

Population-based analysis of ocular *Chlamydia trachomatis* in trachoma-endemic West African communities identifies genomic markers of disease severity

AR Last¹, H Pickering¹, Ch Roberts¹, F Coll², J Phelan², SE Burr^{1,3}, E Cassama⁴, M Nabicassa⁴, HMB Seth-Smith^{5,6,7}, J Hadfield⁵, LT Cutcliffe⁸, IN Clarke⁸, DCW Mabey¹, RL Bailey¹, TG Clark^{2,9}, NR Thomson^{2,5}, MJ Holland¹

¹Clinical Research Department, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom, ²Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom, ³Disease Control and Elimination Theme, Medical Research Council Unit The Gambia, Fajara, The Gambia, ⁴Programa Nacional de Saúde de Visão, Ministério de Saúde Pública, Bissau, Guinea Bissau ⁵Pathogen Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, ⁶Clinical Microbiology, Universitätsspital Basel, Switzerland, ⁷Applied Microbiology Research, Department of Biomedicine, University of Basel, Switzerland, ⁸Molecular Microbiology Group, University of Southampton Medical School, Southampton, United Kingdom, ⁹ Department of Infectious Diseases Epidemiology, London School of Hygiene and Tropical Medicine, Keppel Street, London, United Kingdom.

ABSTRACT

Chlamydia trachomatis (*Ct*) is the most common infectious cause of blindness and bacterial sexually transmitted infection worldwide. Using *Ct* whole genome sequences obtained directly from conjunctival swabs, we studied *Ct* genomic diversity and associations between *Ct* genetic polymorphisms with ocular localization and disease severity in a treatment-naïve trachoma-endemic population in Guinea Bissau, West Africa. All sequences fall within the T2 ocular clade phylogenetically. This is consistent with the presence of the characteristic deletion in *trpA* resulting in a truncated non-functional protein and the ocular tyrosine repeat regions present in *tarP* associated with ocular tissue localization. We have identified twenty-one *Ct* non-synonymous single nucleotide polymorphisms (SNPs) associated with ocular localization, including SNPs within *pmpD* (OR=4.07, $p^*=0.001$) and *tarP* (OR=0.34, $p^*=0.009$). Eight SNPs associated with disease severity were found in *yjfH* (*rImB*) (OR=0.13, $p^*=0.037$), *CTA0273* (OR=0.12, $p^*=0.027$), *trmD* (OR=0.12, $p^*=0.032$), *CTA0744* (OR=0.12, $p^*=0.041$), *glgA* (OR=0.10, $p^*=0.026$), *alaS* (OR=0.10, $p^*=0.032$), *pmpE* (OR=0.08, $p^*=0.001$) and the intergenic region *CTA0744-CTA0745* (OR=0.13, $p^*=0.043$). This study demonstrates the extent of genomic diversity within a naturally circulating population of ocular *Ct*, and the first to describe novel genomic associations with disease severity. These findings direct investigation of host-pathogen interactions that may be important in ocular *Ct* pathogenesis and disease transmission.

KEYWORDS

Chlamydia trachomatis, trachoma, disease severity, genome wide association analysis, single nucleotide polymorphisms, pathogen genomic diversity

INTRODUCTION

The obligate intracellular bacterium *Chlamydia trachomatis* (*Ct*) is the leading infectious cause of blindness (trachoma) and the most common sexually transmitted bacterial infection^{1,2}.

Ct strains are differentiated into biovars based on pathobiological characteristics and serovars based on serological reactivity for the major outer membrane protein (MOMP) encoded by *ompA*³. Serovars largely differentiate biological groups associated with trachoma (A-C), sexually transmitted disease (D-K) and lymphogranuloma venereum (LGV) (L1-3). Despite diverse biological phenotypes, *Ct* strains share near complete genomic synteny and gene content⁴, suggesting that minor genetic changes influence pathogen-host and tissue-specific infection characteristics⁵⁻⁷. All published African ocular *Ct* genomes are situated on the ocular branch within the T2 clade of non-LGV urogenital isolates⁴. Currently there are only 31 published ocular *Ct* genome sequences^{4,9-12}. In particular there appears to be limited genomic diversity between published *Ct* genomes from Gambian and Tanzanian populations⁸.

The pathogenesis of chlamydial infection begins with epithelial inflammation and may progress to chronic immuno-fibrogenic processes leading to blindness and infertility, though many *Ct* infections do not result in sequelae^{13,14}. Strain-specific differences related to clinical presentation have been investigated in trachoma^{8,15,16}. These studies examined a small number of ocular *Ct* isolates from the major trachoma serotypes and found a small subset of genes in addition to *ompA* that were associated with differences in *in vitro* growth rate, burst size, plaque morphology, interferon- γ sensitivity and most importantly, intensity of infection and clinical disease severity in non-human primates (NHPs), suggesting that genetic polymorphisms in *Ct* may

contribute to the observed variability in severity of trachoma in endemic communities⁸.

The obligate intracellular development of *Ct* has presented significant technical barriers to basic research into chlamydial biology. Only recently has genetic manipulation of the chlamydial plasmid been possible, allowing *in vitro* transformation and modification studies, though this remains technically challenging, necessitating alternative approaches^{17,18}.

Whole genome sequencing (WGS) has recently been used to identify regions of likely recombination in recent clinical isolates, demonstrating that WGS analysis may be an effective approach for the discovery of loci associated with clinical presentation⁶. Additionally, a number of putative virulence factors have been identified through WGS analysis and subsequent *in vitro* and animal studies^{5,19-30}. However there are currently no published population-based studies of *Ct* using WGS with corresponding detailed clinical data, making it difficult to relate genetic changes to functional relevance and virulence factors *in vivo*.

There is an increasing pool of *Ct* genomic data, largely from archived samples following cell culture and more recently directly from clinical samples³¹. WGS data obtained directly from clinical samples can be preferable to using WGS data obtained from cell cultured *Ct*, since repeated passage of *Ct* results in mutations that are not observed *in vivo*³²⁻³⁴.

Ct bacterial load is associated with disease severity, particularly conjunctival inflammation, in active (infective) trachoma³⁵. Conjunctival inflammation has previously been shown to be a marker of severe disease and plays an important role in the pathogenesis of scarring trachoma³⁶⁻³⁸. In this study we used principal component analysis (PCA) to reduce the dimensions of clinical grade of inflammation (defined

using the P score from the FPC trachoma grading system³⁹) and *Ct* bacterial load to a single metric to define an *in vivo* conjunctival phenotype in active (infective) trachoma. PCA is a recognised dimension reduction technique used to combine multiple correlated traits into their uncorrelated principal components (PC)⁴⁰⁻⁴², allowing us to examine the relationship between *Ct* genotype and disease severity. These data currently represent the largest collection of ocular *Ct* sequences from a single population from the trachoma-endemic region of the Bijagós Archipelago of Guinea Bissau, and provide a unique opportunity to gain insight into ocular *Ct* pathogenesis in humans.

RESULTS

Conjunctival swabs collected during a cross-sectional population-based trachoma survey on the Bijagós Archipelago yielded 220 ocular *Ct* infections detected by *Ct* plasmid-based droplet digital PCR (ddPCR). Of the 220 *Ct* infections detected, 184 were quantifiable using *Ct* genome-based ddPCR.

We obtained WGS data from 126 using cell culture ($n=8$) or direct sequencing from swabs with SureSelect^{XT} target enrichment ($n=118$), representing the largest cross-sectional collection of ocular *Ct* WGS. Eighty-one of these sequences were subsequently included in the phylogenetic and diversity analyses and 71 were retained in the final genome-wide association (tissue localization (derived from the anatomical site of sample collection) and disease severity) analyses. The quality filtering process is illustrated in *Figure 1* and detailed in Methods. Briefly, we used standard GATK SNP-calling algorithms where $>10x$ mean depth of coverage is defined as a threshold value and performs well in variant calling, is highly sensitive and has a false positive rate of $<0.05\%$ ^{43,44}.

A total of 1034 unique SNP sites were identified within the 126 Bijagós *Ct* genomes relative to the reference strain *Ct A/HAR-13*. Following application of further threshold criteria based on minor allele frequency (MAF) and genome-wide missing data thresholds, we retained only high quality genomic data in the final association analyses (129 SNPs from 71 sequences). There were no significant differences between the 71 retained and the 55 excluded sequences with respect to demographic characteristics, bacterial load, disease severity scores or geographical location (*Table 1*). Clinical and demographic details of the survey participants in whom we did not identify *Ct* infection have been published previously⁴⁵. Of the ten SNPs initially identified within the *Ct* plasmid sequences, none fulfilled the quality filtering criteria and were not retained for the genome-wide association analyses.

Ocular C. trachomatis Phylogeny and Diversity

For the phylogeny and diversity analyses 81 Bijagós *Ct* sequences were included on the basis of quality filtering criteria described in detail in *Figure 1*. SNP-based phylogenetic trees constructed using all 1034 SNPs for sequences above 10x coverage ($n=81$), with 54 published *Ct* reference genomes, are shown in *Figure 2*.

The Bijagós sequences are situated within the T2 ocular monophyletic lineage with all other ocular *Ct* sequences⁴⁶ except those described by Andersson *et al.*¹⁰. However, our population-based collection of ocular *Ct* sequences has much greater diversity at whole genome resolution than previously demonstrated in African trachoma isolates^{4,8}. We used a pairwise diversity (π) metric to compare two populations of ocular *Ct* from regions with similar trachoma endemicity and studies with similar design, sample size and available epidemiological metadata. These data

show much greater genomic diversity in the Bijagós ocular *Ct* sequences ($\pi=0.07167$) compared to the Tanzanian (Rombo) ocular *Ct* sequences ($\pi =0.00047$).

By *ompA* genotyping, 73 of the Bijagós sequences are genotype A and eight are genotype B, supporting their classical ocular nature (*Supplementary Information S1*). The high resolution of WGS data obtained directly from clinical samples captures diversity that may be useful in strain classification, particularly as we found some evidence of clustering at village level, although the very small number of sequences per village means that it is not possible to provide accurate estimates of clustering in this study (*Figure 3*).

Homoplastic SNPs and regions affected by recombination are shown in *Supplementary Information S2 (a)*. Removal of these regions of recombination identified using the pairwise homoplasmy index had no effect on phylogenetic relationships. Additionally, a site-wise log likelihood plot demonstrated that there was no clear genomic region where there was significant lack of confidence in the tree construction due to recombination (*Supplementary Information S2 (b)*). Whether regions containing recombination were included or excluded, tree topology remained essentially identical, indicating that branching order is not affected by the removal of these regions.

Genome-wide analysis of C. trachomatis localization

Candidate genes thought to be involved in or indicative of ocular localization or preference were examined to further characterize this population of ocular *Ct*. Polymorphisms and truncations in the tryptophan operon have previously been implicated in the inability of ocular *Ct* to infect and survive in the genital tract⁵. All sequences contained mutations in *trpA* resulting in truncation. The majority (80/81)

were truncated at the previously characterised deletion at position 533⁵.

Polymorphisms in *trpB* and *trpR* were less common (*Supplementary Information S3*).

The variable domain structure of the translocated actin-recruiting phosphoprotein (*tarP*) has also been implicated in tropism⁴⁷. Ocular strains possess more actin-binding domains (three or four) and fewer tyrosine repeat regions (between one and three). Urogenital strain *tarP* sequences have low copy numbers of both and LGV strain sequences have additional tyrosine repeat regions. In this study, all sequences contain the expected three tyrosine repeat regions and three or four actin-binding domains (*Supplementary Information S3*).

The nine virulence associated polymorphic membrane proteins (Pmp) are variably related to tissue preference with all encoding genes except *pmpA*, *pmpD* and *pmpE* clustering by tissue location²⁰. In this population all phylogenies of the six tropism-clustering *pmps* show that all sequences cluster with other ocular sequences (*Supplementary Information S4*).

Permutation-based re-sampling methods, commonly used in GWAS analyses, were used to account for multiple comparisons⁴⁸⁻⁵¹. 1007 SNPs were tested in 157 *Ct* sequences (*Figure 1*) for association with ocular localization (defined by anatomical site of sample collection), comparing eight ocular, 17 urogenital and 13 LGV strains (*Figure 4(a)*). One hundred and five SNPs were significantly associated with ocular localization ($p^* < 0.05$) of which 21 were non-synonymous (details in *Table 2(a)* and *Supplementary Information S5*). These were within a number of genes known to be polymorphic, genes previously identified as tropism-associated (*CTA0156*, *CTA0498/tarP* and *CTA0743/pbpB*) and virulence factors (*CTA0498/tarP* and *CTA0884/pmpD*). No predicted protein localization was over-represented in the

ocular localization-related SNPs ($p=0.6174$), however early and very-late expressed genes were over-represented ($p=0.0197$).

