‘Fail criteria’ as a method of testing sequence quality, and the relationship between quality and features of the sequenced prokaryotic genome

Anna Blewden

BSc Natural Sciences

Supervisor: Prof Laurence Hurst

2018

Word count: 5,843

Table of contents

Acronyms and abbreviations

CDS	Coding sequence
DDBJ	DNA Data Bank of Japan
EBI	European Bioinformatics Institute
EMBL	The European Molecular Biology Laboratory
ENA	European Nucleotide Archive
INSDC	International Nucleotide Sequence Database Collaboration
NCBI	National Centre for Biotechnology Information

Lay summary

Since the creation of GenBank, the first major online sequence database, in 1982, over 208,452,303 DNA sequences have been made available at the touch of a button (NCBI, 2018a). The applications of DNA sequence data are enormous, from developing diagnostic and therapeutic tools for disease, to helping us discover the answers to evolutionary secrets. With this level of significance, there is a need for strict quality control of sequence data – something that is not provided by major online databases. This research project investigates whether there is any correlation between indicators of sequence quality and features of the prokaryotic genome sequence, with the intention that such understanding can help to guide our approach towards maintaining higher quality sequence data.

To represent quality, this project defines 'fail criteria', a set of qualifiers based on common coding-sequence features, which can be used to filter through prokaryotic sequences to remove those of potential poor quality. Among the findings, there was found to be no significant relationship between quality indicators and complement genes, but a significant relationship was observed between quality and pseudogenes, and between quality and partial coding sequences. Furthermore, no significant correlation was found between sequencing year and quality, but a significant negative correlation was observed between genome size and quality. Results suggest that more investigation should be conducted into the relationship of quality indicators with biological and sequencing-related features of the genome. I also believe that quality assessment in the context of genomic sequences is an area that needs more thorough analysis, particularly in terms of defining quality and standardising approaches.

Abstract

It is crucial that genomic sequences in public databases are kept to a high standard of quality, both in terms of their accuracy to the source organism's genome and the appropriate labelling and annotation of genes. In this report, I define five 'fail criteria' as way of measuring quality of sequences based on features of coding DNA sequences (CDS). These can be used to filter samples of genomic data prior to use in research, so that sequences that are suggestive of poor quality can be discarded and do not interfere with results. The aim of this study was to assess whether there is any relationship between 'fails' as quality assessment criteria and features of the genome – both innate biological features, and features relating to the sequencing process. Using the fail criteria, 58% of prokaryotic genomes in the sampled data were classified as having failed, suggestive of low quality. No significant correlation was observed between year of sequencing and percentage of genes failed, but a weak negative correlation was found between genome size and percentage of genes failed ($r=-0.1517$, $p<0.05$). On a gene-level basis, essentially no relationship was observed complement genes and fail ($r=0.003336$, $p<0.05$), whilst a significant, but weak positive correlation was found between pseudogenes and fail ($r=0.1392$, $p<0.05$) and partial CDS and fail ($r=0.04854$, $p<0.05$). This research suggests that the instance of low quality in sequences can possibly be predicted by certain gene and genomic features, and with more investigation, these findings can be used to develop more accurate approaches to testing quality of sequences prior to publishing.

Introduction

The uses of DNA sequencing are extensive: from important biomedical applications such as the development of diagnostic tests and therapeutic tools, to answering fundamental questions in evolutionary biology.

To facilitate this research, published sequence data are openly accessible via online databases. The three major databases are GenBank (US), ENA (UK), and DDBJ (Japan), which collectively form the International Nucleotide Sequence Database Collaboration (INSDC), and are synchronised daily (EMBL-EBI, 2018). Each database provides submission guidelines for sequences, which must be submitted with correct annotation, such as the source organism and identification of protein coding sequences (CDS), tRNA, and rRNA, and their position within the nucleotide sequence.

With sequence data being provided from research institutions across the world, and their significant role in bioinformatical- and computational-based research, there is a necessity for the submitted data to be of high quality.

'Quality' can be considered in terms of accuracy of sequencing, as well as correct formatting and annotation of flat-file text sequences. This is vital for conducting downstream computational analysis, where programming languages such as Python and Tcl are used to 'parse' these files, the success of which relies on the accuracy of sequence formatting. Improper formatting could lead to script breaks, or incorrect data retrieval (e.g. if non-conventional formatting meant that the language would process the sequence in the incorrect reading frame).

Whilst it is relatively straightforward to ascertain whether a genomic sequence file has been formatted appropriately, particularly when referring to INSDC guidelines, it is harder to conclude whether a sequence has been subject to erroneous technique. Accuracy of genomic sequences, in terms of representing the true DNA sequence of the source organism, is imperative for the validity of results obtained by computational analysis. For example, if a sequence contains bases which have been incorrectly called by sequencing technology, or a region of nucleotides which are not found in the source organism but a consequence of artefactual contamination, this may result in incorrect conclusions being drawn, with no knowledge of the discrepancy. Thus, there is need for a method which identifies sequences that are indicative of poor quality sequencing, based on flat-file sequences alone, so that they can be removed before bioinformatical analyses. In this project, five 'fail' criteria have been set which can be used to identify these sequences of potential poor quality.

To define these 'fails', the core features of a coding DNA sequence had to be considered; the absence of which would then be used to flag sequences which are in breach of these core features. The absence of core defining features of a CDS could occur in nature, especially in the case of mutations which result in the

change of a single nucleotide or frameshift (Cristianini and Hahn, 2006), or due to a range of other reasons not entirely synonymous with 'poor quality'. For example, discrepancies in 'typical' CDS features may be due to artefactual contamination, misreading of bases, improper formatting, or conscious experimental design (for example inclusion of 'partial CDS' regions) (NCBI, 2009). It is therefore important to note that the instance of 'fail' only serves as an indication of potential poor quality, and cannot be interpreted a conclusive or direct reflection of sequence quality.

Regardless of the source of sequence discrepancies – mutation-related, technical, or otherwise – using 'fail' criteria provides an approach for precautionary measures, in which large sets of genomic sequences are filtered and those labelled as 'fail' are isolated and removed. This approach can be considered as an effort to remove sequences indicative of poor quality, ensuring that they will not sabotage results of downstream analysis.

This study investigates whether the instance of 'fail' relates to other features of the gene or genome in question. Gene types including pseudogenes, complementary-strand genes, and 'partial coding sequences' (in the context of gene sequencing) are described in the annotations of genomic sequences, making it possible to qualify the gene with these features alongside instances of fail. Features of the genome include size, number of genes, and individual nucleotide identity. Details of the INSDC database submission can also be assessed to determine if there is any relationship between quality (predicted by instances of fail) with year of sequencing, institution responsible for sequencing, and other descriptors.

These features can be used to investigate whether there is any correlation between gene and genomic features and the instance of fail in any given gene or genome. Any relationships identified could be used to frame future approaches to quality control, and increase the fundamental understanding of DNA sequence properties.

Materials and methods

Defining 'fail' criteria

Fail 1 – not multiple of 3

A sequence of nucleotides containing a whole number of codons, starting with the start codon ATG and ending with a stop codon (TAA, TAG or TGA) is defined as a 'coding DNA sequence' (CDS). These describe features of genes which are translated into proteins (Sieber, Platzer and Schuster, 2018).

As tRNA molecules translate codons of 3 nucleotides, it is necessary for coding sequences to be divisible by 3. Although some variations to this assumption exist in nature (Alberts et al., 2002), bioinformatic scripts often interrogate sequence data in frames of 3 nucleotides, hence gene sequences not divisible by 3 will be discounted in common practice and will therefore be labelled as a 'fail'.

Fail 2 – ambiguous nucleotides

Ambiguous nucleotides are used in DNA sequences where the identity of the nucleotide is not known. They are characterised by non-ATCG letters to represent the possible nucleotide identity; most commonly by N ('a**N**y nucleotide'), R (pu**R**ine), Y (p**Y**rimidine), and several others (Cornish-Bowden, 1985).

Despite possible justifications behind their use in sequence data, the use of non-ATCG nucleotides does not represent the true genetic sequence. When sequences undergo computational analysis, particularly for investigations into nucleotide identity, e.g. GC content, it is not appropriate to include these sequences. Thus, the presence of a non-ATCG character within a CDS will qualify the gene as a 'fail'.

