Reductive Genome Evolution of Obligate Symbiont *Midichloria humanum*: Implications of Gene Loss

Eleonora Petryayeva[1‡] and Vincent Hanlon[2‡]

[1] *Institute for Symbiont Genetics, Galactic University of Coruscant, Galactic City BB8C3PO, Coruscant*
[2] *Department of Sentient Biology, Hutt University of Tatooine, Hutt City ASWR2D2, Tatooine*
[‡] These authors contributed equally to this work

Received 24 March 60 ABY | Accepted April 13 60 ABY

We sequenced and analyzed the genome of the *Midichloria humanum*, a ubiquitous intracellular symbiont, in 34 human subjects covering a range of population sizes from 1,000 to over 14,000 midichlorians per human cell. Midichlorians mediate the relationship between their human hosts and the Force, which confers exceptional physical and cognitive abilities on hosts with per-cell midichlorian populations ('midichlorian count,' mdc) larger than 10,000. Although, quantitatively, the genome of *M. humanum* is more extensively degraded than those of closely related *Rickettsia* species, midichlorians retained many intact genes involved in metabolic pathways, energy production and conversion, as well as nucleotide transport and metabolism. Wide-genome comparison between human groups classified based on mdc revealed that this symbiont continues to lose genes as a result of reductive genome evolution. Human groups with mdc < 10,000 have shown a loss of the DNA replication initiation gene *dnaA* that codes for the protein DNaA. Extensive polymorphisms, single nucleotide polymorphisms (SNPs) and deletion mutations have degraded the *recA* gene implicated in DNA recombination, which remains intact in the closely related species *M. togrutas*. A single mutation in the DNA repair gene *mutY* produced an inactive protein in groups with low mdc, further reducing pathways of DNA repair. The groups with medium and high mdc number retained active MutY protein. Overall, this work demonstrates that *M. humanum* is undergoing a slow genome reduction and that these genomic changes may be related to the continuous decline in birth rate of Force-sensitive children with mdc over 10,000.

Keywords: midichloria, symbiont, genome, shotgun sequencing, gene loss, evolution.

Introduction

Midichlorians are symbiotic bacteria that inhabit the cells of all known eukaryotes. Populations of the human-specific midichlorian species *Midichloria humanum* range in size from a minimum of 1,000 per cell in Force-insensitive individuals to a reported maximum of over 20,000 per cell in the Jedi and Sith lord Anakin Skywalker. Force-sensitivity in humans is not an acquired trait, but rather determined by the population size of midichlorians in each human cell ('midichlorian count,' mdc), with mdc of 10,000 being the minimum threshold for significant Force-sensitivity. Force-sensitivity confers exceptional physical abilities on human hosts, as well as telekinesis, mind-control, and precognition in extreme cases.

Over the last millennium, wars and galactic unrest have led to a substantial decrease in the number of Jedi Knights. The process of becoming a Jedi typically involves intensive training a Force-sensitive human child, that is, a child with mdc > 10,000. An alarming decline in the numbers of Jedi Knights, as well as a *ca.* 8% decrease over the last 50 years in the number of newborn humans with mdc > 10,000 (based on statistical data from Kyber Memory Crystal at Coruscant Temple[1]), prompted the establishment of the Midichlorian Genome Project funded by the Republican Defence Board.

Obligate mutualistic symbionts such as *M. humanum* both depend on and are essential for host survival and reproduction. Such symbiotic relationships are typically vertically transmitted (*i.e.* vertical gene transfer) from parent to offspring and result from a prolonged evolutionary association with the host lineage. Recent genome analysis of various intracellular obligate symbionts has shown that these species have tendency to reduce in genome size during evolution.[2] Generally, smaller genomes have a lower G+C content (< 35%) exhibiting strong bias towards high A+T content and lack genes for the regulation of gene expression, DNA replication, recombination and repair.[2-4] The reduction of genome in obligate symbiont can reach an extreme, resulting in 130-200 Kbp genomes (*e.g. Candidatus* genus).[2]

Because symbiosis between bacteria and hosts relies on mutual benefits, bacteria tend to lose genes that are not essential for their own or their hosts' survival, either by deletion or – less commonly – by gene transfer to the host genome.[4] In the latter case, gene functionality is not lost and bacteria rely on the host for gene expression.[2] Commonly lost genes include those involved in DNA recombination (*dnaA* – replication initiator, *dnaB* – helicase, *dnaG* – primase, *dnaD, dnaI, dnaE*), and DNA repair (*recA, lexA*, *MutY*, and *recN*). Genes that encode repair proteins may be lost spontaneously, but this is expected to result in higher mutation rates over time and further accelerate the pace of genome evolution.

