Supplementary Methods: HbS allele-frequency compilation (district-wise surveys, Uganda)

Scope and inputs

We used of two district level datasets to estimate the HbS allele frequency in Uganda. The analysis unit was the administrative district. All computations were performed on district totals.

Data sources

Kiyaga dataset. District rows with genotype counts for HbAA, HbAS, and HbSS, and near uniform denominators, typically 100 individuals per district.

Ndeezi dataset. District rows with total tested, percent with sickle cell trait, and percent with sickle cell disease. Denominators varied by district.

Combined analytic file. A merged table produced from the two sources after harmonization of district names. Genotype and allele counts were summed by district before computing allele frequencies.

Kiyaga covered 122 districts with 12,009 individuals. Ndeezi covered 112 districts with 97,631 individuals. The combined file covered 128 districts with 109,640 individuals.

Eligibility and unit of analysis

All districts present in either source were eligible. Each row represents one district. If a district appeared in only one source, it was retained with available fields. Region labels were retained when present.

Data extraction

For Kiyaga, genotype counts (n_{AA}, n_{AS}, n_{SS}) and denominators (N) were read directly.

For Ndeezi, the reported trait percent and disease percent were converted to counts using the district total:

$$n_{AS} = \text{round}\left(\frac{\%\text{Trait}}{100} \times N\right), \quad n_{SS} = \text{round}\left(\frac{\%\text{Disease}}{100} \times N\right),$$

then

$$n_{AA} = N - n_{AS} - n_{SS}.$$

Nearest integer rounding was used. If rounding produced $n_{AA} < 0$, the pair (n_{AS}, n_{SS}) was adjusted by one count toward the smaller of the two to enforce $n_{AA} \geq 0$ and $n_{AA} + n_{AS} + n_{SS} = N$.

District name harmonization

District names were lowercased, stripped of whitespace artifacts and hyphens, and matched on a canonical key that removed internal spacing. Known split spellings were mapped to a single form, for example “Bukomans imbi” to “Bukomansimbi” and “Nakapiripi riti” to “Nakapiripirit”. The merge used a one-to-one key. Records without a match were carried forward from their source.

Variable definitions

For each district:

Denominator: $N = n_{AA} + n_{AS} + n_{SS}$

Total alleles: $2N$

S-alleles: $2n_{SS} + n_{AS}$

S-allele frequency:

$$q = \frac{2n_{SS} + n_{AS}}{2N}, \qquad p = 1 - q.$$

For the combined file, genotype counts from the two sources were summed within district. Allele counts and q were recomputed from the summed genotypes, not averaged from source-specific frequencies.

Aggregation and weighting

National summaries were computed by summing genotype counts across districts, computing total and S-allele counts, then evaluating q from these totals. This avoids bias from averaging district frequencies with unequal denominators. Where regional labels were available, person weighted summaries were computed within region:

$$q_{\text{region}} = \frac{\sum_{d \in \text{region}} (2n_{SS,d} + n_{AS,d})}{\sum_{d \in \text{region}} 2N_d}.$$

Unweighted summaries, such as the district median and interquartile range of q, were reported to describe dispersion across districts.

Quality control

Deterministic checks were applied to every district:

$$n_{AA} + n_{AS} + n_{SS} = N,$$

Total alleles $= 2N$

S-alleles $= 2n_{SS} + n_{AS}$

$$0 \leq q \leq 1$$

If a check failed, the dependent quantity was recomputed from the independent components. Plausibility was assessed by comparing observed HbAS and HbSS with Hardy–Weinberg expectations $2pq$ and q^2. No districts were excluded based on this screen. Districts with very small N were flagged for cautious interpretation.

Handling of missing data

If a district lacked a region label, the record was retained with region missing. No imputation of genotype counts or frequencies was performed. Districts with $N = 0$ after harmonization were excluded from pooled estimates.

All counts, allele derivations, merges, and summaries were scripted and are reproducible from the three input files described above.

