Tracing lifestyle adaptation in prokaryotic genomes

Eric Altermann¹,² *

¹ Rumen Microbiology, Animal Nutrition and Health Capability Group, Grasslands Research Centre, AgResearch Limited, Palmerston North, New Zealand
² Riddet Institute, Massey University, Palmerston North, New Zealand

INTRODUCTION

Microbial genomes range in size from the smallest microbial genome known to date of Candidatus Tremblaya princeps with just under 139,000 nt (McCutcheon and von Dohlen, 2011) to large genomes with over 13,000,000 nt such as Sorangium cellulosum SO ce56 (Schneiker et al., 2007). With an average gene size of about 1,000 nt, microbial genomes harbor between 140 and 13,000 genes.

Lifestyle adaptation of microbes due to changes in their ecological niches or acquisition of new environments is a major driving force for genetic changes in their respective genomes. Moving into more specialized niches often results in the acquisition of new gene sets via horizontal gene transfer to utilize previously unavailable metabolites, while genetic ballast is shed by gene loss and/or gene inactivation. In some cases, larger genome rearrangements can be observed, such as the incorporation of whole genetic islands, providing a range of new phenotypic capabilities. Until recently these changes could not be comprehensively followed and identified due to the lack of complete microbial genome sequences. The advent of high-throughput DNA sequencing has dramatically changed the scientific landscape and today microbial genomes have become increasingly abundant. Currently, more than 2,900 genomes are published and more than 11,000 genome projects are listed in the Genomes Online Database⁴. Although this wealth of information provides many new opportunities to assess microbial functionality, it also creates a new array of challenges when a comparison between multiple microbial genomes is required. Here, functional genome distribution (FGD) is introduced, analyzing the diversity between microbes based on their predicted ORFeome. FGD is therefore a comparative genomics approach, emphasizing the assessments of gene complements. To further facilitate the comparison between two or more genomes, degrees of amino-acid similarities between ORFeomes can be visualized in the Artemis comparison tool, graphically depicting small and large scale genome rearrangements, insertion and deletion events, and levels of similarity between individual open reading frames. FGD provides a new tool for comparative microbial genomics and the interpretation of differences in the genetic makeup of bacteria.

Keywords: functional genomics, genome comparison, lifestyle adaptation, horizontal gene transfer, genome plasticity

With the advent of high-throughput DNA sequencing techniques such as automated capillary sequencing, solid-state sequencing¹, or pyrosequencing² or by synthesis³, access to high-quality draft, and complete bacterial genomes has become feasible and is a commonly used technique. Currently 2,943 complete genomes (including eukaryotic organisms) have been reported, with another 184 archael and 5,490 bacterial genome projects in progress⁴. The availability of complete or nearly complete genomes triggered attempts to incorporate selected larger genetic subsets (Makarova et al., 2006; Makarova and Koonin, 2007) or complete genomes (Rohwer and Edwards, 2002; Henz et al., 2005) to infer evolutionary lineages, while less progress has been made in analysis of multiple whole microbial genomes from functional and comparative genomic perspectives. One of the most prominent examples of whole genome comparative analyses is based on Blast Score Ratios (Rasko et al., 2005).

Here, we introduce a new analysis tool, functional genome distribution (FGD). FGD does not attempt to represent the...
The evolutionary path a genome has taken, since different genes will have been acquired by different routes. Instead, FGD investigates the overall levels of similarity between microbial genomes based on amino-acid sequences of the predicted complete ORFeomes. This reflects the impact the evolutionary force has had on genome makeup in the past, resulting in the current level of niche adaptation (Thomson et al., 2003). Thus presence, absence, or modification of individual genes or genetic islands defines the phenotypic potential of a given organism at a given temporal snapshot. The comparison of these ORFeomes to each other ultimately defines the level of similarity of the genomes. This approach then also takes into account important genetic adaptations to specific ecological niches or even to human made environments such as industrial fermentation processes. Such common genotype adaptations might render organisms more similar by FGD analyses than their respective evolutionary heritage would indicate. The presented approach of a FGD is a BLAST-based ORF-position-independent algorithm, implemented in the compACTor software. In the context of FGD analyses the term "functional" is used in the sense of functionality based on sequence and sequence similarity and is not based on annotation classification [i.e., such as implied by COG (Tatusov et al., 2003) or KEGG (Kanehisa, 2002; Kanehisa et al., 2008) databases].

Research in functional genomics (defined as the investigation of gene function by gene inactivation, gene complementation, and in silico analyses) relies heavily on the identification of differences between two or more genomes, identifying differences in the presence or absence of individual genes. The compACTor software also creates all-vs.-all ORFeome distance data files which, in combination with the respective GenBank files of the query microbes and the Artemis comparison tool (ACT; Carver et al., 2005), facilitate visual qualitative comparative in silico analyses of complete and draft phase microbial genomes for the subsequent analysis of changes to gene synteny and operon structures. Furthermore, to identify genes shared and unique between selected clusters a mining tool, FGDfinder, is provided, facilitating rapid identification of relevant gene sets for further in vitro analyses.

**MATERIALS AND METHODS**

**GENERAL WORKFLOW OF THE compACTor SOFTWARE**

A pool of genomes in GenBank format represents the query space for compACTor (Figure 1). Gene models embedded and subsequently retrieved from the GenBank files of each query genome are parsed and the resulting ORFeome is translated into individual amino-acid sequences. The parsing algorithm considers both gene and CDS features for any given start position, while giving preference to CDS entries. Nucleotide sequences are retrieved from the genome sequence and translated into amino-acid sequence. Gene model errors such as unbalanced number of nucleotides, unrecognized amino-acid codons, or multiple stop-codons are reported but treated as non-critical. The translated ORFeome is then used to build query-specific amino-acid BLAST databases (DBs) using formatdb (Altschul et al., 1990). In subsequent runs previously built BLAST databases can be re-used (Figure 1), reducing the overall runtime of the comparator software. The pooled ORFeomes of all query entries are hashed for subsequent all-vs.-all analyses. Levels of similarities are inferred for each predicted query ORF by an all-vs.-all analysis, submitting sequence-database combinations to a non-filtered gapped BLASTP (Altschul et al., 1997) pipeline. Briefly, the deduced amino-acid sequence for each open reading frame (ORF) of each query entry is compared to all subject databases generated from the entry pool. Individual BLAST results are parsed and collated in ORF-specific ASCII result files. The total number of BLAST analyses performed is a direct function of the number of individual ORFeomes and the pool size.

The collated flatfile database is the basis for all subsequent calculations such as the generation of ACT (Carver et al., 2005) pMSP-datafiles, the prediction of putative strain-specific genes, and the generation of FGD trees. It is also possible to add new query entries to existing analyses. The compACTor software tests for the presence of respective ORF-database BLAST results and omits the BLAST pipeline where possible. This significantly reduces runtime requirements and facilitates future expansions of existing comparative analyses.

**GENERATION OF ACT COMPARISON DATA FILES (pMSP)**

Based on the all-vs.-all principle, all genome-genome pair combinations are analyzed. A bi-directional BLASTP approach was implemented for the creation of ORF-specific ACT comparison files. Genome pairs analyzed in respective comparison files are reflected within the filename which features both filenames in the order query-to-subject. Each predicted ORF of the query genome was allowed a maximum of 20 similarity hits to ORFs present in the respective subject genome. ACT allows two quality parameters, namely “Score cutoff” and “Percent ID cutoff” in its comparison file format. These two cutoffs determine the stringency of sequence similarities, and were replaced in the compACTor software output by the alternative parameters of “e-value” ("Percent ID") and "weighted score" ("Score"). The parameter "e-value" is derived from individual e-values which are grouped into empirically determined trust level ranges (Table A1 in Appendix).

