Complete genome sequence of Treponema paraluiscuniculi, strain Cuniculi A: The loss of infectivity to humans is associated with genome decay

David Smajs
Masaryk University

Marie Zobanikova
Masaryk University

Michal Strouhal
Masaryk University

Darina Cejkova
Masaryk University

Shannon Dugan-Rocha
Baylor College of Medicine

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Part of the Medicine and Health Sciences Commons

Recommended Citation
Smajs, David; Zobanikova, Marie; Strouhal, Michal; Cejkova, Darina; Dugan-Rocha, Shannon; Pospisilova, Petra; Norris, Steven J.; Albert, Tom; Qin, Xiang; Hallsworth-Pepin, Kym; Buhay, Christian; Muzny, Donna M.; Chen, Lei; Gibbs, Richard A.; and Weinstock, George M., ,"Complete genome sequence of Treponema paraluiscuniculi, strain Cuniculi A: The loss of infectivity to humans is associated with genome decay." PLoS One., e20415. (2011).
https://digitalcommons.wustl.edu/open_access_pubs/610
Authors
David Smajs, Marie Zobanikova, Michal Strouhal, Darina Cejkova, Shannon Dugan-Rocha, Petra Pospisilova, Steven J. Norris, Tom Albert, Xiang Qin, Kym Hallsworth-Pepin, Christian Buhay, Donna M. Muzny, Lei Chen, Richard A. Gibbs, and George M. Weinstock

This open access publication is available at Digital Commons@Becker: https://digitalcommons.wustl.edu/open_access_pubs/610
Complete Genome Sequence of *Treponema paraluiscuniculi*, Strain Cuniculi A: The Loss of Infectivity to Humans Is Associated with Genome Decay

David Šmajs1*, Marie Zobaniková1, Michal Strouhal1, Darina Čejková1, Shannon Dugan-Rocha2, Petra Pospíšilová1, Steven J. Norris³, Tom Albert4, Xiang Qin2, Kym Hallsworth-Pepin5, Christian Buhay2, Donna M. Munzy2, Lei Chen2,5, Richard A. Gibbs3, George M. Weinstock2,5

1 Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic, 2 Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United States of America, 3 Department of Pathology and Laboratory Medicine, University of Texas Medical School at Houston, Houston, Texas, United States of America, 4 NimbleGen Systems, Inc., Madison, Wisconsin, United States of America, 5 The Genome Sequencing Center, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, United States of America

Abstract

*Treponema paraluiscuniculi* is the causative agent of rabbit venereal spirochetosis. It is not infectious to humans, although its genome structure is very closely related to other pathogenic *Treponema* species including *Treponema pallidum* subspecies *pallidum*, the etiological agent of syphilis. In this study, the genome sequence of *Treponema paraluiscuniculi*, strain Cuniculi A, was determined by a combination of several high-throughput sequencing strategies. Whereas the overall size (1,133,390 bp), arrangement, and gene content of the Cuniculi A genome closely resembled those of the *T. pallidum* genome, the *T. paraluiscuniculi* genome contained a markedly higher number of pseudogenes and gene fragments (51). In addition to pseudogenes, 33 divergent genes were also found in the *T. paraluiscuniculi* genome. A set of 32 (out of 84) affected genes encoded proteins of known or predicted function in the Nichols genome. These proteins included virulence factors, gene regulators and components of DNA repair and recombination. The majority (52 or 61.9%) of the Cuniculi A pseudogenes and divergent genes were of unknown function. Our results indicate that *T. paraluiscuniculi* has evolved from a *T. pallidum*-like ancestor and adapted to a specialized host-associated niche (rabbits) during loss of infectivity to humans. The genes that are inactivated or altered in *T. paraluiscuniculi* are candidates for virulence factors important in the infectivity and pathogenesis of *T. pallidum* subspecies.

Introduction

*Treponema paraluiscuniculi* is a noncultivable species of the genus *Treponema* that causes venereal spirochetosis in rabbits, but that is not infectious to humans [1]. The genome structure and chromosome sequence of *T. paraluiscuniculi* is closely related to other pathogenic species and subspecies of the *Treponema* genus, including syphilis causing spirochete *Treponema pallidum* ssp. *pallidum* [2] and *Treponema pallidum* ssp. *pertene*, the causative agent of yaws.

The presence of spirochetes resembling *T. pallidum* in rabbit genital lesions was reported as early as 1912; and organism was described as *Spirochaeta paralues-cuniculi* (syphilis-like spirochetes in rabbits) by Jacobsthal [3]. The naturally occurring infection of rabbits with *T. paraluiscuniculi* is described in a detailed historical review by Smith and Pesetsky [4], as well as in more recent articles by DiGiacomo et al. [5,6]. The disease is typically sexually transmitted, and results in erythema and edema of the prepuse, vagina, anus, or scrotum, often followed by ulceration and crusting (eschar formation) of the lesion. Infection of the nose, eyelids, lips, and paws can also occur. Intradermal inoculation of rabbits with either *T. paraluiscuniculi* or *T. pallidum* results in erythematous lesions that may undergo ulceration [4–8]. *T. paraluiscuniculi* lesions are noted to be less indurated (raised and hardened) than *T. pallidum* lesions. *T. pallidum* subspecies and *T. paraluiscuniculi* are nearly indistinguishable in terms of morphology, antigen content, and physiology [8–10], consistent with the close genetic relationship among these organisms [2]. However, *T. paraluiscuniculi* and *T. pallidum* cause different diseases with different host specificity.

Rabbit venereal spirochetosis is characterized by localized genital lesions, whereas human syphilis is a multistage, sexually transmitted disease with varied clinical manifestations. Syphilis treponemes can infect virtually any human tissue, causing gummatous, neurologic, and cardiovascular manifestations [11,12]. While there is little evidence for systemic manifestations following *T.
Infection in rabbits, the organism can disseminate and be recovered from lymph nodes months after infection [4,13]. Evidence to date indicates that *T. paraluiscuniculi* is not pathogenic to humans, indicating a basic difference between this organism and the *T. pallidum* subspecies. Two studies involving a total of three volunteers described experimental inoculation of humans [1,14] with rabbit virulent strains of *T. paraluiscuniculi* with negative results. When injected intradermally into human volunteers, rabbit virulent strains of *T. paraluiscuniculi* caused only mild, local erythema and edema (without systemic effects) that disappeared after three weeks [1,14]. In the Graves and Downes study [1], only a limited serological response in the human volunteer was observed. In contrast, intradermally inoculated rabbits exhibited prominent, long-lasting lesions and a strong serological response to *T. paraluiscuniculi*.

