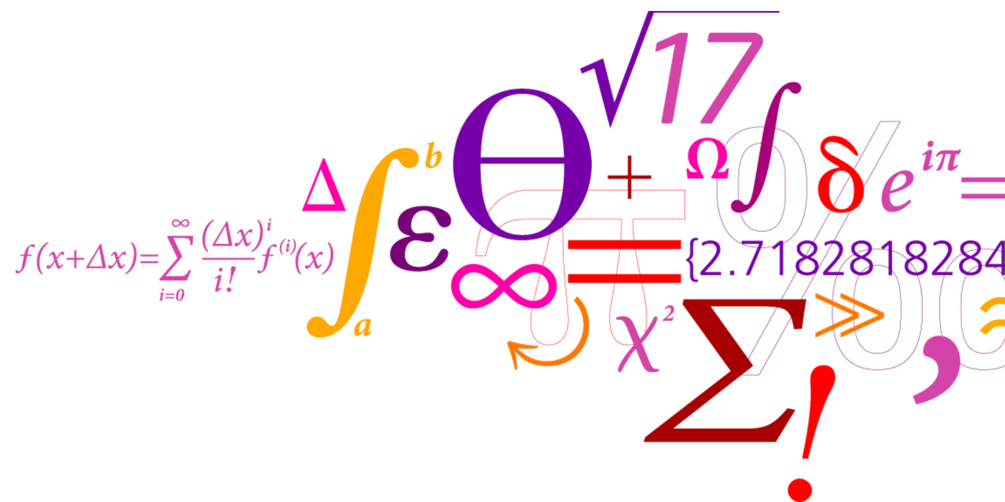
The Perspectives of Whole Genome Sequencing in the Study of VTEC infections

Katrine Grimstrup Joensen



DTU Food
National Food Institute

DTU

My Background

- DTU Food, Division for Epidemiology and Microbial Genomics (Frank Møller Aarestrup and Henrik Hasman)
- SSI, Department of Microbiological Surveillance and Research (Eva Møller Nielsen)
- Ph.d project on "Application of WGS for Diagnostics, Surveillance and Outbreak Detection of Foodborne Pathogens"

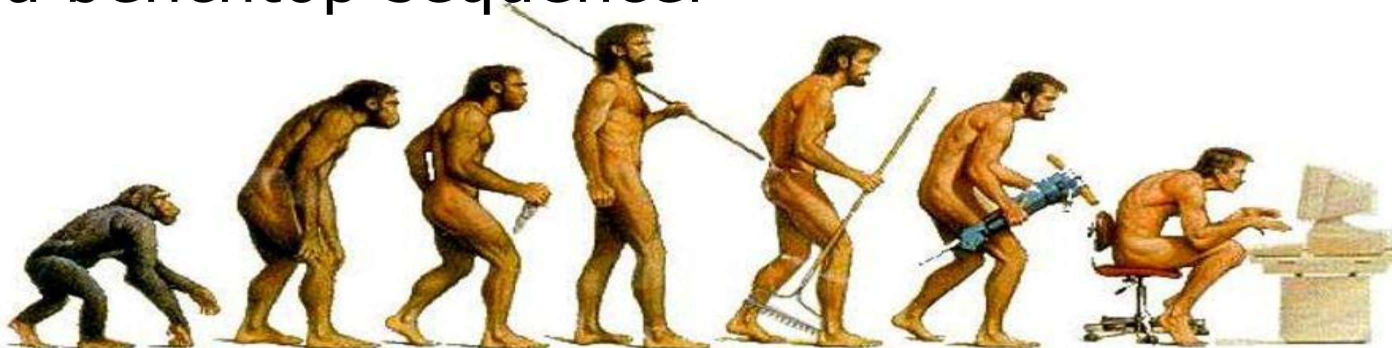
- Current project on Evaluating WGS for Typing of VTEC Infections (with Flemming Scheutz, SSI)