Finding

The Kiyaga dataset contributed 122 districts with 12,009 individuals. The Ndeezi dataset contributed 112 districts with 97,631 individuals. The harmonized combined file covered 128 districts with a total of 109,640 individuals after name matching and aggregation.

In the combined dataset, the national S-allele frequency (q) was 0.07237, calculated from 15,870 S alleles out of 219,280 total alleles. The aggregated genotype distribution was 13.04 percent HbAS and 0.72 percent HbSS, corresponding to about 14,302 individuals with trait and 784 with disease. The district distribution of (q) showed a median of 0.07454, an interquartile range of 0.05169, a minimum of 0.005, and a maximum of 0.13960. These figures describe a broad geographic spread in allele frequency, with many districts clustering between 0.05 and 0.09 and a smaller group extending into the low teens.

Each source summarized on its own produced a consistent national picture. In the Kiyaga file, (q) was 0.0605 with 11.03 percent HbAS and 0.53 percent HbSS, based on near-uniform denominators of about 100 per district. In the Ndeezi file, (q) was 0.0738 with 13.29 percent HbAS and 0.74 percent HbSS, based on variable denominators. The combined file yields national estimates that reflect the larger totals from Ndeezi while retaining the wider geographic coverage of Kiyaga.

The genotype and allele arithmetic were coherent at district level. For each district, S alleles equaled 2×HbSS plus HbAS, and total alleles equaled two times the number tested. At the national level, (q) near 0.07 implies (2pq) near 13 percent and (q^2) near 0.5 percent under Hardy–Weinberg equilibrium. The observed HbAS tracks this expectation closely. HbSS sits modestly above (q^2), which is compatible with rounding in district tables and local departures from equilibrium in smaller strata.

Kiyaga (genotype based, near-uniform N)

- Overall S-allele frequency qqq from aggregated counts: 0.0605
- National HbAS prevalence: 11.03%
- National HbSS prevalence: 0.53%
- District S-allele frequency distribution: median 0.061, IQR 0.055, range 0.000 to 0.152

Ndeezi (percent based, variable N)

- Overall S-allele frequency qqq from derived counts: 0.0738
- National HbAS prevalence: 13.29%

- National HbSS prevalence: 0.74%

Combined (summed counts from both sources)

- Overall S-allele frequency qqq: 0.07237
- National HbAS prevalence: 13.04%
- National HbSS prevalence: 0.72%
- District S-allele frequency distribution: median 0.07454, IQR 0.05169, range 0.005 to 0.13960

Supplementary Methods: Empirical Bayesian Kriging of HbS allele frequency

Objective and input

We produced a continuous surface of district-level HbS allele frequency using Empirical Bayesian Kriging (EBK) in ArcGIS Pro, Geostatistical Analyst. The input was the district point table from the preceding section with fields for coordinates in UTM 36N, the pooled district estimate (p) in the open unit interval, and quality flags. The national boundary defined the processing extent and mask.

Geoprocessing environment

All EBK runs used the following environment settings to keep predictions comparable across iterations: output coordinate system EPSG 32636, processing extent set to Uganda ADM0, mask set to ADM0, and a 2,000 m cell size for the prediction raster. A snap raster was set to the first accepted EBK product to ensure subsequent runs landed on the same grid.

EBK workflow

EBK samples semivariogram parameters from a prior, generates multiple local semivariogram realizations, and averages the resulting local predictions and variances. We used the ArcGIS Pro implementation with local subsets to handle spatial heterogeneity in sampling density. Subdivision and simulations. Subset size 100, overlap factor 1.5, simulations per subset 100. These values stabilize the semivariogram estimation without inflating runtime at national extent. Semivariogram family. K-Bessel in the primary run, Power in a sensitivity run. K-Bessel handled moderate curvature at short lags and provided well-calibrated uncertainty on cross validation. Neighborhood. Standard neighborhood with 8 sectors, minimum neighbors 12, maximum

neighbors 32, fixed search radius 60 km, azimuth 0. Sectoring reduces directional bias in clustered areas and avoids overfitting to a single sector of dense observations.
Coincident points. When the harmonized file retained coincident coordinates, values were averaged before kriging to prevent duplicated support.
Trend. No global trend term was specified because district means already absorb slow spatial drift at the administrative scale. Residual maps were reviewed to confirm no large-scale bias.