Partial immunological cross-protection between *T. paraluiscuniculi* and *T. pallidum* has been observed, as demonstrated by infecting rabbits with one species, inoculating intradermally with the other species 3–6 months later, and observing lesion development for rabbits with one species, inoculating intradermally with the other species 3–6 months later, and observing lesion development for 3–6 months later, and observing the development of lesions reflect the primary genetic differences between these species [8,9,13,15]. Heterogeneity in the paralogous gene families of these organisms have been characterized [16,17], and may be responsible in part for the pathogenic differences and incomplete immunologic cross-protection observed.

Differences in host specificity and in clinical manifestations of these diseases reflect the primary genetic differences between *T. paraluiscuniculi* and *T. pallidum*. In this communication, we report a complete genome sequence of *T. paraluiscuniculi*, strain Cuniculi A and compare this sequence to the published genomes of *T. pallidum* ssp. *pallidum* Nichols, SS14 and Chicago [18–20].

**Results**

*T. paraluiscuniculi* Cuniculi A genomic parameters and annotation

The summarized genomic features of *T. paraluiscuniculi* strain Cuniculi A are shown in the Table 1. The genome size of *T. paraluiscuniculi* Cuniculi A (1,133,390 bp) is 4.6, 5.9 and 6.1 kb smaller than the genome size of the previously published *T. pallidum* ssp. *pallidum* Nichols (1,139,011 bp) [18], Chicago (1,139,281 bp) [20] and SS14 genomes (1,139,457 bp) [19], respectively. Similar whole genome nucleotide diversity (π ± SD) of 0.01029±0.00514, 0.01021±0.00511, and 0.01016±0.00500, was revealed by DnaSP v5 software between Cuniculi A genome and Nichols, Chicago, and SS14 genomes, respectively. The deletions in the Cuniculi A genome were not evenly distributed in the genome and were predominantly localized in tpr loci and the vicinity of these regions [2,16]. The overall gene order in the Cuniculi A was identical to both Nichols and SS14 genomes. Out of 1,133,390 bp of the Cuniculi genome, 1,092,714 bp were aligned with the Nichols genome and this part of the genome contained 8074 single nucleotide replacements and additional 1124 bp changes in 224 individual indels. This corresponds to 99.16% sequence identity between the conserved regions of the Nichols and Cuniculi A genomes. Out of 1016 annotated genes encoding proteins in the Cuniculi A genome, gene function was predicted for 650 genes (64%). In addition to genes present in the Nichols genome, 57 hypothetical genes were newly annotated in the Cuniculi A genome. These genes were annotated in the orthologous sequences and did not represent a new genetic material specific to Cuniculi A. Instead, they represented differences in the recent annotation algorithms compared to the previously used ones [18]. The average gene length of these 57 genes was of 234 nt (median length of 162 nt) indicating markedly shorter gene length compared to average and median gene length of all Cuniculi A genes (1006 and 873 bp, respectively). Seventeen automated gene predictions were omitted in the final annotation mainly due to overlap with other already annotated genes. 75 and 280 genes were annotated shorter or longer, respectively, in the Cuniculi A annotation when compared to the length of orthologous genes in the Nichols and SS14 genome annotations. 47 genes originally annotated in the Nichols genome [18] were not annotated in the Cuniculi A genome as a result of differences in gene prediction criteria. For all of these 47 genes, orthologous sequences were found in the Cuniculi A genome. All of these 47 genes encoded short hypothetical proteins with average gene length of 196.4 bp (median of 130 bp).

| Table 1. Summary of the genomic features of *T. paraluiscuniculi* strain Cuniculi A. |
|--------------------------|---------------------------|
| **Genome parameter**     | **Value**                 |
| Genome size              | 1,133,390 bp              |
| G+C content             | 52.8 %                    |
| No. of predicted genes  | 1070 including 54 untranslated genes |
| Intergenic region length| 62,494 bp (5.5% of the genome length) |
| Average/median gene length| 1006/873                  |
| No. of genes encoded on plus/minus DNA strand | 577/493 |
| No. of genes encoding hypothetical proteins similar to proteins of known function | 650 |
| No. of genes encoding conserved hypothetical proteins | 139 |
| No. of genes encoding hypothetical proteins | 227 |
| No. of pseudogenes or gene fragments | 51 (21 in genes with predicted function and 30 in hypothetical and predicted genes) |
| No. of fused genes       | 52 (resulting in 25 corresponding genes in Cuniculi A, Table S1) |
| No. of tRNA loci         | 45                        |
| No. of rRNA operons      | 2 (6 genes)               |
| No. of other stable RNAs | 3                         |

Gene fragments and fused genes were identified in comparison to the genome sequence of *Treponema pallidum* subspecies *pallidum* Nichols. doi:10.1371/journal.pone.0020415.t001
Alignment of the Cuniculi A genome with the annotated Nichols genes identified 84 orthologous Cuniculi A regions/genes containing internal frameshifts and/or major sequence changes. In the Cuniculi A genome, 63 genes (6.2% of 1016 protein-encoding genes) were annotated in these orthologous loci. Genes with major sequence changes were defined as those causing more than 10 continuous amino acid replacements (or indels) in the corresponding protein sequence or showing more than 20 dispersed amino acid changes in the Cuniculi A protein compared to the Nichols ortholog. Changes in protein length at the N-terminus resulting from predictions of longer Cuniculi A genes were not considered as major sequence changes if there was an existing potential downstream start codon at corresponding gene position as in the Nichols genome. A set of 154 (13.2% of 1016 protein-encoding genes) predicted Cuniculi A genes encoded identical proteins to predicted Nichols proteins and 819 genes (80.6%) encode proteins with one or several amino acid replacements.

