STATENS SERUM INSTITUT

SNP ANALYSIS FROM NGS DATA AT SSI

Kristoffer Kiil, Ph.D
Statens Serum Institut
Denmark

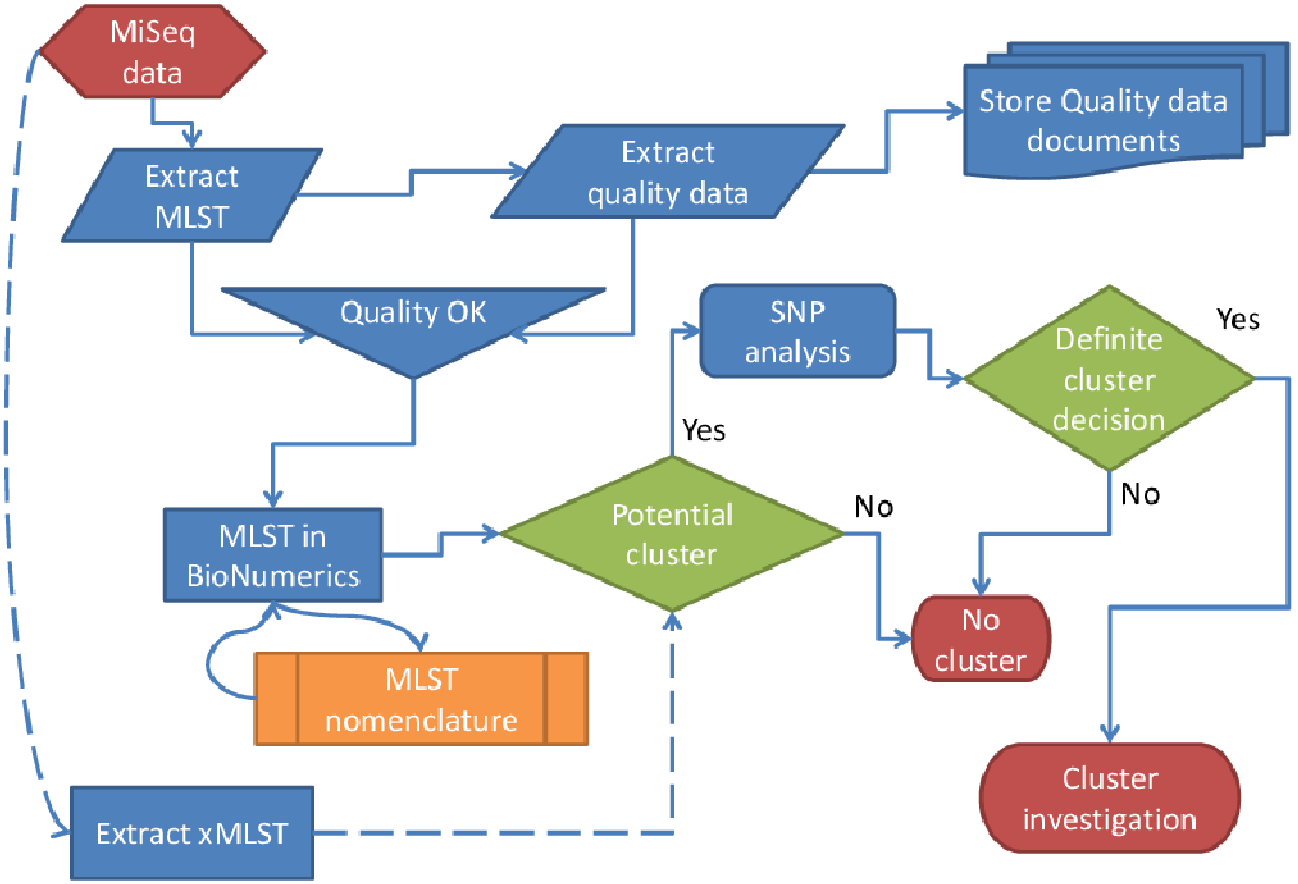
OUTLINE

STATENS SERUM INSTITUT

- SSI SNP cluster detection
  - MLST
  - SNP analysis
- xtMLST
  - why we consider it
  - The listeria case
  - Is it a good idea?
- Can we share this data across borders?