The repertoire of genes that symbiotic bacteria retain during reductive genome evolution is not random. For instance, only 66 protein-coding genes are conserved among members of the *Candidatus* family. These retained genes follow a strong pattern: in addition to a small number of genes involved in information processing (*i.e.* translation, transcription, and replication), these bacteria retain genes necessary for the provision of nutrients to the host. At least 86% of genes in obligate symbionts associated with invertebrates (*Arsenophonus, Candidatus, Buchnera, Wiggleworthia*) are devoted to amino acid synthesis, metabolism and information processing, compared with 37% in free-living *Escherichia coli*.[2]

Some of the genes typically associated with reductive genome evolution are of particular interest: *dnaA, recA*, *mutY*. The replication initiator protein DnaA, which exists in bound forms with ATP or ADP, binds to the *oriC* region of the chromosome which contains a series of five 9-bp sequence repeats known as DnaA boxes.[5] This binding results in the separation of two DNA template strands and initiation of the replication process. Recombinase RecA is a key enzyme for homologous recombination in the SOS-response pathway[6] and was suspected to be a major driving force for genome rearrangement, including large deletions. Its role in reductive genome evolution remains controversial and it was shown that *Salmonella* bacteria could undergo RecA-independent recombination events.[7] The *mutY* gene encodes for MutY protein, a 36-kDa DNA glycosylase that repairs A-G base pair mismatches to C-G.[8] The loss of *mutY* is believed to cause decreasing G+C content of symbiont genome.[9]

Here, we present complete whole-genome shotgun sequencing of *M. humanum*, an obligate symbiont, within four groups that vary in the midichlorian population size (mdc) in human cells. Blood samples were collected from humans living on four planets (Coruscant, Corelian, Devaron, and Tatooine) and human subjects were divided into four groups: Group A (Jedi, mdc > 10,000), Group B (mdc > 10,000), Group C (mdc = 5,000 – 10,000), and Group D (mdc <5,000). The population size of midichloria in isolated human white blood cells was determined using flow cytometry. The complete genome was compared between four groups, as well as with previously sequenced *Midichloria togrutas*[10] from a non-human Jedi host, and with the genome of the last known midichloria ancestor, *Rickettsia prowazekii.*[11] The incidence of mutation and gene loss in all four groups, as well as the continuing process of genome reduction, was assessed using comparative genomic and functional analysis, with a particular focus on the genes *dnaA, mutY,* and *recA.* Genome-wide analysis will improve our understanding of the process of reductive genome evolution in *M. humanum*. Of particular concern are the potential consequences of reductive genome evolution for long-term changes in mdc in humans and the continuing availability of suitable recruits for the Jedi Order.

Methods

Collection of Blood Samples

Blood samples were collected from 280 human volunteers through medical centres of the Transgalactic Medical Research Centre (TMRC) located on Coruscant, Corellian, Devaron, and Tatooine. Each volunteer gave informed signed consent and the study was approved by the TMRC Research Ethics Board. Additionally, four heparinized blood sample from human Jedi were generously provided by the TMRC Genetic Databank. Peripheral blood mononuclear cell (PBMC) samples were obtained by the Ficoll-Hypaque (Amersham Pharmacia Biotech, NJ, CA) standard density gradient centrifugation method. Small aliquots of PBMC were cryopreserved in a freezing medium containing 10% DMSO and 30% FCS in RPMI 1640 (Life Technologies, NY, USA), in liquid nitrogen.