**Gene fusions**

Compared to the published Nichols genome, 52 Nichols orthologs were fused into 25 genes in the Cuniculi A genome (see Table S1). In two cases, three genes annotated in the Nichols genome were fused to one gene in the Cuniculi A genome (TP0006, TP0007, TP0008 and TP0174, TP0175, TP0176). Similar situation was also found in the genome of *T. pallidum* subspecies *pertenue* Samoa D where sequencing revealed fusion of 48 Nichols orthologs into 23 genes (data not shown). An ongoing resequencing of the Nichols genome (P. Pospšilová, personal communication) revealed that most of the observed gene fusions are also present in the Nichols genome indicating that the published Nichols sequence [18] contains dozens of sequencing errors. To test whether these Nichols changes are sequencing errors or intrastrain adaptive mutations, we analyzed 208 nucleotide positions in which the resequenced Nichols genome differs from the published Nichols version [18] in 3 *T. pallidum* strains including Chicago [20] and preliminary DAL-1 and Mexico A whole genome sequences (unpublished data). Out of 208 nucleotide differences found in the resequenced Nichols genome, 179 (86.1%) were present in all 3 tested genomes, 12 (5.8%) were present in one or two other genomes and 17 (3.2%) were specific for the newly sequenced Nichols strain. The presence of majority of nucleotide changes identified in the resequenced Nichols genome in three other *T. pallidum* genomes indicates sequencing errors in the published Nichols genome [18] rather than recently emerged intrastrain adaptive mutations.

**Whole genome fingerprinting and sequencing error rate**

Whole genome fingerprinting was used to assess the overall genome assembly of the Cuniculi A genome. Fingerprinting of the Cuniculi A genome described earlier [2] was extended by additional restriction enzyme analyses to reduce the length of the resulting DNA fragments. The *in silico* restriction mapping was compared to experimentally obtained restriction digest patterns. In the final assembly, 2017 restriction target sites were experimentally verified resulting in average length of assessed DNA fragment of 562 bp. The 2017 restriction target sites corresponded to a total sequence length of 11,702 bp (1.0% of the genome length). Since no discrepancies between *in silico* and experimental restriction analysis were observed, the expected corresponding sequencing error rate in the Cuniculi A genome was set to the order of 10^{-7} or better.

### T. paraluiscuniculi genes encoding identical proteins as corresponding orthologs in the Nichols genome

Altogether, 154 (13.2%) of Cuniculi A genes were found to encode identical proteins to those encoded in the Nichols genome (Table S2 and S3). 35 of these genes (Table S3) encoded proteins of unknown function and the remaining 99 genes encoded proteins involved in the translation of mRNA (32 genes), general metabolism (24 genes), transport (12 genes), flagellar synthesis (11 genes), gene regulation (6 genes) and other functions (14 genes). Conservation of these proteins and their predominant involvement in translation and general metabolism indicate that these genes are housekeeping genes under strong negative selection in the genus *Treponema*. The identical genes encoding proteins of unknown function (35, or 26.1%) may encode proteins needed for as yet unidentified essential functions, including general metabolic processes. Although the median RNA transcript level of these 35 genes in *T. pallidum* ssp. *pallidum* Nichols [21] is lower than that of the 99 identical genes with annotated functions (0.92 and 1.37, respectively), it is close to the median value for all genes with annotated functions (0.93).

**Genes containing frameshifts and/or major sequence changes (MSC) in T. paraluiscuniculi functional gene groups**

*T. paraluiscuniculi* genes were classified into 7 functional groups according to the modified classification used by Fraser et al. [18] (Table 2). A set of 161 genes (9.8% of all protein-encoding genes) encoding proteins involved in general metabolic functions was used as an internal standard for comparisons with genes from other functional groups. The number of genes containing frameshifts and/or major sequence changes (related to Nichols orthologs) was compared to the number of these genes present in the general metabolism group (1.9%, Table 2) and statistically significant differences were found in the group of virulence factors (41.9%, *p* < 0.001), in genes with an unknown function (14.0%, *p* < 0.001), and in genes involved in DNA metabolism (7.8%, *p* = 0.037).

**Hypothetical genes with internal frameshifts and/or major sequence changes in the T. paraluiscuniculi genome**

Altogether, 52 Cuniculi A hypothetical genes corresponding to 54 Nichols orthologs (Table S4) in the *T. paraluiscuniculi* genome showed frameshifts (21 genes), partial or complete gene deletions (7 genes), internal stop codons (2 genes) or multiple nucleotide changes (22 genes). Twelve of these proteins were predicted inner or outer membrane proteins and 9 of them were identified as antigens. These results indicate that the *T. paraluiscuniculi* genome may be undergoing gene degradation and loss relative to *T. pallidum* subsp. *pallidum*, consistent with a decreased requirement for these genes in the rabbit tissue niche inhabited by *T. paraluiscuniculi*. Altogether, for 25 out these 52 Cuniculi A hypothetical genes, the Ka/Ks ratios and predicted type of selection were calculated (Table S4). Most of the genes (20) were found under neutral selection, while 4 genes under purifying and one under positive selection.

**Cuniculi A genes with predicted cell function containing internal frameshifts and/or major sequence changes**

Altogether, 32 Cuniculi A genes with defined or predicted functions were found to contain frameshifts or major deletions
(resulting in 21 pseudogenes), major sequence changes (8 genes), or reverted frameshift mutations (3 genes) (Table 3). Ten out of these 32 genes were tpr genes encoding paralogous proteins with sequence similarity to the major surface protein (Msp) of Treponema denticola [24]. A schematic representation of all tpr genes in the Cuniculi A genome is shown in Fig. 1. In addition to these potential virulence factors, another three proteins encoded by TPCCA_0136 (fibronectin-binding protein), TPCCA_0326 (tp92, outer membrane protein), and TPCCA_0433 (acidic repeat protein, Arp protein) showed indels and major sequence changes. Together with TPCCA_0760 (penicillin-binding protein), these proteins are important treponemal antigens and/or cell envelope structures [25–27]. Another gene potentially involved in cell wall biosynthesis, the Cuniculi A polysaccharide biosynthesis capD gene (TPCCA_0077), contained an internal stop codon predicted to result in gene inactivation. The Cuniculi A transmembrane chemoreceptors (Mcp proteins) either showed major sequence changes (TPCCA_0040 and TPCCA_0488, Table S4) or showed a relatively high number of amino acid replacements (TPCCA_0639, TPCCA_0640; 14 and 10 aa changes, respectively; data not shown).

