Evira

MATKAKERTOMUS
Anniina Jaakkonen

Tutkimus- ja laboratorio-osasto
Elintarvike- ja rehumikrobiologian tutkimusyksikkö

Pvm/Datum/Date
22.6.2016

Dnro/Dnr/DNo

Tweb id:

2nd course on bioinformatics tools for Next Generation Sequencing data mining: use of bioinformatics tools for typing pathogenic *E. coli*

Aika:	16.–17.6.2016
Järjestäjä & paikka:	EU Reference Laboratory for *E. coli* (EU-RL VTEC) Istituto Superiore di Sanitá, Rooma, Italia
Verkkosivu:	http://www.iss.it/vtec/index.php?lang=2&id=248&tipo=22 (kurssimateriaalit tulevat tänne)

EURL VTEC järjesti kansallisille vertailulaboratorioille suunnatun kurssin, jossa esiteltiin erityisesti STEC-bakteerien kokogenomisekvenssidatan analysointiin soveltuvia, selainpohjaisia työkaluja. Osa työkaluista oli käytettävissä EURL:n ylläpitämällä Galaxy-pohjaisella ARIES-alustalla (https://w3.iss.it/site/aries/) ja osa Tanskan teknillisen yliopiston (DTU) genomiepidemiologiakeskuksen (CGE) ylläpitämällä sivustolla (http://www.genomicepidemiology.org/). Kummatkin palvelut ovat ilmaisia, mutta ARIES vaatii rekisteröitymisen. Kurssi koostui luennoista ja tietokoneharjoituksista seuraavilla aihealueilla:

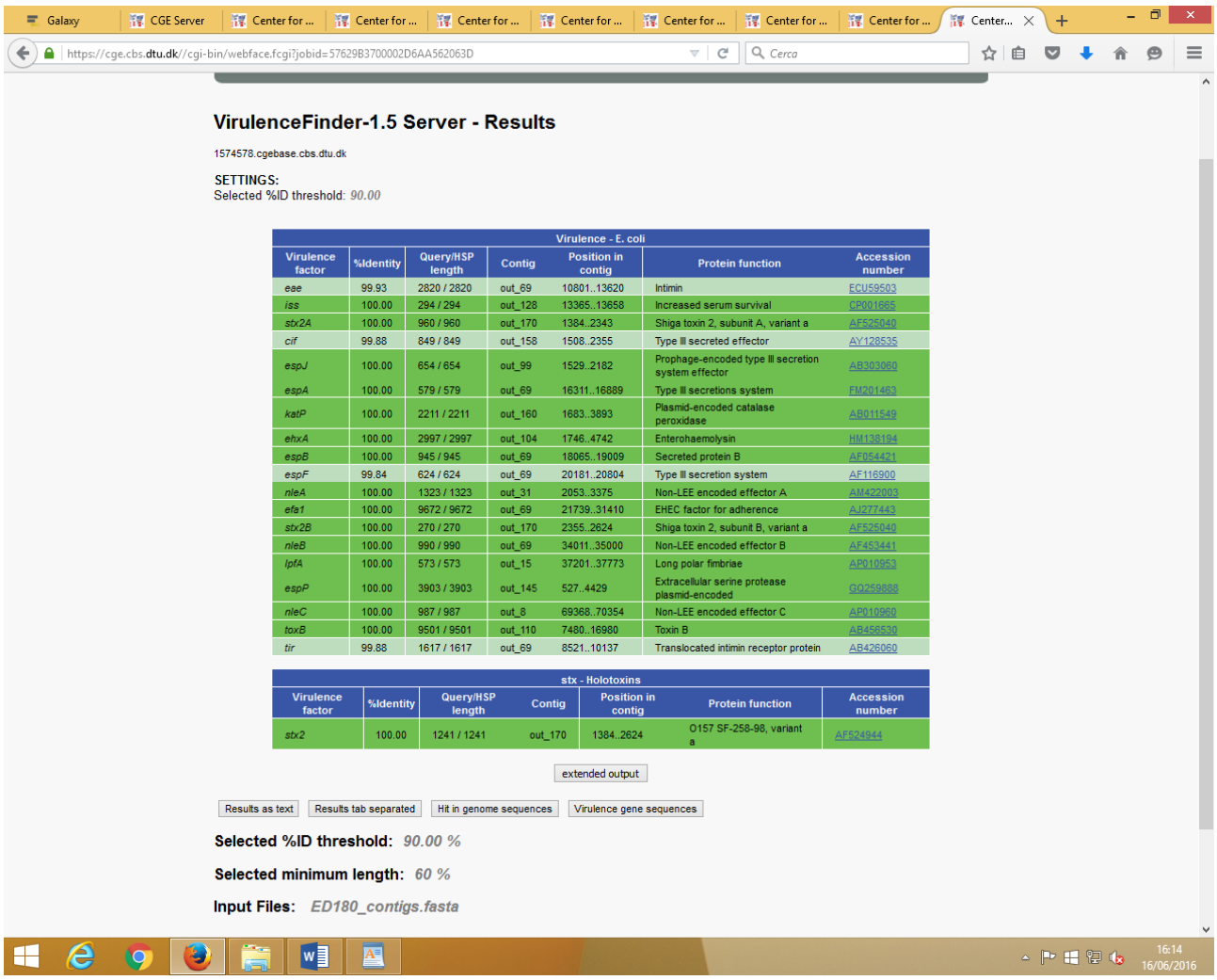
1. EFSA:n ja ECDC:n kokogenomisekvensointiin (WGS) liittyvät aktiviteetit
2. Galaxy-arkkitehtuuri ja ARIES-klusteri
3. Virulotyypitys
4. Serotyypitys
5. wgSNP-tyypitys (whole genome single nucleotide polymorphism)
6. HReVAP (high resolution virulence allelic profiling)
7. MLST ja agMLST (accessory genome multi-locus sequence typing)