Similarly, for the second parameter, a weighted score is calculated, placing additional emphasis on alignment quality and length (Eq. 1). Briefly, the initial alignment-centric local BLAST score is reflecting the maximum level of sequence similarity between a given ORF pair ($\xi$, Eq. 1). To account for changes outside the local alignment and to further assess the alignment quality, the score is then subjected to three penalty blocks, assessing the quality of amino-acid similarity [(P − Id/2) + Id/λ, Eq. 1], the size of gaps (1 − G/A, Eq. 1), and the length of the BLAST alignment in relation to the query ORF length (A − G/L, Eq. 1). These penalty parameters take into account differences in amino-acids sequence alignments which are not or insufficiently covered by the regular BLAST score. This process effectively reduces the level of similarity between both ORFs when the BLAST alignment covers only partial ORF sequences and/or the alignments show significant levels of insertions or deletions and results in a weighted BLAST score $\psi$.

The weighted score is a sub-parameter of the distance calculation between two genomes (see Eq. 2) and allows for stringency shifting in ACT while maintaining continuity to the FGD algorithm.
FIGURE 1 | Graphical workflow of the compACTor software. Schematic representation of the algorithm structure. Symbols used are consistent with standard flowchart icons. Abbreviations: DB, database; pMSPs, ORFeome based MSPCrunch comparison format (Sonnhammer and Durbin, 1994) data files. Oval symbol: data pools, diamond symbol: internal decision points, square boxes: internal processes and functions, hourglass symbol: central parsing algorithm, Multidocument symbol: external flatfile databases created by compACTor.
Functional genome distribution

Distance between two genomes

\[ \delta_{\text{Query} \rightarrow \text{Subject}} = \frac{\text{TORF}_{\text{Query}}}{\sum_{n=1}^n \psi(n)} \times \rho_{\text{ORF}} \]  

with: \( \delta \) = final similarity score between two genomes, \( \psi = \) weighted BLAST score as described in Eq. 1, \( \lambda = \) length of deduced amino-acid query sequence, \( T_{\text{ORFQuery}} = \) total number of ORFs in query genome, \( n = \) current query ORF, and \( \rho_{\text{ORF}} = \) ORF number ratio as described in Eq. 3.

ORFeome ratio between query and subject genome

\[ \rho_{\text{ORF}} = \left( \frac{T_{\text{ORFquery}}}{T_{\text{ORFsubject}}} \leq 1 \right) \lor \left( \frac{T_{\text{ORFsubject}}}{T_{\text{ORFquery}}} \right) \]  

with: \( \rho_{\text{ORF}} = \) ORF number ratio between query and subject genome, \( \leq 1 \), \( T_{\text{ORFquery}} = \) total number of ORFs in query genome, \( T_{\text{ORFsubject}} = \) total number of ORFs in subject genome.

MOTIVATION OF THE FGD ALGORITHM

The FGD algorithm initially investigates the level of similarity for each predicted ORF in a query genome to a subject ORF and is based on the BLAST score. This score describes the overall quality of the best local alignment found for a given query–subject sequence alignment. However, the score alone does not necessarily reflect the overall level of similarity between two sequences outside the local alignment. Also, from a functional genomics perspective, insertion, and deletion events in individual amino-acid sequences may change the properties of the gene product, thus decreasing the level of similarity. Furthermore, the relation of the local alignment to the overall sequence is an important factor. While stretches of highly conserved sequence contribute to a higher BLAST score, the presence of unique sequence outside the local alignment is likely to contribute to changing the properties of respective gene products. Therefore, the maximum level of similarity as described by the BLAST score will be decreased by the FGD algorithm if insertion/deletion events and imbalanced alignment/deduced sequence ratios are detected.

The FGD analysis then summarizes individual similarity scores determined for each ORF in the ORFeome. The presence of unique strain-specific genes and differences in the overall number of predicted ORFs in respective microbial genome pairs contribute to different genotypes and consequently decrease the level of overall functional similarity between genome pairs.

The initial alignment BLAST score is subjected to quality assessments as described above for Eq. 1. In Eq. 2 the resulting weighted score is normalized by the number of deduced amino-acids of the query ORF, resulting in a score per amino-acid \( [\psi/\lambda(n)] \). Individual similarity scores are then are summed up over the query ORFeome and normalized over total number of predicted ORFs, resulting in a weighted score per ORF. Assessing differences in ORFeome sizes by calculating the ratio between query and subject ORFeomes (Eq. 3), the similarity score is adjusted to result in a final genome similarity score \( \delta_{\text{Query} \rightarrow \text{Subject}} \).

Genome similarity scores for each query–subject combination are entered into a distance matrix. A symmetrical distance matrix, based on the geometrical means of each genome–genome pair combination is calculated according to Eq. 4.

Symmetrical distance matrix

\[ \Delta_{\text{Sym}} = \frac{\delta_{\text{Query} \rightarrow \text{Subject}} + \delta_{\text{Subject} \rightarrow \text{Query}}}{2} \]
with: 
\[ \Delta_{\text{sym}} = \text{final score for symmetrical distance matrix between a given genome–genome pair} \]
\[ \delta_{\text{query}} \rightarrow \delta_{\text{subject}} = \text{similarity score between a given genome–genome combination}. \]

Distance matrices of larger datasets become increasingly complex to read directly. Therefore a traditional tree representation was used as a visual aid to the underlying data structure. The resulting symmetrical distance matrix can be imported into software packages such as Mega4 (Tamura et al., 2007). Genome clusters are approximated using the Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm (Sneath and Sokal, 1962). Approximated branch lengths, representing the level of similarity between genomes, are depicted by distance units (du). Although similar in appearance to phylogenetic trees, the resulting functional distribution tree (FDT) depicts a different concept, clustering organisms by their respective ORFeome similarities rather than by their inferred evolutionary line.

**Identification of putative strain-specific genes**

A key interest in comparative functional genomics is the identification of strain and cluster specific and shared gene sets, identifying the genetic differences, and similarities between selected microbial genomes. These gene sets can then be used to investigate potentially unique metabolic capabilities of microbes and may reveal specific genome adaptation to lifestyle changes. Cluster specific genes or genes shared between clusters are identified using the flatfile BLAST database created by compACTor. Query ORFs exhibiting a minimum e-value threshold with respect to a given subject genome are identified and stored if they fulfill the stringency restriction. Stringency between selected clusters can be set by selecting the number of allowable mismatches. A mismatch is defined as an ORF in the query cluster that has at least one ORF–ORF pairing above the minimum e-value threshold to the subject cluster. Objects implemented in the output are: name of the query GenBank file, respective ORF identifier, the name of the subject GenBank file, annotation of the subject ORF, and the corresponding e-value, query ORF length, score, query ORF start and stop positions, and gene, locus_tag, product, and note annotation. Data for each query ORFeome – subject genome combination are collated and summarized in a tab-delimited ASCII file. The individual data files can easily be imported into spreadsheet programs such as MS Excel or OpenOffice Calc and then be further re-grouped and analyzed. For example, grouping by query ORF number quickly identifies the degree to which a respective ORF is unique among the pool of genome entries.

