Insights into Bacterial Genome Composition through Variable Target GC Content Profiling

Scott Mann[1], Jinyan Li [2], Yi-Ping Phoebe Chen[1,2§]

[1]Faculty of Science and Technology, Deakin University, Australia

[2]School of Computer Engineering, Nanyang Technological University, Singapore

[32]ARC Centre of Excellence in Bioinformatics, Australia

[§]Corresponding author: phoebe@deakin.edu.au

Abstract

This study presents a new computational method for GC content profiling based on the idea of multiple resolution sampling. The benefit of our new approach over existing techniques follows from its ability to locate significant regions without prior knowledge of the sequence, nor the features being sought. The use of multiple resolution sampling has provided novel insights into bacterial genome composition. Key findings include those that are related to the core composition of bacterial genomes, to the identification of large genomic islands (in Enterobacterial genomes), and to the identification of surface protein determinants in human pathogenic organisms (e.g. Staphylococcus genomes). We observed that bacterial surface binding proteins maintain abnormal GC content, potentially pointing to a viral origin. This study has demonstrated that GC content holds a high informational worth and hints at many underlying evolutionary processes.

Background

Bacterial genomic GC content has been reported to broadly range between 25% to 75% [1], and recently some organisms have been identified outside of this range, including Carsonella extending below 20% GC. The GC content of a prokaryote genome is a complex property borne out of evolutionary time. For example, free-living bacteria have average GC content higher than obligatory pathogens and symbionts as evidenced across several taxonomic branches [2]. There are many other research works examining the relationship between environment and genome nucleotide composition. As shown in [2], the selection against GC is attributed to the increased energy cost in GTP and CTP synthesis, and the mutational bias toward A/T may therefore confer selective advantage and may impact coding sequence length. The unavailability under stress of cytosine and its tendency to mutate to U/T suggest short coding sequences with longer sequences devoid of cytosine to be a feature of prokaryote genomes [3]. The mechanisms contributing to low GC content in shorter genomes also include the inability to clear uracil from DNA [4] and habitat specificity [5] in conjunction with the increased energy differential in GC synthesis. Recently, environmental considerations toward a relationship between optimal growth temperature and GC genome content have been proposed [6], and then questioned [7] and rebutted [8]. The environmental division between aerobic and anaerobic prokaryotes have shown increased GC content for aerobes [9].

The heterogeneous environment of free-living organisms potentially impacts their genomic GC content via lateral DNA transfer [10, 11]. Pathogenicity islands (PAI), as an important GC content pattern, are tracts of DNA conferring specific pathogenic potential to the host genome. The ‘island’ represents a cluster of genes culminating in a pathogenic potential either integrated into the chromosome or plasmid DNA [12]. Of great interest in this study are the pathogenicity islands for Escherichia coli 536 (O6:K15::H31), an important organism which has been identified and investigated by [13]. The differential GC content of PAIs vs. the background genome arises due to the insert not belonging to the host genome. At the most simplistic level, profiling a genome’s GC content can just indicate genomic islands, but not specifically pathogenicity islands nor the case when the pathogenicity island has similar GC content to the host genome.

Genome reduction and the notion of a ‘core-genome’, essentially the minimal set of coding sequences to sustain life, pose an interesting question: which GC scanning can contribute. In this paper, we investigate genomes categorized as containing low < 35% GC or high > 65% genomic GC. The role of obligate host-environment interrelationships and their impact on GC reduction are of significant interest. From the GC analysis perspective, profiling of low %GC genomes for aberrant regions are theorised to highlight core regions critical to survival. As low %GC is a desirable biological quality, the presence of lower than average GC regions in an already reduced genome may signify regions of more ancient composition or structural importance, for which reductive processes have acted over an extended time. The lack of functional DNA repair enzymes are a feature of

reduced genomes [14]. This poses an interesting conclusion, stating that the default condition for unchecked replicate error is biased toward A+T. Studies have shown A+T richness to be a feature surrounding the replication terminus [15]. Closely associated with A+T biasing near the replication terminus is the translesion repair mechanism [16], a less accurate repair mechanism favouring adenine incorporation. Under the translesion scheme, sequences located toward the replication terminus are hindered by their single copy number for longer durations hence other repair mechanisms (including homologous recombination) are biased against. Interestingly, genes associated with the region proximally preceding the replication terminus evolve at an accelerated rate [17].

From a historical perspective, computational efforts toward genomic segmentation of compositional features has centred around usage and distribution patterns of nucleotides. As summarized by Karlin [18], compositional variances can be analysed through G+C window calculations, dinucleotide bias (often termed genomic signature) and discernable differences in codon and amino acid usage. Detractions from these techniques readily occur when compositional deviance levels are small in magnitude (with respect to the parent background genome) and resolution concerns become apparent through the use of sliding windows of unknown size to define features not fitting the *a priori* consensus length.

Computational models and algorithms generally seek to segment sequences based on identifiable shifts in G+C composition. Several techniques exist beyond simple compositional measures for genome segmentation with variances based on prior threshold setting and resolution capability. The ultimate measure of effectiveness of such algorithms relies on their ability to handle segmentation with the best judgment /refinement / optimization of input parameters. The heterogeneity of target sequence length e.g. PAIs whose length range can extend between 10-200kb [19], poses significant challenges. To begin the background study of more advanced techniques, the reader is advised to keep foremost in their minds, the degree to which decision criteria are optimised, ultimately produces the most desirable outcome. One should also take a view to the generalization of such techniques for profiling features covering gene (1-2 kb) to large lateral tract length scales (~200kb). A segmentation algorithm referred to as 'entropic segmentation' [20] has been proposed to identify regions of abnormal composition for isochore determination at the eukaryote sequence level. Since the statistical foundations call for a confidence level, this key criteria has the potential to limit effectiveness. More recently, entropic methodology has been used to identify non-self genes [21].

Toward targeting pathogenicity islands and lateral transfer regions, a wavelet shrinkage approach has been effectively applied to smooth insignificant G+C variations [22]. Showing similarity with the optimisations made in the entropic segmentation method, wavelet coefficients are eliminated 'shrinkage' with the use of thresholds, effectively removing noise from the data. Outcomes for the method suggest a technique for modeling both small and large features; reported findings included the identification of two putative pathogenicity islands in *N. mengingitidis* [22].

Hidden Markov Models (HMM) have been applied to G+C tract identification [23], with segmentation (coding vs. non-coding) ability in the yeast genome. Toward G+C genome segmentation, the approach concentrated on determining the optimal number of states, a process designed to optimise model effectiveness analogous to the parameter optimisation witnessed in the previous two techniques. The HMM has been used again toward identifying alien DNA, however modeled at the genes codon level [24]. As determined in our analysis, certain genes show invariant nucleotide composition with respect to the surrounding genome, specific steps have been taken in [24], to distinguish them from laterally transferred genomic islands. As with other techniques discussed thus far, the SIGI-HMM utilises several threshold/correction terms (sensitivity term, gap length), to achieve desirable results. The approach utilizing comparative codon usage is thus susceptible to misidentification of HGT event where inserts match the codon usage of the parent.