Cell Labeling and Two-Colour Flow Cytometry Analysis

The population size of midichloria in PBMC samples was determined using flow cytometry. Primary rabbit anti-CD45 monoclonal antibody (EP322Y) and secondary polyclonal anti-rabbit antibody labeled with Alexa Fluor 594 (A594) dye were purchased from Abcam (ON, CA). The midichlorians staining kit – CellLight^{TM} Midichloria-GFP BacMaM 2.0 was from ThermoFisherUniversal (KP, CS). The labeling was performed in a two-step procedure. First, 100 μL of PBMC sample was incubated with CellLight^{TM} Midichloria-GFP stain overnight according to manufacturer's specifications. Cells were washed twice by centrifugation (400g x 5 min) with PBSS buffer (10 mM, pH 7.4, 1% BSA) and resuspended in 100 μL of buffer. Second, for indirect immunostaining of biomarker on human cell membrane, 100 μL of midichloria prestained cells was incubated with 20 μL of 1/5000× dilution of primary antibodies for 4 h at 4^{o}C. Cells were washed twice by centrifugation with 4 mL of HBSS buffer (10 mM, pH 7.3, 1% BSA) at 400g for 5 min at 4^{o}C. Pellet was resuspended in 100 μL of HHBS-BSA and 30 μg of donkey IgG was added to block non-specific adsorption. Then 20 μL of fluorophore-labeled secondary antibody was added and cells were incubated for 30 min at 4^{o}C. Cells were washed twice by centrifugation (400g x 5 min) with 4 mL of HHBSS-BSA buffer containing 0.1% (v/v) saturated solution of LDS-751 in methanol at 4^{o}C.

A two-colour flow cytometry analysis was performed with antibody combinations 1^{o}Ab-CD45 – 2^{o}Ab-A594 and 1^{o}Ab-Midichondria – 2^{o}Ab-A488. All samples were analyzed on Coulter EPICS XL flow cytometer (Beckman Coulter Co.) mounted with two photomultiplier tubes and two lasers (Argon and HeNe). In any typical experiment side-scatter (SS) versus forward-scatter (FS) cytogram (50,000 CD45+ events) were collected per sample and analyzed on EXPO32 software (applied Cytometry Systems, UK). Calibration was performed with fluorospheres (Beckman Coulter Co.).

Midichondrial Genome Isolation and Shotgun Sequencing

Midichlorial DNA (mdDNA) was extracted from PBMC samples using a midichondrial DNA isolation kit and following the supplier protocol. To evaluate the level of DNA contamination, DNA was analyzed by Southern hydridization using digoxigenin oligonucleotide labeling kit (Boehringer Mannheim), with probes that recognize the 16S rRNA, the eukaryotic elongation factor EF1-α, and midichondrial cytochrome oxidase. No human nuclear DNA was detected, and mdDNA purity was estimated at 96%.

Shotgun sequence libraries were generated by mechanical shearing of genomic DNA into small inserts (1.6–2.0 kb) and cloning into pUC18 vector following a two-step ligation method and using pulsed-field gel electrophoresis for all fragments isolation. Random shortgun sequencing was performed to 8.9-fold sequence coverage with 2,104 *M. humanum*-derived sequences (average read length 630 nt). This library required gathering a total of 8,300 sequences. High-throughput sequencing was performed on Illumina HiSeq2500 with HiSeq SBS kit v4 from Illumina (CA, US) at Applied Genomics Research Centre (Coruscant).

Genome Analysis

Open reading frames (ORFs) were identified with metaBEETLE and MEGAN4[12] softwares, and the putative encoded proteins were compared with known sequences database using BLASTP. ORFs smaller than 100 amino acids were not considered a gene, unless there was a similarity detected with BLAST. Amino acid sequence alignment with homologous proteins from genome sequences of other alpha-proteobacteria was done with Clustal Omega[13] and WebPRANK[13]. All tRNA were identified using tRNAscan-SE program.[14] All rRNA and other small RNA were identified by BLASTN searches of intergenic regions against RNA-specifying genes in Genome Information Broker.[15] Potential polymorphisms and single nucleotide polymorphisms (SNPs) were identified with Consed program.[16] In the absence of the *oriC* gene, a diagnostic cluster of DNaA boxes, the putative origin of replication was determined by GC-skew analysis using Oriloc program.[17]

Results and Discussion

Population size of *M. humanum* in Human Cells

Samples of purified human peripheral blood mononuclear cells from 280 volunteers were stained for simultaneous detection of human cells and intracellular midichlorians. The blood cell surface CD-45 biomarker was labeled in a sandwich immunolabeling using primary and secondary antibodies. The midichlorians were labeled using a construct for GFP expression fused to the sequence of E1 alpha pyruvate dehydrogenase. The pattern of distribution of *M. humanum* in human white blood cells was detected using two-colour flow cytometry as shown in Figure 1. All analyzed samples have shown a broad range of midichlorian counts (mdc) in human cells (1,080 – 18,000); however, each individual sample had a discrete number with a variation of < 2%. Similar results were obtained using two-photon confocal microscopy imaging (data not shown). Three subject groups were used to annotate samples based on their mdc (Group B > 10,000 mdc, Group C 5,000–10,000 mdc, and Group D < 5,000 mdc). Ten randomly selected samples within each group with a total of 30 samples were further used for the *M. humanum* genome analysis described below.