**Implementation**

The compACTor and FGDFinder software has been realized as PERL scripts (ActivePerl version 5.8.8.822), using the official NCBI BLAST distribution. The software is freely available for non-commercial use (commercial use requires a license).

The algorithms and BLAST utilities used in compACTor render the application CPU intensive. compACTor is multi-threaded and support multiple CPUs/cores to reduce runtime in a near linear fashion. For the analysis depicted in Figure 2, 39 genomes, ranging from ∼1.7 to ∼6.6 Mbp, were analyzed. A total number of 113,397 ORFs resulted in ∼4,300,000 individual BLAST queries. Artemis Comparison Files, the prediction of putative strain-specific ORFs and the phylogenetic tree were calculated using a collated and parsed BLAST flatfile database of >11 Gb. 1,560 Artemis comparison files were generated, reflecting all possible genome–genome combinations (a datapool of >2.5 Gb).

**Known limitations**

Known limitations of the algorithm at this time are the lack of integrated support for extrachromosomal genetic elements such as plasmids and additional chromosomes. A workaround exists by concatenating DNA sequences of all genetic information present in the respective organism and then performing an automated annotation using pipelines such as GAMOLA (Altermann and Klaenhammer, 2003). The implemented GenBank parser currently ignores joined features. In prokaryotes, joined features such as introns [e.g., the mobilization protein MobA E1 (genomic position 1344187.0.1345428, 1347858.0.1348370) in Lactococcus lactis subsp. cremoris MG1363 (GI:124491690)] are very rare and usually do not interfere with the analyses. Also, identical genomes still incur a distance to each other, resulting in a small branching when visualizing the dissimilarity matrix. A normalization algorithm will be implemented in the next release to address this issue. compACTor and FGDFinder are actively developed and future releases will resolve those issues.

**Results and discussion**

**Evaluation of functional distribution trees**

To assess the FGD concept 35 completed and four draft phase genomes from different phyla, including several archeal genomes, were selected (Table A2 in Appendix). The purpose of this diverse taxonomy is to investigate how FGD places individual genomes into clusters and how similar these genomes are to each other on a functional level. The majority of the genomes selected are members of the family Lactobacillaceae in the order Lactobacillales (12 genomes). To investigate the discriminatory power of the FGD algorithm, the range was expanded and five genomes of the family Bacillaceae (order Bacillales), three genomes of the family Streptococcaceae (order Lactobacillales) and two genomes of the family Leuconostocaceae (order Lactobacillales) were added. Eight more distantly related genomes of the order Clostridiales (class Clostridia) were chosen to broaden the taxonomic selection to different classes. All of these genomes are members of the phylum Firmicutes. Three representatives of the class Gammaproteobacteria, (phylum Proteobacteria) were included to investigate inter-phylum relationships. As a final outlier, six genomes of archael Euryarchaeota were included in the analysis.

Figure 2 represents the functional distribution of the selected taxa within an FDT. To test the influence of early, incomplete draft phase genomes on functional placement, the draft phase sequence of Anaerostipes caccae DSM14662 (Schwiertz et al., 2002) which consisted of ∼1.69 Mbp at the time of analysis was included. FGD penalized the missing genetic information and
Publicly available complete genomes were downloaded in GenBank format from the NCBI genome database. Publicly available draft phase genomes were downloaded in FASTA format, concatenated using a universal spacer-stop-spacer sequence and automatically annotated using GAMOLA (Altermann and Klaenhammer, 2003). The in-house draft phase genome of Butyrivibrio proteoclasticus was assembled into an artificial genome and annotated using GAMOLA (publication in preparation). Predicted ORFeomes of all genomes were subjected to an FGD analysis and the resulting distance matrix was imported into MEGA4. The functional distribution was visualized using the UPGMA method (Sneath and Sokal, 1962). The optimal tree with the sum of branch length = 133.1 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the functional distances used to infer the distribution tree.

set *A. caccae* apart from the Clostridiales as a separate cluster. When the updated genome sequence of *A. caccae* encompassing ~3.6 Mbp was included instead, *A. caccae* shifted its position and clustered with *Ruminococcus obeum* and *Ruminococcus gravis*. 

**FIGURE 2** | Functional genome distribution of 39 taxa. Publicly available complete genomes were downloaded in GenBank format from the NCBI genome database. Publicly available draft phase genomes were downloaded in FASTA format, concatenated using a universal spacer-stop-spacer sequence and automatically annotated using GAMOLA (Altermann and Klaenhammer, 2003). The in-house draft phase genome of Butyrivibrio proteoclasticus was assembled into an artificial genome and annotated using GAMOLA (publication in preparation). Predicted ORFeomes of all genomes were subjected to an FGD analysis and the resulting distance matrix was imported into MEGA4. The functional distribution was visualized using the UPGMA method (Sneath and Sokal, 1962). The optimal tree with the sum of branch length = 133.1 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the functional distances used to infer the distribution tree.
Altermann Functional genome distribution

www.frontiersin.org February 2012 | Volume 3 | Article 48 | 7

Butyrivibrio proteoclasticus (branch length 1.99 du), while showing a deeper branching to Butyrivibrio proteoclasticus (branch length 3.13 du; data not shown). This clearly highlights the necessity of obtaining high-coverage genome sequence data for FGD analyses. However, with high-throughput genome sequencing techniques currently available, initial draft phase genomes usually encompass 85–95% of the genome, thus allowing an initial representative functional placement.

Subsequent analysis of complete genomes included in the FDT revealed the deepest branching (20.6 du) for euryarchaeal genomes. The selected methanogens form two distinct genome clusters (node at 5.3 du), separating Methanococcales burtonii and Methanosarcina mazei (Deppenmeier et al., 2002) from the remaining four taxa. Initial comparisons of habitat, growth temperature, and GC content did not indicate a consistently shared denominator between the two groups. Further analyses will be necessary to determine the intrinsic functional similarities indicated by the FGD approach.

Interestingly, Clostridiales, Bacillales, and a subcluster comprised of Ruminococci and B. proteoclasticus B316 formed a new functional node within the FDT (branch depth from node to Lactobacillales was 7.8 du), combining the taxonomic families into one genome cluster. Although the genomes of members of Clostridiales and Bacillales are still placed into distinct functional groups (internal cluster branch depth was 4 du) and no taxon shuffling was observed between both sub-clusters, it appears that lifestyle adaptation has led to similar genome content, potentially indicative of a high level of HGT between both families or from one family to the other.

A survey of the nine ORFeomes of the Bacillus and Clostridium clusters revealed 154 ORFs that are highly conserved in both groups (e-value threshold 1e-100). As expected, most ORFs could be assigned to central house-keeping functions such as DNA synthesis and repair (21%), tRNA genes and related processes (12%), central metabolism (35%), transcription and translation (6%), cellular processes (9%), and molecule transport (8%). However, besides these central functions, a significant number of ORFs related to sporulation were present and highly conserved in both clusters (8%). Based on the algorithms used, it is reasonable to hypothesize that the presence of these highly conserved sporulation genes may be one of the key drivers for the observed clustering of bacilli and true clostridia.

When compared to the Ruminococcus subcluster a similar conserved gene set was found with the notable absence of most sporulation genes (e-value threshold 1e-60). This is in agreement with the observed non-sporeulating phenotype. It is noteworthy, that two conserved genes involved in sporulation (stage V sporulation protein D, spoVD, and Sporulation initiation inhibitor protein, spoJ) were identified in the Ruminococcus and Butyrivibrio genomes. This may indicate an ongoing genetic loss in response to adaptation to a new environment (rumen) where sporulation is no longer offering an advantage in fitness. The other identified shared genes are likely to be present in most of the other microbial genomes analyzed, and thus would contribute to higher-level genome clustering.