Predictions and uncertainty

The EBK tool returned a prediction raster $\hat{p}(x)$ and a prediction standard error raster $\mathrm{SE}(x)$ on the 2,000 m grid. Both layers were clipped to ADM0 and carried forward. For downstream algebra with PfPR on the 5 km analysis grid, we aligned the EBK prediction to the PfPR template using bilinear resampling after projection checks and then clamped to to eliminate small interpolation overshoots near boundaries.

Cross-validation diagnostics and acceptance criteria

We used the EBK Cross Validation report to assess fit and calibration. Acceptance targets were: mean error near zero, root mean square error minimized and consistent with the prediction standard errors, root mean square standardized error near one, and standardized residuals that are approximately normal. The final run met these criteria, with mean error effectively zero, RMSE about 0.03, RMS standardized near 0.95, and prediction interval coverage near nominal at 90 and 95 percent. The observed versus predicted scatter lay close to the 1:1 line with mild shrinkage at extremes, consistent with EBK regularization. These diagnostics support use of the surface in national planning.

Sensitivity and robustness checks

We ran two focused checks. First, the semivariogram family was switched from K-Bessel to Power while holding all other settings constant. Summary error metrics and maps changed only modestly, and the K-Bessel model slightly improved uncertainty calibration. Second, the neighborhood parameters were perturbed within reasonable bounds, reducing minimum neighbors to 8 in sparse regions and tightening maximum neighbors to 24 in dense clusters. Differences were minor at national scale. We retained the baseline neighborhood for consistency.

Edge handling and masks

All resampling occurred before masking to avoid ringing at the coastline and border. Predictions outside ADM0 were removed before any statistics were computed. For map export and cartography, the 2,000 m raster was used directly. For algebra with PfPR, the aligned 5 km version was used.

Supplementary Methods: Greedy site selection for maximal coverage

Objective and baseline

We formalized siting as a maximal covering location problem on a raster. The baseline to be covered was the non-negative co-risk surface on the 5 km analysis grid,

$$B_0(x) = \mathrm{HbS}(x) \times \mathrm{PfPR}(x),$$

projected to EPSG:32636 and masked to ADM0. Where an existing network was available, an unserved baseline B_0^{unserved} was created by setting to NA all pixels within the chosen service radius around existing facilities, so that demand already inside service areas did not contribute to the objective.

Coverage set and spacing

Coverage is defined as the set of pixels within a fixed service radius r of a selected site, intersected with the national mask. The primary analysis used $r = 25$ km. A minimum spacing rule can be imposed by rejecting a candidate pick that falls within a fixed distance $d_{\min}$ of any previously selected site. In the distance variant used here, $d_{\min} = r$ so spacing is implicit in the coverage mask. In the travel-time variant, coverage sets are derived from a friction surface by cost-distance with an isochrone threshold (for example 60 minutes), and spacing is enforced in minutes.

Greedy heuristic

We used a greedy algorithm that iteratively selects the next site at the pixel with the largest remaining baseline value, then zeros out its covered neighborhood before the next iteration. Let $R^{(0)}(x) = B_0(x)$ denote the residual surface at iteration 0.

For $i = 1, \ldots, K$:

- Locate the pixel

$$x_i^\star = \operatorname{argmax}_x R^{(i-1)}(x).$$

If all values are NA or non-positive, stop.

