Comparative Proteomics

Part 1: A brief history of *E. coli* proteomics

Dave Ussery

DTU course 27105 - Comparative Genomics
Chapter 11 in the textbook

Wednesday, 18 June, 2014
Monday, 16 June:
09:00-11:00 Lectures by Dave - chapters 3 and 7
11:15-12:00 Exercise introduction (Asli and Shinny)
13:00-17:00 Exercise: Genome atlas

Tuesday, 17 June:
09:00-11:00 Lectures by Dave - chapters 8 and 10 from textbook.
11:15-12:00 Exercise introduction (Asli and Shinny)
13:00-17:00 Exercise: BLAST atlas (Asli and Shinny)

Wednesday, 18 June:
09:00-11:00 Lecture by Dave - chapter 11 from textbook.
11:15-12:00 Exercise introduction (Oksana)
13:00-17:00 Exercise: BLAST matrix (Oksana)

Thursday, 19 June:
09:00-11:00 Lectures by Dave - chapters 12 and 14 from textbook.
11:15-12:00 Exercise introduction (Oksana)
13:00-17:00 Exercise: Pan and core genome plots (Oksana)

Friday 20 June:
9:00 - 12:00 Lectures by Dave - Examples of papers using comparative genomics
13:00-17:00 Start work on projects (discussion) (Shinny)
Outline

• Introduction to Proteins
• E. coli diversity
• MetaHit & E. coli in humans
• BLAST matrices
GEN Exclusives

Jun 16, 2014

Medically, Proteomics Advances Will Rival the Genetics Advances of the Last Ten Years

As the number of medically important proteins increases, the impact will be momentous, according to N. Leigh Anderson.

*Mitzi Perdue*
Seventy percent of the decisions made by physicians today are influenced by results of diagnostic tests, according to N. Leigh Anderson, founder of the Plasma Proteome Institute and CEO of SISCAPA Assay Technologies. Imagine the changes that will come about when future diagnostics tests are more accurate, more useful, more economical, and more accessible to healthcare practitioners. For Dr. Anderson, that’s the promise of proteomics, the study of the structure and function of proteins, the principal constituents of the protoplasm of all cells.

In explaining why proteomics is likely to have such a major impact, Dr. Anderson starts with a major difference between the genetic testing common today, and the proteomic testing that is fast coming on the scene. “Most genetic tests are aimed at measuring something that’s constant in a person over his or her entire lifetime. These tests provide information on the probability of something happening, and they can help us understand the basis of various diseases and their potential risks. What’s missing is, a genetic test is not going to tell you what’s happening to you right now.”

He goes on to say that molecular markers, accessible through proteomic testing, do tell what’s happening now. “Defects in proteins underlie most of disease, and that’s because the business end of biology involves proteins. They’re the molecular machines that manage all the chemical reactions in our bodies. It’s proteins that send out signals and provide movement in muscles, and it’s proteins that form a lot of the body’s structures.”

He goes on to say, “If we can learn about how these defects are playing out in individuals, particularly by finding out changes in their numbers and types, there will be a vastly expanded use of proteins in diagnostic tests.”
Lactose

D-galactose

D-glucose
LETTERS TO THE EDITOR

Genetic Mapping of the Regulator and Operator Genes of the Lac Operon

In *Escherichia coli* the genes responsible for the metabolism of lactose are grouped together to form the *lac* operon (Jacob & Monod, 1961), the elements of which are shown in Figure 1. Regulatory control of the structural genes for β-galactosidase (*z*), permease (*y*), and transacetylase (*a*) is effected by the specific interaction between the *i* gene product (repressor) and the operator gene. The repressor has been shown to be a protein of molecular weight 160,000 (W. Gilbert, personal communication) and the nature of the repressor (protein) and operator (DNA) interaction has been studied in some detail (Gilbert & Müller-Hill, 1967).

