Development of a web tool for Escherichia coli subtyping based on fimH alleles: Running title: Development of E. coli fimH sub-typing web-tool

The aim of this study was to construct a valid publicly available method for in silico fimH sub-typing of Escherichia coli particularly suitable for differentiation of fine-resolution subgroups within clonal groups defined by standard multi-locus sequence typing (MLST). FimTyper was constructed as a FASTA database containing all currently known fimH alleles. The software source code is publicly available on https://bitbucket.org/genomicepidemiology/fimtyper, the database freely available at https://bitbucket.org/genomicepidemiology/fimtyper_db, and a service implementing the software available at https://cge.cbs.dtu.dk/services/FimTyperFimTyper was validated on three datasets; (i) containing Sanger sequences of fimH alleles of 42 E. coli isolates generated prior to the current study, (ii) whole-genome sequence data of 243 third-generation cephalosporins-resistant E. coli isolates, and (iii) a randomly chosen subset of 40 E. coli isolates from dataset (ii), which were subjected to conventional fimH sub-typing. The combination of the three datasets enabled an evaluation and comparison of FimTyper on both Sanger sequences and WGS data. FimTyper correctly predicted all 40 fimH sub-types from the Sanger sequences from dataset (i), and successfully analyzed all 243 drafted genomes from dataset (ii). FimTyper sub-typing of the Sanger sequences and WGS data from dataset (iii) were in complete agreement. Additionally, fimH sub-typing was evaluated on a phylogenetic network of 122 ST131 E. coli isolates. There were perfect concordance between the typology and fimH-based sub-clones within ST131 with accurate identification of the pandemic multidrug resistant clonal subgroup ST131-H30. FimTyper provides a standardized tool, as a rapid alternative to conventional fimH sub-typing, highly suitable for surveillance and outbreak detection.
MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads

An increasing amount of species and gene identification studies rely on the use of next generation sequence analysis of either single isolate or metagenomics samples. Several methods are available to perform taxonomic annotations and a previous metagenomics benchmark study has shown that a vast number of false positive species annotations are a problem unless thresholds or post-processing are applied to differentiate between correct and false annotations. MGmapper is a package to process raw next generation sequence data and perform reference based sequence assignment, followed by a post-processing analysis to produce reliable taxonomy annotation at species and strain level resolution. An in-vitro bacterial mock community sample comprised of 8 genera, 11 species and 12 strains was previously used to benchmark metagenomics classification methods. After applying a post-processing filter, we obtained 100% correct taxonomy assignments at species and genus level. A sensitivity and precision at 75% was obtained for strain level annotations. A comparison between MGmapper and Kraken at species level, shows MGmapper assigns taxonomy at species level using 84.8% of the sequence reads, compared to 70.5% for Kraken and both methods identified all species with no false positives. Extensive read count statistics are provided in plain text and excel sheets for both rejected and accepted taxonomy annotations. The use of custom databases is possible for the command-line version of MGmapper, and the complete pipeline is freely available as a bitbucked package (https://bitbucket.org/genomicepidemiology/mgmapper). A web-version (https://cge.cbs.dtu.dk/services/MGmapper) provides the basic functionality for analysis of small fastq datasets.
RUCS: Rapid identification of PCR primers for unique core sequences
Designing PCR primers to target a specific selection of whole genome sequenced strains can be a long, arduous, and sometimes impractical task. Such tasks would benefit greatly from an automated tool to both identify unique targets, and to validate the vast number of potential primer pairs for the targets in silico. Here we present RUCS, a program that will find PCR primer pairs and probes for the unique core sequences of a positive genome dataset complement to a negative genome dataset. The resulting primer pairs and probes are in addition to simple selection also validated through a complex in silico PCR simulation. We compared our method, which identifies the unique core sequences, against an existing tool called ssGeneFinder, and found that our method was 6.5-20 times more sensitive. We used RUCS to design primer pairs that would target a set of genomes known to contain the mcr-1 colistin resistance gene. Three of the predicted pairs were chosen for experimental validation using PCR and gel electrophoresis. All three pairs successfully produced an amplicon with the target length for the samples containing mcr-1 and no amplification products were produced for the negative samples. The novel methods presented in this manuscript can reduce the time needed to identify target sequences, and provide a quick virtual PCR validation to eliminate time wasted on ambiguously binding primers. Source code is freely available on https://bitbucket.org/genomicepidemiology/rucs. Web service is freely available on https://cge.cbs.dtu.dk/services/RUCS. mcft@cbs.dtu.dk. Supplementary data is available at Bioinformatics online.
As whole genome sequence data of microorganisms are becoming easily accessible and cheap to produce, a transformation of the traditional methods used for typing, phenotyping and phylogenetic analysis of microorganisms is on the way. Following the anticipation that most clinical microbiological and food safety laboratories will soon have a sequencer in use on a daily basis, there is a growing need for easy-to-use bioinformatics methods that can quickly convert the sequence data into useful information on, e.g., the type of bacteria, whether it is resistant towards any types of
antibiotics, and whether it is part of an outbreak. The Center for Genomic Epidemiology, which is located at the Technical University of Denmark, has since its beginning in 2010 developed such bioinformatics methods and made them freely available as web-services. These web-services and their use is the focus of this chapter.