Compositional methods and their reliance on training signatures, to some degree impart negative outcomes for overall prediction accuracy whereby smaller genomes offer less comparative background to gauge a defined signature. Similarly, the limitation over the number of genes forming phylogenetics studies limits prediction confidence. The 'Wn' method [25], based on templates of size n ($2<n<=8$) offers the ability to rate gene typicality. Thresholds appear in the methodology as T, where the distribution of typicality scores f, whose derivative f' approaches a constant which is defined as the threshold. In a subsequent work [26], support vector machines were used as an enhanced similarity measure, thus gaining a 10% increase in performance. Using a naïve Bayesian classifier [27], putative HGT sequences of length 400bp could be identified at 85% accuracy. The classifier in this instance was based upon genomic signature. The task specifically addresses solving the origin of HGT segments. Actual HGT events were resolved in *H. influenzae* to *N. meningitis*, whereby sliding windows were required to define criteria for the origin of the insert.

Window based schemes more recently include the straight-forward window-based approach [28] and the windowless approach [29] for G+C content profiling. To overcome some of these limitations, a cumulative G+C profiling technique [29], a windowless approach (Z-curve), has been proposed toward identifying regions of G+C abnormality. The idea of this G+C profiling technique is founded on the independent distribution of purine / pyrimidine, amino / keto and weak / strong hydrogen bonds, the later describing GC / AT distribution.

Outcomes include a clear demarcation of the genomic island where G+C abnormality arises. Simplistic G+C window calculations often prove useful toward gaining an overview level of the genomic nucleotide distribution. G+C window calculations consist of dividing a genome into fixed windows of a defined length *l* and sampled at frequency *f*. Windows can be discrete or overlapping, calculations for each window are evaluated as (G+C)/(A+T+G+C) * 100% and represented in the literature as '%G+C'. Recently an Monte-Carlo approach validating sliding windows was proposed [30]. Window sizing remains an issue in the initial and refinement stages which may be biased by sampling too small a window, hence sensitive to local compositional variations.

Using the umbrella term 'Constrained Nucleotide', techniques including GC skews (G-C/C+G), dinucleotide biasing and aberrant nucleotide-amino acid usage, can be employed to elicit evidence supporting horizontal transfers [18]. Common to these processes is the use of windows in which frequency calculations are made relative to other windows and to the broader genome. The genome signature method [31] and the technique proposed by [18] both detect G+C abnormalities through the shift in expected dinucleotide usage relative to a random null model. Codon usage and the biases introduced via several biological processes including lateral transfers, tRNA species biasing and translational efficiency, transcription effectiveness and transcript stability constitute detectable variances, see [32] for a review. Nucleotide substitution rates 'Q' pose an interesting approach for compositional profiling [33]. Central to this approach is principle of genome specific mutational bias and the resulting nucleotide shift it imparts. Horizontal transfers are thus hypothesized to transfer genes from a donor of differing rate matrix into an acceptor whose rate matrix is detectibly different. The key qualification of this approach is the discrimination ability to gauge incongruence between rate matrices. Using orthologous gene displacement as a case study, such models mandate orthologous genes had shared a common ancestor and the inserted sequence evolved according to a Q different from the acceptor genome under analysis. Simulated experimentation has shown a 10-fold reduction in error rates utilising a multitude of predictor statistics encapsulating phylogenetic tree and compositional (codon 3rd position) data.

Other forms of compositional modeling centring on codon usage include correspondence analysis (CA) [34] and the codon adaptation index (CAI) [35]. Correspondence analysis is the statistical analysis of two and multi-way tabular data (contingency tables). Typical associations could entail modeling codon usage vs. genes, as in [35]. The univariate measure CAI [36] aims to show the direction of synonymous codon bias. As before, tabular calculations of codon usage for a set of genes serve as a starting point with CAI methodology. Interpretation of results indicate the fitness of the gene with respect to the tRNA pool of the genome. High CAI values may indicate a greater level of expression. Such comparisons of CAI vs. a reference set of genes, may highlight the deviance encountered by laterally transferred DNA [37].

Taking the phylogenetics approach, genes are mapped according to their similarities with respect to related and distant genomes. The clustering of a gene to distantly related genomes instead of those within close taxonomic classifications can potentially indicate HGT. Algorithms to address these phylogenetics questions centre on the underlying basis causing the incongruence between gene and species trees [38-40].

Several databases exist to aid similarity searching against putative [41] and known genomic islands [42] and more specifically virulence factors [43] and pathogenicity islands [44]. The strategy of comparing putative transferred regions vs. those of known HGT regions via BLAST has been criticised [45]. The key contention arises from the simplicity by which conclusions are drawn from alignment results. Phylogenetic reconstruction is clouded as convoluting factors including compositional biases in thermophilic bacteria/archaea and the sweeping conclusions drawn across taxonomic domains based on limited analysis (again, dataset availability).

In this paper, we address the challenging issue of window size vs. resolution ability again. We propose an iterative scheme, which analyses a defined resolution at every iteration. The discrete nature of the predictions and non-reliance on any prior training or context, makes our approach extend the utility of GC profiling above other techniques. Our technique is applied for the identification of regions displaying GC aberration. As discussed by [18], %GC windows compared to the parent sequence %GC, represent the standard evaluation approach. In this paper, this evaluation measurement will be revised to focus on the identification of regions displaying aberrant GC composition.

Our findings and results are reported and discussed regarding four aspects: (1) the PAI prevalence in the Enterobacteriales and other free-living organisms' genomes, (2) the conserved high GC content of rRNA in reduced genomes, (3) the determination of dynamic G+C DNA sequences in a reducing genome, and (4) the identification of bacterial capsule adhesion and antigen proteins. In particular, we highlight a detected GC aberrant region from the *Pelagibacter ubique* HTCC1062 chromosome and investigate the basis of bacterial attachment in Staphylococcus.

To evaluate the effectiveness of our method, outcomes reported from other studies, including the most recent cumulative GC profiling technique [46], are compared with ours. Findings from these comparative results suggest that our GC content profiling method has a high informational worth, exceeding standard GC profiling, as our method can discover widely known GC patterns and intricate protein pathogenic determinants. The scale

at which detection operates, enables the identification tRNA scale length sequences up to PAI length features, length is not a factor impeding prediction.

Results

As our method seeks to identify regions of distinct GC composition different from the parent sequence, smaller AT rich genomes were expected to indicate little evidence of substantive genomic islands. For the previously stated reasons, namely biochemical and nutrient efficiency, smaller genomes were predicted to be AT homogenous. However, larger free living genomes whose genomic size reflects its ability to acquire survival pathways and to reproduce in complex environments over evolutionary time were expected to contain regions of aberrant GC content through lateral transfer events.

Pathogenicity Island Detection - Enterobacteriales

As mentioned, pathogenicity islands are genomic clusters of genes that confer specific pathogenic capability to the host genome. For the set of Enterobacterial genomes, success rates of the PAI detection varied, with strong detection in *Escherichia coli* 536, Enterococcus faecalis V583 and *Escherichia coli* O157:H7 str. Sakai. The detection results for *Escherichia coli* 536 are summarized in Table 1. Specifically, Table 1 shows our predicted regions of the 5 PAIs in comparison to the BLAST reference positions. Note that the five clusters in Table 1 are ranked with the 5 highest scoring predictions by our method.(

Table 1. BLAST reference position in comparison to our predicted positions.