Several Cuniculi A genes with frameshifts belonged to the family of genes with predicted regulatory functions including TPCCA_0220 (anti-sigma factor antagonist), TPCCA_0461 (probable transcriptional regulator), TPCCA_0511 (CarD family transcriptional regulator), and TPCCA_0520 (sensor histidine kinase). In contrast to TP0520 encoded in the genome of T. pallidum ssp. pallidum (strains Nichols and SS14) and in the T. pallidum ssp. pertenue (strains Samoa D, CDC-2, Gauthier; M. Zobaniková, D. Čeková, personal communication), TPCCA_0520 of T. paraluiscuniculi encodes full length sensor histidine kinase suggesting possible role of this protein in gene regulation. In addition to the previously mentioned tprA and TPCCA_0520 genes, reversion of frameshift mutation in TPCCA_0812 (fadD), probable long-chain-fatty-acid-CoA ligase was also found in the Cuniculi A genome.

Four genes encoding proteins involved in DNA processing (RecB, RecQ, RecX, and Ssb) also had altered open reading frames in the Cuniculi A strain relative to the Nichols strain. The frameshift mutation in the recQ gene resulted in premature truncation of RecQ, whereas RecX and Ssb proteins were elongated at C-termini. RecB showed multiple amino acid changes. Other genes with major sequence changes involved gldD (TPCCA_0735, glutamate and proline biosynthesis), ushA (TPCCA_0104, bifunctional 5’-nucleotidase/UDP-sugar diphosphatase), TPCCA_0309 (probable polar amino acid ABC-transporter), and mglB (TPCCA_0545, sugar ABC-transporter).

For 11 out of these 32 Cuniculi A genes, selection type was predicted (Table 3). While majority of genes were under neutral (6) or purifying selection (3), two genes including tprL and arp were found under positive selection.

Taken together, genes encoding several potential virulence determinants, including Tpr proteins, other membrane proteins, cell wall biosynthesis proteins, gene regulatory proteins and components of the DNA repair process, are significantly altered in the T. paraluiscuniculi Cuniculi A genome relative to the T. pallidum genomes sequenced to date.

