GBE

Genome Biology and Evolution

SMBE

Distinguishing Microbial Genome Fragments Based on Their Composition: Evolutionary and Comparative Genomic Perspectives

Scott C. Perry, and Robert G. Beiko*

Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada

*Corresponding author: E-mail: beiko@cs.dal.ca.

Accepted: 19 January 2010 **Associate editor:** Emmanuelle Lerat

Abstract

It is well known that patterns of nucleotide composition vary within and among genomes, although the reasons why these variations exist are not completely understood. Between-genome compositional variation has been exploited to assign environmental shotgun sequences to their most likely originating genomes, whereas within-genome variation has been used to identify recently acquired genetic material such as pathogenicity islands. Recent sequence assignment techniques have achieved high levels of accuracy on artificial data sets, but the relative difficulty of distinguishing lineages with varying degrees of relatedness, and different types of genomic sequence, has not been examined in depth. We investigated the compositional differences in a set of 774 sequenced microbial genomes, finding rapid divergence among closely related genomes, but also convergence of compositional patterns among genomes with similar habitats. Support vector machines were then used to distinguish all pairs of genomes based on genome fragments 500 nucleotides in length. The nearly 300,000 accuracy scores obtained from these trials were used to construct general models of distinguishability versus taxonomic and compositional indices of genomic divergence. Unusual genome pairs were evident from their large residuals relative to the fitted model, and we identified several factors including genome reduction, putative lateral genetic transfer, and habitat convergence that influence the distinguishability of genomes. The positional, compositional, and functional context of a fragment within a genome has a strong influence on its likelihood of correct classification, but in a way that depends on the taxonomic and ecological similarity of the comparator genome.

Key words: genome composition, phylogenetic classification, support vector machines, metagenomics.

Introduction

Microbial genomes show dramatic differences in their underlying nucleotide compositions. The average G + C composition in sequenced prokaryotic genomes ranges from 16.6% in the reduced endosymbiont Candidatus *Carsonella ruddii* to nearly 75% in certain Proteobacteria and Actinobacteria. Properties such as oligomer nucleotide signatures (Blaisdell et al. 1986; Brendel et al. 1986; Pietrokovski et al. 1990; Karlin and Burge 1995; Abe et al. 2005), codon usage patterns (Willenbrock et al. 2006), conserved sequence repeats (van Belkum et al. 1998), and structural periodicity (Mrázek 2009) are variable and potentially characteristic of different taxonomic groups of microbes. Variation in these patterns has been tied to selective forces including nitrogen limitation in the environment (Willenbrock et al. 2006) and DNA repair systems (Paz et al. 2006; Rocha et al. 2006). Under certain conditions, these patterns and biases can change rapidly relative to changes in commonly used marker genes; for example, strains of the marine picocyanobacterium *Prochlorococcus marinus* show remarkable G + C content divergence from 30% and 50% despite the presence of very similar 16S rDNA sequences, which is likely due to differences in DNA repair genes (Rocap et al. 2003). The G + C content of these genomes correlates with adaptation to different degrees of light intensity, and the rapid genomic divergence may be tied to the rapid ecological divergence of these clades.

The role of ecology in shaping composition is not yet firmly established and indeed may depend on the type of habitat under consideration. Hypersaline environments impose significant physiological challenges on resident microbes, and there is evidence that taxonomically divergent

genomes show similar amino acid and other compositional biases (Paul et al. 2008). At the genomic level, halophile DNA composition appears to be strongly influenced by increased usage of aspartic and glutamic acid, threonine and valine codons, and possibly also the propensity to transition from B-DNA to a Z-DNA conformation that is stabilized at high salt concentrations (Misra and Honig 1996). Recently, Dick et al. (2009) examined the compositional patterns within a community of acidophilic organisms and found that major and minor lineages could be distinguished, suggesting that genome-specific processes were more important than ecological constraints in acid mine drainage communities. Strikingly, the G + C composition of different genomes from these communities ranged between 35% and 69%, clearly rejecting the idea of an acidophilic genome signature.

Different evolutionary forces give rise to within-genome compositional variation. "Translational" codon usage bias refers to the tendency of highly expressed genes to preferentially use synonymous codons that have high levels of corresponding tRNA in the cell. Genomes that show high levels of translational codon bias can therefore show different compositional patterns between highly expressed and other genes (Carbone et al. 2005). However, these biases do not manifest in all genomes, and in many cases, codon preference is instead driven by compositional or strand bias (Carbone et al. 2003). Exogenous sequences that have been integrated into a genome can also show substantial compositional variation; these variations underpin homology-independent "surrogate" methods for identifying lateral genetic transfer (LGT) events (Ragan 2002; Dufraigne et al. 2005). Bacteriophage genomes tend to be very A + T rich and lysogenized phage can be identified based on their unusual compositional patterns (Karlin 2001; van Passel et al. 2006). Compositional searches are also an essential component of many strategies used to find pathogenicity islands (Hsiao et al. 2003). Within-genome compositional variation can therefore highlight important classes of core and adaptive genes. However, in the context of assigning anonymous sequences to their originating genomes, these variations constitute confounding factors that may contribute to misattribution of certain genomic fragments.

With the rise of environmental shotgun sequencing, genomic patterns have gained prominence as cues by which short sequence reads can be assigned to the appropriate taxonomic unit, ideally strain or species, in a homology-independent way (Abe et al. 2003; McHardy et al. 2007; Manichanh et al. 2008). Sequence assignment is an essential step that precedes the inference of organism-specific regulatory, metabolic, or ecological information. Unsupervised learning approaches to sequence assignment have employed variations of Kohonen's self-organizing map (Abe et al. 2003; Martin et al. 2008; Dick et al. 2009) and statistical correlations between oligonucleotide usage patterns (Teeling et al. 2004). Oligomer (particularly tetranucleotide) profiles and signatures have been successfully used in supervised sequence assignment: prominent examples include a naïve Bayes approach (Sandberg et al. 2001) and PhyloPythia, which performs hierarchical classification of unknown sequences using a large array of support vector machines (SVMs) trained with different subsets of the original data (McHardy and Rigoutsos 2007). Hybrid approaches that use binning augmented with detected phylogenetic marker genes have also been applied to the problem (Chatterji et al. 2008).

The high reported accuracy of classification approaches is encouraging. However, because a great deal of ecological and genomic differentiation can be observed at the species or strain level among microbes (Welch et al. 2002; Rocap et al. 2003; Tettelin et al. 2005), it is essential to understand the taxonomic limitations of these methods, both in theory and in practice. When classification accuracy (CA) is less than 100%, it is important to distinguish fundamentally unclassifiable cases (e.g., recently acquired sequences and slowly evolving traits) from boundary cases that could be distinguished by using a better classifier. Such knowledge can establish or rule out the need for more complex classifiers and encoding schemes. A high global accuracy score achieved on a benchmark data set such as FAMeS (Mavromatis et al. 2007) can still conceal challenging cases that are poorly resolved. The compositional variation within and between genomes has been extensively documented (Bohlin et al. 2008; Mrázek 2009), but it is not clear what sort of challenge these patterns pose to modern classifiers such as SVMs. In this work, we use a relatively simple supervised classification scheme (SVMs trained with tetranucleotide compositional profiles of genomic fragments) to assess the dependence of CA on various measures of genomic similarity and relatedness. We also explore whether there exist compositional or functional subsets of genomes that may be easier or more difficult to distinguish.

Materials and Methods

Genome Sequence Acquisition and Preprocessing The sequences of 774 prokaryotic genomes (see supplementary table S1, Supplementary Material online) were obtained from The National Center for Biotechnology Information (NCBI) via rsync on 28 November, 2008. The 721 bacterial genomes covered a total of 472 unique named species, whereas 49 archaeal species were represented by the 53 available genomes. The average genome size was 3.58 Mbp, ranging from 0.16 Mbp (*C. ruddii*) to 13 Mbp (*Sorangium cellulosum*). The DNA molecules obtained were split into a set of substrings or fragments, each 500 nucleotides in length. We chose to examine fragments 500 nt in length for several reasons:

(i) Fragments of this length are known to be difficult to classify in comparison with sequences 1000 nt or

greater in length (McHardy et al. 2007). Existing binning methods include unsupervised approaches that use fragments of length 500 nt or greater (Sandberg et al. 2001; Abe et al. 2003; Teeling et al. 2004; McHardy et al. 2007; Diaz et al. 2009) and "semisupervised" approaches that can use fragments < 100 nt in length but rely on extrinsic information such as the presence of marker genes or Pfam domains (Chan et al. 2008; Krause et al. 2008).

(ii) These fragments are a realistic approximation of the sequence lengths generated by modern sequencing techniques used in environmental shotgun analysis.

(iii) Many of the forces that can cause compositional variation within a genome such as codon usage bias, introgression of foreign DNA, and differential selection on coding versus noncoding regions will be emphasized in short fragments but smoothed out in longer ones.

