Genome analysis of *Moraxella catarrhalis* strain RH4: a human respiratory tract pathogen

Running title: *M. catarrhalis* RH4 genome sequence

Stefan P.W. de Vries[1], Sacha A.F.T. van Hijum[2,3], Wolfgang Schueler[4], Kristian Riesbeck[5], John P. Hays[6], Peter W.M Hermans[1*], Hester J. Bootsma[1]

[1]Laboratory of Pediatric Infectious Diseases, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, [2]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, [3]NIZO Food Research, Ede, The Netherlands, [4]Intercell AG, Vienna Biocenter 3, Vienna, Austria, [5]Medical Microbiology, Department of Laboratory Medicine, University Hospital Malmö, Lund University, Malmö, Sweden, [6]Department of Medical Microbiology and Infectious Diseases, Erasmus MC, Rotterdam, The Netherlands

*Corresponding author. Mailing address: Laboratory of Pediatric Infectious Diseases, Radboud University Nijmegen Medical Centre, P.O. Box 9101 (Route 224), 6500 HB Nijmegen, The Netherlands. Phone: 31-24-3666406. Fax: 31-24-3666352. E-mail: P.Hermans@cukz.umcn.nl

ABSTRACT

Moraxella catarrhalis is an emerging human-restricted respiratory tract pathogen that is a common cause of childhood otitis media and exacerbations of chronic obstructive pulmonary disease in adults. Here, we report the first completely assembled and annotated genome sequence of an isolate of *M. catarrhalis*: strain RH4, originally isolated from blood of an infected patient. The RH4 genome consists of 1,863,286 nucleotides harboring 1,886 protein-encoding genes. Comparison of the RH4 genome to the ATCC 43617 contigs demonstrated that the gene content of both strains is highly conserved. In silico phylogenetic analyses based on both 16S rRNA and multilocus sequence typing revealed that RH4 belongs to the seroresistant lineage. We were able to identify close to the entire repertoire of known *M. catarrhalis* virulence factors, and mapped the members of the biosynthetic pathways for lipooligosaccharide, peptidoglycan, and type IV pili. A reconstruction of the central metabolic pathways suggests that RH4 relies on fatty acid and acetate metabolism, as the genes encoding the enzymes required for the glyoxylate pathway, tricarboxylic acid cycle, gluconeogenic pathway, non-oxidative branch of the pentose phosphate pathway, beta-oxidation pathway of fatty acids, and acetate metabolism were present. Moreover, pathways important for survival under *in vivo* challenging conditions such as iron-acquisition pathways, nitrogen metabolism, and oxidative stress responses were identified. Finally, we showed by microarray expression profiling that ~88% of the predicted coding sequences are transcribed under in vitro conditions. Overall, these results provide a foundation for future research into the mechanisms of *M. catarrhalis* pathogenesis and vaccine development.

INTRODUCTION

Moraxella catarrhalis is an emerging human-restricted unencapsulated gram-negative mucosal pathogen. For long considered to be a commensal of the upper respiratory tract, the bacterium has now firmly established its position as an etiological cause of otitis media (OM) and exacerbations of chronic obstructive pulmonary disease (COPD). It is the third most common cause of childhood OM after *Haemophilus influenzae* and *Streptococcus pneumoniae* (37,64), being responsible for up to 20% of the cases (64,65). Further, *M. catarrhalis* is the second most common cause of exacerbations of COPD after *H. influenzae*, being responsible for 10 to 15% of the exacerbations, annually accounting for 2 to 4 million episodes in the United States (47,60). Antibiotics are widely used for the treatment of OM, but the high prevalence of this disease and increasing incidence of antibiotic-resistant strains requires development of multivalent vaccines, preferably with protective antigens for all three causative bacterial agents (46).

M. catarrhalis is able to colonize the mucosal surfaces of the middle ear in OM patients and the lower respiratory tract in COPD patients (31,60). Successful colonization of its human host is a complex process which requires the expression of adhesins and the activation of metabolic pathways to overcome specific environmental challenging conditions, such as nutrient limitation (15,53). *M. catarrhalis* also possesses several mechanisms for evasion of the host immune system (15,53), such as the ability to withstand the action of the human complement system. Importantly, most clinical isolates from OM or COPD patients are able to survive complement-mediated killing by normal human serum (66).

Various molecular typing methods, such as 16S rRNA sequencing (8) and multilocus sequence typing (MLST) (71) have shown that the species *M. catarrhalis* can be divided into two distinct phylogenetic lineages, referred to as the serosensitive and seroresistant lineage. The seroresistant, more virulent lineage predominantly contains strains that are resistant to

complement-mediated killing, and is enriched for strains with the ability to adhere to epithelial cells (8,71).

Even though our understanding of the molecular pathogenesis of *M. catarrhalis* has increased over the past years, a complete *M. catarrhalis* genome sequence is undoubtedly expected to serve as a valuable resource to improve our understanding of its biology. At present, however, only the partial genome of *M. catarrhalis* strain ATCC 43617 is available in the scientific literature (59,68). Here, we report the first completely assembled and annotated genome sequence of *M. catarrhalis* strain RH4, which was originally isolated from the blood of an infected patient (14), and compared its gene content to the ATCC 43617 strain. In silico phylogenetic typing revealed that the RH4 strain belongs to the seroresistant lineage. We were able to identify most of the known virulence factors and reconstructed several important biochemical pathways including central metabolic pathways, nitrate metabolism, biosynthesis of lipooligosaccharides (LOS), peptidoglycan, type IV pilus (TFP) biosynthesis, and iron acquisition. In addition, several components were identified that combat oxidative stress. Finally, we confirmed by transcriptional profiling that most of the predicted coding sequences are expressed in vitro.

MATERIALS AND METHODS

Bacterial strain and growth conditions. *M. catarrhalis* strain RH4 was isolated from the blood of an infected patient (14). RH4 was grown on Bacto-brain heart Infusion (Bacto-BHI) agar plates (Difco) in an atmosphere containing 5% CO_2 at 37°C or in broth at 37°C with agitation (200 rpm).

Sequencing of the *M. catarrhalis* RH4 genome. Whole genome sequencing was performed with a hybrid strategy using Roche 454 and Illumina Solexa sequencing by Agowa Genomics (Berlin, Germany). For Roche 454 sequencing, genomic DNA was extracted using the Wizard Genomic Purification kit (Promega), after which a fragment library was prepared and sequenced according to Roche standard procedures. This resulted in a total of 591,043 sequences with an average read length of 224 bp covering approximately 134 Mb, which is more than 70-fold coverage of the total genome. *De novo* assembly using the Roche 454 software Newbler resulted in 44 contigs of over 500 bp, which were aligned with the 41 contigs of the ATCC 43617 strain (deposited at NCBI under patent WO0078968, GenBank accession numbers AX067426 to AX067466) using the gap4 assembler program (Staden Package, Roger Staden, Cambridge). Gap closure by PCR and Sanger sequencing resulted in a contiguous sequence of 1,863,286 bp, which was verified using Solexa sequencing as follows. Genomic DNA was isolated with the Genomic DNA kit (Qiagen), a fragment library (150-200 bp) was prepared according to Illumina's standard genomic DNA library preparation procedure, and the library was sequenced with the Illumina Genome Analyzer II. The data was analyzed on the standard Genome analyzer pipeline, yielding a total of 7,057,256 raw 36-bp reads. The 454-assembled genome was corrected by Solexa short read sequence data using ROAST, an in-house developed tool (van Hijum *et al.*, unpublished). Briefly, Solexa reads were aligned to the assembled genome sequence by BLAT (38). Alignment events of reads to the reference were allowed provided that nucleotide

99 substitutions (single nucleotide polymorphisms; SNPs) or gaps (small insertions or deletions;
100 INDELs) were at least four bp from the end of reads. Only SNPs/INDELs were allowed with a
101 sequence depth of at least six reads unanimously calling a genotype, and with a maximum of one
102 read indicating a different genotype. In total, three SNPs, one insertion and one deletion were
103 corrected in the reference genome sequence.

104 **Annotation.** Open reading frames (ORFs) and initial automated annotation were obtained
105 from both Pedant-Pro and the Institute for Genomics Sciences (IGS) annotation engine
106 (http://ae.igs.umaryland.edu/cgi/ae_pipeline_outline.cgi), with a minimal ORF length of 90
107 nucleotides (nt). The IGS-derived annotation was subjected to manual curation using the Pedant-
108 Pro data, DBGET database (http://www.genome.jp/dbget/), and literature. The putative origin of
109 replication was identified using Ori-Finder (27), and base pair 1 was assigned at the extreme of
110 the CG disparity curve. Functional classification of the predicted ORFs was based on the IGS
111 functional classification, manually improved with data from Clusters of Orthologous Groups
112 (COGs) of proteins and Non-supervised Orthologous Groups (NOGs), both obtained by searches
113 using Signature (18).

