nature genetics

Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing

Simon R Harris[1], Ian N Clarke[2], Helena M B Seth-Smith[1], Anthony W Solomon[3], Lesley T Cutcliffe[2], Peter Marsh[4], Rachel J Skilton[2], Martin J Holland[3], David Mabey[3], Rosanna W Peeling[3], David A Lewis[3,5,6], Brian G Spratt[7], Magnus Unemo[8], Kenneth Persson[9], Carina Bjartling[10], Robert Brunham[11], Henry J C de Vries[12–14], Servaas A Morré[15,16], Arjen Speksnijder[17], Cécile M Bébéar[18,19], Maïté Clerc[18,19], Bertille de Barbeyrac[18,19], Julian Parkhill[1] & Nicholas R Thomson[1]

***Chlamydia trachomatis* is responsible for both trachoma and sexually transmitted infections, causing substantial morbidity and economic cost globally. Despite this, our knowledge of its population and evolutionary genetics is limited. Here we present a detailed phylogeny based on whole-genome sequencing of representative strains of *C. trachomatis* from both trachoma and lymphogranuloma venereum (LGV) biovars from temporally and geographically diverse sources. Our analysis shows that predicting phylogenetic structure using *ompA*, which is traditionally used to classify *Chlamydia,* is misleading because extensive recombination in this region masks any true relationships present. We show that in many instances, *ompA* is a chimera that can be exchanged in part or as a whole both within and between biovars. We also provide evidence for exchange of, and recombination within, the cryptic plasmid, which is another key diagnostic target. We used our phylogenetic framework to show how genetic exchange has manifested itself in ocular, urogenital and LGV *C. trachomatis* strains, including the epidemic LGV serotype L2b.**

C. trachomatis is the most prevalent bacterial sexually transmitted infection worldwide, with an estimated 101.5 million new cases occurring among adults in 2005 (ref. 1). Additionally, ocular *C. trachomatis* is the leading infectious cause of blindness, with >40 million people estimated to be suffering from active disease[2].

C. trachomatis comprises two biovars: the trachoma biovar includes ocular and urogenital strains that are characterized by localized infections of the epithelial surface of the conjunctiva or genital mucosa; strains of the LGV biovar are distinguished by their ability to spread systemically thorough the lymphatic system, causing genital ulceration and bubonic disease[3]. LGV is reported most frequently in Africa, Southeast Asia, South America and the Caribbean and is rare in developed countries[4–6]. However, an epidemic of LGV with atypical presentation and symptoms is currently in progress in Europe and North America, primarily affecting men who have sex with men.

Despite the importance of *C. trachomatis* as a human pathogen, very little is known about the evolution of the strains that cause disease[7]. This is primarily because modern diagnosis is generally based on commercial nucleic acid amplification tests[8,9] rather than culture, where strains would be available for further study. Most of our understanding of the diversity of circulating *C. trachomatis* is based on the primary surface antigen, the major outer membrane protein (MOMP), and the gene encoding MOMP, *ompA*. Typing has traditionally been performed serologically using a number of antibodies

[1]Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]Molecular Microbiology Group, University Medical School, Southampton General Hospital, Southampton, UK. [3]Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [4]Health Protection Agency, Public Health Laboratory Southampton, Southampton General Hospital, Southampton, UK. [5]Sexually Transmitted Infections Reference Centre, National Institute for Communicable Diseases, National Health Laboratory Service, Johannesburg, South Africa. [6]Department of Internal Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [7]Department of Infectious Disease Epidemiology, Imperial College, St. Mary's Hospital Campus, London, UK. [8]Department of Laboratory Medicine and Clinical Microbiology, National Reference Laboratory for Pathogenic Neisseria, Örebro University Hospital, Örebro, Sweden. [9]Department of Laboratory Medicine, Clinical Microbiology, Malmö University Hospital, Malmö, Sweden. [10]Department of Obstetrics and Gynecology, Institute of Clinical Sciences, Malmö University Hospital, Malmö, Sweden. [11]British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada. [12]Department of Dermatology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands. [13]Sexually Transmitted Infections Outpatient Clinic, Infectious Diseases Cluster, Public Health Service Amsterdam, Amsterdam, The Netherlands. [14]Centre for Infection and Immunity Amsterdam, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands. [15]Department of Medical Microbiology of Infection Prevention, Laboratory of Immunogenetics, VU University Medical Center, Amsterdam, The Netherlands. [16]Department of Genetics and Cell Biology, Institute of Public Health Genomics, School for Public Health and Primary Care and School for Oncology & Developmental Biology, Faculty of Health, Medicine & Life Sciences, University of Maastricht, Maastricht, The Netherlands. [17]Geneeskundige en Gezondheidsdienst Amsterdam (GGD; Health Service Amsterdam), Amsterdam, The Netherlands. [18]Université de Bordeaux, Unité Sous Contrat (USC) Mycoplasmal and Chlamydial Infections in Humans, French National Reference Center for Chlamydial Infections, Bordeaux, France. [19]Institut National de la Recherche Agronomique, USC Mycoplasmal and Chlamydial Infections in Humans, French National Reference Center for Chlamydial Infections, Bordeaux, France. Correspondence should be addressed to S.R.H. (sh16@sanger.ac.uk).

Received 5 August 2011; accepted 6 February 2012; published online 11 March 2012; doi:10.1038/ng.2214

against divergent epitopes within the MOMP, but researchers have more recently switched to a genotyping approach that uses the *ompA* sequence. Based on these methods, the two *C. trachomatis* biovars have been subdivided into 15–19 serotypes: the trachoma biovar includes ocular serotypes A–C and the urogenital serotypes D–K, and the LGV biovar includes serotypes L1, L2, L3 and L2b, the last of which is the serotype associated with the current LGV outbreak in Europe and North America. Although *ompA* genotyping provides some further differentiation within these serotypes, it provides little or no detailed information about the nature of the infecting strain or the variation in the remaining 99.88% of the genome.

Attempts have been made to reconstruct the *C. trachomatis* species phylogeny using *ompA* sequences[10], 16S ribosomal RNA (rRNA) gene analysis[11], multilocus sequence typing (MLST) approaches[12–15] and whole-genome sequencing[16–18]. However, there is still no consensus regarding the true evolutionary relationships between the *C. trachomatis* strains. It is generally accepted that *ompA* does not reflect the phylogeny of the species[14,19,20], but the lack of agreement between the phylogenic trees produced from other small gene sets suggests that many other regions of the *C. trachomatis* genome also provide conflicting phylogenetic evidence[19].

Historically, horizontal gene transfer in the chlamydiae was considered unlikely because of their obligate intracellular niche and because co-infections with more than one strain, and therefore opportunities for recombination, are rare. Together, these factors were thought to represent a barrier for effective recombination. However, it has recently become evident that not only do the chlamydiae have all the necessary recombination machinery[21], they can recombine after mixed infection both in tissue culture and the human host[17]. It is probable that this recombination is the cause of the difficulty in resolving the relationships between *C. trachomatis* strains.