Aluksi **EFSA:n ja ECDC:n** edustajat esittelivät WGS-aktiviteettejaan: järjestettyjä kokouksia, julkaistuja asiantuntijalausuntoja ja rahoitettuja hankkeita. ECDC:llä FWD-Next asiantuntijaryhmä on julkaissut asiantuntijalausunnon aiheesta (http://ecdc.europa.eu/en/publications/Publications/food-and-waterborne-diseases-next-generation-typing-methods.pdf) ja julkaisee pian ECDC:n WGS-strategian epidemianselvityksiin. Haasteena on vertailukelpoisen tyypitysdatan saaminen tulevaisuudessa, kun osa kansallisista laboratorioista siirtyy WGS-menetelmiin, kun taas osa pitäytyy vanhoissa menetelmissä. ECDC oli teettänyt kyselyn WGS-menetelmiin siirtymisestä kansallisille laboratorioille ja EFSA suunnitteli toteuttavansa samanlaisen kyselyn EURL:n avustuksella. Tarvittaessa ECDC teettää sekvensointeja kansallisten laboratorioiden lähettämistä kannoista Source Bioscience (UK) -yrityksellä. Toistaiseksi ECDC:llä ei ole kuitenkaan kerätty tai analysoitu STEC-bakteerien WGS-dataa.

Galaxy on avoimen lähdekoodin IT-arkkitehtuuri, jota EURL käyttää **ARIES**-alustassaan. ARIES tarjoaa selainpohjaisen, graafisen käyttöliittymän ja erityisesti STEC-bakteerien WGS-datan analysoimiseen tarkoitettuja työkaluja. ARIES:in tarjoamia työkaluja käyttäjä voi itse automatisoida työvuoksi (pipeline/workflow) visuaali-

sella työkalulla (Workflow canvas). Galaxy-arkkitehtuuri mahdollistaa myös kotitekoisten työkalujen kehityksen, joiden lisäämistä ARIES:iin voi pyytää EURL:ltä. EURL ei kerää eikä varmuuskopioi käyttäjien ARIES:iin lataamaa dataa. ARIES on ollut julkisessa käytössä syyskuusta 2015 lähtien ja sillä on tällä hetkellä 40 käyttäjää (3 Suomessa, 22 Italiassa).

Virulotyypitystä eli virulenssigeenien etsimistä genomista harjoiteltiin kurssilla sekä CGE:n VirulenceFinder-työkalun (kuva 1) että ARIES:in Virulotyper-työkalun (kuva 2) avulla. VirulenceFinder perustuu blastn-hakuun CGE:n virulenssigeenitietokannasta. Se hyväksyy syötteeksi sekä koontisekvenssejä (engl. contig) fasta-tiedostoina että lukusekvenssejä (engl. read) fastq-tiedostoina. Mikäli syötteenä käyttää lukusekvenssejä, työkalu suorittaa *de novo* -koonnin joka tapauksessa ennen blastn-hakua. Virulotyper-työkalu sen sijaan ottaa syötteeksi lukusekvenssejä, jotka se rinnastaa CGE:n tietokannan referenssisekvensseihin (engl. mapping, bowtie2-algoritmi). Rinnastus on nopeampaa kuin blastn-haku. Tulokseksi Virulotyper näyttää parhaan osuman kustakin geenistä (tietokannassa useita alleeleja) ja rinnastuksen kattavuuden (engl. coverage). Mikäli kattavuus on alle 10, geenin läsnäoloon kannattaa suhtautua varauksella. Kattavuus vaihtelee kuitenkin geenikohtaisesti; joidenkin genomialueiden sekvensointi on hankalampaa ja kattavuus jää aina matalaksi (esim. liikkuvat geneettiset elementit kuten *stx*). Lisäksi DNA-eristysmenetelmä vaikuttaa siihen, todetaanko plasmideissa sijaitsevia geenejä. EURL:n kokemuksen mukaan DNA-eristys pylväskitillä säilyttää plasmidit hieman paremmin kuin DNA-eristys saostusmenetelmällä. ”Ei todettu” -tuloksiin on siis suhtauduttava varauksella. Virulotyypityksen varmuutta voi lisätä käyttämällä kumpaakin menetelmää (algoritmia) samalle datalle. Lisäksi hakuarvoja väljentämällä voi hakea esimerkiksi geenejä, joiden sekvensointi on onnistunut vain osittain.



