DNA Research **20**, 273–286, (2013) doi:10.1093/dnares/dst009
Advance Access publication on 9 April 2013

Re-Annotation of Protein-Coding Genes in 10 Complete Genomes of Neisseriaceae Family by Combining Similarity-Based and Composition-Based Methods

Feng-Biao Guo[1,2], Lifeng Xiong[1], Jade L. L. Teng[1], Kwok-Yung Yuen[1,3,4], Susanna K. P. Lau[1,3,4,*], and Patrick C. Y. Woo[1,3,4,*]

Department of Microbiology, The University of Hong Kong, Special Administrative Region, Hong Kong, People's Republic of China[1]; School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China[2]; State Key Laboratory of Emerging Infectious Diseases, The University of Hong Kong, Special Administrative Region, Hong Kong, People's Republic of China[3] and Research Centre of Infection and Immunology, The University of Hong Kong, Special Administrative Region, Hong Kong, People's Republic of China[4]

*To whom correspondence should be addressed. Tel. 852-22554897. Fax. 852-28551241.
E-mail: pcywoo@hkucc.hku.hk (P.C.Y.W.) or skplau@hkucc.hku.hk (S.K.P.L.)

Edited by Prof. Kenta Nakai
(Received 21 September 2012; accepted 8 March 2013)

Abstract

In this paper, we performed a comprehensive re-annotation of protein-coding genes by a systematic method combining composition- and similarity-based approaches in 10 complete bacterial genomes of the family Neisseriaceae. First, 418 hypothetical genes were predicted as non-coding using the composition-based method and 413 were eliminated from the gene list. Both the scatter plot and cluster of orthologous groups (COG) fraction analyses supported the result. Second, from 20 to 400 hypothetical proteins were assigned with functions in each of the 10 strains based on the homology search. Among newly assigned functions, 397 are so detailed to have definite gene names. Third, 106 genes missed by the original annotations were picked up by an *ab initio* gene finder combined with similarity alignment. Transcriptional experiments validated the effectiveness of this method in *Laribacter hongkongensis* and *Chromobacterium violaceum*. Among the 106 newly found genes, some deserve particular interests. For example, 27 transposases were newly found in *Neiserria meningitidis* alpha14. In *Neiserria gonorrhoeae* NCCP11945, four new genes with putative functions and definite names (*nusG*, *rpsN*, *rpmD* and *infA*) were found and homologues of them usually are essential for survival in bacteria. The updated annotations for the 10 Neisseriaceae genomes provide a more accurate prediction of protein-coding genes and a more detailed functional information of hypothetical proteins. It will benefit research into the lifestyle, metabolism, environmental adaption and pathogenicity of the Neisseriaceae species. The re-annotation procedure could be used directly, or after the adaption of detailed methods, for checking annotations of any other bacterial or archaeal genomes.

Key words: the Neisseriaceae family; re-annotation; newly found genes; eliminated non-coding ORFs; newly assigned functions

1. Introduction

The emergence of next-generation DNA sequencing techniques accelerate tremendously the increment of sequences deposited in public nucleotide databases. The wealth of sequence data stimulates wonderful opportunity to understand the biological process of various living species. To achieve this aim, two of the

essential steps are identifying all protein-coding genes and trying to assign their functions. They are jointly named as genome annotation. The quality of the genome annotation is very vital. If one genome could not be annotated accurately but still submitted to the public database, not only subsequent researches based on it may encounter problems but also annotations of after sequenced closely related genomes would be influenced. The annotation errors may propagate and finally affect more and more genomes. Recognizing this serious problem, Ouzounis and Kap[1] appealed to update regularly the genome annotation by latest database and methods.

As for prokaryotes, dozens of genomes have been re-annotated.[2] Three kinds of re-annotations are often performed in sequenced prokaryotic genomes. First, and rather early, falsely predicted protein-coding genes are eliminated from the original annotation using composition-based methods. For example, Wang and Zhang[3] suggested that 172 annotated genes were very unlikely to encode proteins in the genome of *Vibrio cholerae* based on single-nucleotide frequencies. One of the typical re-annotation cases of archaea was associated with *Aeropyrum pernix* K1, in which protein-coding genes were over-annotated up to 60% by the original sequencing institute.[4–6] It is lucky that this major error has been corrected by using proteome approaches and bioinformatics methods.[7–10] *Amsacta moorei* entomopoxvirus may have the most over-annotated protein-coding genes among sequenced viruses.[11] Guo and Yu[12] suggested that ~38 of 294 originally annotated genes did not encode proteins based on the Z-curve method. By using another graphical method, Yu and Sun[13] confirmed this speculation.

Second, some genes may be missed by the original annotation and could be picked up by the *ab initio* gene finding method and further confirmed by the similarity alignment or transcription and/or protein expression proofs. For example, Zhou *et al*.[14] newly added 278 potential genes by the similarity alignment and another 147 by detectable mRNA transcriptions in the genome of *Xanthomonas campestris*. Very recently, Du *et al*.[15] newly added eight potential genes by the similarity-based method in the archaeon *Pyrobaculum aerophilum*.

Assigning functions to hypothetical proteins constitutes the last kind of re-annotation. This type of re-annotation may be performed by using the homology alignment or by functional genomic experiments. For example, 149 hypothetical proteins were assigned detailed functions according to the strict homology information in the genome of *Erwinia carotovora*.[16] A similar method was employed to the genome of *P. aerophilum* and 80 hypothetical proteins were assigned with functional information.[15] Based on cellular fractions and expression profiles under different culture conditions, Okamoto and Yamada[17] provided general functional information for 126 hypothetical proteins in the genome of *Streptococcus pyogenes*.

In this study, we performed all three types of re-annotation in 10 complete genomes of the Neisseriaceae family. As far as our knowledge goes, the only example involved with all three types of re-annotation is the updated annotation in the genome of *P. aerophilum*.[15] Through them, the outdated annotation may be corrected as far as possible. However, alternative approaches may be utilized to obtain similar results with the systematic method used here. Compared with our previous work,[15] here, we used transcriptional analyses to validate the effectiveness of our method to pick up new genes. The Neisseriaceae family belongs to β-proteobacteria. Among the 10 Neisseriaceae strains analyzed in this work, seven belong to the genus *Neisseria* and all can colonize the mucosal surfaces of many animals. *Neiserria meningitidis*, as one of the most common causes of bacterial meningitis, are most virulent in human.[18] *Laribacter hongkongensis* is a recently sequenced bacterium associated with invasive blood stream infections in patients with liver cirrhosis as well as gastroenteritis and traveler's diarrhea.[19–22] Updated annotations of these bacterial strains would help to understand their pathogenicities and environment adaptation capacities.

2. Material and methods

2.1 Data source

Ten complete genomes of the family Neisseriaceae were included in this work. They were *Chromobacterium violaceum* ATCC 12472 (RefSeq accession number: NC_005085), *L. hongkongensis* HLHK9 (NC_012559), *N. gonorrhoeae* FA 1090 (NC_002946), *N. gonorrhoeae* NCCP11945 (NC_011035), *N. lactamica* 020-06 (NC_014752), *N. meningitidis* 053442 (NC_010120), *N. meningitidis* alpha14 (NC_013016), *N. meningitidis* MC58 (NC_0031112), *N. meningitidis* Z2491 (NC_003116) and *Pseudogulbenkiania* sp. NH8B (NC_016002). Among them, two strains of *N. gonorrhoeae* and four strains of *N. meningitidis* and *L. hongkongensis* are pathogens. In fact, dozens of strains in the family Neisseriaceae have been sequenced.[23] However, the NCBI RefSeq project provides curated annotations only for these strains.[24] In this work, we chose the 10 complete genomes to perform re-annotation.

