**Polymorphism in a lincRNA associates with a doubled risk of pneumococcal bacteremia in Kenyan children**

The Kenyan Bacteraemia study group and Wellcome Trust Case Control Consortium 2 (WTCCC2); Anna Rautanen1\*+, Matti Pirinen1\*, Tara C Mills1, Kirk Rockett1,2, Amy Strange1, Anne W Ndungu1,Vivek Naranbhai1, James J Gilchrist1,3, Céline Bellenguez1, Colin Freeman1, Gavin Band1, Suzannah J Bumpstead2, Sarah Edkins2, Eleni Giannoulatou1, Emma Gray2, Serge Dronov2, Sarah E Hunt2, Cordelia Langford2, Richard Pearson1, Zhan Su1, Damjan Vukcevic1, Alex W Macharia4, Sophie Uyoga4, Carolyne Ndila4, Neema Mturi4, Patricia Njuguna4, Shebe Mohammed4, James A Berkley4, Isaiah Mwangi4, Salim Mwarumba4, Barnes S Kitsao4, Brett S Lowe4, Susan C Morpeth4,5, Iqbal Khandwalla4, The Kilifi Bacteraemia Surveillance Group, Jenefer M Blackwell6,7, Elvira Bramon8, Matthew A Brown9, Juan P Casas10,11, Aiden Corvin12, Audrey Duncanson13, Janusz Jankowski14, Hugh S Markus15, Christopher G Mathew16, Colin N A Palmer17, Robert Plomin18, Stephen J Sawcer19, Richard C Trembath20, Ananth C Viswanathan21, Nicholas W Wood22, Panos Deloukas2, Leena Peltonen2, Thomas N Williams4,5,23,24, J Anthony G Scott4,5,25, Stephen J Chapman1,26, Peter Donnelly1,27, Adrian V S Hil1l,28+ & Chris C A Spencer1,27+

\* These authors contributed equally to this work

+ These authors jointly directed this work

A full list of membership of consortia appear at the end of this article.

Corresponding author: Dr. Anna Rautanen, Wellcome Trust Centre for Human Genetics, University of Oxford, Tel: +44 1865 287634, email: anna.rautanen@well.ox.ac.uk

**Affiliations**

1Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.

2Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

3Department of Paediatrics, University of Oxford, Oxford OX3 9DU, UK.

4KEMRI-Wellcome Trust Research Programme, Kilifi 80108, Kenya.

5Nuffield Department of Clinical Medicine, University of Oxford, Oxford OX3 9DU, UK.

6Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK.

7Telethon Institute for Child Health Research, Centre for Child Health Research, The University of Western Australia, Subiaco 6008, Western Australia, Australia.

8National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health at the South London and Maudsley National Health Service (NHS) Foundation Trust and Institute of Psychiatry King's College London, Denmark Hill, London SE5 8AF, UK.

9University of Queensland Diamantina Institute, Translational Research Institute, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland 4102, Australia.

10Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.

11Institute of Cardiovascular Science, University College London, WC1E 6BT, UK.

12Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 2, Ireland.

13Molecular and Physiological Sciences, The Wellcome Trust, London NW1 2BE, UK.

14Associate Deans Office, John Bull Building, Peninsula School of Medicine and Dentistry, PL6 8BU, UK

15Department of Neurology, University of Cambridge, Cambridge, CB2 0QQ, UK

16Division of Genetics and Molecular Medicine, King’s College London, London, SE1 9RT, UK.

17Medical Research Institute, University of Dundee, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK.

18King’s College London, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, De Crespigny Park, London, SE5 8AF, UK.

19Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK.

20London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK.

21NIHR Biomedical Research Centre at Moorfields Eye Hospital NHSFT and UCL Institute of Ophthalmology, London EC1V 2PD, UK.

22Department of Molecular Neuroscience, Institute of Neurology, Queen Square, London WC1N 3BG, UK.

23Department of Medicine, Imperial College, London W21NY, UK.

24INDEPTH Network, Accra, Ghana.

25The London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.

26Oxford Centre for Respiratory Medicine, Churchill Hospital Site, Oxford University Hospitals, Oxford OX3 7LE, UK.

27Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.

28The Jenner Institute, University of Oxford, Old Road Campus Research Building, Oxford OX3 7DQ, UK.

**ABSTRACT**

Bacteremia (bacterial bloodstream infection) is a major cause of illness and death in sub-Saharan Africa but little is known about the role of human genetics in susceptibility. We conducted a genome-wide association study of bacteremia susceptibility in more than 5000 Kenyan children as a part of the Wellcome Trust Case Control Consortium 2 (WTCCC2). Both the blood-culture-proven bacteremia cases and healthy infants as controls were recruited from Kilifi, east coast of Kenya. *Streptococcus pneumoniae* is the most common cause of bacteremia in Kilifi and was thus the focus of this study. We identified a novel association between polymorphisms in a long intergenic non-coding RNA (lincRNA) gene (AC011288.2) and pneumococcal bacteremia and replicated the results in the same population (*P* combined = 1.69x10-9; OR = 2.47, 95% CI = 1.84-3.31). The susceptibility allele is African specific, derived rather than ancestral, and occurs at low frequency (2.7% in controls and 6.4% in cases). Our further studies showed AC011288.2 expression only in neutrophils, a cell type that is known to play a major role in pneumococcal clearance. Identification of this novel association will further focus research on the role of lincRNAs in human infectious disease.

