Genomics update

Sigma factors in a thousand *E. coli* genomes

Helen Cook and David W. Ussery†
Department of Systems Biology, Center for Biological Sequence Analysis, The Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

Summary

Everyone working with bacterial genomics is familiar with the phrase ‘too much data’. In this Genome Update, we discuss two methods for helping to deal with this explosion of genomic information. First, we introduce the concept of calculating a quality score for each sequenced genome, and second, we describe a method to quickly sort through genomes for a particular set of protein families. We apply these two methods to all of the current *Escherichia coli* genomes available in the The National Center for Biotechnology Information database. Out of the 2074 *E. coli/Shigella* genomes listed (June, 2013), only less than half (983) are of sufficient quality to use in comparative genomic work. Unfortunately, even some of the ‘complete’ *E. coli* genomes are in pieces, and a few ‘draft’ genomes are good quality. Six of the seven known sigma factors in *E. coli* strain K-12 are extremely well conserved; the iron-regulating sigma factor FecI (σ19) is missing in most genomes. Surprisingly, the *E. coli* strain CFT073 genome does not encode a functional RpoD (σ70), which is obviously essential, and this is likely due to poor genome assembly/annotation. We find a possible novel sigma factor present in more than a hundred *E. coli* genomes.

How many sequenced *Escherichia coli* genomes are currently in GenBank? The answer to this seemingly simple question is unfortunately not easy. One can of course just go to the NCBI (The National Center for Biotechnology Information) web pages (http://www.ncbi.nlm.nih.gov/genome/browse/) and have a look. But ‘of course’, one should not look at just *E. coli* but also add *Shigella* genomes [which from a genomics perspective, belong to the same species as *E. coli* (Lukjancenko et al., 2010), but for historical reasons are put in a different genus]. At the time of writing, there are 2074 genomes listed as belonging to these two groups on the NCBI web pages of sequenced bacterial genomes. But closer examination reveals that some of these (about 500) are in ‘progress’, and currently, no data is available, leaving about 1500 genomes that, in principle, have sequence data available for comparison. Of these, only a tiny fraction (68) are listed as ‘complete’, although not all of these really are finished, and some of the ‘draft’ genomes are finished and in one piece. Thus, the answer to this simple question is not so simple.

Why would anyone need a thousand *E. coli* genomes? After all, if you’ve seen one *E. coli* genome, isn’t it pretty much the same as another? Some imagine *E. coli* to be clonal, like widgets from a factory, with all copies of the genome exact replicates. However, the observed diversity within *E. coli* is stunning. Several years ago, we have estimated that there are about 45 000 different *E. coli* gene families in the pan-genome (Snipen et al., 2008). Any individual *E. coli* genome sequence has a core of about 3100 genes, which are found in nearly all *E. coli* genomes, and most of these share sequence identity at the level of about 99% (Kaas et al., 2012). But in addition to this core, for any given genome, there is an additional set of another two thousand or so ‘accessory genes’, which are present only in certain types of *E. coli* genomes, and often, these genes come in clusters or genomic islands.

Core problems

The set of conserved core proteins, conserved across all bacterial genomes, is difficult to determine. The original effort to identify the core set of conserved proteins in bacteria found 256 proteins by comparing two fully sequenced genomes (Mushegian and Koonin, 1996). As more genomes were sequenced, the number of core proteins dropped steadily. A decade later, in an analysis based on 191 genomes, 225 proteins (88% of the original)
were no longer part of the core, and only 31 proteins were left (Ciccarelli et al., 2006). Although many accept this number as being correct, it is based on less than 10% of the more than 2000 currently available genomes, and in our opinion, it is already suspiciously low as it contains only about half of the 57 ribosomal proteins conserved across all bacteria (Korobeinikova et al., 2012). When full-length protein methods are extended to compare a thousand bacterial genomes, they give the obviously misleading result that not a single protein is conserved across all genomes (Lagesen et al., 2010).

Because core functions must be conserved, traditional full-length protein comparison methods might not tell the full story. Instead, we prefer methods that compare proteins based on function. Indeed, we find that more than a hundred functional domains are conserved across all bacteria.