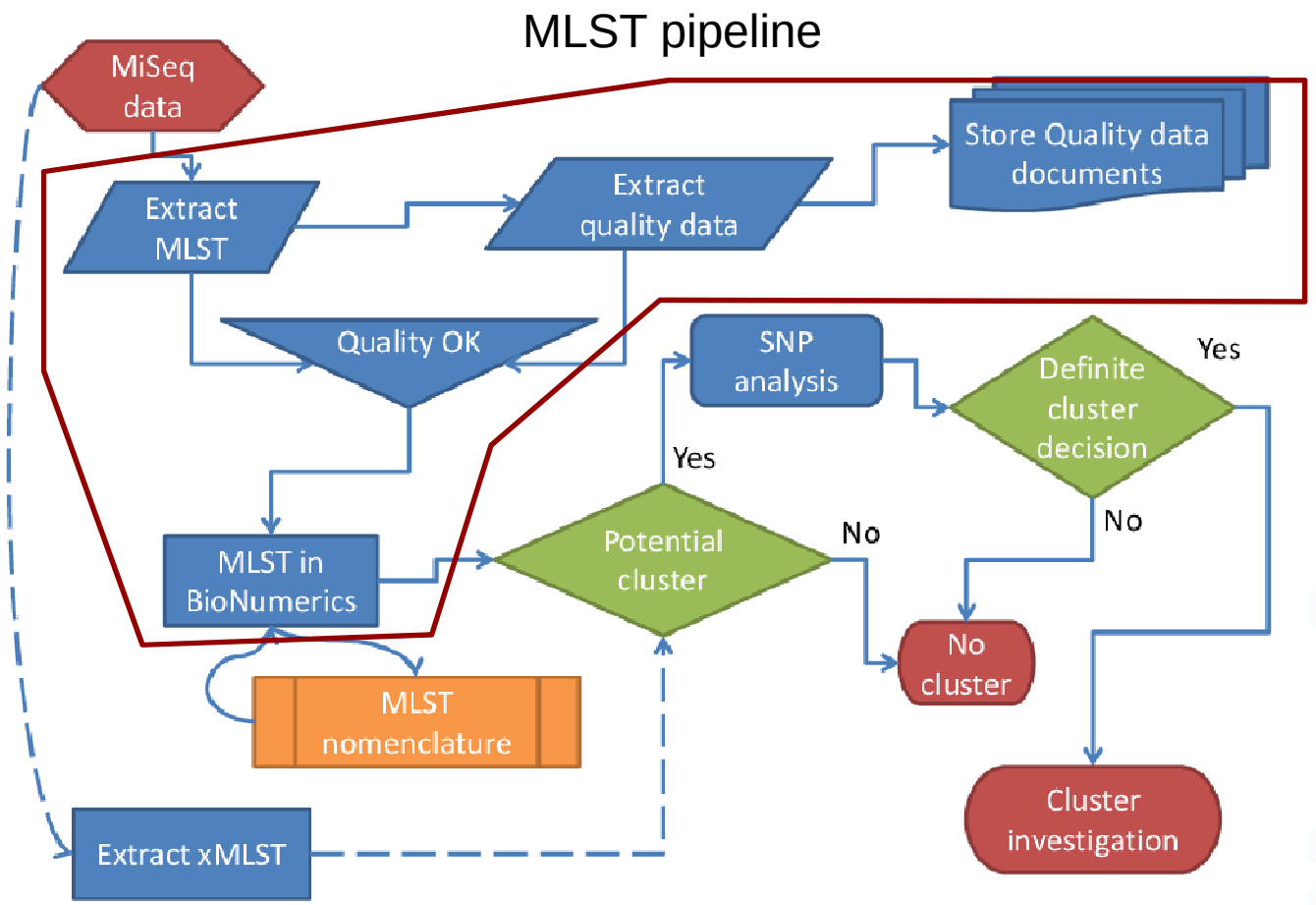
SSI SNP CLUSTER DETECTION

STATENS SERUM INSTITUT



SSI SNP CLUSTER DETECTION

STATENS SERUM INSTITUT



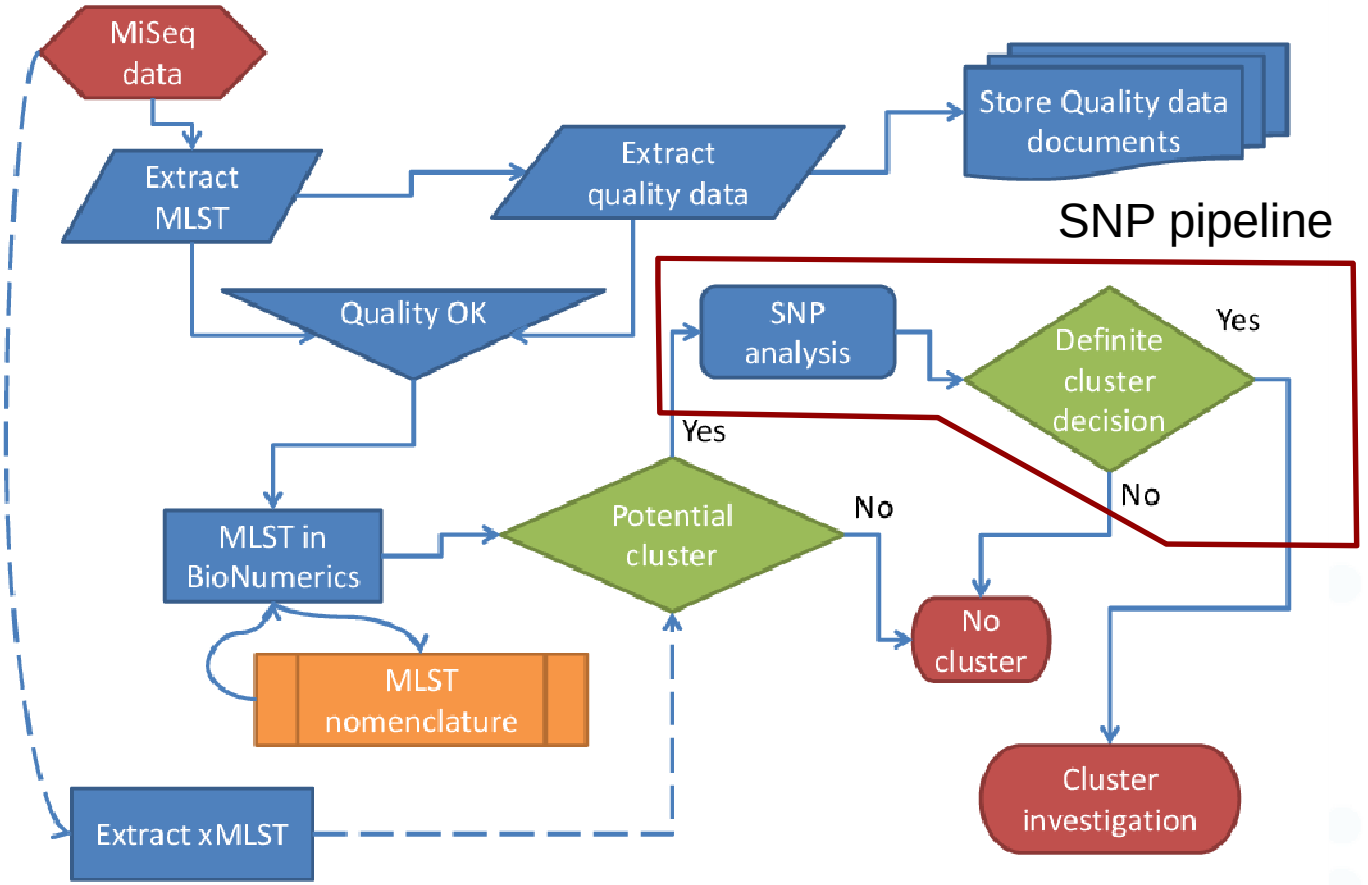
MLST PIPELINE

STATENS SERUM INSTITUT

- Pad MLST allele sequences with 2x50bp from a reference genome
- Align reads to the padded sequences using BWA mem
- Samtools and modified samutils script used to call alleles
- Allele sequences imported into bionumerics to integrate MLST data with other strain data
- Could be replaced with SRST2 (https://github.com/katholt/srst2)

SSI SNP CLUSTER DETECTION

STATENS SERUM INSTITUT



SNP PIPELINE

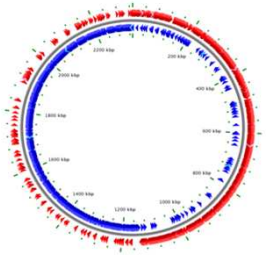
STATENS SERUM INSTITUT

- Align reads to ST specific reference using BWA mem
- Sort and process read alignments using elprep
- Call SNPs against the reference using GATK
- Compare SNP lists from all samples in comparison and create a brutto SNP list
- Force recalling of all SNPs from the brutto list
- Discard low quality SNPs
  - Minimum depth 4
  - Uses maximum likelihood call in diploid model
  - Discards ambiguous SNP calls
- Remove recombination
  - Removing unlikely long stretches of identical SNP profiles

