RESEARCH ARTICLE

# Haplotype heterogeneity and low linkage disequilibrium reduce reliable prediction of genotypes for the -$\alpha^{3.7I}$ form of α-thalassaemia using genome-wide microarray data [version 1; peer review: 1 approved, 1 approved with reservations]

Carolyne M. Ndila[1], Vysaul Nyirongo[2], Alexander W. Macharia[1], Anna E. Jeffreys[3], Kate Rowlands[3], Christina Hubbart[3], George B. J. Busby[3,4], Gavin Band[3,5], Rosalind M. Harding[6], Kirk A. Rockett[3,5*], Thomas N. Williams[1,7*], MalariaGEN Consortium

[1]Department of Epidemiology and Demography, KEMRI-Wellcome Trust Research Programme, Kilifi, PO BOX 230-80108, Kenya
[2]United Nation Statistics Division, United Nations, New York, New York, 10017, USA
[3]Wellcome Centre for Human Genetics, University of Oxford, Oxford, Oxfordshire, OX3 7BN, UK
[4]Centre for Genomics and Global Health, Big Data Institute, University of Oxford, Oxford, Oxfordshire, OX3 7LF, UK
[5]Parasites and Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK
[6]Departments of Zoology and Statistics, University of Oxford, Oxford, Oxfordshire, OX1 3SZ, UK
[7]Department of Infectious Diseases, Imperial College Faculty of Medicine, London, W2 1NY, UK

* Equal contributors

**v1** **First published:** 09 Dec 2020, **5**:287
https://doi.org/10.12688/wellcomeopenres.16320.1
**Latest published:** 22 Sep 2021, **5**:287
https://doi.org/10.12688/wellcomeopenres.16320.2

## Abstract

**Background:** The -$\alpha^{3.7I}$-thalassaemia deletion is very common throughout Africa because it protects against malaria. When undertaking studies to investigate human genetic adaptations to malaria or other diseases, it is important to account for any confounding effects of α-thalassaemia to rule out spurious associations.
**Methods:** In this study we have used direct α-thalassaemia genotyping to understand why GWAS data from a large malaria association study in Kilifi Kenya did not identify the α-thalassaemia signal. We then explored the potential use of a number of new approaches to using GWAS data for imputing α-thalassaemia as an alternative to direct genotyping by PCR.
**Results:** We found very low linkage-disequilibrium of the directly typed data with the GWAS SNP markers around α-thalassaemia and across the haemoglobin-alpha (*HBA*) gene region, which along with a complex haplotype structure, could explain the lack of an association

**Open Peer Review**

**Reviewer Status** ✓ ✓

| | Invited Reviewers | |
|---|---|---|
| | **1** | **2** |
| **version 2** (revision) 22 Sep 2021 | ✓ report | ✓ report |
| **version 1** 09 Dec 2020 | ? report | ✓ report |

1. **Karen G Scheps**, Universidad de Buenos Aires, Buenos Aires, Argentina
CONICET, Buenos Aires, Argentina

2. **Manit Nuinoon**, School of Allied Health

signal from the GWAS SNP data. Some indirect typing methods gave results that were in broad agreement with those derived from direct genotyping and could identify an association signal, but none were sufficiently accurate to allow correct interpretation compared with direct typing, leading to confusing or erroneous results.
**Conclusions:** We conclude that going forwards, direct typing methods such as PCR will still be required to account for α-thalassaemia in GWAS studies.

**Keywords**
Malaria, α-thalassemia, Predictive Models, multinomial regression-model, Classification and Regression Tree, GWAS, haplotypes

Sciences, Walailak University, Nakhon Si Thammarat, Thailand

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding authors:** Kirk A. Rockett (kirk.rockett@well.ox.ac.uk), Thomas N. Williams (twilliams@kemri-wellcome.org)

**Author roles: Ndila CM**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Nyirongo V**: Formal Analysis, Methodology, Writing – Review & Editing; **Macharia AW**: Data Curation, Methodology, Resources, Validation, Writing – Review & Editing; **Jeffreys AE**: Data Curation, Methodology, Resources, Validation, Writing – Review & Editing; **Rowlands K**: Data Curation, Methodology, Resources, Validation, Writing – Review & Editing; **Hubbart C**: Data Curation, Methodology, Resources, Validation, Writing – Review & Editing; **Busby GBJ**: Formal Analysis, Methodology, Writing – Review & Editing; **Band G**: Formal Analysis, Methodology, Resources, Software, Writing – Review & Editing; **Harding RM**: Conceptualization, Investigation, Methodology, Writing – Review & Editing; **Rockett KA**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Williams TN**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing;

**Competing interests:** No competing interests were disclosed.

**Grant information:** The Malaria Genomic Epidemiology Network study of severe malaria was supported by Wellcome [077383] and the Bill & Melinda Gates Foundation through the Foundations of the National Institutes of Health [566] as part of the Grand Challenges in Global Health Initiative. The Resource Centre for Genomic Epidemiology of Malaria is supported by Wellcome [090770; 204911]. This research was supported by the Medical Research Council [G0600718; G0600230; MR/M006212/1]. Wellcome also provides core awards to the Wellcome Centre for Human Genetics [203141] and the Wellcome Sanger Institute [206194]. Sample collection and processing was further supported through a Programme Grant [092654] to the Kilifi Programme from Wellcome. TNW was funded through Senior Research Fellowships from Wellcome [202800; 091758]. CMN was supported by funds from the MalariaGEN consortium and from Wellcome [084538].
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*



**How to cite this article:** Ndila CM, Nyirongo V, Macharia AW *et al.* **Haplotype heterogeneity and low linkage disequilibrium reduce reliable prediction of genotypes for the -α $^{3.7I}$ form of α-thalassaemia using genome-wide microarray data [version 1; peer review: 1 approved, 1 approved with reservations]** Wellcome Open Research 2020, **5**:287
https://doi.org/10.12688/wellcomeopenres.16320.1

**First published:** 09 Dec 2020, **5**:287 https://doi.org/10.12688/wellcomeopenres.16320.1

## Introduction

With recent advances in high-throughput GWAS technologies, a growing number of studies are now being conducted with a view to investigating the contribution of genetics to the risk of a broad range of human diseases. However, such single-nucleotide polymorphism (SNP)-based approaches are imperfect because they only capture a limited picture of total genetic diversity. Specifically, while many SNPs and other short variants across the genome are now routinely accessed by these studies, other important pathogenic variants including larger insertions, deletions and other structural rearrangements are not typically assayed directly. Some of these, such as the thalassaemia-causing mutations of the α-globin (*HBA1-HBA2)* gene region[1], cannot be accurately imputed from current reference panels[2,3], raising a circle of questions about the mutational origin, ancestry and haplotype structure of these variants.

The α-thalassaemias are among the commonest known genetic conditions in humans[4]. They have probably arisen because of an elevated *de novo* mutation rate[5,6] coupled with the fact that they confer a survival advantage against death from *Plasmodium falciparum* malaria[5–7]. As a group, the α-thalassaemias are characterized by the reduced or absent production of the essential α-globin component of normal haemoglobin[4]. α-globin is encoded by a pair of adjacent genes (*HBA1* and *HBA2*) that lie within the haemoglobin-alpha (*HBA*) gene locus on chromosome 16 (Figure 1)[8]. Although many genetic forms of α-thalassaemia have been described worldwide[8], to the best of our knowledge, the 3.7kb Type I deletion ($-\alpha^{3.7I}$) is the only pathogenic variant that occurs at significant frequencies throughout most of Africa[9]. This variant appears to result from a cross-over event between homologous intergenic regions within the *HBA* gene cluster which gave rise to a hybrid gene consisting of the 5' part of *HBA2* and the 3' part of *HBA1*[10]. Although α-globin production is reduced in affected subjects, enough is still synthesized that clinically, both heterozygotes (-α/αα) and homozygotes (-α/-α) are essentially normal[4], while haematologically they display only marginal anaemia and reduced red blood cell volumes[11]. Despite its high allele frequencies and potential importance as a confounder in GWAS studies in African populations, the $-\alpha^{3.7I}$ deletion is not well-captured using regional SNP-based genotype inference methods[2,3]. Notwithstanding this, recovering α-thalassaemia genotypes from GWAS data could be very useful given the amount of data now available worldwide, particularly as it would be difficult or impossible to genotype such samples retrospectively for α-thalassaemia.

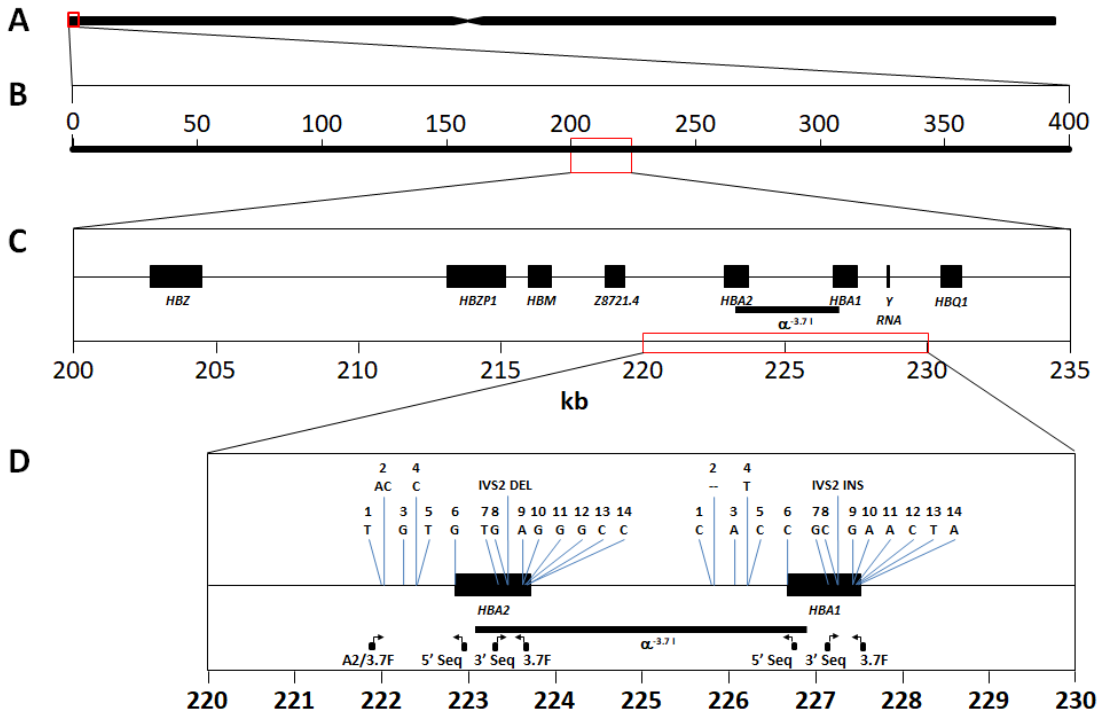



**Figure 1. Schematic of the *HBA* region on human chromosome 16. A**: Representation of human chromosome 16 showing the location of the *HBA* gene region at the p-telomere end (red box). **B**: The 400kb chromosome region spanned by the SNPs used in this study (16:83,000-400,000) approximately centres around the *HBA* gene region (red box). **C**: Chromosome 16:200000-235000 spanning the classical *HBA* gene region, comprising *HBZ* [ζ2], *HBZP1* [ψζ1], *HBM* [ψα1], *HBA2* [α2], *HBA1* [α1] and *HBQ1* [θ1]. The $\alpha^{-3.7I}$ deletion is highlighted between *HBA1* and *HBA2*. **D**: *HBA1* and *HBA2* genes showing the location of the primers used for genotyping and Sanger sequencing (see methods and Extended Data); the region of the $\alpha^{-3.7I}$ deletion; and 15 bases/features that show paralogous differences in the human reference genome between *HBA1* and *HBA2* sequences and used to identify the $\alpha^{-3.7}$ Type I breakpoint (Extended Data).

In this current study we use α-thalassaemia genotype data from a previous study of more than 6,000 children from Kilifi, Kenya[12], along with their corresponding Illumina HumanOmni2.5-4 microarray data[3], to investigate the haplotype structure surrounding the α-thalassaemia deletion variants in this population, and to gain an insight into why inferred α-thalassaemia genotypes were not well captured. We also explored the potential utility of a wide range of other indirect GWAS-based approaches, including the microarray-chip intensity data and haplotype imputation, as an alternative to direct typing of α-thalassaemia, which is technically challenging and adds additional costs.

## Methods

### Study population

Participants were recruits to a case-control study of severe malaria as described in detail previously[12]. Briefly, cases were children presenting to Kilifi County Hospital in Kenya with features of severe falciparum malaria, while controls were children recruited from the surrounding community to a genetic cohort study of childhood illness (the Kilifi Genetic Birth Cohort Study[13]). A subset of 3,036 participants (Table 1) were selected from this study for whom GWAS data and α-thalassaemia sickle (rs334) genotypes were already available from previous studies[3,12].

### Ethics approval and consent to participate

The study was approved by the Kenya Medical Research Institute/National Ethical Review Committee in Nairobi, Kenya (Number: SCC1192), and by the Oxford Tropical Research Ethics Committee in Oxford UK (Number: OXTREC 020-06). Written informed consent for inclusion in this study was given by all participants or their parents.

### Sample genotyping

We used data derived using the Illumina HumanOmni2.5-4 genotyping chip (Illumina, California, USA), which were aligned to Human Genome build GRCh37 and curated as previously

**Table 1. Main characteristics of the studied population and distribution of the chromosomes used in the analysis.**

| Characteristics | Overall | Genotyped at the α-thalassaemia locus | | |
|---|---|---|---|---|
| | | Homozygous ancestral (αα/αα) | Heterozygous derived (-α/αα) | Homozygous derived (-α/-α) |
| All subjects, N (%) | 3036 | 1139 (37.5) | 1474 (48.6) | 423 (13.9) |
| Cases | 1432 (47.2) | 588 (19.4) | 673 (22.2) | 171 (5.6) |
| Controls | 1604 (52.8) | 551 (18.1) | 801 (26.4) | 252 (8.3) |
| Gender, N (%) | | | | |
| Males | 1543 (50.8) | 579 (19.1) | 750 (24.7) | 214 (7.0) |
| Females | 1493 (49.2) | 560 (18.4) | 724 (23.8) | 209 (6.9) |
| Ethnicity, N (%) | | | | |
| Giriama | 1494 (49.2) | 547 (18.0) | 715 (23.6) | 232 (7.7) |
| Chonyi | 946 (31.2) | 342 (11.3) | 482 (15.9) | 122 (4.0) |
| Kauma | 271 (8.9) | 114 (3.8) | 124 (4.1) | 33 (1.1) |
| Others (18) | 325 (10.7) | 136 (4.5) | 153 (5.0) | 36 (1.2) |
| Thalassaemia chromosomes, N | | | | |
| αα | 3752 | 1139 | 1474 | - |
| -α | 2320 | - | 1474 | 423 |
| Sickle (rs334) chromosomes*, N (%) | | | | |
| AA | 2730 (89.9) | 1026 (33.8) | 1317 (43.4) | 387 (12.7) |
| AT | 283 (9.3) | 104 (3.4) | 146 (4.8) | 33 (1.1) |
| TT | 23 (0.8) | 9 (0.3) | 11 (0.4) | 3 (0.1) |

Numbers given are for all individuals included in this study, with percentages shown in parentheses by row.

* The A allele encodes for normal β-globin while the T allele encodes for $\beta^S$-globin such that AT individuals have sickle cell trait and TT individuals have sickle cell disease.

described[3,14,15] and that are already publicly available[16]. Details of how to access the GWAS data package can be found on the MalariaGEN website and in the Data Availability section below. For the analysis of the α-globin region we extracted genotype data and intensity data from the chromosome 16 vcf file (Extended Data Figure 1; Kenya_GWAS-2.5M_b37_chr16_aligned.vcf.gz) into GEN format using QCtool, (options and parameters are shown in Extended Data Figure 1), to which we included lists of excluded samples (Extended Data Figure 1; Kenya_GWAS-2.5M_b37.sample [762/3869 samples]) and SNPs (Extended Data Figure 1; Kenya_GWAS-2.5M_b37_snp_qc.txt [21756/80392 SNPs]) based on information in the QC files provided with the Kenya GWAS dataset package (sample and SNP missingness of ≤ 0.05, allele frequencies of ≥ 0.01 and filtration for curated sample and SNP duplicates) and our own sample requirements (data on clinical status [GWAS data package], data for both α-thalassaemia [Underlying Data2 Table AA_sample_codings], rs334 genotyping [GWAS data package] and gender [GWAS data package]; Extended Data Figure 1; alphathal_sickle_genotypes.csv; Kenya_all_samples.sample; clin_phenotypes.csv). The QCtool outputted a 0-10Mb region of chromosome 16 giving 9723 SNPs for 3107 samples (Extended Data Figure 1). At this point the α-thalassaemia data were merged into the data set using QCtool and then phased into the haplotypes with ShapeIT v2 (options and parameters are shown in Extended Data Figure 1). An African recombination map (Underlying Data2 Table CC_chr16_recombination) was included for this step. Following phasing, the sample set was reduced by a further 71 samples missing HbS (sickle) genotype and/or gender data (used as important covariates in our analysis). After this phasing step, we selected all polymorphisms across the 400kb region from the p-arm telomere (spanning 84870 – 398421bp) of chromosome 16 spanning *HBA2* and *HBA1* where the $\alpha^{-3.7}$ deletion is located (219454–227532: Underlying Data1 and Underlying Data2 Table BB_0-400kb_snp_details). This resulted in a final dataset comprising 179 polymorphisms (178 SNPs and the $\alpha^{-3.7I}$ deletion by direct typing) (Underlying Data2 Table BB_0-400kb_snp_details) for 6072 chromosomes from 3036 individuals (Table 1)[12,17].

## Association analysis

All association analyses were performed using R (Extended Data Table 13) as previously described[12]. The sample size for this study was determined pragmatically, based on the number of samples that were available from our previous study in Kenya[3,15], and the completeness of the GWAS data available (Table 1). Odds ratios for SNP associations with severe malaria were determined by comparison of allele and genotype frequencies among cases and controls, using a fixed-effects logistic-regression model, both with (Underlying Data2 Table JJ_assoc_results_adjusted_hbs) and without (Underlying Data2 Table II_assoc_results_unadjusted) adjustment for the confounding effects of genetic background (using self-reported ethnicity) and rs334 genotype, the SNP responsible for both sickle-cell trait (HbAS) and sickle-cell disease (HbSS), all of which are major potential confounders in the interpretation of such analyses. Gender is also associated with malaria risk both at a genetic and at a social/cultural level, and was also included as a co-variate in our adjusted analyses.

## Linkage disequilibrium (LD) and haplotype frequency estimation

We determined genotype frequencies for α-thalassaemia and all SNPs individually by direct allele counting (Underlying Data2 Table BB_0-400kb_snp_details). Pairwise LDs were estimated by calculating r, $r^2$ and D' metrics using a custom R script (Extended Data Section 7; Underlying Data2 Tables KK_pairwise_R, LL_pairwise_R2 and MM_pairwise_Dprime). Extended haplotype homozygosity (EHH) and bifurcation diagrams were calculated in R using *rehh* (Extended Data Table 13). Haplotype tree structures were analyzed using the R packages *hclust* and *dendrograms* (Extended Data Table 13) and were visualized using *pegas* and *dendextend* (Extended Data Table 13). Haplotype diversity[18] was estimated using with *hap.div* function within *pegas*.