**Sigma factors in *E. coli***

Here, we explore a bit the stunning diversity of *E. coli* by using a functional domain approach (Snipen and Ussery, 2013) to identify sigma factors in the 983 *Escherichia* and *Shigella* genomes with good-enough quality scores for analysis. Further, this method is used to predict novel sigma factors.

Sigma factors are part of the RNA polymerase and are required to initiate transcription (Ishihama, 2012). The sigma factor enables binding of RNA polymerase to DNA, formation of the open complex and the initiation of transcription. Different sigma factors are used to express genes under different conditions such as heat shock and starvation and for specific purposes such as flagella synthesis (Braun et al., 2003; Mahren et al., 2005). The genus *Escherichia* and the closely related genus *Shigella* are known to have seven sigma factors (Paget and Helmann, 2003) listed in Table 1.

The sigma factors shown in Table 1 can be divided into two classes: σ^70^ and σ^54^; the σ^70^ family can be further divided into four broad groups (RpoD, RpoS, RpoH/FliA and RpoE/FecI; see Gruber and Gross, 2003). Genes that are regulated by σ^54^ (RpoN) do not have canonical −10 and −35 regions but instead a −12 and −24 binding site. It is known that FecI, responsible for iron regulation, is not conserved across *E. coli* genomes (Mahren et al., 2005).

**Gathering the genomes and sigma subunits***

Out of the 2074 *E. coli*/*Shigella* genomes listed on the NCBI pages (June, 2013), an initial set of 1558 *Escherichia* and *Shigella* repilcions was obtained from the NCBI genome collection, including complete, draft (also known as Whole Genome Shotgun sequences or ‘WGS’), and sequences from the Sequence Read Archive (SRA). The short sequences from the SRA were assembled using Velvet (Zerbino and Birney, 2008), and for each assembly, the optimal k-mer length was chosen. After removing duplicate and older versions of the genomes, the DNA sequence was extracted. Of these, 57 repilcions yielded no predicted genes with Prodigal (Hyatt et al., 2010); upon examination of these repilcions, it was discovered that 52 of them contained DNA only from small plasmids and not full-length bacterial chromosomal DNA. The remaining five repilcions, while containing data, did not contain DNA sequences, but rather implied links that we could not access at the time. Thus, the draft genome AJWU01000000 and the ‘complete’ genomes with accession CM000662, CM000960, CM001142 and CM001474 were excluded from our analysis, as we could not obtain the DNA sequences from GenBank.

After all of this filtering and throwing out genomes, we were left with a final set of 1475 repilcions that represent 1414 genomes. Plasmids were included with complete genomes (and in the scaffold genomes plasmids would likely be included as well). Many of these genomes were of low quality, and contained large fractions of broken or truncated genes. We assigned a quality score to each genome (as described later) and, for this analysis, decided to only use the 983 genomes with a quality score of greater than 0.2.

**Evaluating size and quality scores of the genomes***

The size of each genome was calculated from the total number of nucleotides of all repilcions. The gene finding program Prodigal (Hyatt et al., 2010) was used on all genomes to consistently find genes, and these genes were run scanned for Pfam domains (Punta et al., 2012) using the PfamA and PfamB HMM (Hidden Markov Model) databases. The number of proteins was calculated as the number of proteins found by Prodigal. A genome quality score was calculated for all genomes based on the number of non-standard bases and the number of pieces (‘contigs’) of each genome sequence. By this measure of genome quality, a fully assembled genome with no non-standard bases will achieve a perfect score of 1.0. A