2.2 Method to pick up missed genes

In each sequenced genome, there are always some bona fide genes that have been missed by the original

annotation. For example, Warren *et al.*[25] uncovered 38 895 intergenic open reading frames (ORFs), readily identified as putative genes by similarity to currently annotated genes, from 1297 prokaryotic replicons based on across-genome alignment. New genes could be confirmed by the similarity alignment or by transcription or proteome analyses. In this work, we first used the ZCURVE[26] program, which is freely available at http://tubic.tju.edu.cn/Zcurve_B/, to pick out all candidate genes that did not have same 5′ terminals with all genes in the original annotation. The candidate may overlap one annotated gene with their sequences but they did not correspond to the same reading frame. Then those candidate new genes would be filtered by blast[27] against the NCBI nr database. If one candidate met the following three conditions, it would be regarded as genuine genes: (i) it had the significant similarity (E-value $< 10^{-20}$, Coverage $> 60\%$ and Identity $> 50\%$) with annotated genes in bacteria beyond the same genus in the database, and it had the similar length with the counterpart (difference $< 20\%$); (ii) it had counterpart in the cluster of orthologous groups (COG) database, i.e. it could be assigned within an existed COG cluster; (iii) it did not overlap annotated genes with any bases or it had a smaller E-value and a higher identity score against functional genes in the other genomes in the case of overlapping. This process has been standardized in this work. For each of the 10 strains, all parameters in the whole process were fixed.

2.3 Method to eliminate over-annotated genes

A composition-based method was used to eliminate over-annotated genes. The method is based on the Z-curve representation of the DNA sequences, which has been successfully used to find genes in various microbes.[3,7,9,12,15,16,28,29] In this analysis, 33 Z-curve variables were adopted,[26] including nine variables of phase-dependent single-nucleotide frequencies[28] and 24 of phase-dependent dinucleotide frequencies.[26] In fact, there are 36 variables denoting phase-dependent dinucleotides.[30] However, long-range correlations between the first and third codon positions tend to be weaker and so 12 variables associated with them were discarded. For details about these variables, refer to Guo *et al.*[26] Besides 33 classifying features, the Fisher linear discrimination algorithm was used to optimally differentiate protein-coding and non-coding sequences, the procedure was as detailed previously.[28] The training set of the classifying model comprised a positive sample set and a negative sample set. The positive sample set was those function-known genes with definite names, e.g. *gyrB* as the name of gene encoding protein of DNA gyrase subunit B. The negative sample was generated by a randomly shuffling sequence of the positive sample and thus destroying its natural structure. After parameters have been trained based on positive and negative samples, all hypothetical proteins would be decided to be genuine genes or falsely annotated non-coding sequences. The latter would be eliminated in the updated annotation. The scatter plot in Figure 1 illustrates the effectiveness of the method to eliminate non-coding ORFs from the collection of hypothetical proteins. A web server has been constructed to check hypothetical genes and eliminate non-coding ORFs in any sequenced bacterial or archaeal genomes, which is freely available at http://147.8.74.24/Zfisher/.

2.4 Method to assign functions to hypothetical proteins

Hypothetical proteins after refining by the above process would be submitted to the nr database. Those with highly significant similarities with function-known genes in the database would be assigned the same functions. To achieve more sensitive results, amino acid (aa) sequences of hypothetical genes were actually aligned against protein sequences translated from the nr nucleotide database.[27] To ensure strict homology, the aligned length covered at least 80% of each gene with the identity of $>70\%$ and the E-value of $<1e-20$. According to the above thresholds, if one hypothetical gene with a translated aa sequence matched two or more proteins with the same functions, then the function information would be transferred to the hypothetical protein.[18]

2.5 Bacterial strains and growth conditions

Laribacter hongkongensis HLHK9 is a clinical isolate in Hong Kong and its complete genome sequence was available recently.[21] It was grown at 37°C, in brain heart infusion (BHI) broth or on BHI agar plates (BD, USA). *Chromobacterium violaceum* ATCC 12472 is a type strain and its complete genome sequence is also available in Genbank.[31] *Chromobacterium violaceum* was cultured in nutrient broth or nutrient agar (Oxoid, England) at 26°C. Unless indicated otherwise, bacteria were cultured to the log-phase for experiment ($\sim$0.6 at OD_{600}).

2.6 Reverse transcription–polymerase chain reaction

The total bacterial RNA was extracted by using the RNeasy mini kit following the manufacturer's instructions (Qiagen, Germany). Genomic DNA was removed