In addition, the ORFeomes were analyzed for predicted genes which are conserved in one genome cluster but not in the other—and vice versa (threshold conserved: 1e-100; threshold unique: 1e-10). Overall, 84 ORFs were identified to be group specific. Seventy of these were found only in genomes assigned to the Bacillus cluster and 14 in genomes in the Clostridium group. Remarkably, genes involved in heme and cytochrome biogenesis (hemE, hemH, hemY, resB, and resC), cytochrome reduction (gerB, gerC, cypD), and cytochrome oxidation (qoxB, cydA, ctaB, ctaD, cydA) were identified, indicating a Bacillus-specific electron transport chain. It is thus tempting to speculate that, functionally, Bacilli are aerobic Clostridia, having acquired the capability of oxidative phosphorylation. Furthermore, a subset of the propionate metabolism pathways identified to be Bacillus-specific. This subset is involved in the conversion of propanoyl-CoA to succinate and succinyl-CoA (pprD, pprB, pccB, succD, succC) and might present an additional energy conversion option for Bacilli which is absent in Clostridia.

Only a few ORFs were identified to be Clostridium cluster specific. Among those a cobyric acid synthase cobQ was identified to be Clostridium specific. CobQ is part of the porphyrin metabolic pathway, involved in converting cobyrinic acid into coenzyme vitamin B12. Notably, a central branching point in this pathway leads to the synthesis of hemes and cytochromes found in Bacilli (see above). Interestingly, CobQ is also absent in the Ruminococcus subcluster, providing further support for the proposed ongoing adaptation to the new rumen environment. In contrast, the Ruminococcus subcluster acquired a number of membrane and sugar utilization (e.g., beta-glucoisidases bgl3A, bgl3B, and bgl3D and an 1-fucose isomerase) which may aid in the adhesion to and degradation of plant fibers in the rumen (threshold conserved: 1e-60; threshold unique: 1e-10; Kelly et al., 2010).

While the function of these genes has been well described in the past, they deliver the proof-of-concept that FGD analyses are able to identify gene sets involved in lifestyle adaptation processes. Because the initial similarity analysis does not rely on existing gene annotation, uncharacterized ORFs (e.g., genes annotated as “conserved hypothetical”) can be identified as potential targets to contributing to respective phenotypes. This is particular important for poorly annotated microbial genomes with a high level of conserved hypothetical ORFs.

In summary, results obtained from the test dataset provide strong support for the usefulness of FGD analyses, by illustrating the ability of the method to draw together groups into common nodes based on shared core (shared by all genomes in a specific cluster) and lifestyle elements, yet distinguishing them into distinct sub-clusters based on relevant genotypic differences and lifestyle adaptation processes. Importantly, FGD subsequently allows identifying gene sets likely to be responsible for the observed clustering, providing meaningful new target selections for functional genomics analyses independent of other means of classification or prior annotation.

Furthermore, distinct placements in genome clustering of Leuconostoc mesenteroides and Oenococcus oeni (Ze-Ze et al., 2000; Makarova et al., 2006; Leuconostocaceae) were observed in the FDT. Both genera can be found epiphytically on fruits, fruit
mashes, and vegetables and are used in industrial and food fermentation processes. Interestingly, Functional Distribution placed *L. casei* and *L. plantarum* into a separate cluster, branching deeply within the Bacillii (*Figure A1* in Appendix). Both genomes are significantly larger than the average *Lactobacillus* genome of 1.8–2.0 Mbp with 2.8 and 3.3 Mbp, respectively. The significantly increased genome size likely reflects a more generalized lifestyle, capable of thriving in a variety of habitats such as raw and fermented dairy products, plants, and the intestinal and reproductive tracts of animals and humans. In contrast, the smaller genomes of other lactobacilli often reflect the more specialized lifestyle to one habitat such as the human or animal gastrointestinal system or specific fermentation processes.

These results may indicate that lifestyle adaptation can lead to a similar genetic makeup of taxa defined as being distinctively different by heritage based phylogeny.

**LIMITS OF THE FGD RESOLUTION**

In the previous example microbial genomes from a wider range of species were investigated. To determine if the algorithm can discriminate strains from the same species, 23 *Chlamydia trachomatis* genomes (host: human), three *Chlamydia muridarum* genomes (host: members of the family Muridae), and one *Chlamydia pneumoniae* genome (host: varied; see *Table A3* in Appendix) were subjected to an FGD analysis. It is interesting to note that until 1999 *C. muridarum*, infecting only members of the family Muridae, was designated as *C. trachomatis* (Everett et al., 1999). *C. pneumoniae*, which can infect a wide variety of different hosts and causes atypical pneumonia, clusters away from both *C. trachomatis* and *C. muridarum* genomes, indicating a different – and possible more flexible – genome makeup (*Figure A2* in Appendix). In contrast to the other *Chlamydia* genomes, *C. pneumoniae* harbors an additional ∼200 kb of genetic information, and a more detailed analysis will be necessary to determine whether gene loss or gene acquisition is the major driving force. Similarly, the *C. muridarum* group is clearly forming its own cluster, albeit indicating a higher-level of similarity to *C. trachomatis* than to *C. pneumoniae*. On this high level, FGD can clearly resolve genomics differences and support observed host specificities (varied hosts – Muridae – human). Within the *C. trachomatis* cluster, three distinct sub-clusters could be identified with little reshuffling observed. Cluster 1 comprises serotypes E, T, I, J and K (and *C. trachomatis* Ds2923), Cluster 2 harbors serotypes A, B, D, and L (and *C. trachomatis* E Sweden2), and Cluster 3 groups two serotype L and one serotype A (*Figure A2* in Appendix). The observed positioning of the serovar D strain *C. trachomatis* Ds2923 into Cluster 1 supports the pairwise alignment of several chlamydial isolates (Jeffrey et al., 2010) which identified the least number of nucleotide substitutions between Ds2926 and E/11023. In contrast, different major groups were identified in this analysis on amino-acid level, a maximum of 20 similar observations. Again, FDTs were approximated and evaluated for each resulting subset (data not shown). In summary, tree topology remained stable and only a swapping of neighboring branches was observed while in no instance shuffling was found for individual genome entries between genome clusters in either Jackknife analysis. This strongly supports the overall stability and discriminatory power of the FGD analysis.

Similarly, a Jackknife analysis was performed for the narrow strain FGD analysis (*Figure A2* in Appendix). Individual entries, individual clusters, and complete serotypes were removed and the resulting tree topology investigated. Removal of individual entries and complete clusters did not change the topology of the FGD tree. Only minor reshuffling was observed within a respective cluster when complete serotypes were removed from the analyses (e.g., removal of serotype L caused a repositioning of *C. trachomatis* D UW3 CX into the serotype D subcluster within Cluster 2.

As expected, the removal of individual entries reduces discriminatory power, resulting in grouping together previously more separated genome clusters, without any entry-reshuffling.

