Supporting Information

Barcoding analysis

Determination of polygenomic, monogenomic, and uninformative barcodes. Barcodes containing more than one position where both alleles were detected were regarded as from polygenomic infections and were omitted from the analysis. Samples missing more than 5 positions in the molecular barcode were also eliminated from our analysis.

Estimation of identity by descent

Genotype calls. A consensus sequence was called for each strain using the GATK Unified Genotyper v1.2.3-g61b89e2 with the following parameters: -A AlleleBalance -stand_emit_conf 0 –output_mode EMIT_ALL_SITES. Poor-quality base calls were removed for GQ less than 30 or QUAL less than 60 or if they called a heterozygous genotype.

Discordance. *P. falciparum* SNPs typically have low minor allele frequency, meaning that most sites are uninformative when comparing two sequences. Accordingly, pairwise discordance was calculated using only those sites with two genotype calls passing quality control, at least one of which was the minor allele. Discordance was defined as the fraction of such sites with discordant genotypes.

Hidden Markov model. To identify specific regions of genomes that were identical by descent, we constructed a Hidden Markov Model comparing pairs of samples. The model has two underlying states describing the sequences being compared: identical (*I*) and non-identical or discordant (*D*). Observations are made at each SNP site (again requiring at least one minor allele call), with the probability of observing concordant or discordant genotypes determined by the *I* versus *D* state, the allele frequency at that site, and the error rate; SNPs were omitted if the preceding SNP was within 10 base pairs. The probability of a transition between states occurring between two SNPs is proportional to the physical distance between them. In more detail:

Initial state probability (start of each chromosome):

$$p(I) = p(D) = 0.5$$

Transition probability between SNP i and $i+1$:

$$p(I_{i+1}|D_i) = p(D_{i+1}|I_i) \equiv p_{tran} = k\rho\Delta d$$

$$p(I_{i+1}|I_i) = p(D_{i+1}|D_i) = 1 - p_{tran}$$

where ρ = estimated recombination rate = 5.8 x 10^{-7} per base pair and

Δd = physical distance between SNPs in base pairs.

The conditional probabilities are as follows:

$$p(discordance|I) = 2\varepsilon\,(1-\varepsilon)$$

$$p(discordance|D) = 2f(1-f)((1-\varepsilon)^2+\varepsilon^2) + 2\varepsilon(1-\varepsilon)((1-f)^2+f^2)$$

$$p(concordance|I) = (1-\varepsilon)^2+\varepsilon^2$$

$$p(concordance|D) = f^2(1-\varepsilon)^2 + (1-f)^2\varepsilon^2 + 2f(1-f)\varepsilon(1-\varepsilon)$$

where f = minor allele frequency

ε = genotyping error rate

The genotyping error rate, estimated from comparisons between ostensibly identical samples, was estimated to be 0.1%. Minor allele frequencies were calculated from the full set of 178 samples. The only remaining free parameter in the model is *k*, the scaling coefficient for the recombination rate. We calculated the maximum likelihood value of *k* for each pair of related samples (those with discordance < 0.85); optimal values ranged from 0.25 to 3.0, except for one (otherwise unremarkable) pair with a value of 11. Conclusions about relatedness turned out to be almost completely independent of *k* for values in this range, however, so we used a single value (*k* = 2, the median value) for all subsequent analyses. The most probable assignment of hidden states was made using the Viterbi algorithm.

Network relatedness (Figure 1)

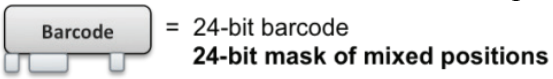
Genetic distances between barcodes were first calculated using the Maximum Composite Likelihood model (Ref. 16). The analysis involved 236 nucleotide sequences. All positions with less than 95% site coverage were eliminated. That is, fewer than 5% alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 19 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.

Figure 1A shows only genetic distances < 0.2, while the sequence data in 1B shows connections if the shared fraction is >= 0.04. Clustering in both parts of the figure was done with the program Gephi, using the "Force Atlas" algorithm, which creates a force between nodes (ranging from attractive to repulsive) that varies with the weight of the edge between them. Parameters used for Force Atlas were as follows: Inertia: 0.1; Repulsion strength: 600; Attraction strength: 2.0; Maximum displacement: 10.0; Auto stabilize: on; Autostab strength: 80; Autostab sensibility: 0.2; Gravity: 10; Attraction distribution: on; Adjust by sizes: on.

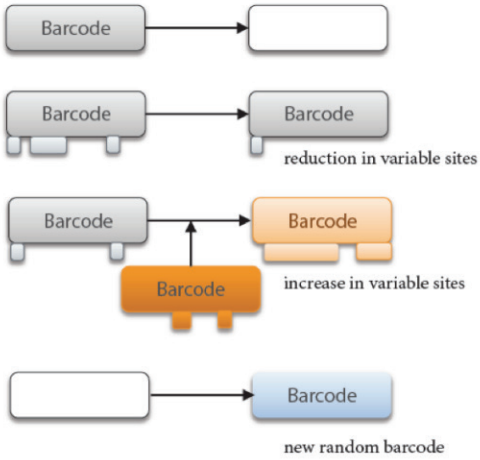
Epidemiological model fitted to barcode data

Model state. The barcode model tracks the state of a parasite population as a collection of individual infections, each of which is characterized by a 24-position barcode. A single

human infection may consist of multiple clones deposited in the same inoculation event, a phenomenon that manifests itself with both variants present at one or more positions in the barcode. The locations of the mixed positions are stored in another 24-bit mask.



Dynamic processes. The state of the system evolves according to four dynamical processes. Existing infections may expire, propagate clonally to a new host, or outcross with another infection. Additionally, infections with random barcodes may be imported into the simulation.