114 Comparative genomics of the RH4 and ATCC 43617 strains was performed by alignment of
115 the RH4 coding sequences (CDSs) to the ATCC 43617 contigs using BLAT (38). CDSs
116 potentially unique for ATCC 43617 were identified by alignment of the ATCC contigs to the
117 RH4 sequence.

118 Pathway analysis was performed using the KEGG Pathway database (36) using KEGG
119 Orthology (KO)-identifiers and was complemented by literature-based pathway reconstructions.
120 KO-identifiers were assigned with the web-based KEGG Automatic Annotation Server (KAAS,
121 http://www.genome.jp/kegg/kaas/) using the bi-directional best hit methods against a set
122 prokaryotic reference genomes (organism codes: prw, par, pcr, abm, aby, aci, ngo, nmc, hso,

123 nma, acb, apl, cvi, hin, hit, hip, and pmu), manually selected based on high abundance in Pedant-
124 Pro BlastP-hits. The acquired list was complemented with KO-identifiers assigned by searching
125 against the prokaryotic reference set. Finally, the RH4 genome was analyzed for the presence of
126 CRISPR repeat regions using CRISPRFinder (29).

127 The subcellular localization of the RH4 proteins was predicted through an extended version of
128 the LocateP software (75) and validated with a highly curated list of *Escherichia coli* proteins.
129 The protein localization prediction was tailored for the Gram-negative *M. catarrhalis* by
130 replacement of specific Gram-positive components with tools suitable for Gram-negative bacteria
131 (Zhou *et al.*,in preparation for publication), such as BOMP (7), CELLO (73), LipoP (35), and
132 SecretomeP (6).

133 Detailed annotation including localization prediction and comparative genomics can be found
134 in supplemental file 1.

135 **In silico phylogenetic typing.** The determination of 16S rRNA type was performed according
136 to Bootsma *et al.* (8). Multilocus sequence typing (MLST) of sequence fragments of eight
137 housekeeping genes was performed according to Wirth *et al.* (71). Allelic sequences were
138 analyzed at http://mlst.ucc.ie/mlst/dbs/Mcatarrhalis and compared to the reference database
139 containing 282 *M. catarrhalis* strains. Lipooligosaccharide (LOS) serotype determination was
140 accomplished by using the method described by Edwards *et al.* (19), and *copB*-typing was
141 performed according to Verhaegh *et al.* (66).

142 **Microarray expression profiling.** Bacteria were grown in BHI medium at 37°C with
143 agitation (200 rpm), harvested by centrifugation at lag phase (OD_{620} = 0.2-0.3), exponential
144 phase (OD_{620} = 1.2-1.4), and stationary phase (OD_{620} = 2.0-2.2), and treated with 2 volumes of
145 RNAprotect bacteria reagent (Qiagen). Total RNA was extracted using the RNeasy Mini kit
146 (Qiagen) after which contaminating genomic DNA was removed by treatment with DNase

147 (DNAfree, Ambion). Total RNA was labeled essentially as described by Ouellet et al. (51).
148 Briefly, 10 µg of RNA was incubated for three hours with 7 µg of 5'Cy3-labeled random
149 nonamers (TriLink Biotechnologies) and Superscript III reverse transcriptase (800 U, Invitrogen)
150 in appropriate reaction conditions (1x First-Strand buffer, 5 mM DTT, 0.33 mM dNTPs, 21 mM
151 actinomycin D (Sigma Aldrich), and 40 U RNaseOut). After first strand synthesis, RNA was
152 degraded by incubating with sodium hydroxide, followed by reaction neutralization with
153 hydrochloric acid. Labeled cDNA was purified and concentrated with CyScribe columns (GE
154 Healthcare Life Sciences) followed by Micron-30 columns (Millipore). 2 µg of labeled cDNA
155 was applied in duplicate to 4x72K custom design NimbleGen arrays. Overnight hybridization at
156 42°C and subsequent washing of arrays was performed according to the manufacturer's
157 instructions. The Nimblegen array contained 1-118 probes per predicted CDS with an average
158 coverage of 15 probes per CDS, probes on both strands within the CRISPR1 and CRISPR2
159 regions (no specific probes could be designed for the putative CRISPR3 region), and 1,103
160 negative control probes with a similar length distribution and CG content as the experimental
161 probes. Array images were acquired with a NimbleGen MS200 scanner, and images were
162 processed with NimbleScan software using the RMA algorithm. Normalized and background-
163 corrected probe signal intensities thus obtained were used to calculate the median expression
164 level of the CDSs. An expression threshold was defined by the median of the $\log_2$ signal intensity
165 of the random control probes plus four times the standard deviation. Expression levels for CDSs
166 were classified as follows: low (+, log2 <10), moderate (++, log2 = 10-12.5), high (+++, log2
167 >12.5), or no expression (4/6 replicates (biological triplicate and technical duplicate) <threshold).
168 Expression data of the RH4 predicted CDSs can be found in supplemental file S1.
169 **Nucleotide sequence accession number and microarray data.** The genomic sequence of *M.*
170 *catarrhalis* RH4 has been deposited in the GenBank database under accession number

171 CP002005. Microarray data have been deposited in the NCBI Gene Expression Omnibus (GEO)

172 database (www.ncbi.nlm.nih.gov/geo/) under GEO Series accession number GSE21632.

RESULTS AND DISCUSSION

General genome features and comparative genomics. The RH4 genome (Fig. 1) consists of 1,863,286 nucleotides (nt), with an overall GC-content of 41.7%. Both length and GC-content are comparable to the unassembled, partial ATCC 43617 genome (~1.9 Mb), which is represented in 41 contigs (59,68). The RH4 genome is predicted to encode 1,964 genes of which 1,886 are protein-encoding genes (Table 1), similar to the 1,761 to 1,849 ORFs predicted for ATCC 43617 (59,68). Of the 1,886 protein encoding genes predicted for RH4, 63.6% could be assigned to a functional category based on similarity to other proteins in public databases, while 172 predicted proteins are of unknown function. The remaining 515 ORFs are classified as either conserved hypothetical proteins (102) or hypothetical proteins (413) (Table 1). The majority (~62%) of the best BlastP hit proteins were found to belong to *Psychrobacter* species (data not shown).

Comparative genome analysis of the RH4 genome and the ATCC 43617 contigs revealed that 416 RH4 coding sequences (CDSs) had an exact match in the ATCC 43617 genome (Table 2). When allowing a maximum of 20% mismatches, 1,252 full length RH4 CDSs were identified in the ATCC 46317 contigs. In total, ~95% (1,793) of the predicted CDSs were found to be conserved between RH4 and ATCC 43617 with a minimum of 80% of the CDS covered. As expected, RH4 genes encoding highly variable virulence factors such as the ubiquitous surface protein A (*uspA*) family (10) showed lower levels of homology with ATCC 43617 sequences. Further, 68 CDSs were found to be unique for RH4, of which 41 are annotated as (conserved) hypothetical. Reversely, using the preliminary annotation of the partial ATCC 43617 genome (68), 14 CDSs were found to be unique for ATCC 43617, among which 12 hypotheticals.