**COMPARATIVE ANALYSIS USING THE ARTEMIS COMPARISON TOOL**

Comparison of the degree of genome similarity between two or more genomes relies on analyses of presence/absence of genes and their respective syntenies in an operon or genome-wide context. In combination with ORFeome based (ORF-to-ORF comparison on amino-acid level, a maximum of 20 similarity hits per ORF is permitted) MSPcrunch (Sonnhammer and Durbin, 1994) comparison format data files (pMSPs) created by the compACTor software and annotated GenBank files, the ACT (Carver et al., 2005) provides an excellent visualization platform for mobility through the entire genome. *Figure 3* illustrates the differences and similarities found between closely related strains such as between members of the acidophilus-complex of the Lactobacilli or between subspecies as well as between less similar genomes of more distantly related microbes. For example, small scale [inversion of a specific gene locus between
FIGURE 3 | ORFeome based comparative ACT visualization of 11 Lactobacillus genomes. Based on the distribution observed in Figure 2, 11 Lactobacillus genomes and their ORFeome similarities were visualized in ACT using pMSP-datafiles. Respective genome designations are indicated on the left hand side of each genome line. Genomes are shown in full and drawn to scale. Genomic nucleotide sequences are represented by gray lines indicating sense and anti-sense strands and position markers are shown in between. Predicted ORFs are shown on each strand in their respective orientation as arrowed boxes. Direct amino-acid similarity between individual ORFs of neighboring genomes are shown as red lines, inverted similarities are indicated by blue lines. Color shadings indicate the level of similarity, the more saturated a similarity line the more conserved are two ORF-pairs. A trust level value of 40 was employed as display threshold to visualize similarity hits below an e-value of $1 \times 10^{-60}$.

Both L. delbrueckii ssp. bulgaricus genomes (Makarova et al., 2006; van de Guchte et al., 2006) and large scale (double inversion $\sim700$ and $\sim150$ kb, respectively) of the terminus of DNA replication between L. gasseri (Azcarate-Peril et al., 2008)
and *L. johnsonii* (Pridmore et al., 2004) genome inversions, deletion, and insertion events between *L. acidophilus* (Altermann et al., 2005) and *L. delbrueckii* ssp. *bulgaricus*, localized gene synteny between *L. delbrueckii* ssp. *bulgaricus* and *L. brevis*, general presence/absence of individual genes and larger synteny trends between *L. casei* and *L. plantarum* can be immediately identified.

**COMPARISON OF THE FGD ALGORITHM TO ALTERNATIVE METHODS**

Although similar algorithms exist to investigate phylogenetic relationships based on whole (or partial) genomes sequences (Snel et al., 1999; Wolf et al., 2001; Henz et al., 2005; Khiripet, 2005; Canchaya et al., 2006; Fuchsman and Rocap, 2006; Berger et al., 2007; Felis and Dellaglio, 2007; Blaiotta et al., 2008), their focus remains mostly to infer a heritage based phylogeny. Furthermore, often only subsets of ORFeomes are chosen for these analyses. These are then analyzed individually (with or without weighting) or as concatenated sequences. Thus an artificial restriction is introduced that may bias the analysis. An example of such a method has been published by Konstantinidis and Tiedje (2005). There, genome information to infer taxonomy of prokaryotes is used, calculating an average amino-acid identity of shared gene subsets. Only few publications investigate the functional relationship of microbial genomes, such as the development of the Blast Score Ratio which analyzes the complete ORFeome but is limited to three genomes at a time (Rasko et al., 2005). Other methods investigating the functional relationship and similarities between gene clusters have been used to address the problem of genomes with different sizes. One example of such a method, GRAST, explores the ongoing genome reductions and rearrangements by identifying clusters of functionally related genes (Toft and Fares, 2006). Subsets of orthogonal gene pairs are identified to determine conserved genetic loci. Similar to Blast Score Ratio analyses, the number of input genomes is limited to two at a time. While the output in part shows similarity to FGD analyses (visual representation of a genome plot and the determination of common and non-common genes), the purpose of this method is distinctly different in specifically identifying genome plasticity trends. A combination of genome analysis and visualization tool, GeneComp, has been published earlier (Yang et al., 2003). GeneComp is able to use different BLAST flavors and then visualize the textual output with varying levels of alignment length stringencies. While this solution offers the advantage of providing a combined analysis and visualization package, a number of limitations exist when compared to FGD. Like BSR and GRAST, GeneComp is restricted to a maximum number of three genomes. Furthermore, the algorithm is sequence based, highlighting genome variations such as repeat regions, insertions, deletions, and rearrangements rather than specific similarities to predicted ORFs.

Non-sequence based methods such as MLST use only a relatively small number of conserved genome loci with the primary aim to establish a highly discriminating (microbial) typing system (Chan et al., 2001; Maiden, 2006; Diancourt et al., 2007). Noteworthy, the ability of FGD analyses to identify cluster-conserved gene sets may provide a high-quality starting point for the selection of MLST targets.

**CONCLUSION**

16S rRNA and other gene subset analyses mainly focus on the determination of the line of descendants of a given gene or organism (Zhang et al., 2009) or on the identification of protein families (Enright and Ouzounis, 2000; Enright et al., 2002; Kelil et al., 2007). Such phylogenetic studies aim to reconstruct the relationship between organisms and are paramount to analyze the (changing) community structures of complex biological ecosystems. While this type of phylogenetic analysis is well accepted and widely used, it does not reflect the respective comprehensive genotypes. In contrast, FGD provides a different view of microbial similarities to each other. The example data set demonstrated the effects of lifestyle adaptation on genome content. FGD has shown the potential to provide new insights into the relationships between microbes from a comparative genomics perspective. The algorithm has already been used in a variety of analyses ranging from microbial (Goh et al., 2011) and archaic (Leahy et al., 2010) genomes to bacteriophage (Lu et al., 2010) which describe the impact of FGD analyses in more detail within their respective scopes.

Functional genome distribution in combination with the graphical visualization in ACT using ORFeome distance files (pMSPs) and functionally annotated GenBank files offers a powerful tool for comparative genomics that allows comparisons of whole genomes within genome space, encompassing heritage based (vertical transmission), lateral gene transfer (HGT), and lifestyle-driven change (adaptation) in a single analysis. Therefore, rather than attempting to reconstruct the evolution of the core genome with its set of commonly shared genes, FGD allows a representation of whole genome similarity at the functional level.

It will be possible to add further functionality to the algorithm, such as the ability to mask defined or dynamically created gene clusters within groups of organisms, thus identifying potentially important genetic elements independent from otherwise overpowering gene sets, such as central house-keeping or metabolism genes.

**AVAILABILITY**

The complete software suite consisting of the compACTor and the FGDfinder software is freely available upon email request for academic use. Commercial use is subject to a license agreement.

**ACKNOWLEDGMENTS**

Special thanks go to Peter Janssen for helpful and critical discussions during the development of the FGD algorithm, the subsequent interpretation of the data and the critical review of the manuscript. I also thank Sinead Leahy, Graeme Attwood, and Bill Kelly for their scientific contributions to this manuscript. This work was funded by the New Economic Research Fund programs C10 × 0314 and PROJ-13812-LNNR-AGR and supported by the Pastoral Greenhouse Gas Research Consortium (P&GRC).

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Evolutionary_and_Genomic_Microbiology/10.3389/fmicb.2012.00048/abstract
REFERENCES

Altermann, E., and Klaenhammer, T. R. (2003). GAMOLA: a new local solution for sequence annotation and analyzing draft and finished prokaryotic genomes. OMICS 7, 161–169.