Markers of disease severity in ocular *C. trachomatis* infection

Using permutation-based resampling methods, eight SNPs were found to be significantly associated with disease severity (*Figure 4(b)*). Seven of these are in coding regions (relative to *Ct A/HAR-13*). Five are present at nucleotide positions 465330 (OR=0.13, $p^*=0.037$), 32779 (OR=0.12, $p^*=0.032$), 875804 (OR=0.10, $p^*=0.024$), 939488 (OR=0.10, $p^*=0.026$) and 1028728 (OR=0.08, $p^*=0.013$) (where p^* is the permuted p-value with a genome-wide threshold of 0.05) representing synonymous codon changes within the genes *yjfH*, *trmD*, *alaS*, *glgA* and *pmpE* respectively. Three further genome-wide significant synonymous SNPs were present at positions 827184 (OR=0.3, $p^*=0.041$) within the predicted coding sequence (CDS) *CTA0744*, 285610 (OR=0.12, $p^*=0.027$) within *CTA0273* and 787841 (OR=0.13, $p^*=0.043$) in the intergenic region between loci *CTA0744-CTA0745* (*Table 2(b)* and *Supplementary Information S6*).

DISCUSSION

This is the first collection of clinical ocular *Ct* WGS from a single trachoma-endemic population to be characterized, enabling us to describe the population diversity of naturally occurring *Ct* in a treatment-naïve population. We used detailed clinical grading combined with microbial quantitation to perform a genome-wide association study (GWAS) and investigated associations between *Ct* polymorphisms with ocular localization and disease severity in trachoma.

Unlike the recently published Australian *Ct* sequences¹⁰, all Bijagós sequences clustered as expected within the T2 ocular clade derived from an urogenital ancestor^{46,52}, each with loci typically associated with ocular tissue localization (*trpA* and *tarP*). Although the Bijagós sequences conform to the classical ocular genotype, the phylogenetic data show greater than expected diversity compared to historical reference strains of ocular *Ct*⁴ and a population of clinical ocular *Ct* sequences obtained from cultured clinical conjunctival swab specimens collected from another African trachoma-endemic population⁵³ (*Supplementary Information S7*). Our use of direct WGS from clinical samples reveals the natural diversity of a population-based collection of endemic treatment-naïve ocular *Ct* infections. This diversity may indicate genome-wide selection for advantageous mutations as demonstrated in other pathogens⁵⁴ or evidence that these are remnants of a previously larger and more diverse population in West Africa.

The apparent village-level clustering provides new evidence that WGS has the necessary molecular resolution to fully investigate *Ct* transmission. Although the number of sequences from each village were very small, overall *Ct* genomic diversity supports our hypothesis of ongoing or recent transmission, since diversity requires mutation, recombination and gene flow. The data from this study demonstrates such mutation and indicates that WGS data may be useful in defining transmission networks and developing transmission maps, which have not been adequately defined using alternative *Ct* genotyping systems. Whole genome mapping has previously been shown to be a useful tool in the analysis of outbreaks and bacterial pathogen transmission^{55,56} and thus has multiple potential applications in epidemiological analysis and transmission studies. However, greater numbers of sequences per village are required to validate this finding.

Such diversity is likely to be representative of recombination present in *Ct*⁵⁷. Genome-wide recombination was common and widespread within these sequences. Extensive recombination has been noted in previous studies, and is thought to be a source of diversification with possible interstrain recombination^{4,57}. Recombination may represent fixation of recombination in regions that are under diversifying selection pressure⁴.

Recently a handful of bacterial GWAS studies have provided insight into the genetic basis of bacterial host preference, antibiotic resistance, and virulence⁵⁸⁻⁶³. Until now, most inferences regarding disease-modifying virulence factors in chlamydial infection have been derived from a limited number of comparative genomic studies where only a few virulence factors were associated with disease severity. Chlamydial genomic association data have previously been used to highlight genes potentially involved in pathoadaptation^{10,64} and tissue localization⁶⁵.

In the current GWAS we found 21 genome-wide significant non-synonymous SNPs associated with ocular localization and eight genome-wide significant synonymous SNPs associated with disease severity.

Confidence that new SNPs identified in the ocular localization GWAS are candidate markers of pathoadaptation is supported by the observation that half of the SNPs identified have previously been described as polymorphic or recombinant within *Ct* and the ocular serovars^{8,66-68}. Genes expressed early in the *Ct* developmental cycle (with a peak at six hours post-infection [HPI]) or very late in the *Ct* developmental cycle (with a peak after 24 HPI) were over-represented, supporting the hypothesis that early events in infection and intracellular growth are crucial events in *Ct* survival and pathogenicity. Amongst the early-expressed genes are *CTA0156* (encoding early endosomal antigen 1 [EEA1]⁶⁹), *CTA0498* (encoding translocated

actin-recruiting phosphoprotein [*tarP*]⁷⁰) and *CTA0884* (encoding polymorphic membrane protein D [PmpD]⁷¹), which have identified roles in entry to and initial interactions with host cells.

The eight disease severity-associated SNPs are within less well-characterized genes, although there is some evidence of correlation of expression in the *Ct* developmental cycle (with a peak between 18 and 24 HPI). Apart from *pmpE*, there is a paucity of published data showing polymorphism in these genes. This suggests that these SNPs may be important in ocular *Ct* pathogenesis, rather than in longer-term chlamydial evolution. Three of these genes are putative *Ct* virulence factors, with functions in nutrient acquisition (*glgA*^{24,28,72}), host-cell adhesion (*pmpE*⁷³) and response to IFN γ -induced stress (*trmD*⁶⁹). Homologues of *alaS*^{74,75} and *CTA0273*^{76,77} are known virulence factors in related Gram-negative bacteria, suggesting that these genes are potentially important in *Ct* pathogenesis.

Transcriptome analysis of chlamydial growth *in vitro* has shown that there is highly upregulated gene expression of *trmD* (encoding a tRNA methyl-transferase) associated with growth in the presence of interferon gamma, thought to be important in the maintenance of chlamydial infection⁶⁹. *yjfh* (renamed *rlmB*) is phylogenetically related to the TrmD family and encodes the protein RlmB, which is important for the synthesis and assembly of the components of the ribosome⁷⁸. In *Escherichia coli*, *Haemophilus influenzae* and *Mycoplasma genitalium*, RlmB catalyses the methylation of guanosine 2251 in 23S rRNA, which is of importance in peptidyl tRNA recognition but is not essential for bacterial growth^{78,79}. *alaS* encodes a tRNA ligase of the class II aminoacyl-tRNA synthetase family involved in cytoplasmic protein biosynthesis. It is not known to have virulence associations in chlamydial infection, but has been described as a component of a virulence operon in *Haemophilus*

*ducreyi*⁷⁴ and *H. influenzae*⁷⁵. The CDS *CTA0273* encodes a predicted inner membrane protein translocase component of the autotransporter YidC, an inner membrane insertase important in virulence in *E. coli*⁷⁶ and *Streptococcus mutans*⁷⁷. This is the first study suggesting that these loci may be important in disease severity and host-pathogen interactions in chlamydial infection. A summary of available literature for these key ocular localization and disease severity-associated SNPs is tabulated in *Supplementary Information S8*. We cannot speculate further on the effect polymorphisms on expression. It is possible that the synonymous disease severity-associated SNPs are linkage-markers for disease-causing alleles that were not included in the GWAS. For both analyses, further mechanistic studies are required to establish causality, validity and to fully understand the nature of the associations presented in this study.

Though we were intrinsically limited to those cases where infection was detectable and from which we were able to obtain *Ct* WGS data, our population-based treatment naïve sample attempts to provide a representative picture of what is observed in ocular *Ct* infection. We acknowledge that there may be *Ct* genotypes that are cleared by the immune system such that we do not capture them a cross-sectional study. We are limited to the small sample size in this study, but attempt to address the issues of statistical power and multiple testing by using a bi-dimensional conjunctival phenotype and permutation-based multivariable regression analysis. To date the majority of published microbial GWAS have sample sizes under 500⁸⁰, including several key studies examining virulence⁵⁹ and drug-resistance⁶⁰ in *Staphylococcus aureus* with sample sizes of 75 and 90 respectively.

The potential of bacterial GWAS has only recently been realized, and despite the limitations with sample size, its use to study *Ct* in this way is particularly

important, since *in vitro* models are intrinsically difficult to develop and it has not been possible to study urogenital *Ct* in the same way due to the lack of a clearly defined *in vivo* disease phenotype. The genomic markers identified in this study provide important direction for validation through *in vitro* functional studies and a unique opportunity to understand host-pathogen interactions likely to be important in *Ct* pathogenesis in humans. The greater than expected diversity within this population of naturally circulating ocular *Ct* and the clustering at village-level demonstrates the potential utility of WGS in epidemiological and clinical studies. This will enable us to understand transmission in both ocular and urogenital *Ct* infection and will have significant public health implications in preventing and eliminating chlamydial disease in humans.

METHODS

Survey, Clinical Examination and Sample collection

Survey, clinical examination and sample collection methods have been described previously^{45,81}. Briefly, we conducted a cross-sectional population-based survey in trachoma-endemic communities on the Bijagós Archipelago of Guinea Bissau. The upper tarsal conjunctivae of each consenting participant were examined, digital photographs were taken, a clinical trachoma grade was assigned and two sequential conjunctival swabs were obtained from the left upper tarsal conjunctiva of each individual using a standardized method⁴⁵. DNA was extracted and *Ct omcB* (genomic) copies/swab quantified from the second conjunctival swab using droplet digital PCR (ddPCR)^{81,82}.

We used the modified FPC (Follicles, Papillary hypertrophy, Conjunctival scarring) grading system for trachoma³⁹. The modified FPC system allows detailed

scoring of the conjunctiva for the presence of follicles (F score), papillary hypertrophy (conjunctival inflammation) (P score) and conjunctival scarring (C score), receiving a grade of 0-3 for each parameter. A single validated grader conducted the examinations and these were verified by an expert grader (masked to the field grades and ddPCR results) using the digital photographs. Grader concordance was measured using Cohen's Kappa, where a Kappa > 0.9 was used as the threshold to indicate good agreement.

Conjunctival inflammation (P score) is known to have a strong association with *Ct* bacterial load in this and other populations^{35,83-86}. For this study we used principal component analysis (PCA) to combine the presence of inflammation (defined by the P score using the FPC trachoma grading system³⁹) with *Ct* bacterial load (defined by tertile cut-offs illustrated in *Supplementary Information S9*)⁸⁷. The conjunctival disease phenotype is a dimension reduction of these two variables defining what we observed in the conjunctiva at the time of sampling (*Figure 5*). Dimension reduction using PCA to define complex disease phenotypes in GWAS is well-recognised, as it allows multiple traits to be included to capture a more complex phenotype and accounts for correlation between traits. This approach therefore may reveal novel loci or pathways that would not be evident in single-trait GWAS, where the full extent of genetic variation cannot be captured⁴⁰.

Preparation of chlamydial DNA for cell culture

For eight specimens, whole genome sequence (WGS) data were obtained following *Ct* isolation in cell culture (from the first conjunctival swab). Briefly, samples were isolated in McCoy cell culture by removing 100µl eluate from the original swab with direct inoculation onto a glass coverslip within a bijou containing

Dulbecco's modified Eagles' Medium (DMEM). The inocula were centrifuged onto cell cultures at 1800rpm for 30 minutes. Following centrifugation the cell culture supernatant was removed and cycloheximide-containing DMEM added to infected cells which were then incubated at 37⁰C in 5% CO₂ for three days. Viable *Ct* elementary bodies (EB) were observed by phase contrast microscopy. Cells were harvested and further passaged every three days until all isolates reached a multiplicity of infection between 50-90% in 2xT25 flasks. Each isolate was prepared and EBs purified as described previously⁸⁸. DNA was extracted from purified EBs using the Promega Wizard Genomic Purification kit according to the manufacturer's protocol⁸⁹.

Pre-sequencing target enrichment

For the remaining specimens ($n=118$), WGS data were obtained directly from clinical samples. DNA baits spanning the length of the *Ct* genome were compiled by SureDesign and synthesized by SureSelect^{XT} (Agilent Technologies, UK). *Ct* DNA extracted from clinical samples was quantified and carrier human genomic DNA added to obtain a total of 3µg input for library preparation. DNA was sheared using a Covaris E210 acoustic focusing unit³¹. End-repair, non-templated addition of 3'-A adapter ligation, hybridisation, enrichment PCR and all post- reaction clean-up steps were performed according to the SureSelect^{XT} Illumina Paired-End Sequencing Library protocol (V1.4.1 Sept 2012). All recommended quality control measures were performed between steps.