Each fragment was converted to a 256-element feature array F by counting the instances of each tetranucleotide X_i, $X = \{AAAA, AAAC, \ldots, TTTT\}$ in the fragment and converting F to an array of frequencies F' by dividing by the number of overlapping tetranucleotides (497 for a fragment of length 500 nt). For each extracted fragment, F' was used to train and test the SVMs. In the statistical literature, frequencies are often symmetrized by summing the counts of reverse complementary tetranucleotides and dividing by two (Karlin et al. 1994); in separate analyses, we performed this symmetrization on each F' to produce arrays of symmetrized frequencies S'. Finally, to test the impact of G + C content on frequencies and distinguishability, each element k of S' was converted to a nucleotide signature based on the underlying frequency of (symmetrized) G + C and A + T:

$$G_k' = \log_2\left(\frac{S_k'}{\frac{1}{2}(f_{k1}f_{k2}f_{k3}f_{k4})}\right),$$

where $f_{k1, \ldots, k4}$ are the symmetrized frequencies of each of the four mononucleotides constituting a given tetramer in the reference molecule (either the current 500-nt fragment or an entire genome, depending on the application; see below).

Other genome-associated data were retrieved as follows: 16S rDNA sequences were retrieved from the Ribosomal Database Project (RDP) Release 10.10 (Cole et al. 2008) by using the appropriate RefSeq accession query. A total of 749 genomes were covered in this fashion, with 47 genomes represented by two or more 16S sequences. The genetic distance between pairs of 16S genes was generated from the RDP reference alignment; in pairwise comparisons, where >1 16S gene was present in at least one genome, the average of all between-genome 16S comparisons was taken. Matrices for comparisons between Bacteria and Archaea were not available, and these comparisons were not included in the accuracy models. For the bacterial/archaeal pairs shown in supplementary table S2 (Supplementary Material online), 16S distances were computed from a ClustalW2 (Larkin et al. 2007) alignment of the sequences from RDP. Functional annotations of genes were obtained from the J. Craig Venter Institute Role Category database (Peterson et al. 2001). The predicted start and end positions of each gene were obtained from the NCBI files.

Analysis of Compositional Patterns For each genome, the average tetranucleotide composition was calculated by summing the individual tetramer counts across all fragments in the genome and then normalizing the tetramer counts by the number of fragments. For each pair of genomes, pairwise tetranucleotide Euclidean (PTE) distances were computed from the 256-element compositional vectors. To examine the compositional similarity among genomes, the 774×774 matrix of PTE distances between genomes was used to construct hierarchical clusters in a manner similar to distance-based genome phylogenies (Snel et al. 1999; Clarke et al. 2002). Similar computations were performed on the symmetrized and G + C-corrected symmetrized frequencies (using mononucleotide frequencies computed from the entire genome), yielding symmetrized PTE (SPTE) and G+C-corrected SPTE (GC-SPTE) distances, respectively. The unweighted pair group method with arithmetic mean (UPGMA) implementation in PHYLIP v3.67 (Felsenstein 1989) was used to infer a rooted tree from the distance matrix. The resulting tree was examined in a parsimony context to examine the congruence of compositional similarity-based clusters and with taxonomy. Given a set of N_t taxonomic units, a perfectly congruent clustering would yield the minimal $T_{min} = N_t - 1$ number of transitions on the tree. Clusterings that intermingled taxonomic units would increase the number of transitions T_{opt} on the tree. We used the consistency index (CI: Farris 1969) to express the fit of taxonomy to the UPGMA clustering: the CI is the ratio T_{min}/T_{opt}, with CI = 1.0 indicating a perfect fit. We evaluated the CI at several NCBI-defined taxonomic levels to assess the degree to which distinct phyla, class, order, genus, and species were split in the UPGMA clustering.

SVM Training and Testing We used the libSVM v2.88 implementation (http://www.csie.ntu.edu.tw/~cjlin/libsvm) on a computer cluster running Rocks Cluster software. We elected to use a simple kernel function for classification: Based on previous reported results (McHardy et al. 2007), we chose a Gaussian radial basis function kernel rather than a linear one. The values of SVM parameters C and γ were set heuristically for each pairwise genome comparison by performing an initial grid search using 500 feature vectors from each pair of genomes. SVM training and testing was performed with a 5-fold cross-validation approach. The set S over all frequency vectors F' was split into five subsets S_1, S_2, S_3, S_4, S_5 of approximately equal size, and each S_n used

as the test set for an SVM trained on the other four subsets. CA, corresponding to the proportion of test set cases that were correctly classified, was used as the principal indicator of the effectiveness of SVM training. In separate trials, the 299,151 pairwise comparisons were performed using raw and symmetrized frequencies. We additionally sampled 500 pairs of genomes for reanalysis using G + C-corrected, symmetrized frequencies: mononucleotide frequencies were computed on both a genome-by-genome and fragment-by-fragment basis.

Statistical Tests and Unsupervised Clustering All statistical tests reported in the manuscript were performed with version 2.8.1 of the R statistical package (http://www.r-project.org). To perform unsupervised clustering, the normalized tetranucleotide frequency vectors for each of the selected outlier genomes were independently clustered using the "kmeans" method provided by R, for $k \in \{2, 3, 4, 5, 6\}$. At each value of k, the $2k$ cluster assignments for each outlier pair were used to designate class labels in the corresponding SVM training file. In total, six SVM training files were generated for each outlier pair; one for each of the 5 values of k utilized in the k-means clustering step, plus a control case where no clustering was used (essentially, $k = 1$).

Grid searches were performed on 1,000-element subsets of each of the SVM training files in order to determine reasonable values of C and γ, and SVM models were subsequently trained and evaluated using 5-fold, leave-one-out cross-validation as previously described. Two CAs were recorded for each model: a strict CA in which correct classification was defined as the SVM's ability to correctly predict a given fragment's cluster assignment and a relaxed CA in which correct classification was defined as the SVM's ability to correctly predict a given fragment's source genome.

Heatmaps were generated to visualize and contrast the compositional profiles of different clusters of fragments. A heatmap is a matrix in which each row corresponds to a single fragment and each column represents a given tetranucleotide, with color intensity for a given matrix entry used to represent the relative frequency of the corresponding tetranucleotide in a given fragment. Full 256-column heatmaps contain a great deal of redundancy, so we opted to use a feature reduction strategy based on principal components analysis (PCA). We subjected the full matrix of frequency values for all fragments of a given genome to PCA using the "psych" package in the R statistical software (http://www.r-project.org), using the varimax rotation option to the principal() function and returning the ten components with the largest eigenvalues. Rather than showing the principal components, which obfuscate the input variables, we chose the tetranucleotide variable with the highest loading for each component for display, leading to heatmaps with a maximum of ten columns. It is important to note that while principal components are orthogonal, variables with high loadings on different components will not necessarily be so. Nonetheless, to the extent that different components capture different elements of composition, we expect many high-loading variables to be poorly correlated.

Results

Compositional Variation between Genomes The UPGMA tree constructed from the PTE matrix (fig. 1) intermingles taxonomic groups, as was previously observed with codon usage bias (Willenbrock et al. 2006) and in a recent hexanucleotide analysis by Bohlin et al. (2009). The deepest division in the tree was between the extremely reduced genome of the endosymbiont Candidatus *C. ruddii* (G + C content = 16.6%) and the other 773 genomes. Another deep branch subtends three genomes: two strains of *Thermus thermophilus* and the thermo- and radiotolerant Actinobacterium *Rubrobacter xylanophilus*. Interestingly, *Deinococcus radiodurans*, another radiotolerant organism which is a member of the same phylum as *Thermus*, associates with a different group of thermophiles including the gram-positive organisms *Symbiobacterium thermophilum* and *Thermobifida fusca*. Many groups in the tree appear to be split or aggregated based on lifestyle rather than genetic relatedness. For example, reduced intracellular organisms (both pathogen and endosymbiont) form several clusters in the tree: one such grouping (Group 1 in fig. 1) includes a subset of γ-proteobacterial insect endosymbionts, Tenericute pathogens, and Candidatus *Sulcia muelleri*, a member of phylum Bacteroidetes. The same grouping with the addition of *C. ruddii* was recovered in the SPTE cluster, whereas clustering based on GC-SPTE distances split this group into its constituent phyla (with the exception of *Sulcia muelleri*, which remained with the other insect endosymbionts). Phylum Aquificae, known for having genetic affinities with several other groups including ε-Proteobacteria and Thermotogae (Beiko et al. 2005; Boussau et al. 2008), is split into two: *Aquifex aeolicus* clusters with a set of Archaea and Thermotogae in Group 2a, whereas *Hydrogenobaculum* sp. YO4AAS1 and *Sulfurihydrogenibium* sp. YO3AOP1 are members of Group 2b, which also includes mesophilic ε-Proteobacteria, three members of genus *Thermoanaerobacter*, and thermophilic genomes from several other lineages. The GC-SPTE further disrupts the Aquificae, separating the two members of Group 2b. Interestingly, the apparent early-branching thermophile *Dictyoglomus thermophilum* loses its Group 2b affinities under the GC-SPTE clustering, instead grouping with *A. aeolicus*, *Metallosphaera sedula*, and several euryarchaeote genera including *Pyrococcus* and *Thermococcus*. The halophilic bacterium *Salinibacter ruber* groups with halophilic and methanogenic Archaea (Group 3), rather than its fellow Bacteroidetes; this grouping is consistent with earlier studies

showing compositional convergence and LGT between halophilic Bacteria and Archaea (Mongodin et al. 2005; Paul et al. 2008). The affinity of *S. ruber* for methanogenic Archaea remains in both the SPTE and GC-SPTE clustering.