## Sanger sequencing to confirm the form of α-thalassaemia present in the Kilifi population

Genotyping of the α-thalassaemia polymorphism was undertaken as described[19] and sequences (forward orientation) for the human reference sequences of the *HBA2/HBA1* (16:221882-227718) region were downloaded from Ensembl GRCh37 to span the forward and reverse PCR primers used to determine the presence or absence of the $\alpha^{-3.7I}$-deletion (Table 2). The *HBA2* region spanning the common forward PCR primer (A2/3.7F; 16:221882-221901 [Table 2], that lies at the beginning of the Y2 box, to approximately 50bp 3' of end of the X-box [16:223809] [Figure 3]) was aligned with an equivalent *HBA1* region between 16:225474 and the *HBA1* reverse PCR primer (3.7-R; 16:227698-227718 [Table 2, Figure 2]). Clustal Omega was used to generate the alignment and manual finishing was used to finalise any misaligned regions and identify features (Figure 2). From the aligned sequences, a number of key differences were identified across the homologous *HBA2* and *HBA1* regions that included seven individual bases, the 7bp IVS2 INDEL and four restriction sites (ApaI, BalI, ApaI and RsaI – 1, 2, 3, 2 base-differences, respectively[10]) at the 3' end of the *HBA* coding region (Figure 2 and Table 3 [Underlying Data2 Table NN_Sanger_Sequence_summary]).

For Sanger sequencing, 25 individuals identified as homozygous for the -$\alpha^{-3.7I}$ deletion (using primers A2/3.7F and 3.7-R [hybrid *HBA2-HBA1*]) and 9 individuals homozygous wild-type (primers A2/3.7F and A2-R [*HBA2*], and *HBA1*_specific_F with A3.7R [*HBA1*]) were selected and used to generate fresh PCR product as previously described[19]. We used the same set of 9 wild-type individuals for sequencing both *HBA1* and *HBA2*.

Amplicon presence and genotype classes were confirmed by running an aliquot of the PCR product by agarose gel electrophoresis. The remaining PCR product was cleaned using the Qiagen PCR clean-up kit (QIAquick PCR purification kit, Qiagen, Crawley, UK) and quantified using picogreen (Quant-iT™, Thermo Fisher Scientific, Loughborough, UK).

**Table 2. Primer details for Sanger sequencing the alpha-thalassaemia deletion region in *HBA1* and *HBA2*.**

| Name | Dir | Genome Position | Gene | Genome Position | Gene Position | Primer Sequence (5'-3') |
|---|---|---|---|---|---|---|
| A2/3.7-F | FWD | 16:221882-221901 | 5' to *HBA2* | | | CCCCTCGCCAAGTCCACCC |
| *HBA1*_specific_F | FWD | | | 16:225771-225792 | *HBA1* | CTCCAGCCGGTTCCAGCTATTG |
| *HBA*-5' SEQ REV | REV | 16:222930-222950 | *HBA2* | 16:226734-226754 | *HBA1* | GGCCTTGACGTTGGTCTTGTC |
| *HBA*-3' SEQ FWD | FWD | 16:223310-223332 | *HBA2* | 16:227114-227136 | *HBA1* | GGACCCGGTCAACTTCAAGGTGA |
| A2-R | REV | 16:223666-223684 | 3' to *HBA2* | | | AAAGCACTCTAGGGTCCAGCG |
| 3.7-R | REV | | | 16:227698-227718 | 3' to *HBA1* | CCCCTCGCCAAGTCCACCC |

α-thalassaemia PCR was undertaken using the standard protocol as described[19]. PCR products were then sequenced using the primers above (see methods).

A2/3.7-F, A2-R and 3.7-R are primers used in the standard PCR assay for $\alpha^{-3.7}$-thalassaemia deletion.

*HBA*-5' SEQ REV and *HBA*-3' SEQ FWD were designed internal to the products to allow more coverage. All primer locations are identified in Figure 2

Amplicons and sequencing primers (Table 2) were prepared according to instructions and sent to Eurofins (Eurofins Genomics UK, Wolverhampton, UK) for processing. In summary;

- PCR amplicons for *HBA2* were generated from 9 individuals who were identified as wild-type for *HBA1* and *HBA2* using primers A2/3.7-F, and A-2R (Table 2).
- PCR amplicons for *HBA1* were generated from 9 individuals who were identified as wild-type for *HBA1* and *HBA2* using primers *HBA1*_specific_-F, and A3.7R (Table 2) – NB: these were the same individuals as for the *HBA2* sequencing.
- PCR amplicons for the $\alpha^{-3.7}$ deletion were generated from 25 individuals who were identified as homozygotes using primers A2/3.7-F and 3.7-R (Table 2).
- Sanger sequencing was undertaken on all amplicons using their respective forward and reverse PCR primers, plus 2 internal primers; HBA-5'.SEQ and HBA-3' (Table 2).

Sequence traces for individuals' samples were inspected and curated using Chromas to generate FASTA files for alignment on Clustal Omega with the human reference sequences (Extended Data Sections 9-12 for pile-ups). Alignments were then manually finished before inspection and the key paralogous bases in the human reference sequence that distinguish *HBA2* from *HBA1* (Figure 2) were identified and recorded (Table 3 [Underlying Data2 Table NN_Sanger_Sequence_summary]).

## Illumina chip intensity data

The Illumina genotyping method creates intensity data at two channels (X and Y), one for each of the alleles, for each feature (typically a SNP)[20]. We identified all features on the Illumina chip used for genotyping the samples (HumanOmni2.5-4v1_H) across the $-\alpha^{3.7kb}$ deletion region that sits between the Z-boxes of *HBA2* and *HBA1* [16:219454-227532] (Underying Data2 Tables BB_0-400kb_snp_details and OO_HBA_region_features, Figure 3), and extracted intensity data for these SNPs across all samples. We also included intensity data for the 5' and 3' SNPs immediately flanking the deletion region that were used in the haplotype analysis (rs2974771 [16:221057] and rs9936930 [16:233272]; Extended Data Sections 4 and 5). Because the intensity data may not always be normally distributed and the group sizes between genotypes may not be even, we made comparisons between genotype intensities for each SNP using the Kruskal-Wallis non-parametric test followed by Dunn's test to compare each pair of genotypes[21,22]. We also computed Cohen's *d* and Hedges' *g* effect size metrics[23–25]. (Underlying Data2 Tables PP and QQ).

## Inferring α-thalassaemia genotypes from intensity data

We took four alternative approaches to inferring α-thalassaemia genotypes.

***A) Direct analysis of intensity distribution***

The first was based directly on the intensity data while the second involved models built on small amounts of data derived by direct genotyping. In the first instance for each of the six chip features located in the deletion region we plotted the distribution of the sum of the X and Y channel intensities as a density function and then selected two clear troughs or shoulders (C1, C2) in the density distributions (Figure 4). Genotypes were then assigned as; homozygous for the derived/deletion allele <= C1; C1 < heterozygous < C2; and homozygous for the ancestral allele >= C2 (Extended Data Sections 4a and 4b).

***B) Hierarchical clustering of intensities***

We next undertook a hierarchical clustering of the data for the six chip features located in the $\alpha^{-3.7I}$ deletion. This was applied using the *heatmap2* function of the *gplots* package

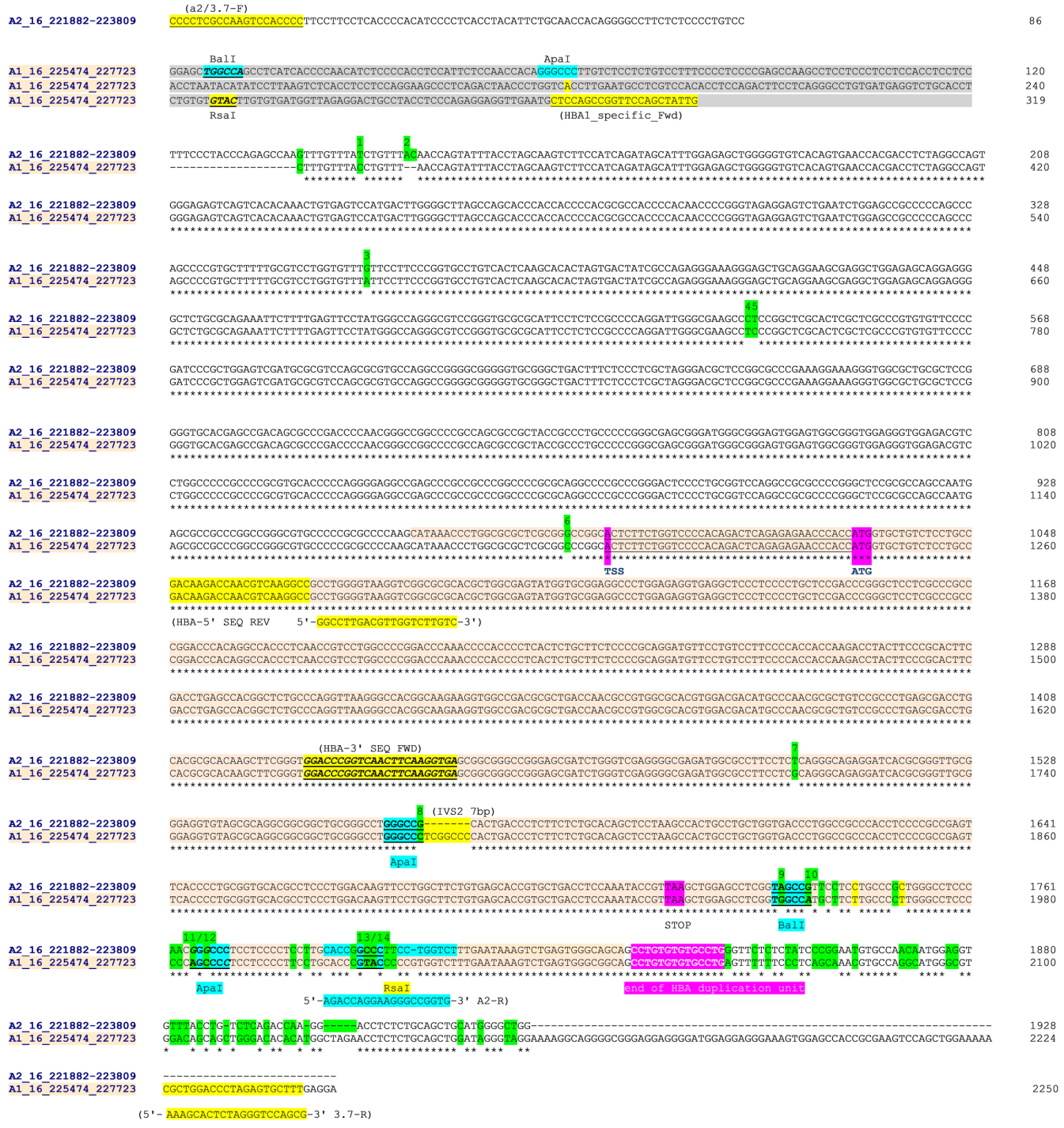



**Figure 2. Sequence alignment from Clustal Omega across *HBA2* and *HBA1* on human chromosome 16 (Sequence data from Ensemble GRCH37).** Sequences were aligned in chromosomal order (*HBA2* [prefix 'A2_16_'] above *HBA1* [prefix 'A1_16_']). The *HBA2* sequences starts at the 5'-most base of the forward PCR primer (A2/3.7-F position 16:221882) in a unique region and 86 bases 5′ of the homologous region with *HBA1*. The *HBA1* sequence starts at position 16:225474 which is 319 bases from the equivalent homologous region with *HBA2*. The *HBA2* sequence ends at position 16:223809 effectively at the end of the homologous region with *HBA1,* while the *HBA1* sequences ends at position 16:22773 and the PCR reverse primer (3.7-R). PCR and sequencing primers (Table 2) are highlighted (A2/3.7-R,HBA-5′-SEQ-REV, HBA-3′SEQ-FWD and 3.7-R) as are key restriction sites[10]. Paralogous differences between *HBA1* and *HBA2* reference sequences are highlighted in green for a set of 14 positions. These were used to help identify the $\alpha^{-3.7}$ deletion type from Sanger sequencing. The *HBA* genic region is coloured and shows the transcription start site (TSS), amino-terminal methionine (ATG) and stop codons (TAA); but not sepearate introns and exons for clarity. The four restriction sites used to distinguish the $\alpha^{-3.7}$ deletion types are identified[10]; the Type I breakpoint was identified as being 5′ of the ApaI/IVS2 sequence; the Type II breakpoint lies between the ApaI/IVS2 and BalI restriction sites; The Type III lies between the RsaI and the 'end of *HBA* duplication unit' (location and identity of *HBA* duplication unit[26]). NB: The ApaI/IVS2 sequence comparison between *HBA2* and *HBA1* has been aligned here to clearly highlight the ApaI restriction site; it may be shown differently in other publications.

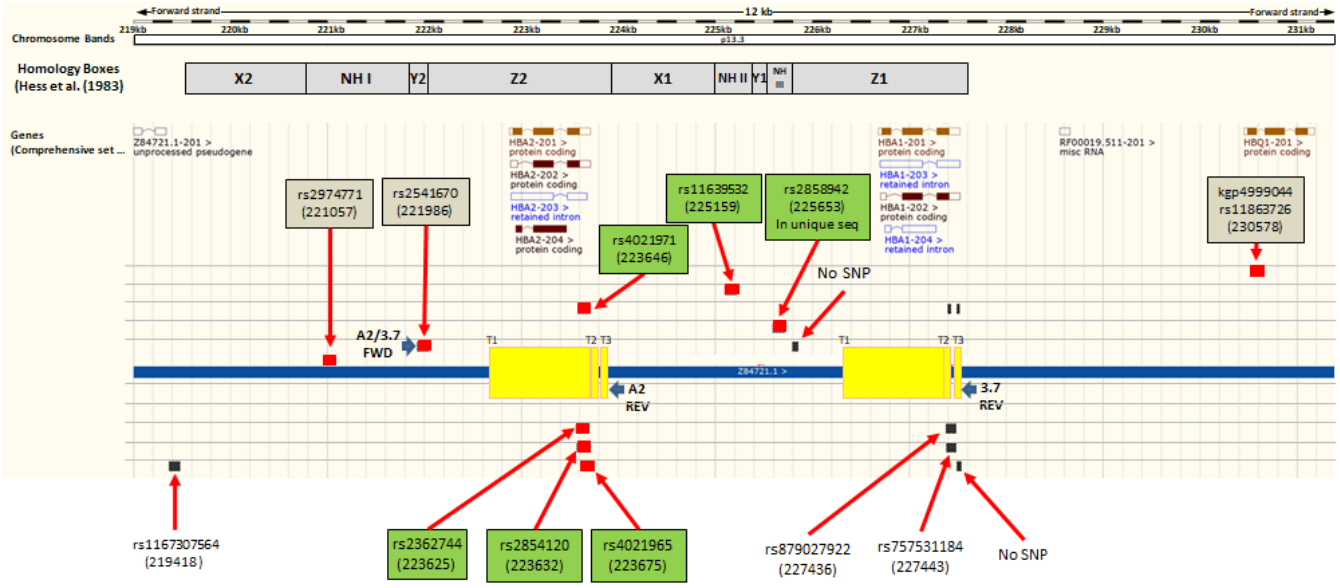



**Figure 3. Map of the *HBA* region on human chromosome 16 identifying Illumina chip features flanking and internal to the α$^{-3.7}$kb deletion.** Ensembl GRCh37 chromosome 16 *HBA* region; Illumina HumanOmni2.5-4 feature match; (red boxes are perfect match of probe with ref sequence; Black boxes are lesser matches; boxes on same row/level are from same probe); SNP name boxes have GRCh37 positions; Green boxes are six features within the deletion region while the three brown boxes are the SNPs immediately flanking the region. Non-highlighted labels are other blast hits. Yellow boxes show regions of breakpoints/crossovers for the three known types of α$^{-3.7}$kb deletion; Homology boxes X, Y and Z indicated as per Hess *et al.*[27].

in R[28]. We created heatmaps for all six deletion features (Figure 5) but as can be seen in panel A, the 2 3'-most features have different intensity profiles to the other four features. We have therefore also created heatmaps for the 5 5'-most features and the 4 5'-most features. In both the latter cases we identified the three most likely clusters corresponding to the three genotype classes (red, αα/αα normal; green, -α/αα heterozygotes; blue, -α/-α homozygotes). The assigned genotypes were extracted and compared with the directly-typed genotypes (Underlying Data2 Table TT_del_hier_cluster_assignment).

***C) Multiple-Regression Model (MRM) and Classification and Regression Trees (CART)***

a) MRM

The MRM extends the binary logistic models when there are more than two outcome categories. This approach was used to model, classify and predict multi-category outcomes conditional on a given set of explanatory variables. For an individual case (denoted as $i$), we want to predict whether the α-thalassaemia genotype is homozygous for α$^{-3.7}$ (homo), heterozygous (het) or homozygous wild-type (norm) given intensities of predictor SNPs rs236744, rs2854120, rs4021971, rs4021965, rs11639532, rs2858942.

Let the tuple $X = \left[x_i^k, y_i^k\right]_{k=1}^{3}$ denote the intensity values for the $i^{th}$ individual at genotype $k$ = 1,2,3 (het, homo, norm respectively) for each SNP. As in other forms of linear regression, the MRM uses a linear function of chosen predictor variables, say:

$$l(T = t_g \mid X) = a_g + \sum_k (b_{gk}x_i^k + c_{gk}y_i^k + d_{gk}x_i^k y_i^k) \quad \text{equation 1}$$

Where

$$Homo_i = a_1 + b_1x_i^1 + c_1x_i^2 + d_1x_i^3 + e_1y_i^1 + f_1y_i^2 + g_1y_i^3 + h_1x_i^1y_i^1 + m_1x_i^2y_i^2 + n_1x_i^3y_i^3 \quad \text{equation 2}$$

$T = t_g$ and denotes α-thalassaemia het, homo or norm genotypes ($g$ = 1,2,3 respectively); $a_g$, $b_{gk}$, $c_{gk}$ and $d_{gk}$ are regression coefficients that are learned from a training dataset (a set of samples with known genotypes).