These five genes are sufficient to allow a complete description of induction and repression of the *lac* system. There are two additional elements which are considered necessary for the complete system: the promoter and operator. Figure 1 shows the positions of these elements.

\[
\text{pro-C} \quad i \quad o \quad z \quad y \quad a
\]

Fig. 1. Genetic map of the regulator and of the *lac* operon. This map is intended to represent order only, and not distance. The mutant *i*$_{99}$, was extensively mapped and was considered to be close to *i*$_{199}$; it has since been shown to be a double mutant (*i* and *i*$_{0}$, J. Beckwith, personal communication).
The Organization of the *LacZYA* Operon

---

**A.**

- **P_{lacI}**
- **P_{lacZYA}**
- **lacO₁**
- **lacO**

DNA

- **mRNA**

- Lacl repressor

- **β-Galactosidase (LacZ)**

- **Lactose permease (LacY)**

- **Thiogalactoside transacetylase (LacA)**

**lacI and lacZYA are separate transcriptional units, each with its own promoter.**
Summary (so far!)

DNA
RNA
Protein

Sequences ➔ Structure ➔ Function
When are two proteins the same??

Query sequence (protein)

Subject sequence (protein)

50% length of query

50% identity of match

scales poorly

0 proteins conserved in a thousand genomes

Kristoffer

Query sequence (protein)

pfam1

pfam2

pfam3

scales linearly

~100 protein families conserved in a thousand genomes
Comparative Genomics

1. Biochemical name
   - beta-galactosidase
2. 3D structure
3. Gene Ontology Terms
   - GO:0000287
   - GO:0003824
   - GO:0004553
   - GO:0004565
   - GO:0005975
   - GO:0005990
   - GO:0008132
   - GO:0009341
   - GO:0016787
   - GO:0016798
   - GO:0030246
   - GO:0031420
   - GO:0046872
4. Pfam domains
   - PF02929
   - PF0703
   - PF02836
   - PF02837
5. EC Numbers
   - EC 3.2.1.23
6. Subsystem involvement
   - Galactosylceramide and Sulfatide metabolism
   - Lactose and Galactose Uptake and Utilization
   - Lactose utilization
   - Lactose and Galactose Uptake and Utilization
   - Lactose utilization
7. Lactose permease
   - GO:0006810
   - GO:0016020
8. Thiogalactoside acetyltransferase
   - GO:0005989
   - GO:0008870
   - GO:0016407
   - GO:0016740
   - GO:0016746
   - GO:0005737
HMM Coverage of 2000 prokaryotic genomes

Number of genes

PfamA  TIGRFAM  Superfamily  Notinhmm

Coverage of 2000 prokaryotic genomes
A Brief History of Escherichia coli Proteomics
Sonicated extract (growth-minimal)

(spots M155-M190)

Master gel of total-protein extract fractionated by 2-DE, electrotransferred to PVDF, and stained with Coomassie-Blue. The protein extract was prepared using the sonication method from W3110 growing in glucose-minimal media. The IEF used a 4:1 ampholyte blend of Serva 5-7:Serva 3-10, and the SDS-PAGE used a 12.5% T, 2.7% C gel. The largest and smallest pI and Mr of the identified 2-DE spots from the gel are shown on the pI and Mr axes. The data for spots M155-M190 are found in Table a2.
### Table 1: The observed N-terminus of 223 *E. coli* genes

<table>
<thead>
<tr>
<th>Locus</th>
<th>Predicted N-terminus&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Observed N-terminus&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>accA</td>
<td>m*slnfldeeqpiael</td>
<td>slnfldeeqpia</td>
</tr>
<tr>
<td>accC</td>
<td>mldkivianrgeia</td>
<td>mldkivianrge</td>
</tr>
<tr>
<td>aceE</td>
<td>m*serfpndvdpietr</td>
<td>serfpndvdpie</td>
</tr>
<tr>
<td>aceF</td>
<td>m*aieikvpdigad</td>
<td>aieikvpdigad</td>
</tr>
<tr>
<td>adk</td>
<td>mriillgapgagkg</td>
<td>mriillgapgag</td>
</tr>
<tr>
<td>agp</td>
<td>mnktliaaavagvllasaqgta*qtveyqylqqvld</td>
<td>qtvpegyqyqyv</td>
</tr>
<tr>
<td>ahpC</td>
<td>m*slintkikpfknqa</td>
<td>slintkikpfkn</td>
</tr>
<tr>
<td>ahpC</td>
<td>(67)y*avstdthfthka(104)</td>
<td>avstdthfthka</td>
</tr>
<tr>
<td>aldA</td>
<td>M*SVPVQHPMYIDGQF</td>
<td>SVPVQ&lt;sup&gt;(KG)&lt;/sup&gt;PMYXD</td>
</tr>
<tr>
<td>araF</td>
<td>mhkftkalaaglaavmsqsama*enklglflvkqpee</td>
<td>enklglflvkqpee</td>
</tr>
<tr>
<td>arcA</td>
<td>mqtphilivedelvt</td>
<td>mqtphilivede</td>
</tr>
</tbody>
</table>
**12. Annotation quality**