Recent attempts to quantify the impact of recombination by reanalyzing published genomes have been limited as there are only a small number of these genomes available[18]. We set out to sample widely from the diversity of clinical strains, and we here present 36 new *C. trachomatis* whole-genome sequences. Using these sequences and those previously published, we provide a detailed reconstruction of the evolutionary history of *C. trachomatis* that identifies and takes into account the effect of recombination. We show conclusively that the epidemic outbreak of L2b strains in Europe was the result of clonal expansion and transmission and provide evidence of recombination in natural clinical strains both within and between biovars. We show the effect that recombination has had on our general understanding of *C. trachomatis* diversity as well as the implications of recombination in the monitoring and epidemiological tracking of infections based on current typing techniques, and we provide evidence of recent serotype switches within circulating clinical strains.

RESULTS

We sequenced 36 *C. trachomatis* genomes from global collections, isolated between 1959 and 2009, comprising 18 LGV, 14 urogenital and 4 ocular strains. We included 12 L2b strains, which originated from the UK, France, Sweden, The Netherlands and Canada, and a further 6 LGV strains from South Africa and the United States, including the historical L1 and L3 strains. The newly sequenced ocular strains were collected in Tanzania in 2000. The urogenital strains were collected in the UK (ten strains), Sweden (three strains) and the United States (one strain). To guard against potential cross contamination of strains and to provide an independently verifiable set of reference genomes, a key element of our experimental design was to use live strains that were cultured individually. All strains are currently held as live stocks, with the exception of three ocular strains that were destroyed after culture. Our analysis also includes 16 previously published genome sequences: 2 LGV (from the UK and United States), 3 ocular (from Egypt, Tanzania and Gambia) and 11 urogenital strains (from the United States and Sweden where information was available), totaling 52 strains that we included in this study (**Supplementary Table 1**).

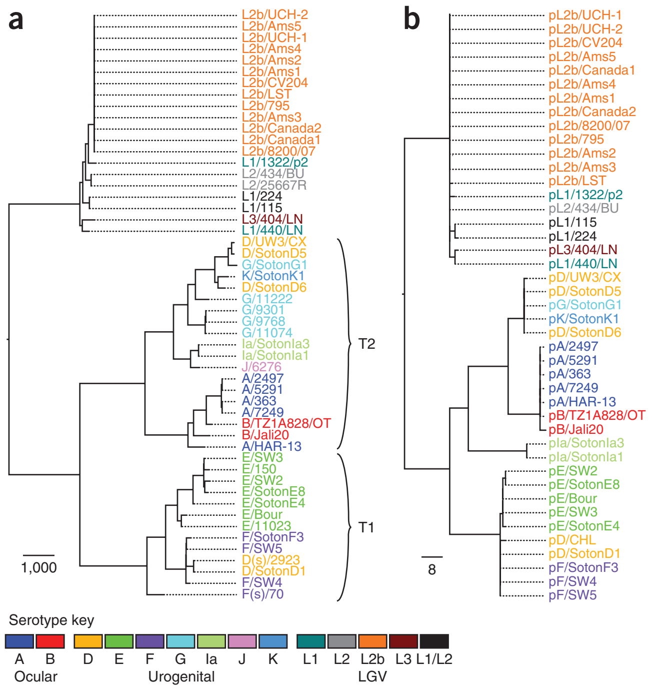

Figure 1 Maximum likelihood reconstruction of the phylogeny of *C. trachomatis* with recombinations removed. (**a**) *C. trachomatis* species phylogeny using the chromosomal sequences of 52 genomes after predicted recombinations have been removed using a previously described method[22]. Bootstrap support for nodes on the tree are shown in **Supplementary Figure 1**. (**b**) Phylogenetic reconstruction of the *C. trachomatis* plasmid after the predicted recombinations have been removed. Strain names are colored by serotype. The scale bar indicates the number of SNPs. Plasmid sequences were not available for all the strains shown in **a**. For comparison, trees without the recombination removal are shown in **Supplementary Figure 2**.

Whole-genome phylogeny of *C. trachomatis*

To establish whether the current understanding of *C. trachomatis* phylogeny is evolutionarily robust, we determined the interrelationships in our strain collection using genome-wide SNPs. To mitigate the effect of homologous recombination between *C. trachomatis* strains on our phylogenetic reconstruction, we used a previously described method to construct our tree[22]. The resulting tree (**Fig. 1a**, see **Fig. 1b** for the plasmid phylogeny, **Supplementary Fig. 1** for bootstrap support values and **Supplementary Fig. 2** for the trees without recombination removed) provides compelling evidence that *ompA* serotyping (**Supplementary Fig. 3a**) does not reflect the evolutionary structure of *C. trachomatis*, with the trachoma serotypes A, B, D, F and G and the LGV serotype L1 all occurring in multiple distinct branches on the tree. This clearly shows that exchange of the whole or part of the *ompA* gene is a natural phenomenon in distinct lineages of *C. trachomatis*. A comparison of the whole-genome phylogeny with phylogenies reconstructed using some of the available *C. trachomatis* MLST schemes (**Supplementary Fig. 3b–d**) shows that multilocus techniques that are based on housekeeping loci show greater congruency with our tree than the *ompA* phylogeny but that these techniques lack resolution.

The whole-genome tree confirms that the species is split into two distinct clades, representing the trachoma and LGV biovars, that are separated by 4,860 SNPs. Using the *Chlamydia muridarum* strain Nigg (previously MoPn) as an outgroup, the root of the tree is located on the branch between the two biovars (**Supplementary Fig. 1**), suggesting that the split between the clades occurred early in the evolutionary history of the species, which is consistent with the conclusions of previous analyses based on 16S rRNA gene sequences[23].

The trachoma clade comprises two lineages (**Fig. 1a**), T1 and T2, that are separated from their common ancestor by 2,374 and 2,228 SNPs, respectively. T1 is composed of clinically prevalent urogenital serotypes, whereas T2 contains most of the rarer urogenital serotypes[24]. All ocular strains form a cluster within T2, indicating that they emerged from a urogenital ancestor (**Fig. 1a**).

Strains within the LGV clade are considerably less diverse than those in the trachoma lineage, as illustrated by the shorter branch lengths within the LGV clade in **Figure 1a**. The 13 strains from the recent L2b outbreak form a tight cluster with maximal bootstrap support (**Supplementary Fig. 1**). The maximum pairwise evolutionary distance between the most variant strains within the L2b serotype is just 19 SNPs. This low level of variation between the strains despite their global distribution shows conclusively that the L2b epidemic[25] is a clonal outbreak that has spread throughout the world.

Our LGV collection also includes two South African strains (L1/115 and L1/224) that could not originally be classified using micro-immunofluorescence with monoclonal antibodies specific to the known MOMP types[26]. Analysis of the *ompA* sequence from these strains showed that the 5′ end (variable segments VS1 and VS2) matches the L1 *ompA* sequence, whereas the 3′ end (variable segments VS3 and VS4) matches the L2 sequence[26]. In our phylogeny, these strains form a clade midway between an L1+L3 clade and a group that includes the L2 and L2b strains. A third South African strain (L1/1322/p2), despite being genotyped as L1, is located on its own branch that is distinct from the other L1 strain used in this study (L1/440/LN). It is evident that these three strains collected in South Africa define two new LGV lineages that are as equally distinct as the accepted LGV serotype clusters. This suggests that LGV diversity is far greater than previously recognized and that additional sampling is necessary.