Fail 3 – non-stop codon at end

A stop codon is required at the end of any CDS as a signal for termination of the translation process: to trigger the dissociation of tRNA, the ribosome, and other translation factors (Nakamura and Ito, 1998). One of three possible stop codons are used: TAA, TAG, and TGA (Alberts et al., 2002).

Although the absence of a stop codon does occur in nature – particularly in cases of 'nonstop' point mutations, or frameshift mutations where the stop codon may not be in-frame within the CDS – stop codons are a requirement in the translation process, and contribute to the very definition of a CDS (Sieber, Platzer and Schuster, 2018). The occurrence of a non-stop codon at the end of a CDS has hence been selected as the third qualifier for a gene 'fail'.

Fail 4 – internal stop codon

An internal stop codon – that is, a stop codon before the 'true' termination codon – results in the premature termination of translation before the protein has been fully synthesised, thus producing an incomplete and often non-functional protein (Alberts et al., 2002).

Internal stop codons can be found in nature as a result of mutation – particularly point-nonsense mutations, or frameshift mutations. However as the absence of internal stop codons is a requirement for productive protein synthesis, presence of one may be indicative of erroneous technique; hence this feature represents the fourth category of fail.

Fail 5 – non-NTG start codon

A CDS must begin with a start codon for the initiator tRNA to begin synthesis (Alberts et al., 2002). In most cases, this is ATG, however uses of alternative start codons have also been reported in certain species, for example in strains of *Escherichia coli* where ATG makes up 83% of start codon use, followed by GTG, TTG, and CTG, with ATT and ATC at much lower frequencies (Belinky, Rogozin and Koonin, 2017; Panicker, Browning and Markham, 2015).

To use the start codon as an indicator of potential low-quality, the CDS fails if the identity of the first codon is not 'NTG', where N is any nucleotide. Whilst this does not take into account all possible variations of prokaryotic start codons, the reduced specificity accounts for the most commonly-used alternatives of ATG (GTG and TTG).

Selected gene features

The gene types considered in this experiment are complementary-strand genes, pseudogenes, and 'partial coding sequence' genes. Complementary-strand genes, referred to henceforth as 'complement genes' (in the context of prokaryotes), are genes located on the complementary strand of DNA, and are hence read in the reverse direction (Cristianini and Hahn, 2006). Pseudogenes are versions of genes that, due to mutation, have lost some functionality in comparison to its functional cognate gene (Vanin, 1985). They are often characterised by internal stop codons (due to point mutation) and shifted reading frames (due to insertion or deletion frameshift mutations).

Finally, 'partial coding sequences' are defined by INSDC databases as DNA features whose sequence is 'partial' or 'incomplete', and therefore not necessarily aligned with the first nucleotide of the actual CDS

(NCBI, 2011). The location of these partial sequences is indicated by the characters < (for 5' end partial features) and > (for 3' end). Although partial coding sequences are usually an intentional consequence of experimental design, for example when using primers designed from an interface of two CDS regions, the impartial nature of the sequence often means that the indicated sequence position is not at the first nucleotide position, which would thus result in the sequence being processed using the incorrect reading frame.

Data sources

Bacterial genomes were downloaded from https://www.ebi.ac.uk/genomes/bacteria.html. These genome files were then filtered to remove any sequences with fewer than 500,000 nucleotides, and of the remaining sequences, one genome per genus was selected (this being the largest genome available per genus).

This resulted in a sample set of 650 bacterial genomes which were parsed by Tcl.

Parsing

The parser script was written in Tool Command Language (Tcl), using text editors BBedit (v12.0.2) and Notepad++ (v7.5.4), and executed in Terminal on MacOS (v1.5), and pulled out relevant information on each genome and gene to be analysed.

Complement genes and pseudogenes were labelled such if the words qualifying these features were found in the respective gene's annotation, whilst partial CDS was labelled if the sequence position contained the character < or >.

The parser script is available in the Appendix.

Statistical analysis

Statistical analysis of results was performed in the R programming language and executed in RStudio (v3.4.4). The ltm package (v1.1-1) was also downloaded for running the point-biserial correlation test (Rizopoulos, 2018). The R scripts are available in the Appendix.

Pearson's product-moment correlation coefficient was used for measuring correlation between continuous variables, however due to the range of variable types in the datasets, algebraic derivatives of Pearson's were also employed to ensure validity and consistency of the analysis.

The datasets contain a range of variable types: 'true' continuous variables, such as percentage of failed genes within a genome, and other 'quantitative but discrete' data, which will be grouped with continuous variables for the purpose of data analysis, such as genome size in nucleotides and number of genes within a genome. The variable 'number of fail categories' is another example of a discrete numeric variable, and will also be considered in the same variable group for the purpose of statistical testing (as the values are intrinsically quantitative rather than ordinal categorical values) (Parab and Bhalerao, 2010).

For analysing the relationship between two dichotomous (binary) variables, for example the presence of a pseudogene with the instance of fail, the Pearson product-moment correlation test was also used. Pearson's Phi coefficient would also be an appropriate test, as it is designed to measure association between two binary variables, however as the Phi coefficient returned would be identical to the Pearson correlation coefficient for these types of variables, Pearson's correlation was selected for consistency (Guilford, 1936).

For measuring correlation between a continuous/'discrete numeric' variable, such as number of genes within a genome, and a dichotomous (binary) variable, such as instance of fail, the point-biserial correlation coefficient was used. The point-biserial coefficient was selected both for suitability of the variables (continuous vs dichotomous), and due to being mathematically equivalent to Pearson's product-moment correlation formula (Henry, Cohen and Cohen, 1977), thus making it an appropriate comparison to other tests. Furthermore, as the point-biserial correlation test does not return a P-value alongside the correlation coefficient, the results of the test have been justified with the P-value obtained from Pearson's test. This has been indicated in the relevant figures.

Results

Results have been divided into genome-level analysis and gene-level analysis. The genome-level analysis sheds light onto the relationship of genome and sequencing features with sequence quality, or 'fail'. The gene-level analysis is then introduced to provide insight into sequence quality and particular gene features.

Fail was represented both as a continuous variable (number of failed genes as a percentage of total gene count) and as a binary variable (fail vs not fail). Expressing fail as a percentage allows comparison of genomes of different sizes (in number of genes). Fail as a binary variable was used to represent the instance of fail in either a single gene or for the entire genome (where a 'failed genome' is a genome containing at least one 'failed gene'), depending on the context of the analysis.

Genome-level analysis

Overview of genome sample data: 58% of genomes failed, with Fail 5 being the most common fail category encountered

The relative number of failed and non-failed genomes were counted: of the 650 total in the sample set, 378 genomes had failed (58%), and 272 had not (**Figure 1**). The relative proportion of each fail category is shown in **Figure 2**. As some genomes are classified with more than one type of fail, some genomes are represented more than once. The most popular fail observed was fail 5 (first codon not NTG), containing 303 of the 378 total failed genomes, with the other fail categories each containing fewer than 136 genomes. Perhaps surprisingly, no single genome or gene sequence met all 5 fail criteria, with the highest number of categories being 4.

Figure 3 shows the distribution of genomes by their year of sequencing, ranging from 2003 to 2015, represented by a bar chart – the sample contains sequences from each year, however some years contain more genomes than others: for example, only 2 genomes were sequenced in 2003, whilst 95 genomes were sequenced in 2012.

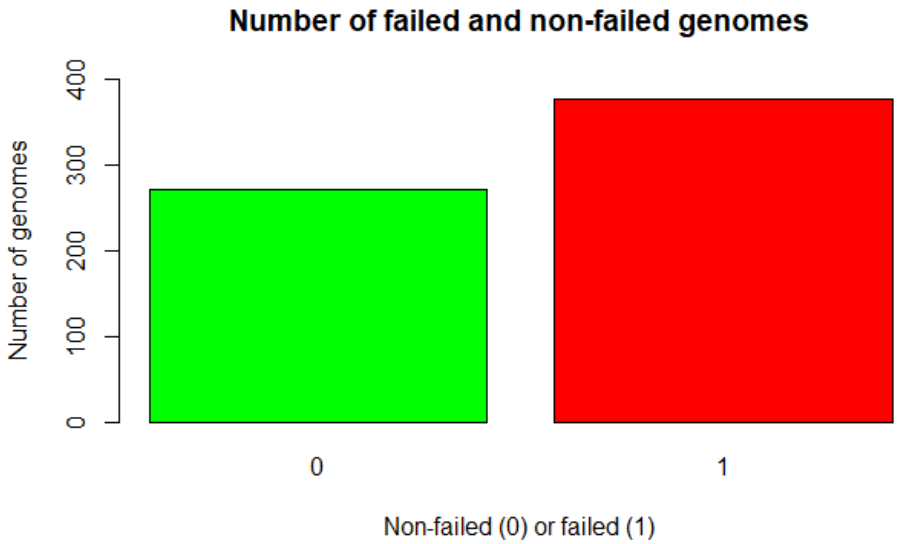


Figure 1. Bar chart showing number of failed genomes (1, red) against number of non-failed genomes (0, green). To be classified as a 'failed' genome, the genome needed to contain at least one gene sequence which had been assigned a fail category.