**INTRODUCTION**

Bacteremia is a common pathway in the progression to death from severe pneumonia, meningitis and sepsis, which together account for an estimated 3 million deaths each year globally in children under the age of 5 years. Even in developed countries the mortality rate from bacteremia remains unacceptably high1-3.The leading bacterial cause of death in young children worldwide is *Streptococcus pneumoniae* (pneumococcus), and 14.5 million episodes of serious pneumococcal disease occur in young children annually4. A key question is why only a proportion of individuals develop invasive disease despite widespread exposure and asymptomatic carriage of bacteria? Host genetic factors play an important role in explaining inter-individual variation in susceptibility to different infectious diseases5. However, the relevant genes for bacteremia susceptibility remain largely unknown..

To identify genetic correlates of bacteremia susceptibility, we conducted a genome-wide association study (GWAS) in Kenyan children, a population with a major disease burden6, as a part of the Wellcome Trust Case Control Consortium 2 (WTCCC2). Bacteremia is a heterogeneous phenotype and immune responses and genetic variants affecting susceptibility are likely to be at least partially pathogen specific. We thus focused on bacteremia caused by *S. pneumoniae*, the most common bacteria found in our study. In addition, all-cause bacteremia was analyzed in order to assess the possible role of genetic risk factors for bacteremia regardless of its etiology.

**SUBJECTS AND METHODS**

**Study Design**

To identify host genetic determinants of susceptibility to invasive pneumococcal disease in African children, we performed a two-stage GWAS of pneumococcal bacteremia in 542 Kenyan children with culture-confirmed disease and 4013 healthy controls. 429 cases and 2677 controls were included in the discovery phase analysis, with 113 cases and 1336 controls included in the replication analysis. To identify determinants of invasive bacterial disease irrespective of the pathogen, we further performed a GWAS of culture confirmed all-cause bacteremia in the same population of Kenyan children (discovery phase, 1536 cases; replication phase, 434 cases). Adopting a Bayesian framework, we considered evidence for shared effects at loci associated with pneumococcal disease and all-cause bacteremia, across pathogens commonly causing bacteremia in this population. Finally, we characterized disease-associated genetic variation identified in the study, analyzing tissue-specific expression of implicated transcripts in immune cell subsets. A detailed study workflow is described in Figure S1.

**Study participants**

All study participants were residents of Kilifi District on the coast of Kenya. Cases were recruited among children younger than 13 years of age who were admitted to Kilifi District Hospital (KDH) in Kenya between 1st August 1998 and 30th October 2010. Blood cultures were investigated from everyone admitted, unless they were admitted for elective procedures or because of minor accidents, using the BACTEC 9050 system. Children with bacteria present in their bloodstream were defined as cases (*Coryneforms* bacteria, *Bacillus* species, coagulase-negative *Staphylococcus*, *Staphylococcus saprophyticus* and Viridans group *Streptococcus* were excluded as contaminants). The annual incidence of bacteremia in Kilifi between August 1998 and July 2002 was estimated to be 505 cases per 100,000 children who were less than 5 years of age6, but the incidence has since decreased7.

Controls were selected among children born consecutively within the same Kilifi region between 1st May 2006 and 30th April 2008 and represent the cases closely in terms of sex, ethnic group, and geographic area of residence. Although the control individuals are part of a birth cohort study and thus aged less than 12 months at the time of recruitment to the study, we have been able to review their follow-up data in terms of development of bacteremia (N=12), mortality (N=49) etc. See further demographic details of cases and controls in Table S2. Table S3 shows the distribution of the most common bacterial isolates identified from bacteremia cases in the discovery and replication sets. The final discovery set included 1536 blood-culture-proven bacteremia cases (of whom 429 were pneumococcal) and 2677 healthy infants as controls. Individuals in the replication set were enrolled during the end of the collection period and included 434 bacteremia cases (of whom 113 were pneumococcal) and 1336 controls.

Ethical approval was granted by the Kenya Medical Research Institute (KEMRI) National Scientific Steering and Research Committees and the Oxford Tropical Research Ethics Committee (OXTREC). Informed consent was obtained from all subjects.

**DNA sample preparation**

Genomic DNA was extracted at the Kenya Medical Research Institute (KEMRI)-Wellcome Trust Collaborative Programme in Kenya, using the QIAamp DNA blood mini kit (Qiagen) and shipped to the Wellcome Trust Centre for Human Genetics, University of Oxford, for further processing. Genomic DNA was whole-genome amplified at the GeneService laboratory with GenomiPhi (GE HealthCare) scaled to amplify 40-50 µg of DNA. Quality of the whole-genome amplified DNA was assessed at the Wellcome Trust Sanger Institute as described elsewhere8 before genotyping.

**Genome-wide genotyping and quality control**

Whole-genome amplified samples from cases and controls were genotyped on the genome-wide Affymetrix SNP 6.0 chip at the Affymetrix service laboratory. Genotypes were called with a modified version of the Chiamo software9 for all samples passing the Affymetrix laboratory quality control measures. Sample QC was performed as described elsewhere8 and details are provided in Table S1 and Figures S1, S2 and S3. Analysis of pairwise allele sharing identified 68 duplicate pairs and 6 triplicates (Figure S4). Phenotypic information suggests that majority of these duplicate and triplicate individuals were unintentionally recruited to the study in Kenya more than once rather than being sample handling problems; therefore one of each of the duplicate pairs or triplicates was included in the analysis (a case rather than a control was included in the analysis; otherwise the sample with a higher call rate was included). First degree relatives (genome-wide IBD sharing probability >0.4; 117 individuals) were removed from the main analysis. The following criteria were used to exclude 102,896 unreliable SNPs: minor allele frequency (MAF) < 1% (50,322 SNPs), info <0.975 (53,419 SNPs), Hardy-Weinberg equilibrium *P* < 1x10-20 (18,288 SNPs), plate effect *P* < 1x10-6 (7,382 SNPs), and SNP missingness >2% (34,430 SNPs). Genotyping cluster plots of each SNP with *P* < 1x10-3 were visually inspected using Evoker10, and SNPs with poor cluster separation were removed. After sample and SNP QC 1536 cases and 2677 controls were analyzed at 790,739 genotyped autosomal SNPs. Three main ethnicities, namely Chonyi, Giriama, and Kauma, were discernible with principal components analysis (PCA) of the genome-wide data (Figure S5).