Recombination is a natural and common occurrence

Considering the traditional view that *Chlamydia* do not recombine, we would expect that, given the small amount of variation identified relative to the size of the genome, very few identical SNPs would have occurred multiple times independently on different branches of the *Chlamydia* tree (homoplasies). However, superimposing the SNP events on the rooted phylogenic tree using PAML[27] showed that homoplasy is common. Of the 17,163 sites in the genome that have a SNP in at least one strain (variant sites), 4,492 (26%) are homoplasic.

Homoplasic SNPs can arise by chance, as a result of a selective pressure for a particular SNP to become fixed in a population or by homologous recombination between divergent strains in which sets of SNPs are effectively imported in one block. In cases where recombination is the cause of homoplasy, it is expected that homoplasic SNPs shared by two distant strains would be clustered along the genome sequence within the regions that have been recombined. Therefore, dense clusters of compatible homoplasies are often used as markers of recombination events. A pair of sites are incompatible if no tree can be drawn in which both sites could be reconstructed without at least one being homoplasic[28]. We applied three compatibility-based recombination detection methods implemented in the PhiPack package[29] to these data: maximum χ^2 (ref. 30), neighbor similarity score[31] and the pairwise homoplasy index[29]. All three methods reported significant P values ($P < 0.05$), indicating that compatibility was significantly higher between closely linked sites, as would be expected if recombination had occurred but is not consistent with a random accumulation of homoplasies or with convergent selection.

To investigate the recombination history, for each node on the tree, we reconstructed the ancestral DNA sequence using PAML[27] and used a scanning statistic[22] to identify regions of the genome in which the node sequence was significantly more similar to a distant branch on the tree than to its direct ancestral node (Online Methods). **Figure 2** shows a reconstruction of recombination events across the *C. trachomatis* tree, where the location of the recombination in the genome of the recipient branch is shown, and the colors indicate the phylogenetic position of the most likely donor. In total, we constructed 267 putative recombination blocks ranging in length from 3–50,141 bp (with a mean length of 4,039 bp) covering 539,409 bp (51%) of the aligned genome length. This provides strong evidence that recombination has occurred many times during the evolution of the species and has not only happened at a few limited hotspots, as has been reported in previous papers[18,32], but, rather, has occurred across a large proportion of the genome.

The SNP density along the length of the genome was not constant, however. Some regions did show increased SNP and homoplasy density, which may have been the result of diversifying selection within the species or homologous recombination that imported preselected SNPs from outside of the species. Of particular note are six regions of the genome in which SNP density, homoplasy density and recombination density were all considerably higher than the average (**Fig. 2**). These regions include *ompA* (region 1), region 2 encoding the polymorphic outer membrane proteins, and region 3, which is predicted to encode four hypothetical proteins (CTL304–CTL307; gene designation from the annotation of L2/434/BU GenBank accession number AM884176) and the 3′ end of *hemN*. Recombination density was also raised around region 4, encoding a putative membrane protein (CTL0399), region 5, spanning the plasticity zone, and region 6, carrying genes predicted to encode a conserved hypothetical protein (CTL0886), a putative exported protein (CTL0887) and a putative virulence protein (*mviN*). Four of these regions (1–3 and 6) also had raised numbers of non-homoplasic SNPs, which may be the result of further recombinations from lineages not represented on our tree.

Evidence of recombination within and between biovars

Within the LGV clade, 5% (46) of the 920 variant sites were homoplasic. We reconstructed 12 recombinations within the clade, all of which were between different LGV serotypes (**Supplementary Fig. 4**) and six of which coincided with regions of the genome that we also found to recombine in other *C. trachomatis* lineages (**Fig. 2**; regions 1, 2 and 3). We identified three recombinations affecting *ompA* that could be interpreted by visualizing the SNP distribution as a genetic 'barcode' (**Fig. 3**). A region of almost 3 kb, from the 3′ end of the gene encoding the translation elongation factor, *tsf*, through the 3′ end of *ompA* to VS2 (**Fig. 3**), seems to have been replaced by homologous recombination between L1/440/LN and L1/1322/p2, explaining why these two strains, which are distinct on the tree, both typed as L1. We identified a second recombination of only 25 bp in the VS2 region of *ompA* as an exchange between the L2 and L2b clades (**Fig. 3**), explaining the reported serotypic results, as the two South African strains, typed as L1-L2 hybrids based on sequence (L1/115 and L1/224; **Fig. 3**), differed from the archetypal L2 sequence in this region. The third recombination was also located in VS2 between L1/404/LN and the L2 clade.

