Supplemental Figures

Figure S1. Scatter plot of the gene density for Genomic Islands (GIs) and randomly sampled regions in *Salmonella* (A), *Staphylococcus* (B) and *Streptococcus* (C) genera. Each point in the scatter plot represents the gene density value for either a GI (class 1, blue coloured) or a randomly sampled region (class 0, red coloured). The *p*-value has been calculated using a two-tailed *t*-test.

Figure S2. Radar diagram illustrating the "importance" of eight structural features under different genus-specific and cross-genus (2 genera) GI models: A) *Staphylococcus*-*Streptococcus*, B) *Salmonella*-*Staphylococcus* and C) *Salmonella*-*Streptococcus*. Each apex in the octagon-like diagram corresponds to one of the eight structural features, while the height of the plot at the corresponding apex is indicative of the actual feature importance.

Figure S3. Radar diagram illustrating the "importance" of eight structural features under different genus-specific and cross-genus (3 genera) GI models: The *Salmonella* complete dataset (A), set1 (B), set2 (C) and set3 (D) are combined with the complete *Staphylococcus* and *Streptococcus* training datasets. The above four cross-genus GI models are shown together under the same diagram for ease of comparison (E).

Figure S4. ROC curve analysis showing the performance of the RVM classifier under different GI structural models and different training and test datasets (cross validation). Each dataset (e.g. Salm) is split into five subsets of approximately equal size; four of the five subsets (e.g. Salm.2, Salm.3, Salm.4 and Salm.5) are used to train an RVM model while the omitted subset (e.g. Salm.1) is used to test the performance of this model. For each dataset (e.g. Salm) this process is repeated five times on non overlapping test sets. The performance of the RVM model for each cross validation is visualized through the ROC curve; the Sensitivity and 1-Specificity of each model is plotted on the Y and X axis of the ROC curve respectively. For the last six "genus-blind" cross validations an RVM model (e.g. SalmTrain-StaphTest) is trained on the complete dataset of a given genus (e.g. Salm) and tested on the complete dataset of another genus (e.g. Staph).

Figure S5. Scatter plot showing the posterior probability of a given region of being a true GI, given the model. Each genus specific dataset (e.g. Salm) is used to train an RVM model (e.g. Salm-train) that is then tested on the dataset of one of the other two genera (e.g. Strep-test). Each point in the scatter plot represents the posterior probability of either a GI (class 1, blue coloured) or a randomly sampled region (class 0, red coloured) of being a true GI given the model. For example in scatter plot A, a model trained on the *Salmonella* dataset was tested on the *Streptococcus* dataset: GIs (blue coloured points) in the test-set were correctly classified with a high probability very close to 1 while randomly sampled regions (red coloured points) in the test-set received on average a much lower probability.

Figure S6. Bar chart illustrating the importance of eight structural features under a *Salmonella*, *Staphylococcus* and *Streptococcus* GI model: Green-coloured bars show the "importance" of every feature, in a (multi-featured) GI model in which all eight features are taken into account ("relative" importance). Orange-coloured bars show the "importance" of each feature, in a (single-featured) GI model with only one structural feature evaluated each time ("absolute" importance).

Figure S7. Venn diagram illustrating the orthologous genes shared between each of the three reference genus and the corresponding outgroup strains: A) 473 *Salmonella*-specific and 1952 core genes (genes shared between the *Salmonella* and the four outgroup strains), B) 688 *Staphylococcus*-specific and 741 core genes, C) 283 *Streptococcus*-specific and 429 core genes.

Figure S8. A) The phylogenetic relationship between the eleven *Salmonella* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values (proportions out of 100) are given for each node. The tree topology is based on the aminoacid sequence of 1952 core gene products shared by the fifteen genomes. B) Phylogenetic tree topologies using the ML (left) and the NJ (right) method, based on the alignment of the 1952 core gene products (top) and the whole chromosome sequences (bottom) of eleven *Salmonella* and four outgroup genomes. C) Differences between the tree topologies (core gene products) given by the ML and the NJ methods are highlighted. The only difference in terms of node topology lies within the Typhimurium lineage. In the ML topology, DT104 and LT2 are grouped together, while in the NJ topology DT104 is grouped together with SL1344. The bootstrap value of 50 supports the observed ambiguity.

Figure S9. A) The phylogenetic relationship between the thirteen *Staphylococcus* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values are given for each node. The tree topology is based on the aminoacid sequence of 741 core gene products shared by the seventeen genomes. B) Phylogenetic tree topologies using the ML (left) and the NJ (right) method, based on the alignment of the 741 core gene products (top) and the whole chromosome sequences (bottom) of thirteen *Staphylococcus* and four outgroup genomes. C) Differences between the tree topologies given by the ML and the NJ methods are highlighted. The likelihood (Ln) for each tree topology, under a JTT model with four categories (4C) of sites and the γ correction (gamma), is provided for each topology. Based on TREE-PUZZLE, the best tree topology is the one given by the NJ method (JTT, 4C, gamma). Based on the tree topology evaluation of *PROML*, the NJ (*Kimura* model) method gives the best tree topology (highest Ln); however the other three topologies are not significantly worse (*p*-value: 0.273, 0.628, 0.157 respectively), suggesting that the observed differences are close to the systematic error of those methods.

Figure S10. A) The phylogenetic relationship between the thirteen *Streptococcus* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values are given for each node. The tree topology is based on the aminoacid sequence of 429 core gene products shared by the seventeen genomes. B)

Phylogenetic tree topologies using the NJ (top) and the ML (bottom) method, for the 429 core gene products of the thirteen *Streptococcus* and the four outgroup genomes. C) Differences in the tree topology given by the ML and the NJ methods are highlighted: based on the tree topology evaluation (TREE-PUZZLE and PROML) the NJ method gives the topology with the highest likelihood (best tree).

Figure S11. Circular map of the *Salmonella* (A), *Staphylococcus* (B) and *Streptococcus* (C) "Mobilome", illustrating the phylogenetic distribution of the putative GIs identified in the three reference lineages (red: presence, pink: partial presence, white: absence). The list of strains (outwards-inwards orientation relative to the map) is embedded at the center of the circular map. The regions are arbitrarily numbered based on the strain first found.

Figure S12. Size distribution of the putative Genomic Islands identified in this analysis for the three reference genera: *Salmonella* (A), *Staphylococcus* (B) and *Streptococcus* (C).

Supplemental Tables

Supplemental Table 1. A list of the genomic (GIs and randomly sampled) regions, identified in the three reference lineages (*Salmonella*, *Staphylococcus* and *Streptococcus*), along with their compositional and structural annotation.

Supplemental Table 2. A list of fifteeen integrase(-like) and 191 phage-related Pfam HMMs used throughout this analysis.