General Genome Properties and Comparison

The genome of *M. humanum* is represented by a single circular chromosome of 378,056 bp with an average G+C content of 29.32%, similar to many endosymbiotic bacteria. No plasmid DNA was found. The variation in the genome size between different groups categorized based on mdc was within 8.6%, as shown in Table 1. In the absence of a diagnostic cluster of DNaA boxes, the putative origin of replication (*oriC*) was assigned to the intergenic region next to the *gidA* gene based on GC- and AT-skew analysis as described previously.[18] We identified in Group A/B a total of 287 putative genes and 7 pseudogenes from genetic mapping of *M. humanum*, all of which had significant database matches, and 246 of which were assigned a biological function. The functional set contains 231 protein-coding genes with an average size of 1,200 nucleotides per gene, one ribosomal RNA, two small RNAs, and 11 tRNAs specifying all 20 amino acids.

Surprisingly, *M. humanum* exhibited strong intrapopulation variation at 634 sites, including polymorphisms, single nucleotide polymorphisms (SNPs) and deletion mutations (indels), as shown in Figure 2A. Although the sample size was small and there is a possibility that a small fraction of observed variations is due to the sequencing error, the distribution of polymorphisms throughout the genome is consistent with natural variation. A total of 45% of observed polymorphisms affected the reading of protein-coding regions leading to potentially deleterious and inactive genes. The pseudogenes showed the most significant variation ($P < 0.001$) with a higher frequency of SNPs which further decrease G+C content and higher frequency of indels most likely due to relaxation against deletional bias, as well as a general loss of DNA repair genes. The low G+C content of endosymbionts is indicative of accumulation of deleterious changes by genetic drifts and mutational bias towards A+T.[2] The variation in the size of indels (1–186 bp) as a result of slipped-strand mispairing was a major factor in the variation of the genome size of *M. humanum* between four groups.

Among 297 identified genes in *M. humanum*, only 78% were conserved across human population with various mdc. Putatively functional copies of 218 genes found in group D represent only 80% of those found in Groups A and B, accounting for genes that were lost entirely from the lineage and genes that have lost functionality but can still be recognized as pseudogenes. Direct comparison of phylogenetic relationship with 272 protein-coding genes found in *M. togrutas* suggests that at least 38 genes have undergone parallel evolution. Three genes that were completely lost in Group D genome are present as pseudogenes in Groups A and B, while *M. togrutas* retained fully functional copies. Two of these genes (COG403, COG470) are involved in RNA modification – queuosine biosynthesis. The relative proportion of conserved genes and pseudogenes between three symbiont families is shown in Figure 2B.

Because *M. humanum* undergoes vertical gene transfer but not horizontal gene acquisition,[19] the gene loss history could be reconstructed since divergence from the last known ancestor, *Rickettsia prowazekii*. Many gene losses determined from comparative analysis occurred *via* inactivation and disintegration of individual genes or to a lesser extent *via* large segment loss of multiple adjacent genes (*e.g. klpM, ygbR, ygbS, ygbT*, and *cysC, -D, -F, -G, -H and –I*).

Functional Analysis of Protein-Coding Genes: Loss and Decay

A systematic comparison of *M. humanum and M. togrutas*[10] genome with its closest fully sequenced relative *Rickettsia prowazekii*[11] revealed that both species have undergone extensive gene loss with respect to 1,177 genes of free-living alpha-proteobacteria. These results are summarized in Table 2. Some of the strongly affected functions include DNA replication, recombination and repair, lipid metabolism, and moderately affected functions include metabolism of nucleotides and amino acids and cell envelope biogenesis (Table 2).

Among all the identified differences in genome makeup between four groups of *M. humanum* and those relative to *M. togrutas*, few were particularly surprising, including DNA replication initiation *dnaA*, DNA recombination *recA*, and DNA repair *mutY* genes.

A common feature of endosymbiont genomes is the loss of the replication initiation mechanism, as indicated by the lack of *dnaA* gene. Suprisingly, the only form of *M. humanum* to retain *dnaA* was found to be in groups A and B where total mdc exceeds 10,000. In contrast, DNA replication of the symbiont in groups B and C is entirely under the control of the host cell. Furthermore, noncanonical *oriC* sites, such as *priA* and *recA,* can provide alternative mechanisms for DNA replication; these, however, are also absent or present only in pseudogene form in the *M. humanum* genome.

