Folkhälsomyndigheten
PUBLIC HEALTH AGENCY OF SWEDEN

Bioinformatic analysis of EHEC WGS data

Ion Torrent NGS workflow for bacteria

DNA extraction from cultured bacteria

Automated prep of 13 libraries using Library Builder

Size selection using Pippin prep

ePCR, enrichment and chip loading using Ion Chef

Sequencing using Ion Torrent PGM

Up to 39 samples can be analyzed in 3 days

EHEC – bioinformatic analysis

~100 Mbp data per sample, 20x coverage

Gene content and subtyping analysis, two approaches:

- De-novo analysis (VirulenceFinder 1.2)
- Reference mapping (CLC Genomics Workbench 7.5)

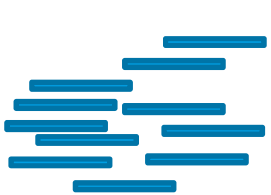
SNP-analysis:

- Core genome SNPs

Approaches we have not yet used:

- cgMLST

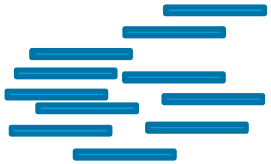
SNP analysis, method



Reads, all samples

assembly

Contigs, pan genome

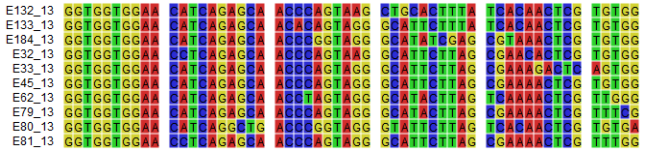


Reads, each sample

mapping

Contigs, with coverage depth

Extract SNPs,
coverage >10x in all samples

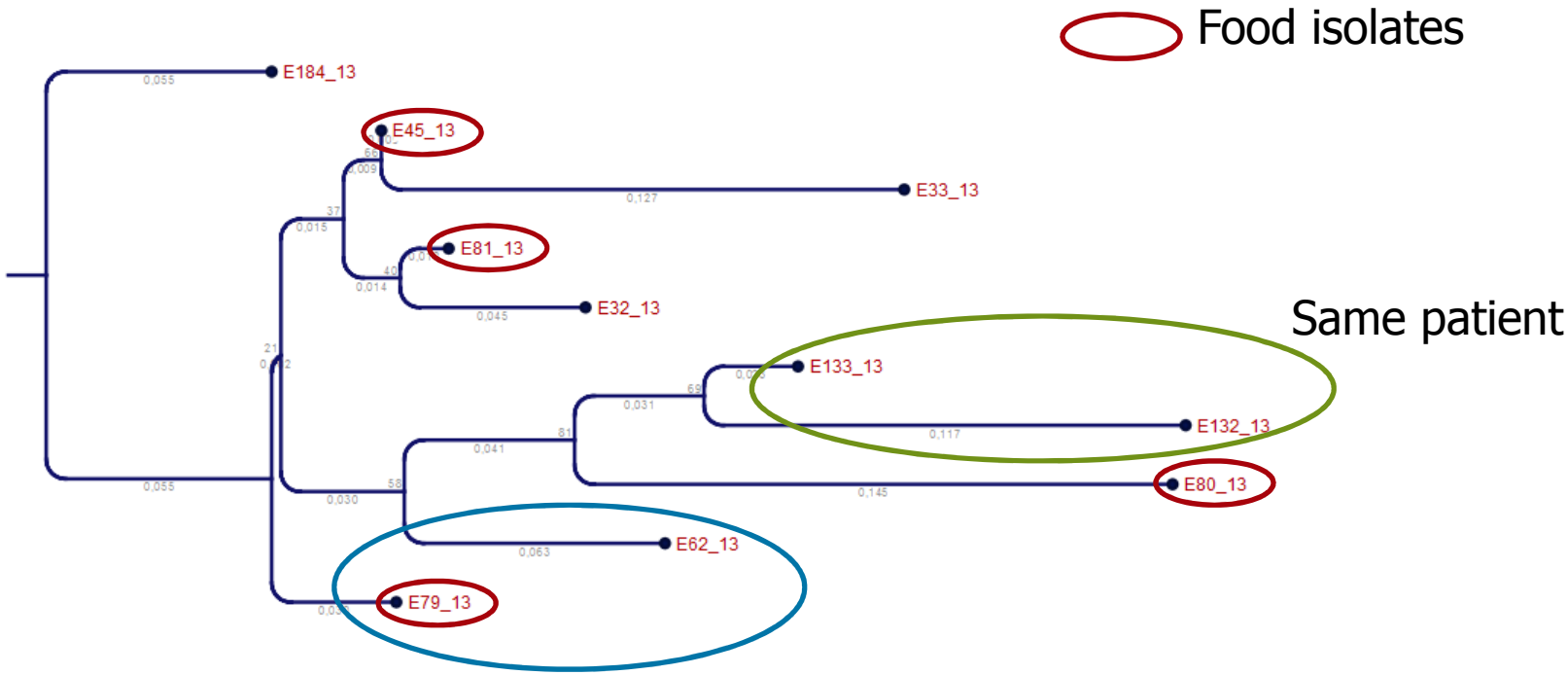


Concatenated SNPs alignment

Core genome SNPs

- A pan genome is built using all the data in the dataset
- Parts of the pan genome in which all samples do not have coverage over a threshold (we used 10x) are excluded. The remainder forms the core genome.
- The reads for each sample are mapped to the core genome and SNPs called. Each site with variability within the dataset is put into a concatenated alignment.
- The alignment is analyzed using phylogenetics (CLC genomics workbench)