**Immunochip genotyping**

Approximately 2000 SNPs out of the total 200,000 SNPs were selected to be included in the ImmunoChip array11 based on the initial association results of the bacteremia analyses. The replication set was genotyped with this array at the Wellcome Trust Sanger Institute. All the samples went through a similar QC process as described above for the discovery samples (Table S1) and 434 cases and 1336 controls passed the QC. After excluding SNPs based on minor allele frequency < 1%, SNP call rate <95% (<99% if MAF <5%), and Hardy-Weinberg equilibrium *P* < 1x10-10, 143,100 SNPs remained for the further analyses. The same ethnicities were detectable by PCA in the replication sample set as in the discovery analysis (Figure S6). As the ImmunoChip genotyping was performed before the imputation, these genotypes were mainly utilized to account for population stratification and relatedness in the later replication analyses.

**Imputation and association analyses**

We performed whole-genome imputation using the 1000 Genomes Phase I data as a reference panel. Genotypes were pre-phased using SHAPEIT12 before imputation with IMPUTE213. Only samples and SNPs passing the QC were included for pre-phasing and imputation. SNPs with potentially unreliable imputation were filtered out based on MAF (< 2%), imputation info value (< 0.8), and Hardy-Weinberg equilibrium (*P* < 1x10-10). 10,996,499 imputed autosomal SNPs that passed the QC were analyzed for additive and genotypic models using SNPTEST214, taking the imputed genotype uncertainty (frequentist score test) and the first two principal components (PCs) of genetic structure into account. The genomic control parameter λ for bacteremia overall and pneumococcal bacteremia after imputation and QC were 1.043 and 1.013, respectively (see the QQ-plots in Figure S7). At associated SNPs, statistical tests were also performed using a linear mixed model that uses genome-wide data to model the pair-wise relatedness among the individuals15.

**Sequenom replication and confirmation of imputation accuracy**

SNPs with *P* < 1x10-5 in the additive model or *P* < 5x10-7 in the genotypic model were directly genotyped in the discovery set to confirm imputation accuracy and in the replication sample set to confirm the associations using two Sequenom iPLEX assays. Five SNPs looked unreliable after inspection of the cluster plots leaving 37 SNPs in the analysis (the cluster plot for the most significant SNP is shown in Figure S8). All of these SNPs had a call rate greater than 95% and the genotype distribution among controls obeyed Hardy-Weinberg equilibrium (P>0.05). After removing the samples that were originally excluded from the discovery and Immunochip analyses, 102 and 80 samples were removed because of the low call rate (<80%) and 7 and 9 samples because of the mismatching gender from the first and second multiplexes, respectively. This left 1514 cases (418 pneumococcal cases) and 2642 controls in the discovery sample set and 407 cases (103 pneumococcal cases) and 1333 controls in the replication analyses. Genotyping of these two iPlexes was performed at the Wellcome Trust Sanger Institute. The functional SNP rs334 in the *HBB* failed the initial assay design, and was therefore genotyped separately using a Sequenom iPlex at the Wellcome Trust Centre for Human Genetics, University of Oxford. The QC measures described above were applied to these samples, leaving 1360 cases and 2644 controls in the discovery set and 389 cases and 1312 controls in the replication set.

Only the samples that were included in either the discovery set or ImmunoChip replication set were included in the final analysis to allow inclusion of the first two PCs in logistic regression analysis using PLINK16 and to model the pair-wise relatedness in a linear mixed model. The combined statistics for the discovery and replication samples were obtained using fixed effects meta-analysis in GWAMA17. The replication data set had 80% power to detect an association (*P* < 0.05) with a common SNP (MAF 0.20) that has an effect size ≥ 1.3, but for more rare SNPs (MAF = 0.05) an effect size ≥1.54 was required (See Figure S9). Therefore we did not have sufficient statistical power to reliably replicate associations with modest effect sizes.

**Approaches to handle relatedness**

The SNPs chosen for replication were also analyzed using a linear mixed model15 which uses genome-wide data to model the pair-wise relatedness among the individuals, and which also included the first two PCs as covariates, to better account for relatedness and possible population structure within the sampled individuals. This was done by including all relatives and also by including only distantly related individuals (r < 0.2).

We further assessed whether the sample set with pneumococcal infection includes more pairs of close relatives than other bacteremia cases or than controls. This was assessed by comparing the observed number of relative pairs with estimated r > 0.025 among the pneumococcus cases to an equal sized set of the rest of the cases or controls that are matched with respect to manual clustering (Figure S11) by resampling 100,000 data sets.

**Bayesian model comparisons**

To compare models of the similarity of effect across bacterial species at identified disease-associated loci, we took a Bayesian approach (for a similar approach see18; 19). The likelihood function is based on multinomial regression with strata corresponding to the controls and each of the seven most common bacterial subgroups (Figure 2; Table S3). Cases infected with more than one of these seven different bacterial species (2.1% of cases) are included in the analysis for each group.