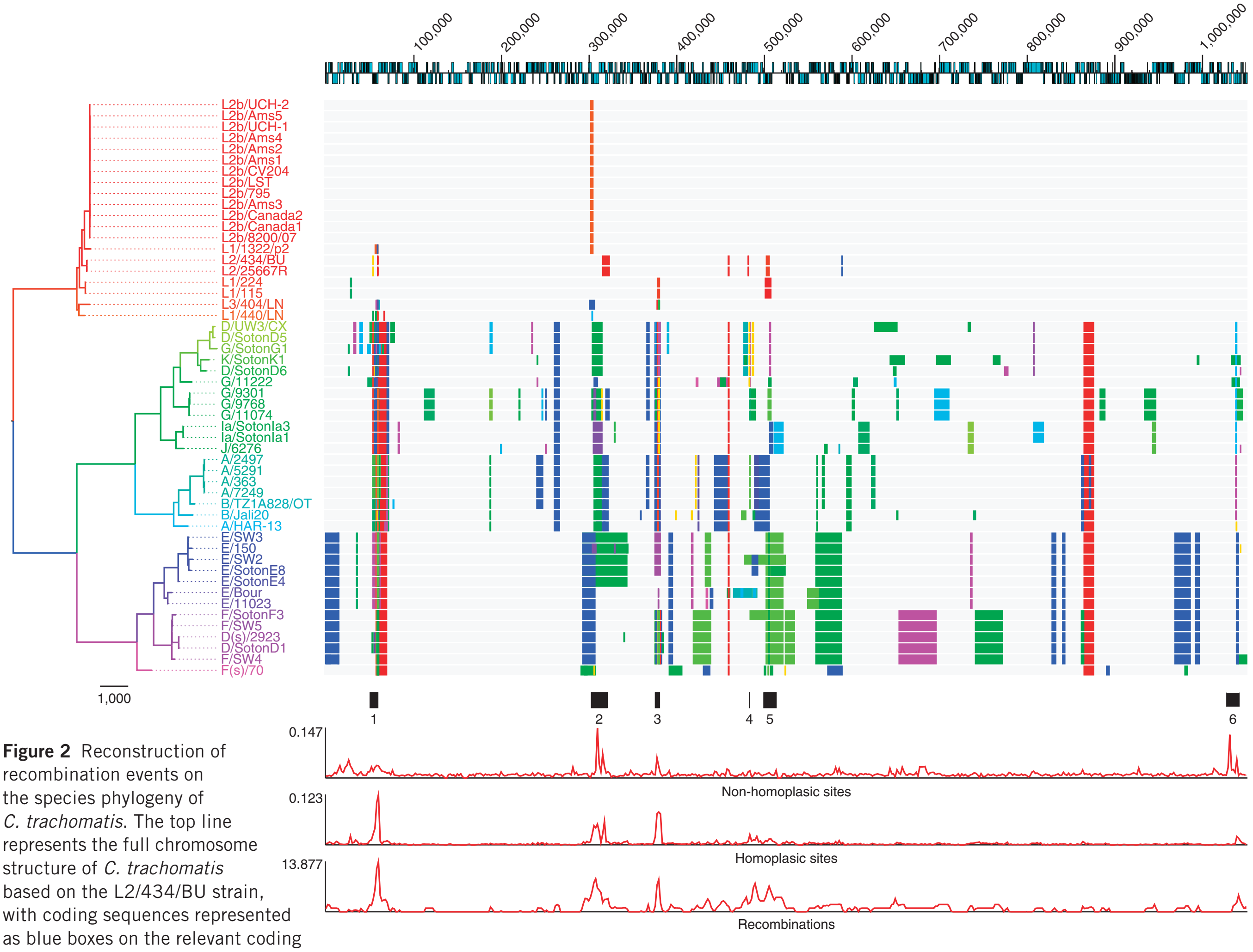

Figure 2 Reconstruction of recombination events on the species phylogeny of *C. trachomatis*. The top line represents the full chromosome structure of *C. trachomatis* based on the L2/434/BU strain, with coding sequences represented as blue boxes on the relevant coding strand. The numbers indicate the position in the genome alignment, beginning at CTL0001 (L2/434/BU GenBank accession code AM884176). Each horizontal track represents the chromosome of a strain in the species phylogeny on the left. Blocks shown on the tracks represent the location of received homologous replacements, with their color corresponding to the color of the donor branch on the tree. Tree branches and taxon names are colored by phylogenetic distance, with more similar colors representing more closely related branches. Regions of interest along the genome are highlighted immediately below the recombination tracks. Shown below are plots of the density of non-homoplasic SNP sites, homoplasic SNP sites and recombination events based on a moving window analysis. The window size used was 2,000 bp.

More notably, there was clear evidence of recombination between the LGV and trachoma biovars, with 24 recombination events being reconstructed. Fourteen of these events affect *ompA*, with a particularly clear example being the similarity of the L3 *ompA* sequence to that of trachoma serotypes Ia, J and K (**Fig. 3**), which was previously noted when the K serotype was first recognized[33]. Outside *ompA*, the clearest interbiovar recombination was a 665-bp region within *recD* in which 47 SNPs have been transferred from the T2 trachoma lineage to the two African LGV strains L1/115 and L1/224 (**Supplementary Fig. 5**).

Within the trachoma biovar, recombination events are far more prevalent than in the LGV biovar. Of the total 15,902 variant sites in the trachoma biovar, 3,367 (26.2%) showed homoplasy, and we reconstructed 162 intrabiovar recombination events (**Fig. 2**) covering 446,917 bp (43%) of the length of the aligned genomes. **Figure 3** highlights the extent of *ompA* swapping between the trachoma strains. Not only have multiple serotype switches occurred, but the SNP barcodes of *ompA* show clearly that serotypes with similar barcodes (sequences) do not necessarily fall near each other on the tree (**Fig. 3**). We also reconstructed 43 recombination events between the ocular and urogenital branches, clearly indicating that DNA exchange has occured between these two biotypes (**Fig. 2**).

Evidence of recent recombination

Our analyses provide evidence that recombination is an ongoing process rather than a purely historical event in the evolution of *C. trachomatis*. This can be clearly illustrated with two examples.

Researchers from a previous study[16] described the genome of a serotype B *C. trachomatis* strain (B/TZ1A828/OT) from Kongwa, central Tanzania. In this study, we included new genomes from four serotype A strains (A/2497, A/363, A/5291 and A/7249) from Rombo, northern Tanzania, which is 550 km from Kongwa. Our phylogenetic reconstruction placed these new strains as the sister group to B/TZ1A828/OT rather than to the reference serotype A strain, A/HAR-13. An analysis of the SNP differences between the genomes of B/TZ1A828/OT and a representative from the new serotype A strains (A/2497) (**Supplementary Fig. 6a**) showed that 434 of the 720 SNPs differentiating the two strains are located in a 10-kb region (4.34% divergence) around *ompA* (from *aspC* to *pbpB*) (**Supplementary Fig. 6b**),