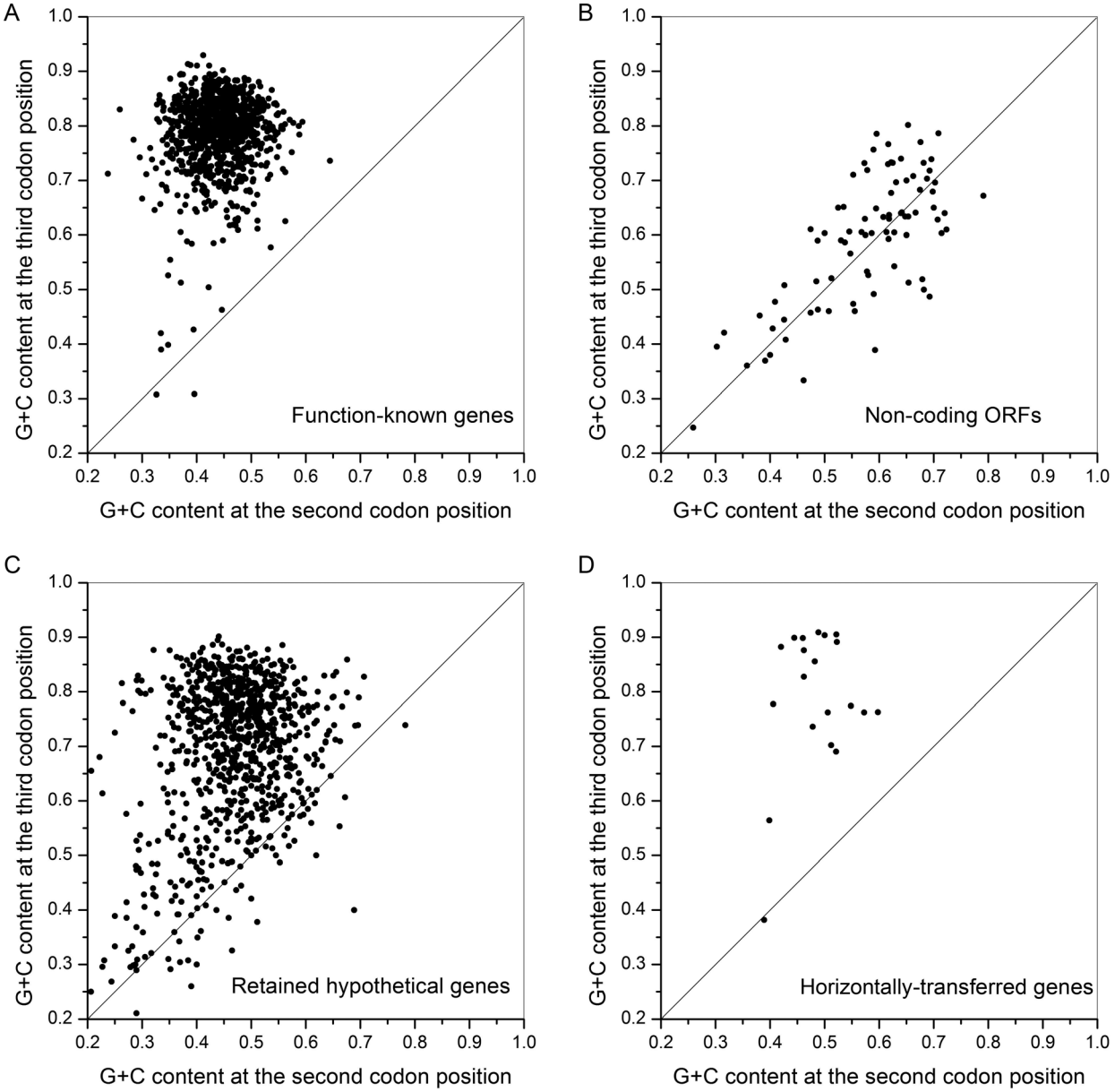

Figure 1. Distribution of GC_2 versus GC_3 for four types of sequences. The *x*-axis indicates the value of GC_2 and the *y*-axis denotes the value of GC_3. (A) For 969 function-known genes in *L. hongkongensis*; (B) for 86 predicted non-coding genes in *L. hongkongensis*; (C) for 915 retained hypothetical genes in *L. hongkongensis* and (D) for 20 horizontally transferred genes in *P. aeruginosa*.

by DNase digestion using RNase-free DNase I (Roche Diagnostic, Switzerland) as described by the manufacturer. Reverse transcription (RT) was performed using Superscript III Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's recommendations. One microlitre of cDNA was used as a template for RT–polymerase chain reaction (PCR) with each specific primer pair. Mock RT–PCR without reverse transcriptase was also conducted as control. Triplicate assays using RNAs extracted in three independent experiments were performed for each target gene.

3. Results

3.1. Eliminated non-coding ORFs and the graphic proof

All *ab initio* gene finders would predict a certain number of non-coding ORFs as potential genes and these predictions constitute false positives of gene annotation.[32] To ensure less species-specific genes be missed from the annotation result, one or more *ab initio* gene finders are necessary in the process of annotating prokaryotic genomes.[33] Often, similarity alignment methods are combined with *ab initio* methods to achieve better results. Because of the

intervention of the latter, it is not inevitable that some non-coding ORFs will appear in the final list of potential genes in every sequenced genome.[34] They may constitute source of annotation errors of closely related genomes and thus should be eliminated from the current annotations. According to the RefSeq[24] annotation for each Neisseriaceae strain, all annotated genes could be classified into three groups. The first group contains only function-known genes with definite names. The last group includes those genes encoding hypothetical proteins. The remaining genes constitute the second group. Obviously, the first group encodes proteins without any uncertainty. Coding potentials of the third group would be doubted to some extent. Therefore, we focus on the third group, hypothetical genes, in this work.

Numbers of genes belong to the first and third groups and genomic characteristics of each strain are listed in Table 1. As can been seen, ratios of the numbers of the first and third class genes to the total gene number vary significantly. For example, *N. gonorrhoeae* FA 1090 has the highest ratio of hypothetical genes, whereas almost the least ratio of function-known genes with definite names. This illustrates that the function annotation in this strain is much poorer compared with the other strains. The variation of the gene ratio is associated with many factors, such as annotation methods, genomic G + C contents and the number of closely related genomes that have been sequenced when annotating the strain.

The training set needed the *Z*-curve method with 33 variables was used to filter over-annotated genes from the RefSeq annotations of 10 Neisseriaceae strains. For each of them, function-known genes with definite names were chosen and shuffled sequences were correspondingly generated. Thus, the training set was obtained. For example, in *L. hongkongensis*, the training set was comprised of 969 function-known genes that correspond to positive samples, and 969 shuffled sequences corresponding to negative samples. Based on the training set, the discriminant model was built. The accuracy of the model based on 5-fold cross-validation is listed in Table 2 for each of the 10 strains. With the model, each hypothetical protein was decided to be a genuine gene or a falsely predicted ORF. Consequently, 86 hypothetical genes were predicted as a non-coding ORF by the *Z*-curve method in the genome of *L. hongkongensis*.

Prediction of 86 hypothetical genes as non-coding is based on the assumption that all protein-coding genes should have similar nucleotide composition in one specific bacterial genome.[28] That is to say, hypothetical genes should have similar composition features with function-known genes in *L. hongkongensis*. If not, they should have been over-annotated as genes. The scatter plot of the nucleotide distribution of 969 function-known genes and 86 predicted non-coding ORFs is shown in Figure 1A and B. As can be seen, non-coding ORFs are distributed far away from function-known genes. In detail, almost all function-known genes lie far above the diagonal and G + C contents of them at the second codon positions are much lower than that at the third codon positions, whereas almost all 86 non-coding ORFs locate around the diagonal, indicating that their G + C contents at the second positions are approximate to that at the third codon positions. The need to encode functional proteins exerts severe constrain on the nucleotide composition of genes.[35] Previous work showed a similar nucleotide distribution of functional genes in seven high G + C prokaryotic genomes.[36,37] Therefore, the distinct nucleotide composition between 86 hypothetical genes and the function-known genes draw them away from being genuine genes. Nucleotide compositions for the 915 retained hypothetical genes are shown in Figure 1C. Different from 86 predicted non-coding ORFs, most of the retained hypothetical genes have a similar distribution of GC_2 versus GC_3 with function-known genes.

The COG database has been widely used during the annotation process of sequenced bacterial genomes.[5,38] Belonging to a COG is believed to be a very reliable evidence of protein-coding genes.[38,39] In *L. hongkongensis*, 954 among 967 function-known genes have been assigned a COG code. However, only 1 of 86 predicted non-coding ORFs has the COG code. Based on the above analyses, these 86 ORFs are very unlikely to encode proteins. COG statistics information for the other nine strains is shown in Table 2. As can been seen, the COG ratio of predicted non-coding ORFs is extremely lower than that of genes with known functions and definite names. Summarily, 7260 among 7426 (97.8%) genes belonging to the first class are assigned with COG codes in the 10 Neisseriaceae genomes. In comparison, only 5 of 418 (1.2%) predicted non-coding ORFs could be assigned to the COG database, indicating that our prediction is much accurate in another sense. These five ORFs with COG codes are likely to constitute falsely predictions of our method because having COG counterparts has been believed to be one of the reliable evidences of encoding proteins. Finally, we only eliminated the remaining 413 hypothetical proteins from the RefSeq annotations in the 10 complete genomes. Details of them are listed in Supplementary Table S1.

As is well known, horizontally transferred genes may also have a different nucleotide composition with core genes to some extent. The DarkHorse database stores horizontally transferred genes in sequenced bacterial genomes. Entries in it are all those predicted by comparative genomes methods

Table 1. Statistical information in genomes of the 10 Neisseriaceae strains

Strains	Published year[a]	Gene density (kb)	Genome size (bp)	Gene number	G + C content (%)	First class gene (ratio)	Third class gene (ratio)
C. violaceum	18 September 2003	0.927	4 751 080	4405	64.8	1494 (33.9%)	1714 (38.9%)
L. hongkongensis	14 April 2009	1.021	3 169 329	3235	62.4	969 (30.0%)	1001 (30.1%)
N. gonorrhoeae FA 1090	15 February 2005	0.930	2 153 922	2002	52.7	266 (13.3%)	800 (40.0%)
N. gonorrhoeae NCCP11945	9 July 2008	1.195	2 232 025	2668	52.4	244 (9.1%)	996 (37.3%)
N. lactamica	16 December 2010	0.888	2 220 606	1972	52.3	744 (37.7%)	657 (33.3%)
N. meningitidis 053442	3 December 2007	0.938	2 153 416	2020	51.7	875 (43.3%)	645 (31.9%)
N. meningitidis Alpha 14	24 July 2009	0.872	2 145 295	1872	52.0	865 (46.2%)	467 (25.0%)
N. meningitidis MC58	10 March 2000	0.908	2 272 360	2063	51.5	806 (39.1%)	810 (39.3%)
N. meningitidis Z2491	30 March 2000	0.874	2 184 406	1909	51.8	459 (24.0%)	669 (35.0%)
Pseudogulbenkiania	8 September 2011	0.926	4 332 995	4012	64.4	704 (17.5%)	831 (20.7%)

[a]Information of published date were extracted from http://www.genomesonline.org/.