<table>
<thead>
<tr>
<th>Gene</th>
<th>Aliases</th>
<th>Function</th>
<th>Wt. (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpoD</td>
<td>σ^70^</td>
<td>Household sigma factor</td>
<td>70 ± 0.06</td>
</tr>
<tr>
<td>rpoS</td>
<td>σ^54^, σ^5^</td>
<td>Stationary phase</td>
<td>38 ± 1.2</td>
</tr>
<tr>
<td>rpoH</td>
<td>σ^32^, σ^4^</td>
<td>Heat shock (cytoplasm)</td>
<td>32 ± 0.02</td>
</tr>
<tr>
<td>rpoE</td>
<td>σ^34^, σ^5^</td>
<td>Heat shock (periplasm)</td>
<td>22±</td>
</tr>
<tr>
<td>rpoN</td>
<td>σ^34^, σ^5^</td>
<td>Nitrogen regulation</td>
<td>54 ± 0.2</td>
</tr>
<tr>
<td>tflA</td>
<td>σ^38^, σ^5^</td>
<td>Flagella biosynthesis</td>
<td>28 ± 0.3</td>
</tr>
<tr>
<td>fecI</td>
<td>σ^19^</td>
<td>Ferric citrate transport</td>
<td>19±</td>
</tr>
</tbody>
</table>

a. For details, see text.
penalty of 0.1 is subtracted for every 1000 non-standard bases per megabase. Further, for each genome a penalty of 0.2 is subtracted for every 25 contigs in excess of its chromosomes and plasmids. Scores cannot drop below 0.1. A series of python scripts were used to compile the results into the GenomeAtlas database (Hallin and Ussery, 2004). Genomes with scores of 0.2 and lower were excluded from further analysis. Two hundred sixty-nine SRA and 162 WGS sequences were removed, leaving a total of 983 genomes (see Table 2).

### Table 2. Genome sources, with the number of genomes gathered, and the number of genomes with score > 0.2 that were included in the final analysis.

<table>
<thead>
<tr>
<th>Source</th>
<th>Genomes</th>
<th>Score &gt; 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI complete</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>NCBI draft</td>
<td>960</td>
<td>798</td>
</tr>
<tr>
<td>Assembled SRA</td>
<td>386</td>
<td>117</td>
</tr>
<tr>
<td>Total</td>
<td>1414</td>
<td>983</td>
</tr>
</tbody>
</table>

Finding the sigma factors

A set of domains was constructed that would serve as a query set for predicting novel sigma factors. The seven known sigma factors were identified in the *E. coli* K-12 MG1655 reference genome (GenBank accession number U00096), and the domains present in these proteins were added to the query set. One PfamB domain (PB000208) was found to occur in RpoH and FliA, and this was included in the query set. This set of domains was compared with the list of domains that was returned from a text search for ‘sigma factor’ in the Pfam database. From this search result, only one additional known sigma factor domain (PF07638), the extracytoplasmic function (ECF) sigma factor (Ho and Ellermeier, 2012), was missing so it was added to the query set. The numbers of known and novel sigma factors found are summarized in Table 3. The first row in this table is read as follows: two genomes with a score of more than 0.2 encoded none of the seven known sigma factors, and 331 of the genomes had no novel sigma factors; of the genomes with a quality score of greater than 0.9, none encoded zero of the conserved known sigma factors – that is, all encoded at least one or more, and 32 of them had no novel sigma factors predicted.

All genomes were searched for proteins containing the query set domains, and any protein that contained at least one query set domain was predicted to be a potential sigma factor. The pan and core were calculated by the following method for genomes with quality scores ≥ 0.9. All sigma factor architectures in one genome were compared with all sigma factor architectures in all other genomes. Identical architectures across all genomes were selected as the core, and all other architectures were added to the pan. The figures were drawn using R (R project, 2013).

### Properties of more than a thousand *E. coli* genomes

The genome quality score for the 1414 genomes varied between 0.1 (worst) and 1.0 (best), and is shown in Fig. 1. For all complete genomes listed on the NCBI web pages, the score was 0.9 or 1.0. (green in Fig. 1). Draft genomes from NCBI had quality scores ranging from 0.1 to 1.0. Genomes assembled from the SRA had quality scores of less than 0.5, as shown in Fig. 1.

The 983 genomes with quality scores above 0.2 varied in size between 2.6 and 6.2 million bp (Mbp), with an average of 5.1 ± 0.3 Mbp. Without excluding the poor quality genomes, all 1414 genomes varied in size between 0.029 and 9.5 Mbp, with an average of 5.1 ± 0.9 Mbp. The genomes at both low and high extremes are genomes that were assembled from the SRA and have low genome quality scores.