The parameters of interest are the genetic effect sizes (bk, k=1,..,7) on a log-odds scale for each of the case cohorts. We first find maximum likelihood estimates (with the corresponding observed information matrix) by including two PCs as covariates in the model, and then compute approximate Bayes factors using a multivariate normal approximation to the likelihood and the prior. The models are defined by prior distributions on the parameters bk:

NULL: bk=0 for all k=1,..,7, i.e. no effects, all case groups are like the control group.

SAME: bk ~ N(0,1) and cor(bi,bj)=1 for all pairs i≠j, i.e. each bk is the same.

REL: bk ~ N(0,1) and cor(bi,bj)=0.96 for all pairs i≠j, i.e., bi and bj are correlated but not necessarily the same.

Additional models are defined following inspection of the observed association at each locus for each pathogen. Bacterial species hypothesized to be associated with a given locus are assumed to have the same non-zero effect with a prior of N(0,1), whereas for other pathogens the effect is 0.

**Quantification of lincRNA expression in primary immune cell subsets**

Previous reports suggest that AC011288.2 encodes a lincRNA and is expressed in white blood cells and placental tissue. To identify which leukocyte population this lincRNA is expressed in, we isolated monocytes, B-cells and Natural Killer (NK) cells from consenting healthy adult Caucasian donors using magnetic activated cell sorting (MACS, Miltenyi), as previously described20. In addition we isolated granulocytes (predominantly neutrophils) using Polymorphoprep (Allere) according to the manufacturer’s instructions from eight individuals. The purity of cell subsets after cell separation was assessed by flow cytometry and was >90% in a representative sample. Viability following sorting was assessed by the Trypan Blue dye exclusion method and observed to be >95% in all cases. Total RNA was extracted using the RNeasy mini kit (Qiagen) or TRIzol (Life technologies) according to the manufacturer's instruction (Qiagen). Total RNA was quantified by Nanodrop and Bioanalyzer for a subset following the manufacturers' instruction (Bioanalyzer RNA 6000 Nano kit, Agilent).

To quantify levels of lincRNA expression we performed quantitative real-time PCR (qPCR) using a relative quantification method. Beta-Actin (ACTB) was selected as a reference gene based on previous reports of its stable expression in neutrophils. Single-strand complementary DNA was synthesized by reverse transcription with the SuperScriptIII First-Strand Synthesis System (Invitrogen). Primers specific to each of the two reported transcripts for the lincRNA AC011288.2 gene were designed: AC011288.2-001 (for- GTCAGAAGCGGGGTTCAAAG, rev-TTTAATTCTTGAGTTCTGCAGGC), AC011288.2-002 (for-GATGCTAAGCCTGGAAACCC, rev- TCCAGCTTCTATTCCCAGAGG). In addition we designed primers to AC006000.5 (for-ACTCCACGTCCCACAGATAC and rev-TGACAGAGTGAGACCCTGTG) but consistent with previous reports that observed no expression in leucocytes, we did not identify any individuals that expressed this transcript and do not describe it further. To avoid potential amplification of genomic DNA, primers were designed to span exons. qPCR was performed using SYBR Green Supermix (BioRad) on a CFX96 Real-Time PCR Detection System (Bio-Rad). Reactions were run in duplicate with 1 cycle at 95 ˚C (10 min), followed by 42 cycles consisting of denaturation at 95 ˚C (10 sec), annealing at 58 ˚C (20 sec) and extension at 72 ˚C (20 sec). Detection of the fluorescent products was carried out at the end of the 72 ˚C extension period. To confirm amplification specificity, the PCR products were subjected to a melting curve analysis and agarose gel electrophoresis. Detection of a fluorescent product after cycle 38 (Ct=38) was considered evidence of expression beneath the confident detection limit based on careful inspection of the melting curves and agarose gel electrophoresis results. Therefore, if Ct values of greater than 38 were obtained, the Ct value was re-assigned to 38, so as to conservatively estimate the highest level of expression beneath the detectable level. On average, the Ct value for AC0011288.2-001 was 33 cycles in neutrophils. Relative gene transcript levels were determined by the [DELTA][C.sub.T] method expressed relative to ACTB. Comparisons of log10 transformed relative expression levels were made using a non-parametric Mann-Whitney test in GraphPad Prism.

**RESULTS**

**Genome-wide association results of directly genotyped discovery and Immunochip replication sets**

We identified several suggestive associations in the directly genotyped discovery data both in pneumococcal bacteremia and bacteremia overall analyses in Kenyan children (Supplementary Table S4), but none of the SNPs reached established criteria for identifying novel associations (*P* < 5x10-8)in a combined analysis after replication.

**Genome-wide association results of bacteremia caused by *Streptococcus pneumoniae* after imputation and replication**

Following genome-wide imputation and quality control, nearly 10 million autosomal SNPs were included in the association analyses of pneumococcal bacteremia (Figure S10). In this analysis of 429 cases and 2677 controls, 17 SNPs in a single region on chromosome 7 were associated with disease at a level exceeding genome-wide significance (*P*<5x10-8), with the peak of that association observed at rs140817150 (*P* imputed = 7.25x10-9; OR = 2.74) (Figure S10 and Figure 1).This novel associated region includes two overlapping long intergenic non-coding RNA (lincRNA) genes AC00600.5 and AC011288.2. The association at rs140817150 was confirmed by direct genotyping (*P* discovery = 3.58x10-7; OR = 2.39) and replication (*P* = 1.16x10-3; OR = 2.72), resulting in a combined OR estimate of 2.47 (95% CI 1.84-3.31) and *P* value of 1.69x10-9 (See Table S5 for a list of all suggestive associations in the analysis of pneumococcal bacteremia and Table S6 for the comprehensive list of associated SNPs in the chromosome 7 top associated region). Direct genotyping of the top imputed SNPs confirmed that imputation was generally accurate (average concordance between imputed and directly genotyped genotypes was 98.3%). In order to protect against spurious associations due to possible cryptic relatedness, the SNPs chosen for replication were also analyzed using the mixed model approach (rs140817150; *P* combined = 1.5x10-10, OR = 2.66; Table S7) and by stratifying individuals on the basis of genetic background across the main ethnic groups (rs140817150; *P* discovery = 1.5x10-7,OR = 2.54; Figure S11). By either approach the evidence for association remained strong. The pneumococcal cases do not have more close relatives than the controls do in any of the four ancestry groups (*P* ≥ 0.22). When compared to other cases, they show elevated levels of relatedness (*P* < 0.05) in Group 1 only, which does not contribute strongly to the observed association signal (Figure S11). After conditioning on the top SNP, no associations were detected with *P* < 10-4 in the region.