In the process of clonal propagation, the number of barcode positions with both alleles present is reduced as a result of small oocyst numbers in the vector and infected hepatocytes in the host. In the model, the number of mixed positions in the next generation $P\left(N_m^{g+1}\right)$ is approximated by drawing from a zero-inflated Poisson distribution with inflated fraction α and otherwise mean fractional reduction β in the number of mixed positions from the previous generation N_m^g:

$$P\left(N_m^{g+1}\right) = \alpha + (1-\alpha)Poisson(\beta N_m^g)$$

For outcrossing of infections acquired from multiple inoculation events, the model chooses two strains at random from the population of parasite infections. The 24 barcode positions are treated as unlinked, such that the progeny have an independent chance at each position to inherit from either parent. As described in the following section, the rate of random outcrossing is related to the fraction of the host population currently infected with parasites from multiple inoculations.

Time dependence of process rate. The daily rate at which existing infections expire (R_{exp}) is a constant parameter of the model.

The rate R_{rep} at which existing infections reproduce, *i.e.* the sum of the clonal propagation and outcrossing rates, is related to the expiration rate by the reproduction

number $R_0(t)$, which varies according to seasonality as well as any year-on-year changes in transmission dynamics. As this simple model has no explicit mechanism for cross-reactive immunity, the reproduction rate is also modified by a limiting term that reduces the acquisition of new infections for individuals with one or more existing infections, such that the mean number of concurrent inoculations $\langle C_{inf} \rangle = N_{infections}(t)/N_{humans}$ is not allowed to exceed five:

$$R_{rep} = R_0(t) R_{exp} \left(1 - \frac{\langle C_{inf} \rangle}{5}\right)$$

The primary motivation for restricting $\langle C_{inf} \rangle$ to five was to limit the computational time required for parameter sweeps including simulations with very high transmission intensity, while maintaining an accurate representation of the dynamics in Thiès, where a maximum of around 50% polygenomic infections were observed.

Seasonal forcing is introduced with the following periodic functional form where n is an even positive integer and ρ is the ratio of the minimum to maximum reproduction number:

$$R_0(t) = R_{0,max}(t) \left[\rho + (1-\rho) \cos^n \frac{\pi t}{365}\right]$$

Longer timescale variation in transmission intensity is intended to capture both the annual variation in climate as well as the effects of increasing (or waning) effectiveness of malaria control strategies, e.g. ITN usage and ACT treatment. Specifically to quantify the significance of patterns observed in Thiès, we parameterized the annual variation by the following piecewise function with three transmission intensities and three inflection points:

$$R_{0,max}(t) = \begin{cases} R_0^a, & t < t_1 \\ R_0^a + (R_0^b - R_0^a)\dfrac{t - t_1}{t_2 - t_1}, & t_1 < t < t_2 \\ R_0^b, & t_2 < t < t_3 \\ R_0^c, & t > t_3 \end{cases}$$

Note that the model parameters $- R_0^a, R_0^b, R_0^c -$ are the reproduction rates at the peak transmission season. As an example, for $R_{0,max} = 2.0$ and $\rho = 0.18$, the minimum reproduction rate is significantly below replacement at 0.36.

In a later section, we will discuss how this piecewise function relates to an alternative strategy of fitting independently varying transmission intensities in each measured year (2006–2013).

The probability P_{out} that a next-generation parasite infection is generated from the outcrossing of at least two existing strains is related to the fraction of humans experiencing multiple inoculations during the infectious period of a single infection. Under the simplification that all humans experience the same risk of inoculation, the fraction of multiply infected individuals can be approximated by a Poisson distribution.

$$P_{out} = \sum_{k \geq 2} \text{Poisson}\left(k, \lambda = \langle C_{inf} \rangle\right)$$

It follows that the rates of outcrossing and clonal propagation are:

$$R_{out} = R_{rep}\, P_{out}$$
$$R_{clone} = R_{rep}\, (1 - P_{out})$$

Initialization. Simulations are initialized according to the specified model input parameters, in particular the number of independent parasite populations to model, the total number of human hosts, the initial number of parasite infections, and the number of unique random barcodes from which to draw those initial infections. Because the evolution of the above-mentioned dynamical processes is a stochastic process with many random-number draws, we also initialize the random seed of the pseudo-random number generator. To accurately address this stochastic variation in the model-calibration procedure described in a following section, each point in model-parameter space was simulated for twenty different random seeds.

Sampling of model state. To compare model simulations to real-world barcode observations, several values are tracked and outputted by the model. The number of infections, unique barcodes, and polygenomic infections are tracked daily. In addition, the barcodes of all infections are recorded in an annual census at the point of peak transmission intensity. For comparison to the annual observations from Thiès, we sample randomly without replacement the same number of barcodes from the simulated output as were analyzed from the clinic in the years from 2006 to 2013.

Calibration of barcode model to data. The parameterization of the model to the data collected in Thiès, Senegal from 2006 to 2013 required two components: establishing a distance measure to compare realizations of the stochastic model to the data; and determining the regions of parameter space that fit the data. For the latter, a computationally efficient method for searching a multi-dimensional parameter space is the Incremental Mixture Importance Sampling (IMIS) algorithm (Ref. 8). Previously, the method has been employed to parameterize the UNAIDS model of HIV (Ref. 9). The method is fundamentally Bayesian and can be generalized by the following equations:

$$Posterior \propto Likelihood \times Prior,$$

$$P(Model\,|Data) = \frac{P(Data|Model)P(Model)}{P(Data)}$$

The IMIS algorithm is an iterative method that finds high-likelihood regions of parameter space, by selecting the next set of sample points based on the likelihood evaluations of previous iterations. The iterative nature of the algorithm allows for efficient convergence to local maxima. A stopping criterion is defined in Ref. 8, indicating the regions near the maxima have been well sampled and characterized.

For the execution of the IMIS algorithm to fit the malaria barcode model to the data, we define the prior distribution for each of the parameters to be uniform. Six parameters are allowed to vary from the model: the expiration rate, the importation rate, the size of the human host population, the initial R_0, middle R_0, and final values of R_0. The respective domains of these parameters are allowed to vary in the range [0.005, 0.03], [0, 0.05], and [750, 1500] to [1, 3], [1, 3], and [1, 3]. Twelve hundred initial points in parameter space were randomly chosen using Latin hypercube sampling (LHS) from the bounded domains.