Island	Reference Position	Predicted Position Range
I	3986088-3965918	3929000...4028170
II	4777771-4815325	4727000...4896341
III	334178-361041	331000...408170
IV	1970000-1972628	1888000...1976170
V	3140691-3177184	3118000...3204170

They also correspond to PAIs I-V in the *Escherichia coli* 536 genome supported by the literature [13, 47].
When compared to GC-Profile, only 2 of the 5 PAIs are detected by 'GC-Profile', whereas our method can locate all the known PAIs in a discrete manner. The effectiveness of the technique presented here also includes non-reliance on input parameters, a detraction of 'GC-Profile' and traditional sliding window approaches. The effect of altering segmentation thresholds in the GC-Profile technique either resulted in too many or too few predictions. To be effective, 'GC-Profile' required a valid and reasonably accurate estimate initially. Without this estimate, the 'GC-Profile' continuous graphical output requires human interpretation to determine regions of significance.

Comparison results with other recent profiling techniques are presented in Table 2.

Table 2. Multiple resolution sampling compared to other techniques, we denote our technique as 'MRS'

Genome	Predictions				Unique Prediction				Prediction Agreement			
	MRS	Wavelet	Z-Curve	Genome Signature	MRS	Wavelet	Z-Curve	Genome Signature	MRS	Wavelet	Z-Curve	Genome Signature
D. radiodurans R1 chr I	22	7	-	-	16	1	-	-	6	6	-	-
H. pylori 26695	5	5	-	-	1	1	-	-	4	4	-	-
H. pylori J99	7	3	-	-	5	1	-	-	2	2	-	-
N. meningitidis Z2491	17	8	-	-	8	0	-	-	9	8	-	-
N. meningitidis MC58	15	10	-	-	6	0	-	-	10	10	-	-
C. glutamicum	7	-	1	-	6	-	1	-	1	-	1	-
V. vulnificus CMCP6 chr. I	7	-	3	10	4 vs Zhang 3 vs vanPassel	-	0	6	3 vs Zhang 4 vs vanPassel	-	3	4
V. vulnificus CMCP6 chr. II	6	-	-	11	4	-	-	9	2	-	-	2

Trends arising from the data presented in Table 2 suggest PAI detection by our MRS method is better than or comparable to other techniques based on GC analysis. In addition to PAI detection, MRS when compared with

the other GC based techniques [22, 29], our method resolves many more genomic features. Presented in [22] are wavelet algorithms and their ability to detect putative pathogenicity islands. For the *Neisseria meningitidis* MC58 chromosome, the putative PAIs at 1428-1455kb and 2222-2232kb were identified in addition to previously identified PAIs [48]. The multiple resolution GC scan in this genome has found additional features. Analysis of the Z2491 chromosome revealed similar findings with the addition of a transposase locus and other core cellular proteins. The *Helicobacter* strains analysed by [22], predictions were in agreement with additional features identified by MRS functioning as DNA restriction or modification enzymes. *Deinococcus radiodurans* R1 chromosome 1 predictions arising from MRS were numerous in comparison to [22], loci of interest included insertion sequences, transposases and additional features. Beyond the predictions by [29] for the *Corynebacterium glutamicum* ATCC 13032 chromosome, our method determined a cell wall / sugar modification locus, rRNA - transposase locus and an integrase locus. The *Vibrio vulnificus* CMCP6 chromosomes I & II revealed several sugar processing proteins on a single loci and a helicase protein on chromosome I. On chromosome II, a plasmid stabilisation - DNA repair locus was identified. A comparison of the results obtained by MRS with those of [31] shows the ability of MRS to determine lateral gene transfer based on %GC content. Using the genomic signature approach, [31] were able to locate regions deviating significantly to MRS. The degree of difference in predictions can be explained by the different approaches used to analyse the genome.

GC regional detection in reduced genomes

Here, we focus on the following three organisms: *Buchnera*, *Staphylococci*, and *Pelagicbacter* each containing low genomic GC content. A reason to carefully study them is their medical and evolutionary importance. We use an example from each genus to demonstrate a key genomic property arising from GC analysis.

The Buchnera aphidicola genome

The *Buchnera aphidicola* species survives as a symbiont within the bacteriocytes of most aphid species [49]. Two strains showing significant GC deviation include *Buchnera aphidicola* str. APS (*Acyrthosiphon pisum*) and *Buchnera aphidicola* str. Sg (*Schizaphis graminum*). In similarity with *Pelagibacter ubique*, the *Buchnera aphidicola* strains are under environmentally limiting conditions. The environment of their host heavily limits symbionts and as such, evolutionary pressures adapt to mold genomic priorities. The genome reduction found in the *Buchnera* strains suggested a close association with its aphid host to the extent of true symbiotic mutual metabolite transfer [5].

However, our analysis on the available *Buchnera aphidicola* strains has identified regions of GC composition deviating from the parent sequence (strains APS & Sg); The genomic region spanning nucleotides 527000...567090 in the *Schizaphis graminum* strain genome and the genomic region spanning nucleotides 527000...571042 in the *Acyrthosiphon pisum* strain genome correspond to rRNA encompassing sequences. To highlight the significance of this result when compared with other techniques, 'GC-Profile' did not determine any region of significance for either genome. Moreover, our GC content profiling method resolved rRNA regions of GC content 0.440 and 0.418 respectively, much greater than the genomic 0.253 and 0.263 respectively. Subsequent BLAST matches coincided with rRNA genes from members of the Enterobacteriales. In addition to rRNA detection, our method has identified a GC deviating region located at 360000-362406 on the *Buchnera aphidicola* str. Bp chromosome [NC_004545.1]. Genbank annotation has defined this region to contain virulence factor MviN, flagellar basal-body rod protein Flgb and flagellar basal-body rod protein FlgC. The presence of a virulence factor in a genome that is in such a stable symbiotic environment was a surprising finding. MviN virulence factor is associated with genomes of other bacteria encoding a membrane protein, including *Salmonella typhimurium* [50], its function has been characterized in the plant pathogen *Pseudomonas viridiflava* [51]. Analysis of four *Buchnera* strains chromosomes revealed` rRNA as the predominant feature detected. The low cluster detection counts of 3, 2, 4, 2 for *Buchnera aphidcola* strains APS, Sg, Bp and Cc respectively indicate very homogenous genomes with respect to nucleotide distribution.

The Staphylococci

The medically significant Staphylococci genus presents several unique features upon genomic GC scans. Confirmatory checks with the PAIDB reveal pathogenicity islands in the genomes under analysis including *aureus* subsp. N315 containing 4, Mu50 containing 5, COL containing 5, MRSA252 containing 3, MSSA476 containing 2, MW2 containing 5 and RF112 containing 2. Other strains including *epidermidis* RP62A containing 2, ATCC 12228 containing 3, *haemolyticus* JCSC1435 containing 0 and *saprophyticus* subsp. *saprophyticus* ATCC 15305 containing 0. With genomes in the order of 2.5-2.9Mb containing %GC of ~32.8%, Staphylococci represent the largest and highest %GC genomes of the low %GC dataset. The large genome and relatively high %GC would indicate the presence of pathogenicity factors, a feature supported by the PAI-DB. GC scans via our method reveal PAI regions and rRNA loci. Other highly significant features were detected, these included surface adherence proteins. The ability for cellular binding serves to enhance the method of