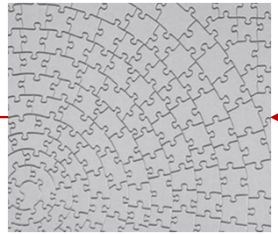
SNP PIPELINE

STATENS SERUM INSTITUT

Filter Reads
<20% nonATGC
<10% Quality<20

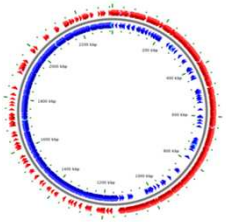


ST specific reference genome

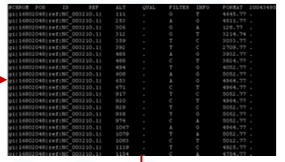


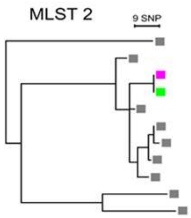
Reference assembly

New reference genome



SNP list

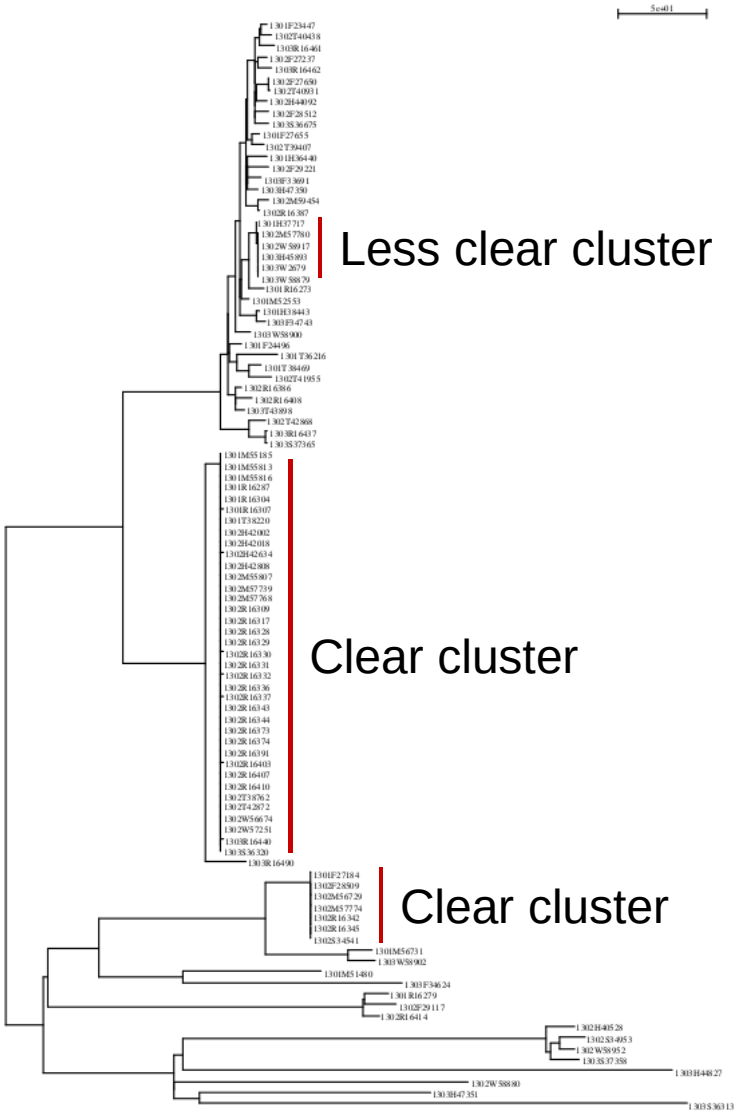




Compare to other SNP lists within ST

EXAMPLE: *S. typhimurium*

STATENS SERUM INSTITUT



- Good resolution
- Fairly clear cluster definitions
- Somewhat computationally intensive
- A few STs comprise most of the strains