*Fig. 3.* Protein length distributions for *Escherichia coli.*

*Fig. 4.* Protein length distributions for *Aeropyrum pernix.* The curves are similar to those in Fig. 3, except that there is no 2D-PAGE data.

---

*Trends in Genetics, 17:425-429, (2002).*
What is a pan-genome?

E. coli pan-genome
~50,000 gene families

Escherichia coli K-12 MG1655

4144 proteins

~3000 E. coli gene families

Lars Snipen
CBS, Department of Systems Biology
2009 *E. coli* genomes, as of 17 June, 2014:

<table>
<thead>
<tr>
<th>Organism/Name</th>
<th>BioSample</th>
<th>BioProject</th>
<th>Group</th>
<th>SubGroup</th>
<th>Assembly</th>
<th>Level</th>
<th>Size (Mb)</th>
<th>GC%</th>
<th>Chromosomes</th>
<th>Plasmids</th>
<th>WGS</th>
<th>Scaffolds</th>
<th>Gene</th>
<th>Protein</th>
<th>Release Date</th>
<th>Modify Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>SAMN02647163</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000597845.1</td>
<td>4.76</td>
<td>51.10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CAUT01</td>
<td>1</td>
<td>4630</td>
<td>4571</td>
<td>2012/06/01</td>
<td>2012/12/04</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666434</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599625.1</td>
<td>4.81</td>
<td>51.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CP007390.1</td>
<td>1</td>
<td>4824</td>
<td>4562</td>
<td>2013/03/24</td>
<td>2014/04/16</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666695</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599645.1</td>
<td>4.88</td>
<td>51.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CP007414.1</td>
<td>1</td>
<td>4915</td>
<td>4607</td>
<td>2013/03/24</td>
<td>2014/04/16</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666696</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599665.1</td>
<td>5.05</td>
<td>50.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CP007392.1</td>
<td>1</td>
<td>4910</td>
<td>4631</td>
<td>2013/03/24</td>
<td>2014/04/16</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666697</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599685.1</td>
<td>5</td>
<td>50.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CP007393.1</td>
<td>1</td>
<td>4846</td>
<td>4580</td>
<td>2013/03/24</td>
<td>2014/04/16</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666698</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599705.1</td>
<td>5.09</td>
<td>50.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CP007394.1</td>
<td>1</td>
<td>4967</td>
<td>4572</td>
<td>2013/03/24</td>
<td>2014/04/16</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666699</td>
<td>PRJNA238952</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599725.1</td>
<td>4.79</td>
<td>50.50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>CAUT01</td>
<td>1</td>
<td>4630</td>
<td>4571</td>
<td>2012/06/01</td>
<td>2012/12/04</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02666809</td>
<td>PRJNA239027</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599755.1</td>
<td>5.29</td>
<td>50.60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFJ01</td>
<td>134</td>
<td>5065</td>
<td>4874</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668081</td>
<td>PRJNA239028</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599745.1</td>
<td>5.06</td>
<td>50.50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFK01</td>
<td>99</td>
<td>4733</td>
<td>4585</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668082</td>
<td>PRJNA239029</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599765.1</td>
<td>4.94</td>
<td>50.50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFL01</td>
<td>168</td>
<td>4621</td>
<td>4478</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668089</td>
<td>PRJNA239030</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599785.1</td>
<td>5.34</td>
<td>50.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFM01</td>
<td>139</td>
<td>5195</td>
<td>5030</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668090</td>
<td>PRJNA239031</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599805.1</td>
<td>4.8</td>
<td>50.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFN01</td>
<td>84</td>
<td>4511</td>
<td>4338</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668091</td>
<td>PRJNA239032</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000599825.1</td>
<td>5.02</td>
<td>50.40</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFO01</td>
<td>102</td>
<td>4948</td>
<td>4778</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668259</td>
<td>PRJNA237732</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_00047795.1</td>
<td>4.61</td>
<td>50.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFP01</td>
<td>102</td>
<td>4948</td>
<td>4778</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02668261</td>
<td>PRJNA237735</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_00047815.1</td>
<td>4.63</td>
<td>50.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JFFQ01</td>
<td>102</td>
<td>4948</td>
<td>4778</td>
<td>2013/04/24</td>
<td>2014/03/24</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02726231</td>
<td>PRJNA244447</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_00081435.1</td>
<td>5.14</td>
<td>50.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JINV01</td>
<td>202</td>
<td>4961</td>
<td>4810</td>
<td>2014/04/05</td>
<td>2014/05/07</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02709452</td>
<td>PRJNA242116</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000690945.1</td>
<td>5.13</td>
<td>50.60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JICF01</td>
<td>146</td>
<td>5017</td>
<td>4848</td>
<td>2014/05/14</td>
<td>2014/05/14</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02709453</td>
<td>PRJNA242116</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000690965.1</td>
<td>5.01</td>
<td>50.50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JICE01</td>
<td>130</td>
<td>4850</td>
<td>4690</td>
<td>2014/05/14</td>
<td>2014/05/14</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>SAMN02709454</td>
<td>PRJNA242116</td>
<td>Proteobacteria</td>
<td>Gammaproteobacteria</td>
<td>GCA_000690985.1</td>
<td>5.44</td>
<td>50.60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>JICG01</td>
<td>194</td>
<td>5248</td>
<td>5609</td>
<td>2014/05/14</td>
<td>2014/05/14</td>
</tr>
</tbody>
</table>