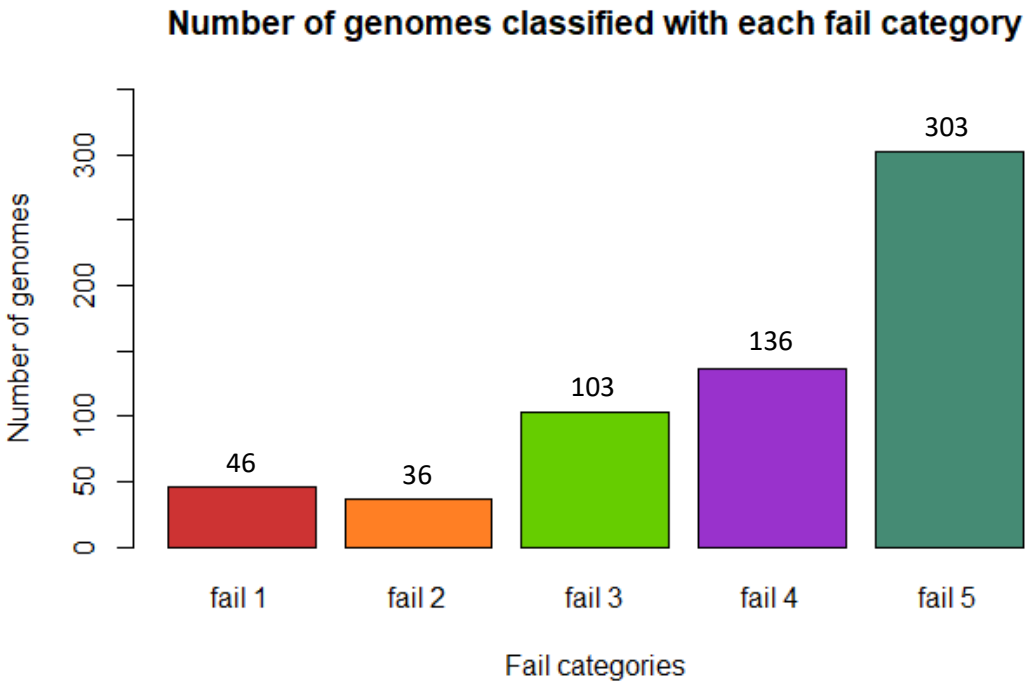


Figure 2. Bar chart showing the number of failed genomes which fall within each fail category – that is, if a genome contains at least one gene with a particular fail category, the genome will fall into that fail category. Occurrences of genomes in each fail category is therefore not mutually exclusive – a single genome can be in multiple fail categories within this bar chart. This chart shows that the 'fail 5' category (non-ATG start codon) is the most frequently occurring fail category, with 303 genomes containing at least one 'fail 5' gene; followed by 'fail 4' category, 'fail 3' category, 'fail 1' category, and finally 'fail 2' category with the fewest number of genomes (36).

Fail categories at the level of genomes was selected rather than at the level of genes, as genes within a single genome are linked in complex ways – and, particularly for the assessment of sequencing 'quality', it may be the case that an instance of fail in one gene due to poor technique is seen in other genes within the same genome, and would therefore not be representative of truly independent genes.

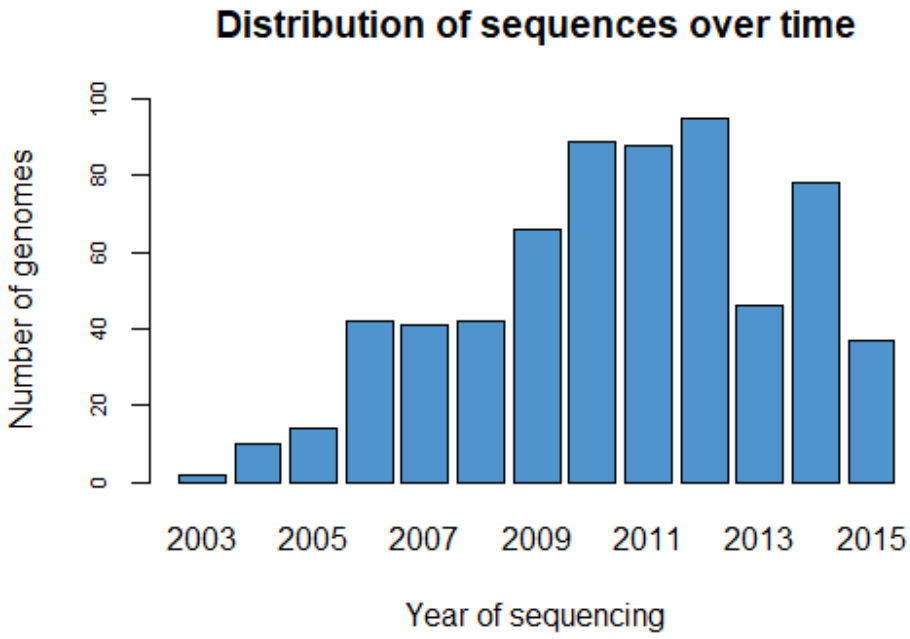


Figure 3. Bar chart showing the sequence year of each of the 650 genomes used for the analysis. The 'oldest' genome in the sample was sequenced in 2003, whilst the 'newest' genome in the sample was sequenced in 2015.

No significant correlation was found between the size of the genome and the year of sequencing

It had been hypothesised that genome size (in nucleotides) would increase over the years, due to the development of new methods and technologies which enabled successful sequencing of larger genomes. However, the p-value was greater than the previously-set 0.05 significance level (Pearson's; r=0.0702, p=0.07368), hence no significant correlation was observed between these two variables (**Figure 4**).