Whole genome sequencing and sequence quality filtering

DNA was sequenced at the Wellcome Trust Sanger Institute using Illumina paired-end technology (Illumina GAI or HiSeq 2000). All 126 sequences passed standard FastQC quality control criteria⁹⁰. Sequences were aligned to the most closely related reference genome, *Chlamydia trachomatis A/HAR-13* (GenBank Accession Number NC_007429.1 and plasmid GenBank Accession Number NC_007430.1), using BWA⁹¹. SAMtools/BCFtools (SAMtools v1.3.1)⁹² and GATK⁹³ were used to call SNPs. We used standard GATK SNP calling algorithms, where >10x depth of coverage is routinely used as the threshold value^{93,94}. This has been shown to be adequate for SNP calling in this context^{43,44,46,94}.

Variants were selected as the intersection dataset between those obtained using both SNP callers and SNPs were further quality-filtered. SNP alleles were called using an alternative coverage-based approach where a missing call was assigned to a site if the total coverage was less than 20x depth or where one of the four nucleotides accounted for at least 80% total coverage⁹⁵. There was a clear relationship between the mean depth of coverage and proportion of missing calls, based on which we retained sequences with greater than 10x mean depth of coverage over the whole genome (81 sequences retained).

Heterozygous calls were removed and SNPs with a minor allele frequency of less than 25% were removed. Samples with greater than 25% genome-wide missing data and 30% missing data per SNP were excluded from the analysis (n=10, 71 sequences retained). The quality assessment and filtering process is shown in *Figure 1*. Detail of WGS data is contained in *Supplementary Information S10*.

Phylogenetic Reconstruction

Samples were mapped to the ocular reference strain *Ct A/HAR-13* and SNPs were called as described above. Phylogenies were computed using RAxML version 7.8.2⁹⁶ from a variable sites alignment using a GTR+gamma model and are midpoint rooted. Recombination is known to occur in *Ct*^{4,6} and can be problematic in constructing phylogeny. We applied three compatibility-based recombination detection methods to detect regions of recombination using PhiPack⁹⁷: the pairwise homoplasmy index (Phi), the maximum Chi2 and the neighbour similarity score (NSS) across the genome alignment. We also examined the confidence in the phylogenetic tree by computing RAxML site-based likelihood scores⁹⁶. Phylogenetic trees were examined adjusting for recombination using the methods described above.

Additionally, sequence data for the tryptophan operon (*CTA0182* and *CTA0184-CTA0186*), *tarP* (*CTA0498*), nine polymorphic membrane proteins (*CTA0447-CTA0449*, *CTA0884*, *CTA0949-CTA0952* and *CTA0954*) and *ompA* (*CTA0742*) were extracted from the 81 ocular *Ct* sequences from Guinea Bissau retained after quality control filtering described above, 48 ocular sequences originating from a study conducted in Kahe village, Rombo District, Tanzania⁵³ and 38 publicly available reference sequences. Phylogenies were constructed as described above.

Polymorphisms, insertions and deletions (INDELs) and truncations for the tryptophan operon were manually determined from aligned sequences using SeaView⁹⁸. Tyrosine repeat regions and actin-binding domains in *tarP* were found using RADAR⁹⁹ and Pfam¹⁰⁰ respectively.

Pairwise diversity

A comparison was made between the two population-based *Ct* sequence data sets from the Bijagós (Guinea Bissau) and Rombo (Tanzania) sequences whereby short read data from the 81 Bijagós sequences and 48 Rombo sequences were mapped against *Ct A/HAR-13* using SAMtools. Within population pairwise nucleotide diversity was calculated using the formula:

$$\pi = 2 \times \sum_{i=1}^n \sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

where n is the number of sequences, x is the frequency of sequences i and j and π_{ij} is the number of nucleotide differences per site between sequences i and j ¹⁰¹. Frequency of sequences was considered uniform within the populations and sites with missing calls were excluded on a per sequence basis.

Genome-Wide Association Analyses

To investigate the association between *Ct* polymorphisms with ocular localization and clinical disease severity, we used permutation-based logistic regression methods, which are powerful and well-recognised tools in GWAS, allowing for adjustment for population structure, age and gender in the model and accounting for multiple testing⁴⁸⁻⁵¹.

We used permutation analyses of 100,024 phenotypic re-samplings, where the distribution of the p-value was approximated by simulating data sets through randomisation under the null hypothesis of no association between phenotype and genotype. Genome-wide significance was determined as $p^* \leq 0.05$, where p^* was defined as the fraction of re-sampled (simulated) data that returned p-values that were

less than or equal to the p-values observed in the data⁸⁷. All analyses were conducted using the R statistical package v3.0.2 (The R Foundation for Statistical Computing, <http://www.r-project.org>) using MASS, GLM and lsr. All R script used for these analyses is contained within *Supplementary Information S11* and is released as a CC-BY open resource (CC-BY-SA 3.0).

Ocular localization

Tissue localization is defined as the localization (or presence) of a detectable *Ct* infection to either the conjunctival epithelium or the urogenital tract. Short read data from the 129 clinical ocular sequences from the pairwise diversity analysis and 38 publicly available reference sequences from ocular (n=8), urogenital (n=17) and rectal (n=13) sites were mapped against *Ct A/HAR-13* using SAMtools. Only polymorphic sites were retained and SNPs were filtered as described above. The final analysis includes 1007 SNPs from 157 sequences, a phylogeny of which is contained within *Supplementary Information S7*. A permutation-based generalized linear regression model was used to test the association between collection site (ocular or urogenital tissue localization) and polymorphic sites. For each SNP the standard error for the t statistic was estimated from the model and used to calculate the odds ratios (OR) and 95% confidence intervals. A χ^2 test was used to determine the association between ocular localization-associated SNPs and both gene expression stage and predicted localization of the encoded proteins. Developmental cycle expression stage for each transcript was based on data and groupings from Belland *et al.*⁶⁹. Predicted localization of expressed proteins was defined using the consensus from three predictions using Cello¹⁰², pSORTB¹⁰³ and LocTree3¹⁰⁴.

Clinical Disease Severity

A permutation-based ordinal logistic regression model was used to test the association between the disease severity score (using the *in vivo* conjunctival phenotype defined previously) and polymorphic sites. The final analysis includes 129 SNPs from 71 sequences derived as described in *Figure 1*. For each SNP the standard error for the t-statistic was estimated from the model and used to calculate the odds ratios (OR) and 95% confidence intervals. Individuals' age and gender were included as a covariate to the regression analysis.

We investigated the effect of population structure on the results of the GWAS analysis using PCA¹⁰⁵. The first three principal components (PC) captured the majority of structural variation but including these in the model had no effect and therefore these were not included in the final model.

We corrected for genomic inflation if the occurrence of a polymorphism in the population was over 90% or there was a minor allele frequency of 3%.

DATA AVAILABILITY

All sequence data is available from the European Bioinformatics Institute (EBI) short read archive. See *Supplementary Information S12* for details and accession numbers.

ACKNOWLEDGEMENTS

AL was funded by the Wellcome Trust through a Clinical Research Training Fellowship (grant number 097330/Z/11/Z). MH and ChR were funded by a Wellcome Trust Program Grant (grant number 079246/Z/06/Z). ChR was funded by the Wellcome Trust Institutional Strategic Support Fund (grant number 105609/Z/14/Z). Work undertaken by the Wellcome Trust Sanger Institute by NRT, HSS and JH was

supported by the Wellcome Trust (grant number 098051 and the Wellcome Trust Institutional Strategic Support Fund (grant number 105609/Z/14/Z)). TGC is funded by the Medical Research Council UK (grant numbers MR/K000551/1, MR/M01360X/1, MR/N010469/1). JP is funded by a BBSRC PhD studentship, and FC and HP were funded by Bloomsbury Colleges Research Fund PhD studentships. We extend thanks to colleagues at the Programa Nacional de Saúde de Visão in the Ministério de Saúde Publica in Bissau, the study participants and dedicated field research team in Guinea Bissau and the MRC Unit The Gambia for their support and collaboration in this work. All communities received treatment for endemic trachoma in accordance with WHO and national policies following the survey.

FOOTNOTES

ARL/RLB/MJH designed the study. ARL/SEB/EC/MN conducted the field study. ARL/ChR/SEB conducted the molecular laboratory work. LTC/IN performed the Chlamydial cell culture. HSS/JH designed and performed the whole genome sequencing and QC. ARL/ChR/HP/FC/TC conducted the GWAS analysis.

HP/JP/SH/JH/HSS supported the phylogenetic analysis.

ARL/HP/MJH/DCWM/TC/NRT wrote the paper. All authors have contributed to and reviewed the manuscript.

COMPETING FINANCIAL INTEREST

All authors declare no competing financial interests.

Table 1. Characteristics of ocular *Chlamydia trachomatis* sequences included in the disease severity association analysis

Sequence ID	Sample ID	Average Depth of Coverage	% Missing Reads*	Gender	Age (years)	Island of Origin	Village Code	Ocular Load**	P Score [‡]
11152_3_1	14344	764	0.35%	M	4	Canhabaque	33	202632	1
11152_3_10	17347	121	0.21%	M	5	Bubaque	17	69093	2
11152_3_11	4422	19	19.95%	F	2	Bubaque	12	68782	2
11152_3_12	11231	68	2.24%	M	0	Soga	43	64036	1
11152_3_13	15631	21	14.93%	F	2	Canhabaque	33	55749	3
11152_3_14	6105	1664	0.05%	F	1	Bubaque	14	55202	3
11152_3_15	12628	191	0.10%	F	12	Canhabaque	29	54651	2
11152_3_16	7524	2065	0.14%	M	10	Canhabaque	35	54539	2
11152_3_17	5016	61	0.44%	F	1	Bubaque	15	46510	2
11152_3_18	1485	44	1.21%	F	4	Canhabaque	27	45929	1
11152_3_19	15554	825	0.06%	F	1	Canhabaque	33	44052	2
11152_3_20	6094	3070	0.00%	F	3	Bubaque	14	42917	2
11152_3_22	5082	51	0.81%	M	6	Bubaque	15	42427	1
11152_3_23	12969	3643	1.81%	F	3	Canhabaque	29	41308	3
11152_3_25	8140	246	0.36%	M	13	Bubaque	20	39816	2
11152_3_26	6083	2746	0.00%	F	23	Bubaque	14	38771	3
11152_3_27	16621	1664	0.00%	M	3	Canhabaque	37	33514	3
11152_3_28	16852	143	0.16%	M	5	Canhabaque	38	31228	2
11152_3_29	16588	53	0.81%	M	6	Canhabaque	37	29991	1
11152_3_3	4180	51	0.92%	M	2	Bubaque	12	140693	2
11152_3_30	7612	107	0.44%	F	3	Canhabaque	35	28528	2
11152_3_31	6985	177	0.10%	M	6	Bubaque	17	27924	2
11152_3_32	4411	24	9.68%	F	1	Bubaque	12	27584	2
11152_3_33	4257	381	0.06%	M	0	Bubaque	12	24033	3

11152_3_34	4400	48	0.98%	M	6	Bubaque	12	23435	2
11152_3_35	15180	571	0.35%	F	7	Canhabaque	33	23254	0
11152_3_36	13596	496	0.06%	M	18	Canhabaque	23	22098	3
11152_3_37	1672	20	18.42%	M	6	Canhabaque	25	21630	3
11152_3_38	5181	81	0.32%	M	4	Bubaque	15	21339	2
11152_3_39	15532	243	0.08%	F	25	Canhabaque	33	21174	2
11152_3_4	8074	150	0.13%	M	4	Bubaque	18	131175	2
11152_3_40	16984	145	0.19%	M	4	Canhabaque	21	20113	1
11152_3_41	1881	37	2.71%	F	1	Canhabaque	32	15963	2
11152_3_42	10032	101	0.16%	M	2	Soga	42	15706	1
11152_3_43	8492	70	2.60%	M	1	Rubane	45	15582	2
11152_3_44	13585	31	4.97%	M	23	Canhabaque	23	15417	3
11152_3_48	7535	61	0.84%	M	18	Canhabaque	35	13439	3
11152_3_5	7095	235	0.44%	F	4	Bubaque	17	105453	3
11152_3_50	6028	46	1.24%	F	4	Bubaque	14	12961	2
11152_3_52	10021	20	16.15%	F	6	Soga	42	11840	1
11152_3_55	12650	59	0.54%	M	6	Canhabaque	29	9001	2
11152_3_57	8965	21	16.60%	M	27	Soga	43	7336	1
11152_3_58	5104	33	3.68%	M	2	Bubaque	15	7203	2
11152_3_6	16599	52	0.73%	M	9	Canhabaque	37	96333	2
11152_3_62	7062	22	13.41%	F	4	Bubaque	17	6986	3
11152_3_63	8778	17	25.47%	F	11	Rubane	46	6760	3
11152_3_66	1892	45	1.25%	F	2	Canhabaque	32	6374	1
11152_3_7	10747	581	1.82%	F	3	Soga	44	82916	2
11152_3_70	13189	25	8.87%	F	3	Canhabaque	24	4703	1
11152_3_74	15499	24	10.49%	M	5	Canhabaque	33	4226	1
11152_3_76	726	417	0.06%	F	3	Canhabaque	26	3753	0