STATENS SERUM INSTITUT

DTU

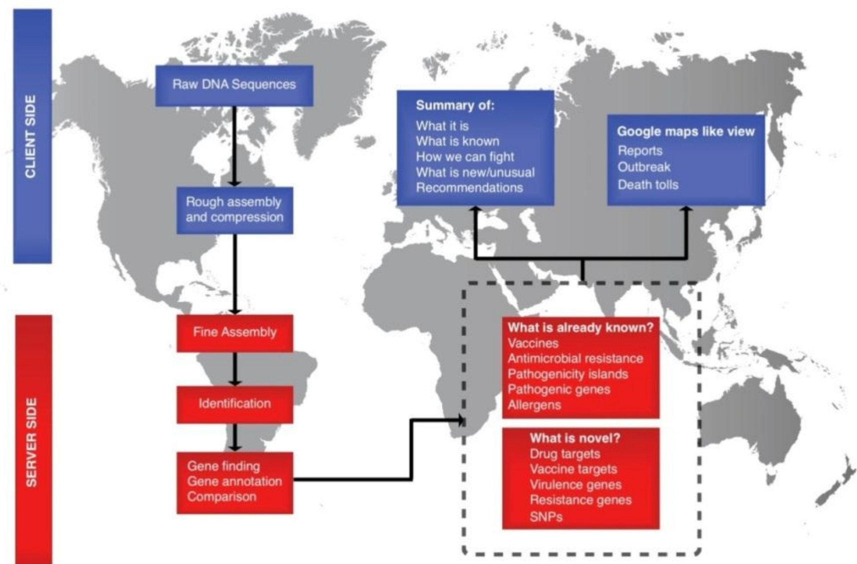
Evolution of typing technologies

- Rapid advancement of NGS holds great promise for improving speed and quality of clinical investigations
- NGS in routine use in public health laboratories (Nationally, Regionally and Globally) for clinical diagnosis and other essential information
- Decreased cost ~100 €/genome
- Future of each microbiological laboratory having a benchtop sequencer



Center for Genomic Epidemiology

- How to assemble, process and handle NGS data in a standardized way for diagnostic and surveillance
- Web-based solutions with central database to simplify WGS information and analysis
- http://www.genomicepidemiology.org/



Center for Genomic Epidemiology

Identification of acquired antimicrobial resistance genes

Ea Zankari[1,2]*, Henrik Hasman[1], Salvatore Cosentino[2], Martin Vestergaard[1], Simon Rasmussen[2], Ole Lund[2], Frank M. Aarestrup[1] and Mette Voldby Larsen[2]

[1]National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; [2]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

*Corresponding author. Tel: +45-35887183; E-mail: east@food.dtu.dk

Received 13 March 2012; returned 26 April 2012; revised 8 June 2012; accepted 13 June 2012

Objectives: Identification of antimicrobial resistance genes is important for understanding the underlying mechanisms and the epidemiology of antimicrobial resistance. As the costs of whole-genome sequencing (WGS) continue to decline, it becomes increasingly available in routine diagnostic laboratories and is anticipated to substitute traditional methods for resistance gene identification. Thus, the current challenge is to extract the relevant information from the large amount of generated data.

Methods: We developed a web-based method, ResFinder that uses BLAST for identification of acquired antimicrobial resistance genes in whole-genome data. As input, the method can use both pre-assembled, complete or partial genomes, and short sequence reads from four different sequencing platforms. The method was evaluated on 1862 GenBank files containing 1411 different resistance genes, as well as on 23 *de-novo*-sequenced isolates.

Results: When testing the 1862 GenBank files, the method identified the resistance genes with an ID=100% (100% identity) to the genes in ResFinder. Agreement between *in silico* predictions and phenotypic testing was found when the method was further tested on 23 isolates of five different bacterial species, with available phenotypes. Furthermore, ResFinder was evaluated on WGS chromosomes and plasmids of 30 isolates. Seven of these isolates were annotated to have antimicrobial resistance, and in all cases, annotations were compatible with the ResFinder results.

Conclusions: A web server providing a convenient way of identifying acquired antimicrobial resistance genes in completely sequenced isolates was created. ResFinder can be accessed at www.genomicepidemiology.org. ResFinder will continuously be updated as new resistance genes are identified.