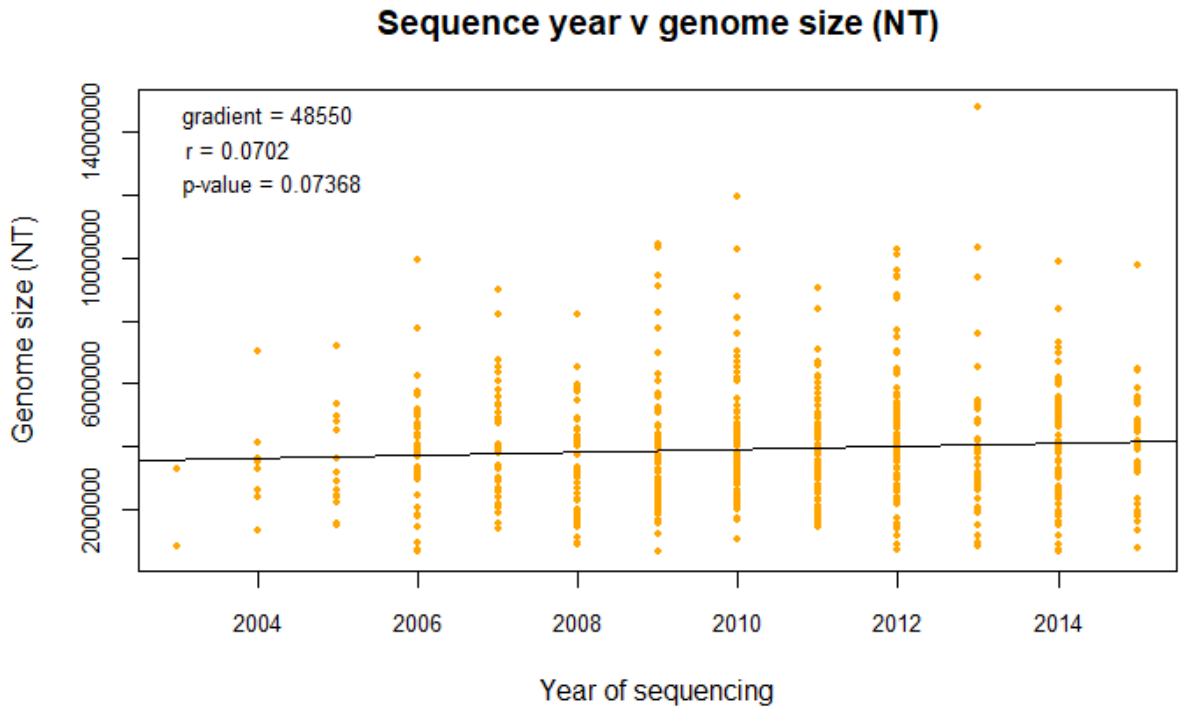
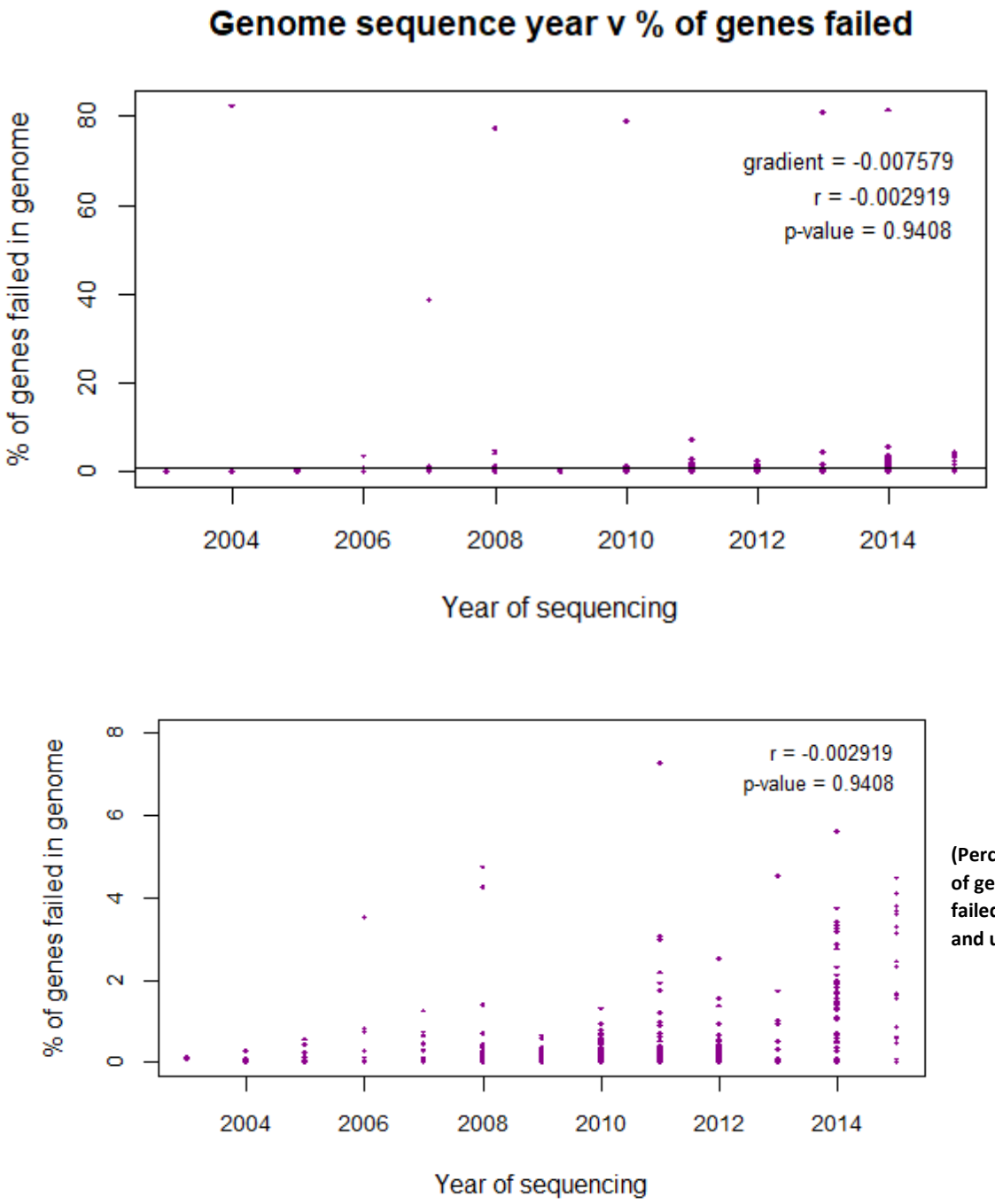


Figure 4. Scatter plot showing the size of genomes (in number of nucleotides) sequenced over the years. No significant correlation was found (Pearson's; p>0.05).

No significant correlation was observed between year of sequencing and percentage of genes failed for that genome

It was hypothesised that there would be a negative correlation between sequencing year and percentage of genes failed, based on the assumption that sequencing technique would improve over time with the development of more efficient sequencing technologies.

However, the results of Pearson's correlation test between these variables were not significant (p>0.05), leading to the rejection of this hypothesis (**Figure 5**). Similarly, there was no significant correlation found between sequencing year and the instance of fail observed as a binary variable (point-biserial; r=0.01833, p=0.6409).



(Percentage of genes failed: 8% and under)

Figure 5. Scatter plots showing the percentage fail of genomes distributed by year of sequencing. There was no significant correlation found (Pearson's; r=0.002919, p>0.05). High p-value (0.9408) likely caused by high proportion of genomes sequenced in later years making it difficult to identify trend over years.

Second graph uses focus of '8% fail and under' to more closely observe distribution of genomes for each year.

Weak negative correlation was observed between genome size and percentage of genes failed

To investigate the relationship between genome size and fail, the correlation between genome size in nucleotides and genome size in number of genes was first tested, with the hypothesis that there would be an extremely strong positive correlation between the two variables. This was observed, with Pearson's correlation test returning r=0.9793 and p<2.2e-16 (see Appendix for graphical representations).

It could therefore be predicted that genome size (in nucleotides) would be related to percentage of genes failed, although there are factors which could lead to positive or negative correlation. In support of positive correlation, as it has been established that number of genes increases with number of nucleotides, it may become more likely that a single gene would be classified as a fail (at least with the assumption that instance of fail is randomly distributed between genes). From the viewpoint of a negative correlation, it could be predicted that researchers only attempt sequencing larger genomes if they have sufficient experience and access to efficient technologies, hence one would observe a decrease of percentage fail as genome size increased (in contrast, this argument is based on the assumption that instance of fail is not randomly distributed, and instead intrinsically linked with sequencing technique).

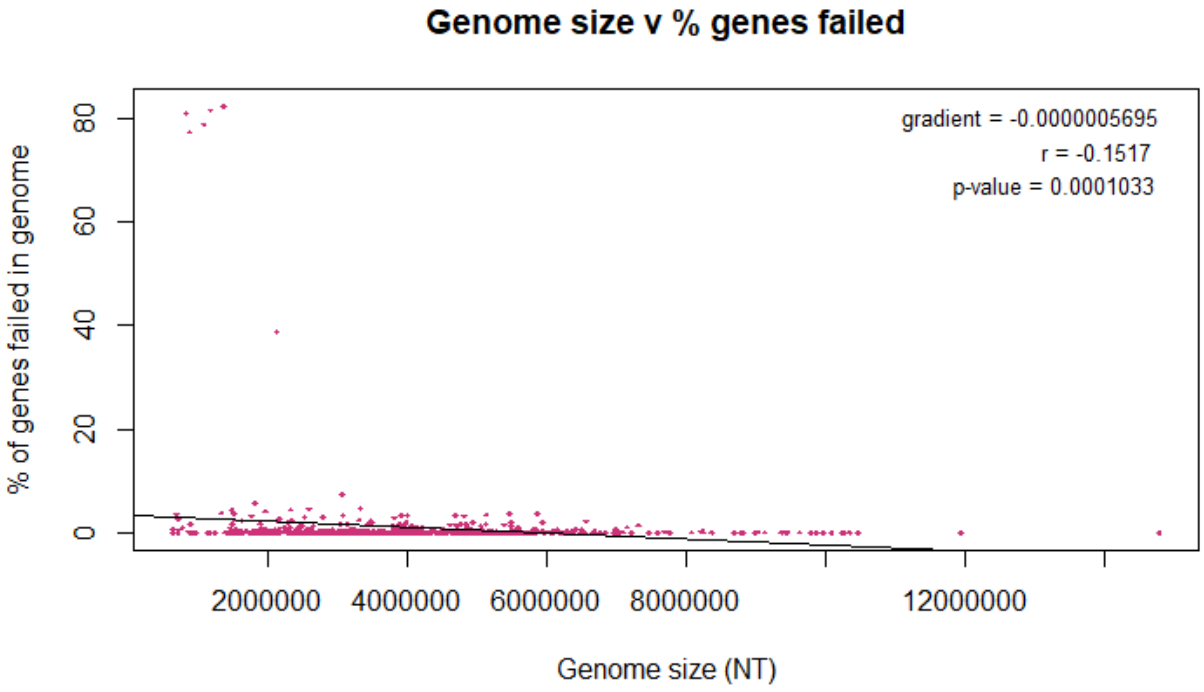
Pearson's correlation test found that there was a weak, but significant, negative correlation between genome size in nucleotides and percentage fail (r=-0.1517, p=0.0001033) (**Figure 6**).