opportunistic action by affixing to susceptible tissue. Genomic %GC scans with our method has resolved a hemagglutinin, several cell wall anchoring proteins, methicillin resistance, map and membrane proteins.
The presence of the cell wall binding proteins in a GC scan was very significant. Firstly, their presence would suggest a level of conservation, secondly a pointer to their origin, since they deviate compositionally from the background genome and thirdly, a correlation with the pathogenic binding ability to specific cell types in the host organism. Subsequent BLASTN [52] analysis of these regions against the 'refseq_genomic' database revealed two important findings. The first finding indicated the cell wall protein of *S. epidermidis* RP62A was derived from a putative Streptococcal origin. The second finding was the resolution of the cell wall component *in S. saprophyticus* subsp. ATCC 15305, the uniqueness of this protein has direct implication for its site of pathogenesis. With a commonly accepted PAI absent from the genome (*S. sapropyticus*), the result was of greater significance. This isolated protein performing specific cellular attachment with the host tissue type was not found in any other organism, related or non-related. The hemagglutinin protein, has been characterised to perform a dual role in binding to human uroepithelia [53] or causing hemagglutination in sheep erythrocytes [54]. As noted [53], substrate specific adhesion is mediated by separate distinct adhesins in *S. aureus*. Unlike *S. aureus*, S. *saprophyticus* has a dual role in hemagglutination and fibronectin binding [55]. The lack of any additional cell wall-anchored proteins, which are abundant in the other Staphylococci, would thus indicate its target specific cell type [56]. The other non-PAI containing *Staphylococcus* genome analysed in this study (*S. haemolyticus* JCSC1435) uncovered many insertion sequences and supports the conclusions of [57] regarding genome plasticity. In summary, the medically important Staphylococci present a heterogenous genome from the %GC viewpoint, containing pathogenicity islands and discernable surface binding proteins. %GC profiling can be used as an indicator of origin as in the *S. epidermidis* finding, however as in the *S. saprophyticus* subsp. ATCC 15305 finding, suggests a unique inclusion and results in attachment to a specific host tissue type Sample predictions are presented in Table 3.

Table 3. Staphylococcus Cell Wall / Capsular protein identification

Genome	Location	Product(s) Function
S. epidermidis RP62A	2317000..2319555	Putative hemagglutinin
S. epidermidis RP62A	2544000-2460110	Putative cell wall anchor protein
S. aureus subsp. aureus COL	59000..61743	surface anchored protein / clumping factor / methicillin resistant surface protein
S. saprophyticus subsp. saprophyticus ATCC 15305	153000..157915	Serine-thronine rich surface antigen / LPXG-motif cell wall anchor domain
S. aureus subsp. aureus N315	2002000..2007497	map protein
S. aureus subsp. aureus N315	717000..719748	membrane protein
S. epidermidis ATCC 12228	2471000..2473440	Ser-Asp rich fibrinogen-binding

The Candidatus Pelagibacter ubique HTCC1062 genome

The information summarised in Table 4, highlights the ability of GC content profiling to determine regions of laterally transferred DNA.

Table 4. Sample subset of Enterobacterial genomic predictions via our method

Genome	Genome %GC	Cluster Count	Significant Clusters	Cluster Score	Genome Pos.	Cluster %GC	Δ%	PAI-DB	Literature PAIs	Significant GC-Profiler
Buchneria										
NC_004061.1	0.253	2	1	9	527000...567090	0.291	0.15	0	-	-
NC_002528.1	0.263	3	1	10	527000...571042	0.294	0.12		-	-
Candidatus										
NC_007205.1	0.297	1	1	23	514000...603797	0.267	-0.10		-	-
NC_008610.1	0.340	6	2	9 9	949000...1021548 1068000...1087137	0.316 0.384	-0.07 0.13		-	-
Escherichia										
NC_008253.1	0.505	20	5	24 12 16 11 10	4727000...4896341 3118000...3204170 1888000...1976170 3929000...4028170 331000...408170	0.472 0.467 0.462 0.469 0.462	-0.07 -0.08 -0.09 -0.07 -0.09	# # # # #	[13, 47]	4728844…4773931 - - 3941033…4001718 -
NC_008563.1	0.505	32	2	16	260000...349406	0.464	-0.08	#	[58]	-

				11	3295000...3392406	0.471	-0.07	#		
NC_002695.1	0.505	29	3	30 13 12	4520000...4718826 3466000...3515956 3679000...3764913	0.473 0.451 0.460	-0.06 -0.05 -0.09	* - -	[37]	2773370…2790867 4588426…4624331 4686135…4693432 5327868…5464674
NC_002655.2	0.504	15	1	15	4596000...4786763	0.471	-0.07	*	[37]	2843583…2861081 4657409…4693314 4755116…4762413 5357853…5494668
NC_007946.1	0.506	28	1	29	4679000...5000608	0.483	-0.05	#	[59]	1071387…1089056 2070120…2100295 2597096…2662406 4771813…4817253 4817254…4842055

# indicates non-existent genome in the PAI-DB, - indicates no match in the PAI-DB for the genome, ′ indicates a match to a candidate PAI in the PAI-DB ′′ indicates a match to a non-candidate PAI in the PAI-DB, emboldened predictions indicate PAIs supported by the literature or PAI-DB

Interestingly, organisms not considered to contain laterally transferred DNA contain clearly defined signals. One such example, the *Pelagibacter ubique* HTCC1062 genome, constitutes a very small 1.3Mb genome [NC_007205.1] belonging to the SAR11 clade [60] which is estimated to account for 25% [60] of all oceanic microbial cells. The genome is recognised to have the minimal macromolecular components for autonomous replication [61]. As stated previously in the paper, GC genomic content is dependent on many factors, with two key, environmental constraints and lateral transfer events. Properties of the *Pelagibacter ubique* HTCC1062 genome include the near absence of redundant and non-functional DNA and the absence of pseudogenes, phage genes, laterally transferred and duplicated genes, whilst maintaining a majority of metabolic pathways [60]. Genomic streamlining [62], a process of genomic refinement to aid replication under selective pressure, may explain the stripping of non-functional DNA from the genome [60]. As expected, the genome has a global GC content of 29.7%. The genomic analysis however has highlighted a region of significantly lower GC content 26.7%. The feature is strong among the predictions with significance at window sizes spanning 1278-82797bp, corresponding to a genomic position of 514000-603797 [NC_007205.1]. Analysis of the genome using 'GC-Profile' resulted in only a single segmentation point (1061902), bearing no relevance to the region discovered by our technique.

The special quality of this genome is its reduced size whilst incorporating core metabolic functionality. The streamlining, reductive evolution in conjunction with its nutrient poor environment [60] are contributing factors toward its low genomic GC content. Analysis of the genomic region using BLASTN [52] and the 'nr' nucleotide database revealed no significant matches, whilst recent investigation have suggested this region to be a hotspot of horizontal gene transfer arising from metagenomic studies [63]. Depicted in Figure 1, we demonstrate the features detected by our algorithm, the contention between genomic streamlining and HGT (typically acting against each other) has been shown as a specific example. We are therefore confident that MRS G+C profiling can indeed highlight such deviations from convention and thus providing worth as an initial stage scanning tool, in this case applicable to incomplete and uncharacterized genomic data.