Multiple genomes representing the same species formed cohesive clusters in the tree, with a few notable exceptions. Species groupings were sometimes intermingled within a cohesive genus such as *Mycobacterium*: such examples likely reflect the combined effects of minimal within-genus variation and species definitions that are ecological (and therefore potentially based on a relatively small number of characteristics) rather than genomic in nature. More rarely, members of a given species were intermingled with other genera: *Psudomonas fluorescens* is split by other members of genus *Pseudomonas* but also the genera *Rhodoferax*, *Polaromonas*, *Bordetella*, and *Dechloromonas*. A more dramatic example is provided by the 12 sequenced genomes of the species *P. marinus*: these genomes are split into subgroups that correspond to low-light (Groups 4a and 4b) and high-light (Group 4c) adaptation, with Group 4a comprising the strains with the largest genomes and greatest similarity to marine *Synechococcus*. The separation of *P. marinus* was also observed in the SPTE clustering, whereas Groups 4b and 4c were merged in the GC-SPTE tree. The GC-SPTE clustering also recovered a much larger grouping of *Synechococcus* (eight strains instead of the two seen in Group 4a).

Computed CIs for six distinct taxonomic levels (supplementary fig. S1, Supplementary Material online) confirmed the splitting of broad taxonomic groupings, with CI at the phylum level = 0.184. Because shallower taxonomic levels are nested within deeper ones, it is perhaps not surprising that CI values increase steadily through class (0.262), order (0.416), family (0.629), and genus (0.888), which indicates increasing cohesion with more specific taxonomic groups. The computed CI for species was >0.99, with rare exceptional cases outlined above. CIs for the SPTE clustering were remarkably similar to those obtained using raw distances, with differences no greater than 0.005 for any taxonomic level. GC-SPTE clustering increased the CI at every level, indicating that correction for G + C bias yielded an improvement in the recovery of phylogenetic signal. Nonetheless, as indicated above, many taxonomic units were still disjoint in the GC-SPTE clustering, and some of the recovered groupings appear to still be driven by ecology and putative LGT rather than by taxonomy alone (Kirzhner et al. 2007).

SVM-Based Classification of Genome Fragments All pairings of the 774 genomes in our data set were used to train and test two-class SVMs, with a 5-fold cross-validation approach used to test the efficacy of training. The resulting set of accuracy scores was compared against a number of indices of genomic divergence. Figure 2 shows the relationship between CA and genetic distance between 16S rDNA for 210,439 pairs of genomes. Genomes with identical 16S sequences were nearly indistinguishable by the trained SVM, but CA increased rapidly with increasing 16S divergence, with 5% divergent 16S sequences yielding CA > 90% on average. The model fit is statistically significant ($P < 2.2 \times 10^{-16}$, $R^2 = 0.74$) and much higher than the fit previously obtained using δ*, a measure of dinucleotide dissimilarity (Coenye and Vandamme 2004). Genomes with 16S divergence of 2–3% yielded accuracies between 61.99% (two *Wolbachia* species, associated with *Drosophila melanogaster* and *Culex quinquefasciatus* Pel) and 99.51% (*P. marinus* strain MIT 9515 vs. *Synechococcus* sp. WH 7803). The relationship between CA and PTE distance yielded a higher goodness of fit ($R^2 = 0.84$: see supplementary fig. S2, Supplementary Material online); with very few exceptions, PTE values > 2.5 yielded CA ≥ 90%. The difference in mean genomic G + C content defines a minimum bound on CA: no comparison between genomes whose compositions differed by >10% yielded a CA less than 90%. In many cases, genomes with identical genomic G + C contents were distinguished with 100% accuracy, showing that tetranucleotide frequencies are not solely defined by nucleotide composition. A similar phenomenon was reported by Teeling et al. (2004) when using tetranucleotide-derived *z* scores. There was also a strong correspondence between taxonomic level and CA (fig. 3). Supplementary figure S3 (Supplementary Material online) shows the CA versus G + C distance relationship separated by taxonomic level. Comparisons between genomes from different families generally yielded CA > 80% (supplementary fig. S3*a*–*e*, Supplementary Material online). Pairs within the same genus or family had CA values covering the entire range between 50% and 100% (supplementary fig. S3*f* and *g*, Supplementary Material online). In total, 88.5% of within-species CA values were <60%, although some exceptional cases had a CA of 90% or greater (overall average CA = 54.3%). As shown in supplementary figure S3*h* (Supplementary Material online), only five species had interspecific comparisons with CA > 75%: *P. marinus* (36 comparisons), *Clostridium botulinum* (12), *Buchnera aphidicola* (5), *Chlorobium phaeobacteroides* (1), and *P. fluorescens* (1).

With few exceptions, the transformation of tetranucleotide frequencies by symmetrization and correction for underlying mononucleotide frequencies had little effect. The 298,589 pairwise comparisons that achieved CA ≥ 50% (eliminating results <50% that had high variance) in both the raw and symmetrized trials had a difference in means of 0.06% (98.20% vs. 98.26%). Over 99.5% (297,162/298,589) of trials had a difference in CA < 1%; of the remainder, 901 cases showed an increase in CA with symmetrization, whereas 526 showed a decrease. Correcting for mononucleotide frequencies on a fragment-by-fragment basis in a subsample of 435 genome pairs with CA ≥ 50% (supplementary fig. S4, Supplementary Material online, panel A) yielded a very close fit to the original model