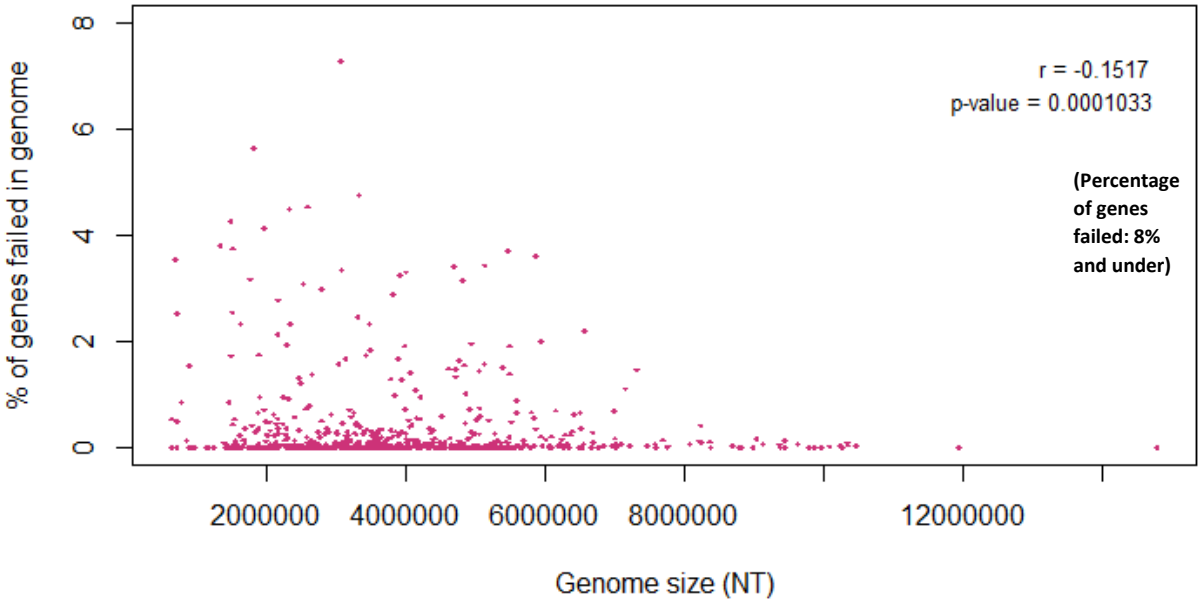
To complement the full-scale graph of genome size and percentage fail, two additional graphs have been provided; one with a focus on '8% fail and under', and one with '0.5% fail and under', in order to help visualise the pattern of the data at the bottom half of the y-axis where the majority of genomes are distributed (although a cluster of outliers are observed at around 80% fail and <2,000,000 nucleotides).

To provide additional depth to this analysis, points representing genomes in the '0.5% fail and under' graph are colour-coded to visualise their sequencing year (blue for 2003-2009 and red for 2010-2015), although the results show little pattern to this distribution.

Despite there being significant correlation found between genome size and percentage fail, there was no significant correlation found between genome size and the instance of fail (point-biserial; r=0.02642, p=0.5013).

Similar findings were also observed when comparing genome size in number of genes with percentage fail, as would be expected due to the strong positive correlation of genome size in nucleotides with number of genes. This graph can also be found in the Appendix.


Genome size v % genes failed
gradient = -0.0000005695
r = -0.1517
p-value = 0.0001033
% of genes failed in genome
0
20
40
60
80
2000000
4000000
6000000
8000000
12000000
Genome size (NT)


r = -0.1517
p-value = 0.0001033
(Percentage of genes failed: 8% and under)
% of genes failed in genome
0
2
4
6
8
2000000
4000000
6000000
8000000
12000000
Genome size (NT)

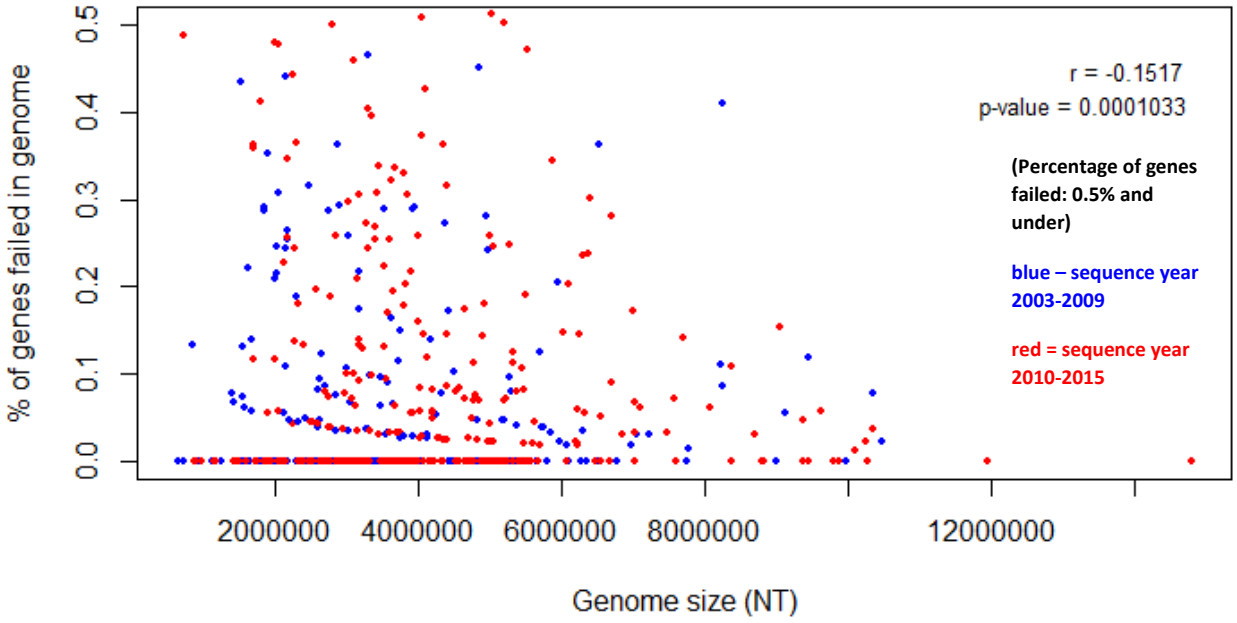


Figure 6. Scatter plots showing genome size (in nucleotides) against the percentage of genes failed for each genome. 'Percentage of genes failed' was calculated by expressing the number of 'failed' genes (genes with at least one fail category) as a percentage of the total number of genes in that genome. Found a statistically-significant weak negative correlation (Pearson's; r=-0.1517, p>0.05).

Second graph uses focus of '8% fail and under' to observe the pattern of distribution.

Third graph uses '0.5% fail and under' to further observe this pattern, whilst using colours to visualise time of sequencing (blue for genomes sequenced between 2003 and 2009, and red for genomes sequenced between 2010 and 2015).

Gene-level analysis

Of the gene types selected for analysis, complement genes were found to be in considerable excess to the other gene types, with 1,142,364 complement out of the total 2,288,663 genes (49.9%); followed by 10,903 failed genes (0.476%), 1068 pseudogenes (0.0467%), and 203 partial CDS (0.00887%) (**Figure 7**). (Note that the instance of a partial CDS is indicated by 'non-specific pos', due to its sequence position being partial or 'non-specific'.)