Genomic variation in *Salmonella enterica* core genes for epidemiological typing

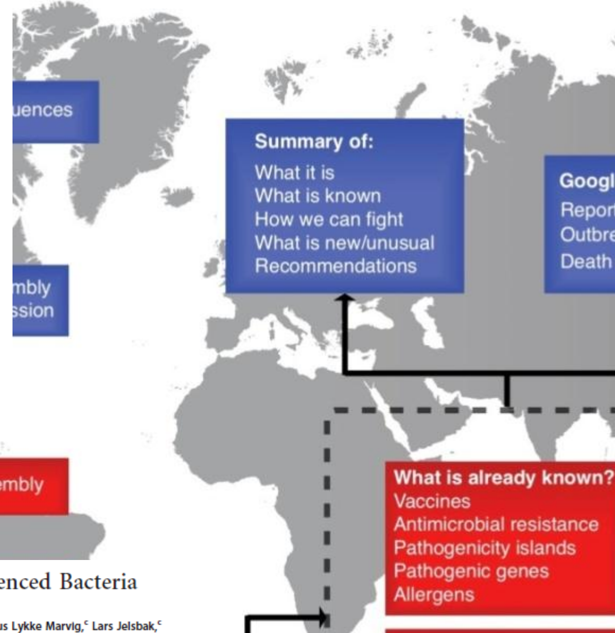
Pimlapas Leekitcharoenphon[1,2], Oksana Lukjancenko[2], Carsten Friis[1], Frank M Aarestrup[1] and David W Ussery[2]*

Abstract

Background: Technological advances in high throughput genome sequencing are making whole genome sequencing (WGS) available as a routine tool for bacterial typing. Standardized procedures for identification of relevant genes and of variation are needed to enable comparison between studies and over time. The core genes–the genes that are conserved in all (or most) members of a genus or species–are potentially good candidates for investigating genomic variation in phylogeny and epidemiology.

Results: We identify a set of 2,882 core genes clusters based on 73 publicly available *Salmonella enterica* genomes and evaluate their value as typing targets, comparing whole genome typing and traditional methods such as 16S and MLST. A consensus tree based on variation of core genes gives much better resolution than 16S and MLST; the pan-genome family tree is similar to the consensus tree, but with higher confidence. The core genes can be divided into two categories: a few highly variable genes and a larger set of conserved core genes, with low variance. For the most variable core genes, the variance in amino acid sequences is higher than for the corresponding nucleotide sequences, suggesting that there is a positive selection towards mutations leading to amino acid changes.

Conclusions: Genomic variation within the core genome is useful for investigating molecular evolution and providing candidate genes for bacterial genome typing. Identification of genes with different degrees of variation is important especially in trend analysis.



Population Genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak

Rene S. Hendriksen,[a] Lance B. Price,[b] James M. Schupp,[b] John D. Gillece,[b] Rolf S. Kaas,[a] David M. Engelthaler,[b] Valeria Bortolaia,[a] Talima Pearson,[c] Andrew E. Waters,[b] Bishnu Prasad Upadhyay,[d] Sirjana Devi Shrestha,[d] Shailaja Adhikari,[d] Geeta Shakya,[d] Paul S. Keim,[b,c] and Frank M. Aarestrup[a]

National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark[a]; Division of Pathogen Genomics, Translational Genomics Research Institute (TGen), Flagstaff, Arizona, USA[b]; Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, Arizona, USA[c]; and National Public Health Laboratory, Kathmandu, Nepal[d]

ABSTRACT Cholera continues to be an important cause of human infections, and outbreaks are often observed after natural disasters, such as the one following the 2010 earthquake in Haiti. Once the cholera outbreak was confirmed, rumors spread that the disease was brought to Haiti by a battalion of Nepalese soldiers serving as United Nations peacekeepers. This possible connection has never been confirmed. We used whole-genome sequence typing (WGST), pulsed-field gel electrophoresis (PFGE), and antimicrobial susceptibility testing to characterize 24 recent *Vibrio cholerae* isolates from Nepal and evaluate the suggested epidemiological link with the Haitian outbreak. The isolates were obtained from 30 July to 1 November 2010 from five different districts in Nepal. We compared the 24 genomes to 10 previously sequenced *V. cholerae* isolates, including 3 from the Haitian outbreak (began July 2010). Antimicrobial susceptibility and PFGE patterns were consistent with an epidemiological link between the isolates from Nepal and Haiti. WGST showed that all 24 *V. cholerae* isolates from Nepal belonged to a single monophyletic group that also contained isolates from Bangladesh and Haiti. The Nepalese isolates were divided into four closely related clusters. One cluster contained three Nepalese isolates and three Haitian isolates that were almost identical, with only 1- or 2-bp differences. Results in this study are consistent with Nepal as the origin of the Haitian outbreak. This highlights how rapidly infectious diseases might be transmitted globally through international travel and how public health officials need advanced molecular tools along with standard epidemiological analyses to quickly determine the sources of outbreaks.

Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria

Mette V. Larsen,[a] Salvatore Cosentino,[a] Simon Rasmussen,[a] Carsten Friis,[b] Henrik Hasman,[b] Rasmus Lykke Marvig,[c] Lars Jelsbak,[c] Thomas Sicheritz-Pontén,[a] David W. Ussery,[a] Frank M. Aarestrup,[b] and Ole Lund[a]

Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark[a]; National Food Institute, Technical University of Denmark, Lyngby, Denmark[b]; and Center for Systems Microbiology, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark[c]

Accurate strain identification is essential for anyone working with bacteria. For many species, multilocus sequence typing (MLST) is considered the "gold standard" of typing, but it is traditionally performed in an expensive and time-consuming manner. As the costs of whole-genome sequencing (WGS) continue to decline, it becomes increasingly available to scientists and routine diagnostic laboratories. Currently, the cost is below that of traditional MLST. The new challenges will be how to extract the relevant information from the large amount of data so as to allow for comparison over time and between laboratories. Ideally, this information should also allow for comparison to historical data. We developed a Web-based method for MLST of 66 bacterial species based on WGS data. As input, the method uses short sequence reads from four sequencing platforms or preassembled genomes. Updates from the MLST databases are downloaded monthly, and the best-matching MLST alleles of the specified MLST scheme are found using a BLAST-based ranking method. The sequence type is then determined by the combination of alleles identified. The method was tested on preassembled genomes from 336 isolates covering 56 MLST schemes, on short sequence reads from 387 isolates covering 10 schemes, and on a small test set of short sequence reads from 29 isolates for which the sequence type had been determined by traditional methods. The method presented here enables investigators to determine the sequence types of their isolates on the basis of WGS data. This method is publicly available at www.cbs.dtu.dk/services/MLST.

A Closer Look at *Bacteroides*: Phylogenetic Relationship and Genomic Implications of a Life in the Human Gut

Fredrik H. Karlsson · David W. Ussery · Jens Nielsen · Intawat Nookaew

Received: 1 October 2010 / Accepted: 14 December 2010 / Published online: 11 January 2011

© Springer Science+Business Media, LLC 2011

Abstract The human gut is extremely densely inhabited by bacteria mainly from two phyla, Bacteroidetes and Firmicutes, and there is a great interest in analyzing whole-genome sequences for these species because of their relation to human health and disease. Here, we do whole-genome comparison of 105 *Bacteroidetes/Chlorobi* genomes to elucidate their phylogenetic relationship and to gain insight into what is separating the gut living *Bacteroides* and *Parabacteroides* genera from other *Bacteroidetes/Chlorobi* species. A comprehensive analysis shows that *Bacteroides* species have a higher number of extracytoplasmic function σ factors (ECF σ factors) and two component systems for extracellular signal transduction compared to other *Bacteroidetes/Chlorobi* species. A whole-genome phylogenetic analysis shows a very little difference between the *Parabacteroides* and *Bacteroides* genera. Further analysis shows that *Bacteroides* and *Parabacteroides* species share a large common core of 1,085 protein families. Genome atlases illustrate that there are few and only small unique areas on the chromosomes of four *Bacteroides/Parabacteroides* genomes. Functional classification to clusters of othologus groups show that *Bacteroides* species are enriched in carbohydrate transport and metabolism proteins. Classification of proteins in KEGG metabolic pathways gives a detailed view of the genome's metabolic capabilities that can be linked to its habitat. *Bacteroides pectinophilus* and *Bacteroides capillosus* do not cluster together with other *Bacteroides* species, based on analysis of 16S rRNA sequence, whole-genome protein families and functional content, 16S rRNA sequences of the two species suggest that they belong to the Firmicutes phylum. We have presented a more detailed and precise description of the phylogenetic relationships of members of the *Bacteroidetes/Chlorobi* phylum by whole genome comparison. Gut living *Bacteroides* have an enriched set of glycan, vitamin, and cofactor enzymes important for diet digestion.