The number of proteins predicted by Prodigal for each genome is plotted in Fig. 2, and coloured and sized by genome quality score. The 983 genomes with quality scores above 0.2 contain between 2570 and 6337 proteins, with an average of just under 5000 proteins (4945 ± 364). Without excluding poor-quality genomes, all 1414 genomes were found to contain between 103 and 14 102 proteins (as annotated by Prodigal), with an average of about the same – around 5000 proteins per genome (4973 ± 1083). The genomes with extremely low and high numbers of proteins have low genome quality scores. We have found previously that PfamA run on a set of 347 *E. coli* genomes will retrieve 95% of the domains found in the *E. coli* pangenome (Snipen and Ussery, 2013), so we expect that we should have good coverage of this set of genomes.

A total of 4 837 791 genes and 7 900 180 Pfam A and B domains were found in the selected non-poor quality genomes.

### Table 3. Number of genomes that encode known and novel sigma factors for genomes with quality score > 0.2, and for genomes with quality score ≥ 0.9.

<table>
<thead>
<tr>
<th>σ Factors</th>
<th>Score &gt; 0.2</th>
<th>Score ≥ 0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Known</td>
<td>Novel</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>331</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>367</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>191</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>457</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>445</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

© 2013 John Wiley & Sons Ltd and Society for Applied Microbiology, *Environmental Microbiology*
983 genomes. A total of 7,571,819 genes and 11,947,848 Pfam A and B domains were found in all 1414 genomes.

Sigma factor domain analysis

Within the *E. coli* K-12 MG1655 reference genome, a search for the query domains returned only the sigma factors. That is, in the reference genome, no other protein contains a domain that is found in the known sigma factors. However, this is not the case across the 983 genomes, as will be shown later.

The main sigma factor, RpoD, contains six conserved functional domains, as shown in Fig. 3A (Malhotra *et al.*, 1996; Bowers and Dombroski, 1999; Bowers *et al.*, 2000; Campbell *et al.*, 2002). In particular, the region 2 domain is responsible for binding to the −10 region of the promoter, and region 4 binds to the −35 region. In the reference K-12 genome, there is significant overlap between the domains that comprise the sigma factor proteins. The seven sigma factors have an average of 3.4 domains each, but they contain only 11 unique domains between them. The architectures for the sigma-70 factors are shown graphically in Fig. 3A. All of the σ70 proteins share the same conserved region (region 2, shown in blue in Fig. 3A), which is responsible for binding to the TATA box upstream of the transcription start site (this is the −10 DNA sequence which is were the DNA melts; Pribnow, 1975a,b). The only sigma factor that does not share any of its three domains with any of the other sigma factors is rpoN, which is not a member of the σ70 family, and may bind to a different part of RNA polymerase than the σ70 domains (Merrick, 1993).
Sigma factor weights

The sigma factors are named based on their weight in kiloDaltons (kDa), for example, the largest and main sigma factor, RpoD ($\sigma^{70}$), has a weight of $70 \pm 0.06$ kDa. RpoE and fecI have identical domain structures and so cannot be distinguished on the basis of domains themselves, but they can be distinguished based on weight. About a third (458) of 1460 observations of the architecture RpoE ($\sigma^{24}$) or FecI ($\sigma^{19}$) occur at 19 kDa, and 914 observations occur at 22 kDa, together comprising 94% of all observations of this architecture. The most common weight for RpoE is 22 kDa, and this is lower than expected from its designation as $\sigma^{24}$. The average weight for fecI is 19 kDa, as expected. The average weights for each sigma factor are shown in Table 1, and the distributions are shown in Fig. 3B as box plots. The distribution for the known sigma factors is quite narrow (with the exception of RpoE and FecI), and the same narrow pattern is seen for several of the novel domains summarized in Fig. 3.