**Genome-wide association results of bacteremia overall after imputation and replication**

In addition to the analysis of a more homogeneous phenotype, pneumococcal bacteremia, all-cause bacteremia overall (Figure S12) was also analyzed. Table S8 summarizes loci with the strongest evidence of association in bacteremia overall after imputation, including the replication results. The only genome-wide significant (p<5x10-8) association signal that replicated was under the genotypic model, which allows independent effects on risk for homozygotes and heterozygotes. It revealed a strong association at the previously identified *HBB* locus (rs113892119, *P* = 5.08x10-13). The region of association included the functional rs334 polymorphism (*P* = 1.33x10-10) that leads to the production of hemoglobin S (HbS)21 and is located 25.6 kb upstream from the most associating SNP rs113892119.

As previously described, rs334 was associated with susceptibility among homozygotes22 (HbSS vs HbAA; directly genotyped combined *P* = 2.66x10-12, OR = 4.9) and with protection from bacteremia among heterozygotes7 (HbAS vs HbAA; directly genotyped combined *P* = 4.67x10-3, OR = 0.77). These same effects are seen in the most common bacterial subgroups (Figure S13).

**Bayesian model comparison results of rs140817150 and rs334 associations with common causes of bacteremia in Kenyan children**

Although the association with rs140817150 was discovered in the pneumococcal bacteremia analysis, we were able to utilize the all-cause bacteremia data at this locus to assess its effect on susceptibility to bacteremia caused by otherspecies (Figure 2)*.* To assess whether the data were consistent with the same effect among cases with different species of bacteremic pathogen we compared models using a Bayesian approach (Figure 2). Assuming all models to be equally likely *a priori*,the most probable model is one in which the susceptibility is confined to pneumococcus, *Acinetobacter* species and *Haemophilus influenzae*. Removing the pneumococcal group, from which the association was ascertained, weakened the evidence for effect heterogeneity. The same effect in all subtypes was found to be the most probable model for rs334 association (heterozygote risk and homozygote protection) in the *HBB* (Figure S13).

**lincRNA expression in primary immune cell subsets**

We assessed AC011288.2 RNA expression in the major leukocyte cell subsets and observed expression only in neutrophils. Expression levels were below the detection limit in monocytes, B-cells and natural killer (NK) cells (Figure 3). To verify that expression of this transcript is constitutive in neutrophils we measured expression in an additional 75 donors, recruited in a separate study23 and observed detectable expression in all 75 donors. We did not observe AC00600.5 expression in any leukocyte subsets.

**DISCUSSION**

We report here a GWAS of bacteremia susceptibility, which is one of the few large-scale GWAS conducted in an African population to date. We identified a novel association between polymorphisms in two overlapping long intergenic non-coding RNA (lincRNA) genes (AC00600.5 and AC011288.2) and pneumococcal bacteremia, the most common cause of bacteremia in our study set. Although immune responses and genetic variants affecting bacteremia susceptibility are likely to be at least partially pathogen specific, we also analyzed the bacteremia overall data set to identify more universal risk factors. The only genome-wide significant hit for bacteremia overall was in a previously reported *HBB*, including the well-known rs334 polymorphism associated with the production of sickle hemoglobin21.

The novel lincRNA risk allele at rs140817150 is derived (as reported on dbSNP) rather than ancestral, its frequency is low (2.7% in controls, 6.4% in pneumococcal bacteremia cases), and according to the 1000 Genomes project data (phase 3) it is polymorphic only in African populations. Consistent with the local recombination landscape and with the expectation that low frequency derived alleles are relatively young, SNPs in linkage disequilibrium with rs140817150 extend over 500 kb (Figure 1). However, Bayesian analysis of the region of association24 in the imputed data suggests there is greater than 95% probability that one of the most associated SNPs (circled in Figure 1) is the causal variant, assuming there is a single causal variant, and it is imputed accurately in our dataset.

The association peak is located in the introns of two separate long intergenic non-coding RNAs (lincRNAs), annotated as AC011288.2 and AC006000.5. The importance of lincRNAs as key regulators of gene expression has only recently been recognized25-27. It has been estimated that the human genome includes at least ten thousand lincRNAs but only a fraction of these has a known function25-27. A recent study aiming to catalogue the function of more than 8000 human lincRNAs reported that lincRNA expression is significantly more tissue-specific than expression of protein coding genes28. AC006000.5 is listed in the catalogue but it is not expressed in any of the studied tissues, whereas AC011288.2 is reported to be expressed only in placenta and white blood cells out of 24 different tissues and cell lines studied. We assessed AC011288.2 RNA expression in leukocyte cell subsets and observed expression only in neutrophils, a cell type that is known to play a major role in pneumococcal clearance29; 30. These results provide an important direction for future functional investigations. Neutrophils express many antimicrobial peptides and proteins that confer both universal and pathogen specific host response31; 32 and it has been shown that absolute neutrophil count is an independent predictor of pneumococcal bacteremia in febrile children33.