Benchtop Sequencing

- Miseq –High data throughput, High quality

~150bp reads
8Gb output

~ 40 h pr. run

- IonTorrent PGM
- ~200bp reads
1Gb output

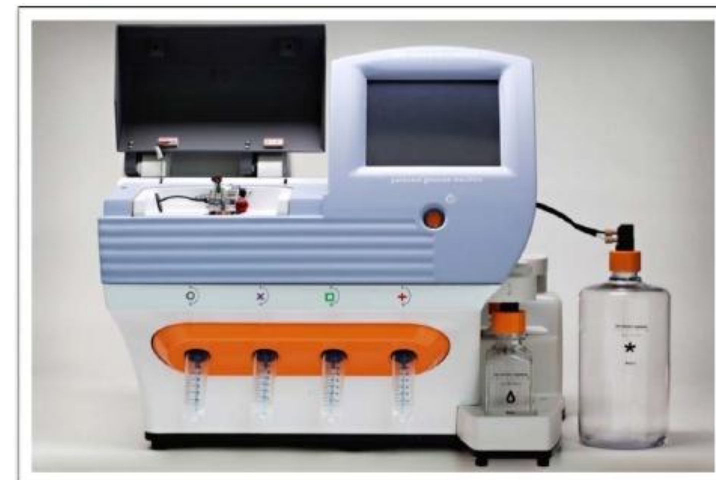
~18 h pr. run

Illumina MiSeq

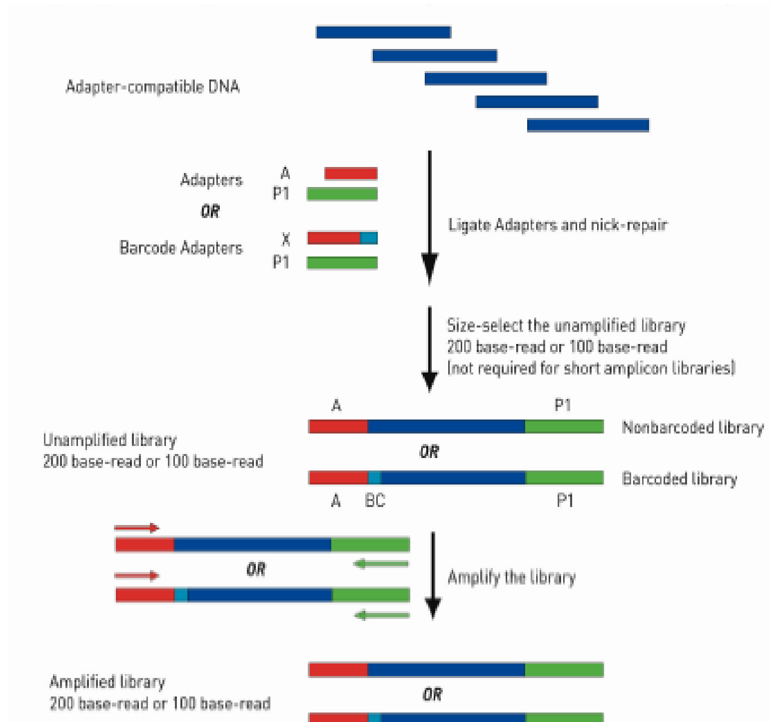
Ion Torrent PGM

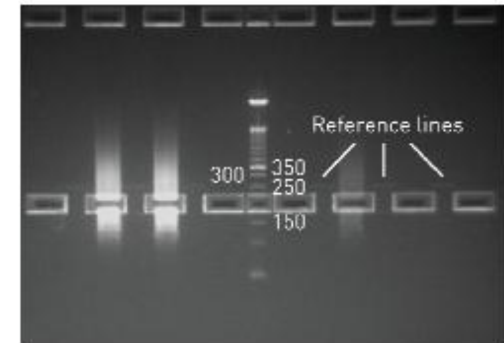
Evaluation of WGS for typing of VTEC infections

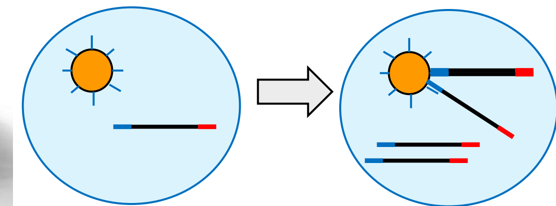
- Real-time project for evaluating the application of WGS for Typing of VTEC
- WGS in parallel with traditional typing at SSI on suspected VTEC isolates
- Sequencing 40 isolates (or for 3 months) on IonTorrent PGM
- Comparing Time, Price and Typing Results