The RH4 genome contains 50 tRNAs representing all twenty amino acids. There are four identical rRNA operons (16S, 23S, and 5S genes) in which the 16S and 23S rRNA genes are interspersed with tRNAs for isoleucine and alanine (Table 1). Genes encoding the RNA

197 polymerase core subunits (α, β, β', ω), the sigma factors σ^{70}, and the alternative σ^{32} were
198 identified, as were the transcription elongation factor GreA, and the transcription
199 termination/anti-termination factors NusA, NusB, Rho, and NusG. Interestingly, the RH4 genome
200 is predicted to encode only 32 transcription factors, 4 two-component regulatory systems, and 2
201 orphan two-component system genes. This raises the possibility that *M. catarrhalis* has other,
202 alternative mechanisms to adapt its gene expression to changing environmental conditions, such
203 as phase-variable expression or non-coding RNA-based regulation. Phase-variable expression has
204 already been described for two *M. catarrhalis* virulence factors, the ubiquitous surface protein A1
205 (UspA1) (40) and *M. catarrhalis* immunoglobulin D binding protein (MID) (45), but a
206 preliminary search for additional homopolymeric tracts known to be involved in phase variation
207 did not lead to the discovery of novel candidates in the RH4 genome. Clearly, further studies are
208 required to elucidate the main mechanism of transcriptional regulation in *M. catarrhalis*.
209 In silico phylogenetic analysis classified RH4 as a 16S rRNA type 1 strain (Fig. S1A),
210 predominantly found in the seroresistant lineage (8). MLST-analysis assigned RH4 to sequence
211 type 128 (Fig. S1B), identical to *M. catarrhalis* GRJ 11, which was isolated from a diseased
212 individual in Salamanca (Spain) and also belongs to the seroresistant lineage (71). In line with
213 this, experimental evidence also demonstrated that RH4 displays a seroresistant phenotype (data
214 not shown).
215 **Expression profiling.** Microarray expression profiling during three phases of in vitro growth
216 in BHI medium (lag, exponential, and stationary) showed that 88.1% of the predicted CDSs were
217 expressed during at least one growth phase, with 84.6% being expressed in all three growth
218 phases. Of the predicted CDSs for which no transcripts were detected, ~81% was annotated as
219 (conserved) hypothetical. Further, we could demonstrate expression for 30 of the 41 (conserved)

hypotheticals CDSs that were not present in the ATCC 43617 strain. Expression of specific genes will be discussed in detail in the subsequent paragraphs.

Virulence factors. Various proteins have been described to play pivotal roles in *M. catarrhalis* pathogenesis (recently reviewed in (15,53)), all but one of which were found to be present in the RH4 genome (Table 3). The ubiquitous surface proteins (UspAs) are among the major virulence factors, with UspA1 mediating binding to epithelial cells and extracellular matrix (ECM) components, and the mutually exclusive UspA2/UspA2H proteins predominantly playing a role in immune evasion (3,39). Determination of the modular structure of the predicted UspA1 and UspA2H proteins (the strain did not possess a *uspA2* gene) revealed the presence of the VEEG-NINNY-VEEG amino acid sequence motif involved in binding to Chang conjunctival cells or fibronectin (10) in UspA1, whereas this motif was absent in UspA2H (Fig. 2). RH4 harbors several other adhesins, namely *M. catarrhalis* immunoglobulin D (IgD) binding protein (MID, also referred to as hemagglutinin; Hag) (22), the *M. catarrhalis* adherence protein (63) and OMP CD (33). Resistance to the action of the human complement system is an important aspect of *M. catarrhalis* virulence. Previous studies have shown that, in addition to the above-mentioned UspA proteins, the *M. catarrhalis* proteins CopB (32), OMP CD (33), and OMP E (48), play a role in serum-resistance, and their corresponding genes are all present in RH4. All virulence factors describe above were found to be expressed at intermediate or high levels during all three phases of in vitro growth sampled. Interestingly, the only known virulence locus absent from RH4 is the *mha* locus encoding filamentous hemagglutinin-like proteins involved in adhesion (4,54). We did identify three ORFs (MCR_0770, MCR_0777, and MCR_0778) harboring a small region of homology (37-71 amino acids) to *mhaB1*, but transcriptional profiling indicated either no (MCR_0778) or low levels (MCR_0770 and MCR_0777) of expression of these loci.

Biosynthesis of cell wall structures. A prominent surface component of *M. catarrhalis* that is generally considered to be an important virulence factor is lipooligosaccharide (LOS). Genes encoding LOS glycosyl transferases (Lgt), enzymes that catalyze the formation of core or branched oligosaccharide chains (43), as well as genes required for biogenesis of the deoxy-D-*manno*-2-octulosonic acid (KDO)-lipid A moiety (reviewed by Raetz *et al.* (55)), were all identified in the RH4 genome (Table S2) Expression was detected for all components of the involved pathways except for *lgt5*, which catalyzes the addition of the terminal α-(1→4)-linked terminal galactose of the core oligosaccharide chain (70) (Table S2). Analysis of the RH4 *lgt*-locus (*lgt5*, *lgt1*, *lgt2b/c*, and *lgt3*) revealed that RH4 is a LOS type B strain (Fig. S1C) (19), which is found exclusively in isolates of 16S rRNA type 1 (66).

The peptidoglycan layer is the main target for β-lactam antibiotics that can be degraded by the BRO β-lactamases produced by *M. catarrhalis* (RH4 expresses the *bro-2* gene; Table 3) (9). The complete set of genes required for the biosynthesis of peptidoglycan (61) were identified in RH4 and found to be expressed in all growth phases (Table S3).

Type IV pili (TFP) fulfill a wide variety of functions including adhesion to epithelial cells, biofilm formation, and motility (44). Biogenesis of TFP is a complex process requiring a large set of proteins (52), which are present in the RH4 genome (Table S4). Overall, low levels of gene expression were detected for most components of the TFP pathway, except for *pilA*, the major pilin subunit, which on average was found to be highly expressed (Table S4).

Protein secretion. Gram-negative bacteria transport proteins from the cytosol across the inner membrane (IM) to the periplasm via one of two protein secretion systems: Sec or Tat (16,41). RH4 was found to contain a complete Sec-machinery, as well as the main components of the Tat-system (Table S5). After their deposition into the periplasmic space, proteins reach their final destination by other means. For instance, outer membrane lipoproteins are inserted

into the outer membrane by the components of the Lol-system (49) (Table S5). Furthermore, several ORFs are predicted to encode components of resistance-nodulation-division (RND) efflux systems, such as components of the Acr and Mtr systems (50), both sharing homology to type I secretion system components (Table S5). Strikingly, no components of the general secretory pathway, a type II secretion system, were found, although the evolutionary-related TFP assembly machinery (62) is entirely present as mentioned above. Even though a homologue of the type III secretion effector (HopJ) was identified (MCR_0582), other components of type III secretion systems were not found, nor did we identify components of type IV, V, and VI secretion systems. An overview of protein secretion components and their detected expression levels is given in Table S5.

Protein localization prediction. A complete genome sequence provides a window of opportunity to discover novel vaccine targets, with surface-exposed and secreted proteins being of special interest in the context of vaccine development. Ruckdeschel *et al.* applied a genome mining approach to the partial genome sequence of *M. catarrhalis* ATCC 43617, which led to the discovery and animal model testing of novel vaccine targets (58,59). An extensive subcellular localization prediction revealed that 134 (7.1%) of the predicted RH4 proteins localize to the outer membrane or are secreted into the extracellular environment. These surface-exposed proteins include vaccine candidates such as the lipid-anchored outer membrane protein Msp22 (59) and OMP E (48). In addition, we identified 35 (1.9%) proteins that are predicted to localize to the cytoplasm, but are possibly secreted via non-classical secretion pathways (i.e. not via the Sec- or Tat-pathway), including for example the autotransporter McaP (42). The subcellular compartment distribution of the predicted RH4 proteome is summarized in Table 4.

Central metabolic pathways. As a respiratory tract pathogen, *M. catarrhalis* is, at least partially, dependent on nutrient availability inside its human host to fulfill its needs for energy

and intermediates for biosynthesis of essential macromolecules. *M. catarrhalis* is reported to be incapable of utilizing exogenous carbohydrates and consequently does not produce acid from carbohydrates (12). In line with this, no genes encoding carbohydrate transport or catabolism components were found in the RH4 genome. Reconstruction of its central metabolism (Fig. 3 and Table S6) showed that while RH4 possesses an incomplete glycolytic pathway, all the enzymes of the gluconeogenic pathway are present, indicating that carbohydrate intermediates can be synthesized. Gluconeogenesis uses phosphoenolpyruvate (PEP) as a starting substrate, which can be generated from tricarboxylic acid (TCA) cycle intermediates. The TCA cycle is supplied with acetyl-CoA via the degradation of fatty acids and acetate assimilation, for which the required genes were identified in RH4. The incomplete TCA cycle, missing both subunits of succinyl-CoA synthetase, can be bypassed by the glyoxylate pathway, which is completely present in RH4. The glyoxylate pathway fulfills an anaplerotic function by supplying the TCA cycle with oxaloacetate (acetyl-CoA acceptor molecule). Furthermore, the enzymes required for oxidative stages of the pentose phosphate pathway are missing, but transaldolase and the enzymes of the non-oxidative branch are present. As mentioned, RH4 possesses all genes required for beta-oxidation of long chain fatty acids (Table S6) (23). Long chain fatty acids are transported across the cell membrane by the combined action of the outer membrane protein FadL and the inner membrane-associated FadD protein, an acyl-CoA synthase. We predict that the highly conserved OMP E protein (48) serves as the FadL homolog in RH4. Two adjacent ORFs (67% identical) appear to encode homologs of FadD (aerobic) or its anaerobic counterpart FadK (13). Interestingly, we did not detect obvious homologs of known fatty acid metabolism regulators such as FadR, although we did identify family members of such proteins such as GntR and DeoR-family regulators (23). Unfortunately, no obvious function could be assigned to these ORFs. Further, we identified two ORFs with homology to the 3-oxoacid CoA-transferase involved in the degradation of short-

315 chain fatty acids (34). Finally, we identified all genes encoding the enzymes required for
316 completion of fatty acid biosynthesis (23) (Table S6).