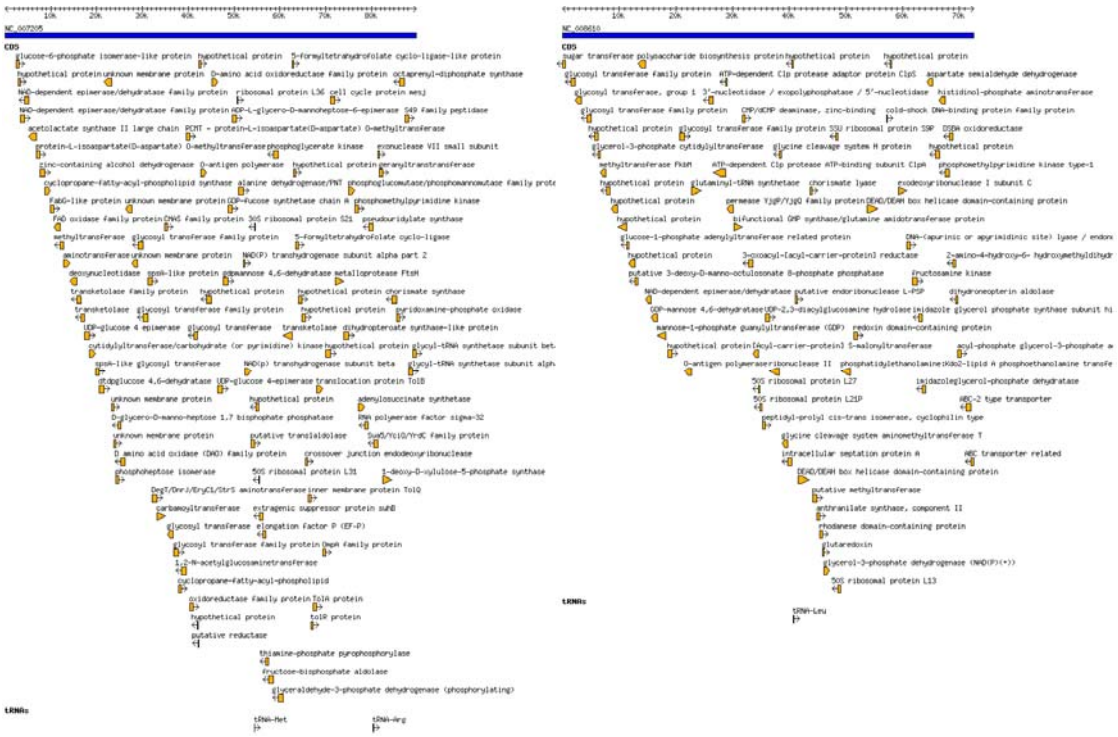

Figure 1. Regions of G+C deviation identified in *Pelagibacter ubique* HTCC1062 (Left) *and Candidatus Ruthia magnifica str*. Cm (Right)

GC regional detection in expansive genomes

Whilst the high %GC genome set consisted of a similar number of chromosomes, the number of predictions and the number of gene products were more than double the low %GC dataset. Organisms presenting significant GC deviation included *Burkholderia* and *Rhodobacter*. The *Burkholderia*, opportunistic pathogens are associated with chronic human conditions including cystic fibrosis. Two regions of interest were identified. The first region corresponded to a unique hypothetical protein on B. sp. 383 chromosome 3 [NC_007509] located at 1139000-1141724. The second region of interest was located on *B. cenocepacia* AU 1054 chromosome 2 [NC_008061] at 436000-438723 (RHS transmembrane protein). *Rhodobacter* form an interesting genus of organisms with respect to their diversity. The multiple chromosome genomic complement, has been studied recently [64] and stated, chromosome II content has rapidly evolved. Results from our analysis revealed important features consistent with rRNA, transposases, membrane transporter proteins and other coding sequences of aberrant G+C composition with respect to the host genome.

Core Chromosomal GC coding sequence identification

Analysis of the relative coding sequence products between the high and low %GC chromosomal sets reveal proteins encoded by high aberrant GC content and their relative occurrence. Figure 2 summarises the findings.

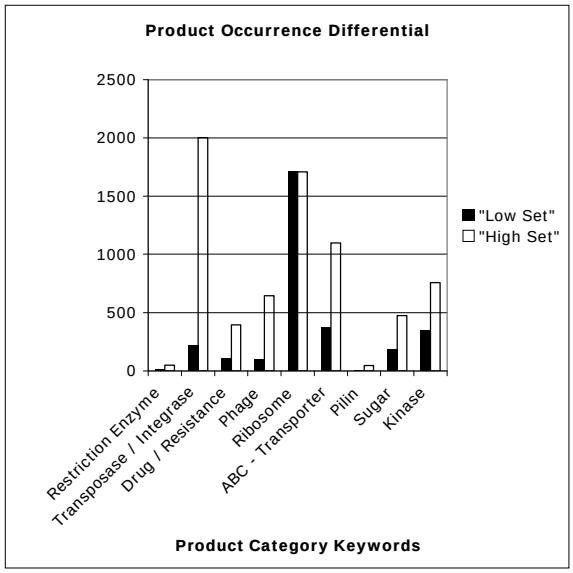

Figure 2. Keyword occurrence differential between high and low chromosome dataset aberrant GC features.

For each dataset, keywords were extracted from the Genbank coding sequence 'product' entries. Two factors explain the observable trends. In comparison to the low %GC set, the high %GC set contains a vastly greater occurrence count for each product except ribosomes. Since high %GC genomes generally constitute large bacterial genomes of free-living organisms, the scope and capability to acquire laterally transferred DNA is increased. Results obtained in this study indicate the low %GC set formed 519 predictions containing 20856 products while the high %GC set formed 1448 predictions containing 58962 products. It can therefore be determined that high %GC genomes contain a greater degree of aberrant GC regions. Of interest are the proteins with the highest occurrences as detailed in Figure 2. The coding sequences identified in Figure 2, represent those typically found in pathogenicity islands. The second feature arising from Figure 2, is the identification of ribosomes. The detection pattern is different to the other product categories as the high and low %GC sets return near identical results. The conclusion drawn from this finding would indicate ribosome coding sequences are of a conserved GC content. In terms of defining core chromosomal coding sequences by GC content, ribosomal sequences are strong candidates.

Evaluation of our algorithm has been shown against previously reported outcomes whilst, further analysis of the sequenced bacterial genomes has shown these findings. (1) In the low GC set, predictions per genome occurred at a ratio of 5:1. (2) In the high GC set, predictions per genome occurred at a ratio of 13:1.

Further classification of the predictions into functional regions are often difficult to gauge as for example, the consensus PAI does not have a defined composition. Utilising Refseq product keywords, categorisation into 4 classes {core, phage, surface, drug resistance} were made. The low GC set resulted in 242, 30, 69 and 15, core, phage, surface and drug resistance clusters. The high GC set resulted in 349, 330, 192 and 42, core, phage, surface and drug resistance clusters. The core clusters contained product keywords including 'ribosomal', 'tRNA', phage grouped, transposon, integrase, surface included adhesins, membrane proteins, transporters, while the drug cluster included, drug / multidrug, antibiotic and resistance keywords.

Discussion

Our method for GC content profiling has revealed many characteristics of genome architecture. By sampling across arbitrary window sizes in a consistent and non-biased scheme, regions of deviant GC composition are determined in a systematic manner. Regions defined as a cluster represent multiple overlapping predictions across multiple resolution scales. This scheme has the advantage of ensuring signal significance. As differences in GC content are not theorized to present sharp and well-defined boundaries down to individual nucleotides, except perhaps in the case of laterally transferred genomic islands, multiple resolution sampling is thus a valid technique. The outcomes include the PAI detection and core genomic constituent identification. PAIs, the laterally transferred genomic islands that confer specific pathogenic potential to the host genome are readily detectible through standard aberrant GC composition analysis. The results achieved for Enterobacteriales have shown that PAI detection seems to be marginal unless the island is well defined with respect to deviance in its genomic GC content. Detection results in the Escherichia genus however were significant.