11152_3_77	7579	105	0.52%	F	5	Canhabaque	35	3468	1
11152_3_78	12089	16	27.78%	F	13	Canhabaque	47	3203	2
11152_3_8	6996	38	2.03%	M	3	Bubaque	17	82614	1
11152_3_88	748	163	0.10%	F	2	Canhabaque	26	1636	0
11152_3_9	10967	20	17.52%	F	2	Soga	44	81124	3
11152_3_92	1463	73	0.30%	F	42	Canhabaque	27	1273	2
13108_1_14	24519	51	2.81%	M	2	Rubane	45	29040	3
13108_1_15	6941	33	1.81%	M	36	Bubaque	17	13155	1
13108_1_7	25124	27	5.27%	M	4	Canhabaque	22	21750	3
13108_1_9	22154	18	20.56%	F	5	Soga	43	14349	1
8422_8_49	2353	39	5.70%	M	11	Canhabaque	35	96889	2
8422_8_50	2366	82	1.08%	M	1	Canhabaque	35	289778	2
9471_4_86	12980	287	1.90%	M	4	Canhabaque	29	85456	1
9471_4_87	15367	215	0.46%	M	1	Canhabaque	33	99064	1
9471_4_88	15543	192	0.11%	F	23	Canhabaque	33	49125	1
9471_4_89	1870	119	0.14%	M	3	Canhabaque	32	158548	3
9471_4_90	2145	111	0.11%	M	15	Canhabaque	32	140297	2
9471_4_91	4158	94	0.14%	M	4	Bubaque	12	63654	1
9471_4_92	4169	85	0.13%	F	3	Bubaque	12	274835	2
9471_4_93	7590	242	0.51%	F	1	Canhabaque	35	128025	3

Sequences ($n=55$) were excluded from the association analysis if there was a) $<10x$ coverage b) $>25\%$ missing reads genome-wide and c) $>25\%$ missing (N) calls at the single nucleotide polymorphism (SNP) locus. Coverage and missing data were correlated and resulted in exclusion of the same samples irrespective of criteria chosen. 71 sequences were retained in the final disease severity analysis. **Ocular *C. trachomatis* load = *omcB* (*C. trachomatis* genome) copies per conjunctival swab measured using droplet digital PCR. *** P score = Conjunctival inflammation score (0-3) using the modified FPC (Follicles, Papillary Hypertrophy, Conjunctival Scarring) grading system for trachoma³⁹.

Table 2. SNPs across the *Chlamydia trachomatis* genome identified using permutation-based genome-wide association analysis for (A) ocular localization (non-synonymous only) and (B) disease severity

(A)

SNP position	ocular allele (%)	uro-genital allele (%)	name A/HAR13	CDS/ NCR	p-value	p*	OR	95% CI	t	SE(t)	MAF	N calls at locus	ocular AA	urogenital AA	
168413	A (61.54)	G (93.33)	CTA_0156	CDS	5E-05	1E-04	21.56	6.11	137.25	4.07	0.75	0.50	0.04	H	R
95863	A (60.47)	G (86.67)	CTA_0087	CDS	7E-05	1E-04	9.56	3.47	33.86	3.98	0.57	0.49	0.02	E	G
785083	A (62.20)	G (96.67)	pbpB	CDS	2E-04	1E-04	45.92	9.34	831.41	3.70	1.03	0.49	0.05	I	V
777345	A (58.59)	G (96.67)	karG	CDS	3E-04	1E-04	40.71	8.29	736.79	3.59	1.03	0.47	0.04	Y	H
156982	C (51.54)	T (90.00)	oppA_1	CDS	4E-04	1E-04	9.44	3.13	40.92	3.54	0.63	0.43	0.02	V	I
637206	A (56.59)	C (96.67)	sctR	CDS	5E-04	1E-04	36.25	7.39	655.80	3.48	1.03	0.45	0.03	K	Q
157069	A (51.54)	G (86.67)	oppA_1	CDS	7E-04	3E-04	6.81	2.48	24.09	3.39	0.57	0.44	0.02	S	P
367095	C (60.77)	T (73.33)	CTA_0348	CDS	1E-03	1E-03	4.23	1.81	10.82	3.20	0.45	0.46	0.01	T	I
544233	A (61.54)	G (73.33)	CTA_0510	CDS	1E-03	3E-04	4.23	1.81	10.82	3.20	0.45	0.46	0.02	R	G
954865	A (59.69)	G (73.33)	pmpD	CDS	2E-03	1E-04	4.04	1.73	10.33	3.10	0.45	0.46	0.04	E	G
969418	C (59.06)	T (73.33)	sucD	CDS	2E-03	1E-04	3.94	1.68	10.07	3.04	0.45	0.46	0.03	T	I
544610	A (61.54)	G (70.00)	atoS	CDS	3E-03	1E-03	3.59	1.56	8.85	2.92	0.44	0.45	0.01	D	G
543548	T (60.63)	C (70.00)	CTA_0508	CDS	5E-03	1E-04	0.29	0.12	0.67	-2.83	0.44	0.45	0.06	F	S
969583	T (58.73)	C (70.00)	sucD	CDS	7E-03	1E-04	0.30	0.12	0.70	-2.72	0.44	0.46	0.04	L	P
44611	C (60.63)	T (66.67)	CTA_0043	CDS	1E-02	1E-04	2.96	1.30	7.10	2.53	0.43	0.45	0.04	A	V
533906	T (74.62)	C (50.00)	CTA_0498	CDS	1E-02	9E-03	0.35	0.15	0.80	-2.51	0.42	0.31	0.01	L	P
295635	G (61.24)	A (63.33)	CTA_0284	CDS	2E-02	1E-04	0.38	0.16	0.86	-2.30	0.42	0.44	0.03	R	K
95527	C (60.77)	T (60.00)	CTA_0087	CDS	5E-02	4E-02	2.24	1.00	5.15	1.94	0.41	0.44	0.01	S	L
413567	A (60.47)	G (60.00)	CTA_0391	CDS	6E-02	1E-04	2.21	0.99	5.08	1.91	0.41	0.44	0.04	V	A
1027490	G (58.91)	T (60.00)	CTA_0948	CDS	7E-02	1E-04	2.13	0.96	4.91	1.83	0.41	0.45	0.01	P	Q
777183	T (58.59)	C (60.00)	karG	CDS	7E-02	1E-04	0.47	0.21	1.06	-1.80	0.41	0.45	0.04	I	V
168413	A (61.54)	G (93.33)	CTA_0156	CDS	5E-05	1E-04	21.56	6.11	137.2	4.07	0.75	0.50	0.04	H	R

(B)

SNP Position	Reference ALLELE	Alternative ALLELE	name A/HAR13	CDS/NCR	Strand	p*	p-value	t	SE(t)	OR	95% C.I. (UL)	(LL)	MAF	N Calls at Locus
1028728	C	T	pmpE	CDS	-	0.013	0.011	-2.550	0.555	0.078	0.026	0.232	0.310	7.042
875804	C	T	alaS	CDS	-	0.024	0.022	-2.298	0.530	0.100	0.036	0.284	0.310	4.225
939488	G	A	glgA	CDS	-	0.026	0.023	-2.273	0.491	0.103	0.039	0.270	0.479	4.225
285610	G	A	CTA_0273	CDS	-	0.027	0.034	-2.123	0.526	0.120	0.043	0.336	0.310	4.225
32779	G	A	trmD	CDS	+	0.032	0.031	-2.160	0.525	0.115	0.041	0.323	0.310	2.817
465330	C	G	yjfH	CDS	-	0.037	0.042	-2.032	0.519	0.131	0.047	0.362	0.310	1.408
787841	A	G	NA	inter	NA	0.038	0.038	-2.074	0.524	0.126	0.045	0.351	0.310	4.225
827184	A	G	CTA_0774	CDS	+	0.041	0.043	-2.020	0.516	0.133	0.048	0.365	0.310	1.408
22049	G	T	ileS	CDS	+	0.057	0.050	-1.962	0.505	0.141	0.052	0.378	0.324	4.225
152011	G	A	NA	inter	NA	0.058	0.050	-1.964	0.505	0.140	0.052	0.377	0.324	4.225
710787	A	C	CTA_0675	CDS	-	0.060	0.052	-1.941	0.517	0.144	0.052	0.396	0.310	4.225
19085	T	C	NA	inter	NA	0.061	0.060	-1.882	0.530	0.152	0.054	0.430	0.296	5.634
388175	G	A	CTA_0368	CDS	-	0.061	0.059	-1.889	0.524	0.151	0.054	0.422	0.296	1.408
696782	A	T	rpoD	CDS	-	0.064	0.062	-1.864	0.511	0.155	0.057	0.422	0.310	1.408
286636	C	T	lgt	CDS	-	0.065	0.061	-1.876	0.511	0.153	0.056	0.417	0.310	0.000
930453	C	T	mutS	CDS	-	0.067	0.061	-1.876	0.511	0.153	0.056	0.417	0.310	0.000
465525	C	T	CTA_0439	CDS	-	0.067	0.062	-1.865	0.472	0.155	0.061	0.391	0.493	1.408
60858	G	A	CTA_0057	CDS	-	0.068	0.070	-1.813	0.512	0.163	0.060	0.445	0.310	1.408
835039	G	A	CTA_0782	CDS	-	0.070	0.061	-1.876	0.511	0.153	0.056	0.417	0.310	0.000
19005	A	G	NA	inter	NA	0.071	0.071	-1.807	0.525	0.164	0.059	0.459	0.296	2.817
4554	A	G	gatB	CDS	+	0.071	0.070	-1.813	0.512	0.163	0.060	0.445	0.310	1.408
303590	C	A	murE	CDS	-	0.072	0.061	-1.876	0.511	0.153	0.056	0.417	0.310	0.000
215130	C	T	gyrA_1	CDS	-	0.072	0.062	-1.864	0.511	0.155	0.057	0.422	0.310	1.408
806382	C	T	CTA_0761	CDS	+	0.073	0.058	-1.896	0.530	0.150	0.053	0.424	0.296	4.225

