In silico identification of *Streptococcus suis* serotype

J. Matiašovic[1], N. Králová[1,2], M. Zouharová[1]
[1] Department of Infectious Diseases and Preventive Medicine, Veterinary Research Institute, Brno, Czech Republic
[2] Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic

Introduction

Serotypes of *Streptococcus suis* are associated with various pathogenicity. The serotype is determinated by the cps locus harboring genes encoding enzymes involved in the synthesis of capsular polysaccharide. Identification of genes specific for particular serotype by PCR is already used for PCR-based serotyping. Here we present a script detecting serotype-specific genes in assembled genome and thus identifying putative serotype of sequenced isolate.

Materials and Methods

The script utilise local installation of NCBI-blast+ software and a database of serotype-specific genes. The database of serotype-specific sequences is modified from the SsuisSerotyping_pipeline (https://github.com/streplab/SsuisSerotyping_pipeline) by adding the genes specific for the new capsular polysaccharide serotypes (NCLs) up to NCL32. After the blast search of sequences similar to serotype-specific genes present in studied genome, the result of the blast search is filtered by score and the database gene most similar to sequence present in genome determine the probable serotype. The script also identify the presence of potential modifying genes altering the final polysaccharide structure and also the presence of recN, the gene predominantly present in *Streptococcus suis sensu stricto* isolates.

```
#!/bin/bash
#run from folder containing Streptococcus suis genomes in multiple fasta format with .fasta suffix
#read name of *.fasta file and blast it against Ssuis_SerotypingX.fasta database. Presence of recN gene will be tested and serotype will be determined.
for x in *.fasta
  do blastn -query "$x" -db /$HOME/genomes/test/db/Ssuis_Serotyping8-NK-V2v3.fasta -task megablast -dust no -outfmt "6 sseqid evalue score ppos" -max_target_seqs 5 -num_threads 10 | sort -k3 -n -r > "$x".blasttest
#take lines from blast results containing recN result and take only first line
grep 'recN' "$x".blasttest | head -1 > "$x".blastrecN
#prepare blastrecN result for final table
cut -f 2,3,4 "$x".blastrecN > "$x".blastrecN2
#take the first line from blast results containing all non recN, non cpsK and non mdresults sorted by score
grep -v 'recN' "$x".blasttest | grep -v '^cpsK.' | grep -v '^md' | sort -k3 -n -r | head -1 > "$x".blastcps
#take the first line from blast results containing modifying genes (md) sorted by score
grep 'md' "$x".blasttest | sort -k3 -n -r | sed -n '1p' > "$x".blastcps2
#take first line from blast results containing any cpsK results sorted by score
grep '^cpsK.' "$x".blasttest | sort -k3 -n -r | head -1 > "$x".blastcpsK
#test if the "$x".blastcps2"" is empty; case it is, add 0
if [ -s ""$x".blastcps2" ]; then
    echo
else
    echo "md not detected   .         .         ." > ""$x".blastcps2"
fi
#make head named after the actual fasta file
echo "$x" | sed 's/.fasta//' > shorthead
#from "$x".blastrecN result take first column, will read presence of recN
rcN=`sed -n 1p "$x".blastrecN| cut -d$'\t' -f 1`
#from "$x".blastcps and "$x".blastcpsK take first column, will read resolution of serotype, 2vs1/2 and 1vs14
cps=`sed -n 1p "$x".blastcps | cut -d$'\t' -f 1`
cpsK=`sed -n 1p "$x".blastcpsK | cut -d$'\t' -f 1`
#cps2=`sed -n 1p "$x".blastcps2 | cut -d$'\t' -f 1`
echo "$x"
echo "$rcN"
echo "$cps"
echo "$cps2"
echo "$cpsK"
echo
case "$rcN" in
  recN)
  recNpresence="recN present"
  ;;
  *)
  recNpresence="recN not detected      .        .        ."
esac
case "$cps" in
 cps-1)
   case "$cpsK" in
   cpsK2cpsK14)
    vys="14"
    ;;
    cpsK1cpsK1/2-T)
     vys="1"
    ;;
    cpsK1cpsK1/2-C)
     vys="1"
    ;;
    *)
     vys="serotype not determined         .        .        .        ."
    ;;
    esac
    ;;
  cps-2)
   case "$cpsK" in
   cpsK2cpsK14)
    vys="2"
    ;;
    cpsK1cpsK1/2-T)
     vys="1-2"
    ;;
    cpsK1cpsK1/2-C)
     vys="1-2"
    ;;
    *)
     vys="serotype not determined         .        .        .        ."
    ;;
  esac
;;
cps-3)
vys="3"
;;
cps-4)
vys="4"
;;
cps-5)
vys="5"
;;
cps-6)
vys="6"
;;
cps-7)
vys="7"
;;
cps-8)
vys="8"
;;
cps-9)
vys="9"
;;
cps-10)
vys="10"
;;
cps-11)
vys="11"
;;
cps-12)
vys="12"
;;
cps-13)
vys="13"
;;
cps-15)
vys="15"
;;
cps-16)
vys="16"
;;
cps-17)
vys="17"
;;
cps-18)
vys="18"
;;
cps-19)
vys="19"
;;
cps-20_S.parasuis)
vys="20_S.parasuis"
;;
cps-21)
vys="21"
;;
cps-22_S.parasuis)
  vys="22_S.parasuis"
  ;;
  cps-23)
  vys="23"
  ;;
  cps-24)
  vys="24"
  ;;
  cps-25)
  vys="25"
  ;;
  cps-26_S.parasuis)
  vys="26_S.parasuis"
  ;;
  .
  .
  ;;
  NK242-NCL32)
  vys="NK242-NCL32"
  ;;
  *)
  vys="serotype not determined   .         .         .         ."
  ;;
esac
#make final table
#test if the "$x".blastcpsK is empty; case it is, add three dosts separated by tab
if [ -s ""$x".blastcpsK" ]; then
    echo
else
    echo ".      .         .         ." > ""$x".blastcpsK"
fi
echo "$recNpresence" > "$x".recNpresence
echo "$vys" > "$x".vysl
cat shorthead "$x".vysl "$x".blastcps "$x".blastcps2 "$x".blastcpsK "$x".recNpresence "$x".blastrecN2 > "$x".vysle
#remove line wrapping and tab column separation
sed ':a;N;$!ba;s/\n/\t/g' "$x".vysle > "$x".merged
done
#make head of result file serotypes.tsv
echo "sample        serotype       cps_gene     cps_evalue   cps_score     cps_ppos      modifying_gene      md_evalue    md_score      md_ppos      cpsk    cpsk_evalue cpsk_score   cpsk_ppos     recN    recN_evalue recN_score   recN_ppos" > hlavicka
cat hlavicka *.merged > serotypesV8.NK-V2v4.tsv
#remove temporary files
rm *.blasttest
rm *.blastcps
rm *.blastcps2
rm *.blastrecN
rm *.blastrecN2
rm *.blastcpsK
rm *.recNpresence
rm shorthead
rm *.merged
rm *.vysl
rm *.vysle
rm hlavicka
exit 0
```