Known and novel sigma factors

Out of the 4 837 791 proteins found in the 983 genomes, 7344 (0.15%) have hits to least one sigma domain. One out of seven (1095) of these proteins do not match the known sigma factor architectures. There were 48 distinct novel architectures found, and 33 of these occur more than once. These architectures are composed of 26 unique domains, 15 found in PfamA (10 of these are query domains, in our models) and 11 in PfamB (that is, the query domain in RpoH and FliA).

Across the genomes with score $> 0.2$, the number of known sigma factors found encoded in a genome is most commonly 6 and ranges from 0 up to 8. When we look only within the 70 genomes that have a genome quality score of 0.9 and greater, the number of known sigma factors per genome is still 6, but the max drops to 7 and the minimum rises to 4.

Across the genomes with score $> 0.2$, the number of novel sigma factors predicted is most often just one and ranges from 0 up to 8. When the search is limited to genomes with a genome quality score of 0.9 and greater, most genomes have none, and the maximum number of novel sigma factors found is 5. See Table 4 for details.

Genomes with the fewest known sigma factors

*Shigella boydii* CDC 3083-94 (CP001063) and *E. coli* CFT073 (AE014075) each encode only four known sigma factors and also encode novel predicted sigma factors.

Table 4. Number of genomes that is missing each sigma factor.

<table>
<thead>
<tr>
<th>$\sigma$ Factor</th>
<th>Score $&gt; 0.2$</th>
<th>Score $\geq 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RpoD</td>
<td>14$^a$</td>
<td>1</td>
</tr>
<tr>
<td>RpoS</td>
<td>68$^a$</td>
<td>6</td>
</tr>
<tr>
<td>RpoH</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>RpoN</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>FliA</td>
<td>41</td>
<td>9</td>
</tr>
<tr>
<td>RpoE or FecI</td>
<td>6</td>
<td>0$^b$</td>
</tr>
</tbody>
</table>

$^a$. Note in some genomes, RpoD and RpoS are found in fragments.

$^b$. Thirty-eight genomes have one copy, and 32 have two copies.

© 2013 John Wiley & Sons Ltd and Society for Applied Microbiology, *Environmental Microbiology*
Shigella boydii CDC 3083-94 encodes RpoD, RpoH, RpoN and RpoE but does not encode RpoS, FliA and FecI. However, the genome does encode fragments of the RpoS protein; the original four domains (see Fig. 3) are now split into two sets of two, PF00140_PF04542 and PF04539_PF04545. Surprisingly, these novel sigma factors occur 22 and 19 times, respectively, in genomes PF04539_PF04545. The E. coli strain CFT073 genome encodes RpoH, RpoN, FliA and RpoE but does not encode RpoD, RpoS or FecI. However, fragments of the six-domain protein RpoD (Fig. 3) can be found encoded in two adjacent genes in the annotated genome, as PF03979_PFO0140_PF04546 and PF04542_PFO04539_PF04545, and similarly, RpoS is found as fragments (PF04539_PF04545 and PFO0140_PF04545). The CFT073 genome encodes the largest number of novel sigma factors of the genomes with a quality score of 0.9 or greater, and it also encodes the novel sigma factor PB011436_PFO08281, which occurs 64 times in the genome set with a score > 0.2.

Genomes with the most novel sigma factors predicted

Three genomes with a quality score of 0.9 and greater have four novel sigma factors predicted. One is S. boydii CDC 3083-94, already mentioned, and the other two are E. coli SMS-3–5 (CP000970), and E. coli P12b (CP002291). These both encode the novel sigma factor PF04542_PFO4545 (occurs 165 times) and the singleton domain PF04545. The genome of E. coli strain P12b additionally encodes the novel domain PF04539_PBP000208_PFO4545 (occurs 11 times) and another singleton domain PF04542. The genome for E. coli strain SMS-3–5 encodes the novel combination PB000208_PFO08281 (occurs seven times) and a predicted sigma factor containing a Pfam domain not present in the original query set of domains PB011436_PFO08281 (64 times).