Table 2. Accuracy based on 5-fold cross-validation and the COG ratio in each strain

Strain	Accuracy of the method (%)	Class 1 with COG (% ratio)	Predicted non-coding ORFs	Predicted non-coding with COG (% ratio)
C. violaceum	100	1457 (97.5)	88	0 (0)
L. hongkongensis	99.90	954 (98.7)	86	1 (1.2)
N. gonorrhoeae FA 1090	100	263 (98.9)	24	1 (4.2)
N. gonorrhoeae NCCP11945	100	241 (98.9)	56	0 (0)
N. lactamica	99.87	725 (97.4)	8	0 (0)
N. meningitidis 053442	99.77	851 (97.3)	37	1 (2.7)
N. meningitidis Alpha 14	99.77	840 (97.1)	25	1 (4.0)
N. meningitidis MC58	99.75	790 (98.0)	48	0 (0)
N. meningitidis Z2491	100	456 (99.3)	12	0 (0)
Pseudogulbenkiania	99.86	683 (97.0)	34	1 (2.9)

and hence are very reliable. However, the information of HGT is not available for *L. hongkongensis* in the DarkHorse. To circumvent this problem, we fall back on the bacterial strain *Pseudomonas aeruginosa* PA01, which also has a high G + C content, and the distribution of GC_2 versus GC_3 for function-known genes has shown to be similar to that of *L. hongkongensis*.[36] Furthermore, as an early sequenced genome, the gene annotation of it is very reliable and so could be used a good reference. For the genome, 22 horizontally transferred genes were extracted from the DarkHorse. As can be seen from Figure 1D, the nucleotide distribution of horizontally transferred genes tends to be similar to that in Figure 1A. Although the DarkHorse genes have basically similar nucleotide compositions with function-known genes, it does not mean the 86 ORFs in Figure 1B are all definitely non-coding. In fact, there may still be the possibility that some of the genes, particularly those with GC_2 between 0.3 and 0.6, are falsely predicted as non-coding because of their very recent transfers. To investigate the possibility of clusters of horizontally transferred genes, we checked the chromosomal locations of these 86 ORFs and found that they not have any cluster pattern. Therefore, it is sure that, at least, most of the 85 eliminated ORFs do not belong to

horizontally transferred genes and are indeed non-coding. Also note that 11 of the 20 horizontally transferred genes have been assigned the COG code and this ratio is much higher than that of the collective of the 85 eliminated ORFs.

3.2. *Missed genes found by joint methods and their functions*

During the process of annotating sequenced genomes, some genuine genes may be missed because the annotators pursue a balance between the number of all annotated genes and that of finding genuine genes and ensure not too high falsely positive predictions.[14,25,40] The method for picking up missed genes was a two-step process by combining the *ab initio* gene finding and the blast search. With this method, we found a varying number of missed genes with potential functions (Table 3). For example, eight new genes were added in *L. hongkongensis*, and details for them are listed in Table 4. RT–PCR analyses validated the transcriptions of all eight sequences (Fig. 2). Out of the 10 newly found genes in the genome of *C. violaceum*, transcriptions of eight ones are validated with RT–PCR and the exceptions are (Table 6) CV_A0007 and CV_A0010 (Fig. 3). All the amplification with genome DNA was positive (data not shown). Therefore, transcriptional analyses illustrated that the method was effective for picking up new genes and had very high accuracy. However, this method is only applicable to picking up genes having homologues in other genomes but not to strain-specific genes. In the other genomes, the method was directly used and without experimental validation.

Table 3. Numbers of revised genes, which contains newly found genes, hypothetical genes with newly assigned functions, eliminated ORFs and disrupted ORFs

Strains	Newly found genes	Newly assigned functions	Eliminated ORFs	Disrupted ORFs
C. violaceum	10	120	88	0
L. hongkongensis	8	20	85	2
N. gonorrhoeae FA 1090	18	400	23	23
N. gonorrhoeae NCCP11945	9	207	56	0
N. lactamica	5	218	8	8
N. meningitidis 053442	6	214	36	19
N. meningitidis Alpha 14	30	214	24	14
N. meningitidis MC58	11	331	48	27
N. meningitidis Z2491	8	299	12	20
Pseudogulbenkiania	1	46	33	0

Among the eight new genes in *L. hongkongensis*, only two do not have any overlapping bases with annotated genes and they are LHK_A0003 and LHK_A0004. The details of the eight genes and their corresponding functions and the details of PCR primers and conditions used in the validation experiments are listed in Tables 4 and 5, respectively. For each of the remaining six overlapping genes, the region spanned by the two PCR primers does not overlap with the annotated genes and thus could reduce false-positive errors for the transcription trace. The six genes are analyzed as follows. LHK_A0001 overlaps the annotated gene LHK_00511. According to the annotation, the two sequences have the same potential function of encoding the phosphoserine phosphatase. LHK_A0001 has the similarity score of the *E*-value of 10^{-49} and Identity of 55% and the similarity score of LHK_00511 is the *E*-value of 10^{-39} and Identity of 65%. We could not decide which of them is the genuine gene based on the scores. LHK_00511 has the length of only 211 bp, which is much shorter than the length (~669 bp) of known genes with the same function in other genomes. LHK_A0001 has the length of 471 bp. Based on the length information, LHK_A0001 is more likely to encode the phosphoserine phosphatase than LHK_00511. LHK_A0002 overlaps the annotated gene LHK_00916 with only 29 bp. LHK_A0002 is predicted to code for site-specific recombinase and LHK_00916 encodes replicase. Because they have so little overlap and have different functions, it is very likely that they are both genuine genes. LHK_A0005 has the potential function of encoding the Na^+-dependent transporter. LHK_A0006 overlaps 61 bp with one functional gene but they locate on two different DNA strands. LHK_A0007 and the annotated *rpsI* gene (LHK_02777) constitute an interesting overlap. After the blast alignment, both of them are found to be significantly similar to the *rpsI* gene in the other genomes. As shown in Figure 4, LHK_A0007 matches the last 141 bp (57 aa) of the other *rpsI* genes and LHK_02777 matches the first 219 bp (73 aa). LHK_A0007 and LHK_02777 as a whole just constitute a complete *rpsI* gene (130 aa). In fact, LHK_A0007 and LHK_02777 are adjacent and overlapping. For the overlapping part, LHK_A0007 has the correct reading frame but the LHK_02777 does not have the correct frame, according to the other *rpsI* genes. Therefore, it is suggested that there has appeared an event of nucleotide insertion/deletion for the *rpsI* gene in *L. hongkongensis*. After point mutation, the single reading frame changed to two different ORFs. In this work, the two generated segments could be transcribed and should have functions. But we do not

Table 4. Details of the eight newly found genes in the genome of *L. hongkongensis*