1018 complete

62 complete

543 scaffolds

413 SRA

760 no data

as of 17 June, 2014

In Figure 3 we show the distribution (presence/absence) of domain sequences over the *E. coli* population. The leftmost bar is the number of ORFans, domain sequence families found in one single genome only. We found 909 ORFans when considering all 347 genomes. The rightmost bar are the 479 core domain sequence families found in at least one copy in all 347 genomes. The distribution is very similar in shape to what we usually see for pangenomes of other data sets and where gene families are computed. Hence, also the domain sequence families are distributed across genomes in this way. The most remarkable result is the large number of ORFans. More than 15% of the domain sequence families found in *E. coli* are seen in only 1 out of 347 genomes. The Pfam-A models are curated and have a built-in threshold for assigning significant hits. However, ABC-transporters (PF00005.22). Among the multi-domain proteins the sequence PF00126.22,PF03466.15 is the most common. This is a domain sequence characteristic of HTH-type transcriptional regulators, and it occurs in more than 30 copies in each of the 347 *E. coli* genomes investigated here. Figure 2 shows that more than 3000 of the domain sequence families are defined by a single domain, while gradually fewer families are defined by multi-domain-sequences. Single-domain proteins also clearly outnumber the multi-domain proteins in the genomes, contrary to what is sometimes claimed. The longest domain sequence contains 25 non-overlapping Pfam-A hits in the same protein, mainly multiple copies of PF05662.9 and PF05658.9, both short repeats as associated with haemaglutinins. This protein is found in 4 of the 347 genomes.