778783	G	A	rrf	CDS	-	0.077	0.075	-1.780	0.502	0.169	0.063	0.451	0.324	2.817
136812	G	A	incF	CDS	+	0.079	0.075	-1.780	0.502	0.169	0.063	0.451	0.324	2.817
169573	G	A	CTA_0156	CDS	+	0.082	0.077	-1.771	0.523	0.170	0.061	0.474	0.310	9.859
956953	C	T	pmpD	CDS	+	0.082	0.072	-1.800	0.523	0.165	0.059	0.461	0.296	2.817
44990	A	G	ruvB	CDS	+	0.087	0.086	-1.718	0.493	0.179	0.068	0.472	0.338	2.817
62140	G	T	sucA	CDS	+	0.091	0.078	-1.760	0.502	0.172	0.064	0.461	0.324	5.634
542521	G	A	CTA_0507	CDS	-	0.092	0.090	-1.696	0.494	0.183	0.070	0.483	0.338	2.817
181019	C	A	CTA_0164	CDS	-	0.095	0.096	-1.666	0.494	0.189	0.072	0.498	0.338	4.225
151156	C	G	CTA_0140	CDS	-	0.096	0.077	1.770	0.502	5.871	2.195	15.703	0.324	4.225
1028728	C	A	pmpE	CDS	-	0.01	0.011	-2.550	0.555	0.08	0.03	0.23	0.31	13.58%
1028728	C	T	pmpE	CDS	-	0.0134	0.0108	-2.5504	0.5550	0.0781	0.0263	0.2317	0.3099	7.0423
875804	C	T	alaS	CDS	-	0.0242	0.0216	-2.2981	0.5295	0.1005	0.0356	0.2836	0.3099	4.2254
939488	G	A	glgA	CDS	-	0.0259	0.0230	-2.2727	0.4906	0.1030	0.0394	0.2695	0.4789	4.2254
285610	G	A	CTA_0273	CDS	-	0.0269	0.0338	-2.1226	0.5264	0.1197	0.0427	0.3359	0.3099	4.2254
32779	G	A	trmD	CDS	+	0.0318	0.0308	-2.1596	0.5248	0.1154	0.0412	0.3227	0.3099	2.8169
465330	C	G	yjfH	CDS	-	0.0370	0.0422	-2.0315	0.5187	0.1311	0.0474	0.3625	0.3099	1.4085
787841	A	G	NA	inter	NA	0.0377	0.0381	-2.0742	0.5236	0.1257	0.0450	0.3506	0.3099	4.2254
827184	A	G	CTA_0774	CDS	+	0.0413	0.0433	-2.0203	0.5164	0.1326	0.0482	0.3648	0.3099	1.4085
22049	G	T	ileS	CDS	+	0.0568	0.0497	-1.9624	0.5052	0.1405	0.0522	0.3782	0.3239	4.2254
152011	G	A	NA	inter	NA	0.0578	0.0495	-1.9642	0.5051	0.1403	0.0521	0.3775	0.3239	4.2254
710787	A	C	CTA_0675	CDS	-	0.0605	0.0523	-1.9409	0.5174	0.1436	0.0521	0.3958	0.3099	4.2254
19085	T	C	NA	inter	NA	0.0608	0.0598	-1.8819	0.5298	0.1523	0.0539	0.4302	0.2958	5.6338
388175	G	A	CTA_0368	CDS	-	0.0610	0.0589	-1.8889	0.5238	0.1512	0.0542	0.4222	0.2958	1.4085
696782	A	T	rpoD	CDS	-	0.0638	0.0623	-1.8643	0.5114	0.1550	0.0569	0.4223	0.3099	1.4085
286636	C	T	lgt	CDS	-	0.0654	0.0606	-1.8764	0.5113	0.1531	0.0562	0.4172	0.3099	0.0000
930453	C	T	mutS	CDS	-	0.0668	0.0606	-1.8764	0.5113	0.1531	0.0562	0.4172	0.3099	0.0000
465525	C	T	CTA_0439	CDS	-	0.0670	0.0622	-1.8650	0.4719	0.1549	0.0614	0.3905	0.4930	1.4085

60858	G	A	CTA_0057	CDS	-	0.0684	0.0698	-1.8134	0.5121	0.1631	0.0598	0.4450	0.3099	1.4085
835039	G	A	CTA_0782	CDS	-	0.0700	0.0606	-1.8764	0.5113	0.1531	0.0562	0.4172	0.3099	0.0000
19005	A	G	NA	inter	NA	0.0710	0.0707	-1.8074	0.5254	0.1641	0.0586	0.4595	0.2958	2.8169
4554	A	G	gatB	CDS	+	0.0713	0.0698	-1.8134	0.5121	0.1631	0.0598	0.4450	0.3099	1.4085
303590	C	A	murE	CDS	-	0.0718	0.0606	-1.8764	0.5113	0.1531	0.0562	0.4172	0.3099	0.0000
215130	C	T	gyrA_1	CDS	-	0.0722	0.0623	-1.8643	0.5114	0.1550	0.0569	0.4223	0.3099	1.4085
806382	C	T	CTA_0761	CDS	+	0.0726	0.0580	-1.8960	0.5297	0.1502	0.0532	0.4241	0.2958	4.2254
778783	G	A	rrf	CDS	-	0.0767	0.0751	-1.7797	0.5021	0.1687	0.0630	0.4514	0.3239	2.8169
136812	G	A	incF	CDS	+	0.0792	0.0751	-1.7797	0.5021	0.1687	0.0630	0.4514	0.3239	2.8169
169573	G	A	CTA_0156	CDS	+	0.0821	0.0765	-1.7712	0.5227	0.1701	0.0611	0.4740	0.3099	9.8592
956953	C	T	pmpD	CDS	+	0.0823	0.0719	-1.7998	0.5226	0.1653	0.0594	0.4605	0.2958	2.8169
44990	A	G	ruvB	CDS	+	0.0871	0.0858	-1.7181	0.4932	0.1794	0.0682	0.4717	0.3380	2.8169
62140	G	T	sucA	CDS	+	0.0914	0.0784	-1.7601	0.5024	0.1720	0.0643	0.4605	0.3239	5.6338
542521	G	A	CTA_0507	CDS	-	0.0916	0.0899	-1.6960	0.4940	0.1834	0.0696	0.4830	0.3380	2.8169
181019	C	A	CTA_0164	CDS	-	0.0953	0.0958	-1.6656	0.4940	0.1891	0.0718	0.4979	0.3380	4.2254
151156	C	G	CTA_0140	CDS	-	0.0955	0.0767	1.7701	0.5019	5.8714	2.1953	15.7035	0.3239	4.2254

- (a) Ocular localization-associated non-synonymous SNPs ($p\text{-value} < 0.1$). Position of the SNPs and name of the impacted are from the *Ct A/HAR13* (GenBank Accession Number NC_007429) genome. 'Allele Percentage' is the percentage of each group where the given allele was present. 'CDS/NCR' identifies whether the SNP was in a coding or non-coding region. 'P*' indicates p-values from 100,024 simulations indicating genome wide significance at $p^* < 0.05$. 'MAF' is the minor allele frequency. 'N Calls at Locus' is the proportion of isolates which had no base called. 'AA' is the amino acid coded for.
- (b) Disease severity-associated SNPs ($p\text{-value} < 0.1$). Disease severity is defined by a composite *in vivo* conjunctival phenotype derived using principal component analysis using ocular *C. trachomatis* load and conjunctival inflammatory (P) score (using the modified FPC (Follicles, Papillary Hypertrophy, Conjunctival Scarring) trachoma grading system³⁹). 'Reference Allele' indicates the reference allele on *Ct A/HAR-13* (GenBank Accession Number NC_007429). 'CDS/NCR' identifies whether the SNP was in a coding or non-coding region. 'P*' = permuted p-value after 100,024 simulations indicating genome wide significance at $p^* < 0.05$. 'T' is the t statistic; SE(T) is the Standard Error of the t statistic. OR is the adjusted Odds Ratio (derived from the t statistic). 95% C.I. = 95% confidence interval of the OR. 'MAF' is the minor allele frequency. 'N Calls at Locus' is the proportion of isolates which had no base called.

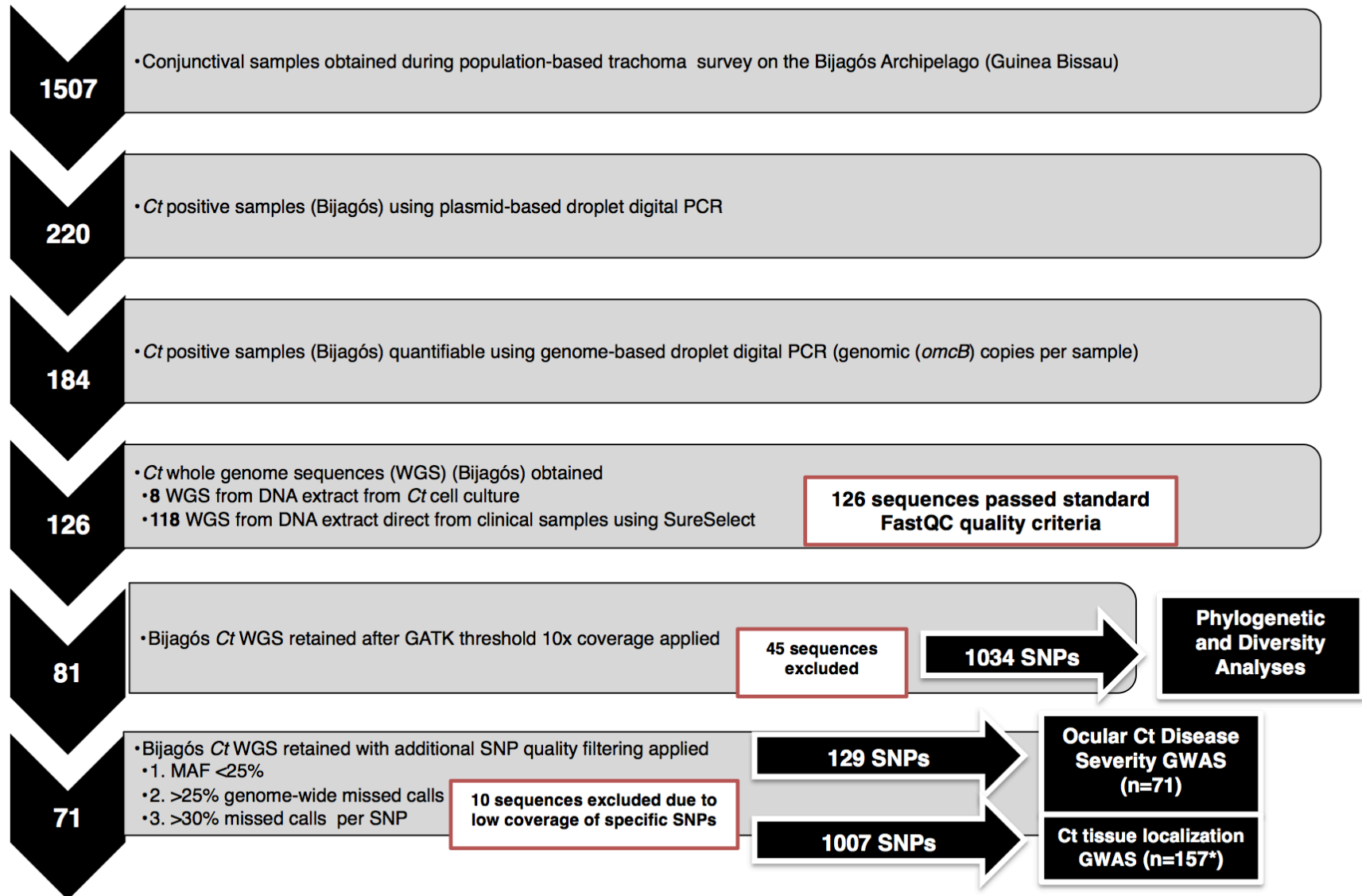
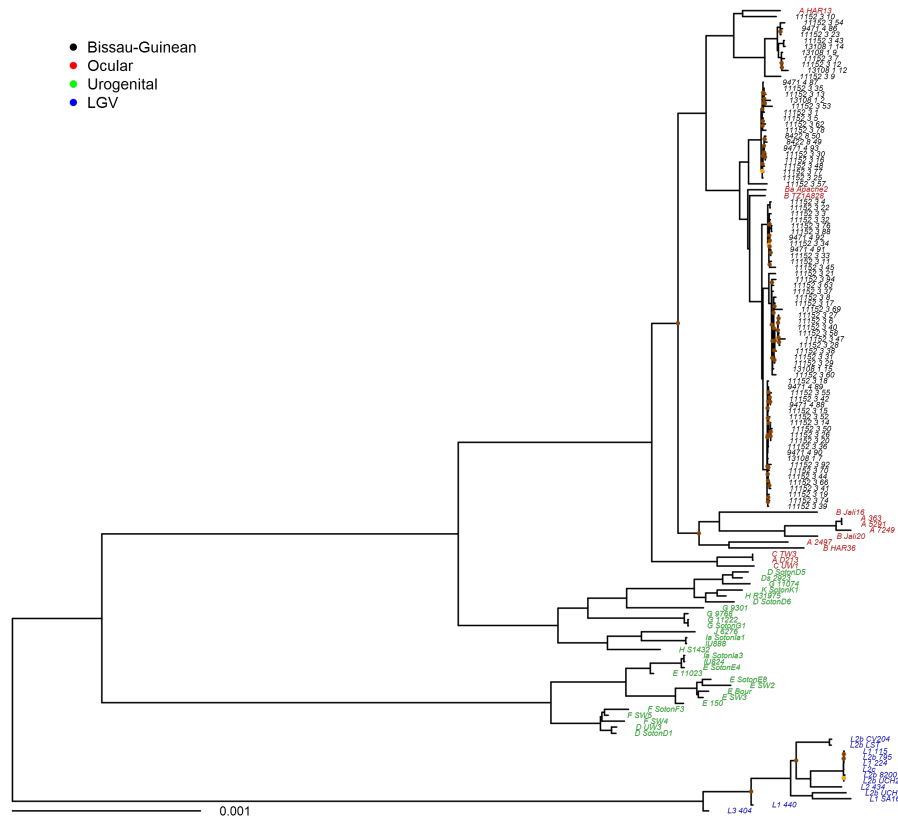