VirulenceFinder-1.5 Server - Results

1574578.cgebase.cbs.dtu.dk

SETTINGS:
Selected %ID threshold: *90.00*

Virulence - E. coli

Virulence factor	%Identity	Query/HSP length	Contig	Position in contig	Protein function	Accession number
eae	99.93	2820 / 2820	out_69	10801..13620	Intimin	ECU59503
iss	100.00	294 / 294	out_128	13365..13658	Increased serum survival	CP001665
stx2A	100.00	960 / 960	out_170	1384..2343	Shiga toxin 2, subunit A, variant a	AF525040
cif	99.88	849 / 849	out_158	1508..2355	Type III secreted effector	AY128535
espJ	100.00	654 / 654	out_99	1529..2182	Prophage-encoded type III secretion system effector	AB303060
espA	100.00	579 / 579	out_69	16311..16889	Type III secretions system	FM201463
katP	100.00	2211 / 2211	out_160	1683..3893	Plasmid-encoded catalase peroxidase	AB011549
ehxA	100.00	2997 / 2997	out_104	1746..4742	Enterohaemolysin	HM138194
espB	100.00	945 / 945	out_69	18065..19009	Secreted protein B	AF054421
espF	99.84	624 / 624	out_69	20181..20804	Type III secretion system	AF116900
nleA	100.00	1323 / 1323	out_31	2053..3375	Non-LEE encoded effector A	AM422003
efa1	100.00	9672 / 9672	out_69	21739..31410	EHEC factor for adherence	AJ277443
stx2B	100.00	270 / 270	out_170	2355..2624	Shiga toxin 2, subunit B, variant a	AF525040
nleB	100.00	990 / 990	out_69	34011..35000	Non-LEE encoded effector B	AF453441
lpfA	100.00	573 / 573	out_15	37201..37773	Long polar fimbriae	AP010953
espP	100.00	3903 / 3903	out_145	527..4429	Extracellular serine protease plasmid-encoded	GQ259888
nleC	100.00	987 / 987	out_8	69368..70354	Non-LEE encoded effector C	AP010960
toxB	100.00	9501 / 9501	out_110	7480..16980	Toxin B	AB456530
tir	99.88	1617 / 1617	out_69	8521..10137	Translocated intimin receptor protein	AB426060

stx - Holotoxins

Virulence factor	%Identity	Query/HSP length	Contig	Position in contig	Protein function	Accession number
stx2	100.00	1241 / 1241	out_170	1384..2624	O157 SF-258-98, variant a	AF524944

extended output

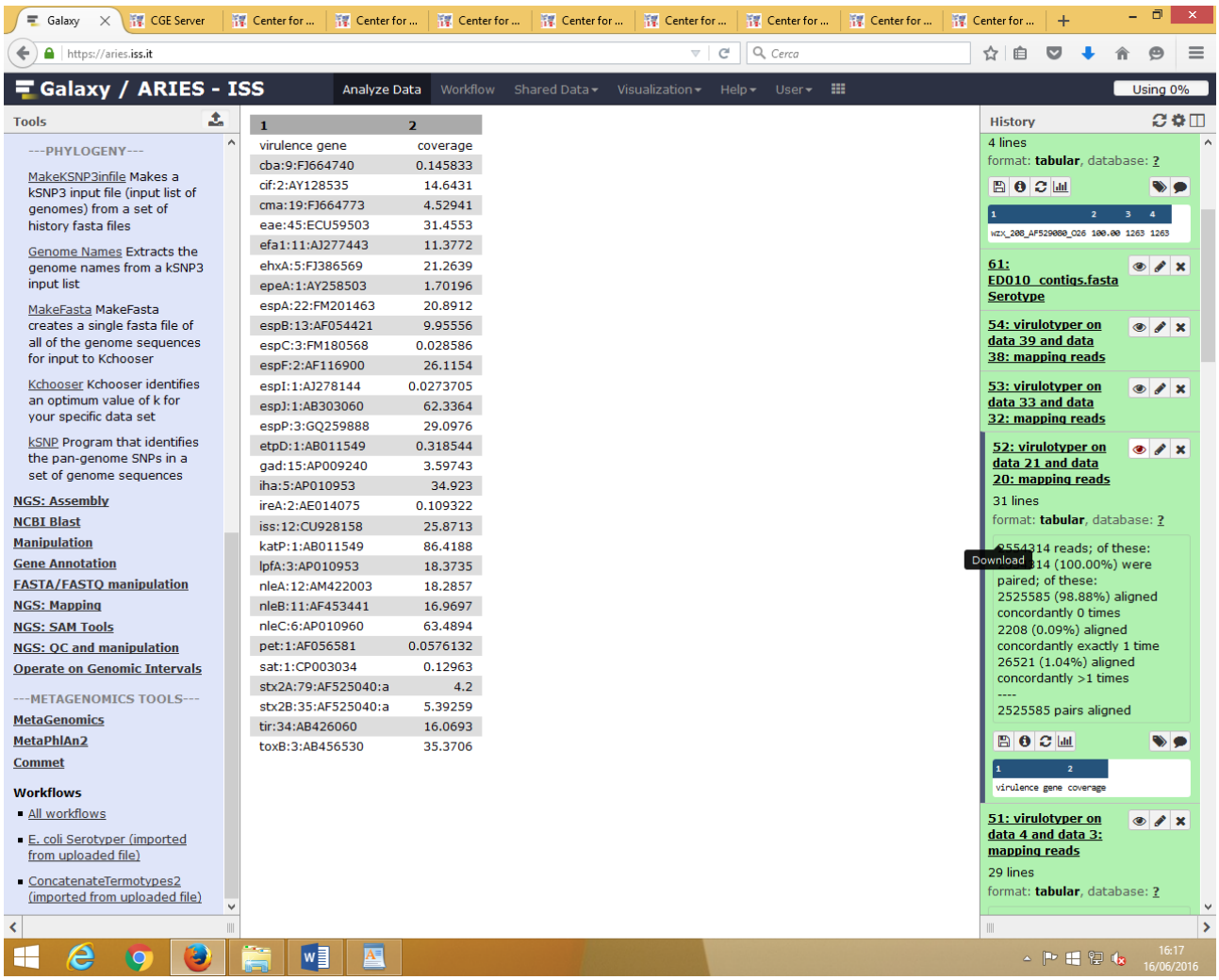
Results as text | Results tab separated | Hit in genome sequences | Virulence gene sequences

Selected %ID threshold: *90.00 %*

Selected minimum length: *60 %*

Input Files: *ED180_contigs.fasta*

Kuva 1. VirulenceFinder (CGE) -tulokset kannasta ED180.