317 Expression profiling showed that all genes of the gluconeogenic, TCA cycle, fatty acid
318 degradation, acetate assimilation, and pentose phosphate pathway are expressed during in vitro
319 growth (Table S6). However, expression of malate synthase, the second enzyme of the glyoxylate
320 pathway, was undetectable at lag and exponential phases of growth, and only at low levels during
321 the stationary growth phase. This suggests that glyoxylate is further processed via other enzymes,
322 e.g. glycerate dehydrogenase (MCR_1529) and phosphoglycolate phosphatase (MCR_0365),
323 which are both expressed at intermediate levels in all growth phases.

324 **Nitrogen metabolism.** A truncated denitrification pathway (reduction of nitrite to nitrous
325 oxide), described by Wang *et al*., has been suggested to provide *M. catarrhalis* with an
326 alternative mechanism to generate energy under low oxygen tension, and to contribute to biofilm
327 formation and *in vivo* virulence. In the diagnostic microbiological laboratory, the reduction of
328 nitrate is one of the differential tests used to confirm the identity of *M. catarrhalis* (12). The
329 genes encoding the nitrate reductase complex (*narGHJI*), nitrite reductase (*aniA*), the nitric oxide
330 reductase (*norB*), and the *narX*/*narL* two-component system were identified in RH4 (Table S7).
331 In addition, a putative regulator of fumarate and nitrate reduction was identified, however, the
332 true function of this FNR/CRP family member remains to be elucidated (69). In addition to the
333 nitrate ABC transport system, encoded by *nrtABCD*, we identified two other candidates that
334 could play a role in nitrate/nitrite transport, namely NarK1, and a putative nitrate/nitrite
335 transporter designated NarK2 (Table S7). Expression levels of the denitrification pathway
336 components was not uniform, e.g. nitrate reductase expression ranged from undetectable to low,
337 whereas nitrite reductase was expressed at high levels during lag and exponential phase (Table
338 S7).

Iron acquisition. Iron is a key nutrient that is functionally involved as a cofactor in various metabolic processes, and is essential for both *M. catarrhalis* and its human host (56). In this respect, the RH4 genome contains many iron acquisition and transport systems (Table S8), including all of the *M. catarrhalis* proteins previously described as being involved in this process: lactoferrin binding protein A and B (17), transferrin binding protein A and B (74), heme utilization protein (25), *M. catarrhalis* hemoglobin utilization protein (26), CopB (1), and the main regulator of iron-responsive genes Fur (24). In addition, iron may be acquired through the degradation of heme, catalyzed by a heme-oxygenase. An iron transport system for transport of Fe^{3+} from the periplasm to the cytosol, encoded by the *fbpABC* gene cluster, was identified by homology to the corresponding locus in *H. influenzae* (2). Further, the *afeABCD* gene cluster was identified in RH4, a cluster proposed to be involved in the acquisition of chelated iron, as described for *Actinobacillus actinomycetemcomitans* and *Yersinia pestis*, and regulated by Fur in those species (5,57). Interestingly, the iron binding components of *fbpABC* and *afeABCD* ABC transporter systems were expressed at higher levels compared to the other components of these systems. Two putative NRAMP homologs, involved in transport of Fe^{2+} and Mn^{2+}, could possibly compete with the host divalent ion transporters of the NRAMP family (21). In addition, RH4 possesses two putative bacterioferritins, intracellular iron-storage proteins important for preventing the presence of free iron (11), both of which were found to be expressed at high levels during all phases of growth. Overall, all iron acquisition and transport systems of RH4 were found to be expressed in vitro, but the importance of the individual pathways remains to be investigated.

Oxidative stress. Inherent to aerobic metabolism is oxidative stress caused by reactive oxygen species (ROS) (20). As mentioned previously, acquisition of iron is essential for growth of *M. catarrhalis*, but it can also be harmful for the bacterium, as iron can react with hydrogen

peroxide, resulting in the formation of hydroxyl radicals via the Fenton reaction (20). The well-studied superoxide dismutase/catalase system is able to counteract the effect of oxidative stress, catalyzing the conversion of superoxide to water and oxygen. RH4 contains and expresses the *sodA* and catalase genes, but lacks the *sodB* gene (Table S9). Catalase production is used during routine identification processes, but has limited differential value because not all strains produce catalase (12). Further, the RH4 genome was found to encode and express several putative bacterial peroxiredoxins (Prx) that catalyze the reduction of peroxide, peroxynitrite, and diverse organic hydroperoxides (ROOH) (72), as well as alkyl hydroperoxide reductase/thioredoxin family members (67). Despite the presence of a large number of antioxidant genes in the RH4 genome, we could not identify homologues of known oxidative stress regulators such as OxyR and SoxR (67).

CRISPR elements. Clustered regularly interspaced short palindromic repeats (CRISPRs) are widespread among genomes of prokaryotic organisms, and are thought to be transcribed and processed into small RNAs that confer resistance to phages (30). CRISPR direct repeats are separated by non-repetitive spacers elements, and are often located near gene clusters encoding CRISPR-associated (Cas) protein family members. Two CRISPR repeat regions, I and II, and one putative repeat region (III) were identified in the RH4 genome. CRISPR repeat region II, localized at nt 30997-28086, is in vicinity of six genes encoding proteins with significant homology to known Cas proteins, showing high similarity to the *Yersinia pestis* (Ypest) subtype (30). The RH4 CRISPR repeat regions (I and II) are characterized by an average repeat length of 28.1 nt (Ypest 28.0 nt) and a spacer length of 32.0 nt (Ypest 32.1 nt). Microarray analysis showed constitutive expression of both CRISPR I and II from the minus strand only (Fig. 4). In addition, intermediate levels of expression of the six *cas* genes were detected during all three growth

386 phases (data not shown). The exact role for the CRISPR system in *M. catarrhalis* remains to be
387 determined.

388 **Concluding remarks.** In this publication, we present the completely assembled and annotated
389 genome of the clinically important bacterial pathogen *M. catarrhalis* strain RH4. Comparative
390 genomics revealed a high degree of similarity and sequence conservation between the RH4 and
391 ATCC 43617 genomes. As RH4, the ATCC 43617 strain is classified as a 16S rRNA type 1
392 strain (8) and MLST-analysis showed that it also belongs to the seroresistant lineage (71). Taken
393 together, the availability of a completely assembled and annotated genome of a clinical isolate of
394 *M. catarrhalis* will facilitate the identification of novel (surface-exposed) vaccine targets, will
395 provide the basis for omics-based research such as transcriptomics and proteomics, and is
396 indispensible for a complete understanding of the biology of *M. catarrhalis*.

ACKNOWLEDGMENTS

This work was supported by the OMVac project under the European Union 6th Framework Program. SdV and HJB are supported by the Vienna Spot of Excellence (VSOE) grant (ID337956). KR is supported by grants from the Anna and Edwin Berger, Marianne and Marcus Wallenberg Foundations, and the Swedish Medical Research Council. The authors wish to thank Wolfgang Zimmermann and Steffen Krüger of Agowa Genomics for assistance in genomic sequencing, Miaomiao Zhou and Roland Siezen for the LocateP analyses, Tilman Todt for assistance in microarray data preparation, and the researchers of the Institute for Genome Science (IGS), in particular Michelle Gwinn Giglio, for the use of their annotation engine.

REFERENCES

1. **Aebi, C., B. Stone, M. Beucher, L. D. Cope, I. Maciver, S. E. Thomas, G. H. McCracken, P. F. Sparling, and E. J. Hansen**. 1996. Expression of the CopB outer membrane protein by *Moraxella catarrhalis* is regulated by iron and affects iron acquisition from transferrin and lactoferrin. Infect. Immun. **64**:2024-2030.

2. **Anderson, D. S., P. Adhikari, K. D. Weaver, A. L. Crumbliss, and T. A. Mietzner**. 2007. The *Haemophilus influenzae* hFbpABC Fe3+ transporter: analysis of the membrane permease and development of a gallium-based screen for mutants. J. Bacteriol. **189**:5130-5141.