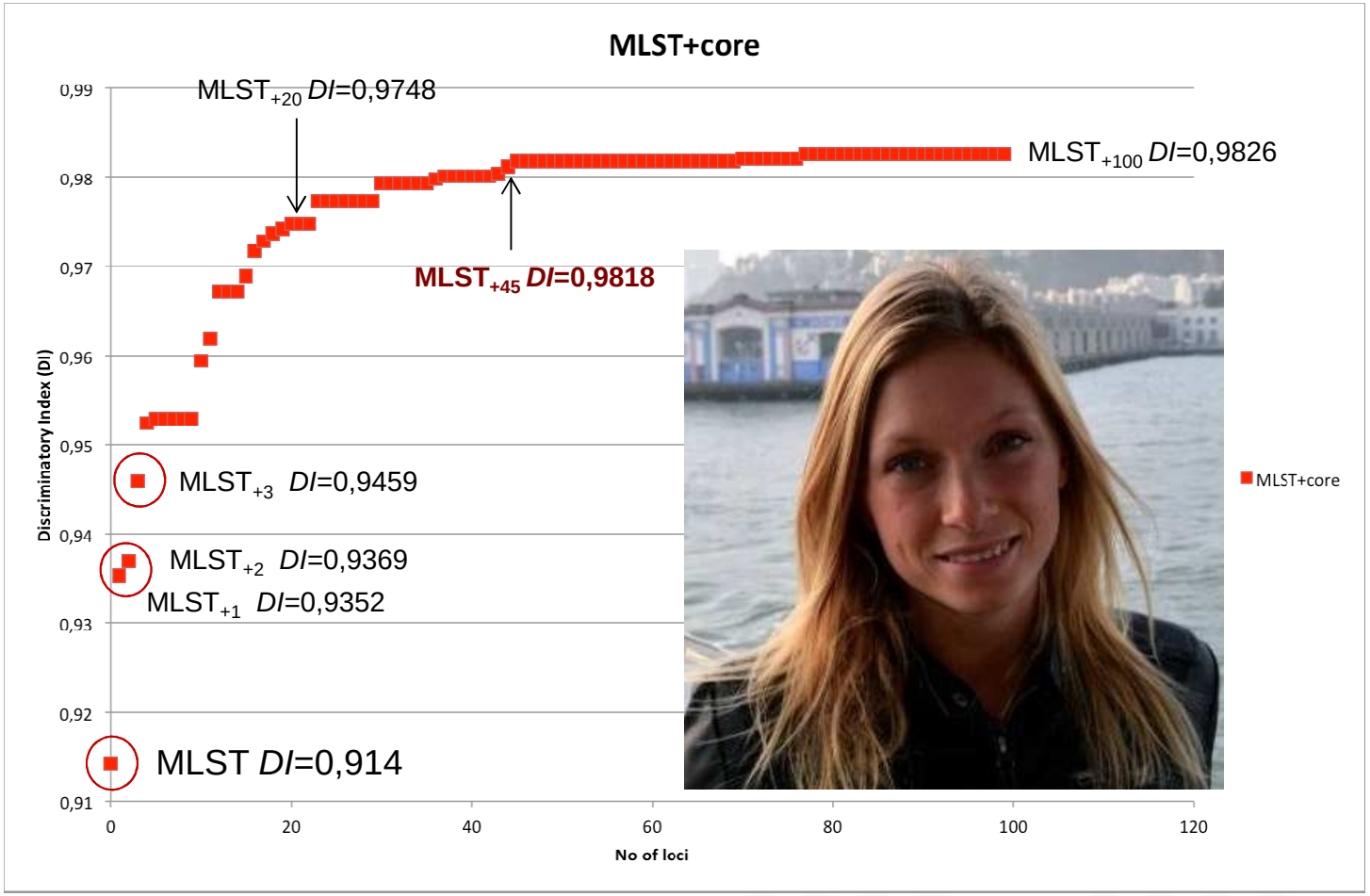
xtMLST

STATENS SERUM INSTITUT

- Could we improve the performance of our workflow
  - Get better discrimination
  - Higher DI = fewer cluster investigations
  - Extend with more loci