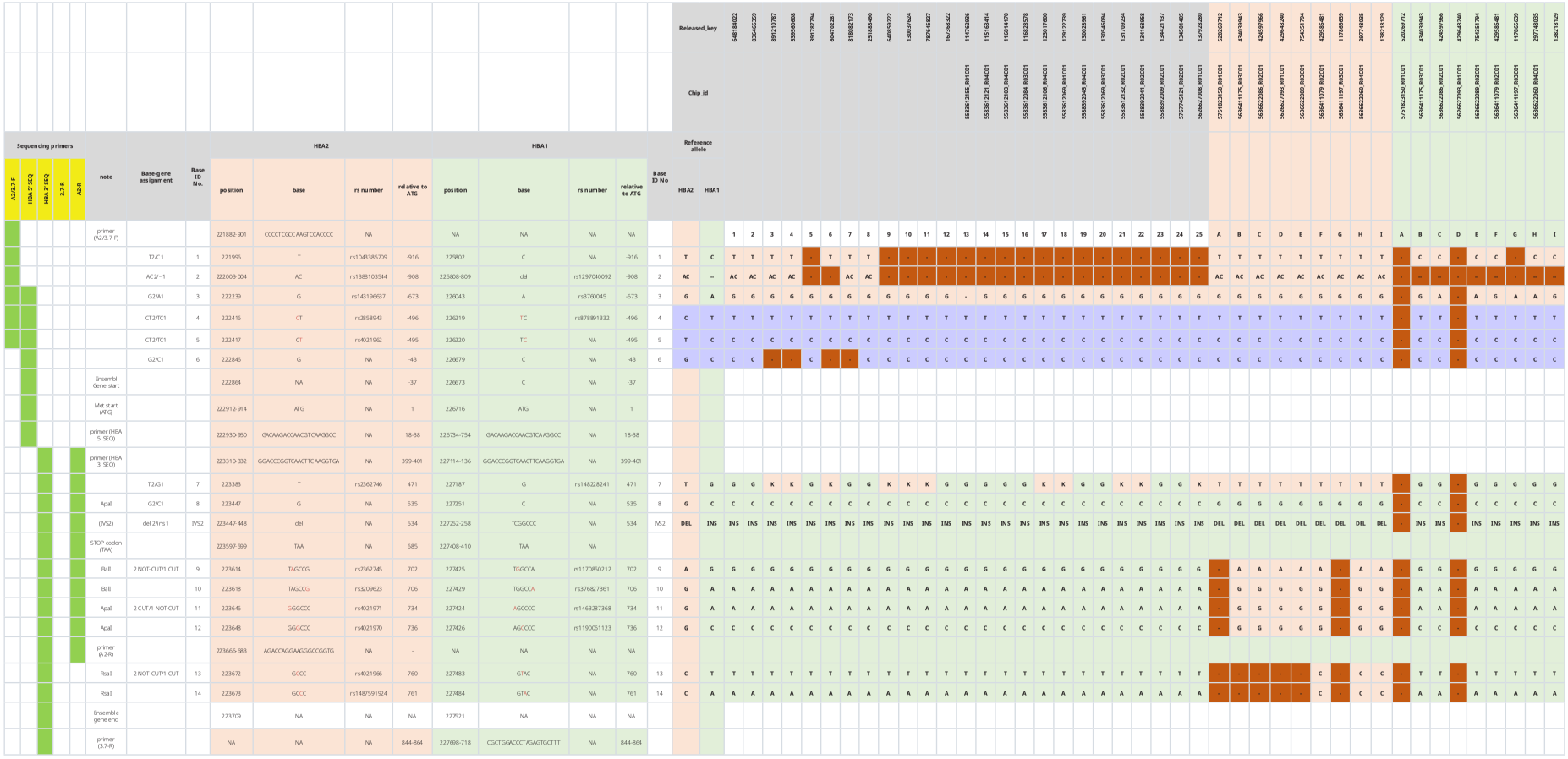
Note that $l(t_g \mid X)$ is a function of learned parameters and intensities that we can use to predict unknown α-thalassaemia genotypes wherever the intensities of each SNP are known. Thus, for given intensities of $X$, and parameters $a_g$, $b_{gk}$, $c_{gk}$ and $d_{gk}$, the probability that α-thalassaemia genotypes $T = t_g$ are het, homo or norm for the $i^{th}$ individual are:

$$\text{a)}\quad p(t = t_1) = \frac{1}{1 + \exp(l(t_2, X)) + \exp(l(t_3, X))}\, i^{th} \quad \text{heterozygote} \quad \text{equation 3}$$

$$\text{b)}\quad p(t = t_2) = \frac{\exp(l(t_2, X))}{1 + \exp(l(t_2, X)) + \exp(l(t_3, X))}\, i^{th} \quad \alpha^{-3.7}\ \text{homozygote} \quad \text{equation 4}$$

$$\text{c)}\quad p(t = t_3) = \frac{\exp(l(t_3, X))}{1 + \exp(l(t_2, X)) + \exp(l(t_3, X))}\, i^{th} \quad \text{wild - type} \quad \text{equation 5}$$

**Table 3. Sanger sequencing of PCR amplicons for the α$^{-3.7}$kb deletion in 25 homozygous α$^{-3.7}$kb and 5 normal individuals in Kilifi, Kenya.**



Individuals were identified following PCR typing for the α$^{-3.7}$kb deletion and selected for Sanger sequencing of the amplicons. From the sequence traces the identity of a set of differentiating bases was recorded (*Base ID No.) as highlighted in Figure 2.

The table rows show features ordered by chromosome position.

Sequencing primers: These columns show the span of the Sanger sequencing typically covered by the named primers.

Note: field identifies specific features across the *HBA2* and *HBA1* region; primers; gene start, stop; restriction enzyme sites[10] distinguishing the α$^{-3.7}$ types I, II and III.

*HBA2*, *HBA1*: these columns identify the chromosomal position, sequence, rsIDs and position relative to the ATG site for *HBA2* and *HBA1* aligned to each other.

Sequenced individuals are shown with their unique IDs (chip_id and released keys that match them to the main datasets).

For individuals, the bases identified are shown and coloured according to the reference sequence (green – *HBA1*, orange – *HBA2*).

Bases in purple were found to be identical across all individuals sequenced rather than divergent as in the reference sequences. Dashes are bases that were not possible to identify due to short sequences. The full table of this data can be found in Underlying Data2 Table NN_Sanger_Sequence_summary.



Wellcome Open Research 2020, 5:287 Last updated: 27 SEP 2021

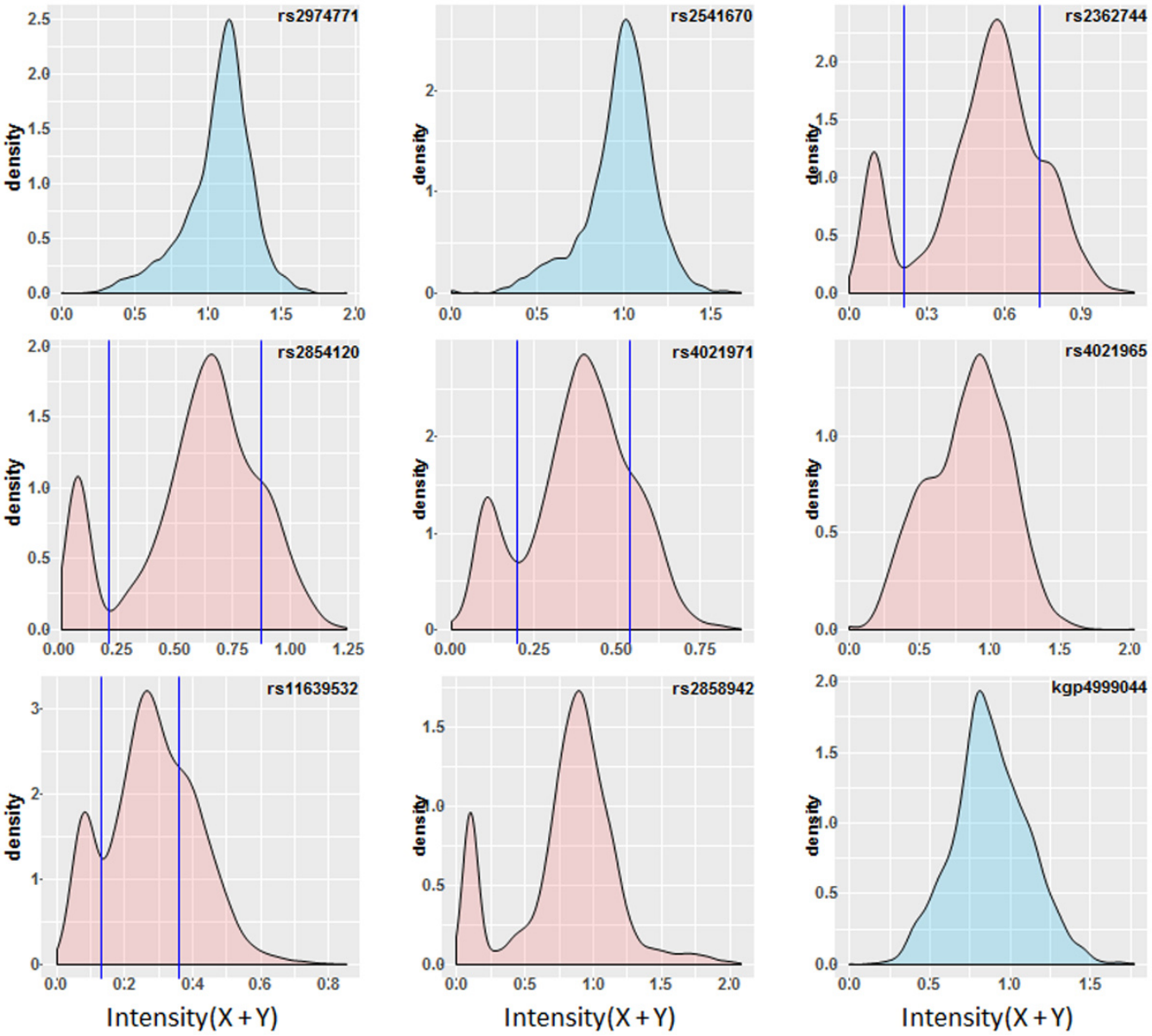
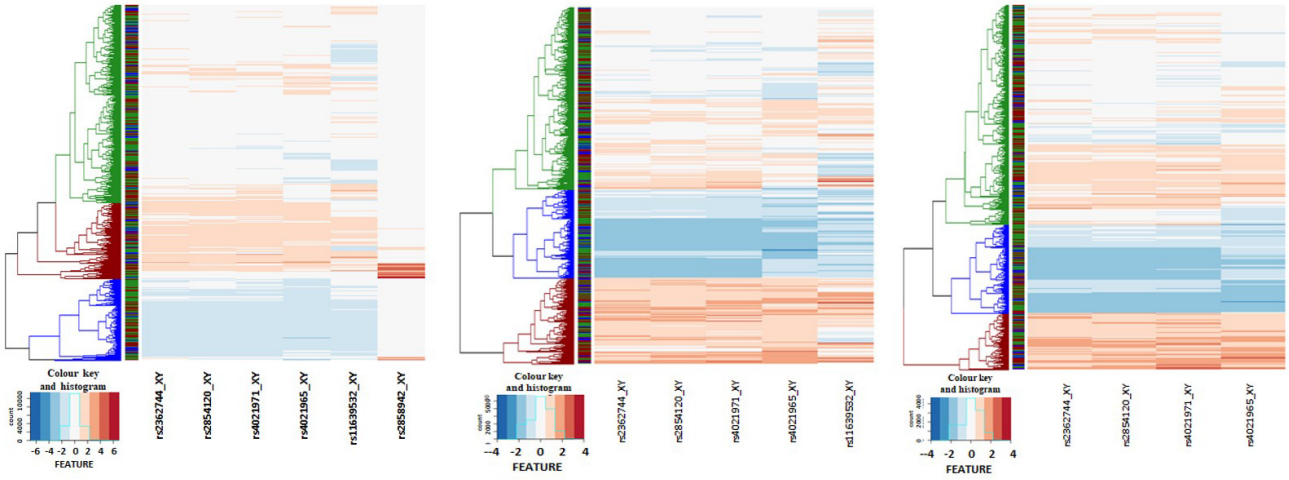



**Figure 4. Chip Intensity (sum[X and Y] channels) density plots for features internal and immediately flanking the $\alpha^{-3.7}$ deletion.** Features filled in blue are flanking to the deletion (rs2972771 and kgp4999044 are present in the haplotype SNP data, rs2541670 did not pass QC). Features filled in red are within the $\alpha^{-3.7}$ deletion. Vertical lines illustrate where both a trough and shoulder are discernable in the distribution and potentially infer the breaks between genotypic groups; rs2362744 0.211 and 0.738; rs2854120 0.215 and 0.869; rs4021971 0.199 and 0.538; rs11639532 0.131 and 0.358.

i. MRM construction

A general process flow is shown in Figure 6A. We first selected the 6 SNP-features located in the $\alpha^{-3.7I}$ deletion region (rs236744, rs2854120, rs4021971, rs4021965, rs11639532, rs2858942) and extracted the intensity data from the Illumina chips. We then made a random selection of $n$=50 to $n$=500 individuals with known α-thalassaemia genotypes (the training set) for each of the SNP-features. These were used to calculate a set of parameters that the model could use to predict genotypes in the remaining (3036 – $n$) individuals. Finally, using these parameters the model assigned a probability to each α-thalassaemia genotype category for each individual. The category with the highest predicted probability was used to define the genotype for that individual. We repeated this 1000 times for each set of $n$ samples to calculate a mean and SD and a predictive power score for each genotype class using the directly typed results as baseline. From the plots of the performances (Extended Data Figure 24A) we were able to identify a




| | NORM | HET | HOMO | NORM | HET | HOMO | NORM | HET | HOMO |
|---|---|---|---|---|---|---|---|---|---|
| Sensitivity (True Positive Rate) | 0.48 | 0.78 | 0.96 | 0.56 | 0.74 | 0.96 | 0.38 | 0.80 | 0.96 |
| Specificity (False Positive Rate) | 0.93 | 0.61 | 0.87 | 0.95 | 0.71 | 0.86 | 0.98 | 0.58 | 0.87 |
| Positive Predictive Value | 0.77 | 0.64 | 0.55 | 0.86 | 0.71 | 0.53 | 0.92 | 0.65 | 0.55 |

**Figure 5. Heatmaps of core deletion feature intensities.** Heatmaps were generated for four, five and six SNP features inside the $\alpha^{-3.7I}$ deletion region. Sample-intensities were clustered as shown by the dendrograms at the left side of each panel and genotypes assigned (Blue: Homo, Green: Het, Red: Norm). In each case the intensities between SNPs were normalised to create a common intensity profile (Colour key and Histogram inset at the bottom left of each panel). Left-hand panel: all six SNP-features; Middle panel: five SNP-features with rs2858942 removed; Right-hand panel: four SNP-features with rs2858942 and rs11639532 removed. The ribbon between the dendrogram and the SNP features show the directly-typed genotype assignments in the same colour scheme.

minimum training set of $n$ = 100 individuals that could be used for further analysis using ROC curves and association analysis (Example run: Underlying Data3).

b) CART

i. Background

CART is a machine-learning approach that has been shown to be particularly useful when analyzing non-linear relationships and it is thought to be more robust than the standard regression models for classification[29]. Furthermore, the approach is simple to understand or interpret and requires little data preparation. It uses "decision trees" to classify new data.

ii. CART Model construction

A general process flow is shown in Figure 6B. We first selected the 6 SNP-features located in the $\alpha^{-3.7I}$ deletion region (rs236744, rs2854120, rs4021971, rs4021965, rs11639532, rs2858942) and extracted the intensity data from the Illumina chips. We then made a random selection of $n$=50 to $n$=500 individuals with known α-thalassaemia genotypes (the training set) for each of the SNP-features. These were used to calculate a set of parameters that the model could use to predict genotypes in the remaining (3036 – $n$) individuals.

We used CART to build a binary tree by subdividing the data at each possible split and choosing the best split that produces the most homogenous sub-groups (as an example Extended Data Figure 23 shows the CART dendrogram and cut-offs applied to the polar cluster plots for each SNP for the full dataset). The predictive variables (the chip feature intensities) included in the model were the same as the MRM approach. The process of growing the tree was stopped when the number of individuals within each node was less than five. We repeated this 1000 times for each set of $n$ (50 – 500) samples to calculate a mean and SD and a predictive power score for each genotype class using the directly typed results as baseline. From the plots of the performances (Extended Data Figure 24B) we were able to identify a minimum training set of $n$ = 100 individuals that could be used for further analysis using ROC curves and association analysis (Example run: Underlying Data3).

***D) Imputation using haplotypes***

An increasingly popular way to infer missing genotype information is by statistical imputation using haplotype structure. This method employs a reference set of haplotypes (such as the 1000 Genomes resource[30–32]) for comparison with the observed test data generated on genome-wide chip arrays. The IMPUTE2 program[33,34], was developed to phase the test data (if not already done so) and compare to the reference data set to find the most probable haplotype matches to infer (impute) any missing polymorphic sites in the data.

Our dataset comprised 178 polymorphisms directly typed on the Illumina HumanOmni2.5-4 for the 0-400kb region on chromosome 16 (see above and Extended Data Section 1) in 3036 individuals with the addition of PCR-directly-typed

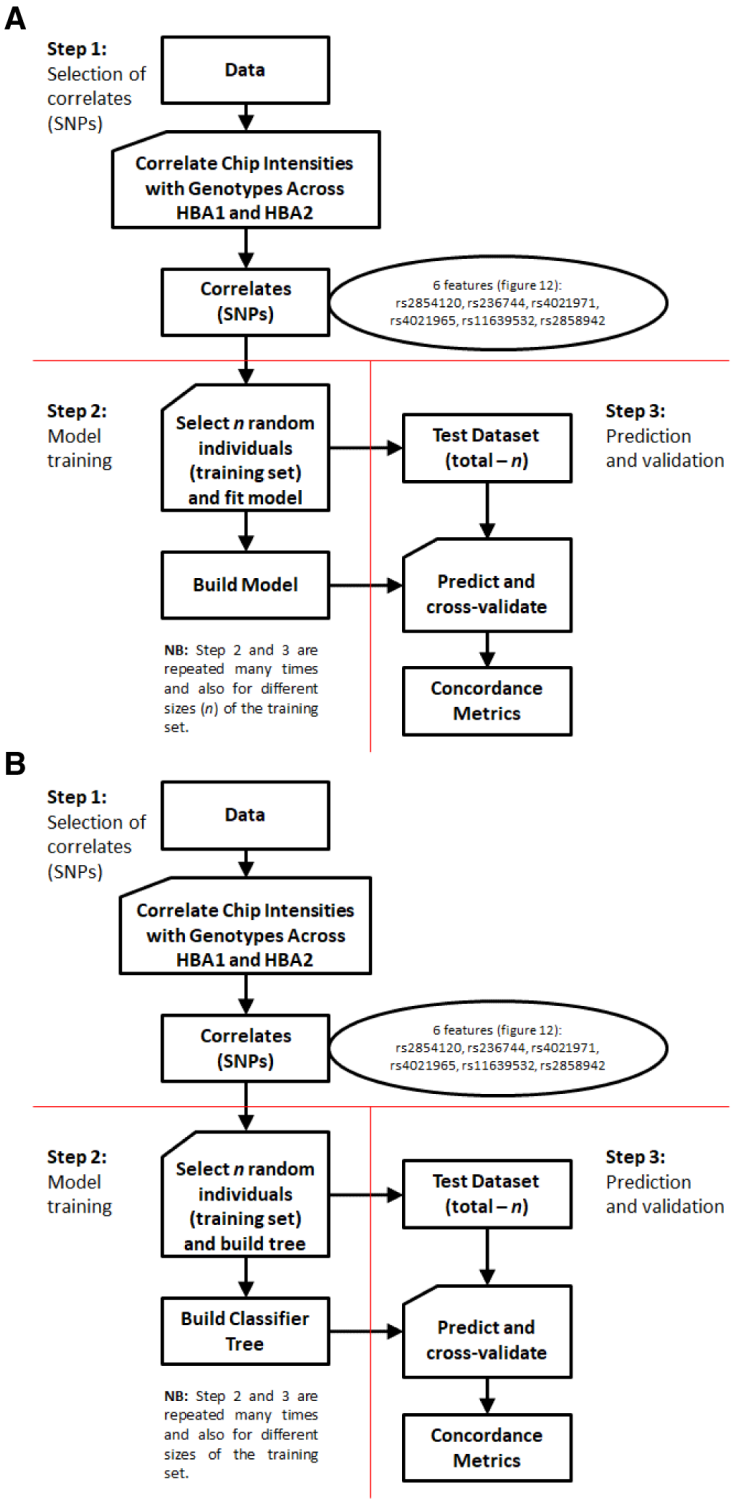



**Figure 6. Multiple-Regression Model (MRM) and Classification And Regression Tree (CART) process flows. A**: MRM process flow. This is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. The model can then be applied to new explanatory variables (i.e. without known genotypes) to predict unknown genotypes **B**: CART process flow. This method builds a binary decision tree (i.e. a series of evaluations based on a single concomitant variable at each point) and aims to split the data such that there is maximal separation of individuals in terms of the variable of interest. At each point, the evaluation of an individual is either positive or negative and the procedure seeks a cut-off point for a range of values of the concomitant variable such that the positive and negative groups contain maximal number of individuals of the same type. These learned series of evaluations can then be applied to a new set of individuals with concomitant variables known (without known types) to predict their unknown types. Here the concomitant variables are the intensities while the "types" are genotypes.

α-thalassaemia genotypes (see above and Extended Data Section 1) for comparison testing.