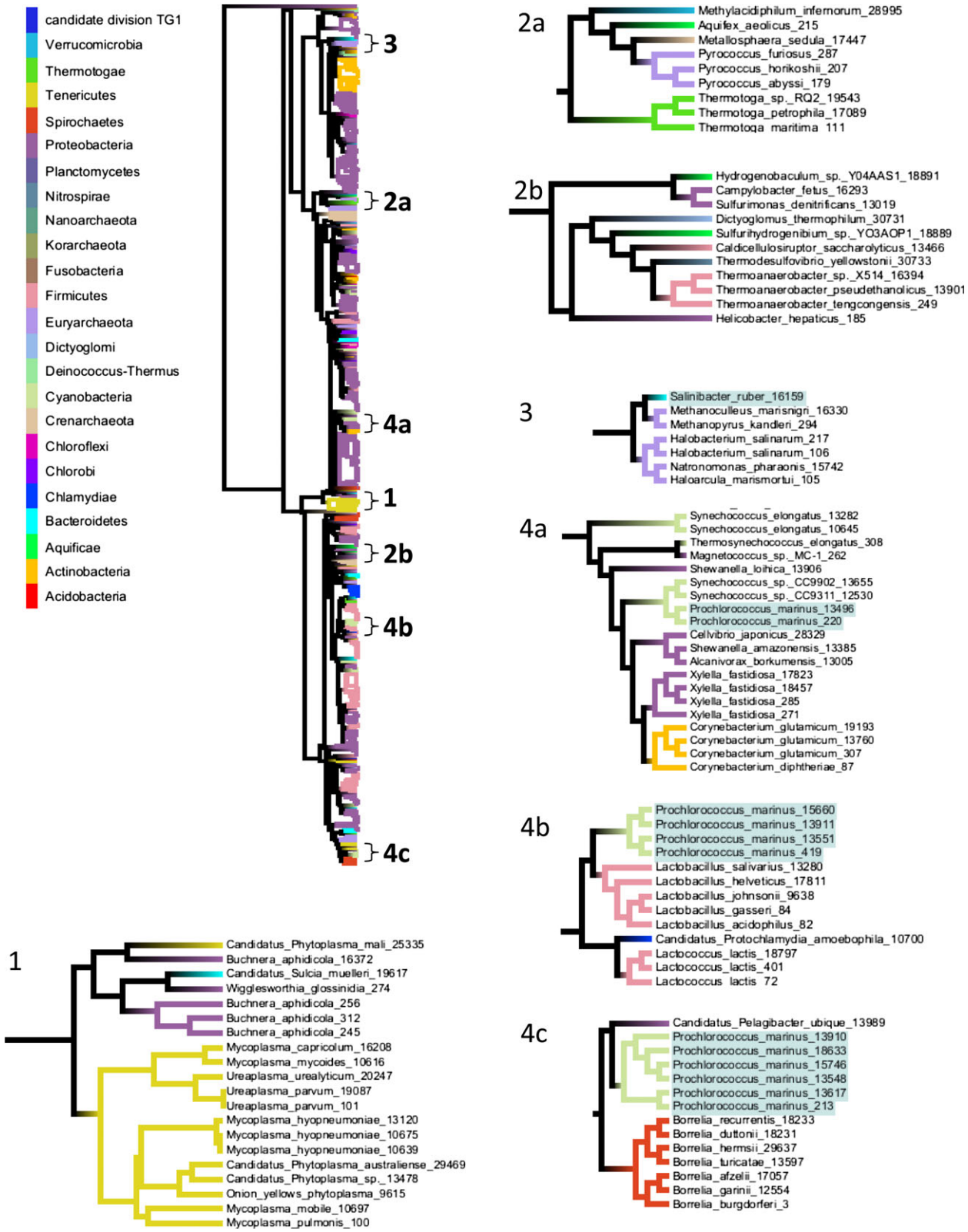

Fig. 1.—Clustering of 774 prokaryotic genomes from a matrix of PTE distances. Edges in the tree are colored according to the legend if their descendant leaves all belong to the same phylum; internal edges that subtend >1 phylum are black. Numbers and letters indicate sets of genomes that are split or merged in ways that are consistent with genome size or habitat. In the detailed subtrees, individual genomes are identified using genus,

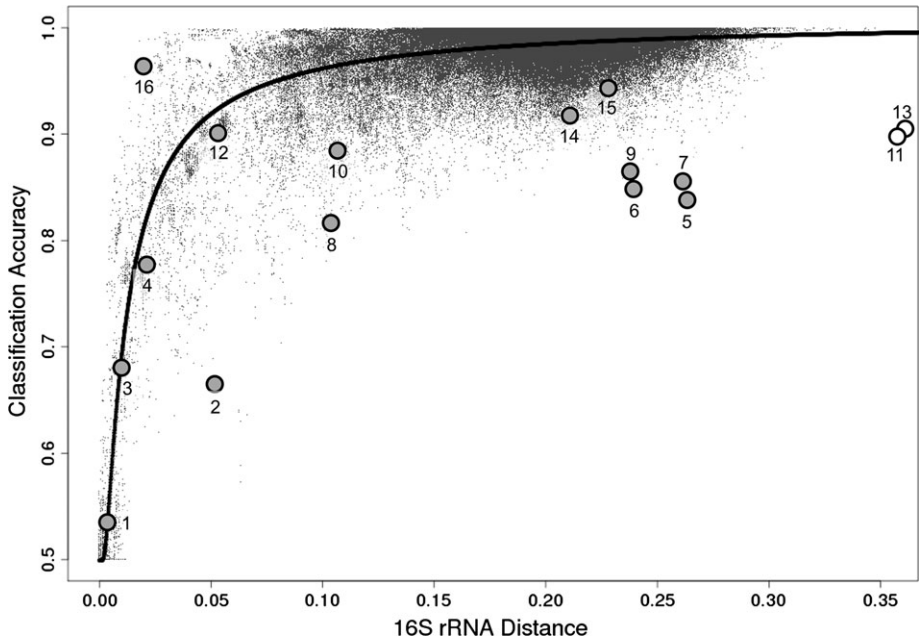

Fig. 2.—CA versus genetic distance between 16S rDNA genes for 210,439 pairs of genomes. The genome pairs listed in supplementary table S2 (Supplementary Material online) are highlighted with large dots and the identifier for each pair; empty circles indicate 16S distances that were computed from ClustalW2 alignments.

($y = 0.98x + 0.497$; $R^2 = 0.99$), and a paired sample *t*-test showed a statistically significant decrease in CA (difference in means = 0.73; $P = 3.2 \times 10^{-26}$). Interestingly, correcting all fragments using the genome-wide mononucleotide frequencies (supplementary fig. S4, Supplementary Material online, panel B) yielded a statistically significant increase in CA (difference in means = 0.63; $P = 1.7 \times 10^{-8}$). This improvement was mostly due to dramatic increases in a few genome pairs: comparisons between the congeners *Methanosarcina barkeri* strain Fusaro/*Methanosarcina acetivorans* C2A, *Mycobacterium tuberculosis* H37Rv/*M. marinum* M, and *M. tuberculosis* H37Rv/*M. ulcerans* Agy99 yielded CA increases of 10.3%, 16.9%, and 18.6%, respectively. More distantly related genome pairs such as *Rhodococcus jostii* RHA1/*Frankia* sp. EAN1pec, *Proteus mirabilis* HI4320/*Yersinia pestis* Angola, and *Enterobacter sakazakii* ATCC BAA-894/*Escherichia coli* APEC O1 improved from ~87% to ~94%. In the absence of a complete comparison of all pairs under this criterion, it is difficult to determine what properties might predispose a pair of genomes to such improved CA under mononucleotide correction. We note, however, that all the pairs highlighted above had large (4–19%) negative residuals relative to the fitted CA versus 16S model.

The model of CA shown in figure 2 predicts the CA for any pair of genomes with a defined 16S distance: outliers with large residual values are interesting because rapid genome mutation or LGT may be the cause of increased or decreased distinguishability (Diaz et al. 2009). The 225 comparisons with a positive residual (i.e., that are more accurate than expected, given the model) >10% relative to the fitted 16S model fall into four categories: 1) interorder comparisons between the picocyanobacterial groups *Prochlorococcus* and *Synechococcus*; 2) intergenus comparisons between *Xanthomonas* and *Xylella*, *Thermococcus* and *Pyrococcus*, and *Herminiimonas* and *Janthinobacterium*; 3)

←
species, and NCBI project ID: these identifiers can be cross-referenced with strain and other information at URL http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. 1: a set of reduced genomes (maximum genome size = 1.1 Mbp) with low genomic G + C content that belong to phyla Bacteroidetes, Tenericutes, and Proteobacteria. 2: dispersal of phylum Aquificae (comprising *A. aeolicus*, *Hydrogenobaculum* sp. YO4AAS1, and *Sulfurihydrogenibium* sp. YO3AOP1) into two distinct groups. Group 2a includes members of phylum Thermotogae including *Thermotoga maritima*, whereas Group 2b includes mesophilic ϵ-Proteobacteria. 3: clustering of *Salinibacter ruber* (highlighted) with haloarchaea and methanogens. 4: splitting of sequenced *Prochlorococcus marinus* genomes (highlighted) into three groups. Group 4a includes the low–light-adapted strains MIT 9313 and MIT 9303, which have relatively large genomes (>2.5 Mbp) in close association with marine *Synechococcus*, Group 4b includes four low–light-adapted strains with genome sizes ~1.8 Mbp and close compositional affinities to lactic acid bacteria and the obligate intracellular endosymbiont Candidatus *Protochlamydia amoebophila*. Group 4c includes the high–light-adapted strains, with the marine α-Proteobacterium Candidatus *Pelagibacter ubique*.

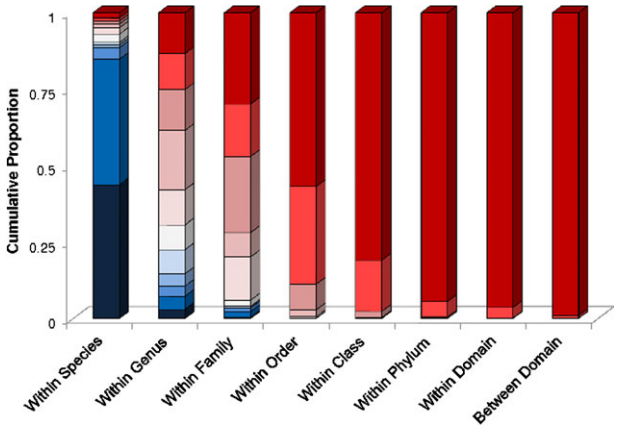

Fig. 3.—CA of all comparisons at each taxonomic level. At each level, CA scores were assigned to 11 bins: a bin for CA 50% and lower, and 10 bins covering intervals of 5% in the range (50%, 100%). Accuracy levels are shown using a color gradient: the deepest blue bar indicates the proportion of comparisons with CA = 50% or less, whereas the top, deep red bar in each column indicates CA > 95%. The lightest colored bar corresponds to 70% < CA ≤ 75%.

interspecies comparisons within a wide array of genera; and 4) intraspecies comparisons involving *P. marinus*, *Pseudomonas putida*, and *P. fluorescens*. The reasons for these large residuals likely differ among genome pairs: for instance, *Prochlorococcus* and *Synechococcus* show remarkable G + C content divergence between 30% and 60% despite the presence of very similar 16S rDNA sequences, which is likely due to differences in DNA repair genes (Rocap et al. 2003). *Xanthomonas* and *Xylella* show considerable genomic collinearity but have acquired a great deal of foreign DNA through different means: *Xanthomonas* contains several dozen insertion sequence-flanked genomic islands, whereas *Xylella* is enriched in phage-related regions (Monteiro-Vitorello et al. 2005). The increased distinguishability of these two genera may be due to consistent compositional differences in these two families of mobile genetic elements.