Another distinct feature of endosymbionts is the loss of DNA recombination (*recA* gene) mechanism. Recombinase RecA is an essential enzyme for homologous recombination with the primary function of deleting large DNA sequences. A search for *recA* sequence revealed that all Group A – D contained highly degraded DNA fragments consistent with RecA. Deletions were found in positions 2389 – 2489. In the *M. togrutas*[10] genome, the *recA* gene was not recognized due to a large deletion despite some possible fragments corresponding to *N*-terminal of RecA. In the symbiont *R. prowazekii* the coding region of RecA for 344- amino acid protein was intact.

Another striking difference found between four groups carrying *M. humanum* symbiont was variation in DNA repair gene, *mutY*. A translated amino acid sequence search of MdH_0017 was homologous to *MutY* found in other bacteria. This gene was found in a putatively active from in the genome of Groups A, B, and C, while pseudogene with mutation "UGA" at codon 154 was found in Group D. This mutation results in the splitting of the open reading frame, making a new stop codon. MutY is composed of two domains: N-terminal and C-terminal, such that DNA substrates bind in the cleft of these two domains. While 3D homology modeling of MutY in Groups A–C produces functional intact protein, the product of split genes observed in Group D results in aberrant protein structure. Considering that almost an entire sequence is still intact further suggests that *M. humanum* is still in the process of reductive genome evolution and may not have reached its final end point. The symbiont genomes represent a progression of genomic changes that originate from host-restricted lifestyle.[2, 3] Tiny compact genomes are a long-term outcome driven by proliferation of mobile elements, chromosome rearrangements, gene inactivation, pseudogene accumulation, and deletions.[3] Over time, pseudogene fragments and mobile elements are removed and gene loss continues until the genome is shrunk to the least size, sometimes even losing genes that are considered essential for cellular life.[2] Co-adaptation with the host is crucial in the process of gene loss.

Conclusions

Based on a hypothesis that the presymbiotic ancestor of *Midichloria* had a large, enterobacterial-like genome containing at least 2,000 – 2,500 genes as was recently noted by Nobuki *et al.*[20], our comparative analysis of two *Midichloria* strains suggests that the vast majority of reductions in genome size (80-85%) occurred soon after the establishment of symbiosis but before diversification of major lineages. Since diversification from *Rickettsia*[20], the genomes of *Midichloria* underwent further reductions of 5-15%. This pattern is consistent with the current view within scientific community that genome size reduction associated with bacterial lifestyle transitions is subject to exponential decay.[21] Our analysis indicates that genome evolution in *Midichloria* is rather degenerate (*cf.* adaptive), based on the mutation bias, erosion of the regulatory system (regulatory genes, promoters, transcription attenuators, and DnaA boxes), ongoing pseudogene formation and gene loss across all functional categories, a bias of gene inactivation towards the terminus of replication, and most importantly – given their deleterious effect on mutation rates – a continuous reduction of the arsenal of repair systems and of the replication machinery. Extensive gene loss and high population-level polymorphism associated with DNA repair, replication, and recombination suggest that *M. humanum* is undergoing a process of genome reduction typical of intracellular endosymbionts. Genetic isolation and small effective population size are major factors in degenerate genome evolution.[22] Typically, such populations are subject to increased genetic drift and reduced efficacy of selection, and as a result show an irreversible accumulation of mildly deleterious mutations and a progressive loss of fitness.[22] The predicted long-term evolutionary outcome of this process, known as Muller's ratchet, are mutational meltdowns and population extinction.[23] For symbionts, the timeframe of this process may depend on a number of factors, including strength of selection on both host and symbiont, and population genetic parameters such as effective population size.[23] The varying extent of genome reduction revealed by the present analysis seems to further suggest that indefinite genomic stasis is unsustainable and a symptom of genome degeneracy.