The disparity between free-living and non free-living organisms' genome composition has been reflected in the number of coding sequences. The free-living Escherichia, Salmonella and Bacillus encode 1500-6000 proteins

in contrast to 500-1000 for obliquely pathogenic bacteria [14]. Of interest is the finding that <40% of the genes identified in the *Escherichia coli* K12 genome have corresponding identities in the EDL933 and CFT073 strains [65]. From Table 4, detection of pathogenicity islands in *Escherichia coli* strains 536, APEC O1, O157:H7, O157:H7 EDL933 and UTI89 have been accomplished along with other genomes of the Enterobacteriales.

Outcomes arising from Table 4 with respect to the Buchnera genomes, have identified regions of aberrant GC content, showing homologous genomic positions as determined by BLASTN [52]. The major contributor to the high %GC in these segments is the rRNA coding sequences. The Buchnera genome is influenced by its symbiotic relationships with the aphid over evolutionary time, leading to its 'reduced' size and structure. The rRNA is a core component of the genome and function of the organism at the molecular level. Conservation of rRNA has been shown and hypothesised to persist outside of genomic compositional pressures [66]. The finding presented via our MRS method suggests this exclusion exists in the genome of two Buchnera strains outside of the powerful influence from genomic reduction.

One of the most successful bacterial clades, SAR11 [67], whose member *Pelagibacter ubique* HTCC1062 has shown abnormal chromosomal GC content in the study. In agreement with the Buchnera studied in this investigation, the genome size and environmental constraints have influenced genome architecture in the Pelagibacter. The small genome size and low %GC content are features associated with reductive evolution. A region of the chromosome (549000..559112, subset excluding rRNA of the prior cluster) was found to contain abnormal GC content of 21% in contrast to the genome %GC of 29.7%.Associated with activated sugar transfer and cell envelope biosynthesis, the coding sequences located in this region represent a peculiar biological feature. Subsequent BLASTN [52] (analysis revealed no significant matches, however the region has been identified as 'hypervariable' [63]. The consequence of locating this region further suggests the role HGT may play in defining genome structure, even within heavily reduced genomes, an ongoing research question [68].

The detection of cell surface and adhesion proteins via GC scanning posed an interesting question. Central to the finding is the question, to what degree and why do these class of proteins present with GC content different from the surrounding genome? Toward answering this question, an analysis of the proteins identified in particular Staphyloccocal species held an explanation. Hemmaglutination proteins and more generally adhesins constitute heavily to the pathogenic potential of an organism. The simplest binding scenario is found in viral protein coats. In the system of viral attachment, hemaglutinin plays a central role as a viral surface glycoprotein. The mechanism of viral attachment is facilitated by hemagglutinin which binds to cell surface receptors on the target cell. The influenza A - hemagglutination binding to sialic-acid-containing cell-surface receptors has been well studied [69].

Owing to the ability for viral DNA to be inserted into host genomes, the transfer of genetic material is one explanation for the presence and divergence of progenitor cell wall attachment proteins in bacteria.

As determined in this study, and well accepted in the community, lateral transfer of genetic material is commonplace between bacteria and accounts for the genomic islands found in many genomes. The role played by genomic GC analysis centres on identifying changes in GC content that may indicate genomic material from a non-host origin. The presence therefore of cell wall attachment proteins in GC scans thus indicates either of two conclusions.

(1) These binding proteins were acquired via lateral transfer events over long stretches of evolutionary time. (2) The compositional nature of the attachment proteins are very highly conserved as in rRNA, as to persist aberrant GC levels relative to the genomic background.

The second conclusion would thus indicate, retention of cellular attachment proteins may potentially serve as a survival strategy. The goal and contribution of this investigation has been developing a computational method to identify GC regions to a more sensitive and biologically relevant grade than other techniques. The outcomes from this process have led to many questions effecting core genome compositional theorems. Conclusions drawn include; the environment shaping to a major degree genomic composition; free-living lifestyles providing models for lateral DNA uptake as in the Enterobacteriales. Identification of aberrant GC elements in already reduced (A+T biased) genomes indicated regions that have implication for underlying ancient genomic processes and newer HGT contributions.

More recently, a comparison of PAI detection tools has revealed a marked disparity between recall and precision [70]. As the definition of a genomic island (GI) is relatively fluid, the approach taken by the analysis centers upon comparing percentage overlap of GI predictions. Datasets used in their comparisons resulted in 118 genomes with 771 predictions in the positive dataset (data containing GI). To contrast existing sequence based approaches contained in this recent comparison, we utilised their published database. Of the 771 predicted islands, the multiple resolution approach (G+C) covered 41% of the islands. This suggested multiple resolution sampling is an effective technique for detecting patterns of genomic G+C variation. These values surpass SIGI-HMM [71] (33%), Centroid [72] (25.7%), DIMOB [73] (35.6%), PAI_IDA (32.2%), but not DINUC [73] (53.3%) or AlienHunter [74] (77%). Alternatively when the multiple resolution algorithm is used in conjunction with codon analysis coverage jumps to 74.5%(GC1), 72.6%(GC2) and 71%(GC3) and when combined (G+C

and codon MRS) 80% coverage, the highest of sequence based tools, refer Table 5. To complement coverage, GI detection rates are presented in Table 5, whereby MRS outperformed all but AlienHunter across the dataset.

Table 5. Multiple resolution sampling compared to other techniques, we denote our technique as 'MRS' Dataset consisted of 118 bacterial chromosomes with annotated GIs according to [70]

Property/ Algorithm	SIGI-HMM	Centroid	DIMOB	PAI_IDA	DINUC	Alien Hunter	MRS G+C	MRS Codon GC1	GC2	GC3	MRS Total
GI Detection	66.3%	33.3%	43.3%	34.5%	67.6%	91.6%	49.3%	75.9%	73.9%	71%	82.9%
GI Coverage	33%	25.7%	35.6%	32.2%	53.3%	77%	41.4%	74.5%	72.6%	68%	80%

Codon usage and observed GC trends have been known for many years, the consensus suggests GC3 (3rd position in the codon), exists as the most redundant position in the codon, hence an indicator of mutational pressure [75]. These mutational biases are thus properties of the organisms' response to environmental conditions [2], an emergent sequence feature that would contrast against those of an appreciably distant host genome in horizontal transfer events. Of the trends identified in our codon analysis, positions one and two of the codon act in slight preference to position three as indicators of GC aberrant genes. Whilst coverage metrics offer a simple guide as to the effectiveness of algorithms, however another metric has to be considered. Specificity, a well-known fact of genomic nucleotide scanning involves informational genes who show marked nucleotide bias, e.g. rRNA and tRNA appearing with experimentation seeking other features such as GIs. Our methodology detects such sequences without distinguishing them from GIs, in addition to other non-GI elements such as phages. This broad scale approach to genomic content profiling does not lend itself to specificity measures based on boundary determination (as underlying window schemes are being employed) nor to strict classification of predictions i.e. (into categories such as GIs, PAIs, phages and informational genes).
Our method may serve as a tool for identifying the variable region of reducing genomes and identifying cell surface adhesion proteins, a pointer to pathogenicity, in addition to PAI detection. With more sequenced genomes, the statements proposed in this paper will be further strengthened. In comparison to other techniques for HGT detection, our method has lessened reliance on prior training parameters. The frequency / codon based methods often rely upon tweaking input parameters to achieve specific desired results. Our approach uses a fixed geometric progression to score windows and standard measures of spread. We therefore make no prior assumptions of the data, hence producing a wide range of result types, PAIs, islands, phages, rRNA, other core coding sequences. In comparison to phylogenetic and database approaches, we do not require a prior dataset to align against. We are therefore confident that our algorithm and its online implementation offers a credible resource for characterising newly sequenced bacterial genomes.