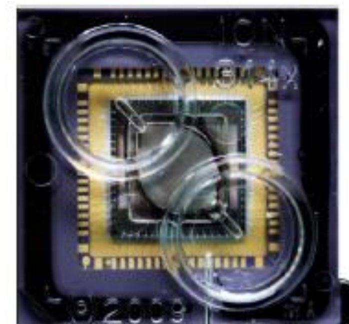


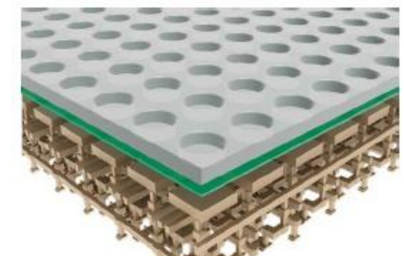
The sequencing workflow



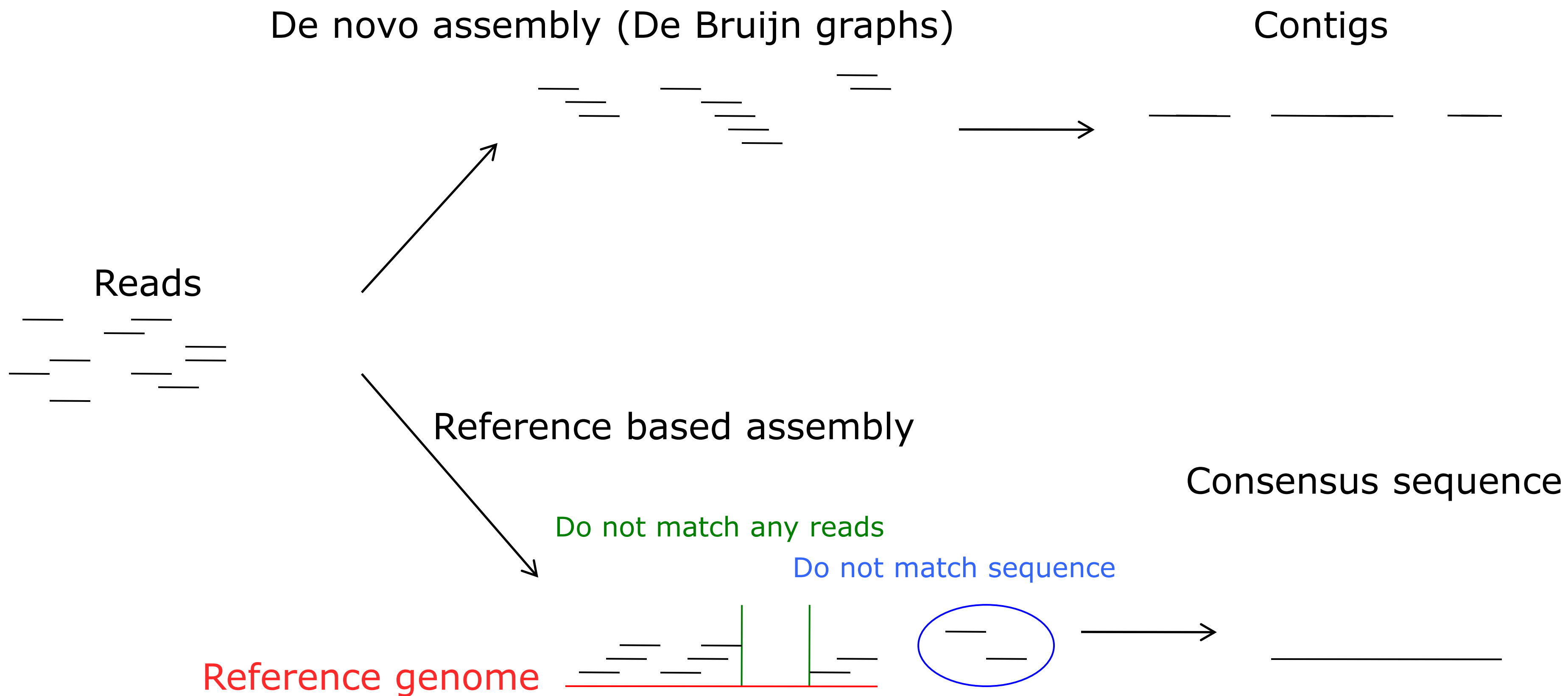




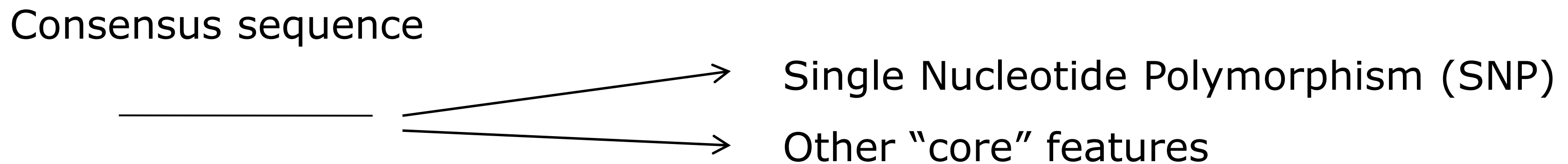




The sequencing workflow



Further data analysis



How do we compare SSI typing results to our WGS results?