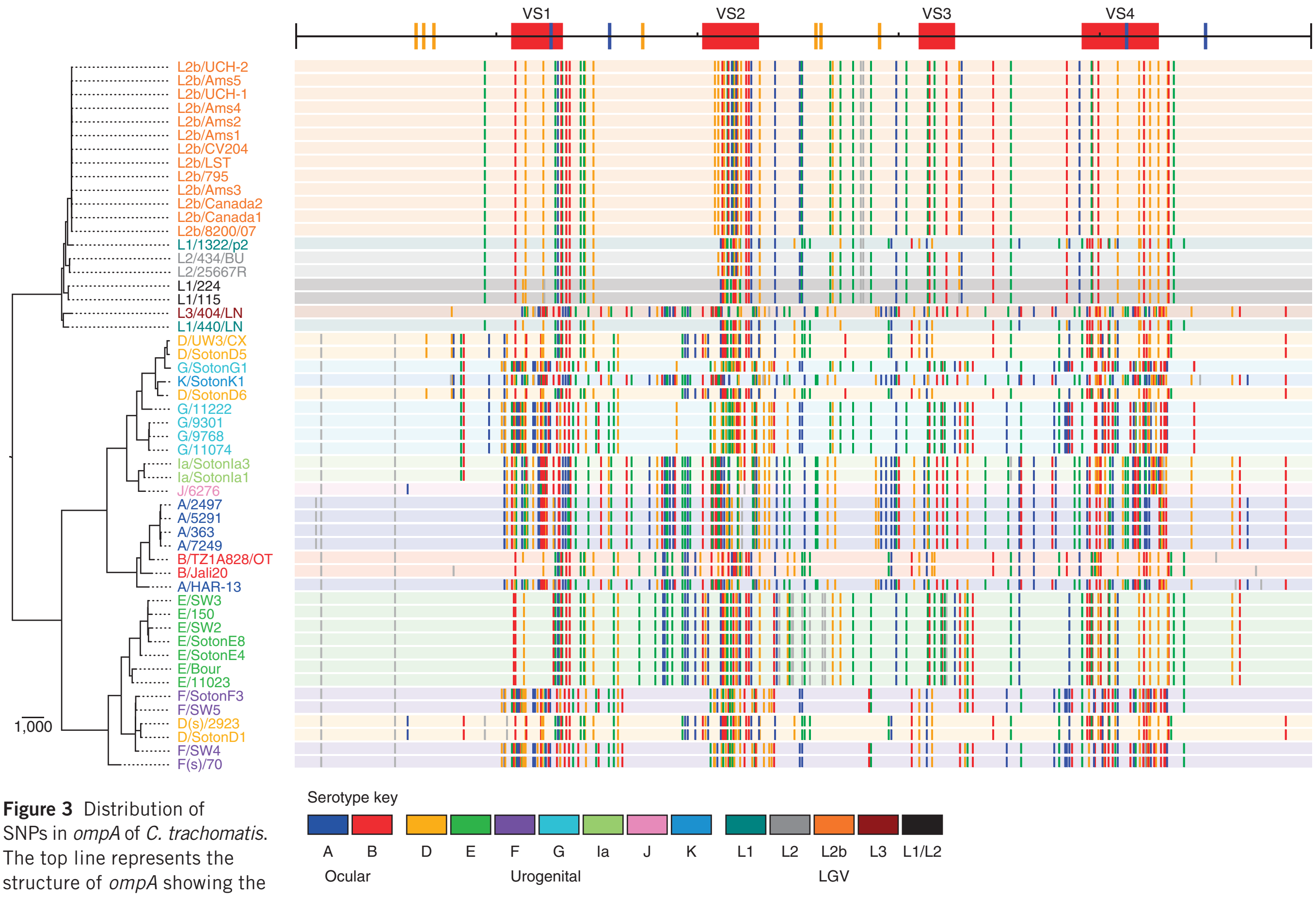

Figure 3 Distribution of SNPs in *ompA* of *C. trachomatis*. The top line represents the structure of *ompA* showing the location of variable regions (VS1–VS4, red blocks) and cysteine residues (with conserved residues shown in blue and non-conserved residues shown in orange). On the left is the species phylogeny of *C. trachomatis* with strain names colored by serotype. Adjacent to each strain name is a track with a background color based on the serotype of the corresponding strain. The c vertical lines along the tracks represent bases that differ from the ancestral sequence (gray, non-homoplasic change), and colored lines represent homoplasic bases (red, A; blue, T; green, C; orange, G). The pattern of the lines provides a barcode of *ompA* similarity between the strains.

such that the *ompA* genes in these species are highly divergent, yet the remaining 99% of the genome differs at only 286 nucleotide sites (0.027% divergence). The converse is true if the reference A/HAR-13 is compared with A/2497. In this comparison, there were 1,242 SNP differences, with only 170 being found in the same 10-kb region in and around *ompA*, and 1,072 being located in the remainder of the genome (**Supplementary Fig. 6a**,**b**). Closer inspection showed that the *ompA* gene itself is almost identical between the genomes of A/2497 and A/HAR-13 (differing by only 5 SNPs), whereas the *ompA* sequence of B/TZ1A828/OT is highly divergent from that of A/HAR-13 (199 SNPs) (**Supplementary Fig. 6c**), explaining why the four new strains are serotype A. This is clear evidence for a recent recombination event in which a divergent *ompA* gene replaced the existing *ompA* sequence and changed the serotype of the circulating clone. Although it is difficult to be certain of the direction of the change (from serotype A to B or vice versa), this example underlines the need to use genome-wide SNPs to be confident of the phylogenetic history of *Chlamydia* strains.

We also found evidence for a recent recombination event within the new-variant (nvCT) Swedish serotype E strain (E/SW2)[34]. A comparison of E/SW2 with another Swedish serotype E strain (E/SW3) showed close similarity along most of the genome length, except for a large number of SNPs in a 30-kb region spanning CTL393 to CTL417 (**Supplementary Fig. 6d**). The sequence of this region in E/SW2 exactly matches the homologous region in D/UW3/CX (**Supplementary Fig. 6d**), indicating a recombination event in E/SW2 from a D/UW3/CX-like donor. The exchanged genes are of unknown function, and a phenotypic comparison of nvCT to other serotype E strains identified no differences[35].

Plasmid phylogeny and recombination

Previously[16] we showed that the cryptic plasmids from a small sample of *C. trachomatis* strains share the same evolutionary history as their chromosomes, suggesting that the plasmid is not (or is rarely) exchanged. Analyzing our much larger dataset of 43 plasmids for evidence of recombination and exchange showed a single difference in phylogenetic structure between the chromosomal and plasmid trees (**Fig. 1**) relating to two serotype Ia strains from Southampton, UK. In the whole-genome analysis, these two strains grouped in the T2 cluster, whereas in the plasmid phylogeny, they formed a distinct branch between the T1 and T2 clades (**Fig. 1b**). Reconstructing SNP events across the plasmid tree showed that the SNPs supporting the positioning of these Ia strains were distributed along the length of the plasmid sequence (**Supplementary Fig. 7**); this provides evidence of the replacement of the Ia lineage plasmid with an equivalent plasmid, potentially from an unsampled *C. trachomatis* lineage. All other relationships were congruent between the plasmid and chromosome trees, suggesting that exchange of whole plasmids between distant *C. trachomatis* strains is indeed rare.

We identified just seven homoplasies on the *C. trachomatis* plasmid tree (**Supplementary Fig. 7**). Applying the methods for recombination detection detailed above, we identified a single recombination event that accounted for six of these homoplasies (**Supplementary Fig. 7**). The recombination region extends from the middle of plasmid coding sequence CDS3 to the middle of CDS5 and can be best explained by homologous recombination of this region between the ocular strains and the T1 clade. By reconstructing the phylogenies of the recombination region and the rest of the plasmid, we can see that the recombined region in the ocular strains clusters in the T1 rather than the T2 clade (**Supplementary Fig. 8**). Alternative explanations are recombination between the Ia plasmid and the T2 clade or the Ia plasmid being a chimera between a T2 plasmid and an unknown donor plasmid. We think the latter is unlikely, as the Ia plasmid does not cluster within a T2 clade in either the backbone or the recombined regions.

Another notable discovery was of a previously unrecognized deletion in the plasmid of LGV strain pL3/404/LN within plasmid CDS1 (**Supplementary Fig. 7**) in a different location than that of the nvCT, providing further evidence that CDS1 is not a stable diagnostic target.

DISCUSSION

Previously, studies of the evolutionary history and diversity of *C. trachomatis* have almost always been based on serological classification or on the sequences of a small number of genes. We have shown that these loci do not necessarily represent the true history of the species or the true relationships between strains, and we showed that an in-depth understanding of the population structure of *Chlamydia* requires the maximum resolution available: whole-genome data. This resolution has allowed us to observe the extent of recombination that has occurred in the population and has raised a number of issues of clinical importance.

Our analysis has confirmed that *C. trachomatis* comprises three distinct lineages. The species seems to have split early into LGV and urogenital clades, with the urogenital clade itself later splitting into two clades termed T1 and T2. Ocular *Chlamydia* is a younger lineage that, within the limits of our sampling frame, seems to have emerged only once from a urogenital ancestor within T2, although more data are needed to confirm this observation. Within the LGV biovar, the emergence of the epidemic L2b lineage is the result of a clonal expansion that probably arose from a single introduction into Europe or North America. We suggest that the lack of variation seen within the L2b genome is indicative of relatively rapid transmission (rather than a lower rate of recombination), as exemplified by the emergence and spread of the Swedish nvCT, which evaded detection by some nucleic acid amplification tests because of a deletion in the first coding sequence of the plasmid. The nvCT and L2b outbreaks have shown that given a selective advantage or lack of competition, a single lineage of *C. trachomatis* can proliferate and spread. However, for lineages with more subtle changes occurring outside either *ompA* or diagnostic primer binding sites, such as drug resistance, it would be more difficult to detect such a clonal expansion using current diagnostic and molecular typing techniques. Fortunately, there have previously been only sporadic reports of antibiotic resistance in clinical strains of *C. trachomatis*[36–39], despite the fact that it is possible to induce such resistance in the laboratory[40].