- If spacing is active and $x_i^\star$ falls within $d_{\min}$ of any previously selected site, set $R^{(i-1)}(x_i^\star) = \text{NA}$ and return to step 1.
- Place a site at $x_i^\star$. Form the coverage mask

$$M_i = \{x\text{:dist}(x, x_i^\star) \le r\} \cap \text{ADM0}.$$

- Record the marginal capture

$$c_i = \sum_{x \in M_i} R^{(i-1)}(x).$$

- Update the residual surface by setting

$$R^{(i)}(x) = \begin{cases} R^{(i-1)}(x), & x \notin M_i \\ \text{NA}, & x \in M_i \end{cases}.$$

The cumulative coverage curve after k sites is

$$C_k = \frac{\sum_{j=1}^{k} c_j}{\sum_x B_0(x)}.$$

We report C_k as a percent and extract milestone counts at 60%, 70%, and 80% capture.

Implementation details

Geometry and units. All rasters were aligned to the 5 km PfPR template, in EPSG:32636. Masks were applied after resampling to avoid edge artifacts. Co-risk values were clamped to to remove interpolation overshoot. Tie handling and reproducibility. Ties in step 1 were broken by the first occurrence in a fixed scan order. A constant random seed was set only if a random tie-breaker was requested, which was not needed in the final runs.

Stopping rule. Runs stopped either when K sites were placed or when $c_i = 0$ indicating that all remaining pixels were NA or zero.

Outputs. Ranked site points in the analysis CRS and in WGS84 for reporting, circular service

areas intersected with ADM0, the residual raster $R^{(K)}$, a coverage curve table with $\{k, C_k\}$, and a site list with marginal and cumulative capture $\{c_i, \sum_{j \le i} c_j\}$.

Variants

Unserved baseline. Replace B_0 with B_0^{unserved} by masking pixels already covered by the existing network at radius r. This focuses placement on gaps and reduces apparent returns when coverage is already dense.

Travel-time coverage. Replace Euclidean buffers with an isochrone mask from a motorized friction surface. The rest of the loop is unchanged. This substitution improves realism in low-access terrain and along corridors.

Soft reduction. Instead of removing the covered burden, multiply it by a factor $\alpha \in [0,1)$ to represent partial relief (for example, $\alpha = 0.2$). This is useful when the program intends to complement rather than fully cover a need from a single site.

Rationale and properties

On non-negative rasters with a union-of-neighborhoods objective, the greedy rule is a standard approximation to the maximal covering location problem. It is fast on national grids, easy to audit, and tends to match the qualitative structure of optimal solutions while providing complete rank orderings for programmatic staging.

Supplementary Methods: Health facility catchment area estimation

Health facility catchment area estimation

Scope and inputs

We delineated service areas for the existing facility network to quantify current reach, define the unserved baseline for siting, and collocate proposed points to named facilities. Inputs were a point feature class of facilities with stable identifiers and attributes, a national motorized friction raster, and the Uganda ADM0 boundary. All layers were projected to WGS 84 UTM Zone 36N (EPSG 32636). Catchment geometry was produced on the 5 km analysis grid used elsewhere so coverage and co-risk accounting are consistent.

Friction calibration and preparation

The friction grid was validated for units and support before any cost accumulation.

Units and range. Values were coerced to minutes per meter. If the raster arrived in minutes per

kilometer, we divided by 1000. Any zeros or negative cells, which would yield nonphysical free travel, were replaced with the 5th percentile of positive values. Extremely high values were winsorized at the 99.5th percentile to prevent isolated spikes from dominating isochrone shapes.
Hydrology and barriers. Permanent water bodies and protected non-traversable areas were rasterized to the friction grid and assigned a large finite cost that reflects realistic detours rather than infinite cost which creates processing holes. Ferries and major bridges, where present, were assigned corridor-specific costs.
Alignment. The friction raster was projected to EPSG 32636, resampled by bilinear interpolation to the 5 km template, then masked to ADM0 after resampling. This order avoids edge ringing and guarantees identical pixel footprints with the co-risk surface.
Sanity checks. We computed the median and interquartile range of the friction values and mapped a 5 percent sample of least-cost paths between random pairs of district centroids to confirm plausible travel-time gradients.