Negative residuals indicate pairs of genomes that are more difficult to classify than would be predicted by the fitted model: 786 pairs had a negative residual value >10%. Interphylum comparisons with a large negative residual included *P. marinus* versus members of genera *Protochlamydia*, *Borrelia*, *Lactobacillus*, and *Streptococcus*, all of which are found in close association with subsets of *P. marinus* in the UPGMA tree. Comparisons among reduced genomes (*Chlamydophila abortus* vs. *Neorickettsia sennetsu* and *S. muelleri* vs. *Buchnera* and *Wigglesworthia*) also yielded large negative residuals; these genomes were also close to one another (e.g., members of Group 1) in figure 1. A large number of distinct genera were implicated in intraphylum comparisons with large negative residuals. Many species within the Gammaproteobacteria and the Enterobacteriaceae in particular have large negative residuals in within-species comparisons, including *E. coli*, *Salmonella enterica*, *Y. pestis*, and *Haemophilus influenzae*. In 68 of 69 cases, comparisons between members of these species yielded CA values <55%. Negative residuals >25% were exclusively associated with comparisons within the genera *Clostridium* and *Lactobacillus*. Large negative residuals may arise from recent LGT that introduced DNA with multiple foreign signatures (possible in the case of *Clostridium* and the Enterobacteraceae), the overrepresentation of slowly evolving informational genes (Rivera et al. 1998) in reduced genomes or unusual evolutionary properties of the 16S gene in either or both organisms in a pair.

Functional Biases of Misclassified Fragments DNA sequences of different functional types have been found to show compositional similarity across many genomes (Blaisdell et al. 1986; Pietrokovski et al. 1990; Nikolaou and Almirantis 2002). To assess whether the distinguishability of a fragment correlates with the function of any encoded genes, we examined a subset of between-genome comparisons for further analysis, including several outliers as well as comparisons at various CA levels with low residuals (supplementary table S2, Supplementary Material online). Role categories (Peterson et al. 2001) can give insights into potential correlations between correct classification of fragments and evolutionary rate (because informational genes tend to evolve slowly) or functional groupings that may be prone to LGT or other selective pressures. Over all sets of comparisons (table 1), we found the tendency of certain functional groupings toward misclassification to be statistically significant ($X^2 = 42.87$, degrees of freedom [df] = 17, $P = 0.0005$). Functional types more likely to be misclassified include "cellular processes," "mobile and extrachromosomal element functions," and "protein synthesis," with the latter showing a strong tendency toward misclassification in congener comparisons. The effect when subrole levels were considered (supplementary table S3, Supplementary Material online) was not significant ($X^2 = 97.0$, df = 94, $P = 0.394$), although the tendency toward misclassification of mobile elements and informational gene-containing segments remained.

Some misclassification patterns were pair specific with potential ecological implications: in the comparison between *P. marinus* MIT 9303 and *P. marinus* AS9601, overall CA was very high (98.2%) but over twice as many fragments in the "Photosynthesis" subrole of proteins were misclassified relative to the null expectation (17 observed vs. 8 expected). *Nitrospira multiformis* and *Nitrosomonas eutropha* are β-Proteobacterial ammonia oxidizers that were distinguished with an accuracy of 89.5%. Two categories of proteins were difficult to distinguish: cellular processes (particularly proteins involved in detoxification) and "transport and binding proteins" (particularly cations and iron-carrying compounds). Given the adaptive importance of these

Table 1
Main Role-Level Summary and Chi-Square Test of Correctly and Incorrectly Classified Fragments from the 16 Genome Pairs Shown in Supplementary Table S2 (Supplementary Material Online)

Main Role	Correct	Incorrect	Expected Correct	χ^2	Trend
Unknown function	12,437	2,375	12,407.64	0.069	+
Protein synthesis	7,292	1,641	7,482.949	4.873	−
Energy metabolism	6,048	1,104	5,991.05	0.541	+
Transport and binding proteins	4,943	924	4,914.638	0.164	+
DNA metabolism	4,883	940	4,877.78	0.006	+
Biosynthesis of cofactors, prosthetic groups, and carriers	4,220	574	4,015.813	10.382	+
Protein fate	3,670	765	3,715.088	0.547	−
Amino acid biosynthesis	3,517	527	3,387.557	4.946	+
Cell envelope	2,473	517	2,504.648	0.400	−
Purines, pyrimidines, nucleosides, and nucleotides	2,278	411	2,252.507	0.289	+
Cellular processes	1,977	538	2,106.752	7.991	−
Regulatory functions	1,865	322	1,831.995	0.595	+
Hypothetical proteins	1,614	245	1,557.237	2.069	+
Transcription	1,441	342	1,493.574	1.851	−
Signal transduction	1,194	297	1,248.973	2.420	−
Central intermediary metabolism	1,038	193	1,031.178	0.045	+
Mobile and extrachromosomal element functions	887	260	960.813	5.671	−
Fatty acid and phospholipid metabolism	835	158	831.811	0.012	+

proteins in the soil environment (Norton et al. 2008), they are good candidates for LGT. *Bradyrhizobium japonicum* and *Mesorhizobium loti*, two nitrogen-fixing soil bacteria, had three functional classes of genes that were more difficult to classify than expected from their 82.1% accuracy score: those encoding proteins associated with mobile DNA (conjugation and prophage), genes involved in nitrogen fixation, and genes related to pathogenesis. Symbiosis genes including those relating to nitrogen fixation are known to reside on "symbiosis islands" that may be readily transferable between species (Sullivan et al. 1995), and it has long been known that symbiosis genes in *B. japonicum* show unusual compositional properties relative to the rest of the genome (Ramseier and Göttfert 1991), so the reduced distinguishability of these genes is consistent with the known evolutionary dynamics of symbiosis in the Rhizobia.

Within this restricted set of pairwise comparisons, we also compared the CA of fragments using two different criteria of coding/noncoding heterogeneity, under the hypothesis that the two types of sequence exhibit different compositional biases (Bohlin et al. 2008; see Dick et al. 2009 for contrasting results). Our noncoding set encompasses structural RNAs and pseudogenes not annotated as protein-coding sequence: separate consideration of pseudogenes would be valuable but would require careful reannotation of all microbial genomes (Lerat and Ochman 2005) and consideration of the degree to which a pseudogene has decayed relative to its original protein-coding state. Because the majority of microbial genome sequence encodes proteins, intergenic sequences might be misclassified because they offer a worse fit to the model learned by the SVM. Two indices of heterogeneity were considered: the number of transition points within a fragment and the length of the longest coding sequence within a given fragment. Consistent with our hypotheses, in a majority of the genome pairs considered (9 of 15: see supplementary fig. S5, Supplementary Material online), the misclassified fragments were significantly more heterogeneous on average, with shorter longest coding sequences and more transition points. Three genome pairs showed the opposite effect: *Borrelia duttonii* Ly versus *B. recurrentis* A1, *P. marinus* strain MIT 9303 versus *P. marinus* strain AS9601, and *Ehrlichia ruminantium* strain Welgevonden v2 versus *Methanosphaera stadtmanae* DSM 3091. Of these three exceptional cases, two involve congeners, whereas the latter involves an unusual intracellular organism with an exceptionally low coding percentage and many tandem repeats (Frutos et al. 2007), and an intestinal archaeon that has many genes of bacterial and eukaryotic origin (Fricke et al. 2006). In the congener comparisons, it is probable that the rapidly evolving intergenic sequences are the first to display sufficient levels of consistent compositional divergence to be distinguishable by the SVM.