IMPUTE2 requires a reference set of haplotypes to infer missing genotypes. However, we decided not to use the 1000 genome reference panel for several reasons;

- The 1000G dataset does not contain any directly typed α-thalassaemia deletion variants, only imputed from the sequence data.
- The 1000G dataset does not have any populations that directly match the populations used in our study (the closest is a Western Kenyan population [LWK – Luhya]).
- There is also evidence that the 1000G imputed α-thalassaemia genotypes are not very reliable[3].

Therefore, to achieve the best possible outcome we decided to use our own data for creating both a reference dataset and for imputing missing genotypes. This entailed using a random sampling and bootstrap methodology (similar to the MRM and CART methods above). For a given run of IMPUTE2 and in line with the 1000 genomes population sizes (and the outcome of the MRM and CART methodologies above) we randomly selected 100 individuals (200 chromosomes) from our sample set to use as the reference set (known haplotypes [option *–h*]). These data were formatted in accordance with IMPUTE2 file policy. The remaining 2936 individuals' data had the α-thalassaemia data removed and identified as missing. These data were then saved in the appropriate IMPUTE2 format (option *–known_haps_g* given our data were already phased [Extended Data Section 1]). Other files required were; a legend file (option *–l*) for the polymorphisms in the dataset (SNP information file); a recombination map (option *-m*) which was taken from the African recombination map as described above and in Extended Data Section 1); a strand file (option *_strand_g*) identifying the orientation of each SNP (i.e. forward or reverse strand; all have already been orientated to forward); an interval for inference (option *–int* our region does not require 'chunking' and can be used in its entirety); an effective population size (option *–Ne*) for which we have used the default setting of 20,000 since we are not using the HapMap or 1000G populations.

An imputation run was then triggered using the following command line:

```
impute2
     -m <recombination map file>
     -h <reference haplotype set - 200 chr
(100 individuals)>
     -l <legend file>
     -known_haps_g <haplotypes with missing
data - 5872 chr (2936 individuals)>
     -strand_g <strand file>
     -int 84870 398421
     -Ne 20000
     -o <output file>
```

Once complete, the output files were inspected and the α-thalassaemia imputation genotype probabilities extracted from the *.output* file. We also stored the directly-typed genotypes of the samples alongside each imputation run.

We undertook 1000 runs of IMPUTE2 using the above method. Upon completion of each run, the data were processed to call α-thalassemia genotypes for each sample based on the IMPUTE2 probabilities using two thresholds (relaxed; 0.7 and strict; 0.9) to allow direct comparison to the typed genotypes. A thousand 'runs' were made and the performance was aggregated across all runs. We measured no-call rates and concordance rates from each of the runs. We also performed association testing, not only for the imputed data but for corresponding set of directly typed data; using the latter to account for variation in the reduced dataset (Extended Data Section 6).

## Results

We used genome-wide data derived from the Illumina HumanOmni2.5-4 genotyping array (Illumina, California, USA), that was collected within a case-control study conducted on the coast of Kenya as described in detail previously[3,12]. The current study included data from 3,036 of the children from the above study (1,432 severe malaria cases and 1,604 community controls)[12] for whom data were also available for α-thalassaemia genotypes, directly typed by PCR[19]. The key characteristics of the samples included in this current study are summarized in Table 1.

### Haplotype structure and LD in the *HBA* region

First, we used statistical phasing to construct haplotype maps for the 400kb region surrounding the -$\alpha^{3.7I}$ deletion. Briefly, we merged the directly typed α-thalassaemia genotype data with genome-wide SNP data (mapped to GRCh37) in a 10Mb region surrounding the -$\alpha^{3.7}$ allele with a view to capturing any potential long-range recombination structure. We phased haplotypes using SHAPEIT2 (see Methods and Extended Data Section 1). We then focused on a 400kb region (chr16: 84870-398421) (chr 16:~225818bp [Figure 1], avoiding the telomere), that included 160kb 5' and 173kb 3' flanking the -$\alpha^{3.7I}$ deletion. This 400kb region contained 178 SNPs (75 SNPs 5' and 103 SNPs 3' to α-thalassaemia) that had a minor allele frequency (MAF) of >1% in this dataset (Underlying Data1 Table BB_0-400kb_snp_details and Extended Data). Although statistical phasing is itself based on a model of haplotype copying which could be problematic for complex mutations, we noted that observed haplotypes were supported by the 423 (14%) homozygous carriers in our sample set (Table 1), suggesting that phasing was generally accurate.

To inspect haplotype structure, we clustered haplotypes separately into -$\alpha^{3.7}$ deletion carriers and non-carriers (Figure 7A and 7B, respectively). On first inspection the haplotypic structure of chromosomes containing the -$\alpha^{3.7I}$ deletion appeared more limited than that of ancestral chromosomes, but this was probably attributable to the different haplotype numbers (2,320 versus 3,752 distinct haplotypes, respectively; [Table 1 and Underlying Data2 Table DD]). Similarly, no obvious differences were seen in the degree of extended haplotype homozygosity (EHH)[35] between ancestral and derived haplotypes (Figure 7C–E and Underlying Data2 Table HH_EHH_values). The EHH peaks

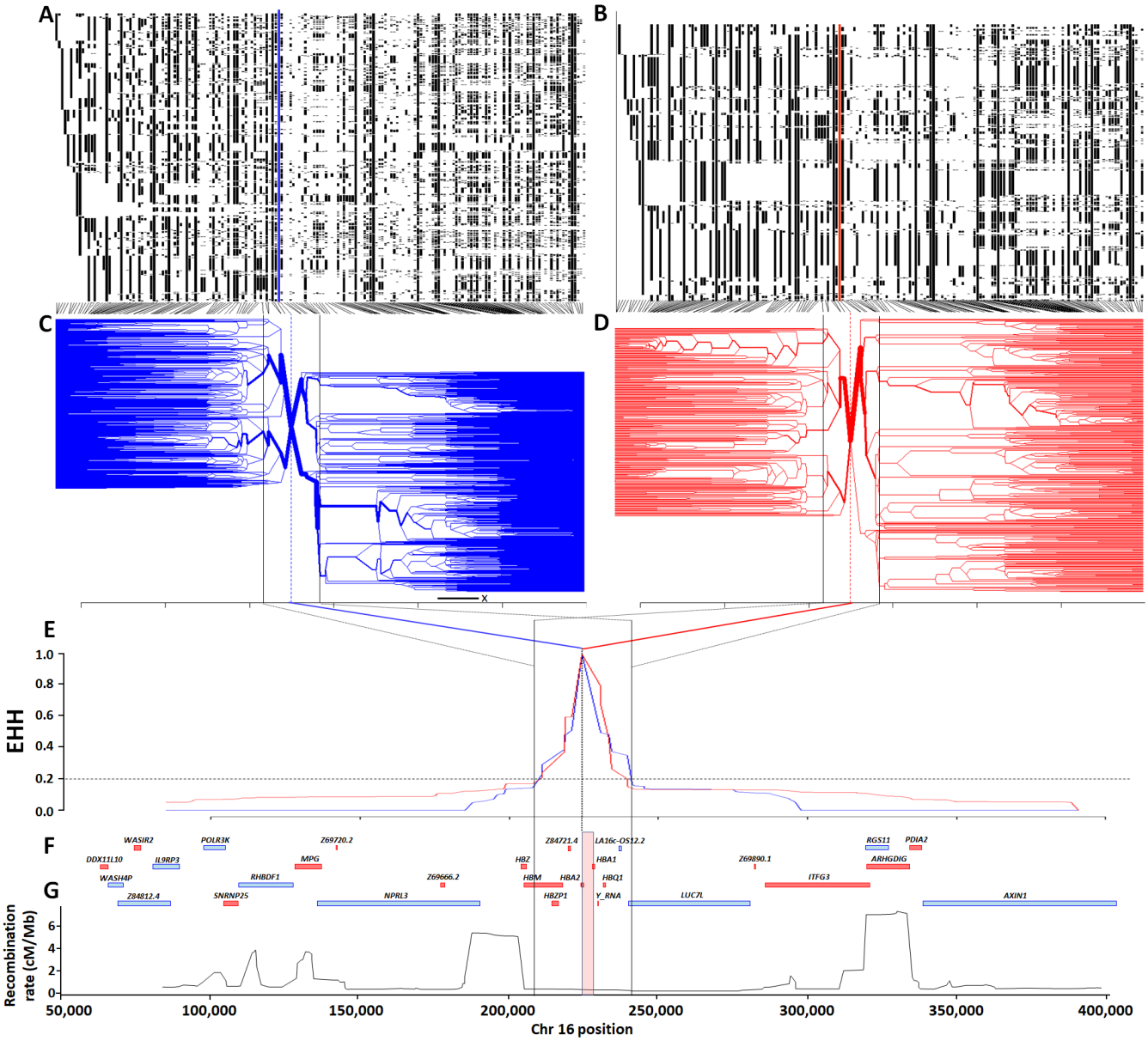



**Figure 7. Haplotype, extended haplotype homozygosity (EHH) and bifurcation diagrams for the *HBA* region in Kilifi, Kenya.** Panels **A** and **B** show haplotype maps for the $\alpha^{-3.7I}$-reference and $\alpha^{-3.7I}$-deletion haplotypes respectively (chromosomes are aligned as rows and SNPs in columns; white = reference and black = alternate), while panels **C** and **D** show corresponding bifurcation diagrams for the $\alpha^{-3.7I}$-reference and $\alpha^{-3.7I}$-deletion haplotypes respectively. Panel **E** shows EHH plots for $\alpha^{-3.7I}$-reference and $\alpha^{-3.7I}$-deletion haplotypes with the deletion allele as the focal point. Panels **F** and **G** show a gene map, and a recombination map based on African sequence data. The red and blue vertical lines in panels **C** and **D** denote the position of the $\alpha^{-3.7I}$–INDEL, as does the pink vertical box in panel **G** which is scaled to the width of the deletion.

of both the derived and ancestral chromosomes showed similar steep decreases to an EHH value of ~0.2, before reaching plateaus on either side of the deletion, each creating a similar region approximately 33kb around the $-\alpha^{3.7I}$ deletion (~16:207909-241210, Underlying Data2 Table HH_EHH_values and Extended Data Section 2). Fifteen SNPs present within this region (with the exception of the *HBZ* [ζ2] gene), span the α-globin gene cluster (*HBZP1* [ψζ1] to *HBQ1* [θ1]), which is bounded at the 5' end by a large 5.4 cM/Mb recombination peak (Figure 7G and Underlying Data2 Table CC_chr16_recombination) in the African recombination map data. However, we saw no such obvious recombination peak at the 3' end, the next large 7.3 cM/Mb peak being situated a further 50kb away from this region (Figure 7G and Underlying Data2 Table

CC_chr16_recombination). This latter peak corresponded with a noticeable change in haplotype bifurcation, suggesting that it might have affected the haplotype structure in both the ancestral and derived haplotypes (Figure 7C and D).

Dendrograms (drawn using the R package Pegas) of haplotypes built across the 400kb region or across the ~33kb region where little recombination was observed (Figures 8A and 8B, and Extended Data Figures 3–9) showed that haplotypes containing the -$\alpha^{3.7I}$ deletion were spread throughout the tree (branches marked in red in Figures 8A and 8B, and Underlying Data2 Tables DD_All_Haplotype_frequencies and EE_Core_Haplotype_frequencies). This was not simply explained by variation between ethnic groups (Figure 9, Extended Data Figures 3–9 and Underlying Data2 Table FF_Ethnic_group_haplotypes), as shown by the track in Figure 8 that shows that the ethnic groups were spread throughout the haplotype tree.

In total, we identified 1,457 haplotypes formed from 179 polymorphisms across the 400kb region, which comprised 1,005 and 452 distinct forms among WT and -$\alpha^{3.7}$ individuals, respectively (Underlying Data2 Table DD_All_Haplotype_frequencies). The haplotypes showed similar frequency distributions in WT and -$\alpha^{3.7}$ subjects, the maximum frequencies of any single haplotype being 2.5% and 8.3%, respectively. We identified 74 ancestral and 48 derived haplotypes within the 33kb core region, the majority having frequencies of <2.5% (Underlying Data2 Table EE_Core_Haplotype_frequencies). Three haplotypes that were present in both the ancestral and derived groupings had frequencies between 2.5% and 17.5%, while one -$\alpha^{3.7I}$ haplotype was observed at a frequency of 22.5%. Finally, we calculated diversity scores[18] which varied from 0.8-1.0 for all haplotype groups, and Tajima's D' statistic[36] which varied from 1.6-3.8 (Underlying Data2 Table GG_haplotype_metrics), showing that the ancestral and derived haplotypes were similar in structure and diversity across this 400kb region.

## The only genetic form of α-thalassaemia found in the Kilifi study population was -$\alpha^{3.7I}$

Given the extensive haplotype diversity around the -$\alpha^{3.7I}$ deletion, we used sequencing to confirm the type of $\alpha^{3.7}$ α-thalassaemia deletion that was present within our study population and to investigate the genetic structure of the region in further detail. We used Sanger sequencing of α-thalassaemia PCR reaction products[19] from twenty-five $\alpha^{-3.7I}$ and nine WT homozygous individuals. We first compared the sequence across the 3' gene region spanning the four restriction sites between the IVS2 sequence and the 3.7-R PCR primer Figure 2) that is commonly used to determine the type of $\alpha^{-3.7}$ deletion present[10]. In all twenty-five $\alpha^{-3.7}$ homozygous samples, the sequences and paralogous bases for both IVS2 and the four restriction sites (Table 3 and bases numbered 8–14 in Figure 2) matched the *HBA1* reference sequences (human genome reference *HBA1* sequence [GRCh37] and the sequences from the *HBA1*-specific

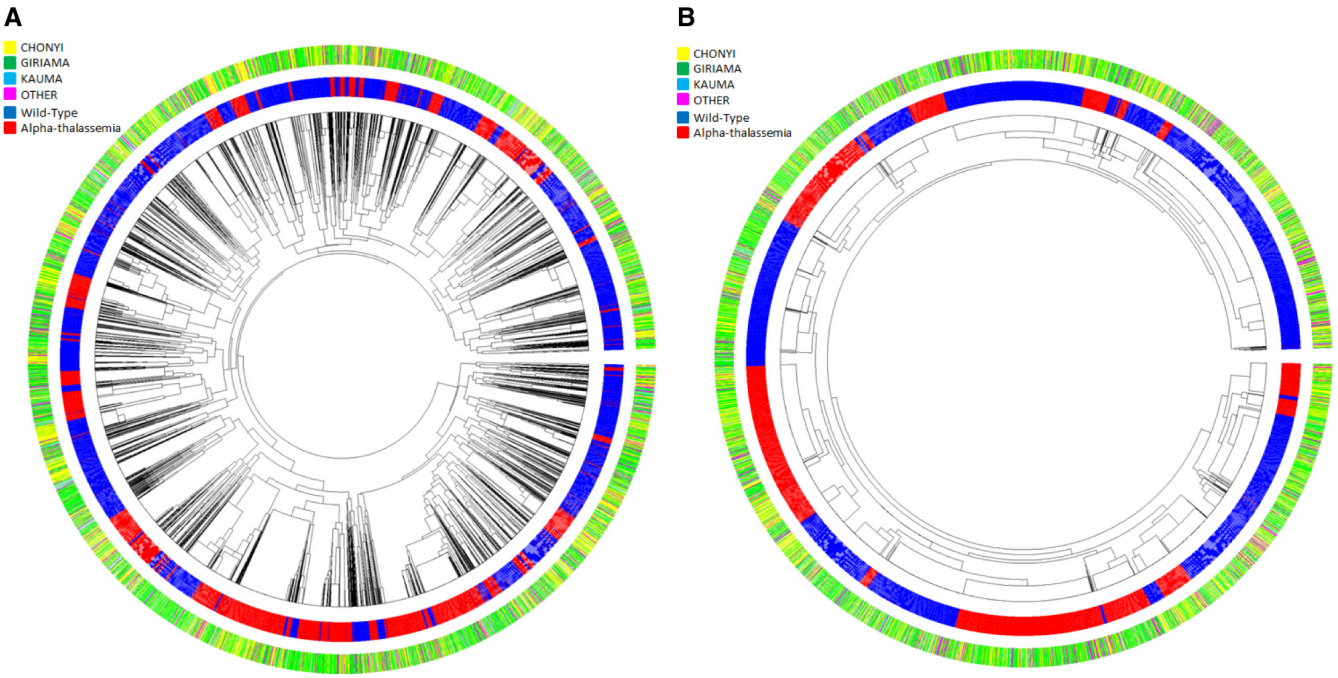



**Figure 8. Circular haplotype dendrograms for polymorphisms across the *HBA* region of chromosome 16 for 6072 chromosomes from coastal Kenyan individuals. A**: Haplotype dendrogram for the full haplotypes of 179 polymorphisms (chr16:84870-398421). **B**: Haplotype dendrogram for the core 33kb region surrounding the $\alpha^{-3.7}$ deletion (16:207909-241210). Centre shows the dendrogram; middle ring shows individual haplotypes (blue = wt, red = α-thalassaemia); outer ring shows the four major ethnic groupings [Yellow = Chonyi, Green = Giriama, Light Blue = Kauma and Magenta = 'other' groups.

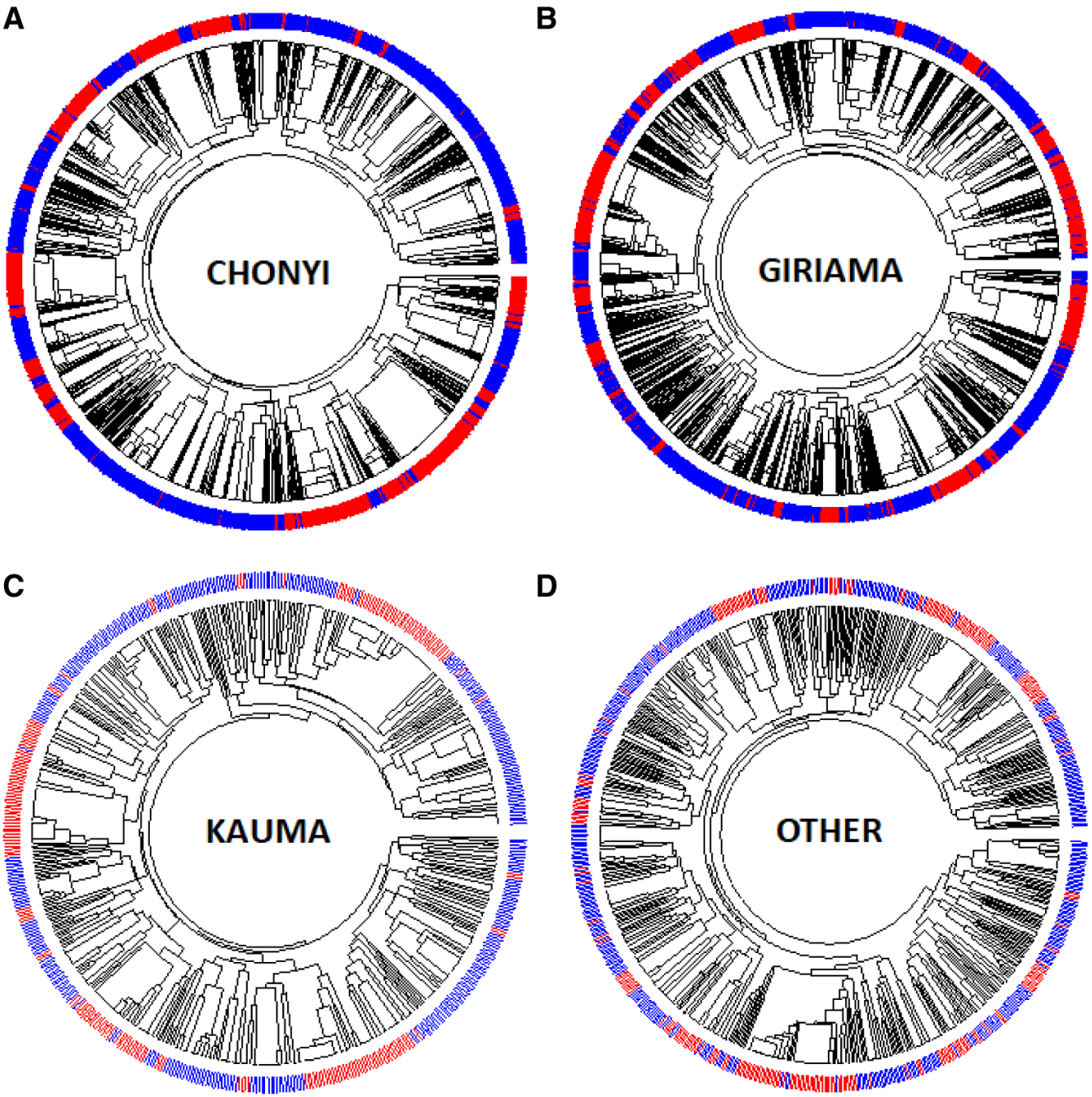



**Figure 9. Circular dendrograms of 'core' *HBA* regional haplotypes (Chr16:207,909-241,210 comprising 178 polymorphisms and the $\alpha^{-3.7I}$ polymorphism) by ethnicity in Kilifi, Kenya.** This core region is defined by the EHH signal > 0.2 (Extended Data Figure S2). **A**: Full haplotypes for 178 SNPs and the $\alpha^{-3.7}$-thalassaemia locus for the Kauma ethnic group. **B**: Full haplotypes for 178 SNPs and the $\alpha^{-3.7}$-thalassaemia locus for the Chonyi ethnic group. **C**: Full haplotypes for 178 SNPs and the $\alpha^{-3.7}$-thalassaemia locus for the Giriama ethnic group. **D**: Full haplotypes for 178 SNPs and the $\alpha^{-3.7}$-thalassaemia locus for other ethnic groups having too few haplotypes per group to display individually. $\alpha^{-3.7I}$-reference (BLUE) and $\alpha^{-3.7I}$-deletion (RED) haplotypes.