**WGS-based surveillance of third-generation cephalosporin-resistant Escherichia coli from bloodstream infections in Denmark**

To evaluate a genome-based surveillance of all Danish third-generation cephalosporin-resistant Escherichia coli (3GC–R Ec) from bloodstream infections between 2014 and 2015, focusing on horizontally transferable resistance mechanisms. A collection of 552 3GC–R Ec isolates were whole-genome sequenced and characterized by using the batch uploader from the Center for Genomic Epidemiology (CGE) and automatically analysed using the CGE tools according to resistance profile, MLST, serotype and fimH subtype. Additionally, the phylogenetic relationship of the isolates was analysed by SNP analysis. The majority of the 552 isolates were ESBL producers (89%), with bla CTX-M-15 being the most prevalent (50%) gene, followed by bla CTX-M-14 (14%), bla CTX-M-27 (11%) and bla CTX-M-101 (5%). ST131 was detected in 50% of the E. coli isolates, with the remaining isolates belonging to 73 other STs, including globally disseminated STs (e.g. ST10, ST38, ST58, ST69 and ST410). Five of the bloodstream isolates were carbapenemase producers, carrying bla OXA-181 (3) and bla OXA-48 (2). Phylogenetic analysis revealed 15 possible national outbreaks during the 2-year period, one caused by a novel ST131/ bla CTX-M-101 clone, here observed for the first time in Denmark. Additionally, the analysis revealed three individual cases with possible persistence of closely related clones collected more than 13 months apart. Continuous WGS-based national surveillance of 3GC–R Ec, in combination with more detailed epidemiological information, can improve the ability to follow the population dynamics of 3GC–R Ec, thus allowing for the detection of potential outbreaks and the effects of changing treatment regimens in the future.
Ratings:
BFI (2018): BFI-level 1
BFI (2017): BFI-level 1
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 4.21 SJR 2.24 SNIP 1.527
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.203 SNIP 1.513 CiteScore 4.06
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 2.303 SNIP 1.772 CiteScore 4.61
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 2.416 SNIP 1.782 CiteScore 4.7
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 2.157 SNIP 1.654 CiteScore 4.35
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 2.235 SNIP 1.745 CiteScore 4.24
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 2.09 SNIP 1.642
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 1.873 SNIP 1.616
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 2.005 SNIP 1.522
Web of Science (2008): Indexed yes
Scopus rating (2007): SJR 1.73 SNIP 1.515
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 1.756 SNIP 1.438
Web of Science (2006): Indexed yes
Scopus rating (2005): SJR 1.744 SNIP 1.506
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 1.416 SNIP 1.465
Web of Science (2004): Indexed yes
Scopus rating (2003): SJR 1.364 SNIP 1.349
Web of Science (2003): Indexed yes
Scopus rating (2002): SJR 1.381 SNIP 1.286
Scopus rating (2001): SJR 1.4 SNIP 1.237
Web of Science (2001): Indexed yes
Scopus rating (2000): SJR 1.092 SNIP 1.262
Web of Science (2000): Indexed yes
Scopus rating (1999): SJR 1.092 SNIP 1.389
Original language: English
DOIs:
10.1093/jac/dkx092
A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance

Recent advances in whole genome sequencing have made the technology available for routine use in microbiological laboratories. However, a major obstacle for using this technology is the availability of simple and automatic bioinformatics tools. Based on previously published and already available web-based tools we developed a single pipeline for batch uploading of whole genome sequencing data from multiple bacterial isolates. The pipeline will automatically identify the bacterial species and, if applicable, assemble the genome, identify the multilocus sequence type, plasmids, virulence genes and antimicrobial resistance genes. A short printable report for each sample will be provided and an Excel spreadsheet containing all the metadata and a summary of the results for all submitted samples can be downloaded. The pipeline was benchmarked using datasets previously used to test the individual services. The reported results enable a rapid overview of the major results, and comparing that to the previously found results showed that the platform is reliable and able to correctly predict the species and find most of the expected genes automatically. In conclusion, a combined bioinformatics platform was developed and made publicly available, providing easy-to-use automated analysis of bacterial whole genome sequencing data. The platform may be of immediate relevance as a guide for investigators using whole genome sequencing for clinical diagnostics and surveillance. The platform is freely available at: https://cge.cbs.dtu.dk/services/CGEpipeline-1.1 and it is the intention that it will continue to be expanded with new features as these become available.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, National Food Institute, Research Group for Genomic Epidemiology
Authors: Thomsen, M. C. F. (Intern), Ahrenfeldt, J. (Intern), Bellod Cisneros, J. L. (Intern), Jurtz, V. I. (Intern), Larsen, M. V. (Intern), Hasman, H. (Intern), Aarestrup, F. M. (Intern), Lund, O. (Intern)
Number of pages: 14
Publication date: 2016
Main Research Area: Technical/natural sciences