Although recombination in *C. trachomatis* was once a controversial concept, it has more recently been shown to occur both in laboratory tissue culture[17] and naturally between clinical strains[17,20,32,41,42]. We extended these observations by showing that recombination is not limited to a few hotspots around the chromosome. Rather than finding true hotspots, it is probable that researchers from previous studies[18,32] simply observed higher rates of fixation of recombinations in genomic regions that are under diversifying selection pressure.

In contrast to a previous analysis[18], our more comprehensive assessment of *C. trachomatis* diversity showed greater exchange between strains with tropism for the same site or tissue, which correlates with the majority of reports of mixed infections[43–47]. However, we identified multiple examples of recombination between strains with tropisms for different tissues, as well as recombination between biovars, suggesting that there are no absolute barriers to genetic exchange. In particular, recombination between ocular and urogenital strains appears to be relatively frequent, again correlating with reports of both cross-site and mixed infections of ocular and urogenital serotypes[47–50].

The clinical importance of our findings is extensive. We have shown clear examples in which the genetic backbone of the strain is unlinked from its serotype. Replacement and chimerism of *ompA* is probably a process of diversification that counteracts the effect of the immune system protecting the host against immediate reinfection, and, as such, it is clear that *ompA*, which is the main chromosomal diagnostic target used to detect *C. trachomatis* infections and to type their strains, is a poor indicator of genetic relatedness within the species. This may explain why there are equally as many studies that have failed to draw any significant association between disease severity and the nature of the infecting strain[51–55] as there are those that have found such an association[44,56–59].

Multilocus typing schemes, particularly those that are based on housekeeping genes under low selective pressure, more closely reflect the genome phylogeny and may prove useful in cases where the ultimate resolution of genome-wide SNP-based techniques is not necessary. However, any scheme based on a small number of loci has the potential to be confused by recombination, meaning that different MLST schemes will differ in resolution and accuracy (**Supplementary Fig. 3b**–**d**). The choice of typing approach will need to be determined on a case-by-case basis, depending on the resolution that is required to answer the question at hand.

Finally, we have shown for the first time, to our knowledge, that there has been some exchange of, and homologous recombination within, the DNA of the cryptic plasmid, providing further evidence of the potential unreliability of the plasmid for diagnostics and typing.

URLs. SMALT, http://www.sanger.ac.uk/resources/software/smalt/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Short reads and assembled genomes and plasmids have been submitted to the European Molecular Biology Laboratory (EMBL) under the accession codes listed in **Supplementary Table 1**.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the core informatics, library-making and sequencing teams at the Wellcome Trust Sanger Institute. S.R.H. is grateful for the opportunity to discuss this project at the Permafrost conference. This work was funded by the Wellcome Trust grant numbers 098051 and 080348. B.G.S. was funded by the Wellcome Trust.