ID	Position	COG	Coverage, *E*-value, Identity	Potential function
LHK_A0001	476554–477024 (+)	COG0560E	92%, 1e−49, 55%	Phosphoserine phosphatase
LHK_A0002	880771–881358 (+)	COG1961L	97%, 2e−72, 60%	Site-specific recombinases
LHK_A0003	1391850–1392488 (−)	COG2869C	98%, 6e−66, 51%	Na+-transporting nicotinamide adenine dinucleotide
LHK_A0004	1570970–1571566 (+)	COG2864C	100%, 1e−110, 81%	Thiosulphate reductase cytochrome subunit B
LHK_A0005	1848723–1850306 (−)	COG0733R	63%, 3e−156, 73%	Na+-dependent transporters of the sodium: neurotransmitter symporter family
LHK_A0006	2282651–2283334 (+)	COG0778C	77%, 3e−73, 66%	Putative Cob(II)yrinic acid a,c-diamide reductase (BluB)
LHK_A0007	2660234–2660422 (−)	COG0103J	96%, 2e−28, 88%	Ribosomal protein S9
LHK_A0008	2875641–2876057 (−)	COG0824R	91%, 7e−55, 64%	Predicted thioesterase

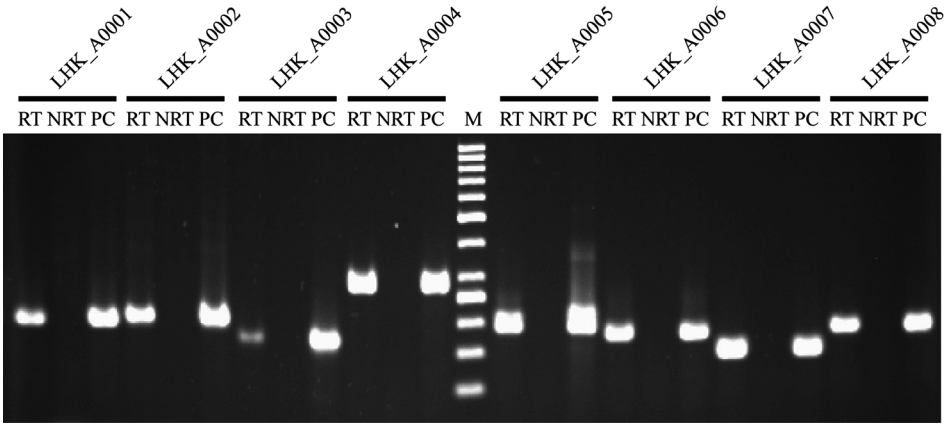

Figure 2. RT–PCR confirmations of eight newly found genes in *L. hongkongensis*. mRNAs corresponding to candidate genes were evaluated by RT–PCR (RT). We used no transcriptase-containing sample as negative control (NRT) and PCR with genomic DNA as a positive control

Table 5. PCR primers and annealing temperature for the eight newly found genes in *L. hongkongensis*

ID	Primer pair	Primer sequence	Annealing temperature
LHK_A0001	LPW20036 LPW20037	GCATGCCGAATTCCTCGAAG TCCGGGCCTTCTTCCAGTTC	60
LHK_A0002	LPW20038 LPW20039	ACGCGCTTTGATTCGGGAAC GCGTTCGCATAACCGTACAG	60
LHK_A0003	LPW20046 LPW20047	TGGCCAATCCGATCGTGAC CCTCCTGAGCGTTTCAAG	55
LHK_A0004	LPW19946 LPW19947	ATTCATCCGTCGTGGCTAAG TGACCACAAGCAGCCACATC	65
LHK_A0005	LPW19950 LPW19951	TGGGCGCCATGATCACCTAC CGGCAGGCATGGTGATGAAG	65
LHK_A0006	LPW19952 LPW19953	TGGCGCTTCATCCGCATCAC TCCGGCATCAGTACCGAGAC	65
LHK_A0007	LPW20545 LPW20546	CATCACCCGTGCCCTGAT CTTGGAGAACTGCTTGCG	60
LHK_A0008	LPW20048 LPW20049	CTGACACCCGGTGCAGTTTC CTGGCGTAATCCACCCAGAC	60

know whether they have the same function of encoding ribosomal protein S9. Finally, LHK_A0008 has possible function of coding for thioesterase.

For the 10 newly found potential genes in *C. violaceum*, there are not so serious cases of overlapping with annotated genes as in *L. hongkongensis*. Either

Table 6. Details of the 10 newly found genes in the genome of *C. violaceum*

ID	Position	COG	Coverage, *E*-value, Identity	Potential function
CV_A0001	486035–487360 (+)	COG0845Q	100%, 0, 72%	Membrane-fusion protein
CV_A0002	487503–489461 (+)	COG0750M	100%, 0, 71%	Membrane-associated Zn-dependent proteases
CV_A0003	1025430–1025810 (+)	COG3536S	89%, 4e−67, 83%	Uncharacterized bacterial conserved region (BCR)
CV_A0004	1026544–1027152 (+)	COG3165S	95%, 6e−88, 66%	Uncharacterized BCR
CV_A0005	1956990–1958684 (+)	COG0654HC	88%, 3e−177, 53%	2-polyprenyl-6-methoxyphenol hydroxylase
CV_A0006	2305072–2305416 (−)	COG3628R	99%, 4e−53, 71%	Phage baseplate assembly protein
CV_A0007	2362585–2363220 (−)	COG0693R	92%, 1e−70, 54%	Intracellular protease/amidase
CV_A0008	4207595–4208386 (+)	COG0614P	85%, 6e−94, 67%	ABC-type cobalamin/Fe3+-siderophores transport systems
CV_A0009	4462117–4463199 (−)	COG0438M	96%, 5e−137, 62%	Predicted glycosyltransferases
CV_A0010	4588140–4588436 (−)	COG1872S	98%, 3e−47, 75%	Uncharacterized ancient conserved region

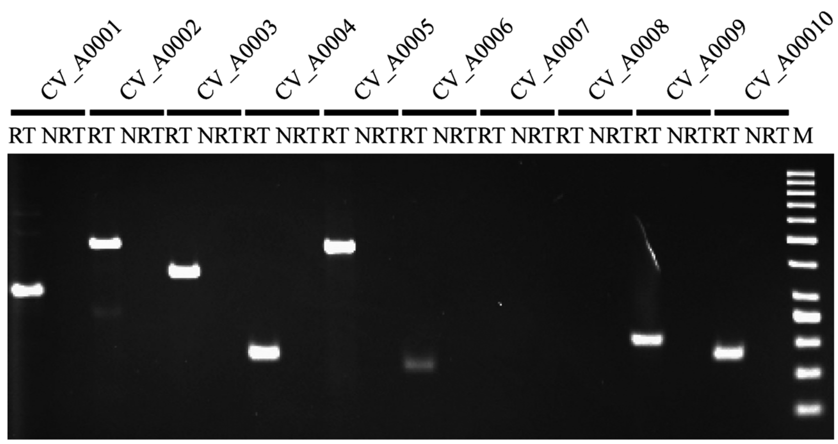

Figure 3. RT–PCR confirmations of newly found potential genes in *C. violaceum*.

```
LHK_02777  1   MNGKYYYGTGRRKSAVARVFMIKGSGKITVNGKPVDEYFARETGRMVIRQPLVLTEHTES  60
               MNGKYYYGTGRRKS+VARVFM KGSG+I VNGKPVDEYFARETGRMVIRQPL  LTEH ES
Subject    1   MNGKYYYGTGRRKSSVARVFMQKGSGQIIVNGKPVDEYFARETGRMVIRQPLALTEHLES  60

LHK_02777  61  FDILVNVTGGGETGPGRCSAPRH  83
               FDI VNV GGGET   G+   A RH
Subject    61  FDIKVNVLGGGET--GQAGAIRH  81
                         73

LHK_A0007  3  AKPGQAGAVRHGITRALIDFSAELKPALSNAGFVTRDAREVERKKVGLHKARRRKQFSKR  62
              + GQAGA+RHGITRALIDFSAELKPALS+AGFVTRDAREVERKKVGL KARR KQFSKR
Subject    61 GETGQAGAIRHGITRALIDFSAELKPALSHAGFVTRDAREVERKKVGLRKARRAKQFSKR  130
                74
```

Figure 4. Matching relationship of aa sequences encoded by LHK_02777 and LHK_A0007 with the RpsI protein in the genome of *Pseudogulbenkiania*. The plot is adapted from the result generated by the NCBI blast application. In the search, the query is LHK_02777 and LHK_A0007, respectively, whereas the rpsI protein constitutes the subject.