MLST+100 CORE GENES

STATENS SERUM INSTITUT



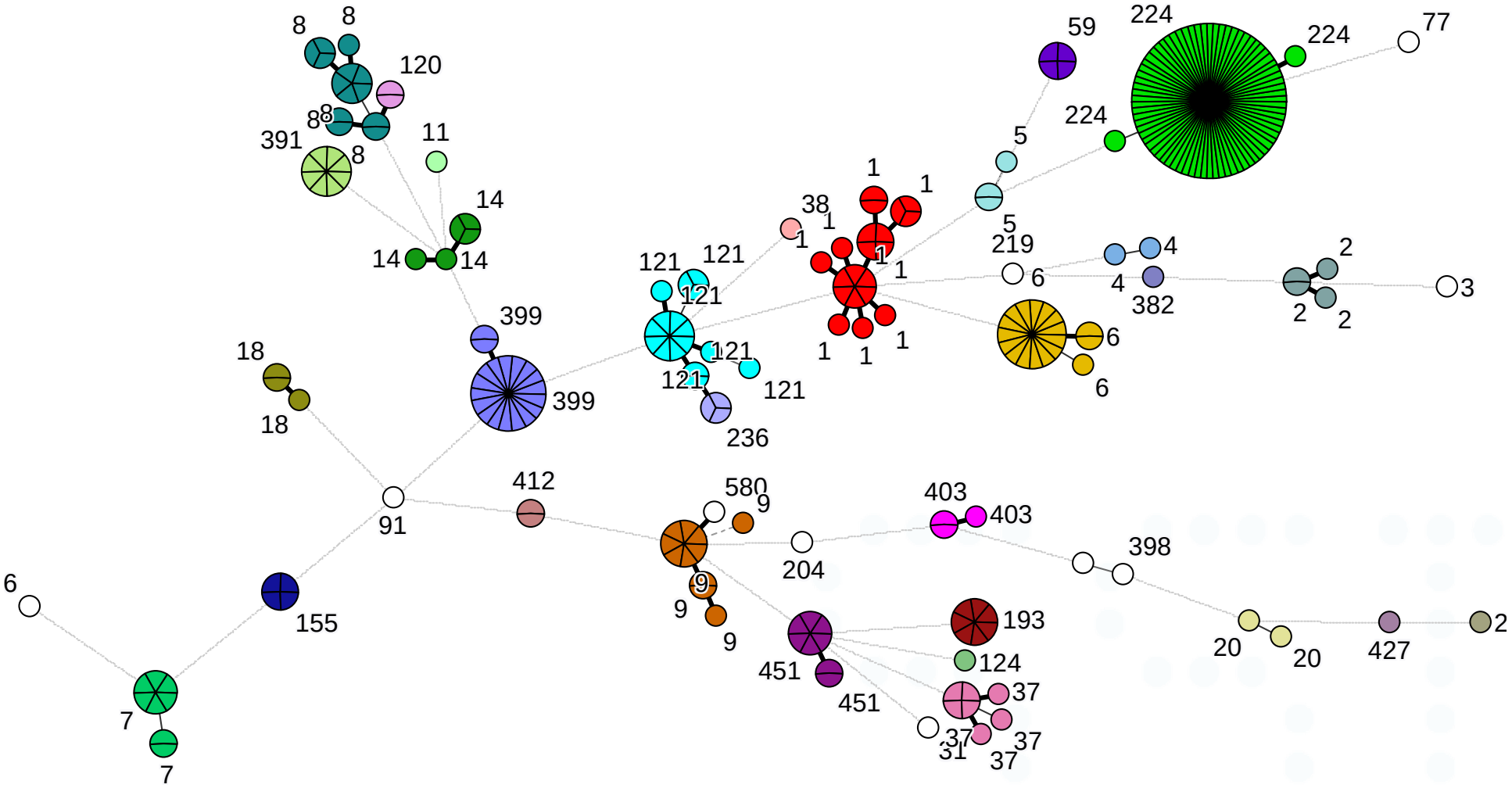
xtMLST

STATENS SERUM INSTITUT

- Could we improve the performance of our workflow
  - Get better discrimination
  - Higher DI = fewer cluster investigations
  - Extend with more loci
- Relatively few additional loci increase DI dramatically
- How does it look when we look at independent surveillance data?