For each selected point in parameter space, twenty realizations of the stochastic model were obtained for the ensemble. The subsequent iterations of the algorithm consist of twenty points in parameter space, each with twenty stochastic realizations of the model. For a more detailed account of the algorithm, see Ref. 8.

The second important calibration component was the definition of a distance metric, based on the following set of lower-dimensional summary statistics:

a. The fraction of monogenomic infections with repeated barcodes, by year;

b. The fraction of monogenomic infections with unique barcodes, by year;

c. The number of repeated barcodes that are repeated exactly twice in a given year, by year;

d. The number of repeated barcodes that are repeated more than twice in a given year, by year;

e. The number of repeated barcodes that persist exactly two, three, four, and greater than four years (allowing for missing years within the interval);

f. The number of repeated barcodes that appear and persist for two or more years in the observation time;

g. The number of repeated barcodes that disappear after persisting for at least two years.

For each of these features, an individual measure of deviation was calculated as the sum of squared differences normalized to the estimated variance. Variance in the simulated data was estimated from multiple stochastic realizations, while uncertainties in the actual data were calculated from binomial statistics assuming different years constituted independent measurements.

As an example, for feature *a* in the above list, the data contain both the mean fraction of polygenomic barcodes and the sample size, thus allowing for a binomial variance (assuming different years are independent measurements) to be computed. For each point in parameter space, twenty realizations were obtained from the model to form an ensemble, providing statistics for the model, *i.e.* mean and variation statistics. The deviation D_j is calculated as the sum of squared differences normalized to the estimated variance for each year:

$$D_j = \sum_i^8 \frac{\left(\mu_{i,sim} - \mu_{i,data}\right)^2}{\sigma_{i,sim}^2 + \sigma_{i,data}^2}$$

where *j* indicates the feature, *i* is an index for the year, μ is the mean, σ^2 is the variance, *sim* indicates the simulations, and *data* indicates the data. Each of the component deviations is constructed similarly except features 5 through 7. These are not indexed by year, but are a single count (features 6 and 7) or four independent numbers (feature 5). From the deviations calculated based on these various features, we construct the pseudo-likelihood as follows:

$$L = \prod_{j=1}^{7} e^{-D_j}$$

Fig. S8 shows the result of the 2400 Latin hypercube sampling (LHS) parameterizations drawn from the three-dimensional parameter space spanning each of the R_0 directions. R_0^a is the value of R0 at the beginning of the multiyear period, R_0^b is the value reached after a linear decrease between 2006 and 2010, and R_0^c is the value from 2012 onwards. The human population, expiration rate, and importation rate were set according to the maximum likelihood point found by the six-dimensional IMIS calibration (1350, 0.022, and 0.01 respectively); see Fig. 4 for a visualization of the output of the IMIS algorithm. Essentially all of the posterior probability distribution requires a significant drop in transmission intensity over the years from 2006 to 2013. Even the most likely scenarios with unchanged transmission intensity are about three orders of magnitude less likely than the maximum-likelihood parameterizations.

Regarding the other dimensions of the parameter-space optimization, the posterior probability space favored an expiration rate with a characteristic generation time between 30 and 100 days – in the expected range from an understanding of the parasite lifecycle – and the results were largely insensitive to the values of importation rate and human host population over the ranges explored. Note that earlier exploratory work had shown that fewer than about 300 humans resulted in a significant chance of stochastic fadeout during the dry season, requiring parameters skewed to higher reproduction rates and longer generation times.

Validation of piecewise R_0 parameterization. The rationale for a piecewise parameterization of $R_{0,max}\,(t)$ was two-fold. First, it allowed a simple framework within which to address the question of whether there was a significant decrease between the first two time intervals and/or a significant increase between the latter two. Second, it minimized the dimensionality of the automated iterative fitting behavior described above, which allowed the concurrent exploration of generation time, population size, and importation rate interactions.

Expanding this approach to characterize the transmission-intensity time series implied by the data, we performed an iterative procedure to resample $R_{0,max}$ independently for each year from 2006 to 2013. Starting from the piecewise profile in Figure 4, subsequent

iterations consisted of selecting a year, running a suite of simulations that perturbed the baseline $R_{0,max}$ in that year, fitting the minimum chi-squared over the range of perturbations, and updating the fitted means and uncertainties as the baseline for the next iteration.

After 30 iterations, the resampling has converged to the distributions shown in Fig. S9. The result is broadly consistent with the piecewise function used in the 6-dimensional fitting that included three free R_0 parameters. The main difference from allowing eight degrees of freedom in the $R_0(t)$ fit is that simulations can match the steeper drop in transmission intensity that is suggested by the data between 2006 and 2007.