### Discussion

The complete genome sequence of T. paraluiscuniculi Cuniculi A was determined by combining the data obtained by Illumina and Sanger sequencing and microarray hybridization approaches. The 454 sequencing data were used as a scaffold for the assembly and the final genome sequence was verified by genomic fingerprinting. This analysis of the Cuniculi A genome revealed striking similarity with other sequenced treponemal genomes (99.16% sequence identity between the conserved regions of the Nichols and Cuniculi A genomes), including identical gene orders, despite the differences in the host specificity and clinical manifestations of infections caused by T. paraluiscuniculi and T. pallidum ssp. pallidum. The Cuniculi A genome size (1,133,390 bp) is about 4.6 and 6.1 kb smaller than the genome size of the previously published T. p. ssp. pallidum Nichols and SS14 genomes, respectively. Additionally, an insertion harboring a tprA-like sequence (1.3 kb) in the intergenic region between TP0126 and TP0127 is present in a subpopulation of the Nichols strain [28] resulting in 5.9 kb difference between Cuniculi A and Nichols genomes. In unpublished studies, the genome of T. pallidum subspecies pertenue Samoa D (1,139,330 bp) was also found to be ~6.0 kb larger than the Cuniculi A genome (D. Čeková, personal communication). As shown by Strouhal et al. [2], the smaller Cuniculi A genome is a result of deletions localized mainly around tpr genes. In addition to decreased genome size, the Cuniculi A genome contained markedly higher number of pseudogenes and gene fragments. In the Nichols genome, only 9 genes containing authentic frameshifts were identified [18]. Besides these 9 genes with authentic
Table 3. *T. paraluiscuniculi* Cuniculi A genes with predicted cell function containing internal frameshifts and/or MSC* compared to the Nichols orthologs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene name</th>
<th>Gene/protein function</th>
<th>Type of change</th>
<th>Functional group/cell function</th>
<th>Ka/Ks ratio (if applicable) and estimation of selection type</th>
<th>Remark/reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPCCA_0009</td>
<td>tprA</td>
<td>Tpr protein</td>
<td>reverted frameshift mutation, MSC</td>
<td>potential virulence factors</td>
<td></td>
<td>[16] authentic frameshift mutation in the Nichols genome</td>
</tr>
<tr>
<td>TPCCA_0117</td>
<td>tprC</td>
<td>Tpr protein</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPCCA_0131</td>
<td>tprD</td>
<td>Tpr protein</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0313</td>
<td>tprE</td>
<td>Tpr protein</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0316</td>
<td>tprF</td>
<td>Tpr protein</td>
<td>partial gene deletion</td>
<td>1.19, neutral selection</td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0317</td>
<td>tprG</td>
<td>Tpr protein</td>
<td>frameshift mutation, partial gene deletion</td>
<td></td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0620</td>
<td>tprI</td>
<td>Tpr protein</td>
<td>gene deleted</td>
<td></td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0621</td>
<td>tprJ</td>
<td>Tpr protein</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0897</td>
<td>tprK</td>
<td>Tpr protein</td>
<td>frameshift mutation, MSC</td>
<td></td>
<td></td>
<td>[16]</td>
</tr>
<tr>
<td>TPCCA_1031</td>
<td>tprL</td>
<td>Tpr protein</td>
<td>MSC</td>
<td>1.64, positive selection</td>
<td></td>
<td>[2,16]</td>
</tr>
<tr>
<td>TPCCA_0136</td>
<td></td>
<td>fibronectin binding protein</td>
<td>MSC</td>
<td>virulence factors</td>
<td></td>
<td>1.05, neutral selection, lipoprotein, [26]</td>
</tr>
<tr>
<td>TPCCA_0326</td>
<td></td>
<td>tp92</td>
<td>outer membrane protein</td>
<td>MSC, gene deletions*</td>
<td>0.57, purifying selection</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0433</td>
<td></td>
<td>arp</td>
<td>treponemal conserved hypothetical protein</td>
<td>MSC, gene insertions</td>
<td>7.37, positive selection</td>
<td>[2]</td>
</tr>
<tr>
<td>TPCCA_0077</td>
<td></td>
<td>capD</td>
<td>capsular polysaccharide biosynthesis protein</td>
<td>internal stop codon</td>
<td>cell structure</td>
<td>0.17, neutral selection</td>
</tr>
<tr>
<td>TPCCA_0040</td>
<td></td>
<td>mcp1</td>
<td>probable methyl-accepting chemotaxis protein</td>
<td>frameshift, MSC</td>
<td>cell processes</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0488</td>
<td></td>
<td>mcp2</td>
<td>methyl-accepting chemotaxis protein</td>
<td>MSC</td>
<td>1.45, neutral selection</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0760</td>
<td></td>
<td>pbp2</td>
<td>penicillin-binding protein</td>
<td>internal stop codon</td>
<td>0.033, purifying selection</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0801</td>
<td></td>
<td>clpA2</td>
<td>S14 family endopeptidase ClpA</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPCCA_0936</td>
<td></td>
<td></td>
<td>probable hemolysin</td>
<td>internal stop codon</td>
<td>0.33, neutral selection</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0220</td>
<td></td>
<td></td>
<td>anti-sigma factor antagonist</td>
<td>frameshift mutation</td>
<td>regulation</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0461</td>
<td></td>
<td></td>
<td>probable transcriptional regulator</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPCCA_0511</td>
<td></td>
<td></td>
<td>CarD family transcriptional regulator</td>
<td>frameshift mutation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPCCA_0520</td>
<td></td>
<td></td>
<td>sensor histidine kinase</td>
<td>reverted frameshift mutation</td>
<td></td>
<td>authentic frameshift mutation in the Nichols genome</td>
</tr>
<tr>
<td>TPCCA_0103</td>
<td></td>
<td>recQ</td>
<td>ATP-dependent helicase</td>
<td>frameshift mutation</td>
<td>DNA replication, repair, recombination</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0310</td>
<td></td>
<td>ssb2</td>
<td>probable single-stranded DNA-binding protein</td>
<td>frameshift mutation</td>
<td></td>
<td>[2]</td>
</tr>
<tr>
<td>TPCCA_0898</td>
<td></td>
<td>recB</td>
<td>exodeoxyribonuclease V beta subunit</td>
<td>MSC</td>
<td>0.48, purifying selection</td>
<td></td>
</tr>
<tr>
<td>TPCCA_1023</td>
<td></td>
<td>recX</td>
<td>recombination regulator RecX</td>
<td>frameshift - missing stop codon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPCCA_0735</td>
<td></td>
<td>gldD</td>
<td>glutamate synthase (NADPH)</td>
<td>frameshift mutation</td>
<td>general metabolism</td>
<td></td>
</tr>
<tr>
<td>TPCCA_0812</td>
<td></td>
<td>fadD2</td>
<td>probable long-chain-fatty-acid–CoA ligase</td>
<td>frameshift mutation</td>
<td></td>
<td>authentic frameshift mutation in the Nichols genome</td>
</tr>
<tr>
<td>TPCCA_0104</td>
<td></td>
<td>ushA</td>
<td>bifunctional 5’-nucleotidase/UDP-sugar diphosphatase</td>
<td>frameshift mutation</td>
<td></td>
<td>lipoprotein</td>
</tr>
<tr>
<td>TPCCA_0309</td>
<td></td>
<td></td>
<td>probable polar amino acid ABC superfamily ATP binding cassette transporter, binding protein</td>
<td>frameshift mutation</td>
<td>transport</td>
<td>[2]</td>
</tr>
<tr>
<td>TPCCA_0545</td>
<td></td>
<td>mglB</td>
<td>sugar ABC superfamily ATP binding cassette transporter, binding protein</td>
<td>partial deletion at 5’ end – start codon missing</td>
<td>0.47, neutral selection</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Cont.

*Major sequence changes are defined in Table 2.
*Ka/Ks ratios were calculated by the MEGA4 software [52] and the selection test was calculated using the Kumar model [51].
*Corresponding protein identified as antigen [22].
*Deleted/changed serine rich regions mediates in some proteins attachment to cell surface [25].
*Fused genes (see Table S1).

doi:10.1371/journal.pone.0020415.t003

Figure 1. A schematic representation of tpr genes in the Cuniculi A and Nichols genomes. Identities at nucleotide levels of Cuniculi A and Nichols genomes are shown. Colors indicate sequence similarities among paralogous tpr genes, i.e. sequence similarities within the T. paraluiscuniculi genome (e.g. tprC and tprD genes are identical). In the Cuniculi A genome, reverted frameshift mutation (in tprA), frameshift mutations (in tprC,D,E,F,G,J,K), deletions (in tprF,G,I) and gene elongation are present (in tprL). The tprF deletion shown in the Nichols genome is based on the tprF sequence taken from T. pallidum ssp. pertenue Samoa D genome (data not shown). In the Cuniculi A and Nichols tprK genes, shorter gene versions (starting with the next available downstream start codon) are expected rather than the presence of frameshift mutation in the Cuniculi A tprK.

doi:10.1371/journal.pone.0020415.g001
frameshifts there are additional 6 pseudogenes in the SS14 genome (5 genes with frameshifts and one with nonsense mutation) [19]. In the genome of Samoa D, there are 13 pseudogenes (M. Zobaniková, personal communication). The number of pseudogenes and gene fragments [31] in the Cuniculi A genome thus markedly exceeds the number in Nichols, SS14 and Samoa D genomes, respectively.