A surprising feature of the polymorphism observed in *M. humanum* is the disproportionate gene loss experienced by Group D populations, that is, populations of midichlorians with mdc < 5000. More extensive genome reduction in groups with low mdc may be attributed to the loss of genes necessary to sustain higher midichlorian population size. Since symbionts can lose beneficial genes if they are not essential for the survival of either the bacteria or the host,[2] it is apparent that genes that sustain high midichlorian population size are non-essential. Although the causal relationships between host Force-sensitivity, mdc, and midichlorian genome reduction are not fully understood, continuing genome reduction may result in decreased mdc. This possibility is especially concerning in light of the long-term decline in the frequency of human individuals with mdc > 10,000 recorded in the Kyber Memory Crystal[1] and the associated decline in potential recruits for the Jedi Order. Further research on the evolutionary relationship between host Force-sensitivity and the molecular evolution *M. humanum* may be required to understand the apparent resistance of high-mdc midichlorian populations to genome reduction and the comparative susceptibility of low-mdc midichlorian populations. Thus, our ongoing studies will focused on mapping out the replication pathways of midichlorians to determine if (1) *M. humanum* genes present in human cells with mdc > 10,000 are the result of a potential "defence" mechanism that resists further genome reduction, or (2) the *M. humanum* genome in high-mdc populations is equally susceptible to mutation and ultimately will decrease mdc below the threshold necessary for Force-sensitivity.

Acknowledgements

We thank Connie OKO and Lilian Skywalker for assistance with genome sequencing and data analysis at the Applied Genomics Research Centre. The authors thank the Transgalactic Medical Research Centre (TMRC), KCRP, and Jedi Institute of Advanced Science for support of this research. The authors also thank Dr. K. Mono, Dr. R. Dooku and two anonymous reviewers for critical reading of the manuscript.

References

[1] Kyber Memory Crystal Archive Database, S. Accessed on 18 March, 60 ABY.

[2] McCutcheon, J. P., and Moran, N. A. (2012) Extreme genome reduction in symbiotic bacteria, *Nat. Rev. Microbiol. 10*, 13-26.

[3] Wernegreen, J. J. (2002) Genome evolution in bacterial endosymbionts of insects, *Nat. Rev. Genet. 3*, 850-861.

[4] Moran, N. A., McCutcheon, J. P., and Nakabachi, A. (2008) Genomics and Evolution of Heritable Bacterial Symbionts, *Annu. Rev. Genet. 42*, 165-190.

[5] Robinson, A., Causer, R. J., and Dixon, N. E. (2012) Architecture and Conservation of the Bacterial DNA Replication Machinery, an Underexploited Drug Target, *Curr. Drug Targets 13*, 352-372.

[6] Lenhart, J. S., Schroeder, J. W., Walsh, B. W., and Simmons, L. A. (2012) DNA Repair and Genome Maintenance in Bacillus subtilis, *Microbiol. Mol. Biol. Rev. 76*, 530-564.

[7] Nilsson, A. I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J. C. D., and Andersson, D. I. (2005) Bacterial genome size reduction by experimental evolution, *Proc. Natl. Acad. Sci. USA 102*, 12112-12116.

[8] Au, K. G., Clark, S., Miller, J. H., and Modrich, P. (1989) Escherichia coli mutY gene encodes an adenine glycosylase active on G-A mispairs, *Proc. Natl. Acad. Sci. USA 86*, 8877-8881.

[9] Kuwahara, H., Takaki, Y., Shimamura, S., Yoshida, T., Maeda, T., Kunieda, T., and Maruyama, T. (2011) Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of Calyptogena clams, *BMC Evol. Biol. 11*, 1-13.

[10] Lan, D., and Ruck, R. (59 ABY) Whole Genome Sequencing of the Symbiont Midichloria Togrutas from a Jedi of the Primary Humanoid Inhabitants of Planet Shili., *Genome Biol. Evol. 7*, 3022-3032.

[11] Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C. M., Podowski, R. M., Naslund, A. K., Eriksson, A.-S., Winkler, H. H., and Kurland, C. G. (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria, *Nature 396*, 133-140.

[12] Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. (2011) Integrative analysis of environmental sequences using MEGAN4, *Genome Res. 21*, 1552-1560.

[13] http://www.ebi.ac.uk/Tools/msa/. Accessed 18 March, 2016.

[14] Schattner, P., Brooks, A. N., and Lowe, T. M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs, *Nucleic Acids Res. 33*, W686-W689.

[15] http://gib-v.genes.nig.ac.jp. Accessed March 18, 2016.

[16] http://www.phrap.org/consed/consed.html. Accesssed March 18, 2016.

[17] http://pbil.univ-lyon1.fr/software/Oriloc/oriloc.html. Acessed March 18, 2016.

[18] Eppinger, M., Baar, C., Raddatz, G., Huson, D. H., and Schuster, S. C. (2004) Comparative analysis of four Campylobacterales, *Nat. Rev. Microbiol. 2*, 872-885.