PCR amplicons from nine Kenyan control individuals). The amplicon region 5' to the IVS2 site spanning the remainder of the *HBA1/2* gene region matched the *HBA2* reference sequences. More specifically, the deletion-amplicon sequences matched the Kenyan *HBA2* sequences; we note that several paralogous base differences in the human reference sequence (Table 3 and bases numbered 4–6 in Figure 2) were identical and invariant between *HBA1* and *HBA2* in the nine Kenyan control samples. From these data we concluded that the $-\alpha^{3.7}$ deletion in these Kenyan samples was Type I[10], but there was also

evidence from variation observed for paralogous base 7 that there may be at least two breakpoint sub-types (Further details are given in Extended Data Section 3).

### A strong signal of association was seen between the -α$^{3.7I}$ allele and severe malaria, but no signals were seen at any of the SNPs in the surrounding region

We were interested to know whether the signal of association between the -α$^{3.7I}$ allele and severe malaria could be replicated by use of any of the SNPs within the surrounding genetic region. Consistent with our previous analyses of the full data set from this case-control study[12], we found strong evidence for associations between the -α$^{3.7I}$ deletion, typed directly by PCR, and all forms of severe *P. falciparum* malaria in this current sub-analysis. The adjusted odds ratio (aOR) for severe malaria overall was 0.78 (95% CI 0.70-0.87; p=11×10$^{-6}$) (Figure 10B and Underlying Data2 Tables II _assoc_results_unadjusted and JJ _assoc_results_adjusted_hbs), while on genotypic analysis, heterozygosity (-α/αα) and homozygosity (-α/-α) were associated with aORs of 0.79 (95% CI 0.67-0.93, P=0.005) and

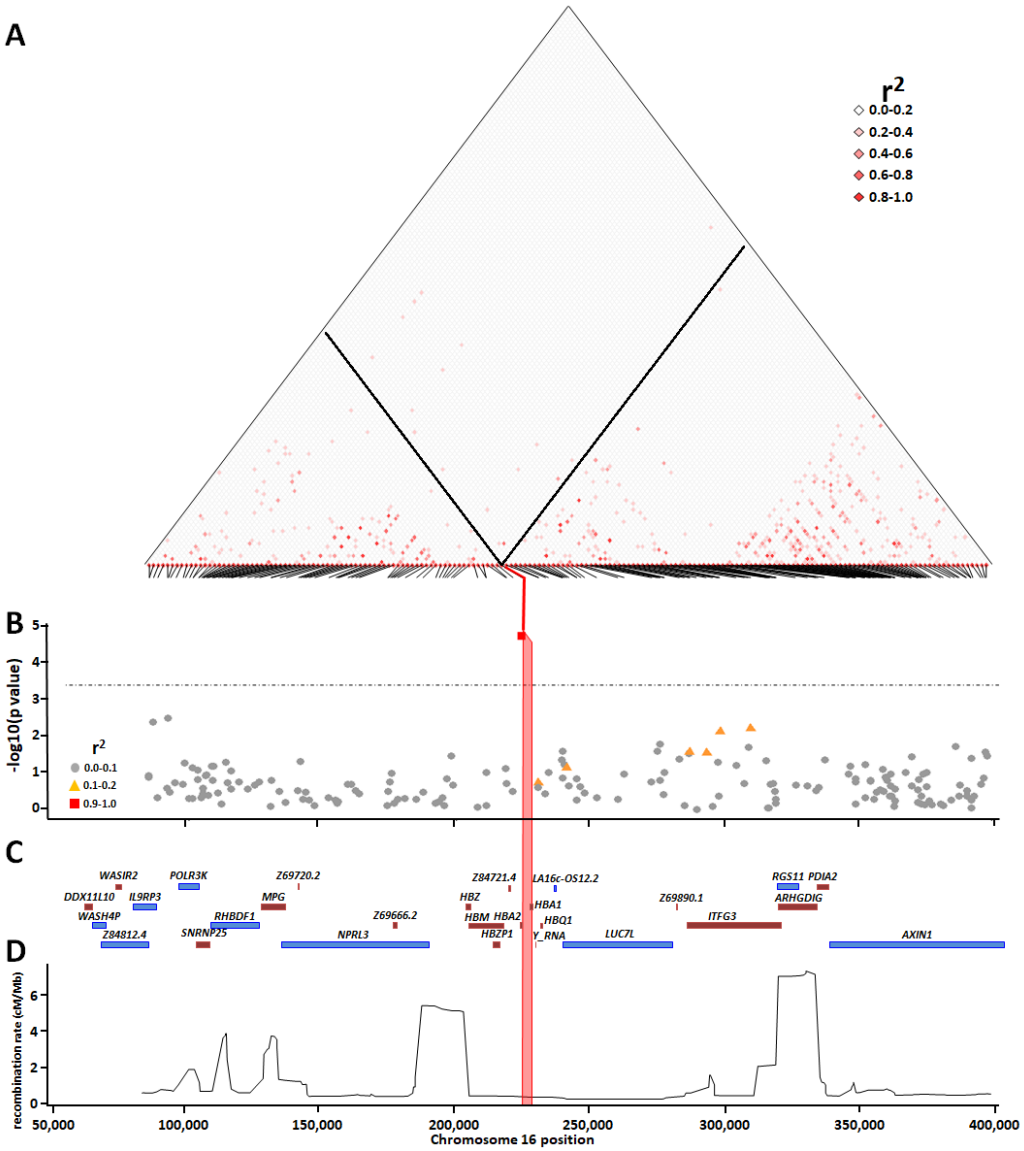



**Figure 10. The association between the α$^{-3.7I}$ deletion and SNPs within the surrounding region with severe malaria.** Panel **A** shows pairwise squared correlations ($r^2$) between genotypes at regional variants. Black lines denote the values for α$^{-3.7I}$ deletion for SNPs within the region. The angled lines at the base of the linkage disequilibrium map identify each SNP and where it aligns to the chromosome. Panel **B** shows a Manhattan plot illustrating the p-values for the associations between individual SNPs within the *HBA* region and severe falciparum malaria; the horizontal dotted line shows the Bonferroni-correction significance threshold (p<0.0003); $r^2$ shows the correlation between α-thalassaemia and the other SNPs. Panels **C** and **D** show the gene map and recombination map, respectively. The pink vertical box in panels **B**, **C** and **D** shows the location of α$^{-3.7I}$ deletion.

0.59 (95% CI 0.47-0.76, P=2.43×10$^{-5}$), respectively. Nevertheless, based on a Bonferroni-corrected significance threshold of P<0.0003, we found no significant associations at any of the 178 SNPs in the surrounding region (Figure 10B). Although two SNPs telomeric to the -$\alpha^{3.7I}$ deletion (rs62031426 [kgp4990237] and rs41340949 [kgp1941708]), were marginally associated (P=0.002 and 0.003, respectively), both were in low LD with the -$\alpha^{3.7I}$ deletion ($r^2$ < 0.01; Figure 10B and Underlying Data2 Table LL_pairwise_R2) and the weak associations at these SNPs were therefore unlikely to have been reflective of α-thalassaemia. As noted by others[2], the absence of associations at SNPs within this region is most likely explained by low LD (Figure 10A and Underlying Data2 Tables KK_pairwise_R, LL_pairwise_R2 and MM_pairwise_Dprime). The maximum $r^2$ values between -$\alpha^{3.7I}$ and any of the surrounding 178 SNPs being 0.081 (rs170058) and 0.16 (kgp4999044 [rs11863726]) in the 3' and 5' regions, respectively (Underlying Data2 Table LL_pairwise_R2).

### Predicting genotypes using haplotype structure with Impute 2

Given the availability within our dataset of both phased data, α-thalassaemia genotypes and haplotypes, we were able to investigate the prediction of α-thalassaemia genotypes using IMPUTE2 software[33]. This method requires a set of reference haplotypes that have been genotyped for α-thalassaemia. To evaluate imputation, we conducted a cross-validation experiment in which we repeatedly selected 100 individuals (200 chromosomes) at random from our total sample set to form a reference panel. We then imputed the α-thalassaemia genotypes for the remaining samples before calculating concordance with the directly typed results. We repeated this exercise 1,000 times with a view to estimating the imputation performance that would be expected if a reference panel for this population were available (Table 4, Extended Data Section 6, and Underlying Data2 Table AD_Impute2_overall_summary).

Overall, depending on the genotype and threshold, we were able to predict the true genotypes with sensitivity of 62–77%, specificity of 88–98% and positive prediction value (PPV) of 83–94% (Table 4 and Table 5) (further details in Extended Data Section 6). The correlation of the imputed genotype calls was 0.82 ± 0.03 and 0.84 ± 0.03 (mean ± SD) for the 70% and 90% genotype-calling threshold, respectively (using called samples only; 2,639 ± 120 and 2,193 ± 212 genotype calls, respectively [mean ± SD]) (Extended Data Figure 31 and Underlying Data2 Table AM_imputation_correlations). Together this correlation and reduced sample size will have the effect of reducing association power.

### α-thalassaemia genotype was correlated with intensity signals for SNPs within the -$\alpha^{3.7I}$ deletion

Given that we were unable to predict α-thalassaemia genotypes with sufficient reliability through the use of LD or imputation methods, we next investigated the potential for an alternative approach – the use of intensity data at SNPs that lie within the deleted region (Underlying Data2 Table OO_intensity_summary). We hypothesized that intensity signals for such SNPs would be reduced in α-thalassaemic subjects and that reductions would be dose-dependent in such a way that they would be greatest in homozygotes. We identified six features (rs2362744, rs2854120, rs4021971, rs4021965, rs11639532 and rs2858942) on the Illumina HumanOmni2.5-4 chip that lay within the -$\alpha^{3.7I}$ deletion, along with three flanking SNPs that served as controls (Figure 11 and Underlying Data2 Table OO_intensity_summary). For the purpose of these analyses, we summed the chip channel intensities (X + Y) to produce a single total intensity value for each individual at each of the six SNPs within the deletion and three flanking SNPs. These SNP data were then plotted as means with 25$^{th}$ and 75$^{th}$ percentiles and outliers on the y-axis with stratification by α-thalassaemia genotype on the x-axis (Figure 11, Extended Data Figures 13 and as scatter plots Extended Data Figure 14). As anticipated, we saw significant step-wise reductions in intensities by α-thalassaemia genotype for all six features within the deletion (Kruskall-Wallis and Dunn's test for multiple comparisons[21]; Extended Data Section 4, Underlying Data2 Tables PP_intensity_comparisons and QQ_intensity_summary), leading us to investigate a range of potential methods for predicting α-thalassaemia genotypes from such data.

### Signal intensities alone were not enough to predict α-thalassaemia genotype with any accuracy

Based on the assumption that most investigators would not have access to direct genotyping, we first created histogram and density plots of the intensity data from the nine SNPs described above, without reference to data on α-thalassaemia genotype (Figure 4). From these data we identified troughs or shoulders for four SNPs (rs2362744, rs2854120, rs4021971, rs11639532) as possible breaks between genotype classes (Figure 4 and Extended Data Table 3 and Underlying Data2 Table RR_intensity_cutoffs) and simply inferred α-thalassaemia genotypes from these groupings (Extended Data Figure 18 and Underlying Data2 Table RR_intensity_cutoffs). The overall ROC curves (Extended Data Figure 19) suggested good prediction of -$\alpha^{3.7I}$ homozygous individuals by this approach when using data from any of the four aforementioned SNPs (78–92% PPV, 93–95% sensitivity and 96–99% specificity; Table 5 and Extended Data Figure 19, Underlying Data2 Table SS_crude_genotype_assignment). However, such data were not useful in distinguishing heterozygotes from WT normal individuals, for which the PPVs were 62 and 93%; sensitivity between 24 and 95%; and specificity between 45 and 99% (Table 5 and Underlying Data2 Table SS_crude_genotype_assignment). In an extension to this method we used hierarchical clustering for the six SNPs described above in the deletion region described above (Figure 5 and Extended Data Figure 20 and Underlying Data2 Table TT_del_hier_cluster_assignment) and due to differences in intensity profiles we also investigated a reduced set of four or five SNPs (Figure 5 and Extended Data Section 4c). Genotypes were assigned based on the three primary tree branches in each model but did not improve performance for

**Table 4. Metrics from Imputation bootstrapping (1,000 runs).**

| metric | Number of correct genotypes | Total number of genotypes | No-call number | Correct Proportion of genotypes | Adjusted-correct proportion of genotypes | Missing proportion of genotypes | Wild-Type Homozygote | | | | $\alpha^{-3.7I}$ Heterozygote | | | | $\alpha^{-3.7I}$ Homozygote | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | sensitivity | specificity | accuracy | PPV | sensitivity | specificity | accuracy | PPV | sensitivity | specificity | accuracy | PPV |
| **70% threshold** | | | | | | | | | | | | | | | | | | |
| **Minimum** | 1645 | 2936 | 54 | 0.560 | 0.746 | 0.018 | 0.582 | 0.826 | 0.796 | 0.582 | 0.486 | 0.787 | 0.710 | 0.486 | 0.256 | 0.910 | 0.889 | 0.256 |
| **1st Qu** | 2173 | 2936 | 206.8 | 0.740 | 0.836 | 0.070 | 0.736 | 0.924 | 0.859 | 0.736 | 0.745 | 0.862 | 0.807 | 0.745 | 0.696 | 0.962 | 0.929 | 0.696 |
| **Median** | 2259 | 2936 | 274.5 | 0.769 | 0.854 | 0.093 | 0.770 | 0.938 | 0.872 | 0.770 | 0.776 | 0.877 | 0.827 | 0.776 | 0.755 | 0.970 | 0.938 | 0.755 |
| **Mean** | 2248 | 2936 | 296.2 | 0.766 | 0.852 | 0.101 | 0.765 | 0.934 | 0.871 | 0.765 | 0.772 | 0.875 | 0.825 | 0.772 | 0.747 | 0.968 | 0.937 | 0.747 |
| **SD** | 123.92 | | 120.15 | 0.040 | 0.025 | 0.040 | 0.050 | 0.020 | 0.020 | 0.050 | 0.050 | 0.020 | 0.030 | 0.050 | 0.090 | 0.010 | 0.010 | 0.090 |
| **3rd Qu** | 2334 | 2936 | 361.2 | 0.795 | 0.869 | 0.123 | 0.800 | 0.948 | 0.883 | 0.800 | 0.807 | 0.891 | 0.844 | 0.807 | 0.809 | 0.976 | 0.946 | 0.809 |
| **Maximum** | 2555 | 2936 | 1027 | 0.870 | 0.911 | 0.350 | 0.895 | 0.976 | 0.922 | 0.895 | 0.870 | 0.933 | 0.891 | 0.870 | 0.912 | 0.992 | 0.965 | 0.912 |
| **90% threshold** | | | | | | | | | | | | | | | | | | |
| **Minimum** | 1170 | 2936 | 215 | 0.399 | 0.771 | 0.073 | 0.453 | 0.851 | 0.754 | 0.453 | 0.302 | 0.829 | 0.639 | 0.302 | 0.134 | 0.932 | 0.876 | 0.134 |
| **1st Qu** | 1791 | 2936 | 589.8 | 0.610 | 0.861 | 0.201 | 0.640 | 0.940 | 0.833 | 0.640 | 0.597 | 0.902 | 0.760 | 0.597 | 0.517 | 0.976 | 0.918 | 0.517 |
| **Median** | 1938 | 2936 | 716 | 0.660 | 0.878 | 0.244 | 0.681 | 0.953 | 0.849 | 0.681 | 0.656 | 0.916 | 0.789 | 0.656 | 0.621 | 0.981 | 0.930 | 0.621 |
| **Mean** | 1919 | 2936 | 742.8 | 0.654 | 0.875 | 0.253 | 0.678 | 0.949 | 0.847 | 0.678 | 0.647 | 0.915 | 0.785 | 0.647 | 0.612 | 0.980 | 0.929 | 0.612 |
| **SD** | 194.51 | | 212.13 | 0.066 | 0.025 | 0.072 | 0.062 | 0.018 | 0.022 | 0.062 | 0.081 | 0.020 | 0.038 | 0.081 | 0.137 | 0.008 | 0.017 | 0.137 |
| **3rd Qu** | 2059 | 2936 | 880 | 0.701 | 0893 | 0.300 | 0.721 | 0.961 | 0.862 | 0.721 | 0.706 | 0.929 | 0.813 | 0.706 | 0.717 | 0.985 | 0.941 | 0.717 |
| **Maximum** | 2436 | 2936 | 1544 | 0.830 | 0.925 | 0.526 | 0.840 | 0.981 | 0.901 | 0.840 | 0.830 | 0.960 | 0.873 | 0.830 | 0.887 | 0.997 | 0.965 | 0.887 |

Adjusted-correct proportion was calculated for each run by adjusting the total number of genotypes by the number of no-called genotypes (Underlying Data2 Table AD_Impute2_overall_summary).