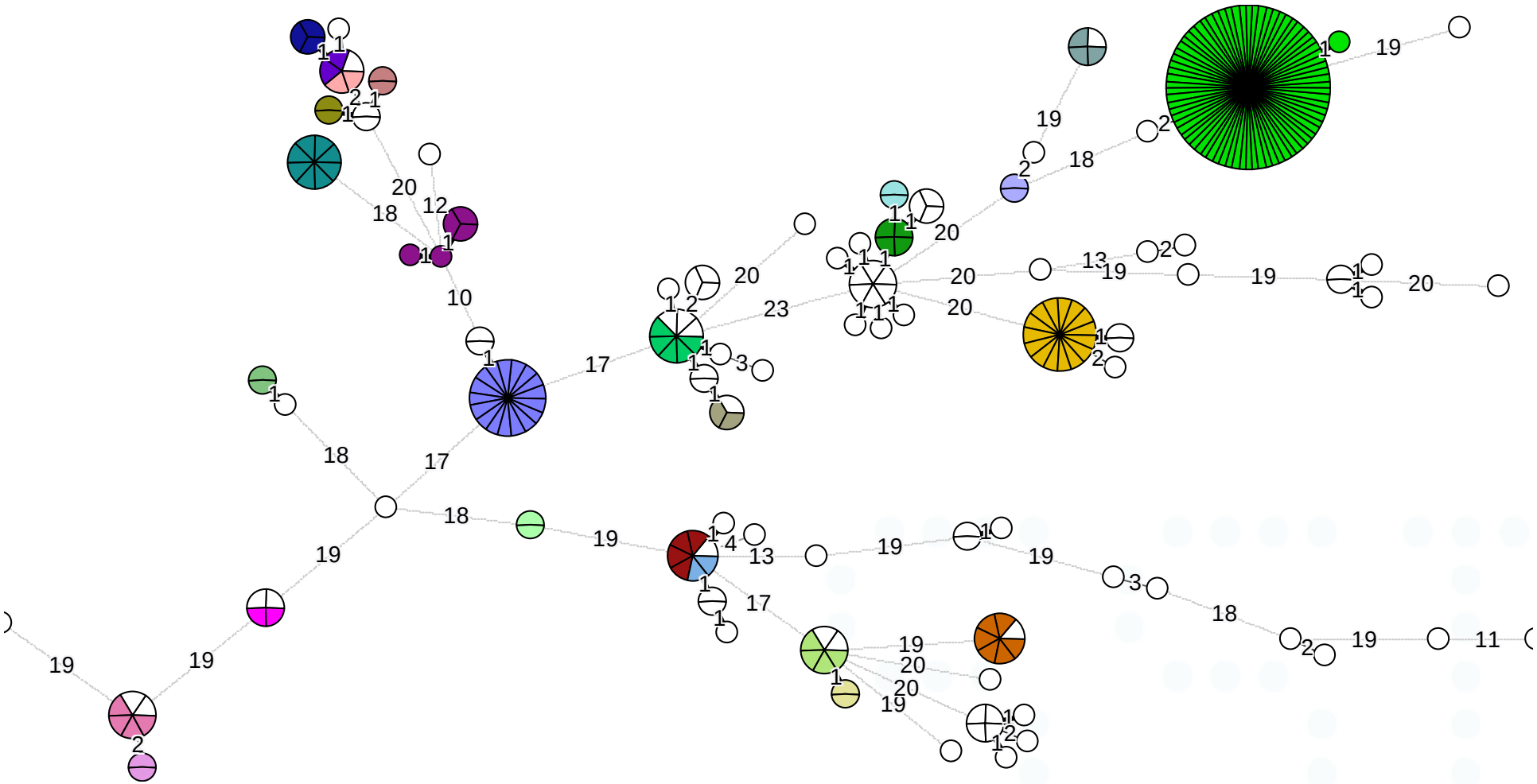
MLST ST

STATENS SERUM INSTITUT



SNP cluster colour

STATENS SERUM INSTITUT



xtMLST

STATENS SERUM INSTITUT

- Decreases the number of SNP cluster investigations
- Results in few false negative clusters
- Developing xtMLST schemes requires some extra work on top of MLST

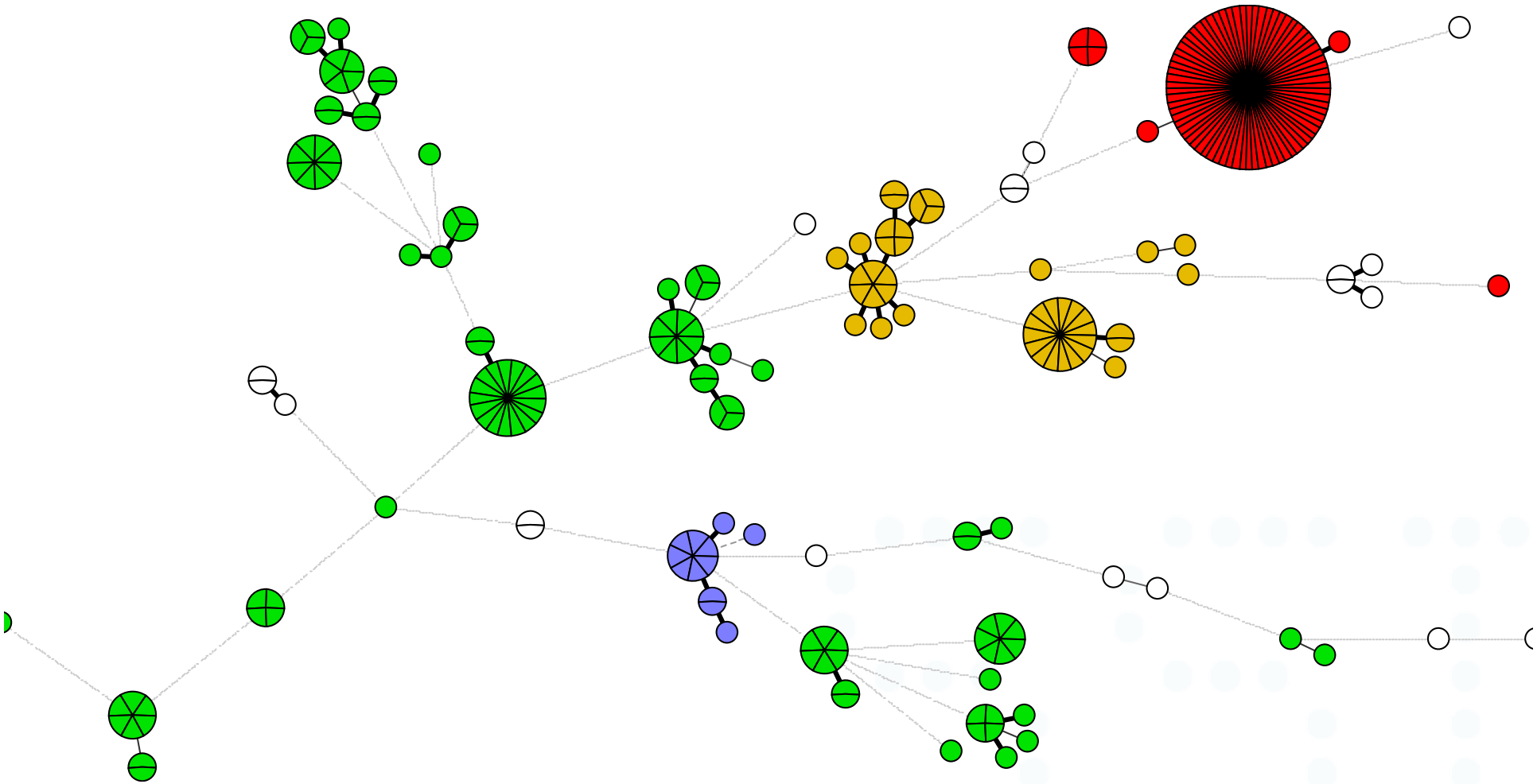
CROSS BORDER SNP EXCHANGE

STATENS SERUM INSTITUT

- Challenges
  - SNP positions must be comparable
  - Data must not take too much space (comparison by email)
  - Distinguishing between no SNP and uncallable SNP
- Solutions
  - Make a shared library of ST specific references
  - Record low quality regions as a bit array, same length as the reference (<1Mb)