Results

The script successfully identify serotype-specific sequence in genome of all reference serotype strains and thus infer serotype. The script also distinguish the serotypes ½ vs. 2 and 1 vs 14. Moreover, the NCL serotypes up to NCL32 can be detected.

Example of output:

sample	serotype	cps_gene	cps_evalue	cps_score	cps_ppos	modifying_gene	md_evalue	md_score	md_ppos	cpsk	cpsk_evalue	cpsk_score	cpsk_ppos	recN	recN_evalue	recN_score	recN_ppos
A1	1-2	cps-2	3.65e-76	150	99.35	md not detected	.	.	.	cpsK1cpsK1/2-T	1.73e-59	120	100.00	recN present	8.00e-159	300	96.43
A3	2	cps-2	3.68e-76	150	99.35	md not detected	.	.	.	cpsK2cpsK14	1.75e-59	120	100.00	recN present	6.30e-179	336	100.00
74	1	cps-1	2.29e-77	153	100.00	md not detected	.	.	.	cpsK1cpsK1/2-T	5.07e-59	120	100.00	recN present	5.68e-166	312	97.62
B3	14	cps-1	1.71e-77	153	100.00	md not detected	.	.	.	cpsK2cpsK14	3.78e-59	120	100.00	recN present	3.73e-179	336	100.00
A4	3	cps-3	1.12e-77	153	100.00	md not detected	.	.	.	.	.	.	.	recN present	5.06e-177	333	99.70
192	4	cps-4	6.79e-75	150	99.35	md not detected	.	.	.	.	.	.	.	recN present	3.15e-167	315	97.92
B5	16	cps-16	1.27e-77	153	100.00	md-NK219-NCL25-2	5.24e-67	135	96.08					recN present	1.43e-160	303	96.73
134	13	cps-13	1.14e-73	147	98.69	md not detected	.	.	.	.	.	.	.	recN not detected	.	.	.
E9	serotype not determined	.	.	.	.	md not detected	.	.	.	.	.	.	.	recN not detected	.	.	.

evalue = Expectation Value. Lower E-values are better: An E-value close to zero (e.g., 1e-100) means that the match is highly significant and very unlikely to have occurred by random chance.
score = Raw Score: This is the fundamental score calculated directly from the alignment. It is determined by the scoring system used for matches, mismatches, and gaps. For example, a match might get a positive score, while a mismatch and a gap opening would receive negative scores (penalties). The raw score is the sum of these values for the entire alignment.
ppos = Positive matches. Indicates the percentage identity of the alignment. For example, if "ppos" is 95%, it means 95% of the aligned bases are identical between the query and the database hit.

Discussion

Whole-genome sequencing is becoming an affordable solution for identifying bacterial isolates. The presented script allows to estimate the potential serotype of *Streptococcus suis* isolate. Due to its dependence only on the freely available software ncbi-blast+, its installation is very simple.

Availability

The script and database are freely available on GitHub at https://github.com/matiajan/seroblast

Reference

Králová N, Fittipaldi N, Zouharová M, Nedbalcová K, Matiašková K, Gebauer J, Kulich P, Šimek B, Matiašovic J. Streptococcus suis strains with novel and previously undescribed capsular loci circulate in Europe. Vet Microbiol. 2024 Sep 26; 298:110265. doi: 10.1016/j.vetmic.2024.110265

Acknowledgements

The work was supported by the Ministry of Agriculture of the Czech Republic (projects RO0523 and TN02000017).