Table 7. PCR primers and annealing temperature for the 10 newly found potential genes in *C. violaceum*

ID	Primer pair	Primer sequence	Annealing temperature
1	LPW21817	CTGGCATTGACCGATGAC	55
	LPW21818	CGAAGCGTTGGGATACAG	
2	LPW21819	CTGTATCGCCTGGTGTTG	45
	LPW21820	GCCCTTGCTCTGCAAATC	
3	LPW21821	TGCCCATGTCAGGACTTG	55
	LPW21822	GAGCTTGTCCAGGTATTG	
4	LPW21823	GATTTGTCGCGGGTGTTC	60
	LPW21824	GAAGCGTTGAACCAGATG	
5	LPW21825	CATGAGGTTAGCCCTTTC	55
	LPW21826	GCCATCGACGAAATACAG	
6	LPW21827	CAGTGCATCCGCATCATC	60
	LPW21828	GCTCCCATTGCCGAATAG	
7	LPW21829	CAGGAAGACCTGTCTTAC	55
	LPW21830	CTGGCAAAGTCCTCTTCC	
8	LPW21835	CGCAGCTGAAGCAGCTGAAG	48
	LPW21836	CGGCTTGAAACCGTTGAG	
9	LPW21837	TTGAGCTACGGCATAGAC	55
	LPW21838	GCCAGCCGTTTCAGATTC	
10	LPW21839	CGTCTGACGCTGCATGTG	55
	LPW21840	GTCGCCGGACAACAATTC	

the overlapping part is shorter than 15 bp or the overlapping gene is annotated as hypothetical protein that has not significant similarity with known genes in the public database. Details of the 10 genes and their corresponding functions and the details of PCR primers and conditions used in the validation experiments are listed in Tables 6 and 7, respectively. The two negative samples with RT–PCR analysis may constitute falsely positive predictions of our method, or alternatively, they are expressed only in special conditions.

In the other nine Neisseriaceae strains, there were also newly found genes with potential functions. In *N. gonorrhoeae* FA 1090, only 1 of 18 newly added genes overlaps the annotated genes and the overlapping part is 20 bp. Among the 18 genes, the function information of four are so detailed to have definite gene names and they encode thiol:disulfide interchange protein DsbD, sulphate ABC transporter permease protein CysU, 23S rRNA (guanosine-2′-*O*-)-methyltransferase RlmB and septum site-determining protein MinD, respectively. In *N. gonorrhoeae* NCCP11945, nine new genes are found. Among them, five do not overlap annotated genes. Four of the five genes have putative important function based on very high similarity and they encode transcription antiterminator (Nusg), 30S ribosomal protein S14 (RpsN), 30S ribosomal protein L30 (RpmD) and translation initiation factor IF-1 (InfA). In *N. lactamica*, none of the five newly added genes overlap annotated genes. Among them, three encode DNA transport competence protein (ComeA), one encodes transposases and one encodes the TonB-dependent receptor. In *Pseudogulbenkiania*, only one gene is added. Interestingly, this gene has the same 5′ terminal with the annotated gene (NH8B_2210) but they do not have the same stop codon. The NH8B_2210 is much longer than the newly added one and they have the same reading frame. According to the RefSeq annotation for NH8B_2210, the original authors seemed to have predicted that there is a stop codon treated as the selenocystein codon. By blast against the public database, the newly added gene is shown to be a more reliable prediction because it has the same length with counterparts in distantly related species where as the annotated does not.

In the four strains of *N. meningitidis*, from 6 to 30 new genes were added and very few of them overlap the annotated genes. The strain alpha14 *N. meningitidis* has the most found new genes and interestingly 27 among the 30 new genes encode transposases. Although their aa identities with known transposases in genomes beyond the *Neisseria* genus tend to be just slightly higher than 50%, the identity is higher than 80% for each of them with counterparts in the same genus. Furthermore, the coverage at each case is greater than 90%. Therefore, predicted functions of encoding transposases for the 27 new genes are reliable. According to the RefSeq annotation, none of the genes code for transposases in this strain. However, 55, 29 and 33 transposases have been annotated in the other three strains of *N. meningitidis*. The lower sensitivity of the *ab initio* gene finder used in the annotation of *N. meningitidis* alpha14 is suggested to be responsible for the missing of transposases. In fact, some of transposases, which aid the integration insertion of genomic islands or single horizontally transferred gene, tend to own abnormal nucleotide composition and are easily missed by composition-based programmes.[41] In addition, some genes with important functions have been added in the *N. meningitidis* genomes, such as the haemoglobin receptor in 053442, transcription elongation factor (*greA*) in alpha14, protein methyltransferase (*hemK*) in Z2491, leucyl aminopeptidase (*pepA*) and allophanate hydrolase subunit 2 in MC58. Details of the 106 newly found genes in the 10 Neisseriaceae strains are listed in Supplementary Table S2.

3.3. *Hypothetical proteins with newly assigned functions*

For the genomes sequenced several years ago, functional information may be outdated.[16,29] Especially, some hypothetical genes may have functional counterparts in current databases,[16] whereas they are

still annotated as hypothetical. This section is aiming to provide functional information for hypothetical genes in the 10 Neisseriaceae strains by using the similarity search method. To ensure reliable function transfer, severe homologous conditions were adopted. After the blast search, varying numbers of hypothetical proteins were assigned functions in the 10 Neisseriaceae genomes (Table 3). Among them, *L. hongkongensis* and *Pseudogulbenkiania* have the least numbers perhaps because their genomes were most recently sequenced. In the two strains of the *N. gonorrhoeae*, the strain FA 1090 has the larger number of assigned functions and it just was sequenced earlier than NCCP11945.[23,42] Among the four strains of the species, *N. meningitidis* MC58 has the largest number and it just was sequenced earliest among them.[23] Therefore, the number of hypothetical proteins with newly assigned functions may be tightly associated with sequenced time of the genome. In fact, if one genome has been sequenced at earlier year, it should have more functional counterparts in the current public database but not existed at that time.[29] The original annotation would be more outdated for earlier sequenced genome in the sense of function information.

Among newly assigned functions, some have been provided definite names (Table 8) and this transferred annotation information should be unquestionable. For example, the hypothetical gene LHK_02863 in *L. hongkongensis* has been assigned not only the function encoding iron-sulphur cluster assembly protein but also the definite name 'IscA'. Some other genes were assigned with detailed functions and would be reliable with this function because stringent homologous criteria were adopted. However, there are still some hypothetical proteins with only general functions, such as membrane proteins, lipoproteins and periplasmic proteins (Table 8). This type of rough function information would give help for the determination of more detailed functions. Interestingly, the number of assigned membrane proteins tends to be larger than periplasmic proteins and much larger than lipoproteins in most cases (Table 8). It is suggested that this information corresponds to their natural ranks existing in the genome. Details of assigned functions for each of the 10 Neisseriaceae strains are listed in Supplementary Table S3.