Wellcome Open Research 2020, 5:287 Last updated: 27 SEP 2021

**Table 5. Summary of the performance of methods using haplotypes or intensities to infer the $\alpha^{-3.7I}$ genotypes from features/SNPs within the $\alpha^{-3.7I}$ deletion region.**

| Methods | Principle | Directly typed (N) | Sensitivity*** | | | Specificity*** | | | Positive Predicative Value*** | | | Overall Association (OR [95%CI] p-value) | Model | Genotypic Association (OR) adjusted | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Norm | Het | Homo | Norm | Het | Homo | Norm | Het | Homo | | | Norm (ref) | Het | Homo |
| Direct genotyping | PCR genotyping | 3036 | NA | | | NA | | | NA | | | 0.78 (0.70-0.87) ($p=1.1x10^{-5}$) | Additive | 1 | 0.79 [0.67-0.93] (p=0.005) | 0.60 [0.47-0.76] ($p=2.43x10^{-5}$) |
| Intensity distributions (4 features)* | Intensity thresholds (AF3;XX) | 0 | 24.1 | 83.1 | 93.1 | 91.3 | 44.7 | 95.8 | 81.4 | 61.8 | 78.2 | 0.76 [0.64-0.9] (p=0.0013) | Homozygous or Heterozygous | 1 | 0.74 [0.62-0.88] ($p=7.2x10^{-4}$) | 0.83 [0.66-1.04] ($p=1.11x10^{-11}$) |
| | | | to | to | to | to | to | to | to | to | to | to | | | to | to |
| | | | 63.2 | 94.8 | 95.2 | 98.8 | 72.9 | 98.6 | 92.6 | 74.4 | 91.8 | 1.74 [1.42-2.13] ($p=5.8x10^{-8}$) | | | 1.87 [1.43-2.44] ($p=4x10^{-6}$) | 1.49 [1.06-1.98] ($p=2.1x10^{-9}$) |
| Hierarchical clustering (5 features) | Hierachical clustering (AF3;TT) | 0 | 56 | 74 | 96 | 95 | 71 | 86 | 80 | 73 | 88 | 1.43 [1.28-1.69] ($p=1.3x10^{-10}$) | Additive | 1 | 1.12 [0.91-1.37] (p=0.05) | 2.02 [1.6-2.55] ($2.2x10^{-9}$) |
| Hierarchical clustering (4 features) | Hierarchical clustering (AF3;TT) | 0 | 41 | 74 | 96 | 93 | 61 | 87 | 73 | 67 | 88 | 1.86 [1.56-2.22] ($p=4.98x10^{-12}$) | Recessive | 1 | 1.24 [1.02-1.49] (p=0.02) | 2.02 [1.62-2.51] ($p=2.5x10^{-10}$) |
| MRM* | Regression (AF3;VV) | 100 | 58 | 75 | 79.8 | 83.1 | 53.4 | 96.6 | 70.1 | 61.7 | 71.6 | 0.79 [0.68-0.92] (p=0.0025) | Heterozygous or Dominant | 1 | 0.8 [0.67-0.95] (p=0.01) | 0.81 [0.65-1.02] (p=0.08) |
| | | | to | to | to | to | to | to | to | To | to | to | | | To | to |
| | | | 73 | 86 | 85 | 89.6 | 75.3 | 98.6 | 79.2 | 76.7 | 90.1 | 1.67 [1.45-1.94] ($p=3.7x10^{-12}$) | | | 1.63 [1.37-1.95] ($p=3.5x10^{-8}$) | 2.61 [2.05-3.33] ($p=6.8x10^{-15}$) |
| CART* | Hierarchical clustering and regression (AF3;UU) | 100 | 68.1 | 45.5 | 68.6 | 76.7 | 61.3 | 96.7 | 62.4 | 61.4 | 67.6 | 0.799 [0.69-0.93] (p=0.004) | Heterozygous, or Dominant | 1 | 0.79 [0.67-0.93] (p=0.01) | 0.81 [0.64-1.02] (p=0.08) |
| | | | to | to | to | to | to | to | To | To | to | to | | | To | to |
| | | | 78.8 | 86.7 | 97.2 | 97.7 | 81.6 | 99.1 | 84.4 | 81.4 | 92.9 | 1.70 [1.47-1.97] ($p=1.19x10^{-12}$) | | | 1.72 [1.44-2.04] ($p=9.8x10^{-9}$) | 1.9 [0.85-1.39] (0.47) |
| IMPUTE2** 70% threshold (means ± SD) | Imputation (1000 runs) (AF3;AG/AI) | 100 | 77 | 77 | 75 | 93 | 88 | 97 | 87 | 83 | 94 | 0.74 ± 0.07 ($p = 1.71x10^{-3} \pm 3x10^{-3}$) | Additive, dominant or recessive | 1 | 0.79 ± 0.04 (p=0.02 ± 0.04) | 0.61 ± 0.04 (p=0.009 ± 0.003) |
| | | | ± | ± | ± | ± | ± | ± | ± | ± | ± | | | | | |
| | | | 5 | 5 | 9 | 2 | 2 | 1 | 2 | 3 | 1 | | | | | |
| IMPUTE2** 90% threshold (means ± SD) | Imputation (1000 runs) (AF3;AH/AJ) | 100 | 68 | 66 | 62 | 95 | 92 | 98 | 85 | 79 | 93 | 0.76 ± 0.08 (p = 0.005 ± 0.010) | Additive, dominant or recessive | 1 | 0.77 ± 0.05 (p=0.03 ± 0.06) | 0.61 ± 0.05 (p=0.004 ± 0.013) |
| | | | ± | ± | ± | ± | ± | ± | ± | ± | ± | | | | | |
| | | | 6 | 8 | 14 | 2 | 2 | 1 | 2 | 4 | 2 | | | | | |

Further details of methods and results can be found in the Extended Data.

* Where several SNPs (as indicated) were tested, then the ranges of results are shown.

** Data are mean ± SD for 1,000 individual runs of IMPUTE2 (70% and 90% refer to the threshold used to assign a genotype).

*** Data are shown as percentages.

MRM, Multiple-Regression Model; CART, Classification and Regression Trees.



Wellcome Open Research 2020, 5:287 Last updated: 27 SEP 2021

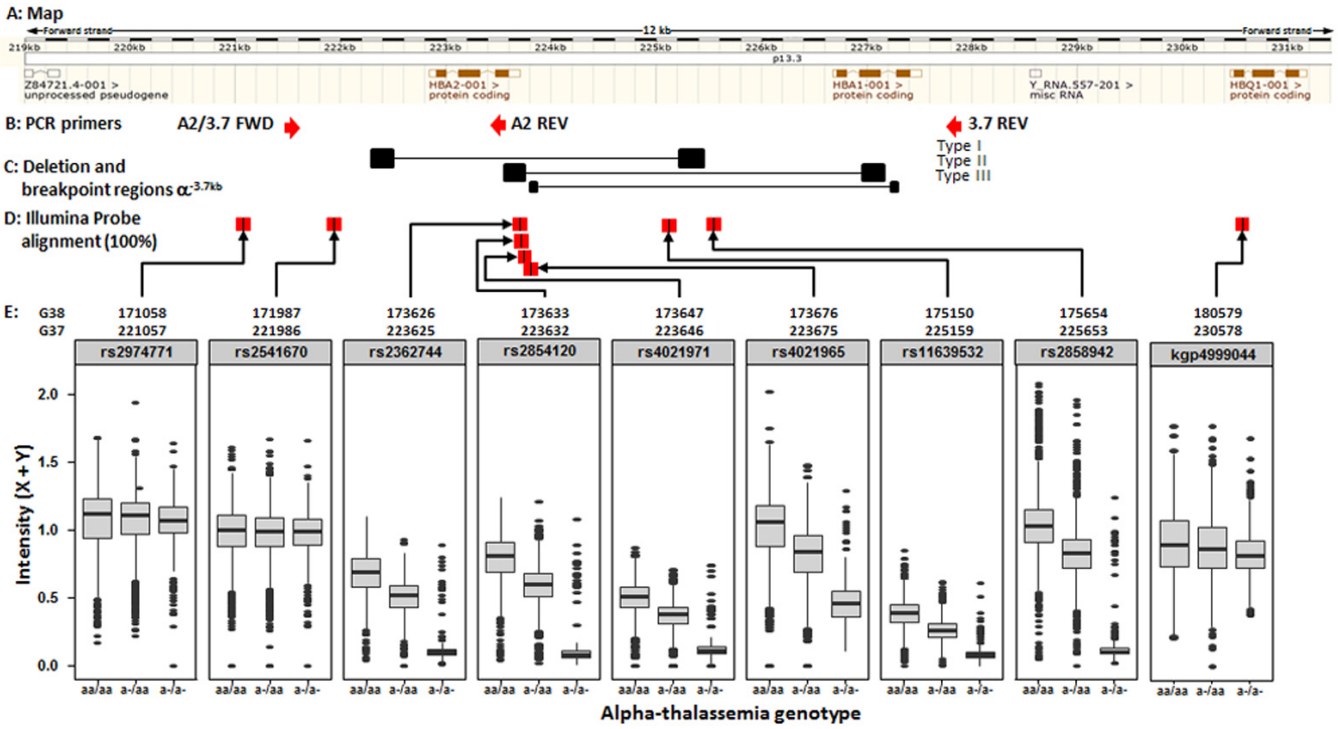



**Figure 11. Intensity plots of the Illumina 2.5M chip features across the α-3.7kb deletion. A**: Map of the Human *HBA2* and *HBA1* region on chromosome 16 (http://www.ensembl.org/index.html), with respect to the forward strand (GRCh37 coordinates). **B**: Primers used for the $\alpha^{WT}$ and $\alpha^{-3.7kb}$ deletion are shown by red arrows below the gene map (A2/3.7 FWD, A2 REV, 3.7 REV). **C**: The $\alpha^{-3.7}$ deletions are shown below the gene map and are highlighted according to Hill *et al.*[37] and data from this study (see main text). **D**: Feature probes with 100% match to the reference sequence are shown below the gene map by red boxes with a black vertical line. **E**: Coordinates for each feature are given for both GRCh37 and GRCH38 coordinates. **F**: Plots of the sum of the X + Y channel intensities (Y-axis) from the chip and PCR-typed genotype (X-axis) for each SNP (αα/αα WT; -α/αα -3.7kb HET, -α/-α -3.7kb HOM).

inferring α-thalassaemia genotypes (80%, 73% and 88% PPV for αα/αα –α/αα and -α/-α, respectively; Table 5 and Underlying Data2 Table TT_del_hier_cluster_assignment).

### α-thalassaemia genotype-prediction was not improved by using modelling approaches to interpret SNP intensity data

Given the poor predictive value of the raw intensity data as is, we investigated the potential utility of an alternative approach using models trained with a subset of directly genotyped samples. We tested two methods - a multinomial regression model (MRM) and a Classification and Regression Tree (CART) model - each of which used both the six SNPs within the deletion individually and all six SNPs combined. For both methods we used a bootstrap approach to test each sample selection 1,000 times using training sets of between 10 and 500 individuals as described in detail in Extended Data Section 5. We found that the minimum training sets that were required to reach a plateau of predicative power of the inferred versus real genotypes included just 100 samples, although the CART method appeared to produce a less stable prediction (Extended Data Figure 24). Depending on which SNP was included, α-thalassaemia homozygotes were predicted with sensitivities and specificities of 58–97% and 97–99%, respectively (Table 6, and Extended Data Figure 24). Predictions from models based on all six SNPs combined did not improve performance.

### We found no consistent concordance between association signals derived from direct and predicted α-thalassaemia genotypes

Finally, although the models above did not predict the true α-thalassaemia genotypes perfectly, we were interested to see whether they might still provide sufficient information to identify the malaria-protective associations that we saw by direct α-thalassaemia genotyping (Figure 10, summarised in Figure 12 and Table 5; and Underlying Data2 Tables JJ_assoc_results_adjusted_hbs, UU_crude_intensity_assoc, VV_del_hier_clustering_assoc, WW_MRM_association_results, XX_CART_association_results, AJ_imp_overall_assoc_70, AK_imp_overall_assoc_90, and AL_imp_overall_Pvals, Extended Data Figures 32, 33, Extended Data Tables 4, 5, 7, 8). Although we identified many significant associations across the predictive models for a number of SNPs, we saw no consistent pattern for either the best inheritance models or the direction of association, both of which were often at odds with those of the true associations (Figure 12 and Table 5). We did note, however, that for genotypic associations, rs11639532 (located in the intra-genic region between *HBA2* and *HBA1*) did give similar results to

**Table 6. Performance of the MRM and CART models for the prediction of α-thalassaemia genotype.**

| Method | SNPID | Genotypes | Sensitivity % (95%CI) | Specificity % (95% CI) | PPV % (95% CI) | NPV% (95% CI) |
|---|---|---|---|---|---|---|
| MRM | rs2362744 | αα/αα | 64.8 (62.0-67.7) | 87.6 (86.1-89.1) | 75.9 (73.4-78.1) | 80.7 (79.3-81.9) |
| | | –α/αα | 82.5 (80.5-84.5) | 71.5 (69.2-73.8) | 73.1 (71.5-74.7) | 81.4 (79.4-83.1) |
| | | –α/–α | 84.8 (81.0-88.2) | 98.5 (97.9-98.9) | 90.1 (86.9-92.5) | 97.5 (96.9-98.0) |
| | rs2854120 | αα/αα | 66.3 (63.5-69.1) | 89.6 (88.1-91.0) | 79.2 (76.8-81.5) | 81.6 (80.3-82.9) |
| | | –α/αα | 85.7 (83.8-87.5) | 75.6 (73.4-77.8) | 76.7 (75.1-78.3) | 85.0 (83.2-86.5) |
| | | –α/–α | 95.1 (92.7-97.0) | 98.6 (98.2-99.1) | 92.3 (89.5-94.4) | 99.2 (98.8-99.5) |
| | rs4021971 | αα/αα | 70.2 (67.4-72.9) | 87.3 (85.8-88.9) | 76.9 (74.6-79.1) | 83.1 (81.7-84.3) |
| | | –α/αα | 82.2 (80.2-84.2) | 74.7 (72.5-77.0) | 75.4 (73.7-77.0) | 81.8 (80.0-83.5) |
| | | –α/–α | 81.4 (77.4-85.1) | 98.3 (97.8-98.8) | 89.0 (85.6-91.6) | 97.0 (96.3-97.5) |
| | rs4021965 | αα/αα | 57.5 (54.5-60.5) | 86.9 (85.4-88.5) | 72.6 (69.9-75.0) | 77.4 (76.1-78.6) |
| | | –α/αα | 80.0 (77.9-82.1) | 53.4 (50.9-56.0) | 61.7 (60.3-63.2) | 74.1 (71.9-76.2) |
| | | –α/–α | 58.2 (53.5-63.1) | 97.5 (96.8-98.1) | 71.6 (65.8-76.8) | 90.5 (89.9-91.2) |
| | rs11639532 | αα/αα | 72.6 (69.9-75.2) | 88.6 (87.1-90.0) | 79.2 (77.0-81.3) | 84.4 (83.1-85.6) |
| | | –α/αα | 82.9 (80.8-84.8) | 74.4 (72.2-76.6) | 75.3 (73.6-76.9) | 82.3 (80.5-83.9) |
| | | –α/–α | 85.9 (71.6-89.0) | 98.1 (97.5-98.6) | 86.8 (83.2-89.8) | 96.1 (95.4-96.7) |
| | rs2858942 | αα/αα | 64.7 (61.8-67.5) | 88.2 (86.7-89.7) | 76.7 (74.2-79.0) | 80.7 (79.4-81.9) |
| | | –α/αα | 80.0 (77.9-82.1) | 53.4 (50.9-56.0) | 61.7 (60.3-63.2) | 74.1 (71.9-76.2) |
| | | –α/–α | 81.0 (53.5-83.1) | 97.5 (96.8-98.1) | 71.6 (65.8-76.8) | 90.5 (89.9-91.2) |
| | Combined | αα/αα | 66.0 (63.2-68.9) | 83.1 (81.4-84.9) | 70.1 (67.8-72.4) | 80.4 (79.0-81.7) |
| | | –α/αα | 75.0 (72.7-77.3) | 72.3 (70.0-74.6) | 71.8 (70.0-73.5) | 75.5 (73.7-77.3) |
| | | –α/–α | 79.8 (75.6-83.6) | 96.6 (95.9-97.3) | 79.8 (76.1-83.1) | 96.7 (96.0-97.2) |
| CART | rs2362744 | αα/αα | 68.1 (53.8-83.7) | 97.7 (77.2-98.9) | 62.4 (59.8-64.7) | 75.9 (74.4-77.1) |
| | | –α/αα | 72.5 (70.2-74.9) | 68.4 (66.1-70.8) | 68.7 (66.8-70.3) | 72.6 (70.6-74.3) |
| | | –α/–α | 97.2 (88.9-97.4) | 99.1 (98.3-99.2) | 92.7 (90.0-94.7) | 98.8 (98.2-99.1) |
| | rs2854120 | αα/αα | 76.5 (59.7-77.5) | 87.6 (85.0-88.2) | 73.9 (71.3-76.1) | 79.6 (78.2-80.7) |
| | | –α/αα | 85.7 (79.7-83.8) | 72.9 (70.6-75.2) | 74.2 (72.3-75.6) | 81.0 (79.2-82.7) |
| | | –α/–α | 97.1 (92.6-97.6) | 98.5 (98.0-99.0) | 91.7 (88.7-93.8) | 99.2 (98.8-99.5) |
| | rs4021971 | αα/αα | 78.8 (61.8-80.6) | 92.7 (90.4-94.0) | 82.6 (80.1-84.7) | 81.4 (80.0-82.5) |
| | | –α/αα | 86.7 (84.9-88.4) | 76.6 (72.6-77.8) | 76.3 (74.6-77.8) | 85.7 (83.9-87.2) |
| | | –α/–α | 96.5 (92.6-97.0) | 97.5 (96.9-98.1) | 86.3 (83.2-89.0) | 99.3 (98.8-99.5) |
| | rs4021965 | αα/αα | 69.0 (61.4-69.7) | 76.7 (74.7-78.7) | 62.4 (60.1-64.5) | 78.4 (76.7-79.6) |
| | | –α/αα | 45.5 (36.7-56.4) | 61.3 (57.8-63.8) | 61.4 (59.5-63.0) | 65.9 (63.8-67.6) |
| | | –α/–α | 68.6 (64.1-69.1) | 96.7 (96.0-97.4) | 67.6 (62.1-72.6) | 91.0 (90.3-91.7) |
| | rs11639532 | αα/αα | 75.0 (72.3-77.5) | 91.7 (90.4-92.9) | 84.4 (82.3-86.4) | 86.0 (84.7-87.2) |
| | | –α/αα | 85.2 (83.3-87.1) | 81.6 (79.6-83.5) | 81.4 (79.7-83.0) | 85.6 (83.8-87.0) |
| | | –α/–α | 97.2 (90.3-98.4) | 96.7 (95.9-97.4) | 82.2 (78.9-85.1) | 98.9 (98.4-99.2) |
| | rs2858942 | αα/αα | 60.8 (56.9-62.7) | 89.6 (88.2-91.0) | 77.7 (75.0-80.0) | 78.8 (77.6-80.0) |
| | | –α/αα | 82.3 (80.3-84.3) | 69.3 (66.9-71.6) | 71.8 (70.0-73.2) | 80.7 (78.8-82.4) |
| | | –α/–α | 95.7 (87.6-96.4) | 96.9 (96.2-97.6) | 82.9 (79.5-85.8) | 98.5 (97.9-98.9) |
| | Combined | αα/αα | 73.2 (70.6-75.9) | 80.1 (78.1-81.8) | 68.7 (66.6-70.8) | 83.3 (81.9-84.7) |
| | | –α/αα | 73.6 (71.2-75.9) | 80.3 (78.1-82.1) | 77.7 (75.9-79.5) | 76.3 (74.7-77.9) |
| | | –α/–α | 93.2 (90.6-95.6) | 98.2 (98.3-99.2) | 92.9 (90.0-94.8) | 98.9 (98.5-99.3) |

Metrics were calculated for α-thalassaemia genotypes inferred from MRM and CART models created using 100 randomly selected individuals as a 'training' set, and repeated 1000 times. PPV = Positive predictive value; NPV= negative predictive value; MRM, Multiple-Regression Model; CART, Classification and Regression Trees.

the directly typed data when using crude intensity cut-offs or MRM or CART models. Similarly, IMPUTE 2 performed reasonably well for genotypic association testing.