Publication information
Journal: P L o S One
Volume: 11
Issue number: 6
Article number: e0157718
ISSN (Print): 1932-6203
Ratings:
BFI (2018): BFI-level 1
BFI (2017): BFI-level 1
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 3.11 SJR 1.201 SNIP 1.092
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 1.414 SNIP 1.131 CiteScore 3.32
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 1.545 SNIP 1.141 CiteScore 3.54
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 1.74 SNIP 1.147 CiteScore 3.94
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 1.945 SNIP 1.142 CiteScore 4.15
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
MHCcluster, a method for functional clustering of MHC molecules

The identification of peptides binding to major histocompatibility complexes (MHC) is a critical step in the understanding of T cell immune responses. The human MHC genomic region (HLA) is extremely polymorphic comprising several thousand alleles, many encoding a distinct molecule. The potentially unique specificities remain experimentally uncharacterized for the vast majority of HLA molecules. Likewise, for nonhuman species, only a minor fraction of the known MHC molecules have been characterized. Here, we describe a tool, MHCcluster, to functionally cluster MHC molecules based on their predicted binding specificity. The method has a flexible web interface that allows the user to include any MHC of interest in the analysis. The output consists of a static heat map and graphical tree-based visualizations of the functional relationship between MHC variants and a dynamic TreeViewer interface where both the functional relationship and the individual binding specificities of MHC molecules are visualized. We demonstrate that conventional sequence-based clustering will fail to identify the functional relationship between molecules, when applied to MHC system, and only through the use of
the predicted binding specificity can a correct clustering be found. Clustering of prevalent HLA-A and HLA-B alleles using MHCcluster confirms the presence of 12 major specificity groups (supertypes) some however with highly divergent specificities. Importantly, some HLA molecules are shown not to fit any supertype classification. Also, we use MHCcluster to show that chimpanzee MHC class I molecules have a reduced functional diversity compared to that of HLA class I molecules.
Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion

Seq2Logo is a web-based sequence logo generator. Sequence logos are a graphical representation of the information content stored in a multiple sequence alignment (MSA) and provide a compact and highly intuitive representation of the position-specific amino acid composition of binding motifs, active sites, etc. in biological sequences. Accurate generation of sequence logos is often compromised by sequence redundancy and low number of observations. Moreover, most methods available for sequence logo generation focus on displaying the position-specific enrichment of amino acids, discarding the equally valuable information related to amino acid depletion. Seq2logo aims at resolving these issues allowing the user to include sequence weighting to correct for data redundancy, pseudo counts to correct for low number of observations and different logotype representations each capturing different aspects related to amino acid enrichment and depletion. Besides allowing input in the format of peptides and MSA, Seq2Logo accepts input as Blast sequence profiles, providing easy access for non-expert end-users to characterize and identify functionally conserved/variable amino acids in any given protein of interest. The output from the server is a sequence logo and a PSSM. Seq2Logo is available at http://www.cbs.dtu.dk/biotools/Seq2Logo (14 May 2012, date last accessed).

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis
Authors: Thomsen, M. C. F. (Intern), Nielsen, M. (Intern)
Pages: W281-W287
Publication date: 2012
Main Research Area: Technical/natural sciences

Publication information
Journal: Nucleic Acids Research
Volume: 40
Issue number: W1
ISSN (Print): 0305-1048
Ratings:
BFI (2018): BFI-level 2
BFI (2017): BFI-level 2
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 2
Scopus rating (2016): CiteScore 9.28 SJR 7.397 SNIP 2.657
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 2
Scopus rating (2015): SJR 7.239 SNIP 2.639 CiteScore 9.48
Web of Science (2015): Indexed yes
snpTree - a web-server to identify and construct SNP trees from whole genome sequence data.

Background
The advances and decreasing economical cost of whole genome sequencing (WGS), will soon make this technology available for routine infectious disease epidemiology. In epidemiological studies, outbreak isolates have very little diversity and require extensive genomic analysis to differentiate and classify isolates. One of the successfully and broadly used methods is analysis of single nucleotide polymorphisms (SNPs). Currently, there are different tools and methods to identify SNPs including various options and cut-off values. Furthermore, all current methods require bioinformatic skills. Thus, we lack a standard and simple automatic tool to determine SNPs and construct phylogenetic tree from WGS data.