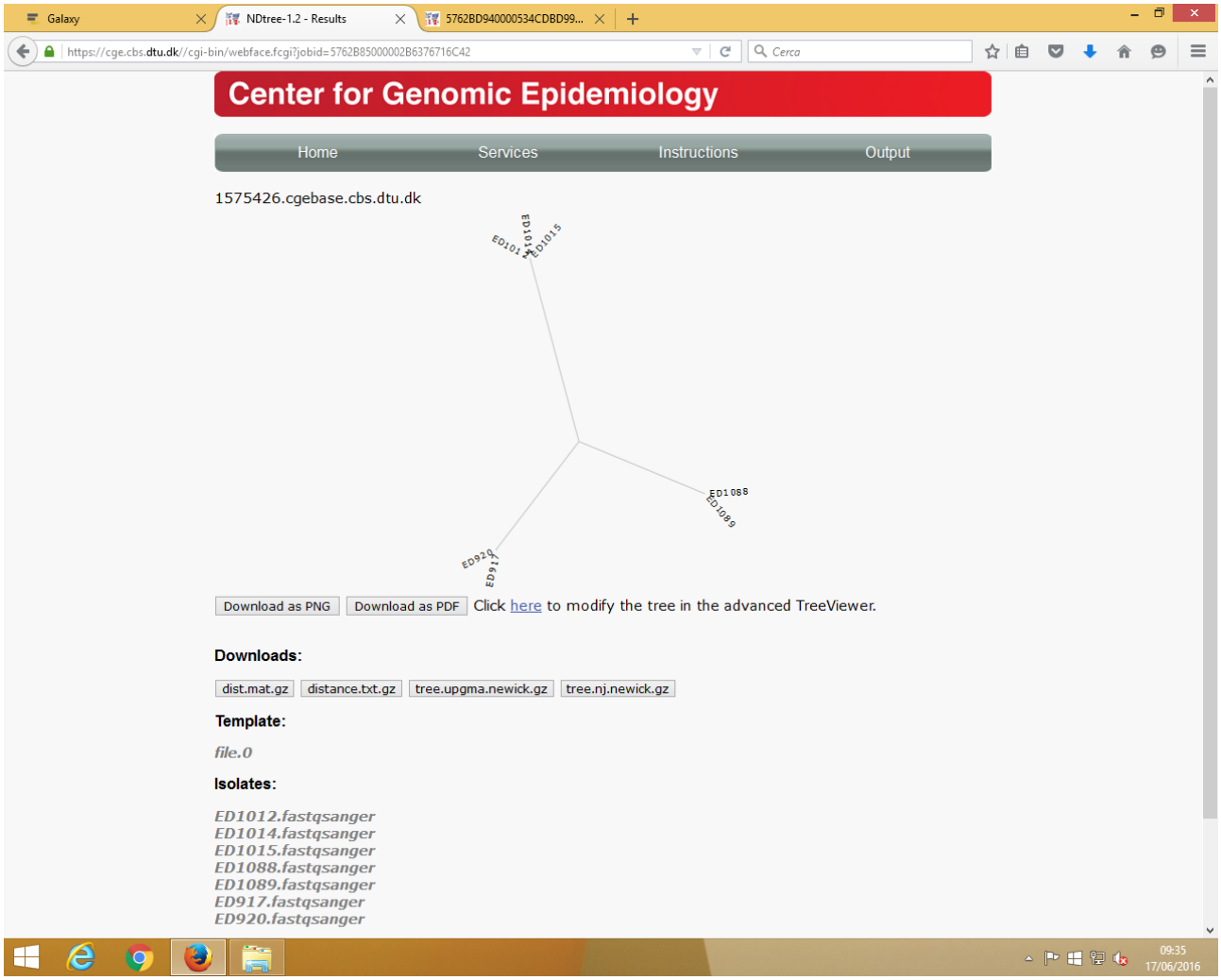


Kuva 2. Virulotyper (ARIES) -tulokset kannasta ED180.