## Discussion

In common with many parts of sub-Saharan Africa, the -$\alpha^{3.7I}$ deletion is present at a high frequency in Kilifi, Kenya, where it not only protects against falciparum malaria, but also affects the risk of other childhood diseases[12,17,38]. Furthermore, in previous studies, we have shown that with regard to its effects on malaria susceptibility, the -$\alpha^{3.7I}$ deletion can also interact with mutations in unrelated genes[39–43]. Taken together, these observations suggest that α-thalassaemia is an important confounder that should be considered when interpreting disease-association studies in Africa. Despite this however, few studies correct for α-thalassaemia because of the current need for direct genotyping. Because an increasing number of GWAS studies are now being conducted in Africa, our main purpose in conducting this study was to investigate whether this might be achievable by the use of indirect GWAS-based data.

In the current study, we confirm the high frequency of α-thalassaemia in the Kilifi population and show that the -$\alpha^{3.7I}$ deletion is the only mutation responsible according to the current classification[10]. This was borne out using 25 samples from subjects with the homozygous deletion that we subjected to Sanger sequencing. While all showed the same sequence that was concordant with having the -$\alpha^{3.7I}$ same deletion type (Type 1[10]), we found evidence that this Type I form had been created by at least two separate breakpoint events, suggesting that the current classification warrants expanding. Our first aim was then to investigate why we had failed to identify the -$\alpha^{3.7I}$ signal in our previous GWAS studies of severe malaria[3,15]. Focusing on a 400kb region spanning the entire α-globin region, we found that LD, both overall and specifically with the -$\alpha^{3.7I}$ locus, was very low. This alone provided a good explanation that was further supported by our observations regarding the haplotype structure across the region. We found that the -$\alpha^{3.7I}$ deletion was distributed throughout the haplotype tree over this entire region, while EHH analysis showed that haplotype structure decayed quickly over ~16kb both 5' and 3' of the -$\alpha^{3.7I}$ locus and that this corresponded closely with the span of the α-globin genes (*HBZ* to *HBQ1*) and a large recombination peak centered on *HBZ*. Similarly, diversity measures were high, and equivalent in both the -$\alpha^{3.7I}$-containing and reference alleles. Our sequencing data, which suggested that there are at least two separate breakpoints creating the -$\alpha^{3.7I}$ type, provided one potential explanation for such haplotype diversity. These breakpoints occur within 1kb 5' to the IVS2 region. Across this region, we found very few differences in the human reference sequences and fewer in the control Kenyan sequences, which makes it difficult to determine more precisely where the breakpoints are or how many other breakpoint events may have occurred in the past.

Multiple lines of evidence now point to a model for the occurrence of -$\alpha^{3.7I}$ deletions, which leads to haplotype patterns that are refractory to imputation. First, it has been observed that -$\alpha^{3.7I}$ occurs with high mutation rates *in vitro*[5,6]. Second, compatible with this high mutation rate, haplotypes carrying -$\alpha^{3.7I}$ deletions occur across the genealogy of our sample of 3,036 individuals. Third, there is evidence from heterozygous yeast and *Arabidopsis* cells that large (from 100's bp to several kb) insertions/deletions can force the formation of loops of the non-paired DNA that can result in aberrant recombination and new haplotypes[44,45]. Further supporting this, we have directly observed evidence that segregating -$\alpha^{3.7I}$ deletions vary in the locations of their deletion breakpoints by up to 1kb. Given that the -$\alpha^{3.7I}$ carrying haplotypes are spread throughout the haplotype tree, this situation is particularly challenging for LD-based imputation approaches, which rely on matching of haplotypes at flanking genetic variants to infer the presence of mutations. The similarity between carrier and non-carrier haplotypes across the tree suggests this may not be easily solvable, even with data from additional reference panels.

Notwithstanding the fact that in agreement with observations from a recent study conducted in Tanzania[2], the sequences identified the Type I $\alpha^{-3.7}$ deletion event in all the -α/-α individuals samples investigated, we found no high LD signals with SNPs in the flanking regions. This makes detection of the α-thalassaemia signal by use of standard GWAS chip data difficult, and probably explains why despite finding a strong signal of association between malaria and α-thalassaemia when typed directly, we found no similar signals at any of the SNPs in the surrounding genetic region.

As a result of this initial conclusion, we next investigated a wide range of alternative approaches to the detection of α-thalassaemia that might be helpful in both measuring and correcting for its effects in African GWAS studies. Beginning with imputation and given the low LD within the region, we were somewhat surprised to find that we were able to predict the -$\alpha^{3.7I}$ deletion with positive-predictive values that ranged from 79–94%, depending on the genotype and statistical model. This was sufficient to allow us to detect an overall signal of association with severe malaria that was similar to that seen using direct genotyping, although the signal detected was considerably weaker, with p-values typically being between 100 and 1000 less significant than those derived from direct genotyping. To see if we could improve on these predictions, we next investigated a range of alternative approaches to typing based on GWAS SNP-chip-intensity data, with a focus on six SNPs from within the -$\alpha^{3.7I}$ deletion. We speculated that the intensity readings for such SNPs would be reduced in samples from affected individuals in a dose-dependent manner, being most marked in homozygotes. While in practice, we found that there was a large degree of overlap in these distributions between individuals of different genotype classes, and that as a consequence the intensity distributions were insufficiently distinct to allow for accurate prediction when interpreted without reference to gold-standard genotype data. Although the intensity distributions were relatively distinct among homozygotes, we found a particularly high degree of overlap in the values of αα/αα and -α/αα subjects, making it difficult to cleanly separate these genotypic groups on the basis of intensity values alone. We therefore attempted to improve our predictions using a variety of model-based approaches. Some of these models required a training set consisting of a small number of directly genotyped samples.

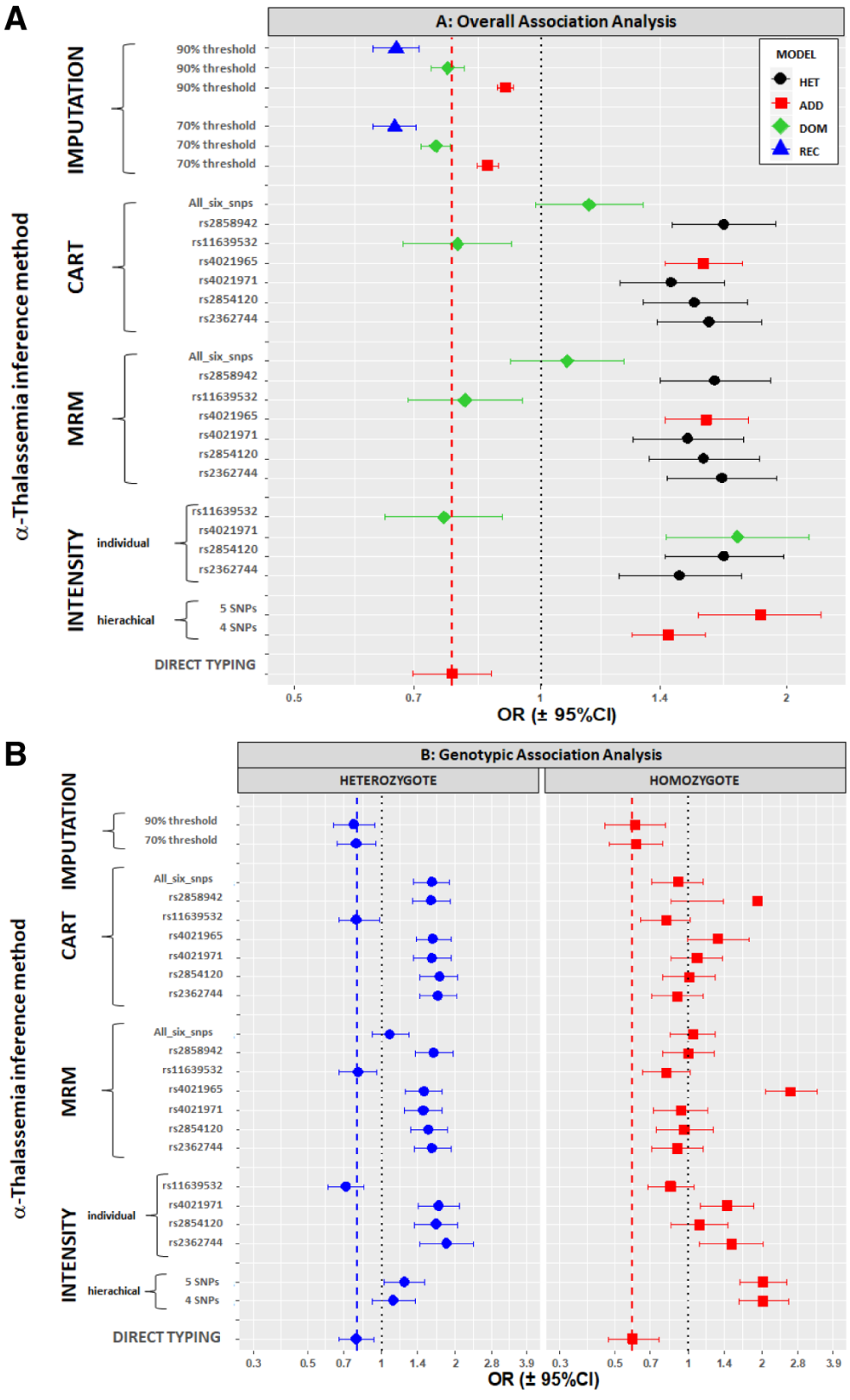



**Figure 12.** Association of α-thalassaemia with severe malaria by an overall test (**A**) and by genotypic tests (**B**). We used various methods to infer α-thalassaemia genotypes and then tested for association with severe malaria as detailed in the y-axis labels (see also main text). For the imputation results, these are mean results across the 1000 runs (see Methods); overall association results are split by the best association model results from each run while genotypic results are for all runs (Extended Data Section 6). The black dashed vertical line shows the no-effect position, while the red or blue vertical dashed lines show the direct typing effects.

Surprisingly we found that 100 samples was sufficient to train such models, potentially making it easier to provide these data for GWAS datasets. Nevertheless, while these approaches seemed encouraging, allowing us to predict the various α-thalassaemia genotypes with sensitivities of 75.6–97.8% and specificities of 85.0–99.2%, the predicted genotypes did not result in consistent signals of association with severe malaria. Although some gave results that were concordant with the true associations, others were highly variable depending on which SNP and model we used. While a number of signals were highly significant, the directions of effect were inconsistent and were often opposite to those derived from direct genotypes (i.e indicating a risk effect rather than protective effect). Moreover, the selection of 100 samples to create the training set was also a factor in the final outcome. These issues mean that including such intensity data to infer α-thalassaemia genotypes in the analysis of GWAS studies would not be helpful. Furthermore, we have explored just one form of α-thalassaemia deletion that is common across Africa as a single 'type'. In other parts of the world, this $-\alpha^{3.7}$ deletion occurs together with Type II and Type III deletions to varying degrees. The GWAS chip SNPs within the $-\alpha^{3.7}$ region may be distributed across the breakpoints depending on the $-\alpha^{3.7}$ type, giving a more complex intensity pattern. There are other α-thalassaemia deletions of varying sizes around the world that delete one or both *HBA1* and *HBA2*, some common, adding to this complexity. It may be that direct genotyping is therefore still the only solution to understanding the associations between this locus and malaria or other diseases.

## Data availability

### Source data

The Illumina dataset generated for the Kenya samples and analysed during this current study is registered and available at European Genome-phenome Archive under "Genome-wide study of resistance to severe malaria in eleven populations" (EGA study ID EGAS00001001311) with the Kenya data specifically in EGA Data Set: EGAD00010000904. These data have a managed data-access policy. Further details about the study, data and data-access can be found on the MalariaGEN website.

### Underlying data

Harvard Dataverse: Replication Data for: Haplotype Heterogeneity and Low Linkage Disequilibrium Reduce Reliable Prediction of Genotypes for the $\alpha^{-3.7I}$ form of α-thalassaemia Using Genome-Wide Microarray Data. https://doi.org/10.7910/DVN/YTXAHR.

This project contains the following underlying data:

- Underlying_DATA1.tab (Illumina 2.5MOmni chip features chr16)
- Underlying_DATA2.tab (Raw and analysis data)
- Underlying_DATA3.tab (Raw and analysis data for MRM and CART example runs)

### Extended data

Harvard Dataverse: Replication Data for: Haplotype Heterogeneity and Low Linkage Disequilibrium Reduce Reliable Prediction of Genotypes for the $\alpha^{-3.7I}$ form of α-thalassaemia Using Genome-Wide Microarray Data. https://doi.org/10.7910/DVN/YTXAHR.

This project contains the following extended data:

- Extended_Data.pdf (Supplementary information, figures, tables and Sanger sequencing alignments for *HBA1*, *HBA2* and the $\alpha^{-3.7kb}$ deletion in Kenyan samples)



### Code availability

Reproducibility code available from: https://www.well.ox.ac.uk/~gav/resources/compute.ld.R and an archived version is included in the Extended_Data.pdf file described above.

### Acknowledgments

We thank the patients and staff of Kilifi County Hospital and the KEMRI-Wellcome Trust Research Programme, Kilifi for their help with this study, and members of the Human Genetics Group in Kilifi for help with sample collection and processing. A full list of the Malaria Genomic Epidemiology Consortium members is provided in the Extended_Data.pdf file and at https://www.malariagen.net/projects/consortial-project-1/malariagen-consortium-members.

## References

1. Flint J, Hill AV, Bowden DK, *et al.*: **High frequencies of alpha-thalassaemia are the result of natural selection by malaria.** *Nature.* 1986; **321**(6072): 744–750.
PubMed Abstract | Publisher Full Text

2. Sepulveda N, Manjurano A, Drakeley C, *et al.*: **On the performance of multiple imputation based on chained equations in tackling missing data of the African alpha3.7 -globin deletion in a malaria association study.** *Annals of Human Genetics.* 2014; **78**: 277–289.

3. Malaria_Genomic_Epidemiology_Network, Band G, Quang SL, *et al.*: **Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania.** *Nat Commun.* 2019; **10**: 5732.
Publisher Full Text

4. Weatherall DJ, Clegg JB: **The thalassaemia syndromes**. Blackwell Scientific Publications, Oxford, 2002.
Publisher Full Text

5. Lam KW, Jeffreys AJ: **Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion.** *Proc Natl Acad Sci U S A.* 2006; **103**(24): 8921–8927.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Lam KW, Jeffreys AJ: **Processes of *de novo* duplication of human alpha-globin genes.** *Proc Natl Acad Sci U S A.* 2007; **104**(26): 10950–10955.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Williams TN, Deepak Gaur CECVSC (ed.),: **Host genetics**. In *Advances in Malaria Research.* John Wiley & Sons, Inc., Hoboken, New Jersey, in press, 2017; 465–494.
Publisher Full Text

8. Flint J, Harding RM, Boyce AJ, *et al.*: **The population genetics of the haemoglobinopathies.** *Baillieres Clin Haematol.* 1998; **11**(1): 1–51.
PubMed Abstract | Publisher Full Text

9. Piel FB, Weatherall DJ: **The α-thalassemias.** *N Engl J Med.* 2014; **371**(20): 1908–1916.
PubMed Abstract | Publisher Full Text

10. Higgs DR, Hill AV, Bowden DK, *et al.*: **Independent recombination events between the duplicated human alpha globin genes; implications for their concerted evolution.** *Nucleic Acids Res.* 1984; **12**(18): 6965–6977.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Williams TN, Maitland K, Ganczakowski M, *et al.*: **Red blood cell phenotypes in the alpha + thalassaemias from early childhood to maturity.** *Br J Haematol.* 1996; **95**(2): 266–272.
PubMed Abstract | Publisher Full Text

12. Ndila CM, Uyoga S, Macharia AW, *et al.*: **Human candidate gene polymorphisms and risk of severe malaria in children in Kilifi, Kenya: a case-control association study.** *Lancet Haematol.* 2018; **5**(8): e333–e345.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Macharia AW, Mochamah G, Uyoga S, *et al.*: **The clinical epidemiology of sickle cell anemia In Africa.** *Am J Hematol.* 2018; **93**(3): 363–370.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Band G, Le QS, Jostins L, *et al.*: **Imputation-based meta-analysis of severe malaria in three African populations.** *PLoS Genet.* 2013; **9**(5): e1003509.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Malaria_Genomic_Epidemiology_Network; Band G, Rockett KA, *et al.*: **A novel locus of resistance to severe malaria in a region of ancient balancing selection.** *Nature.* 2015; **526**(7572): 253–257.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Malaria_Genomic_Epidemiology_Network: **Reappraisal of known malaria resistance loci in a large multicenter study.** *Nat Genet.* 2014; **46**(11): 1197–1204.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Williams TN, Wambua S, Uyoga S, *et al.*: **Both heterozygous and homozygous alpha+ thalassemias protect against severe and fatal *Plasmodium falciparum* malaria on the coast of Kenya.** *Blood.* 2005; **106**(1): 368–371.
PubMed Abstract | Publisher Full Text

18. Nei M, Tajima F: **Genetic drift and estimation of effective population size.** *Genetics.* 1981; **98**(3): 625–640.
PubMed Abstract | Free Full Text

19. Chong SS, Boehm CD, Higgs DR, *et al.*: **Single-tube multiplex-PCR screen for common deletional determinants of alpha-thalassemia.** *Blood.* 2000; **95**(1): 360–362.
PubMed Abstract

20. Staaf J, Vallon-Christersson J, Lindgren D, *et al.*: **Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios.** *BMC Bioinformatics.* 2008; **9**: 409.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Zar JH: **Biostatistical Analysis**, 5th Edition. Pearson Prentice Hall, Upper Saddle River, NJ, 2010.
Reference Source

22. Dunn OJ: **Multiple Comparisons Among Means.** *Journal of the American Statistical Association.* 1961; **56**(293): 52–64.
Reference Source

23. Cohen J: **Statistical power analysis for the behavioral sciences (2nd ed.)**. Academic Press, New York, 1988.
Reference Source

24. Hedges LV, Olkin I: **Statistical methods for meta-analysis**. Academic Press, Orlando, FL, 1985.
Publisher Full Text

25. Cohen J: **A power primer.** *Psychol Bull.* 1992; **112**(1): 155–159.
PubMed Abstract | Publisher Full Text

26. Proudfoot NJ, Maniatis T: **The structure of a human alpha-globin pseudogene and its relationship to alpha-globin gene duplication.** *Cell.* 1980; **21**(2): 537–544.
PubMed Abstract | Publisher Full Text

27. Hess JF, Fox M, Schmid C, *et al.*: **Molecular evolution of the human adult alpha-globin-like gene region: insertion and deletion of Alu family repeats and non-Alu DNA sequences.** *Proc Natl Acad Sci U S A.* 1983; **80**(19): 5970–5974.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Warnes G, Bolker B, Bonebakker L, *et al.*: in press. 2019.