SNP analysis results, small EHEC outbreak



SNP analysis, discussion

High resolution outbreak surveillance

Dependent on a core genome with high similarity to the outbreak strain

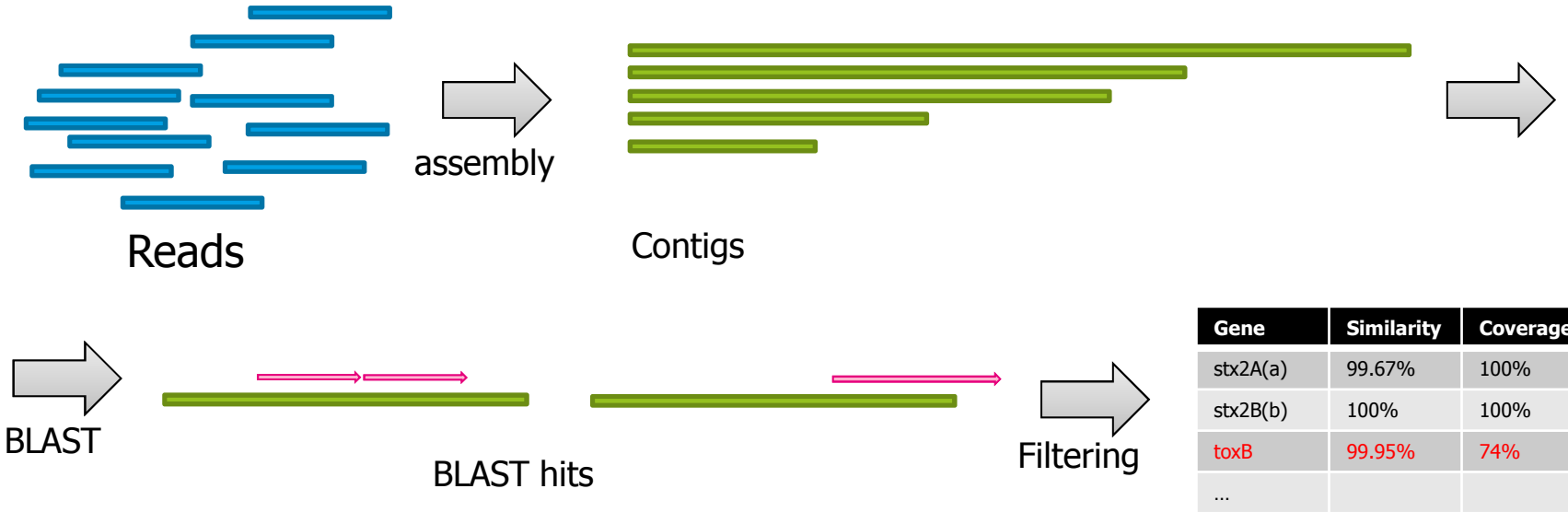
Quite computationally intensive

May underestimate the number of SNPs in low coverage regions of the genome (for example plasmids)

May overestimate the number of SNPs in the case of recombination events

- Can be fixed using some algorithmic changes

Gene content analysis, de-novo based



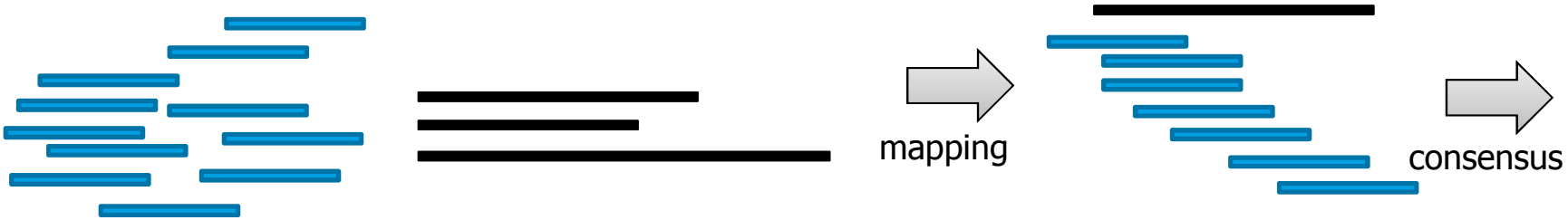
Gene	Similarity	Coverage
stx2A(a)	99.67%	100%
stx2B(b)	100%	100%
toxB	99.95%	74%
…		