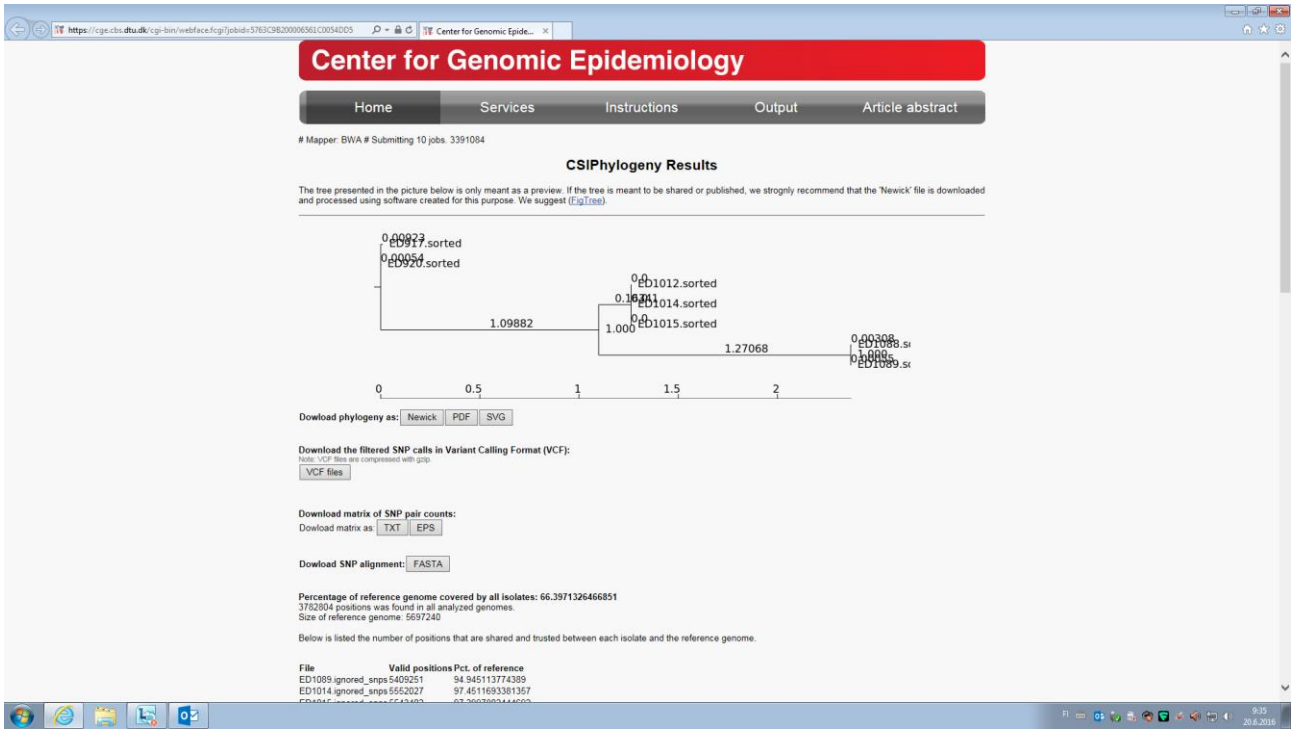
Serotyypitykseen käytettiin CGE:n SerotypeFinder-työkalua ja ARIES:in Serotyper-työkalua. Kummatkin työkalut perustuvat blastn-hakuun samasta CGE:n tietokannasta ja käyttävät syötteenä fasta-koontisekvenssejä. SerotypeFinderiin voi lisäksi syöttää lukusekvenssejä, mutta työkalu suorittaa niille ensin *de novo* -koonnin.

wgSNP-analyysia eli bakteerigenomien fylogeneettistä analyysiä SNP:ien perusteella kokeiltiin kahden CGE:n työkalun (NDtree, CSI Phylogeny) ja ARIES:in ksnp3-työkalun avulla. CGE:n snpTree-työkalu ei ollut käytettävissä kurssipäivänä. CGE:n työkalut perustuvat referenssigenomiin, johon tutkittavien kantojen lukusekvenssit rinnastetaan. Näin ollen referenssin valinta vaikuttaa oleellisesti tulokseen. CGE:n työkaluista yksinkertaisin ja samalla suljetuin systeemi on NDtree (kuva 3), jossa käyttäjä valitsee itse ainoastaan referenssigenomin, eikä voi vaikuttaa muihin parametreihin. SnpTree sallii käyttäjän valita vähimmäiskattavuuden ja etäisyyden SNP:en välillä, mutta ei tee laadunvarmistusta. CSI Phylogeny (kuva 4) sen sijaan tekee myös laadunvarmistusta, mikäli syötteenä on lukusekvenssejä (hyväksyy myös koonteja). Lisäksi se käyttää fylogenian määrittämiseen vain SNP-sijainteja, joiden laatu kantajoukossa on riittävän hyvä. Se hyväksyy syötteeksi myös kantajoukkoja, joista osa on sekvensoitu eri sekvensaattorilla. Kaikki CGE:n työkalut laskevat ja vi-

sualisoivat UPGMA-puun. Neighbor joining -puu on myös saatavilla, mutta sitä varten tulostiedosto täytyy ladata ulkoiseen visualisointiohjelmaan (esim. FigTree). EURL:n kokeilussa NDtree osoittautui epäherkäksi, eikä onnistunut erottelemaan epidemiaan kuulumattomia STEC-kantoja epidemiakannoista. Puusta ei saanut selville etäisyyksiä SNP:ien lukumäärinä. CSI Phylogeny erotteli epidemiakannat muista, muttei onnistunut erottelemaan kantoja saman epidemian sisällä.