3. **Attia, A. S., E. R. Lafontaine, J. L. Latimer, C. Aebi, G. A. Syrogiannopoulos, and E. J. Hansen**. 2005. The UspA2 protein of *Moraxella catarrhalis* is directly involved in the expression of serum resistance. Infect. Immun. **73**:2400-2410.

4. **Balder, R., J. Hassel, S. Lipski, and E. R. Lafontaine**. 2007. *Moraxella catarrhalis* strain O35E expresses two filamentous hemagglutinin-like proteins that mediate adherence to human epithelial cells. Infect. Immun. **75**:2765-2775.

5. **Bearden, S. W. and R. D. Perry**. 1999. The Yfe system of *Yersinia pestis* transports iron and manganese and is required for full virulence of plague. Mol. Microbiol. **32**:403-414.

6. **Bendtsen, J. D., L. Kiemer, A. Fausboll, and S. Brunak**. 2005. Non-classical protein secretion in bacteria. BMC. Microbiol. **5**:58.

7. **Berven, F. S., K. Flikka, H. B. Jensen, and I. Eidhammer**. 2004. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res. **32**:W394-W399.

8. **Bootsma, H. J., H. G. van der Heide, S. van de Pas, L. M. Schouls, and F. R. Mooi**. 2000. Analysis of *Moraxella catarrhalis* by DNA typing: evidence for a distinct subpopulation associated with virulence traits. J. Infect. Dis. **181**:1376-1387.

9. **Bootsma, H. J., H. van Dijk, J. Verhoef, A. Fleer, and F. R. Mooi**. 1996. Molecular characterization of the BRO beta-lactamase of *Moraxella* (*Branhamella*) *catarrhalis*. Antimicrob. Agents Chemother. **40**:966-972.

10. **Brooks, M. J., J. L. Sedillo, N. Wagner, C. A. Laurence, W. Wang, A. S. Attia, E. J. Hansen, and S. D. Gray-Owen**. 2008. Modular arrangement of allelic variants explains the divergence in *Moraxella catarrhalis* UspA protein function. Infect. Immun. **76**:5330-5340.

11. **Carrondo, M. A.** 2003. Ferritins, iron uptake and storage from the bacterioferritin viewpoint. EMBO J. **22**:1959-1968.

12. **Catlin, B. W.** 1990. *Branhamella catarrhalis*: an organism gaining respect as a pathogen. Clin. Microbiol. Rev. **3**:293-320.

13. **Cho, B. K., E. M. Knight, and B. O. Palsson**. 2006. Transcriptional regulation of the *fad* regulon genes of *Escherichia coli* by ArcA. Microbiology **152**:2207-2219.

14. **Christensen, J. J., J. Ursing, and B. Bruun**. 1994. Genotypic and phenotypic relatedness of 80 strains of *Branhamella catarrhalis* of worldwide origin. FEMS Microbiol. Lett. **119**:155-159.

15. **de Vries, S. P., H. J. Bootsma, J. P. Hays, and P. W. Hermans**. 2009. Molecular aspects of *Moraxella catarrhalis* pathogenesis. Microbiol. Mol. Biol. Rev. **73**:389-406.

16. **Driessen, A. J. and N. Nouwen**. 2008. Protein translocation across the bacterial cytoplasmic membrane. Annu. Rev. Biochem. **77**:643-667.

17. **Du, R. P., Q. Wang, Y. P. Yang, A. B. Schryvers, P. Chong, M. H. Klein, and S. M. Loosmore**. 1998. Cloning and expression of the *Moraxella catarrhalis* lactoferrin receptor genes. Infect. Immun. **66**:3656-3665.

18. **Dutilh, B. E., Y. He, M. L. Hekkelman, and M. A. Huynen**. 2008. Signature, a web server for taxonomic characterization of sequence samples using signature genes. Nucleic Acids Res. **36**:W470-W474.

19. **Edwards, K. J., J. M. Schwingel, A. K. Datta, and A. A. Campagnari**. 2005. Multiplex PCR assay that identifies the major lipooligosaccharide serotype expressed by *Moraxella catarrhalis* clinical isolates. J. Clin. Microbiol. **43**:6139-6143.

20. **Farr, S. B. and T. Kogoma**. 1991. Oxidative stress responses in *Escherichia coli* and *Salmonella typhimurium*. Microbiol. Rev. **55**:561-585.

21. **Forbes, J. R. and P. Gros**. 2001. Divalent-metal transport by NRAMP proteins at the interface of host-pathogen interactions. Trends Microbiol. **9**:397-403.

22. **Forsgren, A., M. Brant, M. Karamehmedovic, and K. Riesbeck**. 2003. The immunoglobulin D-binding protein MID from *Moraxella catarrhalis* is also an adhesin. Infect. Immun. **71**:3302-3309.

23. **Fujita, Y., H. Matsuoka, and K. Hirooka**. 2007. Regulation of fatty acid metabolism in bacteria. Mol. Microbiol. **66**:829-839.

24. **Furano, K. and A. A. Campagnari**. 2003. Inactivation of the *Moraxella catarrhalis* 7169 ferric uptake regulator increases susceptibility to the bactericidal activity of normal human sera. Infect. Immun. **71**:1843-1848.

25. **Furano, K. and A. A. Campagnari**. 2004. Identification of a hemin utilization protein of *Moraxella catarrhalis* (HumA). Infect. Immun. **72**:6426-6432.

26. **Furano, K., N. R. Luke, A. J. Howlett, and A. A. Campagnari**. 2005. Identification of a conserved *Moraxella catarrhalis* haemoglobin-utilization protein, MhuA. Microbiology **151**:1151-1158.

27. **Gao, F. and C. T. Zhang**. 2008. Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. BMC. Bioinformatics. **9**:79.

28. **Grant, J. R. and P. Stothard**. 2008. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. **36**:W181-W184.

29. **Grissa, I., G. Vergnaud, and C. Pourcel**. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. **35**:W52-W57.

30. **Haft, D. H., J. Selengut, E. F. Mongodin, and K. E. Nelson**. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. **1**:e60.

31. **Hall-Stoodley, L., F. Z. Hu, A. Gieseke, L. Nistico, D. Nguyen, J. Hayes, M. Forbes, D. P. Greenberg, B. Dice, A. Burrows, P. A. Wackym, P. Stoodley, J. C. Post, G. D. Ehrlich, and J. E. Kerschner**. 2006. Direct detection of bacterial biofilms on the middle-ear mucosa of children with chronic otitis media. JAMA **296**:202-211.

32. **Helminen, M. E., I. Maciver, J. L. Latimer, J. Klesney-Tait, L. D. Cope, M. Paris, G. H. McCracken, Jr., and E. J. Hansen**. 1994. A large, antigenically conserved protein on the surface of *Moraxella catarrhalis* is a target for protective antibodies. J. Infect. Dis. **170**:867-872.

33. **Holm, M. M., S. L. Vanlerberg, I. M. Foley, D. D. Sledjeski, and E. R. Lafontaine**. 2004. The *Moraxella catarrhalis* porin-like outer membrane protein CD is an adhesin for human lung cells. Infect. Immun. **72**:1906-1913.

34. **Jenkins, L. S. and W. D. Nunn**. 1987. Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in *Escherichia coli*: the *ato* system. J. Bacteriol. **169**:42-52.

35. **Juncker, A. S., H. Willenbrock, H. G. Von, S. Brunak, H. Nielsen, and A. Krogh**. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. **12**:1652-1662.

36. **Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi**. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Res. **36**:D480-D484.

37. **Karalus, R. and A. Campagnari**. 2000. *Moraxella catarrhalis*: a review of an important human mucosal pathogen. Microbes. Infect. **2**:547-559.

38. **Kent, W. J.** 2002. BLAT--the BLAST-like alignment tool. Genome Res. **12**:656-664.

39. **Lafontaine, E. R., L. D. Cope, C. Aebi, J. L. Latimer, G. H. McCracken, Jr., and E. J. Hansen**. 2000. The UspA1 protein and a second type of UspA2 protein mediate adherence of *Moraxella catarrhalis* to human epithelial cells in vitro. J. Bacteriol. **182**:1364-1373.

40. **Lafontaine, E. R., N. J. Wagner, and E. J. Hansen**. 2001. Expression of the *Moraxella catarrhalis* UspA1 protein undergoes phase variation and is regulated at the transcriptional level. J. Bacteriol. **183**:1540-1551.

41. **Lee, P. A., D. Tullman-Ercek, and G. Georgiou**. 2006. The bacterial twin-arginine translocation pathway. Annu. Rev. Microbiol. **60**:373-395.