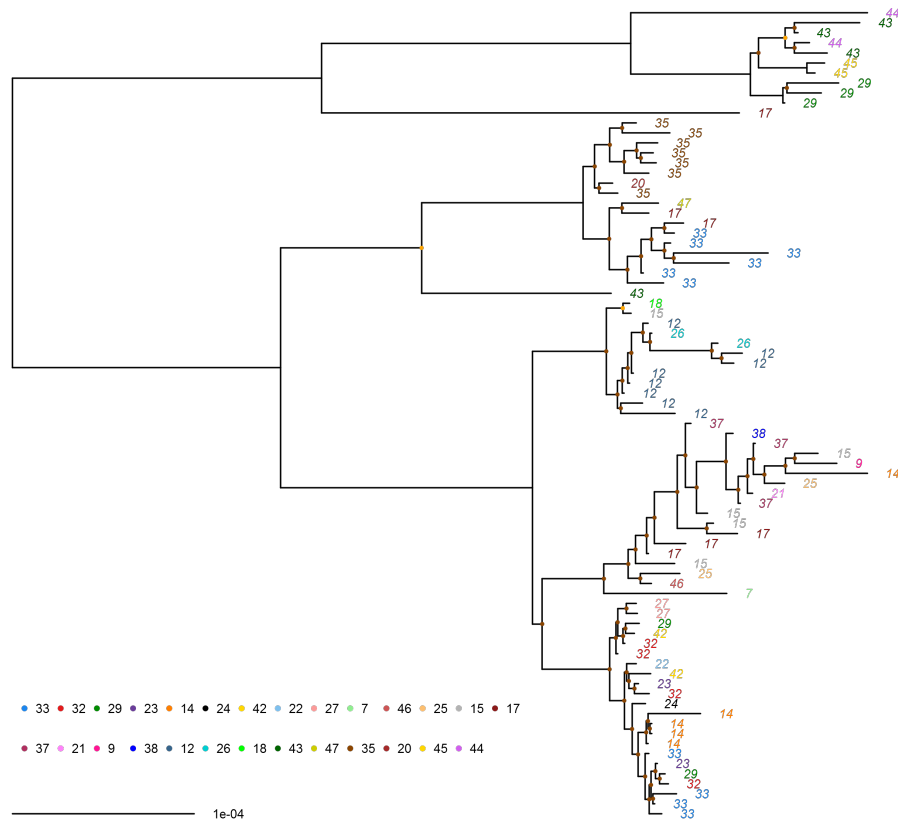
Figure 1. Whole Genome Sequence (WGS) quality filtering processes and threshold criteria for inclusion in analyses. *Ct* DNA detected using droplet digital PCR⁸². WGS data were obtained using SureSelect target enrichment³¹ (or chlamydial cell culture) and Illumina paired end sequencing. FastQC⁹⁰ was used to assess basic WGS quality. SNP alleles were called against reference strain *Ct A/HAR-13* using an alternative coverage-based approach where a missing call was assigned to a site if the total coverage was less than 20x depth or where one of the four nucleotides accounted for at least 80% total coverage⁹⁵. There was a clear relationship between the mean depth of coverage and genome-wide proportion of missing calls, therefore only sequences with greater than 10x mean depth of coverage over the whole genome were retained using the GATK Best Practice threshold^{93,94}. Heterozygous calls were removed and SNPs with a minor allele frequency (MAF) of less than 25% were removed. Samples with greater than 25% genome-wide missing data and 30% missing data per SNP were excluded from the analysis. WGS sequence quality is shown in detail in *Supplementary Information S12*. *n=157 including the 71 Bijagós sequences in addition to 48 Rombo sequences and 38 reference sequences.

Figure 2. Maximum likelihood reconstruction of whole-genome phylogeny of ocular *Chlamydia trachomatis* sequences from the Bijagós Archipelago (Guinea Bissau)



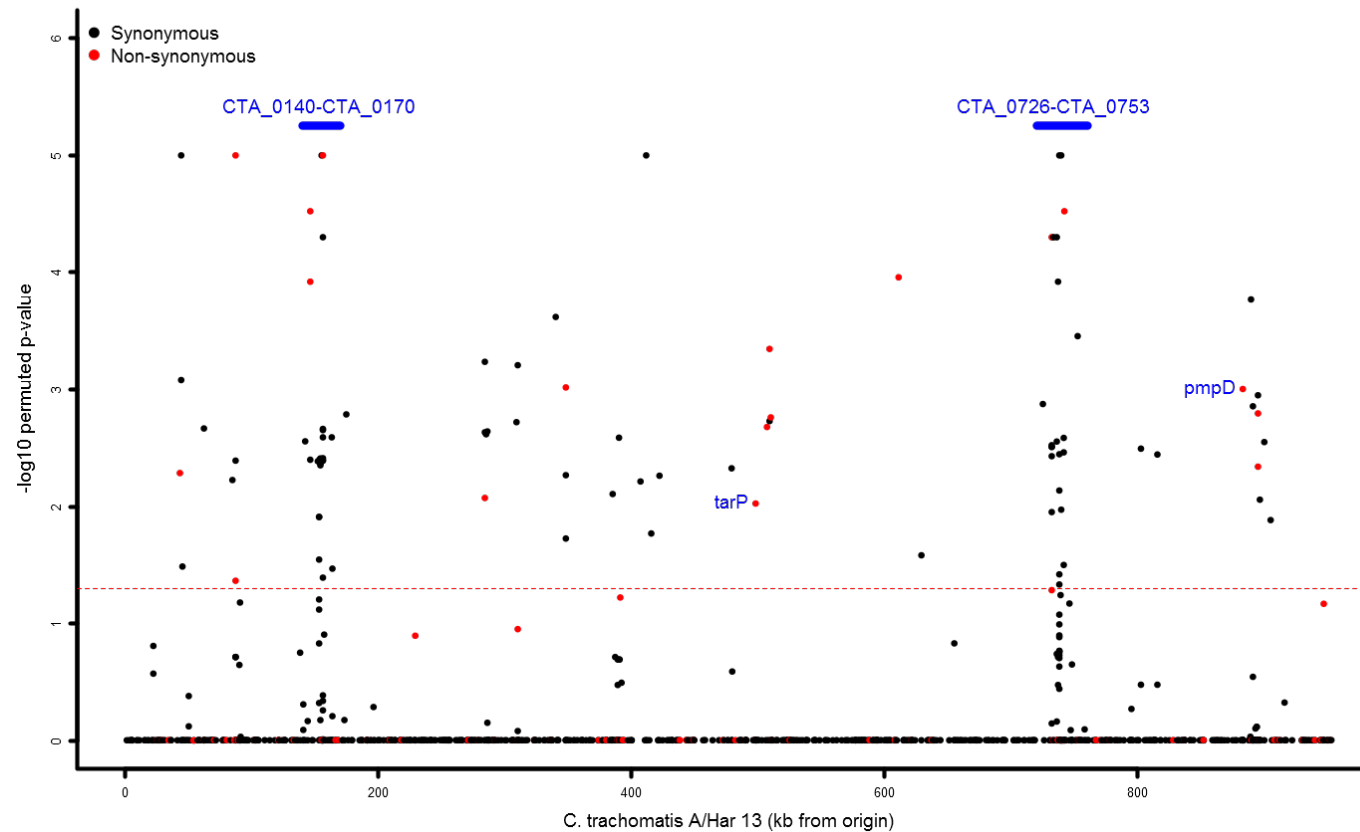
Maximum likelihood reconstruction of the whole-genome phylogeny of 81 *Ct* sequences from the Bijagós Islands and 54 *Ct* reference strains. Bijagós *Ct* sequences (n=81) were mapped to *Ct A/HAR-13* using SAMtools⁹². SNPs were called as described by Harris et al.⁴ Phylogenies were computed with RAXML⁹⁶ from a variable sites alignment using a GTR+gamma model and are midpoint rooted. The scale bar indicates evolutionary distance. Bijagós *Ct* sequences in this study are coloured BLACK and reference strains are coloured by tissue localization (RED=Ocular, GREEN=Urogenital, BLUE=LGV). Branches are supported by > 90% of 1000 bootstrap replicates. Branches supported by 80-90% (ORANGE) and < 80% (BROWN) bootstrap replicates are indicated.

Figure 3. Maximum likelihood phylogenetic tree showing clustering of ocular *Chlamydia trachomatis* sequence types by village

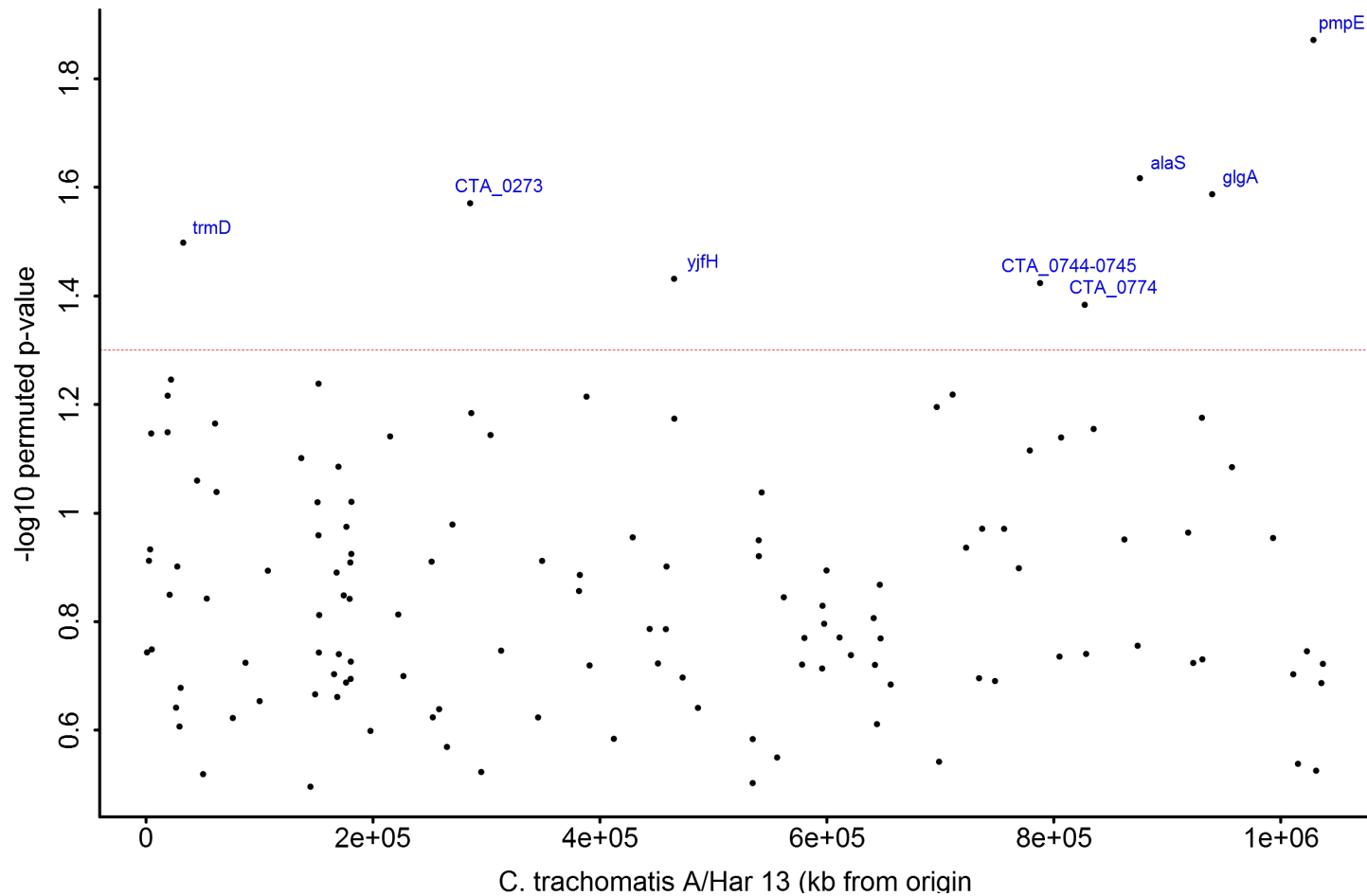


RAxML maximum likelihood phylogenetic reconstruction including all ocular *Ct* sequences retained in the final disease severity association analysis after quality filtering ($n=71$). Ocular *Ct* sequences labelled by village (villages numbered and coloured). Midpoint-rooted and mapped to reference *Ct A/HAR-13*. Branches are supported by > 90% of 1000 bootstrap replicates. Branches supported by 80-90% (ORANGE) and < 80% (BROWN) bootstrap replicates are indicated.