Altermann, E., Russell, W. M, Arczate-Peril, M. A., Barrangou, R., Buck, B. L., Mcauliffe, O., Souther, N., Dobson, A., Duong, T., Callanan, M., Lick, S., Hamrick, A., Cano, R., and Klaenhammer, T. R. (2005). Complete genome sequence of the probiotic lactic acid bacterium Lactobacillus acidophilus NCFM. Proc. Natl. Acad. Sci. U.S.A. 102, 3906–3912.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Arczate-Peril, M. A., Altermann, E., Carver, T. J., Rutherford, K. M., Berrier, M., Argirov, K., Arigoni, F., and Brussow, H. (2007). Analysis of the genome sequence of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J. Mol. Microbiol. Biotechnol. 4, 453–461.

Desiere, F., Mchsan, W. M., Van Sinderen, D., Ferretti, I. J., and Brussow, H. (2001). Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic Streptococci: evolutionary implications for prophage-host interactions. Virology 288, 325–341.

Diancourt, L., Passet, V., Cheruvaxa, C., Garault, P., Smokvina, T., and Brisse, S. (2007). Multilocus sequence typing of Lactococcus casei (L. para-casei) reveals a clonal population structure with low levels of homolo-
gous recombination. Appl. Environ. Microbiol. 73, 6601–6611.

Ensink, A. J., and Ouzounis, C. A. (2000). GeneRAGE: a robust algo-
rithm for sequence clustering and domain detection. Bioinformatics 16, 451–457.

Ensink, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584.

Everett, K. D., Bush, R. M., and Andersen, A. A. (1999). Emended description of the order Chlamydia-
ales, proposal of Porachlamydiaceae fam. nov. and Simkaniacae fam. nov., each containing one mono-
typic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. Int. J. Syst. Bacteriol. 49(Pt 2), 415–440.

Fels, G. E., and Dellaglio, F. (2007). Taxonomy of Lactobacilli and Bifidobacteria. Curr. Issues Intest. Micro-
biota. 4, 44–61.

Fuchsmann, C. A., and Rocap, G. (2006). Whole-genome reciprocal BLAST analysis reveals that plantcommutes do not share an unusually large number of genes with eukarya and archaea. Appl. Environ. Microbiol. 72, 6841–6844.

Goh, Y., Goto, C., O’Hallerty, S., Alterm-
mann, E., and Hutkins, R. (2011). Specialized adaptation of a lactic acid bacterium to the milk environ-
ment: the comparative genomics of Streptococcus thermophilus LMD-9. Microb. Cell Fact. 10, 522.

Henz, S. R., Huson, D. H., Auch, A. F., Nieselt-Struwe, K., and Schuster, S. C. (2005). Whole-genome prokary-
otic phylogeny. Bioinformatics 21, 2329–2335.

James, F. C., and McCulloch, C. E. (1990). Multivariate analysis in ecol-
ogy and systematics: panacea or Pan-
dora’s box? Annu. Rev. Ecol. Syst. 21, 129–166.

Jeffrey, B. M., Suchland, R. J., Quinn, K. L., Davidson, J. R., Stamm, W. E., and Rockey, D. D. (2010). Genome sequencing of recent clinical Chlamydia trachomatis strains identifies loci associated with tissue tropism and regions of apparent recombination. Infect. Immun. 78, 2544–2553.

Kanehisa, M. (2002). The KEGG database. NAR. 30, 44–46.

Kanehisa, M., Araki, M., Goto, S., Hat-
tori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tomisaka, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480–D484.

Kelil, A., Wang, S., Brzezinski, R., and Fleury, A. (2007). CLUSS: clustering of protein sequences based on a new similarity measure. BMC Bioinformatics 8, 266. doi:10.1186/1471-2105-8-266.

Kelly, W. J., Leaby, S. C., Altermann, E., Yee-
mann, C. J., Dunne, J. C., Kong, Z., Pacheco, D. M., Li, D., Noel, S. J., Moon, C. D., Cookson, A. L., and Artwood, G. T. (2010). The glycosome of the rumen bacterium Butyrivirgatus proteoblas-
tes B31(T) highlights adaptation to a polysaccharide-rich envi-
ronment. PLoS ONE 5, e11942. doi:10.1371/journal.pone.0011942.

Khrapov, N. (2005). “Bacterial whole genome phylogeny using proteome comparison and optimal reversal distance,” in Computational Systems Bioinformatics Conference, Stanford.

Konstantinidis, K. T., and Tiedje, J. M. (2005). Towards a genome-based taxonomy for prokaryotes. J. Bacteri-
ol. 187, 6258–6264.

Leaby, S. C., Kelly, W. J., Alterm-
mann, E., Ronimus, R. S., Yeo-
mann, C. J., Pacheco, D. M., Li, D., Kong, Z., Metwali, S., Sang, C., Lambre, S. C. Jansen, P. H., Dey, D., and Attwood, G. T. (2010). The genome sequence of the rumen methanogen Methanobrev-
ibacter ruminantium reveals new possibilities for controlling rumi-
nant methane emissions. PLoS ONE 5, e9826. doi:10.1371/jour-
nal.pone.0008926.

Lu, Z., Altermann, E., Breidt, F., and Kozyrakis, S. (2010). Sequence analysis of Leucoacnic mesenteroides bacteriae bacteriophage [Phi]1-A4 isolated from an industrial vegetable fermenta-
tion. Appl. Environ. Microbiol. 76, 1955–1966.

Maiden, M. C. J. (2006). Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 60, 561–588.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamy-
chev, V., Polouchine, N., Shakhova, V., Grigoriev, I., Leon, Y., Rohrsak, D., Lucas, S., Haug, K., Good-
stein, D. M., Hawkins, T., Pengs
dhya, Y., Welker, D., Hughes, J., Goh, Y., Benson, A., Baldwin, K., Lee, J. H., Diaz-Muniz, I., Dosti, B., Smeianov, V., Wechter, W., Barabote, R., Lorca, G., Altermann, E., Barran-
gou, R., Ganesan, B., Xie, Y., Raw-
sthorne, H., Tamir, D., Parker, C., Breidt, F., Broadbent, J., Hutkins, R., O’Sullivan, D., Steele, I., Unlu, G., Saier, M., Klaenhammer, T., Richard-
son, P., Kozyrakis, S., Weimer, B., and Mills, D. (2006). Comparative genomics of the lactic acid bacte-ia. Proc. Natl. Acad. Sci. U.S.A. 103, 15611–15616.

Makarova, K. S., and Koonin, E. V. (2007). Evolutionary genomics of lactic acid bacteria. J. Bacteriol. 189, 1199–1208.

McCutcheon, J. P., and von Dohlen, C. D. (2011). An interdependent meta-
bolic patchwork in the nested symbi-
omy of myxobacteria. Curr. Biol. 21, 1366–1372.

Nicolas, P., Bessieres, P., Ehrlich, S. D., Maguin, E., and Van De Guchte, M. (2007). Extensive hor-
izontal transfer of core genome genes between two Lactobacillus species found in the gastrointesti-
nal tract. BMC Evol. Biol. 7, 141. doi:10.1186/1471-2148-7-141.

Pridmore, R. D., Berger, B., Desiere, F., Villanova, D., Barretto, C., Pit-
tet, A. C., Zwahlen, M. C., Rou-
yet, M., Tran, T. T., Motellet, R., Mctavish, S., Klaen-
hammer, T., Arigoni, F., and Schell, M. A. (2004). The genome sequence of the probiotic intestinal bacterium Lactobacillus johnsonii NCC 533.
Altermann Functional genome distribution

Frontiers in Microbiology

Rohwer, F., and Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. J. Bacteriol. 184, 4529–4535.