 The closest protein coding genes surrounding the association signal are *ARL4A* and *ETV1* but there is no evidence that the associating lincRNAs regulate these two genes. However, data from a previous expression quantitative trait locus (eQTL) study20 suggest that there are some SNPs in the associating region that function as eQTLs in monocytes (rs1432496) and B-cells (rs2568633) for *PHF14* (PHD finger protein 14), a transcription factor that down regulates *PDGFRα* expression34. However, neither is correlated with our most associated SNP (r2 < 0.01 in 1000 Genomes data). Although the role of lincRNAs in human infections is unknown, recent mouse studies have indicated that some lincRNAs can act in immune cells to regulate host susceptibility to bacterial and viral infections35; 36.

Using the GWAS approach we have identified a novel association between a genetic variation in a lincRNA gene and pneumococcal bacteremia. Furthermore, we have confirmed a previously reported association between *HBB* and bacteremia overall22, with homozygotes associated with strong susceptibility but heterozygotes associated with protection. At both associated loci, the disease-associated alleles are rare in individuals without African ancestry (monomorphic in 1000 Genomes Project data in other than African populations) and exert a large effect on the likelihood of developing bacteremia. These associations have not been reported by earlier GWAS studies of related phenotypes, which is unsurprising as the populations under study have been of European decent and differences in phenotypes are still substantial37-39. The reported SNPs in *FER* [MIM: 176942] that was recently associated with outcome from sepsis due to pneumonia37 or in *CFH* [MIM: 134370] and *CFHR3* [MIM:605336] that have been associated with meningococcal disease38 did not show any evidence of association in the current study (*P*>0.05). Given the likely importance of host-pathogen molecular interactions in bacteremia susceptibility it is plausible that the effect of a risk allele will be dependent on bacterial species. Our data on the lincRNA locus provide initial evidence for this at the bacterial species level, and motivate approaches which stratify host genetic associations by pathogen species, serotype or genotype. Understanding the molecular mechanisms leading to the doubled risk of pneumococcal bacteremia associated with this allele could provide new clues in the pressing search for new therapeutic targets.

**SUPPLEMENTAL DATA**

Supplemental Data include 13 figures and 8 tables and can be found with this article online.

# Acknowledgements

We thank all the study participants and Kilifi District Hospital clinical team and laboratory staff for their involvement in data and sample collection. The principal funding for this study was provided by the Wellcome Trust, as part of the Wellcome Trust Case Control Consortium 2 project (grants 084716/Z/08/Z, 085475/B/08/Z and 085475/Z/08/Z). This work was partially supported by Wellcome Trust Centre for Human Genetics core grant 090532/Z/09/Z. The fieldwork and phenotyping for this study was supported by the Kenya Medical Research Institute (KEMRI) and the Wellcome Trust of Great Britain. A Rautanen was supported by the Wellcome Trust (084716/Z/08/Z) and currently by the European Research Council, M Pirinen is supported by the Academy of Finland (257654), TN Williams and JAG Scott were supported by Senior Research Fellowships from the Wellcome Trust (091758 and 098532 respectively), SJ Chapman was supported by the NIHR Biomedical Research Centre, Oxford, P Donnelly was supported in part by a Wolfson-Royal Society Merit Award, AVS Hill is supported by a Wellcome Trust Senior Investigator Award (HCUZZ0) and an ERC Advanced Grant (294557), and CCA Spencer was supported by a Wellcome Trust Career Development Fellowship (097364/Z/11/Z). This paper was published with the permission of the Director of KEMRI. The authors do not have any conflicts of interest.

**Consortia**

**The Kenyan Bacteraemia study group**

Principal Investigators:

Adrian V S Hill (Chair), Thomas N Williams, J Anthony G Scott, Stephen J Chapman

Key Personnel:

Anna Rautanen, Tara C Mills, Kirk Rockett, Anne W Ndungu, Vivek Naranbhai, Alex W Macharia, Sophie Uyoga, Carolyne Ndila, Neema Mturi, Patricia Njuguna, Shebe Mohammed, James A Berkley, Isaiah Mwangi, Salim Mwarumba, Barnes S Kitsao, Brett S Lowe, Susan C Morpeth, Iqbal Khandwalla

The Kilifi DNA extraction Group:

Alex W Macharia, Sophie Uyoga, Herbert Opi, Carolyne Ndila, Emily Nyatichi, Prophet Ingosi, Barnes Kitsao, Clement Lewa, Johnstone Makale, Adan Mohamed, Kenneth Magua, Mary Njoroge, Gideon Nyutu, Ruth Mwarabu, Metrine Tendwa and Thomas N Williams.

The Kilifi Bacteraemia Surveillance Group:

Ismail Ahmed, Samuel Akech, Alexander Balo Makazi, Mohammed Bakari Hajj, Andrew Brent, Charles Chesaro, Hiza Dayo, Richard Idro, Patrick Kosgei, Kathryn Maitland, Kevin Marsh, Laura Mwalekwa, Shalton Mwaringa, Charles Newton, Mwanajuma Ngama, Allan Pamba, Norbert Peshu, Anna Seale, Alison Talbert and Thomas N Williams.