In the Cuniculi A genome, there are 25 genes representing fusions of 32 Nichols orthologs. In most cases, Nichols orthologs are separated because of sequencing errors in the published whole genome Nichols sequence [18]. Therefore, most of these gene fusions do not represent true differences in the compared genomes. In contrast, missing genes, gene fragments, pseudogenes, and genes encoding proteins with many amino acid changes were considered important differences.

About 13% of the Cuniculi A genes were found to encode identical proteins to those encoded in the Nichols genome, indicating strong conservation of protein sequence in these proteins. Most of these genes encoded housekeeping proteins and likely represent a set of highly conserved treponemal genes important for basic treponemal cellular functions. The 35 genes encoding identical hypothetical proteins in both genomes are also candidates for important cellular functions. Indeed, the functions of three of these proteins were predicted recently: TP0950 as a protein involved in translation, TP0772 as a transcriptional regulator and TP0941 as a regulator of motility [29].

Comparison of number of genes containing frameshifts and/or major sequence changes (when compared to Nichols orthologs) within the Cuniculi A functional gene groups revealed a high percentage of these genes in the group of virulence factors, in genes involved in DNA metabolism, and in a group of genes with unknown function. These findings suggest that the accumulation of changes in genes encoding predicted virulence factors and genes with unknown functions (some of them are potential candidates for virulence factors, see below) is the reason for the loss of T. paraluiscuniculi infectivity to humans. Moreover, affected genes involved in DNA replication, repair, and recombination could suggest their possible role in the acceleration of T. paraluiscuniculi evolution.

Fifty-two hypothetical genes with internal frameshifts and/or major sequence changes were found in the T. paraluiscuniculi genome (Table 2 and S4). The data available from the previously published transcriptome analysis that mapped the Nichols strain gene expression levels during experimental rabbit infection [21] clearly showed that these genes are actively transcribed during infection. Moreover, the median transcript levels of these genes was considerably higher than the median gene expression rate of all genes of unknown function (1.46 versus 0.86) indicating that these genes are likely to represent true genes in the Nichols genome playing an important role during infection. These hypothetical genes are therefore candidates for important virulence factors of T. pallidum and should attract interest in future syphils research. In T. paraluiscuniculi, most of the genes (where the type of selection could be calculated) belonging to this group were found under neutral selection suggesting a genetic inactivity of these genes. The one positively selected gene, TP0031, could represent gene involved in adaptation of T. paraluiscuniculi to rabbits. In addition, 12 of hypothetical genes with internal frameshifts and/or major sequence changes were predicted to be inner or outer membrane proteins. T. pallidum whole-genome antigen screen, which tested 882 gene products [22], identified 106 antigens recognized by rabbit antibodies obtained from infected rabbits. Out of these 106 identified antigens, 9 antigens correspond to the group of 52 Cuniculi A hypothetical proteins that have frameshifts or major sequence changes. Interestingly, the group of predicted membrane and outer membrane proteins and the group of identified antigens did not overlap, probably as a result of problematic recombinant production of membrane proteins in E. coli. In addition, 22 hypothetical proteins were identified to interact with proteins of known function including predominately cell wall structures and antigens, regulatory and metabolic proteins [29]. Altogether, more than half of these proteins are likely involved in cell wall structure, and several others might be involved in gene regulation.

The most affected group of genes in the Cuniculi A genome was the family of paralogous tpr genes. Although the precise role of individual tpr genes in the treponemal infection remains unclear, there is an expanding evidence of the role of tpr genes in treponemal pathogenicity and host specificity. In the genome of T. pallidum subsp. pallidum Nichols, there are 12 paralogous tpr genes [18]. It was shown that Tpr proteins induced an antibody response during infection, and exhibit heterogeneity both within and between the T. pallidum subspecies and strains examined [30–32]. Tpr proteins are thought to be involved in pathogenesis and/or immune evasion. A model of gene conversion-driven antigenic variation of TprK during experimental infection was proposed [33]. Differences between Treponema species and subspecies in tpr gene content and expression are thought to be important determinants of pathogenesis and immunogenicity [16,17]. Out of 12 tpr genes in the Cuniculi A genome (Fig. 1), only four are intact: tprA, tprB, tprH and tprL. The remaining 8 genes contain frameshifts and/or deletions. Interestingly, the tprA gene in the Nichols genome contains an authentic frameshift resulting in inactive tprA gene. This also applies for the SS14 genome [19]. In contrast, tprA in the Cuniculi A (and also in the Samoa D genome; D. Čejková, personal communication) did not contain this frameshift mutation and appears to be functional. In addition to a regular copy of tprK, T. pallidum contains a tprK-like sequence localized in the 1.3 kb insertion present in a part of treponemal population [28]. A similar situation also applies for T. p. pertenue genomes, where this site also showed intraspecies heterogeneity (data not shown). The region containing the tprK-like sequence was not found in the T. paraluiscuniculi [2]. The major differences observed in tpr genes between T. paraluiscuniculi and T. pallidum suggest their role in the host range specificity. In fact, only tprA, tprB, tprH, tprL genes appear to be functional in the Cuniculi A genome. In contrast, tprC, tprD, tprE, tprF, tprG, tprI, tprJ, tprK were affected in the Cuniculi A genome. The tprE, tprG, tprJ genes were shown to be variably expressed in individual T. pallidum clinical isolates with guanosine homopolymers in promoter regions modulating their gene expression [34]. Interestingly, six tpr genes including tprC, tprD, tprE, tprF, tprI, tprJ out of 8 affected were recently predicted as genes encoding rare outer membrane proteins (OMPs) in T. pallidum [35]. These findings suggest the role of Tpr rare OMPs in T. pallidum infectivity to humans. Out of 4 functional tprA, tprB, tprH, tprL genes, the tprBL were predicted to encode rare OMPs (T. paraluiscuniculi TprL had a predicted signal sequence), suggesting that OMPs may be important also during T. paraluiscuniculi infection of rabbits.