3.4. Disrupted ORFs

During the process of genome re-annotation for the 10 Neisseriaceae strains, we found an interesting phenomenon. In a specific genome, two adjacent ORFs, which are predicted as genes by the ZCURVE program, have the same function with known genes in the other species. The total length of the two ORFs just corresponds to that of the known counterparts or just a little different from them. Based on similarity scores, both of the two ORFs should be predicted as genes with the corresponding function. However, either of them is much shorter than the functional counterpart and could not constitute a homologue in the sense of length information. Perrodou *et al.*[43] stated that unrecognized frameshifts, in-frame stop codons and sequencing errors often led to interrupted coding sequences. Very recently, Sharma *et al.*[44] performed a pilot study on bacterial genes with disrupted ORFs. Their results indicated that many disrupted genes likely utilized the non-standard decoding mechanisms: programmed ribosomal frameshifting and programmed transcriptional realignment. Given that our recognized adjacent ORFs have identical functions, they should originate from the disruption of a longer gene by a frameshift or an in-frame stop codon and rarely by the sequencing error. Numbers of disrupted ORFs are listed in Table 3 and details of them are illustrated in Supplementary Table S4. Totally, 111 disrupted ORFs (or partial genes) were identified in six strains of the genus *Neisseria* and two in *L. hongkongensis*.

Table 8. Among hypothetical gene with assigned functions, the numbers of genes with definite names, those encoding membrane proteins, lipoproteins and periplasmic proteins

Strains	With definite name	Membrane protein	Lipoprotein	Periplasmic protein
C. violaceum	29	5	1	1
L. hongkongensis	9	3	0	0
N. gonorrhoeae FA 1090	105	72	25	31
N. gonorrhoeae NCCP11945	34	25	6	8
N. lactamica	39	25	6	21
N. meningitidis 053442	30	21	4	19
N. meningitidis Alpha 14	19	67	16	26
N. meningitidis MC58	47	77	26	17
N. meningitidis Z2491	78	33	8	18
Pseudogulbenkiania	7	9	5	0

According to Sharma *et al*.,[44] some of the disrupted ORFs still encode proteins. Consistent with this, both of the disrupted ORFs LHK_A0007 (newly predicted) and LHK_02777 (originally annotated) are found to be transcribed in *L. hongkongensis* cells by the PCR experiment. Further wet experiment on functions of the other disrupted ORFs would be interesting.

4. Discussions

The genome annotation would be outdated with the functional information when several years passed since the sequencing. Latest database may contain new functional genes which were not yet assigned functional information when the analyzed genome was sequenced.[29] Those newly added functional information would provide the source of function transfer for some hypothetical genes in the analyzed genome. Sometimes, a few genes missed by the original annotation may be found with the similarity alignment.[25] In this work, a total of 2069 hypothetical genes were assigned with general or detailed functions based on homology search against latest database. Among the 10 Neisseriaceae strains, *N. gonorrhoeae* FA 1090 and *N. meningitidis* MC58 and Z2491 were sequenced and annotated earlier.[23] Correspondingly, the numbers of newly assigned functions in them are the largest. Therefore, the time from the sequencing is longer and the function annotation is more likely to be outdated. As for newly found genes, all 10 strains except *N. meningitidis* alpha14 contain a small number. In fact, it is rather difficult to find genes missed by the original annotation using the similarity-based method. When annotating the genome originally, predicted genes, particularly those having similar counterparts, would be retained as far as possible. Therefore, newly identified genes usually are only those having counterparts newly added in the database. As an exception, as many as 30 genes were newly found in the strain of *N. meningitidis* alpha14 and 27 of them correspond to transposases. We speculate that the less sensitivity of the *ab initio* programme, which has been used in the original annotation, for the external genes leads to the missing of the 27 genes associated with insertion function.

Besides the published date, the quality of the original annotation is another determining factor of the extent of errors.[33] As an example, *N. gonorrhoeae* NCCP11945 has the highest gene density according to the RefSeq annotations among the seven *Neisseria* strains. Correspondingly, the largest number of hypothetical genes has been excluded from this strain. When annotating this genome, only one *ab initio* gene finder was used.[42] Usually, two or more *ab initio* programmes are used in annotating the other bacterial genomes. Based on the very abnormal gene density of the strain, we suppose that there may still be non-coding ORFs, besides the 56 ones excluded by us. Besides the annotation method and sequencing time, some other factors also cause the different extent of errors. For example, genes in high G + C bacterial genomes are shown to be difficultly predicted with high accuracy and this is caused by the fact that many long ORFs appear in this type of genomes.[32] In addition, the genome could be difficultly predicted accurately if there are only distantly related species in the public databases.[33]

We should note that there still probably are non-coding ORFs and missed genes after the present updated annotation in the 10 Neisseriaceae strains. To assure reliable results, we picked up only those sequences with very higher similarities when identifying missed genes and we used the method with lower specificity when excluding non-coding ORFs. In addition, some thoroughly new genes, which have not significant similarities with any known genes, may be found by combing *ab initio* programme and wet experimental validation. However, the 106 newly found genes are those having similar counterparts based on latest databases. Also note that different detailed methods may be used to check the annotations of the other bacterial genomes, although the present systematic method has shown to be effective and reliable. For example, the recently developed *ab initio* gene finder Prodigal,[32] which has lower false positives could replace or jointly used with ZCURVE 1.0. Here, we used only the latter because the use of both of them generates basically consistent results (data not shown). The RPGM program,[39] which is also based on the graphical representation of the DNA sequence, could be chosen as an alternative to eliminate non-coding ORFs.

Supplementary data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Acknowledgements: The authors are grateful to the editor and the anonymous reviewers for their valuable suggestions and comments, which has led to the improvement of this article. We thank Mr Zhong-Shan Cheng for helpful discussions. We also thank Miss Annette Y. P. Wong for helps in constructing the website.**:** *Conflict of Interest statement*. None declared.

Funding

This work was supported by the Hong Kong Scholars Program (XJ2011005), National Natural

Science Foundation of China (31071109) and the programme for New Century Excellent Talents in University (NCET-11-0059).