**Wellcome Trust Case Control Consortium 2**

Management Committee

Peter Donnelly (Chair), Ines Barroso (Deputy Chair), Jenefer M Blackwell, Elvira

Bramon, Matthew A Brown, Juan P Casas, Aiden Corvin, Panos Deloukas, Audrey

Duncanson, Janusz Jankowski, Hugh S Markus, Christopher G Mathew, Colin NA

Palmer, Robert Plomin, Anna Rautanen, Stephen J Sawcer, Richard C Trembath,

Ananth C Viswanathan, Nicholas W Wood

Data and Analysis Group

Chris C A Spencer, Gavin Band, Céline Bellenguez, Colin Freeman, Garrett

Hellenthal, Eleni Giannoulatou, Matti Pirinen, Richard Pearson, Amy Strange, Zhan

Su, Damjan Vukcevic, Peter Donnelly

DNA, Genotyping, Data QC and Informatics Group

Cordelia Langford, Sarah E Hunt, Sarah Edkins, Rhian Gwilliam, Hannah Blackburn,

Suzannah J Bumpstead, Serge Dronov, Matthew Gillman, Emma Gray, Naomi

Hammond, Alagurevathi Jayakumar, Owen T McCann, Jennifer Liddle, Simon C

Potter, Radhi Ravindrarajah, Michelle Ricketts, Matthew Waller, Paul Weston, Sara

Widaa, Pamela Whittaker, Ines Barroso, Panos Deloukas.

Publications Committee

Christopher G Mathew (Chair), Jenefer M Blackwell, Matthew A Brown, Aiden

Corvin, Chris C A Spencer

**The Kilifi Bacteraemia Surveillance Group:**

Ismail Ahmed, Samuel Akech, Alexander Balo Makazi, Mohammed Bakari Hajj, Andrew Brent, Charles Chesaro, Hiza Dayo, Richard Idro, Patrick Kosgei, Kathryn Maitland, Kevin Marsh, Laura Mwalekwa, Shalton Mwaringa, Charles Newton, Mwanajuma Ngama, Allan Pamba, Norbert Peshu, Anna Seale, Alison Talbert, Thomas N Williams

**WEB RESOURCES**

Online Mendelian Inheritance in Man (<http://www.omim.org>)

1000 Genomes Project data (http://browser.1000genomes.org/index.html)

# References

1. Lillie, P.J., Allen, J., Hall, C., Walsh, C., Adams, K., Thaker, H., Moss, P., and Barlow, G.D. (2012). Long-term mortality following bloodstream infection. Clin Microbiol Infect.

2. Wyllie, D.H., Crook, D.W., and Peto, T.E. (2006). Mortality after Staphylococcus aureus bacteraemia in two hospitals in Oxfordshire, 1997-2003: cohort study. BMJ 333, 281.

3. Laupland, K.B., Svenson, L.W., Gregson, D.B., and Church, D.L. (2011). Long-term mortality associated with community-onset bloodstream infection. Infection 39, 405-410.

4. O'Brien, K.L., Wolfson, L.J., Watt, J.P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E., Mulholland, K., Levine, O.S., and Cherian, T. (2009). Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. Lancet 374, 893-902.

5. Sorensen, T.I., Nielsen, G.G., Andersen, P.K., and Teasdale, T.W. (1988). Genetic and environmental influences on premature death in adult adoptees. N Engl J Med 318, 727-732.

6. Berkley, J.A., Lowe, B.S., Mwangi, I., Williams, T., Bauni, E., Mwarumba, S., Ngetsa, C., Slack, M.P., Njenga, S., Hart, C.A., et al. (2005). Bacteremia among children admitted to a rural hospital in Kenya. N Engl J Med 352, 39-47.

7. Scott, J.A., Berkley, J.A., Mwangi, I., Ochola, L., Uyoga, S., Macharia, A., Ndila, C., Lowe, B.S., Mwarumba, S., Bauni, E., et al. (2011). Relation between falciparum malaria and bacteraemia in Kenyan children: a population-based, case-control study and a longitudinal study. Lancet 378, 1316-1323.

8. Barrett, J.C., Lee, J.C., Lees, C.W., Prescott, N.J., Anderson, C.A., Phillips, A., Wesley, E., Parnell, K., Zhang, H., Drummond, H., et al. (2009). Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet 41, 1330-1334.

9. Consortium, W.T.C.C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-678.

10. Morris, J.A., Randall, J.C., Maller, J.B., and Barrett, J.C. (2010). Evoker: a visualization tool for genotype intensity data. Bioinformatics 26, 1786-1787.

11. Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. Arthritis Res Ther 13, 101.

12. Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat Methods 9, 179-181.

13. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5, e1000529.

14. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet 11, 499-511.

15. Pirinen, M., Donnelly, P., and Spencer, C.C. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann Appl Stat available online.

16. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.

17. Magi, R., and Morris, A.P. (2010). GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11, 288.

18. Band, G., Le, Q.S., Jostins, L., Pirinen, M., Kivinen, K., Jallow, M., Sisay-Joof, F., Bojang, K., Pinder, M., Sirugo, G., et al. (2013). Imputation-based meta-analysis of severe malaria in three african populations. PLoS Genet 9, e1003509.

19. Bellenguez, C., Bevan, S., Gschwendtner, A., Spencer, C.C., Burgess, A.I., Pirinen, M., Jackson, C.A., Traylor, M., Strange, A., Su, Z., et al. (2012). Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. Nat Genet 44, 328-333.

20. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet 44, 502-510.

21. Ingram V.M. (1957) Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180, 326-8.

.

22. Williams, T.N., Uyoga, S., Macharia, A., Ndila, C., McAuley, C.F., Opi, D.H., Mwarumba, S., Makani, J., Komba, A., Ndiritu, M.N., et al. (2009). Bacteraemia in Kenyan children with sickle-cell anaemia: a retrospective cohort and case-control study. Lancet 374, 1364-1370.