Figure 1 Complete and draft genomes. The box and whisker plots illustrate the differences between completed and draft genomes in this study. The left panel shows that the 56 complete genomes are somewhat smaller in size measured in megabases. This is most likely due to unresolved overlaps between the contigs in the draft genomes. The middle panel contains the number of unique genes predicted by the three gene finders after the elimination of all partial predictions (lacking start or stop codon). Notice the large number of predicted genes in virtually all cases, annotated *E. coli* genomes usually have 4500–5500 genes. Among the draft genomes some genomes have very few predicted genes, seen as circles. The rightmost panel shows the number of predicted gene with at least one Pfam-A hit. Except from four draft genomes with extremely few genes, the differences between complete and draft genomes are now ignorable.
However, the notion of the core is difficult, since we require a core domain sequence to be present in every single genome. Due to the uncertainties in gene prediction and computation of sequence clusters, such a crisp limitation of the core is unfortunate. From the binomial mixture model we get a much better and more robust picture by considering the estimated selection probabilities behind every mixture component. In Figure 7 we have visualized the E. coli pangenome as a pie chart, where the colors indicate the selection probabilities. Using the Bayesian Information Criterion (BIC) we found that 12 components was optimal for the current data set. This means E. coli domain sequences can be grouped into 12 distinct sectors with respect to how often they appear in the genomes. The core genes is one of these types, having selection probability 1.0 since they occur in every genome (darkest blue sector). Notice also the large sector of domain sequences with a selection probability of 0.988. Even the third darker blue sector has a selection probability of 0.966.

We repeated the computations using the strict E-value cutoff in the HMMER3 software. The sample pangenome then reduces to 4745 (from 5724) observed domain sequence families, since only very significant (E < 10^-10) Pfam-A hits are now considered. The binomial mixture model gives a pangenome size estimate of 4973, which still means a coverage around 95%. The Heaps law analysis, however, results in a closed population, with a = 1.01 and an estimate of its size at 4876 domain sequences. This illustrates how the choice of cutoffs in the sequence clustering may change a result completely. The Heaps law analysis is extremely sensitive to the number of ORFans in the data set. The sample core size is 479 domain sequences, and the binomial mixture model predict the population core to be 462.

Figure 6 Genomes in functional space. Each dot correspond to a genome plotted in the two first principal component directions of the E. coli functional space defined by the presence/absence of domain sequence families. There are four large subsets of genomes in the data set, and these dots are marked with colors, see figure legend. The first principal component accounts for 11% of the total data variation, and the second component 8%. Only relative positions of the genomes (dots) are important, the absolute scores on each axis lacks interpretation.
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.
Table 1. Sigma factors in *E. coli*, with estimated molecular weights of the proteins averaged over 983 genomes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Aliases</th>
<th>Function</th>
<th>Wt. (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpoD</td>
<td>$\sigma^{70}$</td>
<td>Household sigma factor</td>
<td>$70 \pm 0.06$</td>
</tr>
<tr>
<td>rpoS</td>
<td>$\sigma^{38}$, $\sigma^S$</td>
<td>Stationary phase</td>
<td>$38 \pm 1.2$</td>
</tr>
<tr>
<td>rpoH</td>
<td>$\sigma^{32}$, $\sigma^H$</td>
<td>Heat shock (cytoplasm)</td>
<td>$32 \pm 0.02$</td>
</tr>
<tr>
<td>rpoE</td>
<td>$\sigma^{24}$, $\sigma^E$</td>
<td>Heat shock (periplasm)</td>
<td>$22^a$</td>
</tr>
<tr>
<td>rpoN</td>
<td>$\sigma^{54}$, $\sigma^N$</td>
<td>Nitrogen regulation</td>
<td>$54 \pm 0.2$</td>
</tr>
<tr>
<td>fliA</td>
<td>$\sigma^{28}$, $\sigma^F$</td>
<td>Flagella biosynthesis</td>
<td>$28 \pm 0.3$</td>
</tr>
<tr>
<td>fecn</td>
<td>$\sigma^{19}$</td>
<td>Ferric citrate transport</td>
<td>$19^a$</td>
</tr>
</tbody>
</table>
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 where 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).
4 Genomics update

**Fig. 1.** Genome quality score for all 1414 genomes by source ranging from 0.1 (worst) to 1.0 (best). The vertical line separates those genomes with score ≤ 0.2 from the 983 genomes included in the analysis.
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.

**Fig. 3.** A. Outline of the functional domains of $\sigma_{70}$ proteins. A shaded box represents each Pfam functional domain. Note that region 2 (PF04542, blue box) is conserved across all $\sigma_{70}$ proteins.