29. Schilling C, Mortimer D, Dalziel K, *et al.*: **Using Classification and Regression Trees (CART) to Identify Prescribing Thresholds for Cardiovascular Disease.** *Pharmacoeconomics.* 2016; **34**(2): 195–205.
PubMed Abstract | Publisher Full Text

30. Birney E, Soranzo N: **Human genomics: The end of the start for population sequencing.** *Nature.* 2015; **526**(7571): 52–53.
PubMed Abstract | Publisher Full Text

31. 1000 Genomes Project Consortium, Auton A, Brooks LD, *et al.*: **A global reference for human genetic variation.** *Nature.* 2015; **526**(7571): 68–74.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Sudmant PH, Rausch T, Gardner EJ, *et al.*: **An integrated map of structural variation in 2,504 human genomes.** *Nature.* 2015; **526**(7571): 75–81.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.** *PLoS Genet.* 2009; **5**(6): e1000529.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Howie B, Fuchsberger C, Stephens M, *et al.*: **Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.** *Nat Genet.* 2012; **44**(8): 955–959.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Sabeti PC, Reich DE, Higgins JM, *et al.*: **Detecting recent positive selection in the human genome from haplotype structure.** *Nature.* 2002; **419**(6909): 832–837.
PubMed Abstract | Publisher Full Text

36. Tajima F: **The effect of change in population size on DNA polymorphism.** *Genetics.* 1989; **123**(3): 597–601.
PubMed Abstract | Free Full Text

37. Hill AV, Wainscoat JS: **The evolution of the alpha- and beta-globin gene clusters in human populations.** *Hum Genet.* 1986; **74**(1): 16–23.
PubMed Abstract | Publisher Full Text

38. Wambua S, Mwangi TW, Kortok M, *et al.*: **The effect of α$^+$-thalassaemia on the incidence of malaria and other diseases in children living on the coast of Kenya.** *PLoS Med.* 2006; **3**(5): e158.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Wambua S, Mwacharo J, Uyoga S, *et al.*: **Co-inheritance of α$^+$-thalassaemia and sickle trait results in specific effects on haematological parameters.** *Br J Haematol.* 2006; **133**(2): 206–209.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Williams TN, Mwangi TW, Wambua S, *et al.*: **Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait.** *Nat Genet.* 2005; **37**(11): 1253–1257.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Opi DH, Swann O, Macharia A, *et al.*: **Two complement receptor one alleles have opposing associations with cerebral malaria and interact with α$^+$ thalassaemia.** *Elife.* 2018; **7**: e31579.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Opi DH, Ochola LB, Tendwa M, *et al.*: **Mechanistic Studies of the Negative Epistatic Malaria-protective Interaction Between Sickle Cell Trait and α$^+$ thalassemia.** *EBioMedicine.* 2014; **1**(1): 29–36.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Atkinson SH, Uyoga SM, Nyatichi E, *et al.*: **Epistasis between the haptoglobin common variant and α$^+$thalassemia influences risk of severe malaria in Kenyan children.** *Blood.* 2014; **123**(13): 2008–2016.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Kearney HM, Kirkpatrick DT, Gerton JL, *et al.*: **Meiotic recombination involving heterozygous large insertions in Saccharomyces cerevisiae: formation and repair of large, unpaired DNA loops.** *Genetics.* 2001; **158**(4): 1457–1476.
PubMed Abstract | Free Full Text

45. Sun X, Zhang Y, Yang S, *et al.*: **Insertion DNA Promotes Ectopic Recombination during Meiosis in Arabidopsis.** *Mol Biol Evol.* 2008; **25**(10): 2079–2083.
PubMed Abstract | Publisher Full Text

# Open Peer Review

## Current Peer Review Status: ? ✓

**Version 1**

Reviewer Report 28 July 2021

https://doi.org/10.21956/wellcomeopenres.17939.r44833



✓ **Manit Nuinoon**
School of Allied Health Sciences, Walailak University, Nakhon Si Thammarat, Thailand

The authors applied SNP genotypes spanning 400 kb of the entire a-globin gene cluster region from GWAS data for imputing 3.7 kb type 1 deletion (α-thalassaemia 2). The GWAS chip SNPs within the -$\alpha^{3.7I}$ deletion region distributed across the breakpoints resulting in a more complex intensity pattern and low linkage disequilibrium (LD). Therefore, it is challenging to apply GWAS chip SNPs for imputing α-thalassaemia. The authors used Sanger sequencing as direct genotyping (reference method) to confirm the form of α-thalassaemia and SNP genotype spanning around -$\alpha^{3.7I}$ deletion region. Several models were constructed to predict α-thalassaemia genotypes according to single SNP or combined SNPs resulting in different predictive performances. This paper is well written and addresses the detail of each step, and the study design was apparent. My suggestions are as follows:

1. According to the MRM and CART models (Table 6), is it possible to select the same ethnicity for calculating the predictive performance?

2. ROC curve (AUC calculation) is recommended for depicting the predictive performance of single SNP or combined SNPs.

3. Regarding the α-globin gene cluster is located close to the telomere, please put more information about how the telomeric region of this gene cluster affects a more complex intensity pattern in the discussion part.

4. Regarding the main text and several figures, two formats of the 3.7 kb type 1 deletion are found in this article (-$\alpha^{3.7I}$ deletion and $\alpha^{-3.7I}$ deletion). Would you please use the same format throughout the manuscript (commonly-used format is -$\alpha^{3.7I}$ deletion)?

5. According to the title of Table 3, the number of normal individuals should be changed from 5 to 9?

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Thalassaemia and hemoglobinopathies

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 26 Aug 2021

**Kirk Rockett**, University of Oxford, Oxford, UK

We thank the reviewers' for their positive and helpful feedback and are pleased to make revisions accordingly as per our comments to their questions below.

We have therefore added further detail to our manuscript with positional references to the published PDF version.

**1. According to the MRM and CART models (Table 6), is it possible to select the same ethnicity for calculating the predictive performance?**

Thank you. We did not separate the different ethnicities for the MRM and CART as these methodologies rely on the intensity pattern of the SNP features rather than the surrounding genetic structure and therefore should not be affected by a person's origin (unless it is sample related in some way). Rather the data should be affected more by experimental methodology, reagents, samples and the microarray. Moreover, one has to balance this with the situation where some ethnic groups may have insufficient numbers to make any methodology viable, particularly when using the data to generate a reference panel as well (as in our group labelled OTHER which is a mixture of 18 different self-reported groups each with too few individuals to ethically identify them or provide sufficient power alone. Obviously one has to review the results with that in mind).

However for this very reason, amongst others, we have made all the data available to allow it to be tested by interested parties.

This comment also hints at the main problem with microarray data and that is its quality, whether due to the specificity of the individual array features (for example the feature for SNP rs2541670 [Extended Data section 4] and main text figure 4) may behave differently in WT chromosomes versus deletion chromosomes, or sample quality (this data set did have 762 sample excluded for various reasons during its initial QC [see methods]) and even the distribution of SNP intensities across the genome that can affect the genotype calling confidence (as alluded to in point 3 below and in 3 new added figures).

**2. ROC curve (AUC calculation) is recommended for depicting the predictive performance of single SNP or combined SNPs.**

Thank you and all this data is included in the Extended Data file and data package.

**3. Regarding the α-globin gene cluster is located close to the telomere, please put more information about how the telomeric region of this gene cluster affects a more complex intensity pattern in the discussion part.**

Thank you and this is clearly an omission on our part. We have now provided new figures in the Main and Extended Data (Main Figure 11 and Extended Data Figures 12-14) along with accompanying text in both the Main Text and Extended Data text.

***Main Text: page 18, section; 'α-thalassaemia genotype was correlated with intensity signals for SNPs within the -α3.7I deletion'***

*"Given that we were unable to predict α-thalassaemia genotypes with sufficient reliability through the use of LD or imputation methods, we next investigated the potential for an alternative approach – the use of intensity data at SNPs that lie within the deleted region (Underlying Data2 Table OO_intensity_summary). For the purpose of these analyses, we therefore summed the chip channel intensities (X + Y) to produce a single total intensity value for each individual across each SNP across chromosome 16. We first plotted these intensities averaged over 100kb bins for the whole chromosome (Figure 11A) showing that there was an end-of-chromosome reduction in intensities (from ~ 1.25 to ~ 0.75). At the individual SNP level the mean intensity for the 0-1Mb 5' telomere region where the HBA region lies (Figure 11B) was ~0.75-1.*

*NEW FIGURE 11 showing the SNPs intensities across the full chromosome and within the 5' telomere and HBA regions.*

*The HBA region lies between 200 and 235kb from the 5' telomeric end where the mean intensities were ~ 0.75 (Figure 11B and C, and Extended Data Figure 12-14). While this may be sufficient to have some impact on the predicative performance of the data we We hypothesized that intensity signals for such SNPs within the deletion region would still be reduced in α-thalassaemic subjects and that reductions would be dose-dependent in such a way that they would be greatest in*

*homozygotes. We identified six features (rs2362744, rs2854120, rs4021971, rs4021965, rs11639532 and rs2858942) on the Illumina HumanOmni2.5-4 chip that lay within the -α$^{3.7I}$ deletion, along with three flanking SNPs that served as controls (Figure 12 and Underlying Data2 Table OO_intensity_ summary). For the purpose of these analyses, we summed the chip channel intensities (X + Y) to produce a single total intensity value for each individual at wach of the six SNPs within the deletion and three flanking SNPs. These Sum(X+Y) data for each individual and each of the 6 SNPs data were then plotted as means with 25 th and 75 th percentiles and outliers on the y-axis with stratification by α-thalassaemia genotype on the x-axis (Figure 12, Extended Data Figures 16 and as scatter plots Extended Data Figure 17). As anticipated, we saw significant step-wise reductions in intensities by α-thalassaemia genotype for all six features within the deletion (Kruskall-Wallis and Dunn's test for multiple comparisons[21] ; Extended Data Section 4, Underlying Data2 Tables PP_intensity_comparisons and QQ_intensity_summary), leading us to investigate a range of potential methods for predicting α-thalassaemia genotypes from such data."*

**Extended Data Section 4, new introductory paragraph; Illumina Chip Intensities**

*"We first extracted the intensities for all features from the chromosome 16 vcf file for the 3036 samples used in this study and then plotted the mean of the Sum(X+Y) channels for these samples for each SNP against chromosomal position (Figure 12A). This showed that the feature average intensities across the chromosome varied from ~0.25 to 2. In the smoothed plot and 100kb windowed plots (Figure 12B, C and D) it can be seen that the general intensity across the chromosome is ~1.25 while at the telomeres (start-10Mb and 85Mb to end), the intensities drop gradually to ~0.75 and 1 respectively. This is not unusual for such microarray data as was noted in the MalariaGEN study[3] that generated this GWAS data. Focussing in on the region of interest, Figure 13 shows the first 1Mb region at the 5' telomere and Figure 14 the region across the HBA gene cluster (200kb to 235kb). These plots show the variation between SNP intensities, which are all reduced (~0.75) in comparison with the central region (~1.0) of the chromosome intensities. Notwithstanding this, it is important to review the distribution of sample intensities for each SNP."*

**4. Regarding the main text and several figures, two formats of the 3.7 kb type 1 deletion are found in this article (-α$^{3.7I}$ deletion and α$^{-3.7I}$ deletion). Would you please use the same format throughout the manuscript (commonly-used format is -a$^{3.7I}$ deletion)?**

Thank you and we have checked, corrected, and standardised throughout the manuscript depending on the context used. When referring to the 3.7kb deletion collectively or where the type is unknown we use -α$^{3.7}$, while the suffix I is added when specifically referring to the Type I variant (and similarly for Type II and Type III).
These have been corrected accordingly through both the Main text and Extended_Data text.

**5. According to the title of Table 3, the number of normal individuals should be changed from 5 to 9?**

Thank you, this has been done.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 20 July 2021

https://doi.org/10.21956/wellcomeopenres.17939.r44831



?

**Karen G Scheps**
[1] Departamento de Microbiología, Inmunología, Biotecnología y Genética, Cátedra de Genética, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Buenos Aires, Argentina
[2] CONICET, Buenos Aires, Argentina

GWAS studies are a powerful tool to contribute to the knowledge of the bases of common and rare diseases. However, many studies do not contemplate the effects of CNVs, which could bias the obtained results. This matter is particularly relevant when studying the association of risk of severe malaria in African populations since α-thalassemia (thal) mutations are prevalent and the long ancestral history of this group has led to low linkage disequilibrium.

The authors proposed a very straightforward analysis flow to infer α-thal genotypes based on the data obtained by GWAS –based approaches. The design was very clear and the partial results obtained in each analysis provide valuable information for the field since it could potentially be applied for retrospective analysis of other GWAS-obtained data.

I have a few questions for the authors and I believe that the answers could be included in the manuscript as well.

1. Could it be possible to analyze the GWAS-obtained data investigated in this study with any of the platforms developed for genomic profiling of CNVs, such as Beadstudio or PennCNV?

2. For the approach of imputation using haplotypes, it is understandable why 1000G reference panel should not be used. Would you consider using the African/African-American genomes included in gnomAD v3.1?

3. Regarding the Chip Intensity (sum[X and Y] channels) density plots featured in Figure 4, is it possible that the lack of troughs or shoulders for SNPs rs4021965 and rs2858942, altered the chances of discriminating the different α-thalassemia (thal) genotypes with this

approach? When analyzing the intensity of these SNPs in the different α-thalassemia (thal) genotypes it would seem that at least the homozygous state could be discriminated

**Is the work clearly and accurately presented and does it cite the current literature?**
Yes

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Genetics of hemoglobinopathies

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 26 Aug 2021

**Kirk Rockett**, University of Oxford, Oxford, UK

We thank the reviewers' for their positive and helpful feedback and are pleased to make revisions accordingly as per our comments to their questions below.

We have therefore added further detail to our manuscript with positional references to the published PDF version.

**1. Could it be possible to analyze the GWAS-obtained data investigated in this study with any of the platforms developed for genomic profiling of CNVs, such as Beadstudio or PennCNV?**

Thank you for this suggestion. This was certainly something that we had considered but we wanted to make the main focus of the manuscript about the haplotype structure and association with malaria rather than an exhaustive comparison of methodologies. Our study was therefore not designed to try and compare and contrast all available methods,

nor identify any failings or provide suggestions for correcting them, but rather provide a description of the intensity data in the region and along with haplotype structure try and provide some explanation why prediction/imputation methods may struggle to properly call the alpha-thalassaemia locus, and therefore drawing the wrong conclusions about the effect and significance in malaria. We do provide details of where all the relevant data can be found for interested parties to test these methods of their own designs.

We also note that this region has already been commented on using the data from the 1000 genomes study where such methods were used to call the alpha-thalassaemia locus in several populations with similar findings to our current manuscript (Malaria Genomic_Epidemiology_Network, Band G, Quang SL, et al. (2019) Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nat Commun.; 10: 5732).

**Main Text Page 6, new introductory paragraph for section; 'Inferring α-thalassaemia genotypes from intensity data'**

*"Given that we know the location and breakpoints for the α-thalassaemia deletion under investigation, we have looked at the SNPs and intensities directly to determine in an a priori fashion how well they infer the genotypes for the -α$^{3.7}$ deletion. We note that other predicative/inference methods are available in addition to those we have used, some designed to identify/discover non-SNP features using intensity data. Accordingly, we have provided all the information and data necessary for others to test alternative methodologies if interested. "*

**Main Text Page 23, Discussion paragraph 5; we have modified and added a section at the beginning of this paragraph to comment on our choice of methodologies for looking at the imputation of genotype from intensity data.**

*"As a result of this initial conclusion, and knowing the location, type, and extent of the deletion, we next investigated a range of alternative approaches to describe and detect the α-thalassaemia deletion using our microarray intensity data, that might be helpful in both measuring and correcting for its effects in African GWAS studies. We note that a number of methodologies exist to detect such features genomewide* de novo *and that have been used to impute INDELS into reference panels. Moreover, this was commented on in the supplementary data of a previous study[3] where the 1000 genomes imputation reference panel was used (the -α$^{3.7}$ deletion is imputed as panel ID: EM_DL_DEL34404), in which the authors found a reduced confidence in predicting the deletion genotypes. Thus, rather than provide an exhaustive comparison of methods, we focussed on methodologies directly using the data for features known to be within the deletion sequence or using the genetic structure of the our population. "*

**2. For the approach of imputation using haplotypes, it is understandable why 1000G reference panel should not be used. Would you consider using the**

**African/African-American genomes included in gnomAD v3.1?**

We agree that other reference panels may be more appropriate to use and we did allude to this in our text, but have now made this clearer. However, through our work with MalariaGEN we note that even that may not be sufficient. Furthermore from our data where we are using the very same population and data as both reference panel and test data, the predictions/imputations were still not sufficient to provide complete confidence in the analyses compared with using directly typed data. We have added to the comments in our discussion on this point (see point 1 above), and to the methods section. However we have made all the data available and therefore this can be tested given an individual's data and circumstances. But our results do suggest caution in that even with some level of correctly predicted/imputed genotypes an incorrect outcome may result with a high level of significance leading to spurious conclusions.

**Main Text Page 13, and Extended Data page 52, section; 'Imputation using Haplotypes'**

*"IMPUTE2 requires a reference set of haplotypes to infer missing genotypes. However, we decided not to use the 1000 genome reference panel for several reasons;*

- *The 1000G dataset does not contain any directly typed α-thalassaemia deletion variants, only imputed from the sequence data.*
- *The 1000G dataset does not have any populations that directly match the populations used in our study (the closest is a Western Kenyan population [LWK – Luhya]).*
- *There is also evidence that the 1000G imputed α-thalassaemia genotypes are not very reliable[3].*

*Other reference panels are becoming more widely available, some of which also include additional populations of African origin. While we decided not to use these for similar reasons to those described above for the 1000 Genomes data, we do not rule out the possibility that some could prove more suitable for imputing haplotypes into Kenyan populations.*

*Instead we used our own data for creating both a reference dataset and for imputing missing genotypes, as this should have provided the best opportunity to predict/impute the α-thalassaemia genotypes."*

**3. Regarding the Chip Intensity (sum[X and Y] channels) density plots featured in Figure 4, is it possible that the lack of troughs or shoulders for SNPs rs4021965 and rs2858942, altered the chances of discriminating the different α-thalassaemia (thal) genotypes with this approach? When analyzing the intensity of these SNPs in the different α-thalassaemia (thal) genotypes it would seem that at least the homozygous state could be discriminated**

Thank you, yes that is exactly what we believe and tried to describe. We have reviewed our text and re-worded the appropriate text to make these points clearer.

***Main Text Page 18, section; 'Signal intensities alone were not enough to predict α-***

***thalassaemia genotype with any accuracy'***

*"Based on the assumption that most investigators would not have access to direct genotyping, we first created histogram and density plots of the intensity data from the nine SNPs described above, without reference to data on α-thalassaemia genotype (Figure 4). From these data we identified troughs or shoulders to separate the different genotype classes; SNPs rs4021965 and rs2858942 did not show a trough or shoulder to help identify heterozygotes from WT normal individuals, although they did show some discrimination of the homozygotes. But four SNPs (rs2362744, rs2854120, rs4021971, rs11639532) did show possible breaks between each genotype class (Figure 4 and Extended Data Table 3 and Underlying Data2 Table RR_intensity_cutoffs) and we focused on these to infer α-thalassaemia genotypes from these groupings ..."*

***Competing Interests:*** No competing interests were disclosed.

Author Response 26 Aug 2021

**Kirk Rockett**, University of Oxford, Oxford, UK

We thank the reviewers' for their positive and helpful feedback and are pleased to make revisions accordingly as per our comments to their questions below.

We have therefore added further detail to our manuscript with positional references to the published PDF version).

My apologies but this is an additonal revision based on reviewer #1's comment #2 ommitted from my original post.

**2. For the approach of imputation using haplotypes, it is understandable why 1000G reference panel should not be used. Would you consider using the African/African-American genomes included in gnomAD v3.1?**

**Main Text Page 25, Final section of the discussion.**
"*...Similarly, although an ever increasing number of 'relevant' reference panels are now being produced, we would still urge a degree of caution in their use without first determining their usefulness in comparison to direct typing from the target population. To add to this, we have only explored a single form of the -α$^{3.7}$ deletion in a population where this appears to be the only deletion that is present. In other parts of the world, the -α$^{3.7I}$ deletion occurs together with Type II and Type III deletions to varying degrees. The breakpoints for these alternative deletions are close to but different to that in the Type I. The GWAS chip SNPs within the -α$^{3.7}$ region may therefore be distributed across the breakpoints depending on the -α$^{3.7}$ type, giving a more complex intensity pattern. Moreover, there are other α-thalassaemia deletions of varying sizes around the world that can involve one or both of the HBA genes (HBA1 and HBA2). Some of these are common, and some also overlap the -α$^{3.7}$ deletion region, adding to this complexity. It may be that direct genotyping is therefore still the only solution to understanding the associations between this locus and malaria or other diseases."*

***Competing Interests:*** No competing interests were disclosed.