Schneiter, S., Perlov, O., Kaiser, O., Gerth, K., Alich, A., Altmeier, M. O., Bartels, D., Bode, B., Beyer, S., Bode, E., Bod, H. B., Bolten, C. J., Choudhuri, J. V., Doss, E., Elnakady, Y. A., Frank, B., Gaigalat, L., Goesmann, A., Groeger, C., Gross, F., Jelsbak, L., Kalinowski, J., Kegler, C., Krause, L., Knauber, T., Konietzny, S., Kopp, M., Krause, L., Krug, D., Linke, B., Mahmud, T., Martinez-Arias, R., Mehardy, A. C., Mera, M., Meyer, F., Mornmann, S., Munoz-Dorado, J., Perez, J., Pradella, S., Rachid, S., Rasko, D. A., Myers, G. S., and Ravel, J. (2007). Complete genome sequence of the myxobacterium, K. J., Welch, R. D., Wenzel, S. C., Suen, G., Treuner-Lange, A., Velicer, G. J., Vorholter, F. J., Weissman, K. J., Welch, R. D., Wenzel, S. C., Whitworth, D. E., Wilhelm, S., Wittmann, C., Blocker, H., Puhler, A., and Muller, B. (2007). Complete genome sequence of the mycobacterium Sorangium cellulosum. Nat. Biotechnol. 25, 1281–1289.

Schwiertz, A., Hold, G. L., Duncan, S. H., Grubh, B., Collins, M. D., Lawton, P. A., Flint, H. J., and Blaut, M. (2002). Anaerostipes haumii gen. nov., sp. nov., a new saccharolytic, acetate-utilising, butyrate-producing bacterium from human faeces. Syst. Appl. Microbiol. 25, 46–51.

Sneath, P. H., and Sokal, R. R. (1962). Numerical taxonomy. Nature 193, 855–860.

Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.

Sonnhammer, E. L., and Durbin, R. (1994). A workbench for large-scale sequence homology analysis. Comput. Appl. Biosci. 10, 301–307.

Sousa, C., De Lorenzo, V., and Cebolla, A. (1997). Modulation of gene expression through chromosomal positioning in Escherichia coli. Microbiology 143( Pt 6), 2071–2078.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Tatusov, R., Fedorova, N., Jackson, D., Jacobs, A., Kropotkin, B., Koonin, E., Krylov, D., Mazumder, R., Melnikov, S., Nikolskaya, A., Rao, B. S., Smirnov, S., Svedlov, A., Vasudevan, S., Wolf, Y., Yin, J., and Natale, D. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41. doi:10.1186/1471-2105-4-41

Thomson, N., Bentley, S., Holden, M., and Parkhill, J. (2003). Fitting the niche by genomic adaptation. Nat. Rev. Microbiol. 1, 92–93.

Toft, C., and Fares, M. A. (2006). GRAST: a new way of genome reduction analysis using comparative genomics. Bioinformatics 22, 1551–1561.

van de Gucht, M., Penaud, S., Grimaldi, C., Barbe, V., Bryson, K., Nicolau, P., Robert, C., Ota, S., Mangenot, S., Coudoulx, A., Loux, V., Dervyn, R., Bossy, R., Bolotin, A., Batto, J. M., Walunas, T., Gibrat, J. F., Bessieres, P., Weissbach, J., Ehrlich, S. D., and Maguin, E. (2006). The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution. Proc. Natl. Acad. Sci. U.S.A. 103, 9274–9279.

Wolf, Y., Rogozin, I., Grishin, N., Tatusov, R., and Koonin, E. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1, 8. doi:10.1186/1471-2148-1-8

Yang, J., Wang, J., Yao, Z.-J., Jin, Q., Shen, Y., and Chen, R. (2003). GenomeComp: a visualization tool for microbial genome comparison. J. Microbiol. Methods 54, 423–426.

Ze-Ze, L., Tenreiro, R., and Paveia, H. (2008). The Oenococcus oeni genome: physical and genetic mapping of strain GM and comparison with the genome of a ‘divergent’ strain, PSU-1. Microbiology 146(Pt 12), 3195–3204.

Zhang, H., Zhong, Y., Hao, B., and Gu, X. (2009). A simple method for phylogenomic inference using the information of gene content of genomes. Gene 441, 163–168.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 08 November 2011; accepted: 30 January 2012; published online: 21 February 2012.

Citation: Altermann E (2012) Tracking lifestyle adaptation in prokaryotic genomes. Front. Microbiol. 3:48. doi: 10.3389/fmicb.2012.00048

This article was submitted to Frontiers in Evolutionary and Genomic Microbiology, a specialty of Frontiers in Microbiology. Copyright © 2012 Altermann. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.
APPENDIX

FIGURE A1 | Functional distribution tree of 17 closely related taxa within the class Bacilli. The tree represents a subset of the one shown in Figure 2. Predicted ORFeomes of all genomes were subjected to an FGD analysis and the resulting distance matrix was imported into MEGA4. The functional distribution was visualized using the UPGMA method (32). The tree is drawn to scale, with branch lengths in the same units as those of the functional distances used to infer the distribution tree.
FIGURE A2 | A functional distribution tree comprising of 23 Chlamydia trachomatis genomes (host: human), three C. muridarum genomes (host: members of the family Muridae), and one C. pneumoniae genome were used to investigate genome similarities and the FGD-power-of-resolution on strain level. Entries in red depict Chlamydia trachomatis serotypes A–C (trachoma), entries in black represent serotypes D–K (sexually transmitted pathovars) and entries in green show serotype LGV (L1–L3; lymphogranuloma venereum). Chlamydia muridarum entries are shown in blue and Chlamydia pneumoniae is depicted in gray. Functional clusters and subclusters are indicated by square brackets.
### Table A1 | e-Value range based trust levels.

| e-Value range                          | Trust value |
|----------------------------------------|-------------|
| <0.1                                   | 0           |
| 0.1 ≤ e-value > 1e−10                  | 1           |
| 1e−10 ≤ e-value > 1e−40                | 10          |
| 1e−40 ≤ e-value > 1e−50                | 20          |
| 1e−50 ≤ e-value > 1e−60                | 30          |
| 1e−60 ≤ e-value > 1e−70                | 40          |
| 1e−70 ≤ e-value > 1e−80                | 50          |
| 1e−80 ≤ e-value > 1e−90                | 60          |
| 1e−90 ≤ e-value > 1e−100               | 70          |
| 1e−100 ≤ e-value > 1e−110              | 80          |
| 1e−110 ≤ e-value > 1e−120              | 85          |
| 1e−120 ≤ e-value > 1e−130              | 90          |
| 1e−130 ≤ e-value > 1e−160              | 95          |
| 1e−160 ≤ e-value ≥ 0                   | 100         |
Table A2 | Genomes used for assessment of functional genome distribution.