Methods

GC content profiling is a computational process that determines the frequency of the guanine and cytosine nucleotide in a genomic sequence.

Multiple Resolution Sampling

The window size reflects the resolution of sampling. This study does not hold any assumptions regarding the size of the features under analysis. Toward determining an optimal window size (resolution) to screen a genomic sequence for GC variability, an iterative self-reducing progression is proposed, following the general form

$$a_n = a_1 r^{n-1} \quad (1)$$

where a_1 represents the genome length with r the common ratio ½ in this study, and an the window size for the iteration n. This geometric progression starting at the genome length and iteratively halving the window size provides a sound basis for population sampling. The stop criteria for the progression is set as when $a_n < 1000$. Given that large windows have lower resolution and small windows are very susceptible to statistical fluctuations, there exists a point in which the decreasing window size will approximate the required resolution. In this manner, a non-biased sampling set can be achieved. In this work, the shifting offset between two consecutive windows is set as 1000bps.

At every particular iteration, we calculate the %GC score for every window (the window score), and calculate their average (the mean %GC score at this iteration). Then, we transform every individual window score s into a log-odds score according to

$$\ln \frac{s \in S}{\overline{S}} \quad (2)$$

where S is the set of the %GC window scores, and $\overline{S}$ is the mean %GC score in this iteration. The effect of taking the log-odds score relative to the mean window score $\overline{S}$ forms an association test with an asymptotic normal distribution. By equation (2), our method yields a distribution of the log-odds scores for all the windows. We accept all windows whose log-odds score is 3 fold away from the standard deviation. This gives us only the most significantly deviating (peak) windows from the mean. We combine overlapping peak windows into a bigger window.

For successively smaller window sizes, spurious signals cloud true signals. To identify significant signals over arbitrary window sizes, a clustering technique is developed in this work. The central idea is to merge similar and overlapping windows crossing all the iterations. Specifically, we combine all overlapping windows (for all window sizes) into a cluster, and then determine the mean %GC score for all the component windows in this cluster. The most 'deviant GC window' in the cluster is the one that deviates most from the mean.

In summary, the pseudocode is presented below.

Input: FASTA genome sequence
Processing:
For each window size in the progression
    Score the windows at 1Kb intervals
    Compute log-odds score over the windows
    Filter peak windows based on ≥3 standard deviations
    Combine adjacent peak windows
Merge overlapping windows under all window sizes into clusters
Determine the most deviant GC window within every cluster
Output: List of significant clusters with component window size predictions enumerated

Using the 4639675bp Escherichia coli K12 [NC_000913.2] chromosome as an example, the following sets would be generated. Window size set w[] = (4639675, 2319837, 1159918, 579959, 289979, 144989, 72494, 36247, 18123, 9061, 4530, 2265, 1132) bp. Window score set $w_s[w_n][t]$. The 2D set containing n rows with each row populated with scores *t* according to (w_1 - w_n) / 1000bp. As an example,w_7 window size 72494, t = (4639675-72494) / 1000 = 4567 windows for which %GC is obtained. Log-odds calculations follow $\ln(t_n/\text{mean}(t))$. A three standard deviation threshold is applied to the log-odds distribution. Those window values exceeding 3σ are held and a check is performed to merge neighbouring predictions if they also fall outside of 3σ. The final processing step iterates over the windows size set w as an index into the w_s set. Windows showing values above 3σ at each window size are grouped to form a 'cluster'. The resultant cluster is a region of genomic GC variation whereby deviant GC composition is detected at multiple resolution scales. The most deviant window is defined as the value furthest from the mean %GC for the cluster.

Datasets Used for Evaluation

The genomes of prokaryote organisms span a wide GC composition. To elucidate the differences between such genomes, a division based on compositional GC was sought. Two datasets were formed and included genomes with chromosomal GC content less than 35% (101) forming the 'low GC set' whilst chromosomes containing greater than 65% GC (108) formed the 'high GC set'. Genomic data was sourced from the NCBI Genomes resource (ftp://ncbi.nlm.nih.gov/genomes/Bacteria). Benchmark datasets for PAI detection were obtained from the NCBI Genbank database with accessions [AJ488511.1, AJ494981.1, X16664.4, AF135406.1, and AJ617685.1] corresponding to PAI (I-V) in the source genome Escherichia coli 536 [NC_008253.1]. To place our results in context of prior work, the PAI-DB [44] for pathogenicity island lookup was utilised.

Authors' contributions

Scott Mann developed and carried out experimental design, execution and data collection and initial paper development in consultation with Phoebe Chen. All authors whom have approved the final version carried out paper drafting and revisions.

Acknowledgements

The work in this paper was partially supported by Australian Research Council Grants LP0349235 and LX0560616.

References

1. Sueoka N: On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A* 1962, 48:582-592.
2. Rocha EPC, Danchin A: Base composition bias might result from competition for metabolic resources. *Trends in Genetics* 2002, 18(6):291-294.
3. Xia X, Wang H, Xie Z, Carullo M, Huang H, Hickey D: Cytosine Usage Modulates the Correlation between CDS Length and CG Content in Prokaryotic Genomes. *Mol Biol Evol* 2006, 23(7):1450-1454.
4. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH: The complete sequence of the mucosal pathogen Ureaplasma urealyticum. *Nature* 2000, 407(6805):757-762.
5. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H: Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature* 2000, 407(6800):81-86.
6. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G: Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett* 2004, 573(1-3):73-77.
7. Wang HC, Susko E, Roger AJ: On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem Biophys Res Commun* 2006, 342(3):681-684.
8. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G: Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun* 2006, 347(1):1-3.
9. Naya H, Romero H, Zavala A, Alvarez B, Musto H: Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol* 2002, 55(3):260-264.
10. Jain R, Rivera MC, Lake JA: Horizontal gene transfer among genomes: The complexity hypothesis. *PNAS* 1999, 96(7):3801-3806.
11. Ochman H, Davalos LM: The nature and dynamics of bacterial genomes. *Science* 2006, 311(5768):1730-1733.
12. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: Molecular Biology of the Cell, 4 edn: Garland Science; 2002.
13. Dobrindt U, Blum-Oehler G, Nagy G, Schneider G, Johann A, Gottschalk G, Hacker J: Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic Escherichia coli strain 536. *Infect Immun* 2002, 70(11):6365-6372.
14. Moran NA: Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 2002, 108(5):583-586.
15. Daubin V, Perriere G: G+C3 structuring along the genome: a common feature in prokaryotes. *Mol Biol Evol* 2003, 20(4):471-483.
16. Deschavanne P, Filipski J: Correlation of GC content with replication timing and repair mechanisms in weakly expressed E.coli genes. *Nucleic Acids Res* 1995, 23(8):1350-1353.
17. Sharp PM, Shields DC, Wolfe KH, Li WH: Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* 1989, 246(4931):808-810.
18. Karlin S: Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* 2001, 9(7):335-343.
19. Schmidt H, Hensel M: Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* 2004, 17(1):14-56.
20. Oliver JL, Bernaola-Galvan P, Carpena P, Roman-Roldan R: Isochore chromosome maps of eukaryotic genomes. *Gene* 2001, 276(1-2):47-56.
21. Azad RK, Lawrence JG: Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res* 2007, 35(14):4629-4639.
22. Lio P, Vannucci M: Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics* 2000, 16(10):932-940.
23. Peshkin L, Gelfand MS: Segmentation of yeast DNA using hidden Markov models. *Bioinformatics* 1999, 15(12):980-986.
24. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R: Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 2006, 7:142.
25. Tsirigos A, Rigoutsos I: A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res* 2005, 33(3):922-933.
26. Tsirigos A, Rigoutsos I: A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res* 2005, 33(12):3699-3707.
27. Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J: Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res* 2001, 11(8):1404-1409.
28. Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K *et al*: Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. *Genome Res* 2003, 13(7):1572-1579.