Compositional Variation within Genomes Although compositional patterns tend to be more variable between than within genomes, within-genome variation can have a significant impact on constructed models (Suzuki et al. 2008). We considered the impact of unsupervised clustering of genome fragments based on their 256-element tetranucleotide profile. We used a *k*-means clustering approach with $k = 2, 3, 4, 5, 6$ to separate fragments into compositional clusters, then assessed the functional breakdown of these clusters to determine whether functional groupings or

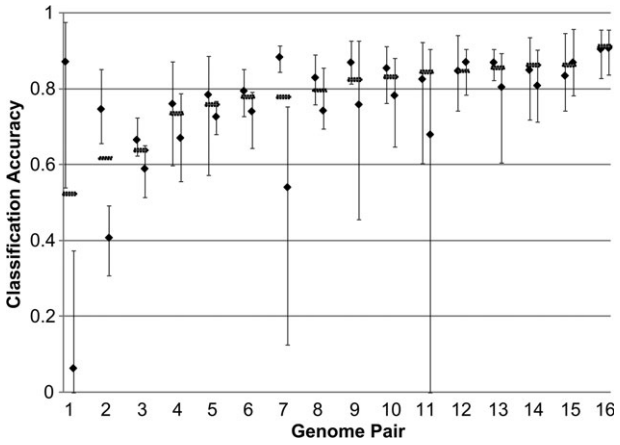

Fig. 4.—Mean and range of cluster accuracies for 16 genome pairs. The CA of each cluster was computed in the strict sense, with any assignment to another cluster deemed a misclassification. Minimum, mean, and maximum accuracy scores are shown using bars and diamonds, whereas gray rectangles indicate the overall CA for that comparison when $k = 6$.

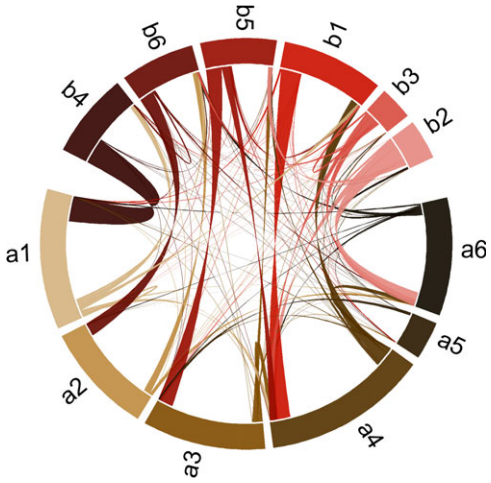

Fig. 5.—Visualization of cluster misclassification between *Sulcia muelleri* (b1–b6) and *Buchnera aphidicola* strain Cc (a1–a6). The thickness of the ribbon emanating from the most counterclockwise (e.g., at the left of cluster b5) position of the cluster indicates the proportion of that cluster that was misclassified. The ribbon connected to the most clockwise position of each cluster indicates the number of other fragments that were mistakenly given this cluster assignment by the SVM.

difficult-to-classify sequences associate preferentially with certain compositional clusters. For each k-way clustering, we trained SVMs to distinguish the genome pairs shown in supplementary table S2 (Supplementary Material online), with a total of $2k$ possible classification assignments. CA was evaluated under two criteria: a stringent case in which a given fragment from cluster i was considered to be misclassified if it was predicted to belong to one of the other $2k - 1$ classes and a relaxed case where fragments only needed to be associated with one of the k clusters from the correct originating genome. Under both the stringent and relaxed criteria, CA tended to decrease with increasing k (supplementary table S4, Supplementary Material online), suggesting that the trained SVM was able to directly model compositional variation within a genome, and the explicit breakdown into clusters was detrimental to the accuracy of the model.

Different clusters are classified with varying degrees of accuracy. When $k = 6$ (fig. 4), over 50% of the fragments in some clusters are misclassified in cases where the two genomes are very similar (for instance, *B. duttonii* vs. *B. recurrentis*) or when clusters are small (e.g., clusters with 55 or fewer fragments from *S. mulleri*). In comparisons between a larger and a smaller genome, a majority of fragments from the smaller genome may be assigned to the larger one, yielding a CA $< 50\%$ for the smaller genome. Even well-distinguished genomes generated clusters that differed by ~10% in CA. When fragments from a particular cluster were misclassified, they were often preferentially assigned to a single cluster from the other genome. Confusion matrices showing these misclassification relationships were visualized using Circos (Krzywinski et al. 2009): in the comparison between *S. mulleri* and *B. aphidicola* strain Cc, fragments from *B. aphidicola* cluster 1 were most frequently misassigned to *S. mulleri* cluster 4 (fig. 5). The most extreme cases were *S. mulleri* clusters 2 and 3, which had only 6/48 and 15/40 fragments correctly classified (CA = 12.5% and 37.5%), with 30 of 42 misclassified cases assigned to *B. aphidicola* clusters 4 or 6. Some clusters were assigned with high levels of accuracy: only 4 of 46 fragments from *B. aphidicola* cluster 5 were assigned incorrectly. In the comparison of two strains of *P. marinus*, 4 of 6 clusters from each genome were classified with accuracy >90% and were always associated with the correct genome if not the correct cluster within that genome. Between 8% and 11% of fragments from the remaining two compositional clusters of each genome were assigned to the incorrect genome (fig. 6).

The unsupervised clustering ($k = 6$) of the extremely biased genomes of *S. mulleri* and *B. aphidicola* yielded four strong clusters that segregated based on strand affinity and G + C content (fig. 7*a* and supplementary table S5, Supplementary Material online). *Sulcia muelleri* clusters 2 and 3 had G + C contents that were even less than those of the four main clusters, and their frequent misclassification appears to arise from greater affinities for the extreme compositions of clusters from *B. aphidicola*. For example, fragments from *S. mulleri* cluster 2 (average G + C = 17.3%) were frequently misclassified as *B. aphidicola* cluster 6: the average G + C content of the 19 fragments mistakenly assigned to this cluster (15.3%) was a close match to its average G + C content of 14.6%. The 11 fragments assigned to *B. aphidicola* cluster 4 (average G + C = 23.1%) had a much higher G + C content of 20.3%.

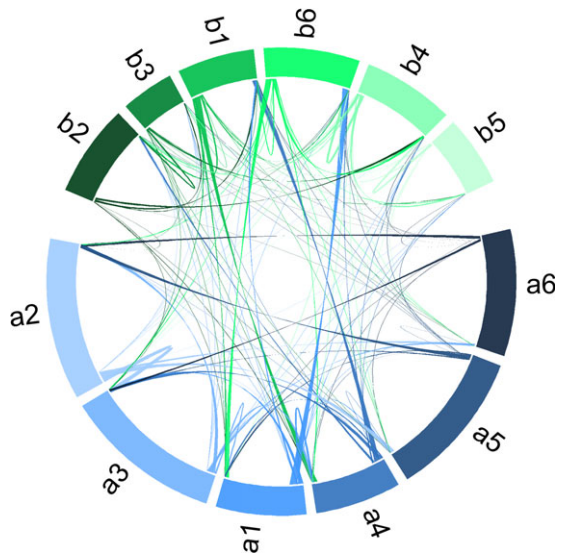

Fig. 6.—Visualization of cluster misclassification between *Prochlorococcus marinus* strains MIT 9303 (a1–a6) and AS9601 (b1–b6). The thickness of the ribbon emanating from the most counterclockwise (e.g., at the left of cluster b2) position of the cluster indicates the proportion of that cluster that was misclassified. The ribbon connected to the most clockwise position of each cluster indicates the number of other fragments that were mistakenly given this cluster assignment by the SVM.

Buchnera aphidicola cluster 5 appears to be a similar outlier relative to the whole genome, but its mean G + C content of 13.3% produced very little overlap with clusters from *S. mulleri* and no misclassifications. A nucleotide BLAST comparison of all fragments from these two genomes against the reference database yielded match profiles that were biased toward other low G + C organisms, particularly *P. marinus*, *C. botulinum*, and *C. difficile*, and confirmed the outlier status of clusters 5 and 6 from *B. aphidicola* and 2 and 3 from *S. mulleri*.

Clusters 1 and 4 recovered from *P. marinus* MIT 9313 (fig. 7*b* and supplementary table S6, Supplementary Material online) have average fragment G + C contents that are considerably lower than the rest (39.6% and 41.6% vs. 48.0%, 54.1%, 54.2%, and 55.1%) and cover a relatively small fraction (22.8%) of the genome. We used a sliding-window approach to assess the extent to which fragments in these clusters tended to form localized groupings in the genome. Any contiguous regions of length $\leq$5 kbp in the genome that had at least 8 of 10 fragments assigned to clusters 1 and/or 4 were considered to be a local group; 63.2% of all fragments satisfied this criterion. These fragments encoded many genes with viral functions (e.g., Crp regulatory proteins, porins), likely candidates for recent transfer (e.g., multidrug resistance proteins), and proteins known to exist on cyanophage (e.g., high-light induced proteins). When compared against the reference genomic database using BLAST, many of these proteins showed best nonself hits to either low-G + C strains of *P. marinus* (particularly *P. marinus* CCMP1375) or noncyanobacterial genomes. The relatively restricted taxonomic distribution of these proteins is suggestive of recent transfers into *P. marinus* MIT 9303, with the lack of amelioration leading to difficulty in correctly assigning the corresponding genome fragments, particularly when trying to distinguish this genome from that of low-G + C strains of *P. marinus*.