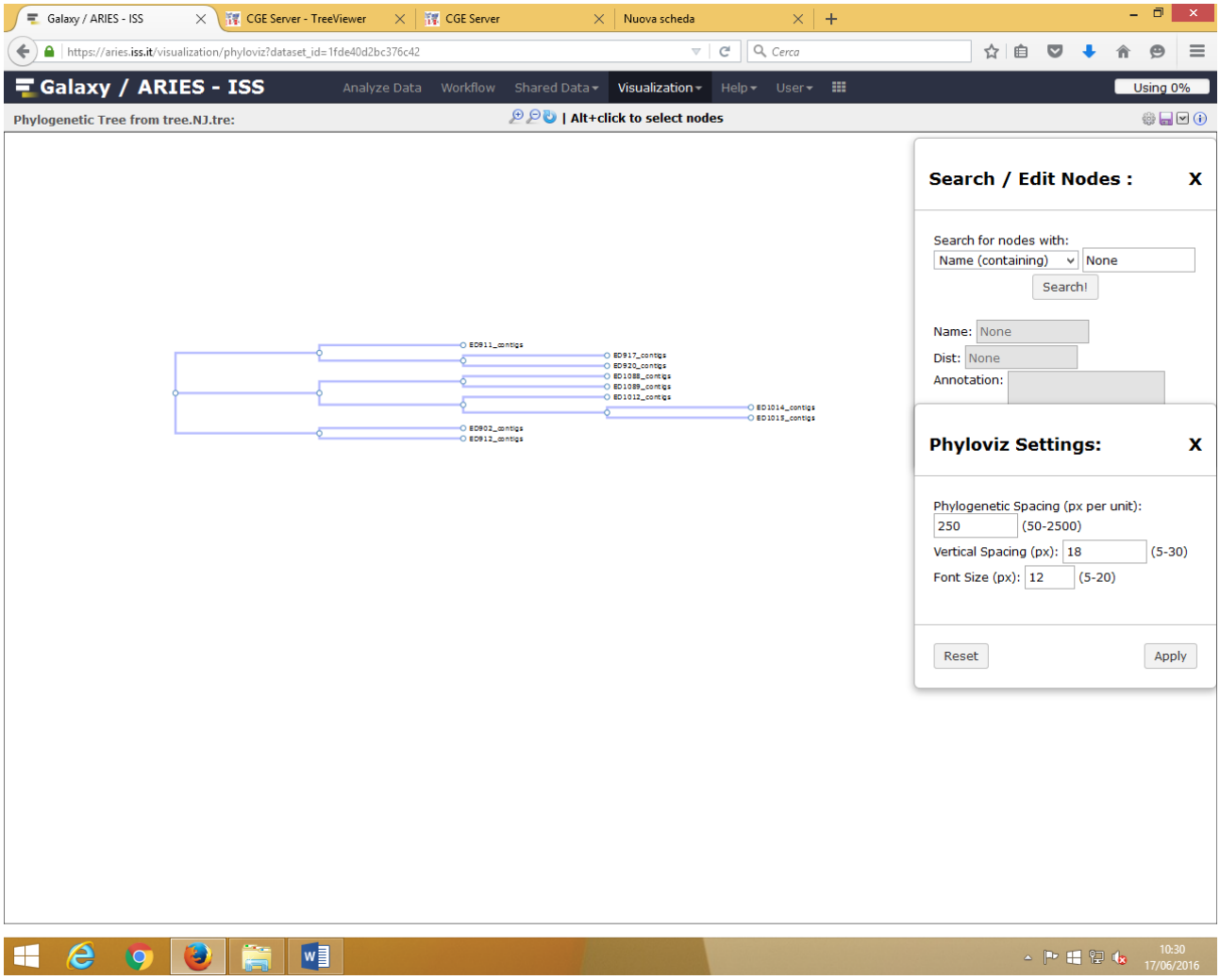


Kuva 3. NDtree-tulos.



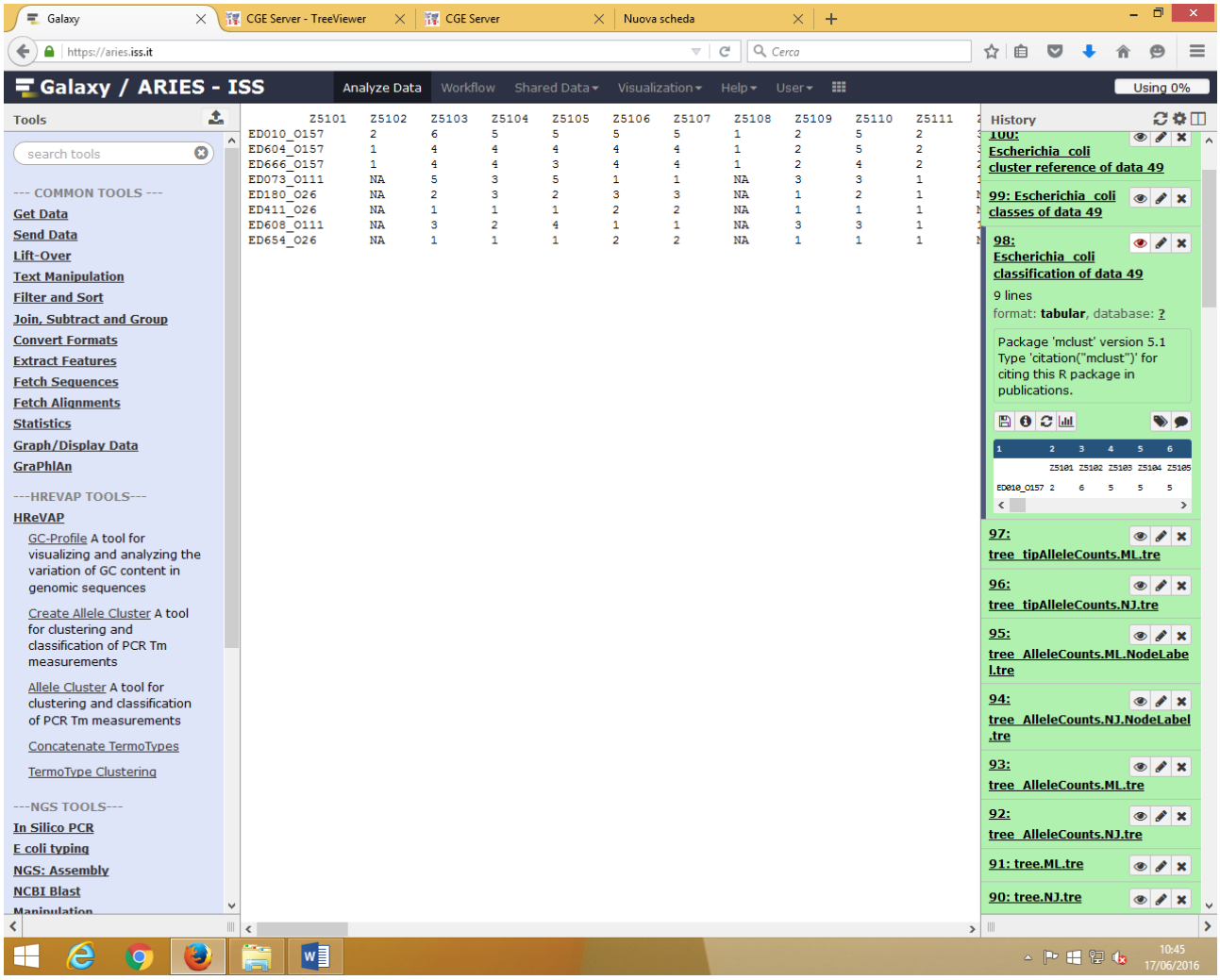
Kuva 4. CSI Phylogeny -tulos. Tuloksista (tulossivulla alempana) näkee lisäksi hylättyjen SNP-sijaintien lukumäärät.