| Designation* | Domain/family | Genome size [bp] | ORFeome size | Accession number |
|--------------|---------------|------------------|-------------|-----------------|
| Lactobacillus plantarum WCFS1 | Bacteria/Lactobacillaceae | 330827 | 3051 | AL935263 |
| Lactobacillus brevis ATCC 367 | Bacteria/Lactobacillaceae | 2291220 | 2314 | CP000416 |
| Pediococcus pentosaceus ATCC 25745 | Bacteria/Lactobacillaceae | 1832387 | 1847 | NC_008525 |
| Lactobacillus sakei subsp. sakei 23K | Bacteria/Lactobacillaceae | 1884661 | 1886 | CR936503 |
| Lactobacillus casei ATCC 334 | Bacteria/Lactobacillaceae | 2895264 | 2909 | CP000423 |
| Lactobacillus salivarius UCC118 | Bacteria/Lactobacillaceae | 1827111 | 1738 | CP000233 |
| Lactobacillus reuteri F275 | Bacteria/Lactobacillaceae | 1999618 | 1944 | CP000705 |
| Lactobacillus johnsonii NCC 533 | Bacteria/Lactobacillaceae | 1992676 | 1857 | AE017198 |
| Lactobacillus gasseri ATCC 33323 | Bacteria/Lactobacillaceae | 1894360 | 1811 | CP000413 |
| Lactobacillus acidophilus NCFM | Bacteria/Lactobacillaceae | 1993561 | 1979 | CP000033 |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 | Bacteria/Lactobacillaceae | 1856951 | 2040 | CP000412 |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365 | Bacteria/Lactobacillaceae | 1864998 | 2218 | CR954253 |
| Bacillus cereus ATCC 14579 | Bacteria/Bacillaceae | 5411109 | 5490 | AE016871 |
| Bacillus thuringiensis serovar konkukian str. 97-27 | Bacteria/Bacillaceae | 5237682 | 5168 | AE017355 |
| Bacillus pumilus SAFR-032 | Bacteria/Bacillaceae | 3704465 | 3737 | CP000613 |
| Bacillus licheniformis ATCC 14580 | Bacteria/Bacillaceae | 4222645 | 4379 | AE017333.1 |
| Bacillus subtilis subsp. subtilis 168 | Bacteria/Bacillaceae | 4214630 | 4106 | AL009126 |
| Streptococcus pneumoniae R6 | Bacteria/Streptococcaceae | 2039615 | 2046 | NC_003098 |
| Streptococcus pyogenes SSI-1 | Bacteria/Streptococcaceae | 1894287 | 1861 | BA000034 |
| Streptococcus thermophilus LMD-9 | Bacteria/Streptococcaceae | 1856368 | 2003 | CP000419 |
| Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293 | Bacteria/Leuconostocaceae | 2038396 | 2073 | NC_005311 |
| Oenococcus oeni PSU-1 | Bacteria/Leuconostocaceae | 1780517 | 1864 | NC_005282 |
| Clostridium perfringens ATCC 13124 | Bacteria/Clostridiaceae | 3256803 | 3107 | CP000246 |
| Clostridium botulinum ATCC 3502 | Bacteria/Clostridiaceae | 3886916 | 3648 | AM413317 |
| Clostridium kluyveri DSM 555 | Bacteria/Clostridiaceae | 3964618 | 3926 | CP000673 |
| Clostridium difficile ATCC 9869 | Bacteria/Clostridiaceae | 4290252 | 4390 | AM190355 |
| Ruminococcus obeum ATCC 29174 | Bacteria/Lachnospiraceae | 3626304 | 4176 | AAAB000000000; draft |
| Anaerostipes caccae L1-92 | Bacteria/Lachnospiraceae | 1691947 | 1582 | ABA09000; draft |
| Ruminococcus gravis ATCC 29149 | Bacteria/Lachnospiraceae | 3501963 | 3913 | AAYG00000000; draft |
| Butyrivibrio proteoclasticus B316 | Bacteria/Lachnospiraceae | 3936787 | 3477 | Unpublished draft |
| Pseudomonas aeruginosa PA7 | Bacteria/Pseudomonadaceae | 6583339 | 6371 | CP000744 |
| Escherichia coli O157:H7 | Bacteria/Enterobacteriaceae | 5528445 | 6006 | AE005174 |
| Escherichia coli K12 K-12 | Bacteria/Enterobacteriaceae | 463975 | 4403 | NC_000913 |
| Methanococcales burtionii DSM 6242 | Archaea/Methanosarcinaceae | 2575032 | 2446 | NC_007955 |
| Methanosarcina mazei Goe1 | Archaea/Methanosarcinaceae | 4096345 | 3371 | NC_000901 |
| Methanocaldococcus janaschii DSM 2661 | Archaea/Methanocaldococcaceae | 1664970 | 1862 | NC_000909 |
| Methanospirillum stadtmanae DSM 3091 | Archaea/Methanobacteriaceae | 1767403 | 1588 | NC_007861 |
| Methanobrevibacter smithii ATCC 35061 | Archaea/Methanobacteriaceae | 1853160 | 1795 | NC_009615 |
| Methanothermobacter thermotrophicus DeltaH | Archaea/Methanobacteriaceae | 1751377 | 1918 | NC_009116 |

*Color coding used in the table corresponds to the color scheme shown in Figure 2 and Figure A1.
Table A3 | Genomes used for assessment of functional genome distribution on strain level.

| Designation* | Serotype | Host | Genome status | Accession number |
|--------------|----------|------|---------------|------------------|
| Chlamydia trachomatis A2497 | A | Human | Complete | 347974781 |
| Chlamydia trachomatis A HAR-13 | A | Human | Complete | 76788711 |
| Chlamydia trachomatis B TZ1A828 | B | Human | Complete | 231272648 |
| OT Chlamydia trachomatis B Jali20 OT | B | Human | Complete | 231273667 |
| Chlamydia trachomatis D UW3 CX | D | Human | Complete | 15604717 |
| Chlamydia trachomatis D-LC | D | Human | Complete | 297749010 |
| Chlamydia trachomatis D-EC | D | Human | Complete | 297748130 |
| Chlamydia trachomatis Ds2923 | D | Human | Complete | 22356784 |
| Chlamydia trachomatis E Sweden2 | E | Human | Complete | 289525045 |
| Chlamydia trachomatis E 150 | E | Human | Complete | 298434583 |
| Chlamydia trachomatis E 11023 | E | Human | Complete | 298438301 |
| Chlamydia trachomatis F 70 | F | Human | Complete | 22244350 |
| Chlamydia trachomatis F 70s | F | Human | Complete | 22244349 |
| Chlamydia trachomatis G 11074 | G | Human | Complete | 296437374 |
| Chlamydia trachomatis G 9301 | G | Human | Complete | 297139873 |
| Chlamydia trachomatis G 9768 | G | Human | Complete | 296435514 |
| Chlamydia trachomatis G 11222 | G | Human | Complete | 296436438 |
| Chlamydia trachomatis J 6276 | J | Human | Complete | 222444352 |
| Chlamydia trachomatis J 6276a | J | Human | Complete | 222444351 |
| Chlamydia trachomatis L2-434 Bu | L | Human | Complete | 166153973 |
| Chlamydia trachomatis L2b UCH1 proctitis | L | Human | Complete | 352981305 |
| Chlamydia trachomatis L2c | L | Human | Complete | 339625373 |
| Chlamydia trachomatis L2tet1 | L | Human | Complete | 301334996 |
| Chlamydia muridarum MopnTet14 draft | Muridae | Complete | 311788820 |
| Chlamydia muridarum Nigg | Muridae | Complete | 29337300 |
| Chlamydia muridarum Weiss.cb | Muridae | Draft | NC_002620.2 |
| Chlamydia pneumoniae | Varied | Complete | 340215159 |