Results
Here we introduce snpTree, a server for online-automatic SNPs analysis. This tool is composed of different SNPs analysis
suites, perl and python scripts. snpTree can identify SNPs and construct phylogenetic trees from WGS as well as from assembled genomes or contigs. WGS data in fastq format are aligned to reference genomes by BWA while contigs in fasta format are processed by Nucmer. SNPs are concatenated based on position on reference genome and a tree is constructed from concatenated SNPs using FastTree and a perl script. The online server was implemented by HTML, Java and python script.

The server was evaluated using four published bacterial WGS data sets (V. cholerae, S. aureus CC398, S. Typhimurium and M. tuberculosis). The evaluation results for the first three cases was consistent and concordant for both raw reads and assembled genomes. In the latter case the original publication involved extensive filtering of SNPs, which could not be repeated using snpTree.

Conclusions
The snpTree server is an easy to use option for rapid standardised and automatic SNP analysis in epidemiological studies also for users with limited bioinformatic experience. The web server is freely accessible at http://www.cbs.dtu.dk/services/snpTree-1.0/.

General information
State: Published
Organisations: National Food Institute, Division of Epidemiology and Microbial Genomics, Center for Systems Microbiology, Department of Systems Biology, Center for Biological Sequence Analysis, Division of Microbiology and Risk Assessment
Authors: Leekitcharoenphon, P. (Intern), Kaas, R. S. (Intern), Thomsen, M. C. F. (Intern), Rundsten, C. F. (Intern), Rasmussen, S. (Intern), Aarestrup, F. M. (Intern)
Number of pages: 8
Pages: S6
Publication date: 2012
Main Research Area: Technical/natural sciences

Publication information
Journal: B M C Genomics
Volume: 13
Issue number: Suppl 7
ISSN (Print): 1471-2164
Ratings:
BFI (2018): BFI-level 1
BFI (2017): BFI-level 1
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 4.05 SJR 2.065 SNIP 1.122
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): SJR 2.287 SNIP 1.172 CiteScore 4.3
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Scopus rating (2014): SJR 2.297 SNIP 1.205 CiteScore 4.18
Web of Science (2014): Indexed yes
BFI (2013): BFI-level 1
Scopus rating (2013): SJR 2.141 SNIP 1.174 CiteScore 4.39
ISI indexed (2013): ISI indexed yes
Web of Science (2013): Indexed yes
BFI (2012): BFI-level 1
Scopus rating (2012): SJR 2.181 SNIP 1.225 CiteScore 4.61
ISI indexed (2012): ISI indexed yes
Web of Science (2012): Indexed yes
BFI (2011): BFI-level 1
Scopus rating (2011): SJR 2.271 SNIP 1.197 CiteScore 4.38
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
BFI (2010): BFI-level 1
Scopus rating (2010): SJR 2.109 SNIP 1.038
Web of Science (2010): Indexed yes
BFI (2009): BFI-level 1
Scopus rating (2009): SJR 2.181 SNIP 1.015
Web of Science (2009): Indexed yes
BFI (2008): BFI-level 1
Scopus rating (2008): SJR 2.067 SNIP 1.005
Web of Science (2008): Indexed yes
Scopus rating (2007): SJR 1.846 SNIP 1.04
Web of Science (2007): Indexed yes
Scopus rating (2006): SJR 1.431 SNIP 0.895
Web of Science (2006): Indexed yes
Scopus rating (2005): SJR 1.994 SNIP 0.894
Web of Science (2005): Indexed yes
Scopus rating (2004): SJR 1.852 SNIP 0.617
Web of Science (2004): Indexed yes
Scopus rating (2003): SJR 1.909 SNIP 0.561
Web of Science (2003): Indexed yes
Scopus rating (2002): SJR 1.342 SNIP 0.532
Scopus rating (2001): SJR 0.132 SNIP 0.016
Original language: English
Electronic versions:
1471_2164_13_S7_S6.pdf
DOIs:
10.1186/1471-2164-13-S7-S6
Links:
Source: dtu
Source-ID: n:oai:DTIC-ART:pubmed/377518828::25251
Publication: Research - peer-review › Journal article – Annual report year: 2012

Projects:

Rapid computational identification and elucidation of infectious disease outbreak
Department of Bio and Health Informatics
Period: 15/04/2014 → 29/07/2017
Number of participants: 6
PhD Student:
Thomsen, Martin Christen Frølund (Intern)
Supervisor:
Aarestrup, Frank Møller (Intern)
Main Supervisor:
Lund, Ole (Intern)
Examiner:
Petersen, Bent (Intern)
Dallman, Tim (Ekstern)
Hansen, Dennis Schrøder (Ekstern)

Financing sources
Source: Internal funding (public)
Name of research programme: Forskningsrådsfinansiering
Project: PhD