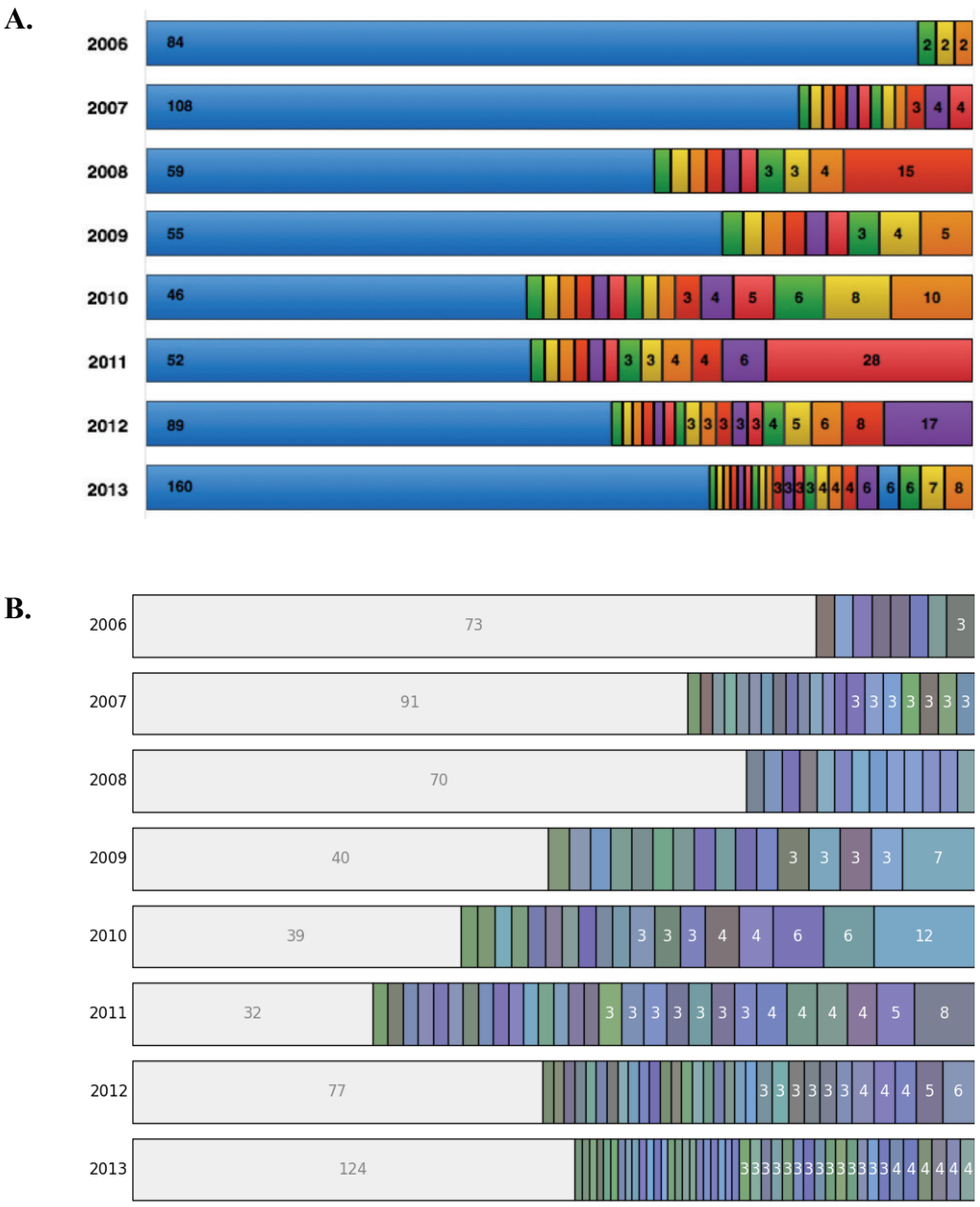


Fig. S1. Unique and repeat barcodes. **A.** Observed number of unique barcodes (blue) and of barcodes repeated in subsets (colors) for 2006–2013 in Thiès, Senegal. Data for 2006–2011 also in Ref. 5. **B.** Model output of unique (gray) versus repeated (color) barcodes for 2006–2013 using maximum likelihood estimates of free parameters in the model.

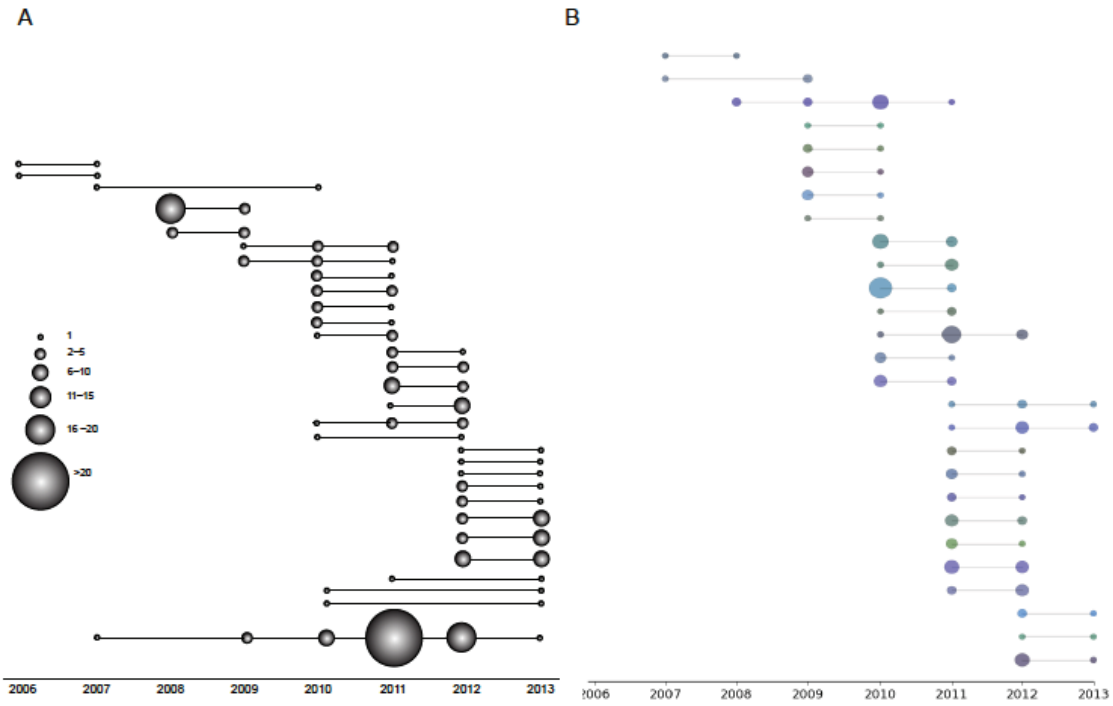


Fig. S2. Persistence of repeated barcodes across years 2006–2013 in Thiès, Senegal. **A.** Observed patterns of persistence. Size of sphere keyed to number of isolates observed. Data for 2006–2011 also in Ref. 5. **B.** Model output of patterns of persistence for 2006–2013 using maximum likelihood estimates of free parameters in the model.