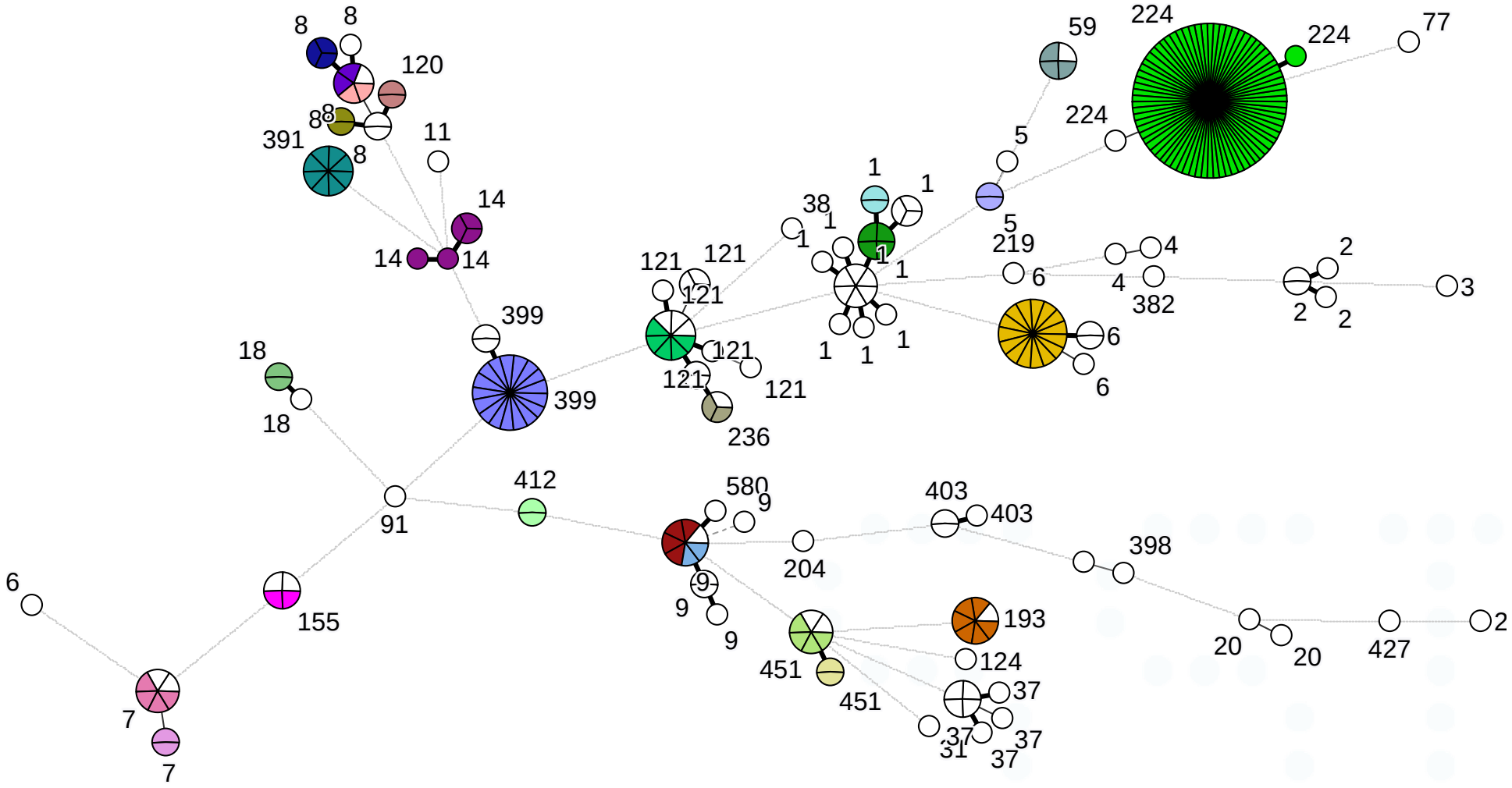
Molecular serotype colour (n=264)

STATENS SERUM INSTITUT



SNP cluster colour MLST ST labels

STATENS SERUM INSTITUT



MLST w/ distances

STATENS SERUM INSTITUT