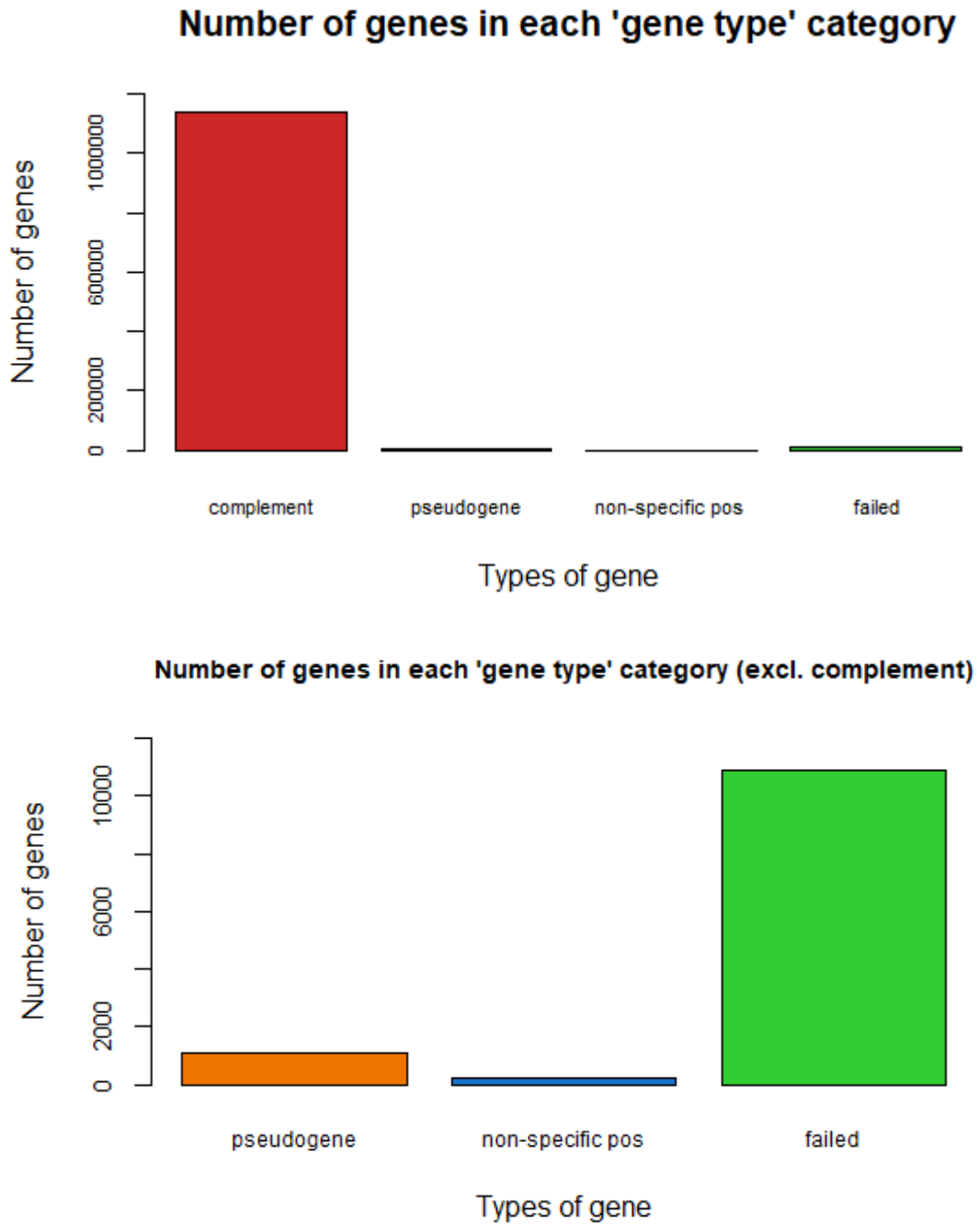


Figure 7. Bar chart showing number of genes within each 'gene type' category. The gene types considered are 'complement', 'pseudogene', 'non-specific position' (that is, a partial nucleotide sequence not positioned at the start nucleotide of the CDS), and 'failed' gene. In the second chart, the 'complement' category was omitted in order to compare the size of the remaining three categories.

No significant relationship was found between complement genes and pseudogenes, or complement genes and partial CDS

No significant correlation was observed between the instance of a complement gene and a pseudogene (Pearson's; r=-0.00008424, p=0.8986). Similarly, there was no significant relationship found between complement genes and partial coding sequences (Pearson's; r=0.000991, p=0.134). It therefore cannot be said that there is a link between instances of complement genes and pseudogenes, or complement and partial sequences.

Contingency tables are provided for reference (**Table 1**, **Table 2**).

Table 1. Contingency table of complement genes and pseudogenes. The instances of complement genes and pseudogenes are represented by binary variables, where '1' refers to the presence of a complement gene / pseudogene, and '0' refers to its absence (i.e. a gene that is not a complement gene / pseudogene).

	Complement gene		
Pseudogene	0	1	Total
0	1145762	1141833	2287595
1	537	531	1068
total	1146299	1142364	2288663

Table 2. Contingency table of complement genes and partial CDS genes. The instances of complement genes and partial CDS are represented by binary variables, where '1' refers to the presence of a complement gene / partial CDS, and '0' refers to its absence (i.e. a gene that is not a complement gene / partial CDS).

	Complement gene		
Partial sequence	0	1	Total
0	1146208	1142252	2288460
1	91	112	203
total	1146299	1142364	2288663

Essentially no relationship was found between partial CDS and pseudogenes

An extremely small, though statistically significant, correlation coefficient was found between the instance of partial CDS and pseudogenes (Pearson's; r=0.001945; p=0.00326). Despite the p-value being below the significance level of 0.05, the value of r is essentially 0, which suggests there is little if any relationship between a gene being both a pseudogene and a partial CDS.

However, due to the sample of genomes containing a very small number of partial sequences (with only 1 gene being positive for both a pseudogene and a partial CDS; see **Table 3**), repeating this test with a larger sample size would help to rule out the absence or presence of any relationship.

Table 3. Contingency table of pseudogenes and partial CDS genes. The instances of pseudogenes and partial CDS are represented by binary variables, where '1' refers to the presence of a pseudogene / partial CDS, and '0' refers to its absence (i.e. a gene that is not a pseudogene / partial CDS).

	Pseudogene		
Partial sequence	0	1	total
0	2287393	1067	2288460
1	202	1	203
Total	2287595	1068	2288663

Essentially no relationship was found between complement genes and instance of fail, or complement genes and number of fail categories

A statistically significant but extremely small correlation coefficient was found between complement genes and failed genes (Pearson's; r=0.003336, p=0.0000004486). As Pearson's r is close to 0, it can be concluded that there is essentially no relationship between complement genes and instances of fail. See **Table 4** for reference.

Table 4. Contingency table of complement genes and instances of fail. The instances of complement genes and fail are represented by binary variables, where '1' refers to the presence of a complement gene / fail, and '0' refers to its absence (i.e. a gene that is not a complement gene / failed gene).

	Instance of fail		
Complement gene	0	1	total
0	1141101	5198	1146299
1	1136659	5705	1142364
total	2277760	10903	2288663

Similarly, there was also little relationship found between instances of a complement gene and number of fail categories encountered by the gene (point-biserial; r=0.002969, p=0.00000707 from Pearson's test).

Essentially no relationship was observed between complement genes and instances of particular fail categories

Pearson's correlation test was also used to determine whether there was any relationship between complement genes and the instances of each particular fail category 1-5; the results of these tests are provided in **Table 5** below.

Of these tests, only two were statistically significant: complement genes with fail 3 (non-stop codon at end), and complement genes with fail 5 (first codon not NTG). However, both of these gave negligible correlation coefficients, hence it can be concluded that in this sample, there is essentially no correlation between these two sets of variables.

Table 5. Summary of correlation values for complement v particular fail categories. Correlation coefficient r and p-value both obtained from Pearson's correlation test.

Fail category	r	p-value
Complement v fail 1	-0.00035	0.5968
Complement v fail 2	-0.00034	0.6098
Complement v fail 3	-0.00475	6.62E-13
Complement v fail 4	0.000823	0.2129
Complement v fail 5	0.006407	3.25E-22

A significant relationship was observed between both pseudogenes and fail, and pseudogenes and number of fail categories encountered

It was hypothesised that, out of the three descriptive gene features in question, pseudogenes would have the strongest relationship with instances of fail due to the intrinsic definition of a pseudogene being a gene characterised by some loss of functionality (Vanin, 1985).