ARIESin ksnp3 (kuva 5) perustuu referenssigenomin sijaan k-mereihin eli työkalu etsii fasta-koontisekvensseistä k:n pituisia nukleotidisekvenssejä, joista se sitten etsii SNP:ejä. Näin työkalu muodostaa SNP-pangenomin tutkittavalle kantajoukolle ja estimoi fylogeneettisiä puita (neighbor joining ja maximum likelihood) näiden SNP:ien perusteella. Ennen ksnp-työkalun käyttöä luodaan syötetiedostot MakeKSNP3infile- ja MakeFasta-työkalujen avulla sekä optimoidaan k (eli maksimoidaan uniikkien k-merien osuus keskipituisessa koontisekvensissä) Kschooser-työkalun avulla. Tulospuun voi visualisoida suoraan ARIES:issa Phyloviz-työkalun avulla tai ladata ulkoiseen visualisointiohjelmaan. Ilman referenssigenomia toimivasta menetelmästä on etua erityisesti STEC-genomien fylogeniaa arvioitaessa, sillä STEC-kantojen monimuotoisuus hankaloittaa referenssigenomin valintaa. Varsinkaan täydellisiä, suljettuja genomisekvenssejä ei ole saatavilla harvinaisemmille serotyypeille. SNP:eihin ja referenssigenomiin perustuvat menetelmät soveltuvat parhaiten keskenään samankaltaisten kantojen fylogeneettiseen analyysiin.

