2011 German *Escherichia coli* outbreak: Alignment-free whole-genome phylogeny by feature frequency profiles

M.K. CHEUNG, Lei LI, Wenyan NONG & H.S KWAN

School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

Correspondence: H.S. KWAN: hskwan@eservices.cuhk.edu.hk

Introduction:

Accuracy of SNP-based whole-genome phylogeny reconstruction relies heavily on quality of sequence alignment which is particularly hindered by poorly assembled genomes. Alignment-free methods might provide additional insights. Here, we constructed a whole-genome phylogeny of 9 outbreak isolates against existing *E. coli* genomes using the alignment-free feature frequency profile (FFP) method (Sims *et al.* 2009). In addition, we looked for gene elements that distinguish the outbreak group from the other *E. coli* strains and possibly accounted for the emergence of the outbreak isolates using the distinguishing feature (DF) analysis.

Datasets:

1. Genome sequences of 30 *E. coli* isolates from NCBI, and
2. Genome sequences of 9 outbreak isolates, TY2482 from BGI, LB226692 from Life Technologies, 5 isolates from HPA, and 2 isolates from Göttingen (Sequence files downloadable from the Github crowdsourcing site: 21-06-2011, https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki/Sequence-reads).

Methods:

i) Phylogenetic analysis:

1. Genome sequences were converted into an RY (purine/pyrimidine)-coded form.
2. Overlapping features, *l*-mers of length 24, were counted over each of the whole genomes.
3. Forward and reverse complement features were considered equivalent.
4. Only core features which were present in all 39 isolates were extracted.
5. Features occurring more than 3 times in any of the isolates were removed.
6. Simple cumulative distances were calculated among all feature states in an unordered manner.
7. A neighbor-joining tree was plotted with the resulting distance matrix using MEGA5 (Tamura *et al.* 2011).

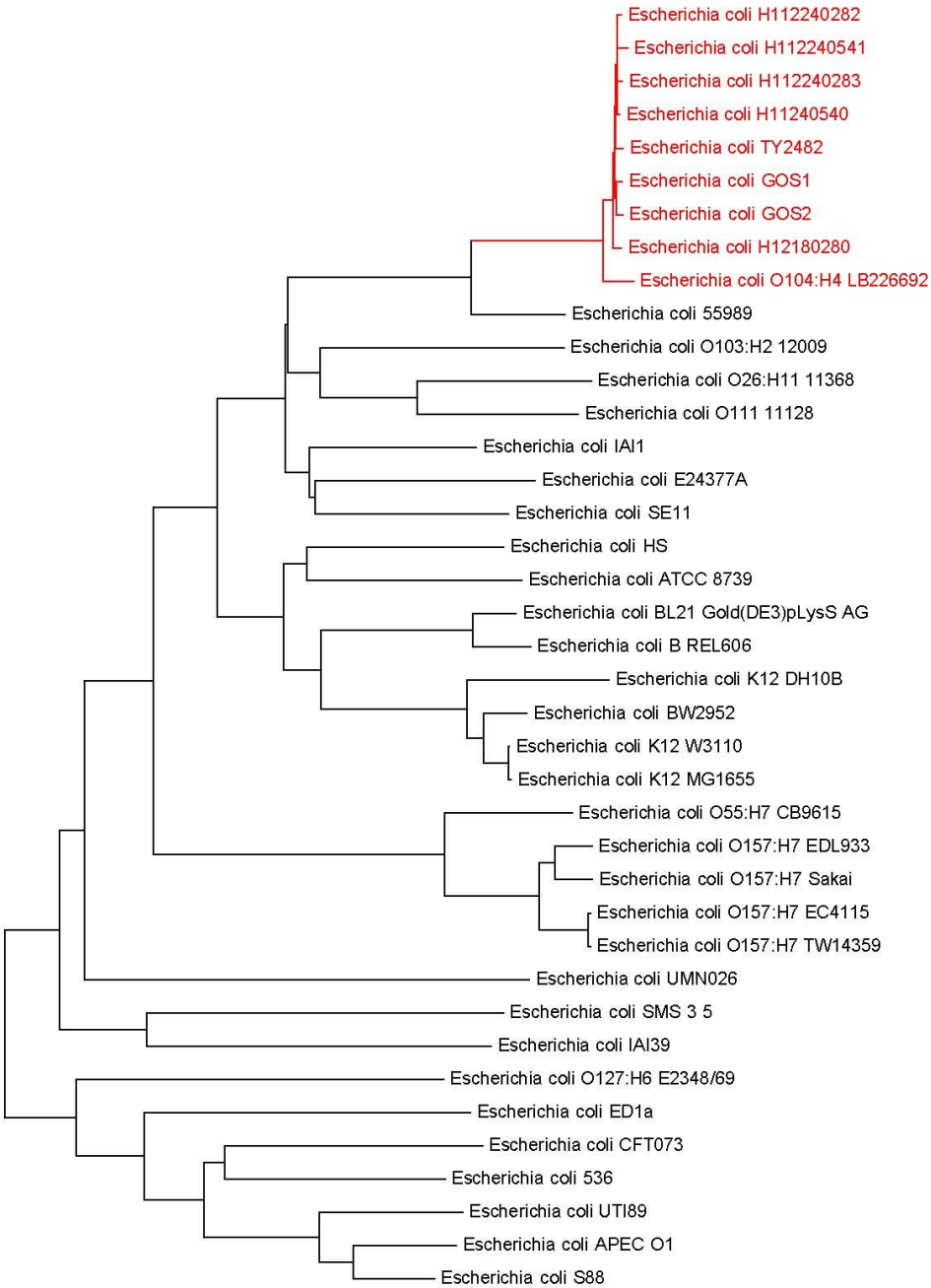
ii) Distinguishing feature analysis:

1. From the unfiltered feature matrix, features that appear in all the 9 outbreak isolates but not in i) 55989 or ii) the 30 NCBI *E. coli* strains were identified as DFs.
2. The DFs were mapped to TY2482.
3. Whenever a DF falls within the boundary of a coding sequence, the respective gene is given a hit. Those with high hits (>50) are referred to as "characteristic genes".

For details, please refer to Sims & Kim (2011).

Results:

Figure 1. Phylogenetic tree produced using the current alignment-free method. Outbreak isolates were highlighted in red.



Our tree generally agrees with that reported by Konrad Paszkiewicz & Kat Holt built using a SNP-based approach (15-06-2011, http://bacpathgenomics.wordpress.com/2011/06/15/snp-base-phylogeny-confirms-similarity-of-e-coli-outbreak-to-eaec-ec55989/), both revealing a high similarity among the outbreak isolates and 55989 being the most closely related isolate sequenced thus far. However, **we should note that the genetic difference between 55989 and the outbreak isolates in our FFP tree is greater than that in the tree built based on SNPs.**

Table 1. Characteristic genes identified in the 9 outbreak isolates against 55989.

Location	Gene	Hits
chromosome	Putative uncharacterized protein	1209
chromosome	DNA transfer protein from phage	867
chromosome	Phage DNA transfer protein	699
chromosome	Antirepressor protein	405
pTY2	Putative uncharacterized protein	280
chromosome	Predicted lipoprotein	272
chromosome	Putative DNA transfer protein	242
chromosome	Phage terminase small subunit	189
chromosome	Putative head DNA stabilization protein	182
pTY2	Uncharacterized protein repA4	164
pTY2	Hypothetical protein IPF_393	138
chromosome	Phage regulatory protein	132
pTY2	Site-specific recombinase	132
pTY2	ORF 153 Hypothetical protein	130
chromosome	Terminase large subunit	125
chromosome	Packaged DNA stabilization protein from phage	82
chromosome	Predicted tail tip fiber protein	54
chromosome	Transposase insH for insertion sequence element IS5R	52
chromosome	Predicted tail fiber protein	52

Table 2. Characteristic genes identified in the 9 outbreak isolates against the 30 NCBI *E. coli* strains.

Location	Gene	Hits
chromosome	Phage DNA transfer protein	166
chromosome	Putative uncharacterized protein	139
pTY2	Putative uncharacterized protein	135
pTY2	Hypothetical protein IPF_393	74
pTY2	Uncharacterized protein repA4	74
chromosome	Predicted lipoprotein	53

Being the most closely-related strain sequenced so far, 55989 has been the target for direct comparison with the outbreak isolate(s) for quite some period of time. The presence of probably more than one prophages in the genome of outbreak isolates but not 55989 has been revealed. From Table 1 above, we could see that most of the characteristic genes belong to or are related to chromosomal phage elements, providing yet another piece of evidence for the presence of prophage(s), of which one is probably shiga toxin secreting, in the outbreak isolates when compared to 55989.

Due to constraints of methods used and/or data availability during the time of analysis, comparisons between all 9 outbreak isolates against all NCBI *E. coli* strains available thus far is limited. Features that are unique to the outbreak group but not in any of the other *E. coli* genomes have been analyzed here and the **results suggested the presence of a totally novel or highly divergent prophage element which is unique to the outbreak isolates** (Table 2). Some more novel or highly divergent yet unknown proteins seem to contribute to the emergence of the outbreak group as well.

Concluding remarks:

Using an alignment-free phylogenetic approach, we further confirm 55989 being the most closely related isolate to the outbreak isolates sequenced thus far, again showing an EAEC-origin of the outbreak isolates. However, the genetic difference suggested that 55989 is not a direct parent of the outbreak isolates. The outbreak group seems to harbor an additional unique prophage element and some unique yet uncharacterized proteins that could not be found in any other *E. coli* genomes sequenced so far.

References:

1. Sims GE, Jun S-R, Wu GA, Kim S-H (2009) **Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions**. PNAS 106: 2677-2682.
2. Sims GE, Kim S-H (2011) **Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs)**. PNAS 108: 8329-8334.
3. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** Mol Biol Evol (In Press).