Cost distance and isochrone construction

For each facility point we computed a cumulative cost surface in minutes using the calibrated friction raster. Isochrones were extracted by thresholding the cumulative surface at 60 minutes and vectorizing the result.
Processing details:
Neighborhood and boundaries. Cost accumulation used an 8-neighbor scheme with diagonal movement scaled by $\sqrt{2}$. Accumulation was constrained to ADM0 to prevent cross-border shortcuts.
Vectorization and cleaning. Binary masks at 60 minutes were polygonized, holes smaller than one analysis cell were removed, and multipart pieces were dissolved per facility identifier. Geometries were validated, self-intersections were fixed, and slivers below one cell were dropped.
Union coverage. A dissolved union of all 60 minute polygons yielded the national coverage mask used to compute the share of land and baseline burden already served.

Travel-time Voronoi partition for unique assignment

Where a unique assignment was required, each analysis cell was assigned to the facility with minimum accumulated travel time, then clipped at 60 minutes. The result is a travel-time

Voronoi partition. These polygons support one-to-one joins for collocation and simplify district accounting. Overlaps from the raw isochrones do not propagate into this partition.

Two-step floating catchment areas, where access scores were needed

When an access metric was required rather than a hard boundary, we computed a 2SFCA index using the same 60 minute reach.

Step 1, facility ratio. For each facility f, we computed

$$R_f = \frac{S_f}{\sum_{j \in \mathcal{D}(f)} D_j \, w(d_{fj})},$$

where S_f is a capacity proxy, D_j is demand in cell j, and w is an impedance weight. Binary w for 0 to 60 minutes was the default. A stepped weight at 0 to 30 to 60 minutes was used in sensitivity.

Step 2, demand access. For each cell i,

$$A_i = \sum_{f \in \mathcal{F}(i)} R_f \, w(d_{fi}).$$

We reported access on the same 5 km grid for comparability with co-risk.

In this study, 2SFCA supported sensitivity checks and mapping only. Collocation used the travel-time Voronoi partition to preserve unique assignment.

Collocation of proposed sites to facilities

Proposed points were intersected with the travel-time Voronoi polygons to obtain a named facility for each point that fell inside any catchment. The workflow was designed to be auditable and to handle edge cases cleanly.

Triage. Each proposed point received one label before assignment:

a) clean when strictly inside one polygon,

b) ambiguous when intersecting multiple polygons or within a small tolerance of a boundary,

c) inside, long connector when inside a very large polygon but far from its representative facility point, and

d) unassigned when outside all polygons. Boundary proximity was measured as the minimum distance from the point to the polygon boundary converted to meters, with a default tolerance of one analysis cell.

Assignment rules. Clean and inside, long-connector points were assigned to the facility that owns the polygon. Ambiguous points were resolved to the nearest facility centroid by travel time, provided the time was less than or equal to 60 minutes. Unassigned points were offered to the nearest facility by travel time if reachable within 60 minutes, otherwise they were retained as candidates for new sites.

Connectors and QA. For every match we computed a straight-line connector length in kilometers and, when the least-cost path was available, a travel-time connector in minutes. Points with long connectors, for example above the 95th percentile of matched connectors, were flagged for manual review since they often indicate very large polygons or irregular friction artifacts.

Quality control and diagnostics

We enforced three classes of checks before downstream use.

Geometry and CRS. All layers were checked for identical CRS and valid polygon topology. The catchments layer was dissolved by facility ID and re-validated to remove self-intersection and dangling nodes.

Coverage plausibility. We computed national counts of covered and uncovered cells, and visualized the union coverage against the Euclidean 25 km buffers to show the expected asymmetry along corridors and across high-friction areas.

Sensitivity. We repeated isochrone extraction at 45 and 75 minutes on a 10 percent stratified sample of facilities to confirm that the coverage footprint scaled smoothly and that the Voronoi partition was stable under modest changes in the threshold.