[19] Lawrence, J. G. (1999) Gene transfer, speciation, and the evolution of bacterial genomes, *Curr. Opin. Microbiol. 2*, 519-523.

[20] Nabuki, U., Uluçkan, Ö., Graña, O., Keller, J., and Busse, B. (59 ABY) The Force: Origin and Evolution of Midichloria, *Science and Universe 8*, 330ra337-330ra337.

[21] Fitzpatrick, B. M. (2014) Symbiote transmission and maintenance of extra-genomic associations, *Front. Microbiol. 5, 46*.

[22] van Ham, R. C. H. J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J. M., Jiménez, L., Postigo, M., Silva, F. J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Morán, F., and Moya, A. (2003) Reductive genome evolution in Buchnera aphidicola, *Proc. Natl. Acad. Sci. USA 100*, 581-586.

[23] Pettersson, M. E., and Berg, O. G. (2006) Muller’s ratchet in symbiont populations, *Genetica 130*, 199-211.

Figure 1.

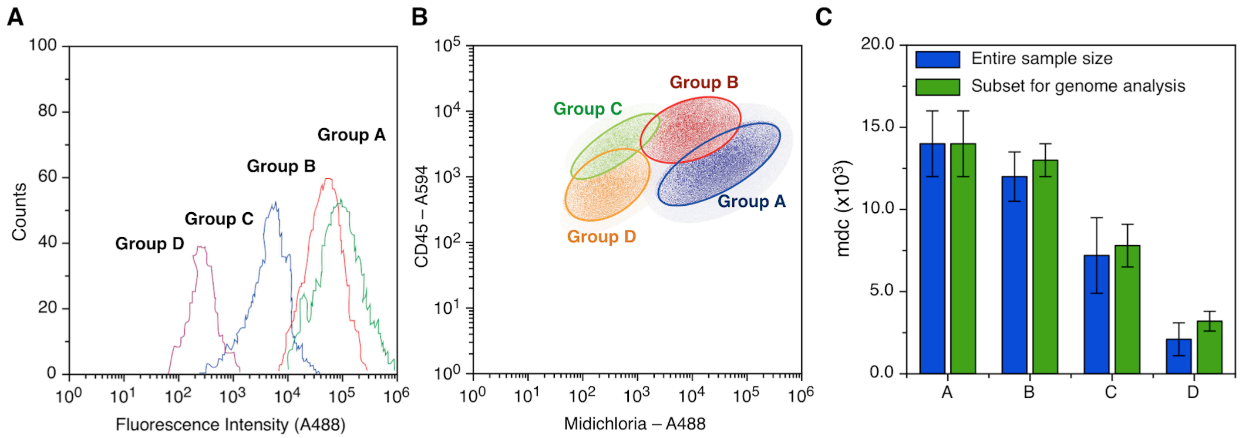

Figure 1. Two-colour flow cytometry showing human blood cells and the intracellular symbiont *M. humanum*. (A) representative flow cytometry histogram showing the relative distribution of mdc in human cells; (B) dot blot corresponding to data shown in (A); (C) four groups assigned based on mdc used in this work. The entire sample size for groups B-D is 280 and subsets each contained ten samples; for group A only four samples were analyzed in total.

Figure 2.