Genomes with the fewest sigma factors

Combining both novel and known sigma factors, the minimum found encoded in any genome is five. This occurs for the genomes E. coli K-12 substr. MDS42 (AP012306) and E. blattae DSM 4481 (CP001560), both of which have genomes smaller than the average (3,976,195 and 4,158,725 bp, and 3,621 and 3,825 proteins, respectively). Both genomes are missing FliA and FecI, and encode no novel sigma factors.

Genomes missing known sigma factors

Not a single sigma factor – including RpoD – was conserved across all 1414 genomes. The number of missing known sigma factors per genome is summarized in Table 4. However, as we saw earlier for S. boydii strain CDC 3083-94 and E. coli strain CFT073, fragments of the sigma factors are often found even if all domains for the sigma factor are not found in the same protein.

Two of the six highly scoring genomes that are missing RpoS (E. coli strain 0127:H6 E2348/69 (FM180568) and S. boydii CDC 3083-94 encodes two proteins with half of RpoS (the first half is PF00140_PFO4542; second half is PF04539_PF04545). The first half occurs 22 times in genomes with scores > 0.2, and the second half occurs 19 times. The proteins appear next to each other but in opposite orders in the two genomes. Escherichia albertii TW11588 (draft accession number AEMF00000000) encodes the first three quarters of RpoS (PF00140_PFO04542_PFO4539, 28 times) and the last domain as a singleton (PF04545, 196 times). The genome for E. coli IA11 (CU928160) encodes the last three quarters of RpoS (PF04542_PFO04539_PFO4545, 18 times) with the first domain as a singleton (PF00140, 5 times). Escherichia coli strain IHE3034 (CP001969) also encodes the last three quarters of rpoS but does not encode the first domain as a singleton.

RpoE and FecI

Of genomes with a quality score of 0.9 or greater, the domain structure for RpoE or FecI is found to be encoded once in 38 genomes, and twice in 32 genomes. Of genomes with a quality score greater than 0.2, RpoE or FecI is encoded three times in one genome, E. coli E128010 (ADUO000000 WGS), which encodes two copies of FecI and one copy of RpoE. RpoH is the most commonly over-represented, occurring twice in six genomes.

Novel sigma factor architectures

Many of the novel sigma factors are equivalent to an existing sigma factor but lacking one domain. One hundred sixty-five genomes with score > 0.2 encode the novel sigma factor PF04542_PFO4545, which is RpoH minus its middle PfamB domain. Because this architecture contains the TATA box-binding domain PF04542, which appears to be unique to sigma factors, it seems reasonable to suspect that this might be reflective of a ‘real’ sigma factor not previously characterized. Further, this architecture is found only in a specific subset of serotypes and in similar strains.

Each domain in the query set is found as a singleton (the only domain found in the protein) except for the PfamB domain PB000208 and the ECF domain, which is found in no protein in any of the 983 genomes.

Out of the 59 plasmids, only 10 encoded novel sigma factors, and none encoded known sigma factors. Two
architectures were found: PF00239_PB000208 (occurs 102 times) and the singleton PF08281 (213 times) occurred on plasmids twice and eight times, respectively. Of genomes with a genome quality score of 0.9 or greater, 497 sigma factors were predicted in total, with 98% being found on chromosomes.

The PF04542_PF04545 combination is over-represented in the KTE strains of E. coli (45 of 170 genomes encode this domain) with $P = 0.0013$, in the BCE strains of E. coli (6 of 10 genomes encode this domain) with $P = 2.1 \times 10^{-5}$, and is encoded in all three O111 serotypes and both O11 serotypes.

To examine the distribution of sigma factor families across the E. coli genomes, we construct the pan-core plot for the predicted and known sigma factors over genomes with quality score $\geq 0.9$, as shown in Fig. 4. The pan contains 22 protein families, and the core contains three proteins: RpoN, RpoH, and the architecture representing RpoE or FecI. In Fig. 4, the second copy of the architecture RpoE or FecI is lost first, followed by FliA, then RpoS, and finally RpoD is missing in the last genome.

Genome quality

Many of the genomes assembled from the SRA were not of high quality. They deviated significantly in length and in number of proteins from known good-quality genomes. The genome quality score is a useful measure of overall quality of the genome and enabled filtering the genomes on an objective measure.