Figure 4. Single Nucleotide Polymorphisms on the *Chlamydia trachomatis* genome associated with (A) ocular localization and (B) disease severity at genome-wide significance



(A) Ocular localization-associated SNPs across the *C. trachomatis* genome. 1007 SNPs were identified in coding and non-coding regions and were included in permutation-based linear regression models in the *Ct* genome-wide association analysis. The threshold for genome-wide significance is indicated by the dashed line ($p^* < 0.05$). The y-axis shows the $-\log_{10}$ p-value. A $-\log_{10}$ p-value of 1.3 is equivalent to a permuted p-value of 0.05 ($p^* < 0.05$). Synonymous (BLACK) and non-synonymous SNPs (RED) are indicated. Regions informative for ocular localization and genes of interest are labelled in BLUE.



(B) Disease severity-associated SNPs across the *Ct* genome. From 129 SNPs identified in coding and non-coding regions, SNPs associated with the disease severity phenotype at genome-wide significance are identified using permutation-based ordinal logistic regression models adjusting for age in the *Ct* genome-wide association analysis. The threshold for genome-wide significance is indicated by the dashed line ($p^* < 0.05$). The y-axis shows the $-\log_{10}$ p-value. A \log_{10} p-value of 1.3 is equivalent to a permuted p-value of 0.05 ($p^* < 0.05$). Genes significantly associated with disease severity are labelled in **BLUE**.

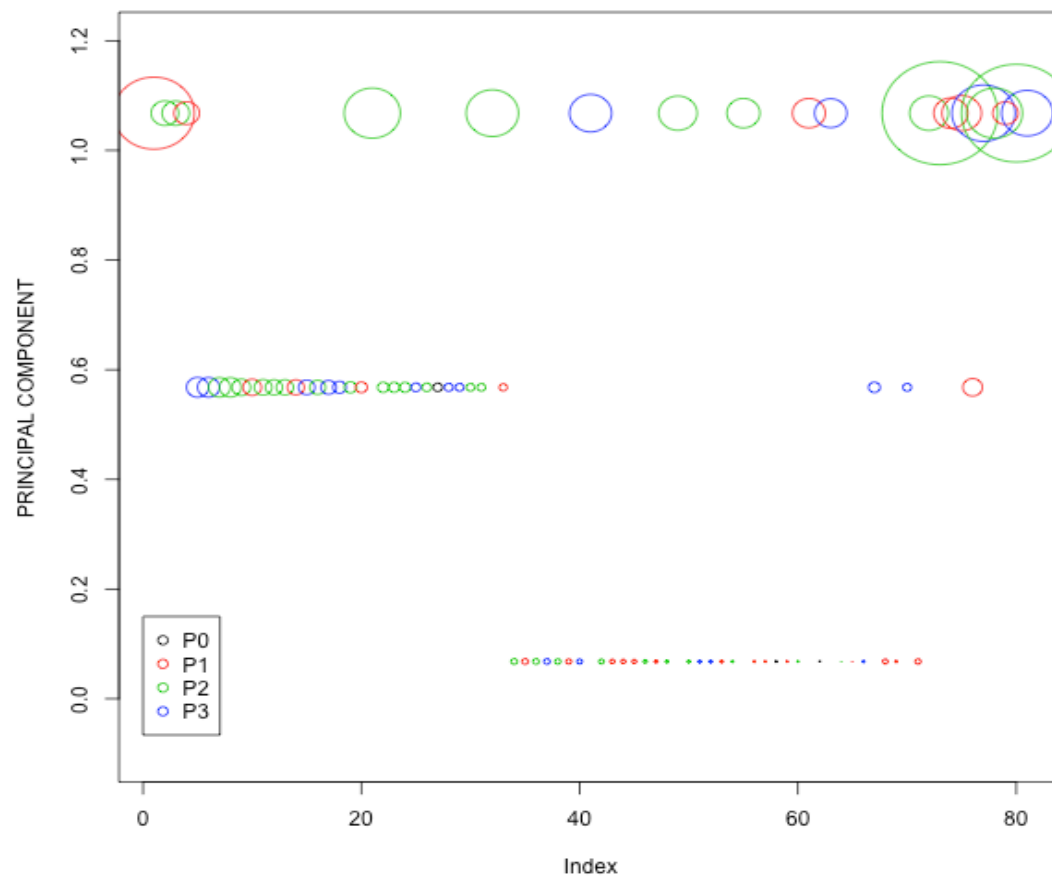


Figure 5.

Composite *in vivo* conjunctival disease severity phenotype in ocular *Chlamydia trachomatis* infection

A composite *in vivo* phenotype was derived using principal component analysis (PCA) for dimension reduction of two phenotypic traits: a disease severity score (using the P score value) and *C. trachomatis* load (where *C. trachomatis* load was log transformed and cut-offs determined from the resulting density plot (See *Supplementary Information S9*)). Each circle represents an individual infection (represented on the x axis (Index), n=81). Circle size reflects *C. trachomatis* load and circle colour reflects inflammatory P score (P0-P3) defined using the modified FPC (Follicles, Papillary Hypertrophy, Conjunctival Scarring) grading system for trachoma³⁹

REFERENCES

1. Hu VH *et al.* Epidemiology and control of trachoma: systematic review. *Trop Med Int Health* **15**(6), 673-91 (2010).
2. World Health Organisation. Sexually Transmitted Infections (STIs) (2016). <http://www.who.int/mediacentre/factsheets/fs110/en/>
3. Rodriguez P *et al.* Typing of *Chlamydia trachomatis* by restriction endonuclease analysis of the amplified major outer membrane protein gene. *J Clin Microbiol* **29**, 1132-1136 (1991).
4. Harris SR *et al.* Whole genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nature Genet* **44**(4), 413-s1 (2012).
5. Caldwell HD *et al.* Polymorphisms in *Chlamydia trachomatis* tryptophan synthase genes differentiates between genital and ocular isolates: Implications in pathogenesis and infection tropism. *J Clin Invest* **111**, 1757-69 (2003).
6. Jeffrey BM *et al.* Genome sequencing of recent clinical *Chlamydia trachomatis* strains identifies loci associated with tissue tropism and regions of apparent recombination. *Infect Immun* **78**, 2544-2553 (2010).
7. Nunes A, Borrego MJ, Gomes JP. Genomic features beyond *C. trachomatis* phenotypes: what do we think we know? *Infect Genet Evol* **16**, 392-400 (2013).
8. Kari L *et al.* Pathogenic diversity among *Chlamydia trachomatis* ocular strains in non-human primates is affected by subtle genomic variations. *J Infect Dis* **197**, 449-56 (2008).
9. Butcher RMR *et al.* Low prevalence of conjunctival infection with *Chlamydia trachomatis* in a treatment-naive trachoma-endemic region of the Solomon Islands. *PLoS Negl Trop Dis* **10**(10), e0005051 (2016).
10. Andersson P *et al.* *Chlamydia trachomatis* from Australian Aboriginal people with trachoma are polyphyletic composed of multiple distinctive lineages. *Nature Commun* **7**, 10688 (2016).
11. Feng L *et al.* Survey, culture and genome analysis of ocular *Chlamydia trachomatis* in Tibetan boarding primary schools in Qinghai Province, China. *Front Cell Infect Microbiol* **6**, 207 (2017).
12. Borges V *et al.* Complete genome sequence of *Chlamydia trachomatis* ocular serovar C strain TW-3. *Genome Announc* **2**, e01204-13 (2014).
13. Darville T and Hiltke T. Pathogenesis of genital tract disease due to *Chlamydia trachomatis*. *J Inf Dis* **201**(Supplement_2), S114-S225 (2010).
14. Haggerty CL *et al.* Risk of sequelae after *Chlamydia trachomatis* genital infection in women. *J Inf Dis* **201**(Supplement_2), S134-S135 (2010).
15. Bailey RL *et al.* Molecular epidemiology of trachoma in a Gambian village. *Br J Ophthalmol* **78**, 813-17 (1994).
16. Andreasen AA *et al.* *Chlamydia trachomatis ompA* variants in trachoma: what do they tell us? *PLoS Negl Trop Dis* **2**(9), e306 (2008).
17. Wang Y *et al.* Development of a transformation system for *Chlamydia trachomatis*: restoration of glycogen biosynthesis by acquisition of a plasmid shuttle vector. *PLoS Pathog* **7**(9), e1002258 (2011).
18. Wang Y *et al.* Transformation of a plasmid-free genital tract isolate with a plasmid vector carrying a deletion in CDS6 revealed that this gene regulates inclusion phenotype. *Pathogens Dis* **67**(2), 100-103 (2013).
19. Longbottom D *et al.* Molecular cloning and characterization of genes coding for the highly immunogenic cluster of 90 kilodalton envelope proteins from

- the *Chlamydia psittaci* subtype that causes abortion in sheep. *Infect Immun* **66**, 1317-1324 (1998).
20. Gomes JP *et al.* Polymorphisms in the nine polymorphic membrane proteins of *Chlamydia trachomatis* across all serovars: evidence for serovar Da recombination and correlation with tissue tropism. *J Bacteriol* **188**, 275-286 (2006).
 21. Rockey DD, Heinzen RA, Hackstadt T. Cloning and characterization of a *Chlamydia psittaci* gene coding for a protein localized in the inclusion membrane of infected cells. *Mol Microbiol* **15**, 617-26 (1995).
 22. Hefty PS, Stephens RS. *Chlamydia trachomatis* type III secretion system is encoded on ten operons preceded by a sigma 70-like promoter element. *J Bacteriol* **189**, 198-206 (2007).
 23. Carlson JH *et al.* *In vivo* and *in vitro* studies of *Chlamydia trachomatis* TrpR:DNA interaction. *Mol Microbiol* **59**(6), 1678-91 (2006).
 24. O'Connell CM *et al.* Toll-like receptor 2 activation by *Chlamydia trachomatis* is plasmid dependent, and plasmid-responsive chromosomal loci are coordinately regulated in response to glucose limitation by *C. trachomatis* but not by *C. muridarum*. *Infect Immun* **79**, 1044-1056 (2011).
 25. Hackstadt T, Scidmore-Carlson MA, Shaw EI, Fischer ER. *Chlamydia trachomatis* IncA protein is required for homotypic vesicle fusion. *Cell Microbiol* **1**, 119-130 (1999).
 26. Nelson DE *et al.* Inhibition of Chlamydiae by primary alcohols correlates with strain specific complement of plasticity zone phospholipase D genes. *Infect Immun* **74**(1), 73-80 (2006).
 27. Carlson JH, Hughes S, Hogan D. Polymorphisms in the *Chlamydia trachomatis* cytotoxin locus associated with ocular and genital isolates. *Infect Immun* **72**(12), 7063-72 (2004).
 28. Carlson JH *et al.* The *Chlamydia trachomatis* plasmid is a transcriptional regulator of chromosomal genes and a virulence factor. *Infect Immun* **76**, 2273-83 (2008).
 29. Frazer LC *et al.* Plasmid-cured *Chlamydia caviae* activates TLR2-dependent signaling and retains virulence in the guinea pig model of genital tract infection. *PLoS One* **7**(1), e30747 (2012).
 30. Song L *et al.* *Chlamydia trachomatis* plasmid-encoded *pgp4* is a transcriptional regulator of virulence-associated genes. *Infect Immun* **81**(3), 636 (2013).
 31. Christiansen MT *et al.* Whole genome enrichment and sequencing of *Chlamydia trachomatis* directly from clinical samples. *BMC Infect Dis* **14**, 591 (2014).
 32. Borges V *et al.* Effect of long-term laboratory propagation on *Chlamydia trachomatis* genome dynamics. *Infect Genet Evol* **17**, 23-32 (2013).
 33. Borges V *et al.* *Chlamydia trachomatis* *in vivo* to *in vitro* transition reveals mechanisms of phase variation and downregulation of virulence factors. *PLoS One* **10**(7), e0133420 (2015).
 34. Bonner C *et al.* *Chlamydia trachomatis* virulence factor CT135 is stable *in vivo* but highly polymorphic *in vitro*. *Pathog Dis* **73**(6), ftv043 (2015).
 35. Burton MJ *et al.* Which members of a community need antibiotics to control trachoma? Conjunctival *Chlamydia trachomatis* load in Gambian villages. *Invest Ophthalm Vis Sci* **44**(10), 4215-22 (2003).