AUTHOR CONTRIBUTIONS

S.R.H. assembled, aligned and analyzed the data and wrote the paper. I.N.C. jointly conceived of the project with N.R.T. and provided samples. H.M.B.S.-S. performed experiments, carried out analyses of the data and helped write the paper. L.T.C., P.M., R.J.S., M.J.H., D.M., R.W.P., D.A.L., M.U., K.P., C.B., R.B., H.J.C.d.V., S.A.M., A.W.S., C.M.B., A.S., M.C. and B.d.B. collected and cultured samples. B.G.S. and J.P. helped interpret the data and write the paper. N.R.T. conceived of and ran the project and wrote the paper.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. World Health Organization, Department of Reproductive Health and Research. *Prevalence and Incidence of Selected Sexually Transmitted Infections.* (World Health Organization, Geneva, Swizerland, 2011).
2. Mariotti, S.P., Pascolini, D. & Rose-Nussbaumer, J. Trachoma: global magnitude of a preventable cause of blindness. *Br. J. Ophthalmol.* **93**, 563–568 (2009).
3. Burgoyne, R.A. Lymphogranuloma venereum. *Prim. Care* **17**, 153–157 (1990).
4. Behets, F.M. *et al.* Chancroid, primary syphilis, genital herpes, and lymphogranuloma venereum in Antananarivo, Madagascar. *J. Infect. Dis.* **180**, 1382–1385 (1999).
5. Mabey, D. & Peeling, R.W. Lymphogranuloma venereum. *Sex. Transm. Infect.* **78**, 90–92 (2002).
6. Viravan, C. *et al.* A prospective clinical and bacteriologic study of inguinal buboes in Thai men. *Clin. Infect. Dis.* **22**, 233–239 (1996).
7. Clarke, I.N. Evolution of *Chlamydia trachomatis*. *Ann. NY Acad. Sci.* **1230**, E11–E18 (2011).
8. Gaydos, C.A. Nucleic acid amplification tests for gonorrhea and *Chlamydia*: practice and applications. *Infect. Dis. Clin. North Am.* **19**, 367–386, ix (2005).
9. Fredlund, H., Falk, L., Jurstrand, M. & Unemo, M. Molecular genetic methods for diagnosis and characterisation of *Chlamydia trachomatis* and *Neisseria gonorrhoeae*: impact on epidemiological surveillance and interventions. *APMIS* **112**, 771–784 (2004).
10. Nunes, A., Borrego, M.J., Nunes, B., Florindo, C. & Gomes, J.P. Evolutionary dynamics of *ompA*, the gene encoding the *Chlamydia trachomatis* key antigen. *J. Bacteriol.* **191**, 7182–7192 (2009).
11. Pudjiatmoko, Fukushi, H., Ochiai, Y., Yamaguchi, T. & Hirai, K. Phylogenetic analysis of the genus *Chlamydia* based on 16S rRNA gene sequences. *Int. J. Syst. Bacteriol.* **47**, 425–431 (1997).
12. Klint, M. *et al.* High-resolution genotyping of *Chlamydia trachomatis* strains by multilocus sequence analysis. *J. Clin. Microbiol.* **45**, 1410–1414 (2007).
13. Pannekoek, Y. *et al.* Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria *Chlamydia trachomatis*. *BMC Microbiol.* **8**, 42 (2008).
14. Brunelle, B.W. & Sensabaugh, G.F. The *ompA* gene in *Chlamydia trachomatis* differs in phylogeny and rate of evolution from other regions of the genome. *Infect. Immun.* **74**, 578–585 (2006).
15. Dean, D. *et al.* Predicting phenotype and emerging strains among *Chlamydia trachomatis* infections. *Emerg. Infect. Dis.* **15**, 1385–1394 (2009).
16. Seth-Smith, H.M. *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).
17. Jeffrey, B.M. *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).
18. Joseph, S.J., Didelot, X., Gandhi, K., Dean, D. & Read, T.D. Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. *Biol. Direct* **6**, 28 (2011).
19. Ikryannikova, L.N., Shkarupeta, M.M., Shitikov, E.A., Il'ina, E.N. & Govorun, V.M. Comparative evaluation of new typing schemes for urogenital *Chlamydia trachomatis* isolates. *FEMS Immunol. Med. Microbiol.* **59**, 188–196 (2010).
20. Millman, K.L., Tavare, S. & Dean, D. Recombination in the *ompA* gene but not the *omcB* gene of *Chlamydia* contributes to serovar-specific differences in tissue tropism, immune surveillance, and persistence of the organism. *J. Bacteriol.* **183**, 5997–6008 (2001).
21. Stephens, R.S. *et al.* Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754–759 (1998).
22. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
23. Stephens, R.S. *Chlamydiae* in evolution: a billion years and counting. in *Proceedings of the Tenth International Symposium on Human Chlamydial Infections.* (eds. Schachter, J. *et al.*) 3–12 (Antalya, Turkey, 2002).
24. Suchland, R.J., Eckert, L.O., Hawes, S.E. & Stamm, W.E. Longitudinal assessment of infecting serovars of *Chlamydia trachomatis* in Seattle public health clinics: 1988–1996. *Sex. Transm. Dis.* **30**, 357–361 (2003).
25. Nieuwenhuis, R.F., Ossewaarde, J.M., van der Meijden, W.I. & Neumann, H.A. Unusual presentation of early lymphogranuloma venereum in an HIV-1 infected patient: effective treatment with 1 g azithromycin. *Sex. Transm. Infect.* **79**, 453–455 (2003).
26. Hayes, L.J. *et al.* Evidence for naturally occurring recombination in the gene encoding the major outer membrane protein of lymphogranuloma venereum isolates of *Chlamydia trachomatis*. *Infect. Immun.* **62**, 5659–5663 (1994).
27. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
28. LeQuesne, W. A method of selection of characters in numerical taxonomy. *Syst. Biol.* **18**, 201–205 (1969).
29. Bruen, T.C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
30. Smith, J.M. Analyzing the mosaic structure of genes. *J. Mol. Evol.* **34**, 126–129 (1992).
31. Jakobsen, I.B. & Easteal, S. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput. Appl. Biosci.* **12**, 291–295 (1996).
32. Gomes, J.P. *et al.* Evolution of *Chlamydia trachomatis* diversity occurs by widespread interstrain recombination involving hotspots. *Genome Res.* **17**, 50–60 (2007).
33. Kuo, C.C., Wang, S.P., Grayston, J.T. & Alexander, E.R. TRIC type K, a new immunologic type of *Chlamydia trachomatis*. *J. Immunol.* **113**, 591–596 (1974).
34. Unemo, M. & Clarke, I.N. The Swedish new variant of *Chlamydia trachomatis*. *Curr. Opin. Infect. Dis.* **24**, 62–69 (2011).
35. Unemo, M. *et al.* The Swedish new variant of *Chlamydia trachomatis*: genome sequence, morphology, cell tropism and phenotypic characterization. *Microbiology* **156**, 1394–1404 (2010).
36. Misyurina, O.Y. *et al.* Mutations in a 23S rRNA gene of *Chlamydia trachomatis* associated with resistance to macrolides. *Antimicrob. Agents Chemother.* **48**, 1347–1349 (2004).
37. Jones, R.B., Van der Pol, B., Martin, D.H. & Shepard, M.K. Partial characterization of *Chlamydia trachomatis* isolates resistant to multiple antibiotics. *J. Infect. Dis.* **162**, 1309–1315 (1990).
38. Lefevre, J.C., Lepargneur, J.P., Guion, D. & Bei, S. Tetracycline-resistant *Chlamydia trachomatis* in Toulouse, France. *Pathol. Biol. (Paris)* **45**, 376–378 (1997).
39. Somani, J., Bhullar, V.B., Workowski, K.A., Farshy, C.E. & Black, C.M. Multiple drug-resistant *Chlamydia trachomatis* associated with clinical treatment failure. *J. Infect. Dis.* **181**, 1421–1427 (2000).
40. Binet, R. & Maurelli, A.T. Fitness cost due to mutations in the 16S rRNA associated with spectinomycin resistance in *Chlamydia psittaci* 6BC. *Antimicrob. Agents Chemother.* **49**, 4455–4464 (2005).
41. Brunham, R. *et al. Chlamydia trachomatis* from individuals in a sexually transmitted disease core group exhibit frequent sequence variation in the major outer membrane protein (*omp1*) gene. *J. Clin. Invest.* **94**, 458–463 (1994).
42. Gomes, J.P., Bruno, W.J., Borrego, M.J. & Dean, D. Recombination in the genome of *Chlamydia trachomatis* involving the polymorphic membrane protein C gene relative to *ompA* and evidence for horizontal gene transfer. *J. Bacteriol.* **186**, 4295–4306 (2004).
43. Jurstrand, M. *et al.* Characterization of *Chlamydia trachomatis omp1* genotypes among sexually transmitted disease patients in Sweden. *J. Clin. Microbiol.* **39**, 3915–3919 (2001).
44. van Duynhoven, Y.T., Ossewaarde, J.M., Derksen-Nawrocki, R.P., van der Meijden, W.I. & van de Laar, M.J. *Chlamydia trachomatis* genotypes: correlation with clinical manifestations of infection and patients' characteristics. *Clin. Infect. Dis.* **26**, 314–322 (1998).
45. Moncan, T., Eb, F. & Orfila, J. Monoclonal antibodies in serovar determination of 53 *Chlamydia trachomatis* isolates from Amiens, France. *Res. Microbiol.* **141**, 695–701 (1990).
46. Wagenvoort, J.H., Suchland, R.J. & Stamm, W.E. Serovar distribution of urogenital *Chlamydia trachomatis* strains in The Netherlands. *Genitourin. Med.* **64**, 159–161 (1988).
47. Barnes, R.C., Suchland, R.J., Wang, S.P., Kuo, C.C. & Stamm, W.E. Detection of multiple serovars of *Chlamydia trachomatis* in genital infections. *J. Infect. Dis.* **152**, 985–989 (1985).
48. Hanna, L., Thygeson, P. & Jawetz, E. Elementary-body virus isolated from clinical trachoma in California. *Science* **130**, 1339–1340 (1959).
49. Spaargaren, J. *et al.* Analysis of *Chlamydia trachomatis* serovar distribution changes in the Netherlands (1986–2002). *Sex. Transm. Infect.* **80**, 151–152 (2004).
50. Dean, D. & Stephens, R.S. Identification of individual genotypes of *Chlamydia trachomatis* from experimentally mixed serovars and mixed infections among trachoma patients. *J. Clin. Microbiol.* **32**, 1506–1510 (1994).
51. Machado, A.C. *et al.* Distribution of *Chlamydia trachomatis* genovars among youths and adults in Brazil. *J. Med. Microbiol.* **60**, 472–476 (2011).
52. Batteiger, B.E. *et al.* Correlation of infecting serovar and local inflammation in genital chlamydial infections. *J. Infect. Dis.* **160**, 332–336 (1989).
53. Millman, K. *et al.* Population-based genetic and evolutionary analysis of *Chlamydia trachomatis* urogenital strain variation in the United States. *J. Bacteriol.* **186**, 2457–2465 (2004).
54. Persson, K. & Osser, S. Lack of evidence of a relationship between genital symptoms, cervicitis and salpingitis and different serovars of *Chlamydia trachomatis*. *Eur. J. Clin. Microbiol. Infect. Dis.* **12**, 195–199 (1993).
55. Lysén, M. *et al.* Characterization of *ompA* genotypes by sequence analysis of DNA from all detected cases of *Chlamydia trachomatis* infections during 1 year of contact tracing in a Swedish County. *J. Clin. Microbiol.* **42**, 1641–1647 (2004).
56. Sturm-Ramirez, K. *et al.* Molecular epidemiology of genital *Chlamydia trachomatis* infection in high-risk women in Senegal, West Africa. *J. Clin. Microbiol.* **38**, 138–145 (2000).
57. Geisler, W.M., Suchland, R.J., Whittington, W.L. & Stamm, W.E. The relationship of serovar to clinical manifestations of urogenital *Chlamydia trachomatis* infection. *Sex. Transm. Dis.* **30**, 160–165 (2003).
58. Gao, X. *et al.* Distribution study of *Chlamydia trachomatis* serovars among high-risk women in China performed using PCR-restriction fragment length polymorphism genotyping. *J. Clin. Microbiol.* **45**, 1185–1189 (2007).
59. van de Laar, M.J. *et al.* Differences in clinical manifestations of genital chlamydial infections related to serovars. *Genitourin. Med.* **72**, 261–265 (1996).