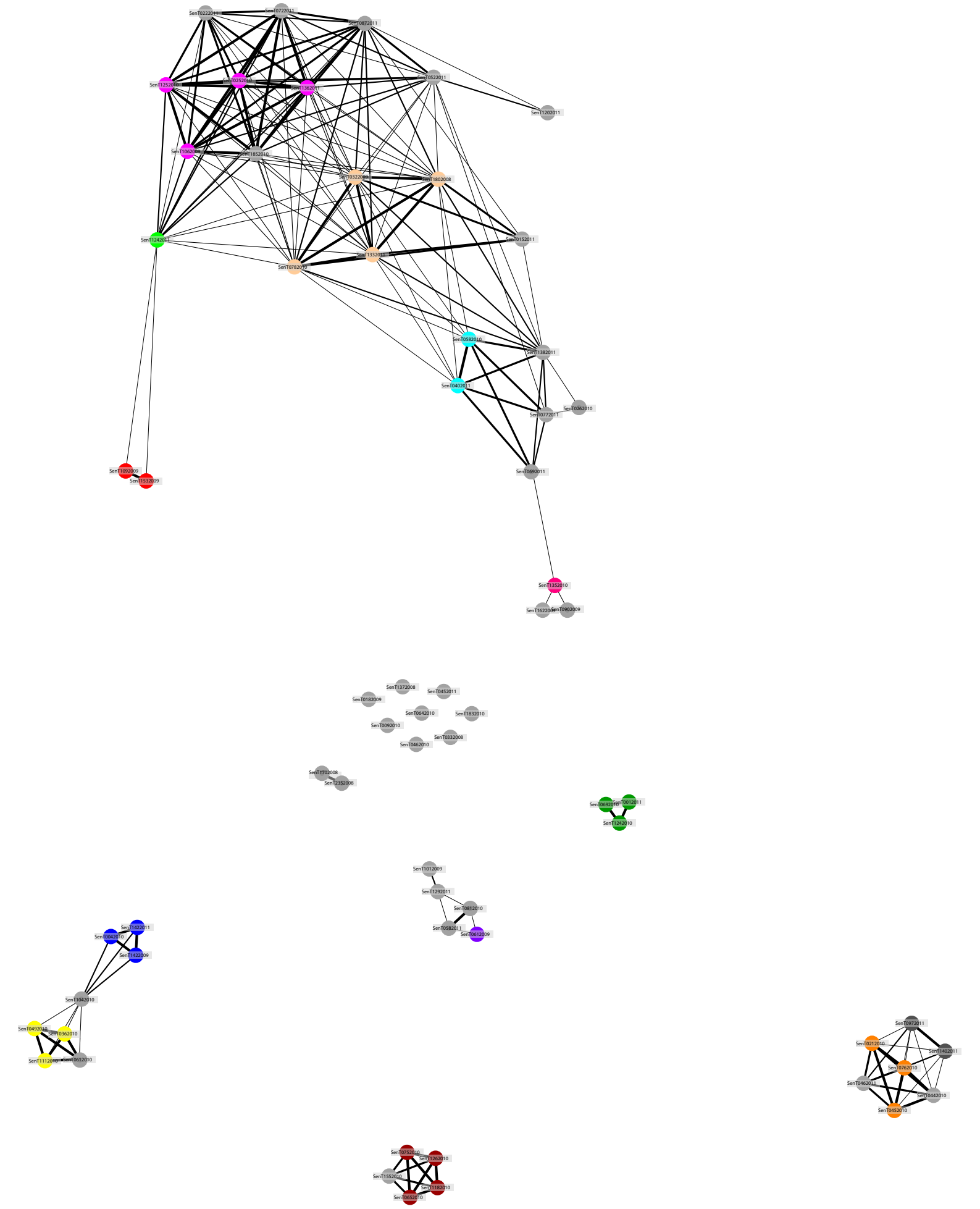


Fig. S3. Samples included in the networks in Fig. 1A.