Discussion

Absent certain confounding factors, the majority of genome pairs could be distinguished with >95% accuracy, even though the fragment size we chose leads to high compositional variance and poses significant problems to classifiers (McHardy et al. 2007). Although our accuracy scores are not directly comparable to the multiclass classifiers that are used in metagenomic analysis, they clearly demonstrate which subproblems in a multiclass setting will diminish the overall accuracy. With few exceptions, congeners were very difficult to distinguish. This observation is consistent with previous compositional analyses done at the genus and species level (Coenye and Vandamme 2004; van Passel et al. 2006), which have shown a statistically significant correlation between indices of compositional similarity and the similarity of marker genes such as 16S. Exceptions to this trend include paired strains of *P. marinus*, *B. aphidicola*, and *C. botulinum*: these genomes are members of the same named species but have diverged very rapidly and possess very different patterns of nucleotide usage and/or gene content. Conversely, members of the same species may show considerable variation in nucleotide sequence or gene content and yet be indistinguishable based on their global usage properties: for example, different genomes of *E. coli* could not be distinguished at all in spite of their considerable genomic and ecological divergence (Welch et al. 2002). Comparisons between noncongeners yielded unexpectedly low CA in cases where taxonomic units are clearly not reflective of evolutionary relatedness (e.g., *Escherichia* vs. *Shigella*) (Lan et al. 2004) and when genomes have extreme compositional biases, as in the case of *S. mulleri* versus *B. aphidicola*. Compositional convergence has been noted for distantly related genomes due to habitat convergence (Bohlin et al. 2009) or "crowding" of nucleotide signature space (Mrázek 2009), and our clustering analysis supports this as well. However, in the majority of cases, this convergence does not impede classification. Symmetrization and mononucleotide correction yielded no significant improvement in CA, except when the mononucleotide corrections were carried out using the genome-wide nucleotide frequencies. Although this result is worthy of deeper investigation, we note that such a scheme is impracticable for the analysis of metagenomic fragments because the correct

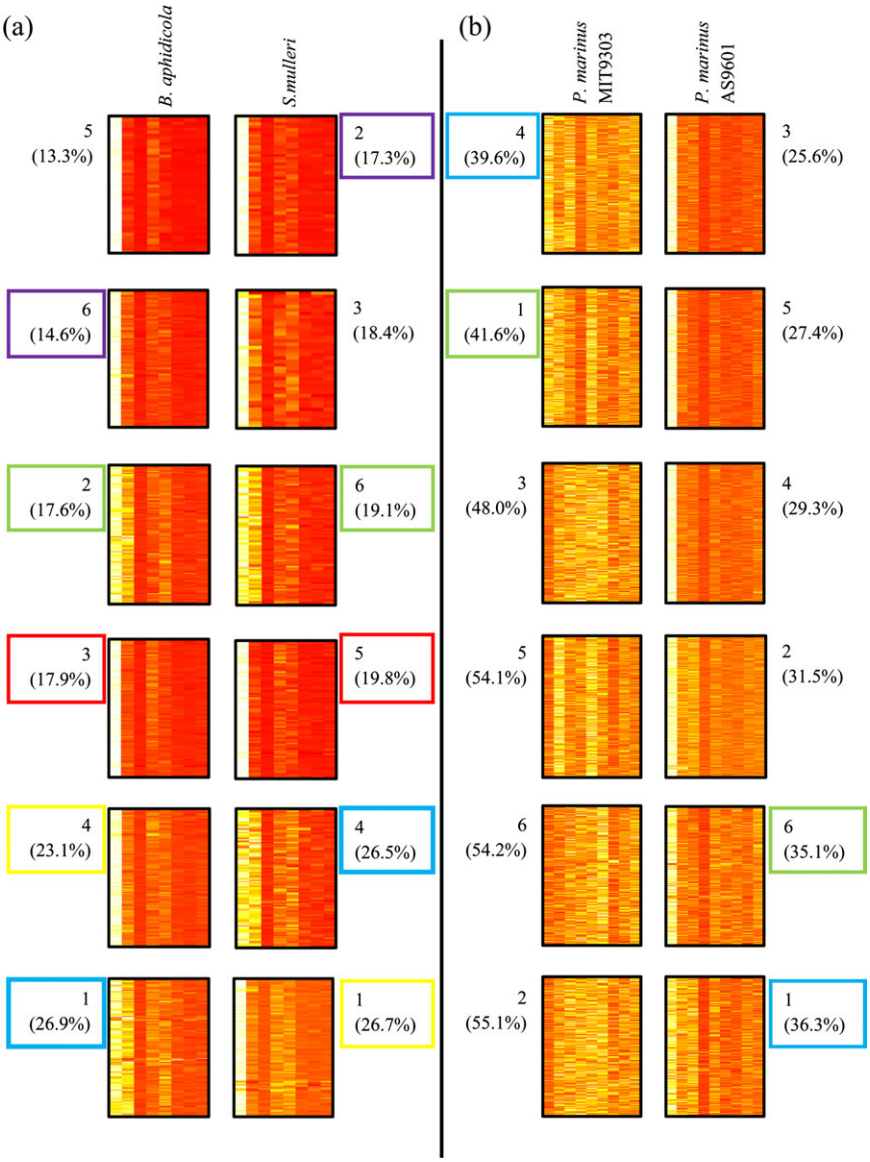

Fig. 7.—Heatmaps showing the relative frequency of representative tetranucleotides in six unsupervised clusters of fragments from two pairs of genomes. Each individual heatmap corresponds to one numbered cluster of sequences from a given genome, with each row showing the frequency profile for an individual fragment. The color gradient ranges from red (tetranucleotide is absent from a given fragment) through orange and yellow to white (tetranucleotide frequency is maximal given the data set). The mean G + C content for each cluster is indicated in parentheses, whereas colored borders indicate paired clusters that are frequently conflated by the SVM, corresponding to thick connecting edges in figures 5 and 6. (*a*) *Buchnera aphidicola* versus *Sulcia muelleri*, with heatmap columns corresponding to the frequencies of AAAT, TTTC, GCCG, AAAC, TGTA, AACG, GCCA, ATCG, and TTGA. (*b*) *Prochlorococcus marinus* MIT 9303 versus *P. marinus* AS9601, with heatmap columns corresponding to frequencies of AATT, CCAA, CTTC, CGGC, AGCA, CTGG, GTAG, GCAT, GGAT, and ATCA. Although the tetranucleotides with highest loadings on the first ten principal components were chosen to illustrate compositional variation, only nine appear in (*a*) because the tetranucleotide TGTA had the highest loading on both components 5 and 6.

mononucleotide frequencies for the originating genome of a particular fragment will generally not be known.

Functional analysis of misclassified segments identified an overrepresentation of mobile elements and informational genes, particularly those implicated in protein synthesis. Mobile elements will likely show a compositional pattern that differs from that of the host genome and potentially the other genome in the pair, leading to an arbitrary classification choice. Most informational genes are highly constrained by interactions with other proteins and evolve slowly (Rivera et al. 1998), which can potentially lead to slower divergence of nucleotide usage patterns. Fragments that contained a mixture of coding and noncoding genomic sequence were typically more difficult to classify accurately

than were pure coding sequences, except when genomes were very similar, in which case intergenic sequence might be the principal distinguishing trait. This tendency could be exploited in metagenomic analysis, by using compositional variation in fragments that contain intergenic sequence to distinguish closely related but ecologically distinct strains.

Important genes may be incorrectly attributed if their composition is not reflective of their host genome or is strongly G + C or A + T biased. This could be potentially confounding in the analysis of habitats that contain multiple G + C-rich or G + C-poor genomes such as insect bacteriomes (McCutcheon and Moran 2007), although in such cases read depth may be helpful in distinguishing genomes based on their relative abundance. The misclassification of low-G + C islands in *P. marinus* strain MIT 9303 is similar to the noted misattribution of likely transfers into the genome of *Thermoplasma acidophilum* (Diaz et al. 2009). More generally, genomic islands often contain important adaptive genes that are among the most important genes to assign correctly in a metagenomic analysis. In performing this analysis, we chose to include all fragments from a genome when training the SVM, as opposed to implementing a filtering phase to exclude genes that are atypical in their composition. A consequence of this is that trained models may not reflect only the "core" compositional pattern of a genome but will also include signatures of introgressed and otherwise unusual genes. Our results above indicate that SVMs are capable of modeling complex mixtures of compositional patterns, so the inclusion of atypical sequences should not harm the ability to correctly identify fragments that exhibit the core compositional pattern. Furthermore, introgressed genes may be undergoing amelioration, in which case they will exhibit a mixture of the compositional patterns from both the donor and recipient genomes: including such fragments will increase the likelihood of correctly binning such genes in a metagenomic sample.