23. Naranbhai, V., Fairfax, B.P., Makino, S., Humburg, P., Wong, D., Ng, E., Hill, A.V., and Knight, J.C. (2015). Genomic modulators of gene expression in human neutrophils. Nature communications 6, 7545

24. Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S., Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet 44, 1294-1301.

25. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res 22, 1775-1789.

26. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature 489, 101-108.

27. Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science 338, 1435-1439.

28. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011 Sep 15;25(18):1915-27 doi: 101101/gad17446611 Epub 2011 Sep 2.

29. Gingles, N.A., Alexander, J.E., Kadioglu, A., Andrew, P.W., Kerr, A., Mitchell, T.J., Hopes, E., Denny, P., Brown, S., Jones, H.B., et al. (2001). Role of genetic resistance in invasive pneumococcal infection: identification and study of susceptibility and resistance in inbred mouse strains. Infect Immun 69, 426-434.

30. Brinkmann, V., Reichard, U., Goosmann, C., Fauler, B., Uhlemann, Y., Weiss, D.S., Weinrauch, Y., and Zychlinsky, A. (2004). Neutrophil extracellular traps kill bacteria. Science 303, 1532-1535.

31. Cederlund, A., Agerberth, B., and Bergman, P. (2010). Specificity in killing pathogens is mediated by distinct repertoires of human neutrophil peptides. J Innate Immun 2, 508-521.

32. Nguyen, Q.T., Nguyen, T.H., Ju, S.A., Lee, Y.S., Han, S.H., Lee, S.C., Kwon, B.S., Yu, R., Kim, G.Y., Lee, B.J., et al. (2013). CD137 expressed on neutrophils plays dual roles in antibacterial responses against Gram-positive and Gram-negative bacterial infections. Infect Immun 81, 2168-2177.

33. Kuppermann, N., Fleisher, G.R., and Jaffe, D.M. (1998). Predictors of occult pneumococcal bacteremia in young febrile children. Ann Emerg Med 31, 679-687.

34. Kitagawa, M., Takebe, A., Ono, Y., Imai, T., Nakao, K., Nishikawa, S., and Era, T. (2012). Phf14, a Novel Regulator of Mesenchyme Growth via Platelet-derived Growth Factor (PDGF) Receptor-alpha. J Biol Chem 287, 27983-27996.

35. Gomez, J.A., Wapinski, O.L., Yang, Y.W., Bureau, J.F., Gopinath, S., Monack, D.M., Chang, H.Y., Brahic, M., and Kirkegaard, K. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. Cell 152, 743-754.

36. Carpenter, S., Atianand, M., Aiello, D., Ricci, E.P., Gandhi, P., Hall, L.L., Byron, M., Monks, B., Henry-Bezy, M., Lawrence, J.B., et al. (2013). A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. Science. 341, 789-92.

37. Rautanen, A., Mills, T.C., Gordon, A.C., Hutton, P., Steffens, M., Nuamah, R., Chiche, J.D., Parks, T., Chapman, S.J., Davenport, E.E. et al. (2015). Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. Lancet Respir Med 3, 53-60.

38. Davila, S., Wright, V.J., Khor, C.C., Sim, K.S., Binder, A., Breunis, W.B., Inwald, D., Nadel, S., Betts, H., Carrol, E.D., et al. (2010) Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet 42, 772-6.

**FIGURE LEGENDS**

**Figure 1.** Signal of association around rs140817150 (additive model) in the discovery analysis of pneumococcal bacteremia. Imputed SNPs are shown as circles and directly genotyped SNPs as triangles with colors indicating the correlation (r2 in 1000 Genomes data) with rs140817150. A set of SNPs that contains the causal SNP with greater than 95% probability is ringed with circles. Annotated genes (blue) and lincRNAs (red) are shown in the bottom panel along with the fine-scale recombination rate.

**Figure 2.** LincRNA association (rs140817150; additive model) with the main bacterial infections. A) Log transformed combined odds ratios and 95% confidence intervals of directly genotyped discovery and replication samples. The dotted line represents the log OR of 0 (OR of 1; no difference between cases and controls). The values of point estimates and standard errors (in parentheses) are also given. Bacterial infections: PNEUM=*Streptococcus pneumoniae* (pneumococcus); ACINET=*Acinetobacter* species; HAEMOPH= *Haemophilus influenzae*; ECOLI=*Escherichia coli*; SALMON= *Salmonella* (non-typhoidal); STREPBH=*Streptococcus beta haemolytic*; SAUR=*Staphylococcus aureus*.

B) The posterior probabilities on the models of association: no effect in any subtype (NULL), same effect in all subtypes (SAME), related effects across subtypes (REL) or a same non-zero effect only in PNEUM, ACINET and HAEMOPH (P+A+H), PNEUM and ACINET (P+A) or PNEUM (P). (See Methods). Models are a priori assumed to be equally likely. Bayes factors, which compare the evidence (marginal likelihood) between any pair of models, can be calculated as the ratio of the posterior probability assigned to each model as reported under each bar of the plot.

**Figure 3.** LincRNA AC011288.2 expression measured in neutrophils, monocytes, B-cells and NK cells. A)) Quantitative PCR of AC011288.2-002 in primary leucocyte subsets. To conservatively estimate the highest level of expression beneath the detectable level, CT values greater than 38 were re-assigned to 38, and normalized to **β**-actin expression. B)) Unadjusted cycle number of amplification is shown. Filled squares () denote detection of a fluorescent product after cycle 38 (Ct=38), the limit of confident detection being based on careful inspection of the melting curves. Similar results for AC011288.2-001 were obtained (data not shown).

 P-values denote the significance of the relative expression levels of AC011288.2-001 in neutrophils compared to other cell types (Mann-Whitney test).