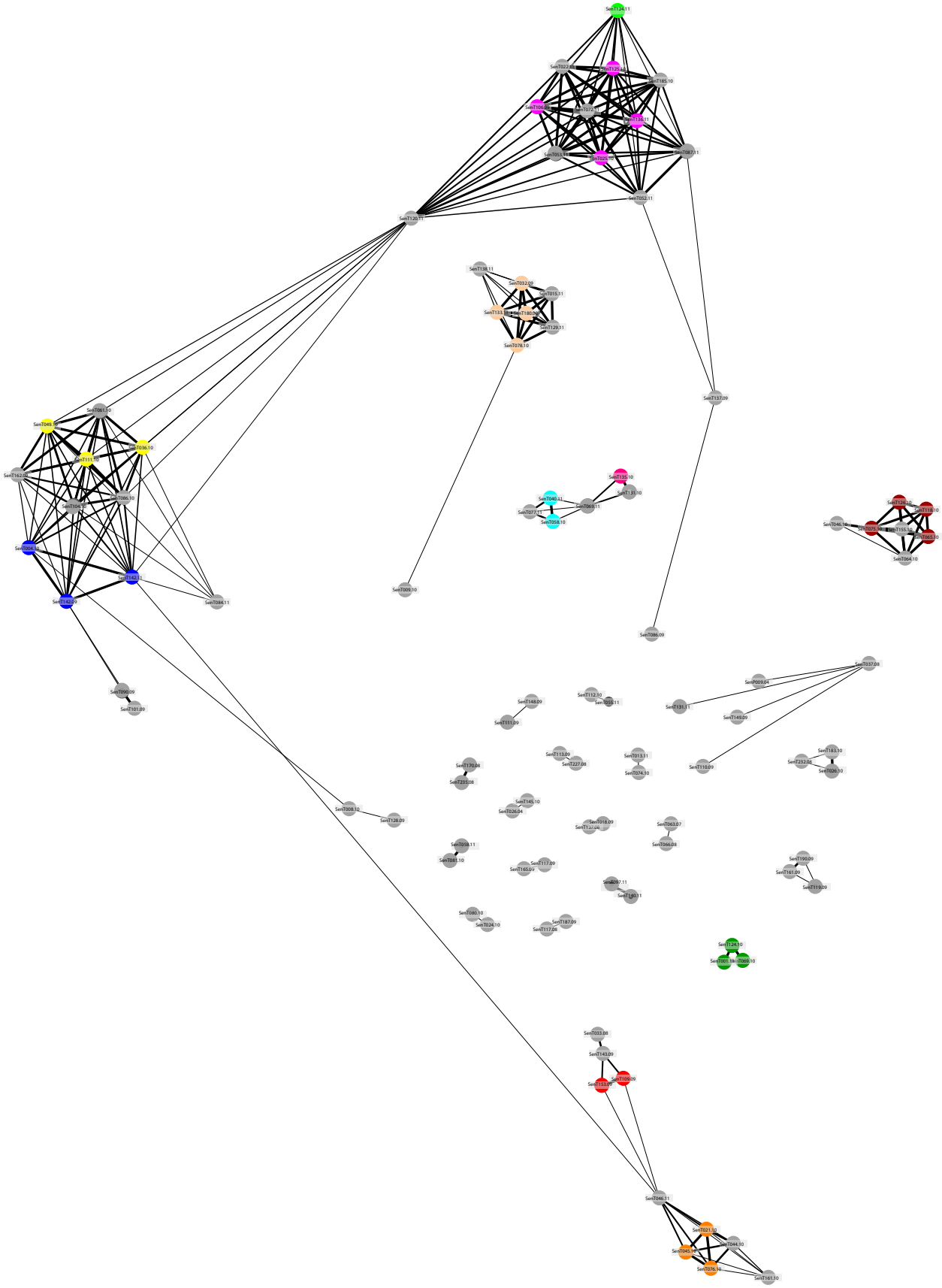


Fig. S4. Samples included in the networks in Fig. 1B.

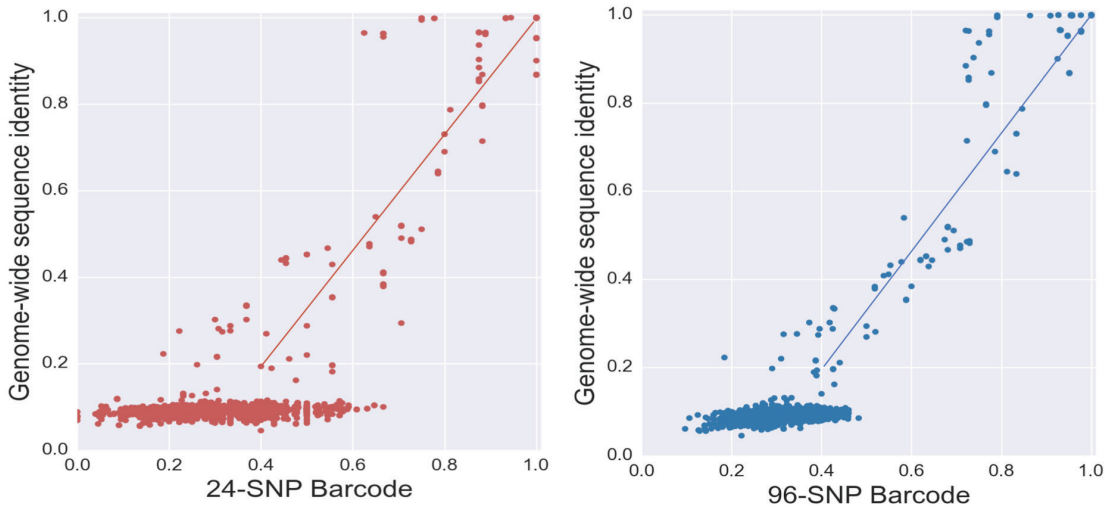


Fig. S5. Comparison of genome-wide sequence identity versus barcode identity for barcodes optimized for 164 fully sequenced samples from Thiès, Senegal. **A.** 24-SNP barcode. **B.** 96-SNP barcode. SNPs in the optimal barcodes are chosen based on SNPs with the highest frequency of the minor allele.