42. **Lipski, S. L., C. Akimana, J. M. Timpe, R. M. Wooten, and E. R. Lafontaine**. 2007. The *Moraxella catarrhalis* autotransporter McaP is a conserved surface protein that mediates adherence to human epithelial cells through its N-terminal passenger domain. Infect. Immun. **75**:314-324.

43. **Luke, N. R., S. Allen, B. W. Gibson, and A. A. Campagnari**. 2003. Identification of a 3-deoxy-D-manno-octulosonic acid biosynthetic operon in *Moraxella catarrhalis* and analysis of a KdsA-deficient isogenic mutant. Infect. Immun. **71**:6426-6434.

44. **Luke, N. R., J. A. Jurcisek, L. O. Bakaletz, and A. A. Campagnari**. 2007. Contribution of *Moraxella catarrhalis* type IV pili to nasopharyngeal colonization and biofilm formation. Infect. Immun. **75**:5559-5564.

45. **Mollenkvist, A., T. Nordstrom, C. Hallden, J. J. Christensen, A. Forsgren, and K. Riesbeck**. 2003. The *Moraxella catarrhalis* immunoglobulin D-binding protein MID has conserved sequences and is regulated by a mechanism corresponding to phase variation. J. Bacteriol. **185**:2285-2295.

46. **Murphy, T. F.** 2009. Vaccine development for *Moraxella catarrhalis*: rationale, approaches and challenges. Expert. Rev. Vaccines. **8**:655-658.

47. **Murphy, T. F., A. L. Brauer, B. J. Grant, and S. Sethi**. 2005. *Moraxella catarrhalis* in chronic obstructive pulmonary disease: burden of disease and immune response. Am. J. Respir. Crit. Care Med. **172**:195-199.

48. **Murphy, T. F., A. L. Brauer, N. Yuskiw, E. R. McNamara, and C. Kirkham**. 2001. Conservation of outer membrane protein E among strains of *Moraxella catarrhalis*. Infect. Immun. **69**:3576-3580.

49. **Narita, S., S. Matsuyama, and H. Tokuda**. 2004. Lipoprotein trafficking in *Escherichia coli*. Arch. Microbiol. **182**:1-6.

50. **Nikaido, H.** 1996. Multidrug efflux pumps of Gram-negative bacteria. J. Bacteriol. **178**:5853-5859.

51. **Ouellet, M., P. D. Adams, J. D. Keasling, and A. Mukhopadhyay**. 2009. A rapid and inexpensive labeling method for microarray gene expression analysis. BMC. Biotechnol. **9**:97.

52. **Pelicic, V.** 2008. Type IV pili: e pluribus unum? Mol. Microbiol. **68**:827-837.

53. **Perez Vidakovics, M. L. and K. Riesbeck**. 2009. Virulence mechanisms of *Moraxella* in the pathogenesis of infection. Curr. Opin. Infect. Dis. **22**:279-285.

54. **Plamondon, P., N. R. Luke, and A. A. Campagnari**. 2007. Identification of a novel two-partner secretion locus in *Moraxella catarrhalis*. Infect. Immun. **75**:2929-2936.

55. **Raetz, C. R., C. M. Reynolds, M. S. Trent, and R. E. Bishop**. 2007. Lipid A modification systems in Gram-negative bacteria. Annu. Rev. Biochem. **76**:295-329.

56. **Ratledge, C. and L. G. Dover**. 2000. Iron metabolism in pathogenic bacteria. Annu. Rev. Microbiol. **54**:881-941.

57. **Rhodes, E. R., A. P. Tomaras, G. McGillivary, P. L. Connerly, and L. A. Actis**. 2005. Genetic and functional analyses of the *Actinobacillus actinomycetemcomitans* AfeABCD siderophore-independent iron acquisition system. Infect. Immun. **73**:3758-3763.

58. **Ruckdeschel, E. A., A. L. Brauer, A. Johnson, and T. F. Murphy**. 2009. Characterization of proteins Msp22 and Msp75 as vaccine antigens of *Moraxella catarrhalis*. Vaccine **27**:7065-7072.

59. **Ruckdeschel, E. A., C. Kirkham, A. J. Lesse, Z. Hu, and T. F. Murphy**. 2008. Mining the *Moraxella catarrhalis* genome: identification of potential vaccine antigens expressed during human infection. Infect. Immun. **76**:1599-1607.

60. **Sethi, S. and T. F. Murphy**. 2008. Infection in the pathogenesis and course of chronic obstructive pulmonary disease. N. Engl. J. Med. **359**:2355-2365.

61. **Smith, C. A.** 2006. Structure, function and dynamics in the mur family of bacterial cell wall ligases. J. Mol. Biol. **362**:640-655.

62. **Thanassi, D. G. and S. J. Hultgren**. 2000. Multiple pathways allow protein secretion across the bacterial outer membrane. Curr. Opin. Cell Biol. **12**:420-430.

63. **Timpe, J. M., M. M. Holm, S. L. Vanlerberg, V. Basrur, and E. R. Lafontaine**. 2003. Identification of a *Moraxella catarrhalis* outer membrane protein exhibiting both adhesin and lipolytic activities. Infect. Immun. **71**:4341-4350.

64. **Verduin, C. M., C. Hol, A. Fleer, H. van Dijk, and A. van Belkum**. 2002. *Moraxella catarrhalis*: from emerging to established pathogen. Clin. Microbiol. Rev. **15**:125-144.

65. **Vergison, A.** 2008. Microbiology of otitis media: a moving target. Vaccine **26 Suppl 7**:G5-10.

66. **Verhaegh, S. J., A. Streefland, J. K. Dewnarain, D. J. Farrell, A. van Belkum, and J. P. Hays**. 2008. Age-related genotypic and phenotypic differences in *Moraxella catarrhalis* isolates from children and adults presenting with respiratory disease in 2001-2002. Microbiology **154**:1178-1184.

67. **Wang, G., P. Alamuri, and R. J. Maier**. 2006. The diverse antioxidant systems of *Helicobacter pylori*. Mol. Microbiol. **61**:847-860.

68. **Wang, W., L. Reitzer, D. A. Rasko, M. M. Pearson, R. J. Blick, C. Laurence, and E. J. Hansen**. 2007. Metabolic analysis of *Moraxella catarrhalis* and the effect of selected *in vitro* growth conditions on global gene expression. Infect. Immun. **75**:4959-4971.

69. **Wang, W., A. R. Richardson, W. Martens-Habbena, D. A. Stahl, F. C. Fang, and E. J. Hansen**. 2008. Identification of a repressor of a truncated denitrification pathway in *Moraxella catarrhalis*. J. Bacteriol. **190**:7762-7772.

70. **Wilson, J. C., P. M. Collins, Z. Klipic, I. D. Grice, and I. R. Peak**. 2006. Identification of a novel glycosyltransferase involved in LOS biosynthesis of *Moraxella catarrhalis*. Carbohydr. Res. **341**:2600-2606.

71. **Wirth, T., G. Morelli, B. Kusecek, A. van Belkum, C. van der Schee, A. Meyer, and M. Achtman**. 2007. The rise and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. Genome Res. **17**:1647-1656.

72. **Wood, Z. A., E. Schroder, H. J. Robin, and L. B. Poole**. 2003. Structure, mechanism and regulation of peroxiredoxins. Trends Biochem. Sci. **28**:32-40.

73. **Yu, C. S., Y. C. Chen, C. H. Lu, and J. K. Hwang**. 2006. Prediction of protein subcellular localization. Proteins **64**:643-651.

74. **Yu, R. H. and A. B. Schryvers**. 1993. The interaction between human transferrin and transferrin binding protein 2 from *Moraxella* (*Branhamella*) *catarrhalis* differs from that of other human pathogens. Microb. Pathog. **15**:433-445.

75. **Zhou, M., J. Boekhorst, C. Francke, and R. J. Siezen**. 2008. LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC. Bioinformatics. **9**:173.