Our oligonucleotide-based approach uses a standard SVM implementation and will likely be outperformed to some extent by more complex supervised machine learning methods (Sandberg et al. 2001; McHardy et al. 2007; Diaz et al. 2009). Nonetheless, our results highlight situations where more explicit modeling of the expected properties of difficult cases is worthwhile and could potentially improve the accuracy of any classifier; additionally, we have identified examples (such as the identification of compositional "bins" within genomes) where such modeling is likely of no benefit. In identifying several different confounding factors impacting on classification, we highlight the importance of reporting what aspect of the classification problem (e.g., close relatives, atypical sequences, functional subcategories) is addressed by an improved classifier. As currently implemented, most classifiers treat a metagenomic fragment as a homogeneous stretch of sequence: this averaging may increase the tendency of a heterogeneous fragment to be misclassified. Sequence segmentation approaches have been successfully applied in other settings (Keith 2008) and will be of value in the analysis of metagenomic data as well. Additionally, although homology search (e.g., best BLAST hit) can be a poor surrogate for the taxonomic identity of a given fragment, learning the most probable functional role of an encoded gene can help to provide a measure of confidence in the predicted assignment.

Beyond the prospect of better model-based approaches, it is clear that there are fundamental limits to classification, particularly between closely related organisms. The semisupervised learning approach of CompostBin (Chatterji et al. 2008) and S-GSOM (Chan et al. 2008) represents a promising direction, with taxonomic cues provided by confidently assigned marker genes. This approach could be further extended to take into account knowledge of likely origins of transferred genes, particularly low-G + C prophage sequences, by mining the rapidly growing set of sequenced genomes (including bacteriophage genomes) for information about frequent exchange partners (Beiko et al. 2005; Dagan et al. 2008). In addition to this, if a small number of distinct genes in a metagenomic sample are classified as belonging to a given genome, but marker genes are not present for that genome, then we might surmise that those genes are in fact derived from a different genome that is known to be present in the sample. Here again, reference databases could be exploited to identify likely associations.

Supplementary Material

Supplementary figures S1–S5 and supplementary tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank Andrew Wong for help with initial SVM trials and Donovan Parks for providing a version of his tree visualization software. This work was supported by the Natural Sciences and Engineering Research Council of Canada (342478); Genome Atlantic; and the Canada Foundation for Innovation (203592). R.G.B. acknowledges the support of the Canada Research Chairs program.

Literature Cited

Abe T, et al. 2003. Informatics for unveiling hidden genome signatures. Genome Res. 13:692–702.

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T. 2005. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res. 12:281–290.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A. 102:14332–14337.

Blaisdell BE. 1996. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci U S A. 83:5155–5159.

Bohlin J, Skjerve E, Ussery DW. 2008. Investigations of oligonucleotide usage variance within and between prokaryotes. PLoS Comput Biol. 4:e10000057.

Bohlin J, Skjerve E, Ussery DW. 2009. Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering. BMC Genomics. 10:487.

Boussau B, Guéguen L, Gouy M. 2008. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol. 8:272.

Brendel V, Beckmann JS, Trifonov EN. 1986. Linguistics of nucleotide sequences: morphology and comparison of vocabularies. J Biomol Struct Dyn. 4:11–21.

Carbone A, Képès F, Zinovyev A. 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol Biol Evol. 22:547–561.

Carbone A, Zinovyev A, Képès F. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics. 19:2005–2015.

Chan C-KK, Hsu AL, Halgamuge SK, Tang S-L. 2008. Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics. 9:215.

Chatterji S, Yamazaki I, Bai Z, Eisen JA. 2008. CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. In: Research in Computational Molecular Biology. Berlin (Germany): Springer. p. 17–28.

Clarke GDP, Beiko RG, Ragan MA, Charlebois RL. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol. 184:2072–2080.

Coenye T, Vandamme P. 2004. Use of the genomic signature in bacterial classification and identification. Syst Appl Microbiol. 27:175–185.

Cole JR, et al. 2008. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37:D141–D145.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105:10039–10044.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. 2009. TACOA—taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics. 10:56.

Dick GJ, et al. 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10:R85.

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 33:e6.

Farris JS. 1969. A successive approximations approach to character weighting. Syst Zool. 18:374–385.

Felsenstein J. 1989. PHYLIP—phylogeny inference package (Version 3.2). Cladistics. 5:164–166.

Fricke WF, et al. 2006. The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. J Bacteriol. 188:642–658.

Frutos R, Viari A, Vachiery N, Boyer F, Martinez D. 2007. Ehrlichia ruminantium: genomic and evolutionary features. Trends Parasitol. 23:414–419.

Hsiao W, Wan I, Jones SJ, Brinkman FSL. 2003. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics. 19:418–420.

Karlin S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol. 9: 335–343.

Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11:283–290.

Karlin S, Ladunga I, Blaisdell BE. 1994. Heterogeneity of genomes: measures and values. Proc Natl Acad Sci U S A. 91:12837–12841.

Keith JM. 2008. Sequence segmentation. Methods Mol Biol. 452: 207–229.

Kirzhner V, Paz A, Volkovich Z, Nevo E, Korol A. 2007. Different clustering of genomes across life using the A-T-C-G and degenerate R-Y alphabets: early and late signaling on genome evolution? J Mol Evol. 64:448–456.

Krause L, et al. 2008. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 36:2230–2239.

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR. 2004. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. Infect Immun. 72:5080–5088.

Larkin MA, et al. 2007. ClustalW and ClustalX version 2. Bioinformatics. 23:2947–2948.

Lerat E, Ochman H. 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Res. 33:3125–3132.

Manichanh C, et al. 2008. A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. Nucleic Acids Res. 36:5180–5188.

Martin C, Diaz NN, Ontrup J, Nattkemper TW. 2008. Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. Bioinformatics. 24:1568–1574.

Mavromatis K, et al. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 4:495–500.

McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci U S A. 104:19392–19397.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 4:63–72.

McHardy AC, Rigoutsos I. 2007. What's in the mix: phylogenetic classification of metagenome sequence samples. Curr Opin Microbiol. 10:499–503.

Misra VK, Honig B. 1996. The electrostatic contribution to the B to Z transition of DNA. Biochemistry. 35:1115–1124.

Mongodin EF, et al. 2005. The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. Proc Natl Acad Sci U S A. 102:18147–18152.

Monteiro-Vitorello CB, et al. 2005. *Xylella* and *Xanthomonas* mobil'omics. OMICS. 9:146–159.

Mrázek J. 2009. Phylogenetic signals in DNA composition: limitations and prospects. Mol Biol Evol. 26:1163–1169.

Nikolaou C, Almirantis Y. 2002. A study of the middle-scale nucleotide clustering in DNA sequences of various origin and functionality, by means of a method based on a modified standard deviation. J Theor Biol. 217:479–492.

Norton JM, et al. 2008. Complete genome sequence of *Nitrosospira multiformis*, an ammonia-oxidizing bacterium from the soil environment. Appl Environ Microbiol. 74:3559–3572.

Paul S, Bag SK, Das S, Harvill ET, Dutta C. 2008. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 9:R70.

Paz A, Kirzhner V, Nevo E, Korol A. 2006. Coevolution of DNA-interacting proteins and genome "dialect". Mol Biol Evol. 23:56–64.

Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. 2001. The comprehensive microbial resource. Nucleic Acids Res. 29:123–125.

Pietrokovski S, Hirshon J, Trifonov EN. 1990. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. J Biomol Struct and Dyn. 7:1251–1268.
Ragan MA. 2002. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett. 201:187–191.
Ramseier TM, Göttfert M. 1991. Codon usage and G + C content in Bradyrhizobium japonicum genes are not uniform. Arch Microbiol. 156:270–276.
Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A. 95:6239–6244.
Rocap G, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature. 424:1042–1047.
Rocha EPC, Touchon M, Feil EJ. 2006. Similar compositional biases are caused by very different mutational effects. Genome Res. 16: 1537–1547.
Sandberg R, et al. 2001. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. Genome Res. 11: 1404–1409.
Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nat Genet. 21:108–110.
Sullivan JT, Patrick HN, Lowther WL, Scott DB, Ronson CW. 1995. Nodulating strains of *Rhizobium loti* arise through chromosomal symbiotic gene transfer in the environment. Proc Natl Acad Sci U S A. 92:8985–8989.
Suzuki H, Sota M, Brown CJ, Top EM. 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res. 36:e147.
Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol. 6:938–947.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102:13950–13955.
van Belkum A, Scherer S, van Alphen L, Verbrugh H. 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev. 62:275–293.
van Passel MWJ, Bart A, Luyf ACM, van Kampen AHC, van der Ende A. 2006. Compositional discordance between prokaryotic plasmids and host chromosomes. BMC Genomics. 7:26.
Welch RA, et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A. 99:17020–17024.
Willenbrock H, Friis C, Juncker AS, Ussery DW. 2006. An environmental signature for 323 microbial genomes based on codon adaptation indices. Genome Biol. 7:R114.

Genome Biology and Evolution SMBE