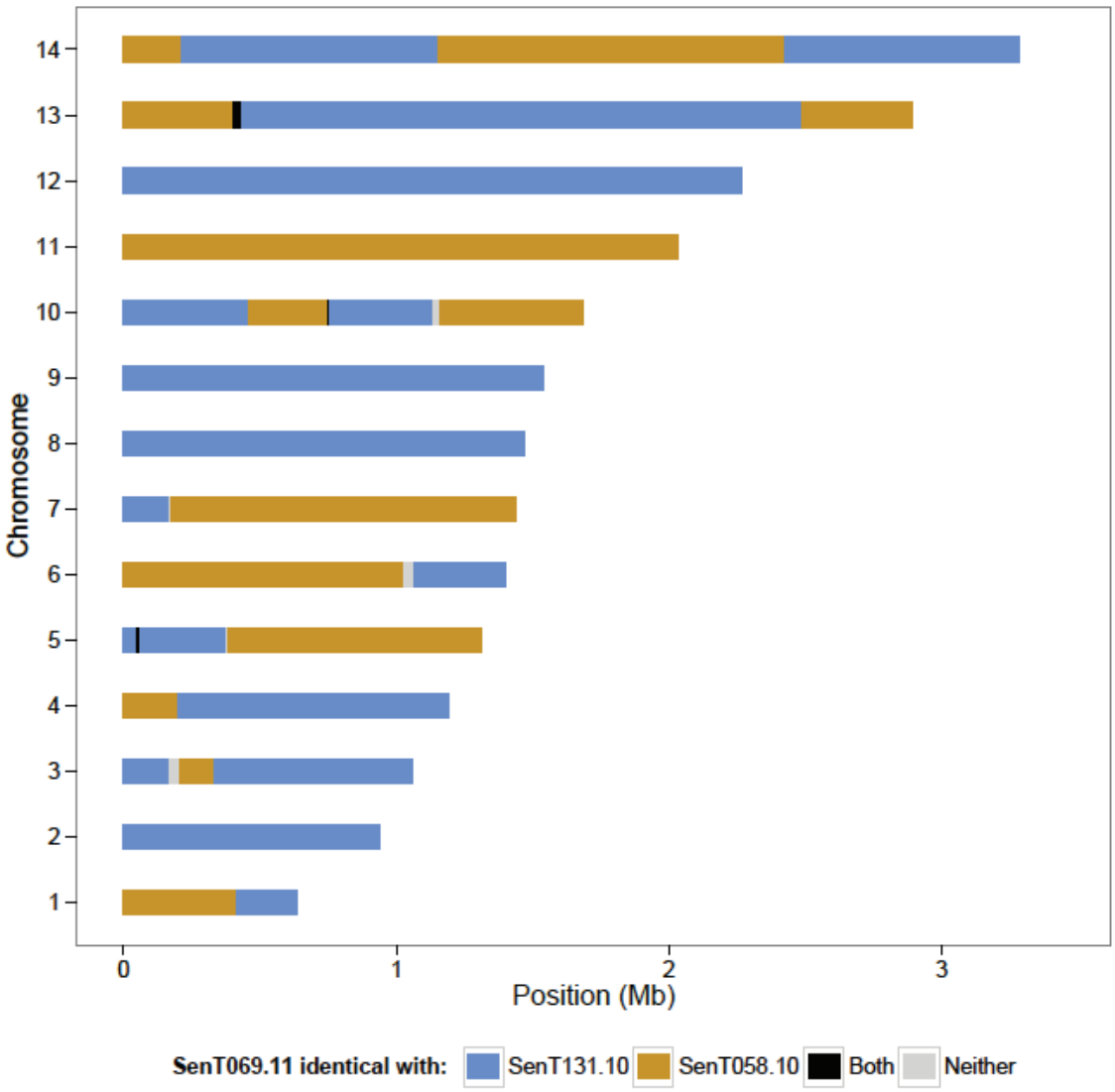


Fig. S6. Parental origin of genomic segments for sample SENT069.11. Colored segments show identity by descent (as assigned by a hidden Markov model) to the two parental types (blue and orange), to both types (black), or to neither (gray).

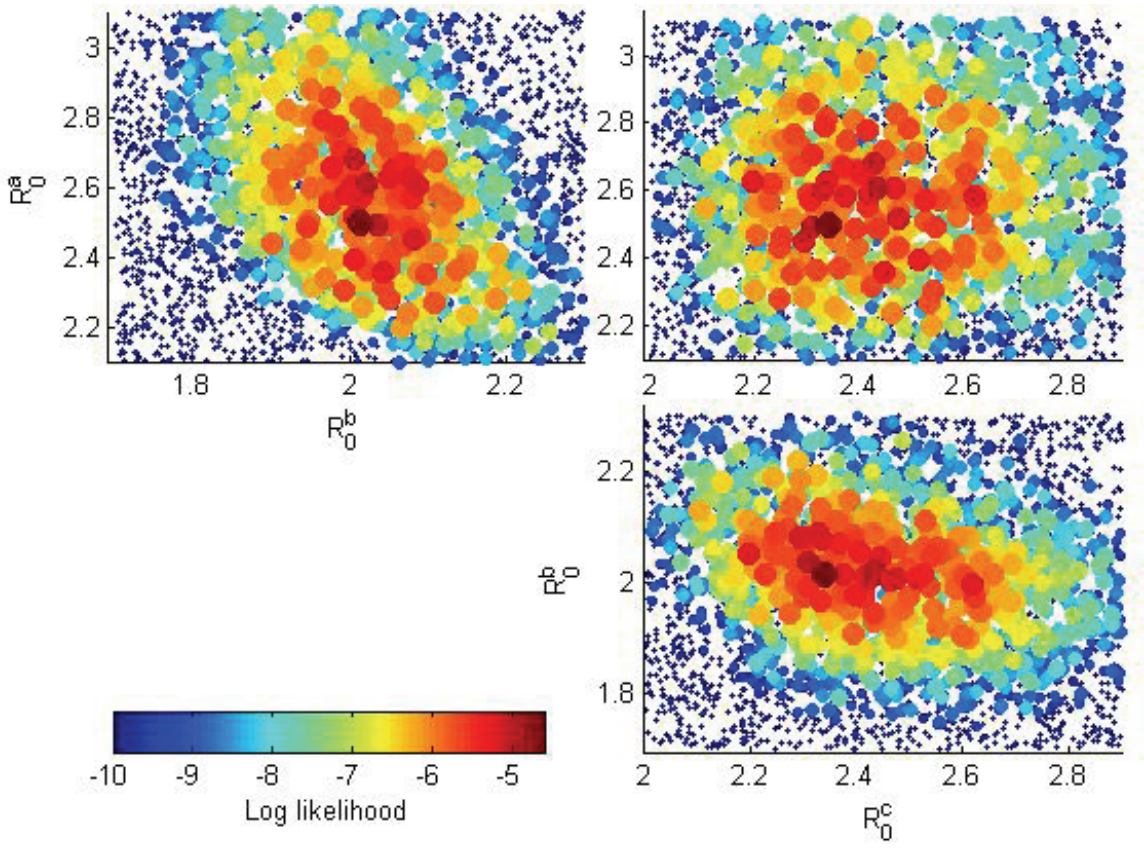


Fig. S7. The likelihood evaluated at 2400 points drawn from a three-dimensional parameter space using Latin hypercube sampling. The three parameters allowed to vary are the three reproductive numbers (R_0) corresponding to initial maximum rate of increase (R_0^a) decreasing linearly to a minimum (R_0^b) and after rebound (R_0^c). The human population, expiration rate, and importation rate were set according to the maximum likelihood point found by the six-dimensional IMIS calibration (1350, 0.022, and 0.01 respectively). The colors of the points indicate the magnitude of likelihood, as indicated by the color bar. The three plots illustrate each of the two-dimensional projections of the three-dimensional parameter points.