B. Distribution of different sigma factor proteins across the 983 E. coli genomes. The first column is the number of genomes that encode a given sigma factor, and the last column represents the distribution of the average mass of the proteins.
Transcription factors conserved in *E. coli* genomes
A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin1*, Ruiqiang Li1,*, Jeroen Raes2,3, Manimozhiyan Arumugam2, Kristoffer Solvsten Burgdorf1, Chaysvanh Manichanh1, Trine Nielsen4, Nicolas Pons5, Florence Levrenez6, Takuji Yamada7, Daniel R. Mende2, Junhua Li1,7, Junming Xu1, Shaochuan Li1, Dongfang Li1,8, Jianjun Cao1, Bo Wang1, Huiqing Liang1, Huisong Zheng1, Yinlong Xie1,9, Julien Tap9, Patricia Lapage1, Marcelo Bertalan9, Jean-Michel Batto9, Torben Hansen1, Denis Le Paslier10, Allan Linneberg11, H. Bjørn Nielsen11, Eric Pelletier1, Pierre Renault4, Thomas Sicheritz-Ponten9, Keith Turner12, Hongmei Zhu1, Chang Yu1, Shengtong Li1, Min Jian1, Yan Zhou1, Yingrui Li1, Xuqing Zhang1, Songyong Li1, Nan Qin1, Huanming Yang1, Jian Wang1, Søren Brunak14, Joel Doré1, Francisco Guarner1, Karsten Kristiansen13, Oluf Pedersen1,14, Julian Parkhill1,15, Jean Weissenbach16, MetaHIT Consortium1, Peer Bork2, S. Dusko Ehrlich1 & Jun Wang1,13

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent bacterium (most frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

It has been estimated that the microbes in our bodies collectively make up to 100 trillion cells, tenfold the number of human cells, and suggested that they encode 100-fold more unique genes than our own genome. The majority of microbes reside in the gut, have a profound influence on human physiology and nutrition, and are crucial for human life. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity. To understand and exploit the impact of the gut microbes on human health and well-being it is necessary to decipher the content, associated with bowel diseases or obesity. Applied to the human gut, such studies have already generated some difference is especially marked among infants. Although this study, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut, and indicates that many bacterial species are shared by different individuals from the United States or Japan. To get a broader overview of the human gut microbial genes we used the Illumina Genome Analyser (GA) technology to carry out deep sequencing of total DNA from faecal samples of 124 European adults. We generated 576.7 Gb of sequence, almost 200 times more than in all previous studies, assembled it into contigs and predicted 3.3 million unique open reading frames (ORFs). This gene catalogue contains virtually all of the prevalent gut microbial genes in our cohort, provides a broad view of the functions important for bacterial life in the gut, and indicates that many bacterial species are shared by different individuals from the United States or Japan.
Too much data.....

"....present shock keeps us suspended in a state of constant disarray, and causes us to prioritize the recent over the relevant and the new instead of the most important...."

The New York Times, 26 Sept 2013
Figure 1

Deep sequencing and *de-novo* gene assembly

Abundance profiles of gene catalogue

Canopy clustering by co-abundance

MGS augmented assembly

Non-redundant gene catalogue

---

**Nature Biotechnology, in the press, June, 2014**

H. Bjørn
Figure 2

A

CAGs

Genes contained

B

CAG size (genes, log)

Essential genes
Phage-like CAGs
Transposases, integrases & recombinases
CRISPR
Restriction endonucleases
DNA methylation
Glycosyltransferases
Dependency-associated

Nature Biotechnology, in the press, June, 2014
Figure 5

Composition of human feces (rough average, based on 16S rRNA studies)
Fecal bacteria are difficult to classify.

Even phyla are uncertain (using BLAST) for most species.

Pathogens?

>400 metagenome samples, 10Tbytes, 9 Million genes

Nature Biotechnology, in the press, June, 2014
Composition of human feces (rough average, based on 16S rRNA studies)
Known taxa in 742 'Meta-genomic species'

based on BLAST of proteins

Pfam of taxa-specific genes

- species: 550
- phyla: 114
- unknown: 99

Oksana

H. Bjørn

CBS, Department of Systems Biology

DTU course 27105  Comparative Genomics
Questions:

● What determines the function of a protein?

● Three genes are encoded on the same operon - do they have the same function?

● What are some advantages of using functional domains?

● What fraction of fecal bacteria contains *E. coli*?