- ResFinder for detecting antibiotic resistance genes
- MLST finder and snpTree to determine relationships between isolates
  (new method for detecting variants)

- Constructing VirulenceFinder -a database containing important virulence genes
  So far for VTEC and *Enterococcus*
- PFGE and Serotyping

DTU

Construction of VirulenceFinder

Center for Genomic Epidemiology

Username
Password
New login | Login

Home | Services | Instructions | Output

NEWS: Dear all CGE service users, Show

VirulenceFinder 1.0

Browse | Remove | Clear

Uploads

Total files: 0 (N/A).

Select configuration
Select multiple items, with Ctrl-Click (or Cmd-Click on Mac)
VTEC
Enterococcus

Select threshold for %ID
100 %

Select type of your reads
Ion Torrent

Submit | Clear fields

- VTEC virulence gene variants *stx1*, *stx2* and *eae*

  …and other *E.coli* virulence genes

K88ab	***eae***	*flmA*	*mchB, mchC,*	*sepA*
aggR	*eatA*	*flnA*	*mchF*	*sfaS*
astA	*efa1*	*hlyB, hlyC,*	*mcmA*	*sigA*
bfpA	*ehx*	*hlyD, hlyE*	*nleA, nleB, nleE*	*sta1A,*
cba	*epeA*	*ipaH9.8*	*paa*	*sta1B*
cdtB	*espJ, espP*	*ireA*	*perA*	*stb*
celb	*etpD*	*iroN*	*pet*	***stx1***
cfaC	*f17A*	*iss*	*prfB*	***stx2***
cif	*fanA*	*katP*	*rpeA*	*subA*
cma	*fasA*	*lngA*	*saa*	*toxB*
cnf	*fim41a*	*lpfA*	*sat*	*tsh*
eaaA	*fliC*	*lthA*	*senB*	*virF*