Some of the genomes marked as complete in the NCBI database also contain dubious annotations. Notably, E. coli strain ABU 83972 (CP001671) and E. coli 'clone D ii' (CP002211) do encode a full-length functional RpoD protein, but it is annotated as a hypothetical protein in the GenBank file. Escherichia coli strain CFT073 does not encode a full length, functional RpoD as a complete protein, but instead, it encodes one protein that contains the first three domains of RpoD and another that contains the last three domains. Only the first of these proteins is annotated as RpoD in the GenBank file for CFT073 (GenBank accession number AE014075).

Conserved sigma factors

No encoded sigma factor is conserved across all genomes with quality score $> 0.2$, but three sigma factors (RpoN, RpoH, and the architecture representing RpoE or FecI) are conserved within genomes with quality score $\geq 0.9$. These findings raise the number of non-conserved sigma factors in E. coli to four, and reveal a more complex nature of sigma factors in E. coli than has previously been appreciated.

Most of the genomes analysed encode six known sigma factors, and most also encode at least one protein that has been identified as a novel sigma factor. Although a large number of sigma factors were predicted, the same architectures and domains recur. Within all genomes with a score $> 0.2$, only an additional five PfamA domains were found in the proteins that were retrieved by this method.

Because this method is based on the already known $\sigma^{70}$ and $\sigma^{64}$ domains, it is only expected to discover potential new sigma factors in these classes, not completely new classes of sigma factors that may bind to different parts of the RNA polymerase.

It is unlikely that a genome could be missing RpoD and still be viable; thus, it is possible that the fragments of RpoD that were found in CFT073 is due to a genome assembly problem or (less likely) these fragments could dimerize and together perform the action of RpoD. Similar reasoning could account for the genomes that are missing RpoS but encode its domains as fragmented proteins.

A potential new sigma factor

The weights of most of the known sigma factors fall within a narrow range indicating that they represent the same protein. The one exception is the architecture that represents RpoE and FecI, which shows spikes at both 22 and 19 kDa.

The domain PF04545, present in many of the predicted novel sigma factors, may be resulting in false-positives, as can be seen from the very wide range of protein weights that it is found in as a singleton.

However, the architecture PF04542_PF04545 is the most frequently occurring non-known, non-singleton architecture to be found, and it occurs over a narrow range of weights. Because the weights of these observations fall within a narrow range and the proteins are quite similar in sequence, it is likely that they do represent a single protein. PF04542_PF04545 contains only sigma factor domains, and so is likely to be interacting with RNA polymerase. Additionally, it is significantly over-represented in the KTE, BCE, O11 and O111 E. coli strains. This is potentially a new $\sigma^{70}$ sigma factor encoded in a subset of strains, with molecular weight 28 ± 0.6 kDa.

Conclusions

There are literally thousands of E. coli genome sequences available for comparison. However, the first (and important) step on the path to high-throughput, large-scale comparison is to know which genomes to throw away – that is, to separate out the low-quality genomes and build a solid data set to work with. Another important
step is to use protein functional domains to allow fast comparisons of thousands of genomes. We have compared the presence of sigma factor proteins across a thousand *E. coli* genomes, with only a few database queries, taking a few seconds of computational time. Not only were 'problem genomes' easily spotted (such as a complete *E. coli* genome lacking a functional RpoD protein) but also a potential novel sigma factor, which is

Fig. 4. Pan-core plot of sigma factors (predicted) from genomes with quality score of 0.9 or 1.0. The pan is 22 genes, and the core is 3.
present only in certain serotypes and strains, has been identified. The validity of this prediction will of course need to be tested experimentally in the lab, and at this stage, the genes regulated are unknown.

Acknowledgements

The authors would like to thank Tammi Vesth for use of her version of the Pfam database customized for complete bacterial genomes. We would also like to thank Salvatore Cosentino for access to his short read archive assemblies. Finally, we thank Oksana Lukjancenko for generating the data for the pan-core figure. This work was supported in part by grant 09-067103/DSF from the Danish Council for Strategic Research.

References