There appear to be important differences between T. paraluiscuniculi Cuniculi A and other treponemal strains with regard to DNA recombination genes. In the Cuniculi A strain, the mutation in recQ resulted in a predicted RecQ protein without C-terminal, HDCR domain [36]. Although the precise role of this domain remains unclear, the C-terminal RecQ domain binds DNA [37]. The predicted sequence of RecB (important in the RecBCD pathway) has a high number of amino acid changes, and the predicted RecX and Ssp proteins have extended C-terminal sequences. There are also differences in the location of
recX in Cuniculi A relative to other bacterial species and treponemal strains. In E. coli, the RecX protein was shown to inhibit some RecA-mediated functions [38]. In contrast, in an exclusive human pathogen, N. gonorrhoeae, RecX enhances RecA activity [39]. Similarly to situation in N. gonorrhoeae, recX in Cuniculi A is not located downstream of (and overlapping with) recA but is in a different genome locus; this difference indicates that RecX expression and function may be more similar to those in N. gonorrhoeae. Taken together, genetic changes in the recB, recQ, and recX genes are consistent with observed increased genetic diversity in the Cuniculi A genome due to an ineffective/modified DNA repair and homologous recombination pathways. Homologous recombination as a major DNA repair process occurs frequently in the bacterial genomes and results in gene conversion. Gene conversion is an important mechanism of evolution of paralogous genes [40], and may be affected in the Cuniculi A genome.

We also examined other genetic differences to assess their potential role in pathogenesis patterns in T. paraluiscuniculi. In contrast to TPASS_0520 in T. p. ssp. pallidum and several other treponemes, TPCCA_0520 of T. paraluiscuniculi encodes full length sensor histidine kinase suggesting possible role of this protein in gene regulation. tprA and TPCCA_0520 genes are two out of four examples of intact genes present in the Cuniculi A genome but containing frameshift mutations in the ssp. pallidum. Several Cuniculi A membrane chemoreceptor proteins (Mcp’s) contained major sequence changes (TPCCA_0040, TPCCA_0498, TPCCA_0639, TPCCA_0640) when compared to their Nichols counterparts. Chemotactic proteins thus seem to be one of the most divergent proteins in the Cuniculi A genome, which may correlate with altered chemotaxis signaling patterns.

Several bacterial envelope components contained major sequence changes. Arp protein is of unknown function but it contains repeated predicted fibronectin-binding immunogenic domains [27]. Moreover, diverse repeats were shown to be associated with sexual transmission route of treponemal pathogens [41]. Interestingly, the atpA gene was found under a strong positive selection in the Cuniculi A genome, supporting its potential role in rabbit infection. TPCCA_0136, fibronectin- and laminin-binding protein is an outer membrane protein showing both inter- and intra-subspecies variable sequences. Immunization with recombinant protein delayed ulceration but did not prevent infection or the formation of lesions [26]. The immunization with recombinant Tp92 partially protected rabbits from subsequent T. pallidum challenge [25]. Sequence changes in genes encoding important antigens are one of the most probable reasons for changes in pathogenicity and host specificity. In Rickettsia prowazekii, the capD gene codes for an epimerase involved in capsular polysaccharide biosynthesis. The Cuniculi A capD was inactive. Since exopolysaccharides are important bacterial virulence factors, capD mutation in the Cuniculi A may be one of the reasons for decreased virulence of this strain. Beside these genes, altered genes with predicted regulatory functions (TPCCA_0220, TPCCA_0461, and TPCCA_0511) suggested differences in the gene regulatory network in the Cuniculi A genome. Changed regulatory network in the Cuniculi A genome and the resulting down- or upregulation of individual genes could be added to potential reasons for observed decreased virulence of Cuniculi A strain.

Taken together, the decreased size of the genome, marked increase in number of pseudogenes, affected genes involved in cell envelope biosynthesis and structure and multiple genetic changes in the proteins involved in DNA recombination, cell signaling and gene regulation appear to be the major reasons for narrower host specificity. Downsizing of the genome and accumulation of pseudogenes is common for bacteria adapting to simpler host-associated niches [42]. The loss of infectivity of T. paraluiscuniculi to humans may represent such a process. On the other hand, adaptation of T. paraluiscuniculi to rabbits resulting in more efficient infection of this host could be a result of additional changes that may include positively selected tprL, arp, and TP0031 genes and/or a number of dispersed mutations throughout the T. paraluiscuniculi genome. T. paraluiscuniculi thus appears to be treponeme in the process of adaptation to a single host (rabbit) and therefore is likely to be a descendant of pallidum- or pertenue-like ancestors rather than the opposite.

Materials and Methods

Isolation of T. paraluiscuniculi chromosomal DNA

T. paraluiscuniculi strain Cuniculi A was initially isolated from an infected rabbit by Drs. Paul Hardy and Ellen Nell; the strain was kindly provided by Dr. Sheila A. Lukehart at the University of Washington. Organisms were propagated by intratesticular inoculation of rabbits, extracted, and purified by Hypaque gradient centrifugation as described previously [18,43]. Genomic DNA was prepared according to the protocol published earlier [18].

DNA sequencing

The Cuniculi A genomic DNA (2.1 μg) was used for sequencing-by-synthesis (based on pyrosequencing) using GS20 sequencing machine (454 Life Sciences Corporation, Branford, CT, USA). Sequencing resulted in 398 individual trimmed contigs with a total contig size of 1,133,704 bp (average contig length of 2848 bp, contig length ranging from 102 to 22,217 bp). A subset of 330 individual contigs showed hits to published Nichols treponemal DNA [18], covering 1,128,602 bp and leaving 4,788 bp unsequenced. The number of individual reads in these 330 contigs was 204,765 representing total read length of 20,443,023 bp. The corresponding sequencing coverage for the Cuniculi A genome was 18.04. Most of contigs not related to treponemes showed similarity to rabbit sequences, probably as a result of contamination of Cuniculi A DNA by rabbit DNA during preparation of chromosomal DNA.