607
608 # TABLES

609 TABLE 1. General genome features and functional classification of the predicted RH4 proteins

General genome features	Number or %	
Genome size	1,863,286	
GC content (%)	41.7	
Number of genes	1964	
Protein encoding ORFs	1886	
Average ORF size (bp)	850	
Coding density (%)	86.1	
rRNAs	12	
tRNAs	50	
Functional class	Number or %	% of total
Amino acid biosynthesis	77	4.0
Purines. pyrimidines. nucleosides. and nucleotides	36	1.9
Fatty acid and phospholipid metabolism	40	2.1
Biosynthesis of cofactors. prosthetic groups, and carriers	90	4.7
Central intermediary metabolism	33	1.7
Energy metabolism	140	7.4
Transport and binding proteins	148	7.8
DNA metabolism	78	4.1
Transcription	35	1.8
Protein synthesis	141	7.4
Protein fate	89	4.7
Regulatory functions	31	1.6
Signal transduction	11	0.6
Cell envelope	168	8.8
Cellular processes	61	3.2
Mobile and extrachromosomal element functions	21	1.1
Unknown function	172	9.0
Hypothetical proteins	515	27.1
Disrupted reading frame	16	0.8

610
611 TABLE 2. Genome comparison of *M. catarrhalis* RH4 and ATCC 43617

% RH4 CDS in alignment	% mismatch in alignment	Number	% of total
100	0	416	22.1
100	<20	1252	66.4
>80	<20	125	6.6
60-80	<20	10	0.5
40-60	<20	7	0.4
0-40	<20	8	0.4
No match	-	68	3.6

612

613
614 TABLE 3. Major virulence factors of *M. catarrhalis* RH4

RH4 locus	Gene	Common name and general function	Note	Expression per growth phase in BHI		
				Lag	Exp	Stat
MCR_1198	*uspA1*	Ubiquitous surface protein A1; functions in adhesion to epithelial cells and ECM components, inhibition of NF-KB pro-inflammatory response via CEACAM binding, serum resistance, binding of complement regulatory proteins, and biofilm formation	9 poly(G) repeats upstream of the ORF, reported to correlate with low expression (40)	+++	+++	+++
MCR_0329	*uspA2H*	Ubiquitous surface protein A2H; functions in adhesion to epithelial cells, serum resistance, and biofilm formation		+++	+++	+++
MCR_0617	*mid/hag*	Immunoglobulin D binding protein/hemagglutinin; functions in adhesion to epithelial cells and IgD binding	3 triplets of G-residues at 5' end within the ORF, reported to correlate with high expression (45)	+++	++	+++
MCR_0419	*mcaP*	*Moraxella catarrhalis* adherence protein; functions in adhesion to epithelial cells		+++	+++	+++
MCR_1698	*ompCD*	Outer membrane protein CD; functions in adhesion to epithelial cells and middle ear mucin	RFLP pattern 1 (BseJI restriction fragments of ~200 and 100 nt), which is almost exclusively found in 16S type 1 isolates (66)	+++	+++	+++
MCR_0492	*copB*	Outer membrane protein CopB; involved in serum resistance and important for in vivo survival	RFLP type II (NsaI restriction fragments of 332 and 187 nt), which almost exclusively exists in 16S type I isolates (66)	+++	+++	++
MCR_0858	*ompE*	Outer membrane protein E; involved in serum resistance	Putative homolog of long chain fatty acid transporter FadL	+++	+++	+++
MCR_1591	*bro-2*	Beta-lactamase family protein BRO-2		++	++	++

615 Abbreviations: ECM, extracellular matrix; ORF, open reading frame; RFLP, restriction fragment

616 length polymorphism; nt, nucleotides.

617 TABLE 4. Subcellular compartment distribution of the predicted RH4 proteins

Subcellular localization	Number	% of total
Cytoplasm – Intracellular	1284	68.1
Cytoplasm (possibly secreted via non-classical pathways) - Intracellular or secreted	35	1.9
Inner membrane - Multi-transmembrane	257	13.6
Inner membrane - Multi-transmembrane (lipid modified N-terminus)	4	0.2
Inner membrane - C-terminally anchored (with CS)	1	0.1
Inner membrane (Periplasm) - N-terminally anchored (no CS)	124	6.6
Inner membrane (Periplasm) - N-terminally anchored (with CS)	14	0.7
Inner membrane (Periplasm) - Lipid anchored	12	0.6
Periplasm - Secreted (with CS)	22	1.2
Outer membrane - Lipid anchored	53	2.9
Outer membrane - β-barrel protein (no CS)	19	1.0
Outer membrane - β-barrel protein (with CS)	23	1.2
Outer membrane (no clear SP found but have β-barrel) - Intracellular or β-barrel	20	1.1
Extracellular - Secreted (with CS)	18	1.0

618 Abbreviations: CS, cleavage site; SP, signal peptide.

619

620
621 **LEGENDS TO FIGURES**

622 FIG. 1: Circular representation of the *M. catarrhalis* RH4 genome. Gene coordinates are given in
623 Mbp. From inside to outside, the circles represent the GC skew, CG content, genes on the
624 forward strand, and genes on the reverse strand. The color-coding of the two outermost rings
625 represent the functional class of the gene, and corresponds to the colors in Table S1. The figure
626 was generated using CGview (28).

627

628 FIG. 2: Modular arrangement and functional domains of UspA1 (A) and UspA2H (B) of RH4.
629 Both proteins are drawn to scale. The various sequence cassettes and the known functional
630 domains are indicated. The VEEG repeats of UspA2H are marked with an asterisk because they
631 only partially match the repeat motif. Abbreviations: CTER1, C-terminal region 1; CTER2, C-
632 terminal region 2.

633

634 FIG. 3: Predicted central metabolism of *M. catarrhalis* RH4. The glycolysis pathway is
635 incomplete, missing the key-enzymes phosphofructokinase and pyruvate kinase. The enzymes of
636 glycolysis are expected to be involved in the gluconeogenic pathway, of which all enzymes were
637 found to be present. Fatty acid degradation and acetate assimilation are probably of high
638 importance, as they serve to supply the tricarboxylic acid (TCA) cycle with acetyl-CoA. The
639 glyoxylate pathway is completely present and allows bypass of the TCA cycle, which is missing
640 both subunits of succinyl-CoA synthetase.

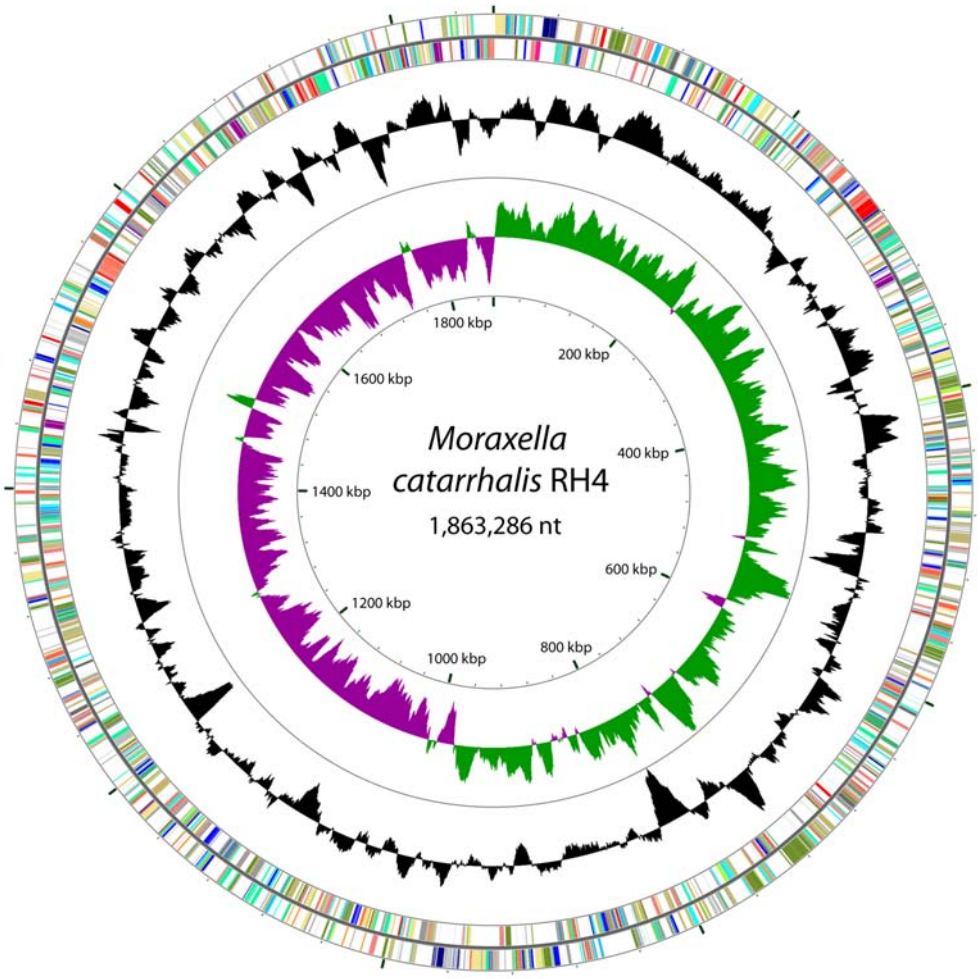
641

642 FIG. 4: Expression profiling of CRISPR I and II chromosomal regions. Exponential growth phase
643 $\log_2$ probe signal intensities of both strands (open circles, reverse strand probes; closed circles,