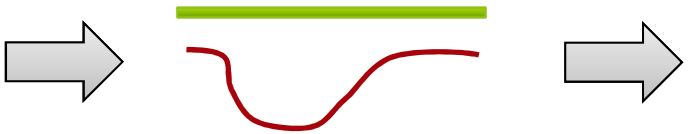
De-novo assembly is performed

BLAST search for a library of reference genes in the genome

Cutoffs for gene coverage and similarity is used to filter the hits

Gene content analysis, reference mapping based



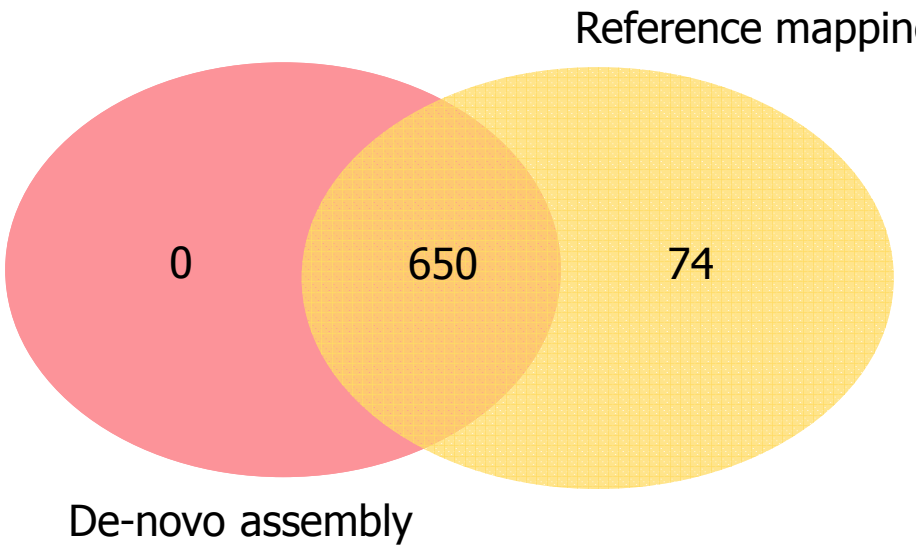


Gene	Similarity	Coverage	Depth
stx2A(a)	99.67%	100%	30x
stx2B(b)	100%	100%	15x
toxB	99.95%	100%	5x
…			

The raw data is mapped to a reference set of target genes

Cutoffs for gene coverage and similarity is used to filter the hits

Gene content analysis, results (40 samples)



Gene	Description
astA	EAST-1 heat-stable toxin
cba	Colicin B
cma	Colicin M
eae	Intimin
ehxA	Enterohaemolysin
espA	Type III secretion system
espB	Secreted protein B
espJ	Prophage-encoded type III secretion system effector
espP	Extracellular serine protease plasmid-encoded
etpD	Type II secretion protein
gad	Glutamate decarboxylase
iha	Adherence protein
iss	Increased serum survival
katP	Plasmid-encoded catalase peroxidase
nleA	Non-LEE encoded effector A
nleB	Non-LEE encoded effector B
nleC	Non-LEE encoded effector C
prfB	P-regulated fimbriae regulatory gene
toxB	Toxin B

- The genes with the most misses in the de-novo analysis were toxB (15/40 found) and iss (27/40 found)
- toxB generally had low coverage depth (5-10x) while iss had high coverage depth (40-50x)
- The average coverage depth in the chromosome was ~20x

Gene content analysis, discussion

De-novo method

Allows for data compression by storing the data as draft genomes

After the initial de-novo assembly, analysis is fast

Cannot handle multiple copies of *stx1* or *stx2* genes with different subtypes[1]

Reference method

Much less sensitive to lower data quality and lower depth of coverage

Not affected by de-novo assembler misassemblies and contig breaks

1) Ashton et al 2014.
Insight into Shiga toxin genes encoded by *Escherichia coli* O157 from whole genome sequencing

Gene content analysis, detailed results for one sample

E1003_97

Reference sequence	Average coverage	Rel coverage	Length	Cov length	Cov %	VirFind
astA	7,41	0,331840573	117	115	98,29%	1
cba	0	0	1536	0	0,00%	0
cma	0	0	816	0	0,00%	0
eae	26,4	1,18226601	2805	2808	100,11%	1
ehxA	9,53	0,426780116	2997	3000	100,10%	1
espA	24,14	1,081056874	579	580	100,17%	1
espB	22,33	1	939	939	100,00%	1
espJ	24,31	1,088669951	654	657	100,46%	0
espP	12,52	0,560680699	3903	3910	100,18%	1
etpD	7,49	0,335423197	1758	1767	100,51%	1
gad	27,18	1,217196597	1401	1402	100,07%	1
iha	25,53	1,143304971	2091	2091	100,00%	1
iss	41,62	1,863860278	294	295	100,34%	0
katP	8,58	0,384236453	2211	2216	100,23%	1
nleA	25,3	1,133004926	1326	1326	100,00%	1
nleB	22,32	0,999552172	981	982	100,10%	1
nleB	25,65	1,148678907	990	992	100,20%	1
nleC	29,03	1,300044783	993	993	100,00%	1
prfB	21,76	0,974473802	882	882	100,00%	1
stx2A	50,38	2,256157635	960	960	100,00%	1
stx2B	47,13	2,110613524	270	270	100,00%	1
tir	18,56	0,831168831	1677	1679	100,12%	1
toxB	8,68	0,388714734	9510	9560	100,53%	0
Median	22,33					