A significant correlation was found between pseudogenes and failed genes (Pearson's; r=0.1392, p<2.2e-16), suggesting that there is a weak relationship between a gene being a pseudogene and being classified as a fail (see **Table 6** below).

Table 6. Contingency table of pseudogenes and instances of fail. The instances of pseudogenes and fail are represented by binary variables, where '1' refers to the presence of a pseudogene / fail, and '0' refers to its absence (i.e. a gene that is not a pseudogene / failed gene).

	Instance of fail		
Pseudogene	0	1	Total
0	2277171	10424	2287595
1	589	479	1068
total	2277760	10903	2288663

A significant relationship was also found between pseudogenes and number of fail categories encountered (point-biserial; r=0.1619, p<2.2e-16 obtained from Pearson's).

This relationship between pseudogenes and number of fail categories can be examined closer by testing the correlation of pseudogenes with particular fail categories – results of which are contained in **Table 7** below. Results were significant (p<2.2e-16) for all except pseudogene v fail 2 (letters other than ATCG;

p>0.05). Of the remaining significant results, the strongest correlation was seen in pseudogene v fail 1 (not multiple of 3; r=0.1739), followed by fail 3 (non-stop codon at end; r=0.1386), then fail 4 (internal stop codon; r=0.09471) and finally fail 5 (first codon not NTG; r=0.0747).

Table 7. Summary of correlation values for pseudogene v particular fail categories. Correlation coefficient r and p-value both obtained from Pearson's correlation test. p-value of '0' corresponds to p<2.2e-16.

Fail category	r	p-value
Pseudogene v fail 1	0.1739	0
Pseudogene v fail 2	0.00096	0.1465
Pseudogene v fail 3	0.1386	0
Pseudogene v fail 4	0.09471	0
Pseudogene v fail 5	0.07474	0

Weak significant correlation was found between partial CDS with instance of fail, and partial CDS with number of fail categories

It was hypothesised that there would be a correlation between partial coding sequences and certain instances of fail – specifically, the fail categories relating to number of nucleotides and codon identity (fails 1, 3, 4, 5).

It was indeed found that there is a weak, but significant, relationship between partial CDS and instance of fail (Pearson's; r=0.04854, p<2.2e-16), with values illustrated in **Table 8** below.

Table 8. Contingency table of partial CDS and instances of fail. The instances of partial CDS and fail are represented by binary variables, where '1' refers to the presence of a partial sequence / fail, and '0' refers to its absence (i.e. a gene that is not a partial CDS / failed gene).

	Fail		
Partial sequence	0	1	Total
0	2277630	10830	2288460
1	130	73	203
Total	2277760	10903	2288663

There was also a significant relationship observed between partial CDS and number of fail categories (point-biserial; r=0.07088, p<2.2e-16 obtained from Pearson's) – the correlation coefficient of which indicates is weak, but still marginally stronger than for the relationship of partial CDS with binary instance of fail.

When considering the relationship between partial CDS and particular fail categories, the correlation of partial sequence with each individual fail category is significant (**Table 9**). The strongest correlation is seen between partial CDS with fail 3 (non-stop codon at end; r=0.09164), followed by fail 5 (first codon not NTG; r=0.06202), with the remaining fail categories having comparatively weaker correlations of under r=0.02 – the strongest being fail 4 (internal stop codon; r=0.01898), then fail 1 (not multiple of 3; r=0.01754), and finally fail 2 (letters other than ATCG; r=0.01472). This trend fits with the hypothesis that instances of fails 1, 3, 4 and 5 in particular would correlate with partial CDS; however these results show that fail 2 also had some correlation, despite being the weakest of the five categories.

Table 9. Summary of correlation values for partial CDS v particular fail categories. Correlation coefficient r and p-value both obtained from Pearson's correlation test. p-value of '0' corresponds to p<2.2e-16.

Fail category	r	p-value
Partial sequence v fail 1	0.01754	4.19E-155
Partial sequence v fail 2	0.01472	6.10E-110
Partial sequence v fail 3	0.09164	0
Partial sequence v fail 4	0.01898	2.33E-181
Partial sequence v fail 5	0.06202	0

Discussion

One cannot make the assumption that 'fail' is synonymous with poor quality: that genes and genomes which are qualified as having 'failed' are necessarily those of lower quality compared to the non-failed sequences.

Limitations to this method of fail characterisation can be illustrated by two examples: the outcome of a false positive, and the outcome of a false negative.

False positive: The instance of labelling a sequence as a 'fail' due to reasons unrelated to the sequencing process, technique, annotation, or otherwise avoidable circumstances; but instead, may have flagged fail criteria due to mutation (e.g. insertion that results in frameshift). This may result in discarding a genome due to risk of poor quality when, actually, the sequence is of 'high quality', in the sense that it is truly representative of its organism.

False negative: When a sequence of low quality – due to poor technique, incorrect base calling, artefactual contamination, or incorrect annotation – is able to 'slip through' the fail tests unnoticed by not flagging any of the specific criteria. This would result in incorrect sequence data polluting the sample of genomes going onto downstream analysis.

It was hypothesised that the instance of a complement gene should not affect fail/sequence quality, particularly as parsing has accounted for its opposite-strand nucleotide sequence. This is consistent with the results seen ($r=0.003336$, $p=0.0000004486$).

However, in the case of pseudogenes, the very definition of the word is a gene characterised by some degree of non-functionality; often caused by a mutation that results in a premature stop codon (Vanin, 1985). One would therefore expect correlation between pseudogenes and fail, especially those which relate to codon identity. This was reflected in the results (significant relationships of $0.074<r<0.17$ for all categories except fail 2).

This correlation suggests that genes may often be mischaracterised as 'fails' (false positive outcome). It may be that the pseudogene sequence is completely accurate in the context of the organism, but is identified as potentially low-quality and discarded regardless. The decision to include pseudogenes in downstream analysis would depend on the intentions of the researcher and their definition of 'quality'. Additionally, removing all pseudogenes from a sample intended to represent only functional genes would

not account for the phenomenon of 'pseudo-pseudogenes', which have been found to code for functional and biologically-important proteins (Prieto-Godino et al., 2016).

Although the analysis of pseudogenes returned some statistically significant results, the relative proportion of pseudogenes to total genes in the sample was unexpectedly low (0.0467%), particularly compared to findings in literature (Lerat and Ochman, 2005). This may be due to modelling errors when identifying the pseudogene qualifier from annotation, or underrepresentation of pseudogenes in sequencing annotation – it has been suggested that, due to the difficulty of identifying transcript structures that represent presence of a pseudogene, certain CDS may not be correctly identified as such (Wright et al., 2016). NCBI Submission Guidelines also advise submitters that are unsure of pseudogene presence to annotate the feature as "non-functional due to frameshift" – therefore there are inconsistencies in pseudogene-related annotation guidelines (NCBI, 2017).

Partial CDS, unlike complement and pseudogenes, are not a defined 'type' of biological gene, but a feature of sequence data. They do not necessarily arise from poor technique, as the decision to produce partial sequences is often part of the design experiment (e.g. by using a primer overlapping two CDS regions) (NCBI, 2017). It is technically possible to 'fix' some partial sequences by identifying overlapping regions of sequence, although this is not required by submitters.

A consequence of including partial sequences is that nucleotide identity-related fail criteria will be met if the partial CDS does not begin at the first nucleotide of the actual CDS. This would especially be the case if the partial CDS position resulted in the script reading the sequence in the incorrect frame. All correlation tests between partial CDS and fail categories returned significant results supporting this hypothesis, but correlation was much weaker than expected – more investigation is required to understand why this may be the case.

Levels of significance and validity in this project could be improved by increasing the sample size and improving the distribution of sequencing dates. This would help determine any trend between genome size or percentage fail with sequence year. It would be important to include sequences pre-dating the early 2000s, as sequencing methods have improved dramatically since their conception in 1970 (Cristianini and Hahn, 2006).

The results indicate that sequence quality could be predicted by certain genomic features, which could be used to develop approaches to maintain higher-quality sequences. Prior to publishing online, it may be useful to analyse genes and genomes displaying certain features (e.g. those containing high numbers of pseudogenes) to test whether quality levels are expected, or if further quality assessment should be conducted.