36. Conway DJ *et al.* Scarring trachoma is associated with polymorphisms in TNF-alpha gene promoter and with increased TNF-alpha levels in tear fluid. *Infect Immun* **65**(3), 1003-6 (1997).
37. West SK *et al.* Progression of active trachoma to scarring in a cohort of Tanzanian children. *Ophthalmic Epidemiol* **8**(2-3), 137-44 (2001).
38. Burton MJ *et al.* Pathogenesis of progressive scarring trachoma in Ethiopia and Tanzania: Two cohort studies. *PLoS Negl Trop Dis* **9**(5), e0003763 (2015).
39. Dawson CR, Jones BR, Tarizzo ML. Guide to trachoma control in programs for the prevention of blindness; Geneva: World Health Organization (1981).
40. Reid JS *et al.* A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape. *Nature Comms* **7**, 13357 (2016).
41. Yang J *et al.* Conditional and joint multiple SNP analysis of GWAS summary statistics identifies additional variance influencing complex traits. *Nat Genet* **44**, S1-3 (2012).
42. Aschard H *et al.* Maximising the power of principal component analysis of correlated phenotypes in genome wide association studies. *Am J Hum Genet* **94**;662-676 (2014).
43. Cheng AY, Teo YY, Ong RT. Assessing single nucleotide variant detection and genotype calling on whole genome sequenced individuals. *Bioinformatics* **30**(12), 1707-13 (2014).
44. Gudbjartsson DF *et al.* Large-scale whole genome sequencing of the Icelandic population. *Nat Genet* **47**(5), 435-44 (2015).
45. Last AR *et al.* Risk Factors for Active Trachoma and Ocular *Chlamydia trachomatis* Infection in Treatment-naïve Trachoma-hyperendemic Communities of the Bijagós Archipelago, Guinea Bissau. *PLoS Negl Trop Dis* **8**(6), e2900 (2014).
46. Pickering H *et al.* Genome-wide profiling of humoral immunity and pathogen genes under selection identifies immune evasion tactics of *Chlamydia trachomatis* during ocular infection. *Sci Rep* (2017) (in press).
47. Lutter EI *et al.* Phylogenetic analysis of *Chlamydia trachomatis* TARP and correlation with clinical phenotype. *Infect Immun* **78**(9), 3678-88 (2010).
48. Che R *et al.* An adaptive permutation approach for genome-wide association study. *BioData Mining* **7**, 9 (2014).
49. Dudoit S *et al.* Multiple hypothesis testing in microarray experiments. *Statist Sci* **18**(1), 71-103 (2003).
50. Sham PC and Purcell SM. Statistical power and significance testing in large scale genetic studies. *Nat Rev Genet* **15**(5), 335-46 (2014).
51. Fairfax BP *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* **44**(5), 502-510 (2012).
52. Hadfield J *et al.* Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. *Genome Res* **27**, 1-10 (2017).
53. Solomon AW *et al.* Strategies for control of trachoma: observational study with quantitative PCR. *Lancet* **362**(9379), 198-204 (2003).
54. Pepperell CS *et al.* The role of selection in shaping diversity of natural *Mycobacterium tuberculosis* populations. *PLoS Pathog* **9**(8), 10.1371 (2013).

55. Fey PD *et al.* Assessment of whole genome mapping in a well-defined outbreak of *Salmonella enterica* serotype Saintpaul. *J Clin Microbiol* **50**(9), 3063-5 (2012).
56. Gilchrist CA, Turner SD, Riley MF, Petri WA Jr, Hewlett EL. Whole genome sequencing in outbreak analysis. *Clin Microbiol Rev* **28**(3), 541-63 (2015).
57. Gomes JP *et al.* Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots. *Genome Res* **17**(1), 50-60 (2007).
58. Sheppard SK *et al.* Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad Sci U S A* **110**, 11923–11927 (2013).
59. Laabei M *et al.* Predicting the virulence of MRSA from its genome sequence. *Genome Res* **24**(5), 839-49 (2014).
60. Alam MT *et al.* Dissecting vancomycin intermediate resistance in *Staphylococcus aureus* using genome-wide association. *Genome Biol Evol* **6**, 1174–1185 (2014).
61. Chewapreecha C *et al.* Comprehensive Identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet* **10**, e1004547 (2014).
62. Maury M *et al.* Uncovering *Listeria monocytogenes* hypervirulence by harnessing its biodiversity. *Nature Genet* **48**, 308-313 (2016).
63. Earle SG *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiol* **1**, 16041 (2016).
64. Borges V, Gomes JP. Deep comparative genomics among *Chlamydia trachomatis* lymphogranuloma venereum isolates highlights genes potentially involved in pathoadaptation. *Infect Genet Evol* **32**, 74-88 (2015).
65. Ferreira R *et al.* *In silico* scrutiny of genes revealing phylogenetic congruence with clinical prevalence or tropism properties of *Chlamydia trachomatis* strains. *G3* **5**(1), 9-19 (2014).
66. Carlson JH, Porcella SF, McClarty G. Comparative genomic analysis of *Chlamydia trachomatis* oculotropic and genitotropic strains. *Infect Immun* **73**(10), 6407-18 (2005).
67. Brunelle BW, Nicholson TL, Stephens RS. Micro-array based genomic surveying of gene polymorphisms in *Chlamydia trachomatis*. *Genome Biol* **5**(6), R42 (2004).
68. Borges V, Nunes A, Ferreira R, Borrego MJ, Gomes JP. Directional evolution of *Chlamydia trachomatis* towards niche-specific adaptation. *J Bacteriol* **194**(22), 6143-53 (2012).
69. Belland RJ *et al.* Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. *Proc Natl Acad Sci USA* **100**(14), 8478-83 (2003).
70. Clifton DR *et al.* A Chlamydial type III translocated protein is tyrosine-phosphorylated at the site of entry and associated with the recruitment of actin. *Proc Natl Acad Sci USA* **101**(27), 10166-71 (2004).
71. Kari L *et al.* *Chlamydia trachomatis* polymorphic membrane protein D is a virulence factor involved in early host-cell interactions. *Infect Immun* **82**(7), 2756-62 (2014).

72. Da Cunha M *et al.* Identification of type THREE secretion substrates of *Chlamydia trachomatis* using *Yersinia enterocolitica* as a heterologous system. *BMC Microbiol* **17**(14), 40 (2014).
73. Becker E and Hegemann JH. All subtypes of the Pmp adhesion family are implicated in chlamydial virulence and show species-specific function. *Microbiologyopen* **3**(4), 544-56 (2014).
74. Gangaiah D *et al.* Carbon storage regulator A contributes to the virulence of *Haemophilus ducreyi* in humans by multiple mechanisms. *Infect Immun* **81**(2), 608-17 (2013).
75. Hogg JS *et al.* Characterisation and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical non-typeable strains. *Genome Biol* **8**, R103.1-R103.8 (2007).
76. Samuelson JC *et al.* YidC mediates membrane protein insertion in bacteria. *Nature* **406**(6796), 637-41 (2000).
77. Palmer SR *et al.* YidC1 and YidC2 are functionally distinct proteins involved in protein secretion, biofilm formation and cariogenicity of *Streptococcus mutans*. *Microbiology* **158**(7), 1702-12 (2012).
78. Michel G *et al.* The structure of the RlmB 23S rRNA methyltransferase reveals a new methylation fold with a unique knot. *Structure* **10**(10), 1303-15 (2002).
79. Lovgren JM, Wikstrom PM. The *rlmB* gene is essential for formation of Gm2251 in 23S rRNA but not for ribosome maturation in *Escherichia coli*. *J Bacteriol* **183**(23), 6957-60 (2001).
80. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: Lessons from human GWAS. *Nat Rev Genet* **18**(1), 41-50 (2017).
81. Last A *et al.* Plasmid copy number and disease severity in naturally occurring ocular *Chlamydia trachomatis* infection. *J Clin Microbiol* **52**(1), 324-7 (2014).
82. Roberts Ch *et al.* Development and evaluation of a next generation digital PCR diagnostic assay for ocular *Chlamydia trachomatis* infections. *J Clin Microbiol* **51**(7), 2195-2203 (2013).
83. Burton MJ *et al.* Conjunctival Chlamydial 16S Ribosomal RNA Expression in Trachoma: Is Chlamydial Metabolic Activity Required for Disease to Develop? *Clin Infect Dis* **42**, 463-70 (2006).
84. Faal N *et al.* Conjunctival FOXP3 expression in trachoma: Do regulatory T cells have a role in human ocular *Chlamydia trachomatis* infection? *PLoS Med* **3**(8), e266 (2006).
85. Derrick T *et al.* Inverse relationship between micro RNA-155 and -184 expression with increasing conjunctival inflammation during ocular *Chlamydia trachomatis* infection. *BMC Infect Dis* **16**, 60 (2016).
86. Last A *et al.* Spatial clustering of high load ocular *Chlamydia trachomatis* infection in trachoma: a cross-sectional population-based study. *Pathog Dis* **75**(5), (2017).
87. Suo C *et al.* Analysis of multiple phenotypes in genome wide genetic mapping studies. *BMC Bioinformatics* **14**, 151 (2013).
88. Skipp P, Robinson J, O'Connor CD, Clarke IN. Shotgun proteomic analysis of *Chlamydia trachomatis*. *Proteomics* **5**(6), 1558-73 (2005).
89. Seth-Smith HMB *et al.* Co-evolution of genomes and plasmids within *Chlamydia trachomatis* and the emergence in Sweden of a new variant strain. *BMC Genomics* **10**, 239 (2009).

90. Andrews S (2010). FastQC: A quality control tool for high throughput sequence data (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>).
91. Langmead B *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.3 (2009).
92. Li H *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
93. McKenna A *et al.* The Genome Analysis Tool Kit: A MapReduce framework for analysing next generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
94. Cheng AY, Teo YY, Ong RT. Assessing single nucleotide variant detection and genotype calling on whole genome sequenced individuals. *Bioinformatics* **30**(12), 1707-13 (2014).
95. Coll F *et al.* PolyTB: A genomic variation map for *Mycobacterium tuberculosis*. *Tuberculosis* (Edinb.) **94**(3), 346-54 (2014).
96. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* **57**, 758-771 (2008).
97. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665-2681 (2006).
98. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**(2), 221-114 (2010).
99. Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**(2), 224-37 (2000).
100. Sonnhammer EL, Eddy SR, Durbin R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28**(3), 405-20 (1997).
101. Nei M, Masatoshi N, Wen-Hsiung L. Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases. *Proc Natl Acad Sci* **76** (10), 5269-73 (1979).
102. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacterial by support vector machines based on n-peptide compositions. *Protein Science* **13**(5), 1402-6 (2004).
103. Yu NY *et al.* PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**(13), 1608-15 (2010).
104. Goldberg T *et al.* LocTree3 prediction of localization. *Nucleic Acids Res* **42**(Web Server Issue), W350-5 (2014).
105. Thomas A *et al.* Effect of linkage disequilibrium on the identification of functional variants. *Genet Epidemiol* **35**(Suppl 1), S115-9 (2011).
106. Misaghi S *et al.* *Chlamydia trachomatis*-derived deubiquinating enzymes in mammalian cells during infection. *Mol Microbiol* **61**(1), 142-50 (2006).
107. Le Negrate G *et al.* ChaDub1 of *Chlamydia trachomatis* suppresses NF-KappaB activation and inhibits 1-KappaB-alpha ubiquitination and degradation. *Cell Microbiol* **10**(9), 1879-92 (2008).
108. Li Z *et al.* Induction of protective immunity against *Chlamydia muridarum* intravaginal infection with a chlamydial glycogen phosphorylase. *PLoS ONE* **7**(3), e32997 (2012).

109. Lu C *et al.* *Chlamydia trachomatis* GlgA is secreted into host cell cytoplasm. *PLoS One* **8**(7), e68764 (2013).
110. Swanson KA *et al.* *Chlamydia trachomatis* polymorphic membrane protein D is an oligomeric autotransporter with a higher-order structure. *Infect Immun* **77**, 508-516 (2009).
111. Werhl W *et al.* From the inside out – processing of the chlamydial autotransporter Pmp D and its role in bacterial adhesion and activation of human host cells. *Mol Microbiol* **51**(2), 319-34 (2004).
112. Nunes A, Gomes JP, Karunakaran KP, Brunham RC. Bioinformatic analysis of *Chlamydia trachomatis* PmpE, PmpF, PmpG and PmpH as potential vaccine antigens. *PLoS ONE* **10**(7), e0131695 (2015).
113. Tan C *et al.* The polymorphic membrane protein gene family of the chlamydiae. *Chlamydia Genomics and Pathogenesis*. Ed: Bavoil PM, Wyrick PB. Horizon Bioscience; Norfolk, UK. pp 195-218 (2006).
114. Liu X *et al.* Identification of *Chlamydia trachomatis* outer membrane complex proteins by differential proteomics. *J Bacteriol* **192**(11), 2852-2860.