 30(24):3506–3514, Dec. 2014.

[35] A. F. Sander, T. Lavstsen, T. S. Rask, M. Lisby, A. Salanti, S. L. Fordyce, J. S. Jespersen, R. Carter, K. W. Deitsch, T. G. Theander, A. G. Pedersen, and D. E. Arnot. DNA secondary structures are associated with recombination in major Plasmodium falciparum variable surface antigen gene families. *Nucleic acids research*, Nov. 2013.

[36] A. Smit, R. Hubley, and P. Green. *RepeatMasker Open-4.0*, 2013.

[37] S. Steinbiss, F. Silva-Franco, B. Brunk, B. Foth, C. Hertz-Fowler, M. Berriman, and T. D. Otto. Companion: a web server for annotation and analysis of parasite genomes. *Nucleic acids research*, 44(W1):W29–W34, July 2016.

[38] L. Tattini, R. D'Aurizio, and A. Magi. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Frontiers in Bioengineering and Biotechnology*, 3:92, 2015.

[39] A. Trampuz, M. Jereb, I. Muzlovic, and R. M. Prabhu. Clinical review: Severe malaria. *Critical care (London, England)*, 7(4):315–323, Aug. 2003.

[40] I. Turner, K. V. Garimella, Z. Iqbal, and G. McVean. Integrating long-range connectivity information into de Bruijn graphs. *Bioinformatics (Oxford, England)*, 34(15):2556–2565, Aug. 2018.

[41] S. K. Volkman, P. C. Sabeti, D. DeCaprio, D. E. Neafsey, S. F. Schaffner, D. A. Milner, J. P. Daily, O. Sarr, D. Ndiaye, O. Ndir, S. Mboup, M. T. Duraisingh, A. Lukens, A. Derr, N. Stange-Thomann, S. Waggoner, R. Onofrio, L. Ziaugra, E. Mauceli, S. Gnerre, D. B. Jaffe, J. Zainoun, R. C. Wiegand, B. W. Birren, D. L. Hartl, J. E. Galagan, E. S. Lander, and D. F. Wirth. A genome-wide map of diversity in Plasmodium falciparum. *Nature genetics*, 39(1):113–119, Dec. 2006.

[42] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PloS one*, 9(11):e112963, Nov. 2014.

[43] D. Walliker, I. Quakyi, T. Wellems, T. McCutchan, A. Szarfman, W. London, L. Corcoran, T. Burkot, and R. Carter. Genetic analysis of the human malaria parasite Plasmodium falciparum. 236(4809):1661–1666, June 1987.

[44] T. E. Wellems, L. J. Panton, I. Y. Gluzman, V. E. do Rosario, R. W. Gwadz, A. Walker-Jonah, and D. J. Krogstad. Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. *Nature*, 345(6272):253–255, May 1990.

[45] T. E. Wellems, A. Walker-Jonah, and L. J. Panton. Genetic mapping of the chloroquine-resistance locus on Plasmodium falciparum chromosome 7. 88(8):3382–3386, Apr. 1991.

[46] B. Xue and L. He. An expanding universe of the non-coding genome in cancer biology. *Carcinogenesis*, 35(6):1209–1216, Apr. 2014.

[47] X. Zheng-Bradley, I. Streeter, S. Fairley, D. Richardson, L. Clarke, P. Flicek, and the 1000 Genomes Project Consortium. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *GigaScience*, 6(7):1–8, May 2017.

[48] M. M. Zilversmit, E. K. Chase, D. S. Chen, P. Awadalla, K. P. Day, and G. McVean. Hypervariable antigen genes in malaria have ancient roots. *BMC Evolutionary Biology*, 13(1):110, May 2013.

[49] D. Zwillinger. *CRC Standard Mathematical Tables and Formulae, 32nd Edition*. CRC Press, June 2011.