Parallel to pyrosequencing, an Illumina (Illumina, San Diego, CA, USA) sequencing approach was employed using the Genome Analyzer sequencing machine. Illumina reads (36 bp each) were assembled into 726 contigs using Velvet short read assembler [44]. Total number of reads (3,053,564) represented total read length of 109,928,304 bp (97x coverage). Some of the 726 contigs overlap by few bp and therefore the number of gaps dropped to the number 475, representing total gap length of 33,634 bp.

Sanger sequencing of Cuniculi A sequences was used to assess the quality of DNA sequencing and to finish the whole genome sequence. Approximately 150 PCR products, generated with primers used for other treponemal genome projects were sequenced to provide a comparison to the Illumina- and 454-generated sequences. All 50 discrepancies were in the 454 sequence results, and included 43 false insertions, 3 false deletions and 4 substitutions. Twenty-four out of 46 indels were found in homopolymeric regions. Based on these results, we considered the Illumina sequences to be more accurate and utilized them for generation of the complete genome sequence.

An additional approach, the CGS strategy [19] was used for determination of the Cuniculi A genome sequence. Oligonucleotide arrays of 29-mers derived from the Nichols sequence,
covering both strands, were hybridized separately to fluorescently-labeled Nichols and Cuniculi A genomic DNA. Equal hybridization signals in both preparations of genomic DNA indicated identical sequences, whereas decreased hybridization with the Cuniculi A DNA occurred in regions with sequence differences or indels. This information was used to help ‘fill in’ the sequence gaps between Illumina contigs. All discrepancies in gap regions were resolved by traditional Sanger sequencing.

Whole genome fingerprinting

Whole genome fingerprinting [45] results for the Cuniculi A genome [2] were used for verification of the genome assembly. Briefly, primers designed for *T. pallidum* subsp. *pallidum* and template DNA from *T. paraluiscuniculi* Cuniculi A were used to produce large, 5 to 28 kb amplicons spanning the entire genome; these were then digested with multiple enzymes to provide a macro restriction map. The Cuniculi A genomic sequence was used for simulated restriction digest *in silico* and these data were compared with experimentally obtained data. Altogether, 19 individual restriction enzymes were used including *Ace I* (194 verified restriction target sites), *Ace II*, *BamHI I* (222), *Cla I* (107), *EcoRI I* (157), *EcoRV V* (200), *Hind III* (250), *Kpn I* (112), *Mlu I* (277), *Mse I* (8), *Nco I* (61), *Nde I* (1), *Nhe I* (14), *Rsa II* (20), *Sac I* (96), *Sph I* (25), *Sso I* (15), *Xho I* (68) or *Xho I* (191) enzymes (NEB) either alone or in combinations. Three enzymes, *BamHI I*, *EcoRI I* and *Hind III*, were used for restriction analysis of all amplicons. The use of other enzymes was optional depending on length of restriction fragments and the availability of restriction target sites. To ascertain the experimental error of WGF, the lengths of 250 individual DNA fragments in 5 fragment intervals (50 fragments per interval) including 0.2–0.5 kb, 0.5–1 kb, 1–2 kb, 2–3 kb, and 3–4 kb, respectively, were measured from agarose gels by AlphaView Software Version 3.0 (Alpha Innotech, San Leandro, CA) and calculated from *in silico* data. The average error for each interval represented an average difference between experimental and calculated fragments sizes. The average errors were calculated to 10.9 bp, 16.8 bp, 22.3 bp, 39.9 bp and 52.5 bp in 0.2–0.5 kb, 0.5–1 kb, 1–2 kb, 2–3 kb, and 3–4 kb intervals, respectively. A set of the smallest available 722 fragments (covering the entire Cuniculi A genome), with length ranging between 0.2 and 4.0 kb (average length of 1712 bp) and covering both strands, were hybridized separately to fluorescently-labelled DNA restriction fragments and the availability of restriction target sites.

**References**

3. Jacobsoth E (1929) Untersuchungen über eine syphilisa¨hnliche Spontanerkran-

**DNA sequence analyses and statistical analyses**

DNA sequence analyses were performed using the *DnaSP* software, version 5.10 [50]. Whole genome nucleotide diversity (θ) between Cuniculi A and individual sequenced *T. pallidum* genomes including Nichols, Chicago and SS14 strains was calculated, respectively. The number of synonymous substitutions per a synonymous site (Ks), the number of nonsynonymous substitutions per a nonsynonymous site (Ka), the Ka/Ks ratios, and the codon-based test for estimation of selection type were calculated using the Kumar model [51] and the MEGA4 software [52]. Statistical significance of the number of genes containing frameshifts and/or major sequence changes (MSC) in *T. paraluiscuniculi* functional gene groups was calculated using standard methods derived from the binomial distribution, including the two-tailed test. For statistical calculations, STATISTICA program, version 8.0, (StatSoft, Tulsa, OK, USA) was used.

**Nucleotide sequence accession numbers**

The nucleotide sequences reported in this study were deposited in the GenBank under the accession number CP002103.

**Supporting Information**

Table S1  Genes fused in the *T. paraluiscuniculi* Cuniculi A genome when compared to the previously annotated *T. pallidum* subsp. *pallidum* Nichols genome [18].

Table S2  99 genes encoding identical proteins in *T. paraluiscuniculi* Cuniculi A and *T. pallidum* subsp. *pallidum* Nichols genomes.

Table S3  35 genes of unknown function encoding identical proteins in *T. paraluiscuniculi* Cuniculi A and *T. pallidum* subsp. *pallidum* Nichols genomes.

Table S4  *T. paraluiscuniculi* Cuniculi A genes with unknown cell function containing internal framshifts and/or major sequence changes (MSC) compared to the Nichols orthologs.

**Acknowledgments**

We like to acknowledge Donna Villasana, Zhangwan Li and Jing Liu for technical assistance in DNA sequencing.

**Author Contributions**

Conceived and designed the experiments: DS GMW. Performed the experiments: DS MZ MS DC PP SD-R. Analyzed the data: TA DS MZ MS DC PP SD-R. Contributed reagents/materials/analysis tools: DS GMW RAG SJN TA. Wrote the paper: DS SJN.
49. Sur unfinished.