Other details relating to the sequencing process itself could also be considered, such as characterising the relationship of fail with the research institution responsible. This was initially considered for analysis, however preliminary investigation showed the format for institution address was inconsistent between genomes, with minor variations in text format making it difficult to group institutions together – this in itself could reflect low levels of annotation quality, and hence should be investigated further. A web-crawler could be written to look up the publication of the sequence, and retrieve details such as method of sequencing, processing software, and quality control (if used, or disclosed) to determine any correlation with quality. Another level of analysis to consider is categorising genes and genomes by the number of fail criteria met, and seeing if this behaves similarly to percentage fail in relationship with genomic features.

There are also limitations to the definition of fail criteria. For example, defining fail 5 as the first codon not being NTG is a generalisation: research has revealed that some strains of *E.coli* utilise non-NTG start codons, such as ATT and ATC (Panicker, Browning and Markham, 2015). The criteria could be altered to take into account these identified alternatives, however rare.

The fail 2 category (ambiguous nucleotides) can be assumed to reflect sequence quality, due to the four nucleotides ATCG making up the entirety of the genetic code. The notion that nucleotide identity reflects quality is also used by online databases, for example NBCI who refer to sequences with over 50% ambiguous nucleotides as “low-quality sequences” that are subsequently rejected by automatic processing (NCBI, 2018b). However, this category is the only ‘fail’ criteria taken into account by INSDC submission guidelines, which some computational biologists consider to be too relaxed for ensuring high standards of quality (Pible et al., 2014; Underwood and Green, 2011).

Whilst there are many quality control algorithms available for authors to assess the quality of their sequence, such as Illumina, FastQC, and Phred (Endrullat et al., 2016), these have limited focus on certain aspects such as quality of base-calling and, crucially, these are not absolute requirements of sequences submitted to INSDC databases. Although INSDC guidelines do claim that they will mark any sequences whose accuracy they are unable to verify as “unverified” (Benson et al., 2017), in addition to running basic algorithms to remove sequences of over 50% ambiguous nucleotides (NCBI, 2018b), they state that including any quality score files alongside the submission is “optional” (NCBI, 2018c).

By not enforcing evidence of quality, INSDC are relying on the submitter to perform quality checks on their own sequences, and there is little quality control above the level of annotation standards (Pible et al., 2014). As a result, database end-users have no knowledge of the quality of the sequences they are using.

Criteria for quality assessment have been developed by others who recognise the necessity for improving quality standards of databases; some approaches resulting in the generation of an overall 'genome quality score' (Land et al., 2014) which can be used as "cut-offs" to filter genomes before downstream computational analysis (Wanchai et al., 2017). According to their criteria, Land et al. (2014) found that over 80% of prokaryotic genome sequences available on GenBank were of insufficient quality.

The importance of improving quality standards for online databases is well-recognised within the bioinformatics field, with increasing efforts being made to accurately quantify and qualify sequence standards. Investigating quality and fail criteria in the context of gene and genome features may help us to identify where to look in continuing these efforts, ultimately helping us to secure higher standards of sequencing in order to increase the accuracy and validity of conclusions from bioinformatical approaches.

Acknowledgements

I thank Prof Laurence Hurst and Ruth Richards for invaluable wisdom, support, and guidance throughout this project.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., 2002. *Molecular Biology of the Cell*. 4th ed. New York: Garland Science.

Belinky, F., Rogozin, I. and Koonin, E., 2017. Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions. *Scientific Reports*, 7(1), pp.12422.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W., 2017. GenBank. *Nucleic Acids Research*, 45(D1), pp.D37-D42.

Cornish-Bowden, A., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9), pp.3021-3030.

Cristianini, N. and Hahn, M.W., 2006. *Introduction to Computational Genomics: A Case Studies Approach*. New York: Cambridge University Press.

EMBL-EBI, 2018. *Submitting assembled and annotated sequences* [Online]. Hinxton: EMBL-EBI. Available from: https://www.ebi.ac.uk/ena/submit/sequence-submission [Accessed 10 April 2018].

Endrullat, C., Glökler, J., Franke, P. and Frohme, M., 2016. Standardization and quality management in next-generation sequencing. *Applied & Translational Genomics*, 10, pp.2-9.

Guilford, J.P., 1936. *Psychometric Methods*. New York: McGraw-Hill.

Henry, N.W., Cohen, J. and Cohen, P., 1977. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. *Contemporary Sociology*, 6(3), pp.320.

Land, M.L., Hyatt, D., Jun, S., Kora, G.H., Hauser, L.J., Lukjancenko, O. and Ussery, D.W., 2014. Quality scores for 32,000 genomes. *Standards in Genomic Sciences*, 9, pp.20.

Lerat, E. and Ochman, H., 2005. Recognizing the pseudogenes in bacterial genomes. Nucleic Acids Research, 33(10), pp.3125-3132.

Nakamura, Y. and Ito, K., 1998. How protein reads the stop codon and terminates translation. *Genes to Cells*, 3(5), pp.265-278.

NCBI, 2009. *Submission of Annotation Using a Table* [Online]. Bethesda, Maryland: NCBI. https://www.ncbi.nlm.nih.gov/projects/Sequin/table.html [Accessed 29 March 2018].

NCBI, 2011. *The GenBank Submissions Handbook* [Online]. Bethesda, Maryland: NCBI. Available from: https://www.ncbi.nlm.nih.gov/books/NBK51157/ [Accessed 30 March 2018].

NCBI, 2017. *Prokaryotic Genome Annotation Guide* [Online]. Bethesda, Maryland: NCBI. Available from: https://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/ [Accessed 30 March 2018].

NCBI, 2018a. *GenBank and WGS Statistics* [Online]. Bethesda, Maryland: NCBI. Available from: https://www.ncbi.nlm.nih.gov/genbank/statistics/ [Accessed 11 May 2018].

NCBI, 2018b. *GenBank Submission Portal Wizards* [Online]. Bethesda, Maryland: NCBI. Available from: https://submit.ncbi.nlm.nih.gov/genbank/help/ [Accessed 2 May 2018].

NCBI, 2018c. *WGS Frequently Asked Questions* [Online]. Bethesda, Maryland: NCBI. Available from: https://www.ncbi.nlm.nih.gov/genbank/wgsfaq/ [Accessed 26 April 2018].

Panicker, I.S., Browning, G.F. and Markham, P.F., 2015. The Effect of an Alternate Start Codon on Heterologous Expression of a PhoA Fusion Protein in Mycoplasma gallisepticum. *PLoS ONE* [Online], 10(5). Available from: https://www.ncbi.nlm.nih.gov/pubmed/26010086 [Accessed 1 May 2018].

Parab, S. and Bhalerao, S., 2010. Choosing statistical test. *International Journal of Ayurveda Research*, 1(3), pp.187-191.

Pible, O., Hartmann, E.M., Imbert, G. and Armengaud, J., 2014. The importance of recognizing and reporting sequence database contamination for proteomics. *EuPA Open Proteomics*, 3, pp.246-249.

Prieto-Godino, L.L., Rytz, R., Bargeton, B., Abuin, L., Arguello, J.R., Peraro, M.D. and Benton, R. 2016. Olfactory receptor pseudo-pseudogenes. *Nature*, 539(7627), pp.93-97.

Rizopoulos, R., (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, *Journal of Statistical Software*, 17(5), pp.1-25. Available from: http://www.jstatsoft.org/v17/i05/ [Accessed 21 March 2018].

Sieber, P., Platzer, M. and Schuster, S., 2018. The Definition of Open Reading Frame Revisited. *Trends in Genetics*, 34(3), pp.167-170.

Underwood, A. and Green, J., 2011. Call for a Quality Standard for Sequence-Based Assays in Clinical Microbiology: Necessity for Quality Assessment of Sequences Used in Microbial Identification and Typing. *Journal of Clinical Microbiology*, 49(1), pp.23.

Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. *Annual Review of Genetics*, 191(1), pp.253-272.

Wanchai, V., Patumcharoenpol, P., Nookaew, I. and Ussery, D., 2017. dBBQs: dataBase of Bacterial Quality scores. *BMC Bioinformatics*, 18(14), pp.483.

Wright, J.C., Mudge, J., Weisser, H., Barzine, M.P., Gonzalez, J.M., Brazma, A., Choudhary, J.S. and Harrow, J., 2016. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature Communications*, 7, pp.11778.

Appendices

See USB for Appendices.