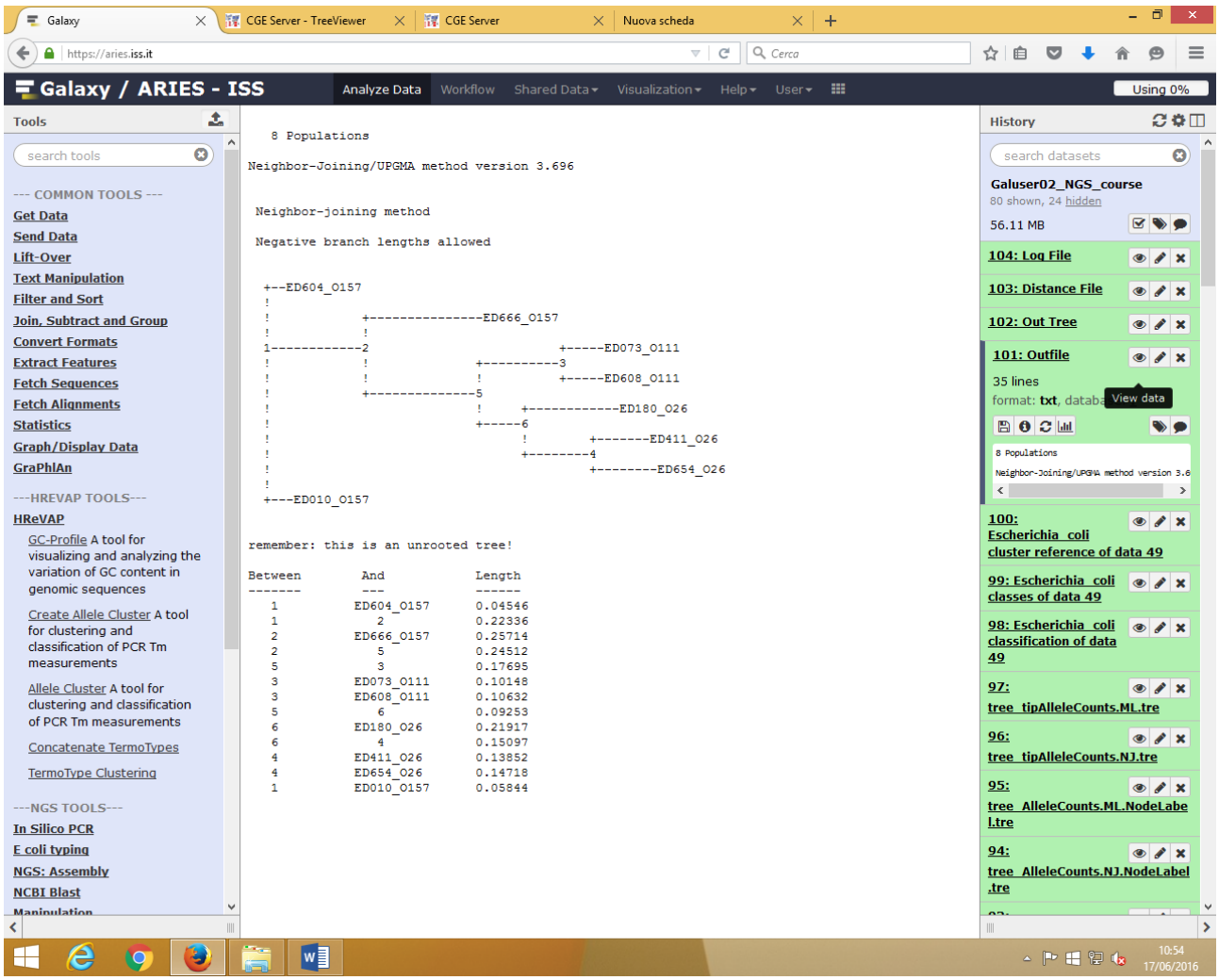


Kuva 5. Ksnp3 (ARIES) -tulos, Phyloviz-visualisointi.

HReVAP on EURL:n kehittämä, ARIES:ssa saatavilla oleva analyysimenetelmä virulenssigeenien (91 geeniä, jotka sijaitsevat LEE, OI-57 ja OI-122 -saarekkeissa) alleelien määrittämiseen ja STEC-kantojen fylogeneettiseen ryhmittelyyn alleeliprofiilin perusteella (http://www.ncbi.nlm.nih.gov/pubmed/26941726). Menetelmä perustuu alleelien erilaisiin PCR-sulamislämpötiloihin ja on kaksivaiheinen: ensin luodaan alleelitaulukko (kuva 6) sulamislämpötilojen perusteella (työkalu: Allele Cluster) ja sitten suoritetaan alleeliprofiilien klusterointi (työkalu: TermoType Clustering, kuva 7). Alleelitaulukon alleeliluokat voidaan joko luoda tutkittavalle kantajoukolle alusta asti (työkalu: Create Allele Cluster), käyttää aiemmin luotua luokittelua tai systeemireferenssiä. Klusterointi tuottaa neighbor joining -puun. EURL:n kokeiluissa menetelmä on onnistuneesti erotellut seroryhmät toisistaan ja erottelukyky on riittänyt myös kantojen välisten erojen havaitsemiseen seroryhmien sisällä. Erottelukyky on ollut ksnp3-analyysin tasoa. Tosin HReVAP:in erottelukyky on heikompi LEE (*eae*) -negatiivisilla kannoilla.



Kuva 6. HReVAP-alleelitaulukko kahdeksalle kannalle (rivi: kanta, sarake: geeni, solu: alleelinumero). Esimerkiksi geenillä Z5102 on kyseisessä kantajoukossa 6 alleeliluokkaa, alleelinumerot 1–6.



Kuva 6. HReVAP-klusterointitulos.

agMLST on ARIES:ssa saatavilla oleva MLST-protokolla, joka perustuu virulenssigeeneihin. Etuna SNP-analyyseihin verrattuna on se, että MLST mahdollistaa yhteisen nomenklatuurin eli uusia alleeleja voidaan lisätä skeemaan ilman vaikutusta aiempiin analyysituloksiin. AgMLST:n tietokanta pohjautuu CGE:n virulenssigeenitietokantaan, ja toistaiseksi EURL kuratoi uusien alleelien lisäystä tietokantaan manuaalisesti. AgMLST hyväksyy syötteeksi koontisekvenssejä, jolloin tietokantahaku perustuu blastn-algoritmiin. Jos syötteenä käytetään lukusekvenssejä, ne rinnastetaan kaikkien tietokannassa olevien alleelien kanssa. Eri alleeliksi määritellään yhden SNP:n ero, indelejä hakualgoritmi ei huomioi. Mikäli alleelit ovat eripituisia, huomioidaan pisin alleeli. Harjoitustyössä käytettiin koontisekvenssejä, jotka piti syöttää työkaluun kanta kerrallaan ja sitten yhdistää tulokseksi saadut kantojen alleelitaulukot Concatenate agMLST Profiles -työkalun avulla. Alleeliprofiilit klusteroitiin käyttäen TermoType Clustering -työkalua, jota käytettiin myös HReVAP-analyysissä. Menetelmä vaikutti siis hieman kotikutoiselta. EURL myönsi, että tällä hetkellä agMLST käyttää yksinkertaisia luokittelualgoritmeja, joita on tarkoitus kehittää tulevaisuudessa. Perinteistä seitsemään geeniin pohjautuvaa MLST-protokollaa kokeiltiin myös kurssilla ja sitä varten ARIES:issa on työkalu SRST2.