VirulenceFinder

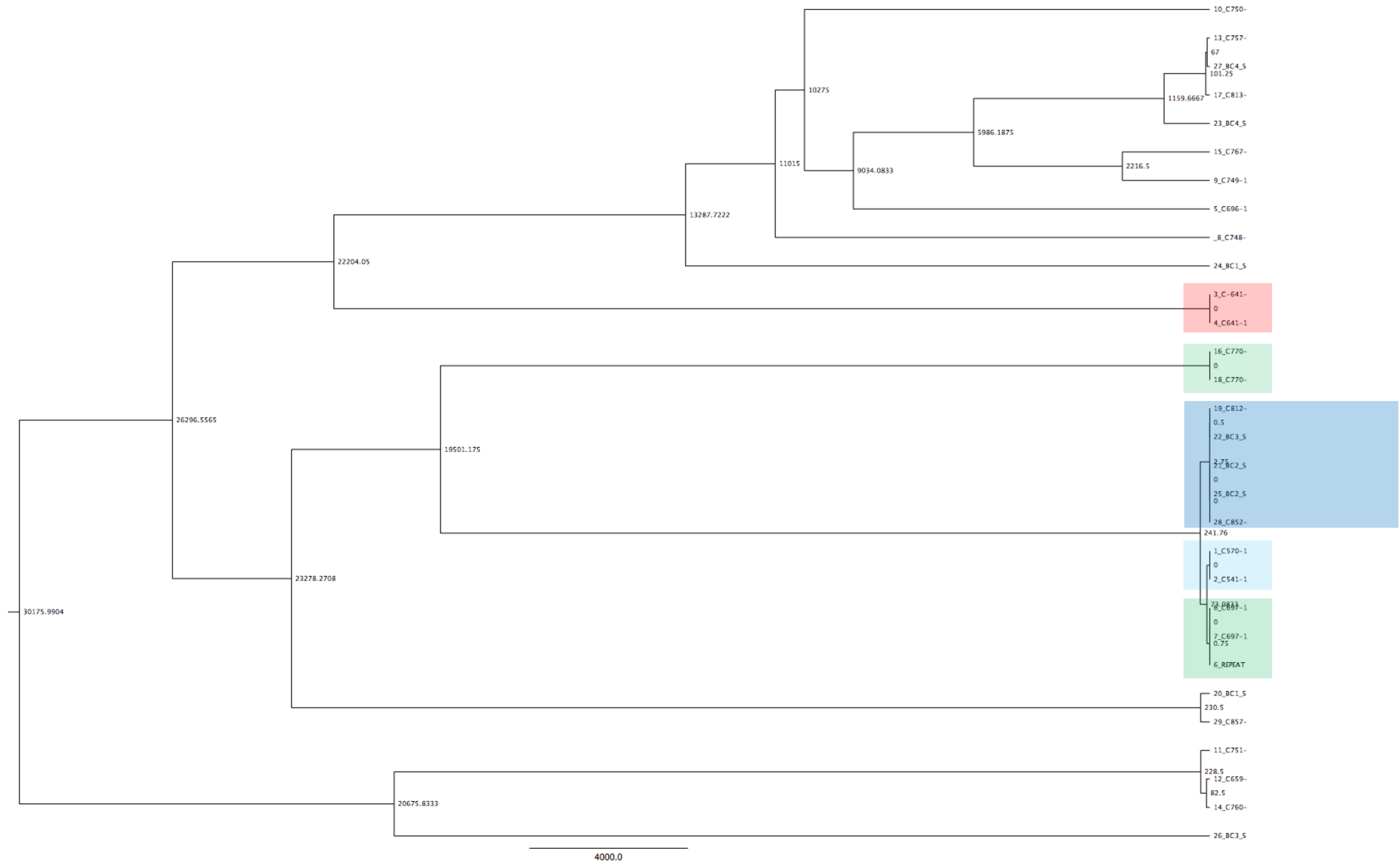
Virulence - VTEC						
Virulence factor	%Identity	Query/HSP length	Contig	Position in contig	Protein function	Accession number
eae	100.00 %	2805/2805	NODE_13_length_146599_cov_22.571287	135909..138713		AE005174
ehx	100.00 %	2997/2997	NODE_349_length_28606_cov_14.518423	17588..20584		AB011549
espJ	100.00 %	654/654	NODE_92_length_33015_cov_20.320915	31744..32397		AE005174
espP	99.97 %	3903/3903	NODE_69_length_11528_cov_15.707668	7491..11393		AF074613
etpD	100.00 %	1929/1929	NODE_349_length_28606_cov_14.518423	4345..6273		AF074613
fliC	100.00 %	1758/1758	NODE_27_length_20338_cov_20.148294	7285..9042		NC013941
flmA	100.00 %	159/159	NODE_33_length_16006_cov_17.142883	1843..2001		NC007414
hlyA	100.00 %	2997/2997	NODE_349_length_28606_cov_14.518423	17588..20584		NC_002128
hlyB	100.00 %	2121/2121	NODE_349_length_28606_cov_14.518423	20634..22754		NC002128
hlyC	100.00 %	516/516	NODE_349_length_28606_cov_14.518423	17071..17586		NC012487
hlyD	100.00 %	1440/1440	NODE_349_length_28606_cov_14.518423	22758..24197		NC002128
katP	100.00 %	2211/2211	NODE_61_length_3768_cov_15.060775	911..3121		X89017
nleB	100.00 %	981/981	NODE_1_length_138466_cov_19.294975	137377..138357		NC013941
nleE	100.00 %	675/675	NODE_72_length_133044_cov_19.993483	128370..129044		NC013353
paa	100.00 %	747/747	NODE_267_length_8958_cov_25.055927	6201..6947		NC013361
stx1	100.00 %	1227/1227	NODE_194_length_4028_cov_22.959534	1524..2750		AB015056
stx2	100.00 %	1241/1241	NODE_257_length_3396_cov_26.901060	2070..3310		AB290938
toxB	99.99 %	9510/9510	NODE_33_length_16006_cov_17.142883	5645..15154		NC017907

extended output

stx1: PERFECT MATCH, Identity: 100.00%, HSP/Length: 1227/1227

DTU

WGS Typing Results



WGS Typing results

Evaluating WGS for typing of VTEC infections

- Time:
  ~18 hours from DNA to sequence data (2.5 normal working days)
  ~8 hours hands-on time

- Price: ~500 € pr. isolate in materials

Acknowledgements

- DTU Food, Division for Epidemiology and Microbial Genomics: Frank Møller Aarestrup, Henrik Hasman, Ole Lund and Rolf Sommer Kaas

- SSI, Department of Microbiological Surveillance and Research: Eva Møller Nielsen and Flemming Scheutz

Questions?