29. Zhang R, Zhang CT: A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. *Bioinformatics* 2004, 20(5):612-622.
30. Chatterjee R, Chaudhuri K, Chaudhuri P: On detection and assessment of statistical significance of Genomic Islands. *BMC Genomics* 2008, 9:150.
31. van Passel MW, Bart A, Thygesen HH, Luyf AC, van Kampen AH, van der Ende A: An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics* 2005, 6:163.
32. Ermolaeva MD: Synonymous codon usage in bacteria. *Current issues in molecular biology* 2001, 3(4):91-97.
33. Hamady M, Betterton MD, Knight R: Using the nucleotide substitution rate matrix to detect horizontal gene transfer. *BMC Bioinformatics* 2006, 7:476.
34. Hill MO: Correspondence analysis: a neglected multivariate method. *Appl Statist* 1974, 23(3):340-354.
35. Perriere G, Thioulouse J: Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res* 2002, 30(20):4548-4555.
36. Sharp PM, Li WH: The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987, 15(3):1281-1295.
37. Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB, Blattner FR: Molecular evolution of a pathogenicity island from enterohemorrhagic Escherichia coli O157:H7. *Infect Immun* 1998, 66(8):3810-3817.
38. Boc A, Makarenkov V: New Efficient Algorithm for Detection of Horizontal Gene Transfer Events. In: *Algorithms in Bioinformatics.* Heidelberg: Springer; 2003: 190-201.
39. Suchard MA: Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics* 2005, 170(1):419-431.
40. Hein J: Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical biosciences* 1990, 98(2):185-200.
41. Garcia-Vallve S, Guzman E, Montero MA, Romeu A: HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* 2003, 31(1):187-189.
42. Mantri Y, Williams KP: Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res* 2004, 32(Database issue):D55-58.
43. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q: VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 2005, 33(Database issue):D325-328.
44. Yoon SH, Park YK, Lee S, Choi D, Oh TK, Hur CG, Kim JF: Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res* 2007, 35(Database issue):D395-400.
45. Kurland CG, Canback B, Berg OG: Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A* 2003, 100(17):9658-9662.
46. Gao F, Zhang CT: GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res* 2006, 34(Web Server issue):W686-691.
47. Schneider G, Dobrindt U, Bruggemann H, Nagy G, Janke B, Blum-Oehler G, Buchrieser C, Gottschalk G, Emody L, Hacker J: The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic Escherichia coli strain 536. *Infect Immun* 2004, 72(10):5993-6001.
48. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ *et al*: Complete Genome Sequence of a Virulent Isolate of Streptococcus pneumoniae. *Science* 2001, 293:498-506.
49. Wixon J: Featured Organism: Reductive Evolution in Bacteria: *Buchnera* sp., *Rickettsia Prowazekii* and *Mycobacterium Leprae*. *Comparative and Functional Genomics* 2001, 2(1):44-88.
50. Carsiotis M, Stocker BA, Weinstein DL, O'Brien AD: A Salmonella typhimurium virulence gene linked to flg. *Infect Immun* 1989, 57(11):3276-3280.
51. Czelleng A, Bozso Z, Ott PG, Besenyei E, Varga GJ, Szatmari A, Kiraly L, Klement Z: Identification of virulence-associated genes of Pseudomonas viridiflava activated during infection by use of a novel IVET promoter probing plasmid. *Curr Microbiol* 2006, 52(4):282-286.
52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402.
53. Meyer HG, Wengler-Becker U, Gatermann SG: The hemagglutinin of Staphylococcus saprophyticus is a major adhesin for uroepithelial cells. *Infect Immun* 1996, 64(9):3893-3896.
54. Meyer HG, Muthing J, Gatermann SG: The hemagglutinin of Staphylococcus saprophyticus binds to a protein receptor on sheep erythrocytes. *Med Microbiol Immunol* 1997, 186(1):37-43.

55. Gatermann S, Meyer HG: Staphylococcus saprophyticus hemagglutinin binds fibronectin. *Infect Immun* 1994, 62(10):4556-4563.
56. Kuroda M, Yamashita A, Hirakawa H, Kumano M, Morikawa K, Higashide M, Maruyama A, Inose Y, Matoba K, Toh H *et al*: Whole genome sequence of Staphylococcus saprophyticus reveals the pathogenesis of uncomplicated urinary tract infection. *Proc Natl Acad Sci U S A* 2005, 102(37):13272-13277.
57. Takeuchi F, Watanabe S, Baba T, Yuzawa H, Ito T, Morimoto Y, Kuroda M, Cui L, Takahashi M, Ankai A *et al*: Whole-genome sequencing of staphylococcus haemolyticus uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species. *J Bacteriol* 2005, 187(21):7292-7308.
58. Kariyawasam S, Johnson TJ, Nolan LK: The pap operon of avian pathogenic Escherichia coli strain O1:K1 is located on a novel pathogenicity island. *Infect Immun* 2006, 74(1):744-749.
59. Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P *et al*: Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proc Natl Acad Sci U S A* 2006, 103(15):5977-5982.
60. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M *et al*: Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 2005, 309(5738):1242-1245.
61. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ: Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 2002, 418(6898):630-633.
62. Mira A, Ochman H, Moran NA: Deletional bias and the evolution of bacterial genomes. *Trends Genet* 2001, 17(10):589-596.
63. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ: Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biology direct* 2007, 2:27.
64. Choudhary M, Zanhua X, Fu YX, Kaplan S: Genome analyses of three strains of Rhodobacter sphaeroides: evidence of rapid evolution of chromosome II. *J Bacteriol* 2007, 189(5):1914-1921.
65. Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J *et al*: Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc Natl Acad Sci U S A* 2002, 99(26):17020-17024.
66. Pace NR: Structure and synthesis of the ribosomal ribonucleic acid of prokaryotes. *Bacteriol Rev* 1973, 37(4):562-603.
67. Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ: SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* 2002, 420(6917):806-810.
68. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K *et al*: The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 2007, 5(3):e77.
69. Varki A, National Center for Biotechnology Information (U.S.): Essentials of glycobiology. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.
70. Langille MG, Hsiao WW, Brinkman FS: Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 2008, 9:329.
71. Merkl R: SIGI: score-based identification of genomic islands. *BMC Bioinformatics* 2004, 5:22.
72. Rajan I, Aravamuthan S, Mande SS: Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics* 2007, 23(20):2672-2677.
73. Hsiao W, Wan I, Jones SJ, Brinkman FS: IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* 2003, 19(3):418-420.
74. Vernikos GS, Parkhill J: Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* 2006, 22(18):2196-2203.
75. Muto A, Osawa S: The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* 1987, 84(1):166-169.