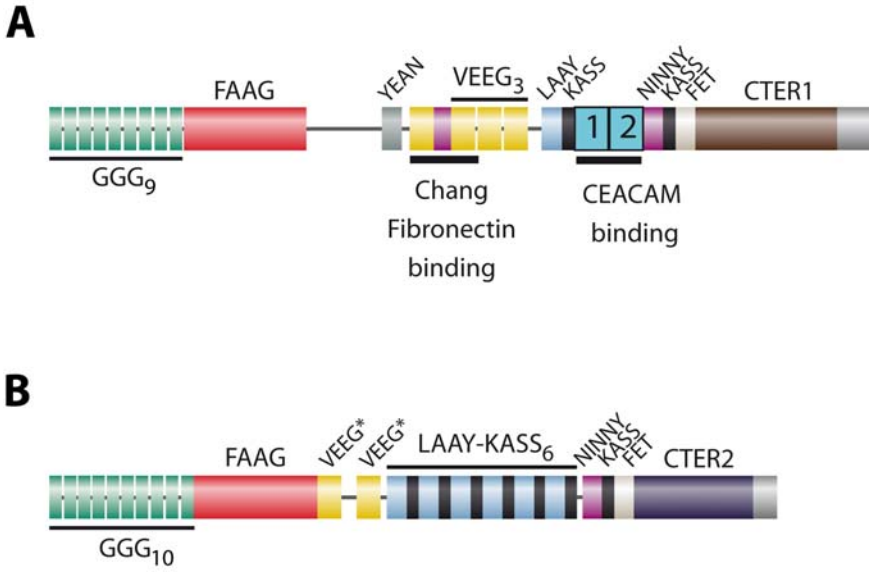
644 forward strand probes) demonstrates reverse strand expression of both CRISPR I and II regions.
645 Localization of the CRISPR1 and CRISPR2 regions is indicated below the graph. Comparable
646 results were obtained during the lag and stationary growth phases (data not shown).

FIGURES

Figure 1

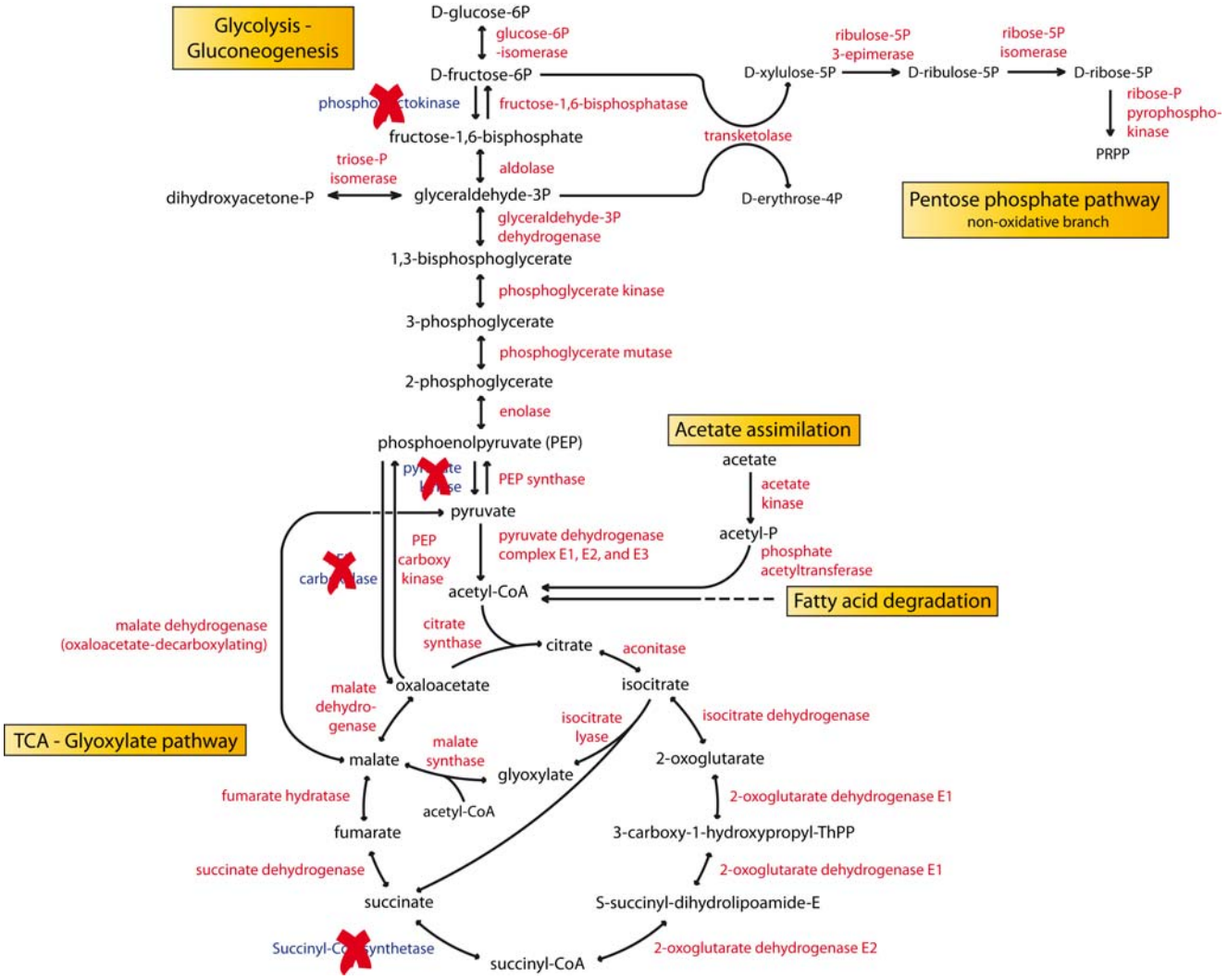


650 Figure 2



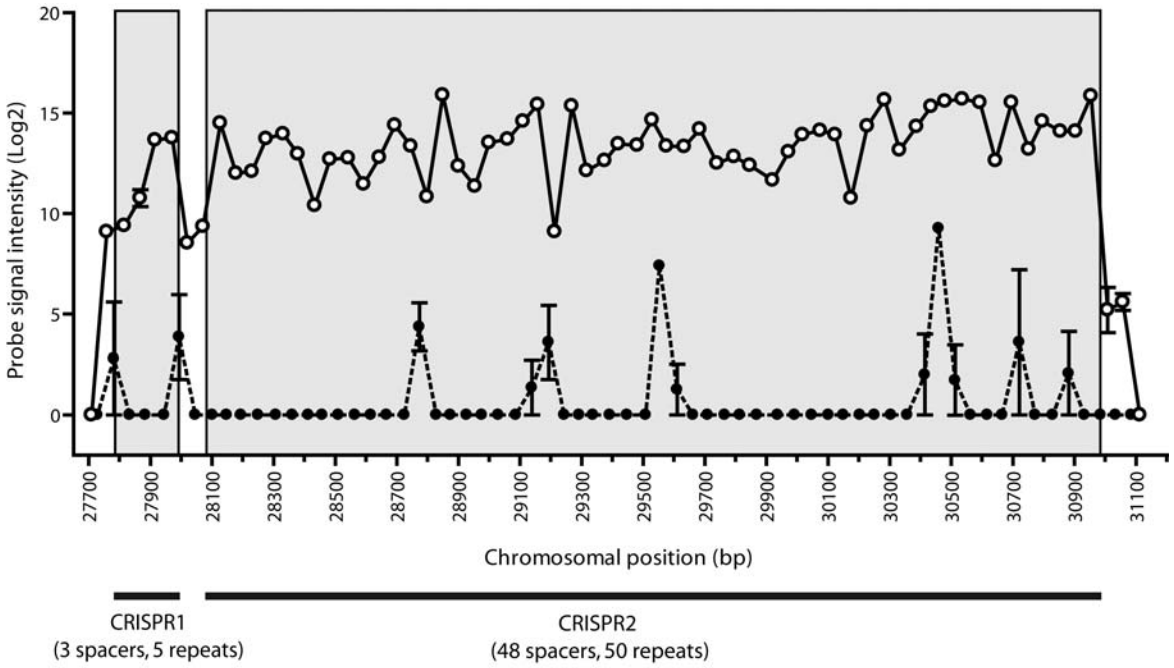
651

652 Figure 3



653

654 Figure 4



655