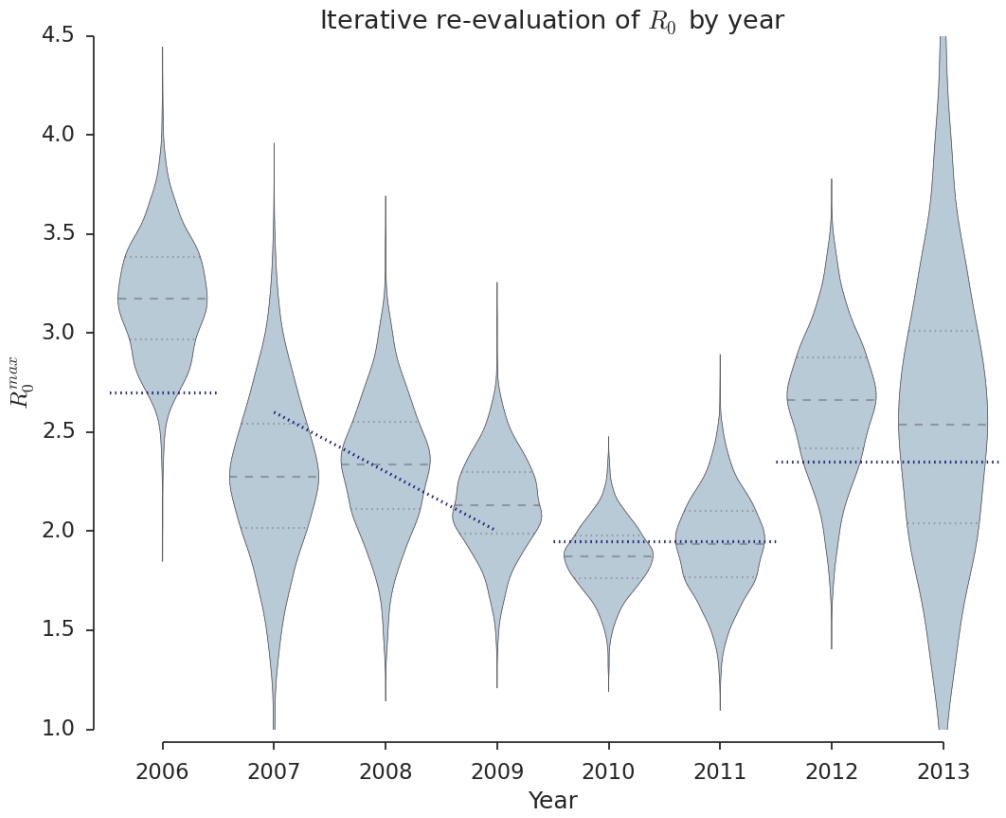


Fig. S8. The distributions of annual maximum reproduction rate fitted by iterative resampling of rates in each of the eight years. The global best-fit parameters are represented by the central gray dashed lines, while the credible intervals around the best fit ($\chi^2 < \chi^2_{min} + 1$) are shown as gray dotted lines. For comparison, the piecewise function from Fig. 4 is shown as blue dotted lines.

Table S1. Variance effective population size estimated from barcode alleles of parasites sampled from Thiès, Senegal, 2006–2013

	Likelihood estimates†	
	MLE ‡	95% Confidence interval
2006–2007	402	(46, infinity)
2007–2008	18	(9, 36)
2008–2009	32	(14, 106)
2009–2010	17	(8, 38)
2010–2011	10	(6, 18)
2011–2012	20	(10, 40)
2012–2013	38	(17, 89)

† Methods as well as data for 2006–2011 from Ref. 5. These estimates count each sample as independent irrespective of its barcode. An alternative estimate can be obtained by counting each barcode subset only once in each year. The latter estimates are larger than those in the table by a factor of about 10.

‡ Maximum likelihood estimate

Table S2. Barcode model parameters

Variable	Description	Value
	Duration of simulation in days	5475
	Initial seed of pseudo-random number generator	6
R_0^a	Maximum R0 before reduction year	2.7
R_0^b	Maximum R0 between plateau and rebound years	1.95
R_0^c	Maximum R0 after rebound year	2.35
t1	Initial year of linear drop in transmission	7
t2	Year of plateau after transmission reduction	10
t3	Initial year of rebounded transmission	12
	Heterogeneity in parasite propagation	CONSTANT
R_{exp}	Daily rate of infection expiring	0.022
R_{imp}	Daily rate of importing random barcodes	0.01
ρ	Fraction of maximum transmission at seasonal minimum	0.18
n	Power of seasonality oscillation	2
α	Probability of losing all mixed-allele positions per generation	0.2
β	Mean reduction in mixed-allele sites per generation	0.5
	Number of initial populations	1
	Number of parasite barcodes at initialization	1200
	Number of unique random barcodes at initialization	1000
	Number of humans corresponding to simulated barcodes	1350

Table S3. Analysis of malaria incidence per person in Thiès and in all Senegal excluding Thiès, 2006–2013.

Parameter estimates for the model: Relative incidence per person = $a \times \mathrm{Exp}(-bx) + cx$, using incidence from Thiès, Senegal, 2006–2013, normalized to the incidence per person in 2006 (when incidence = 0.114386 per person).

	Estimate	Standard Error	t Statistic	P-Value
a	3.24585	1.48588	2.18446	0.0806577
b	1.2014	0.404983	2.96655	0.0312813
c	0.0427244	0.00968897	4.40959	0.00695921

Parameter estimates for the model: Relative incidence per person = $a \times \mathrm{Exp}(-bx) + cx$, using incidence from all Senegal excluding Thiès, 2006–2013, normalized to the incidence per person in 2006 (when incidence = 0.132497 per person).

	Estimate	Standard Error	t Statistic	P-Value
a	2.1068	0.56716	3.71465	0.0137883
b	0.714727	0.19928	3.58654	0.0157655
c	0.01978	0.010895	1.81551	0.129153

Dataset 1. Barcode data from Thiès, Senegal, 2006–2013. Accompanying file Dataset 1_Senegal.csv. These data have also been deposited in EuPathDB (PlasmoDB.org).

Dataset 2. Barcode data from Malawi, 2009–2010. Accompanying file Dataset 2_Malawi.csv. These data have also been deposited in EuPathDB (PlasmoDB.org).

Dataset 3. Genome sequence accession numbers. Accompanying file Dataset 3_Pfalciparum_SRA_SenegalMalawi.csv.