References

1. Ouzounis, C.A. and Karp, P.D. 2002, The past, present and future of genome-wide re-annotation, *Genome Biol.*, **3**, COMMENT2001.
2. van den Berg, B.H., McCarthy, F.M., Lamont, S.J. and Burgess, S.C. 2010, Re-annotation is an essential step in systems biology modeling of functional genomics data, *PLoS One*, **5**, e10642.
3. Wang, J. and Zhang, C.T. 2001, Identification of protein-coding genes in the genome of *Vibrio cholerae* with more than 98% accuracy using occurrence frequencies of single nucleotides, *Eur. J. Biochem.*, **268**, 4261–8.
4. Kawarabayasi, Y., Hino, Y., Horikawa, H., et al. 1999, Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1, *DNA Res.*, **6**, 83–101, 145–52.
5. Natale, D.A., Shankavaram, U.T., Galperin, M.Y., Wolf, Y.I., Aravind, L. and Koonin, E.V. 2000, Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs), *Genome Biol.*, **1**, RESEARCH0009.
6. Bocs, S., Danchin, A. and Medigue, C. 2002, Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes, *BMC Bioinformatics*, **3**, 5.
7. Guo, F.B., Wang, J. and Zhang, C.T. 2004, Gene recognition based on nucleotide distribution of ORFs in a hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1, *DNA Res.*, **11**, 361–70.
8. Yamazaki, S., Yamazaki, J., Nishijima, K., et al. 2006, Proteome analysis of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1, *Mol. Cell. Proteomics*, **5**, 811–23.
9. Guo, F.B. and Lin, Y. 2009, Identify protein-coding genes in the genomes of *Aeropyrum pernix* K1 and *Chlorobium tepidum* TLS, *J. Biomol. Struct. Dyn.*, **26**, 413–20.
10. Yu, J.F., Jiang, D.K., Xiao, K., Jin, Y., Wang, J.H. and Sun, X. 2012, Discriminate the falsely predicted protein-coding genes in *Aeropyrum Pernix* K1 genome based on graphical representation, *MATCH Commun. Math. Comput. Chem.*, **67**, 845–66.
11. Bawden, A.L., Glassberg, K.J., Diggans, J., Shaw, R., Farmerie, W. and Moyer, R.W. 2000, Complete genomic sequence of the *Amsacta moorei entomopoxvirus*: analysis and comparison with other poxviruses, *Virology*, **274**, 120–39.
12. Guo, F.B. and Yu, X.J. 2007, Re-prediction of protein-coding genes in the genome of *Amsacta moorei entomopoxvirus*, *J. Virol. Methods*, **146**, 389–92.
13. Yu, J.F. and Sun, X. 2010, Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence, *J. Comput. Chem.*, **31**, 2126–35.
14. Zhou, L., Vorhölter, F.J., He, Y.Q., et al. 2011, Gene discovery by genome-wide CDS re-prediction and microarray-based transcriptional analysis in phytopathogen *Xanthomonas campestris*, *BMC Genomics*, **12**, 359.
15. Du, M.Z., Guo, F.B. and Chen, Y.Y. 2011, Gene re-annotation in genome of the extremophile *Pyrobaculum aerophilum* by using bioinformatics methods, *J. Biomol. Struct. Dyn.*, **29**, 391–401.
16. Chen, L.L., Ma, B.G. and Gao, N. 2008, Reannotation of hypothetical ORFs in plant pathogen *Erwinia carotovora* subsp. *atroseptica* SCRI1043, *FEBS J.*, **275**, 198–206.
17. Okamoto, A. and Yamada, K. 2011, Proteome driven re-evaluation and functional annotation of the *Streptococcus pyogenes* SF370 genome, *BMC Microbiol.*, **11**, 249.
18. Tettelin, H., Saunders, N.J., Heidelberg, J., et al. 2000, Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58, *Science*, **287**, 1809–15.
19. Yuen, K.Y., Woo, P.C., Teng, J.L., Leung, K.W., Wong, M.K. and Lau, S.K. 2001, *Laribacter hongkongensis* gen. nov., sp. nov., a novel Gram-negative bacterium isolated from a cirrhotic patient with bacteremia and empyema, *J. Clin. Microbiol.*, **39**, 4227–32.
20. Woo, P.C., Lau, S.K., Teng, J.L., et al. 2004, Association of *Laribacter hongkongensis* in community-acquired gastroenteritis with travel and eating fish: a multicentre case-control study, *Lancet*, **363**, 1941–7.
21. Woo, P.C., Lau, S.K., Tse, H., et al. 2009, The complete genome and proteome of *Laribacter hongkongensis* reveal potential mechanisms for adaptations to different temperatures and habitats, *PLoS Genet.*, **25**, e1000416.
22. Kim, D.S., Wi, Y.M., Choi, J.Y., Peck, K.R., Song, J.H. and Ko, K.S. 2011, Bacteremia caused by *Laribacter hongkongensis* misidentified as *Acinetobacter lwoffii*: report of the first case in Korea, *J. Korean Med. Sci.*, **26**, 679–81.
23. Pagani, I., Liolios, K., Jansson, J., et al. 2012, The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata, *Nucleic Acids Res.*, **40**, D571–9.
24. Pruitt, K.D., Tatusova, T. and Maglott, D.R. 2007, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.*, **35**, D61–5.
25. Warren, A.S., Archuleta, J., Feng, W.C. and Setubal, J.C. 2010, Missing genes in the annotation of prokaryotic genomes, *BMC Bioinformatics*, **11**, 131.
26. Guo, F.B., Ou, H.Y. and Zhang, C.T. 2003, ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes, *Nucleic Acids Res.*, **31**, 1780–9.
27. Altschul, S.F., Wootton, J.C., Gertz, E.M., Agarwala, R., Morgulis, A., Schaffer, A.A. and Yu, Y.K. 2005, Protein database searches using compositionally adjusted substitution matrices, *FEBS J.*, **272**, 5101–9.
28. Zhang, C.T. and Wang, J. 2000, Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve, *Nucleic Acids Res.*, **28**, 2804–14.
29. Gao, N., Chen, L.L., Ji, H.F., et al. 2010, DIGA—a database of improved gene annotation for phytopathogens, *BMC Genomics*, **11**, 54.

30. Gao, F. and Zhang, C.T. 2004, Comparison of various algorithms for recognizing short coding sequences of human genes, *Bioinformatics*, **20**, 673–81.
31. Brazilian National Genome Project Consortium. 2003, The complete genome sequence of **Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability, *Proc. Natl. Acad. Sci. USA*, **100**, 11660–5.
32. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. 2010, Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics*, **11**, 119.
33. Richardson, E.J. and Watson, M. 2013, The automatic annotation of bacterial genomes, *Brief Bioinform*, **14**, 1–12.
34. Salzberg, S.L. 2007, Genome re-annotation: a wiki solution? *Genome Biol.*, **8**, 102.
35. Zhang, C.T. and Chou, K.C. 1994, A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences, *J. Mol. Biol.*, **238**, 1–8.
36. Chen, L.L. and Zhang, C.T. 2003, Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages, *Biochem. Biophys. Res. Commun.*, **306**, 310–7.
37. Guo, F.B. 2007, The distribution patterns of bases of protein-coding genes, non-coding ORFs, and intergenic sequences in *Pseudomonas aeruginosa* PA01 genome and its implications, *J. Biomol. Struct. Dyn.*, **25**, 127–33.
38. Natale, D.A., Galperin, M.Y., Tatusovm, R.L. and Koonin, E.V. 2000, Using the COG database to improve gene recognition in complete genomes, *Genetica*, **108**, 9–17.
39. Yu, J.F., Xiao, K., Jiang, D.K., Guo, J., Wang, J.H. and Sun, X. 2011, An integrative method for identifying the over-annotated protein-coding genes in microbial genomes, *DNA Res.*, **18**, 435–49.
40. Luo, C., Hu, G.Q. and Zhu, H. 2009, Genome reannotation of *Escherichia coli* CFT073 with new insights into virulence, *BMC Genomics*, **10**, 552.
41. Hayes, W.S. and Borodovsky, M. 1998, How to interpret an anonymous bacterial genome: machine learning approach to gene identification, *Genome Res.*, **8**, 1154–71.
42. Chung, G.T., Yoo, J.S., Oh, H.B., Lee, Y.S., Cha, S.H., Kim, S.J. and Yoo, C.K. 2008, Complete genome sequence of *Neiserria gonorrhoeae* NCCP11945, *J. Bacteriol.*, **190**, 6035–6.
43. Perrodou, E., Deshayes, C., Muller, J., et al. 2006, ICDS database: interrupted CoDing sequences in prokaryotic genomes, *Nucleic Acids Res.*, **34**, D338–43.
44. Sharma, V., Firth, A.E., Antonov, I., Fayet, O., Atkins, J.F., Borodovsky, M. and Baranov, P.V. 2011, A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment, *Mol. Biol. Evol.*, **28**, 3195–211.