ONLINE METHODS

Cell culture, DNA extraction and sequencing. The strains and sources of *C. trachomatis* used in this work are summarized in **Supplementary Table 1**. Cell culture and DNA extraction was performed as previously described[16] for all strains except A/363, A/5291 and A/7249, which were extracted from a single 24-well plate using 1 N NaOH followed by neutralization with Tris. The genome of A/2497 was sequenced to a depth of 12× coverage derived from plasmid pUC18 (insert size, 0.7 kb) small-insert libraries using dye terminator chemistry on ABI3700 automated sequencers. End sequences from larger-insert plasmid (pMAQ1, 9–12 kb insert size) libraries were used as a scaffold. Sequencing, assembly, finishing and checking of this genome was performed as described[60]. The genome sequence of strain L2b/UCH-1 was improved with Illumina GAII data (2,877,472 37-bp paired end reads) using iCORN[61]. All other genomes were sequenced using the Illumina Genome Analyzer (Illumina) as summarized in **Supplementary Table 1**. Additionally, PCRs were performed on genomic DNA to confirm the sequences of the repetitive regions within *hctB* and *tarp* and at the plasmid origin using Platinum Pfx (Invitrogen) at an annealing temperature of 60 °C using the primers HC2_f, HC2_r, tarp_f, tarp_r, pori_f and pori_f (**Supplementary Table 2**). PCR products were sequenced using the primers above plus the primers tarp_s1- tarp_s7 (**Supplementary Table 2**).

Assembly and alignment. Illumina reads were assembled using velvet v1.0.12 (ref. 62). Because of the small size and non-repetitive nature of the *C. trachomatis* genome, most assemblies produced only a small number of contigs that could be scaffolded by hand using the genome of L2/434/BU[63] as a reference (GenBank accession code AM884176), with the circular genome cut at the origin immediately upstream of *hemB*. Manual insertion of *hctB* and *tarp* gene sequences resulted in Improved High Quality Draft Sequences[64] made up of 1–11 contigs (**Supplementary Table 1**). Plasmid assemblies were completed by PCR across the origin to give a single contig. To double check the assemblies and ensure than none of our cultures was made up of mixed populations, raw data was mapped back against the assemblies using SMALT (see URLs) to check for heterozygosity. All genome and plasmid sequences have been deposited at EMBL under the accession codes listed in **Supplementary Table 1**. Complete genomes were aligned with progressiveMauve[65] using the collinear option. The resulting alignment was manually checked, and misalignments were improved.

To allow the phylogenetic tree to be rooted, the sequence of *C. muridarum* strain Nigg (GenBank accession code AE002160.2) was added to the alignment. Because of the divergent nature of this sequence, it was not possible to align it to the rest of the sequences with progressiveMauve. Instead, it was fragmented *in silico* and then mapped to the *C. trachomatis* alignment using SMALT.

Phylogenetic reconstruction. Because of limitations of assembly of repetitive regions using short-read data, repetitive regions of the consensus of aligned genomes were identified using REPuter[66] and were excluded from phylogenetic analysis. A phylogenetic reconstruction of the alignment data was carried out using RAxML v7.0.4 (ref. 67) using a Generalized Time Reversible model of evolution with a γ correction for among-site rate variation with four rate categories. To reduce the effect of recombination on the phylogeny, a previously described iterative recombination removal method[22] was used with 100 bootstrap replicates calculated on the final tree to provide a measure of support for the relationships identified.

Ancestral sequence reconstruction and recombination detection. Ancestral sequences were reconstructed onto each node of the phylogeny using PAML[27]. From these ancestral sequences, SNPs were reconstructed onto branches of the tree. To identify recombination in the data, we applied a moving window approach similar to that previously used[22]. Researchers from the previous study identified regions of high SNP density on a given branch of the *Streptococcus pneumoniae* PMEN1 phylogeny and hypothesized these to represent recombination events from a source outside of the PMEN1 lineage. Here the same principal was applied exclusively to homoplasic SNPs to identify regions of homologous replacement within the tree. First, homoplasic SNPs were identified for each branch of the tree using the PAML reconstruction of ancestral sequences. The DNA sequences on each branch (the potential recipient) were then compared, in turn, to the corresponding bases along all other branches on the tree to identify potential donors. For all pairwise comparisons where at least three shared homoplasies were found, a moving window approach was used to identify regions where these occurred at a higher density than would be expected if they were randomly acquired. The lengths of these regions were then refined using a spatial scanning statistic as previously described[22], which alters the length of the region until the recombination likelihood is maximized. Once the likelihood of all the clusters was calculated for all recipient branches, SNPs produced by the cluster with the highest likelihood were removed, and the process repeated iteratively until no significant homoplasy clusters were identified.

60. Parkhill, J. *et al.* Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502–506 (2000).
61. Otto, T.D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**, 1704–1707 (2010).
62. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
63. Thomson, N.R. *et al.* Chlamydia trachomatis: genome sequence analysis of lymphogranuloma venereum isolates. *Genome Res.* **18**, 161–171 (2008).
64. Chain, P.S. *et al.* Genomics. Genome project standards in a new era of sequencing. *Science* **326**, 236–237 (2009).
65. Darling, A.E., Mau, B. & Perna, N.T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
66. Kurtz, S. & Schleiermacher, C. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**, 426–427 (1999).
67. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).

doi:10.1038/ng.2214