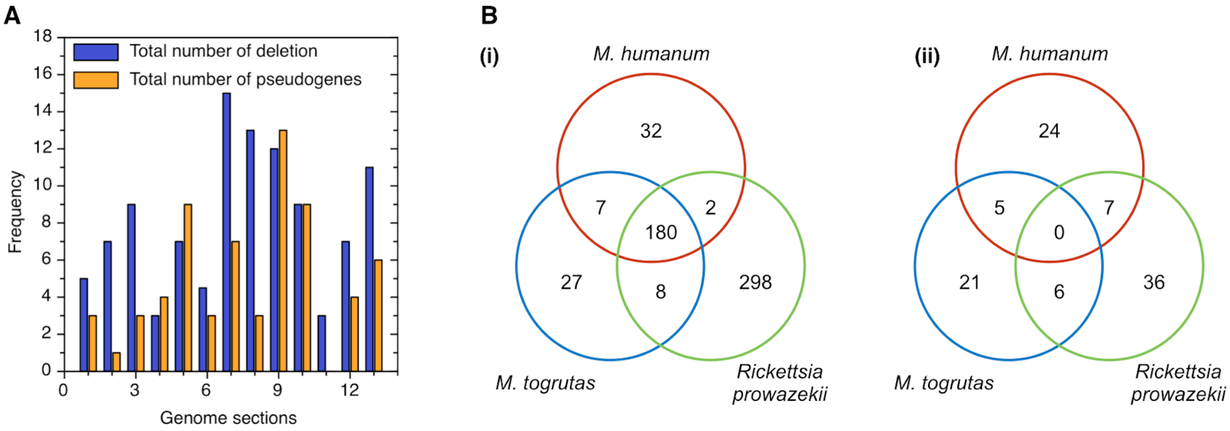

Figure 2. (A) Relation between number of events in gene loss and pseudogene formation as a function of genomic position of affected genes. All 297 genes were numbered from the point of putative origin of replication and binned into 13 sections (*ca.* 23 genes per section). (B) Venn diagram showing number of conserved and unique genes and pseudogenes. (i) based on putative functional protein-coding genes, and (ii) pseudogenes.

Table 1.

Table 1. Comparison of *Midichloria* and proteobacterial endosymbiont genomes

	M. humanum				*R. prowazekii*†	*M. togrutas*§
	Group A	Group B	Group C	Group D		
	mdc* >10,000	> 10,000	5,000-10,000	<5,000		>10,000
	(Jedi)					(Jedi)
Chromosome, bp	378,056	378,056	362,643	345,543	1,111,523	381,163
Plasmids, bp	0	0	0	0		2 (6,381)
G+C content, %	29.32	29.32	28.79	26.46	29.1	25.3
Total gene number	297	297	261	233		294
Protein-coding genes	268	268	241	218	835	272
rRNAs	1	1	1	1	3	2
tRNAs	2	2	2	2	33	2
Pseudogenes	7	7	9	12	14	5
Protein-coding regions, %	85	85	89	86	76	81

*mdc – *M. humanum* count number per human host cell.

[†,§] Data was taken from ref. [10, 11]

Table 2.

Table 2. Gene conservation, loss and degradation in *M. humanum* and *M. togrutas**

Function	*M. humanum*				*M. togrutas*
	Group A	Group B	Group C	Group D	
Translation	98/12/1	98/12/1	96/12/3	95/13/3	101/10/1
Transcription	10/13/1	10/13/1	10/13/1	8/14/2	16/16/1
DNA replication, recombination, and repair	32/20/2	32/20/2	30/21/3	29/22/1	34/18/2
Posttranslational modifications	21/26/3	21/26/3	20/26/4	20/26/4	26/22/4
Signal transduction mechanisms	5/16/0	5/16/0	5/16/0	5/16/0	5/16/0
Amino acid transport and metabolism	23/28/0	23/28/0	23/28/0	23/28/0	28/28/0
Energy production and conversion	45/12/0	45/12/0	45/12/0	45/12/0	45/12/0
Coenzyme metabolism	26/37/2	26/37/2	26/37/2	26/37/2	26/37/2
Nucleotide transport and metabolism	28/12/0	28/12/0	28/12/0	28/12/0	28/12/0
Sugar transport and metabolism	8/12/1	8/12/1	8/12/1	8/12/1	8/12/1
Lipid metabolism	11/32/1	11/32/1	11/32/1	11/32/1	11/32/1
Metabolites biosynthesis	2/8/0	2/8/0	2/8/0	2/8/0	2/8/0
Cell division and chromosome partitioning	8/6/1	8/6/1	6/6/3	6/6/3	12/6/1
Cell envelope biogenesis, outer membrane	26/15/1	26/15/1	26/15/1	28/13/1	26/15/1
Cell motility	0/32/0	0/32/0	0/32/0	0/32/0	0/32/0
Inorganic ion transport and metabolism	12/25/1	12/25/1	12/25/1	12/25/1	12/25/1
Defence mechanism	2/7/1	2/7/1	2/7/1	2/7/1	2/7/1
Intracellular trafficking and secretion	18/2/1	18/2/1	18/2/1	18/2/1	18/2/1
General function prediction	31/84/2	31/84/2	31/84/2	30/84/3	37/80/2
Function unknown	18/94/1	18/94/1	18/94/1	18/94/1	18/94/1

*Gene analysis is performed with respect to 1,177 genes of free-living alpha-proteobacteria. In each column, the first number represents number of retained genes, the second number lost genes and the third number indicates number of pseudogenes.