Align Contigs to Scaffolds

Noëlle Schenk

July 25, 2018

Note : this script requires `translate_markers.Rmd` to run (in order to create the data.table `duplicate_suspects`).

The mapping of marker sequences to the newly scaffolded *P.exserta* genome (from Bionano) indicated a possibility for duplicate Contigs within the genome. In 135 cases, the marker seqence (201 bp) mapped equally well to a Super-Scaffold and a Contig from the "not scaffolded" file.

To look into this, an alignment of the candidate duplicate sequences is done.

```
duplicate_suspects <- readRDS("duplicate_suspects.RDS")
ds <- matrix(data = NA, nrow = 0, ncol = 3)
for(i in unique(duplicate_suspects$query_id)){
  ds <- rbind(ds, c(i, t(duplicate_suspects[which(duplicate_suspects$query_id == i), "subject_id"])))
}
ds <- as.data.frame(ds)
write.table(ds, "dupctgs/input.txt", row.names=F, col.names = F, sep = ' ',quote=F)
```

ncbi-blast

1. Extract both sequences from the blast database by entry name and save them in separate files
2. align sequences with minimap2 and save output file

Example with 1 sequence:

```
blastdbcmd -db P.exserta.opticalmap.v1.fasta -entry Peex113Ctg02834_obj -out Peex.fasta
blastdbcmd -db P.exserta.opticalmap.v1.fasta -entry Super-Scaffold_14460 -out SuperSc.fasta
~/minimap2/minimap2 -x ava-pb SuperSc.fasta Peex.fasta > dupcontgs/ovlp.paf
```

Script for all 135 sequences:

use on command line as `./scriptname.sh inputfile.txt` where `scriptname.sh` is the name of the script below, and `input.txt` is the file generated above. The script extracts sequences from the database (reference genome), aligns them and generates a plot to visualize the result. Visualization of the .paf files was performed with miniasm.

```
#!/bin/bash
while IFS='' read -r line ; do
    w1=$(echo $line | cut -f1 -d' ')
    w2=$(echo $line | cut -f2 -d' ')
    w3=$(echo $line | cut -f3 -d' ')
    echo "bla" $w1 "bli" $w2 "blu" $w3
    blastdbcmd -db /home/exserta/Documents/master_project_noelle/data/optical_mapping_raw/P.exserta.opti
    blastdbcmd -db /home/exserta/Documents/master_project_noelle/data/optical_mapping_raw/P.exserta.opti
    echo "now minimap comes"
    ~/minimap2/minimap2 -x ava-pb $w3.fasta $w2.fasta > $w1.ovlp.paf
    ~/miniasm/minidot $w1.ovlp.paf > $w1.eps
    echo "done?"
done < "$1"
rm *.fasta
rm *.paf
```

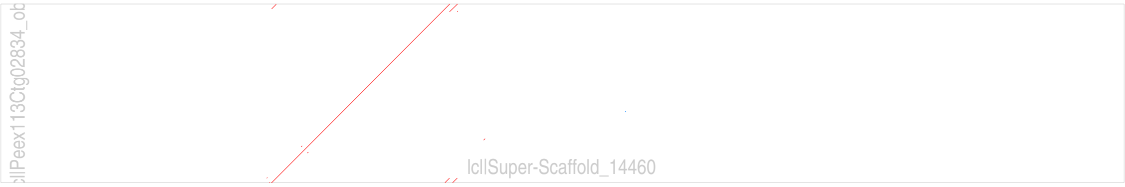

Figure 1: long alignment with overlap

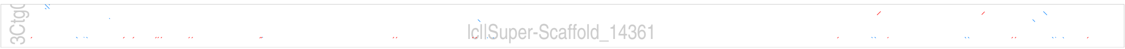

Figure 2: no duplicate

Visual inspection

The results of the visual inspection are documented in the table `visual_inspection.csv`, but incomplete as automatic inspection was performed.

Inspection

In the `.paf` file, the alignments are given by length. In order to be very restrictive in deleting, only the longest alignment is considered and an identity of 99% is required for a contig to be deleted. In any case, the list of putative duplicates can be reconsidered at any time.

Details: for a contig to be deleted, the following requirements need to be met: * The alignment length is at least 99% of the query length (=contig length) * The number of matches in the alignment is at least 99% of the alignment length (only 1% mismatches allowed)

```
# to output the contigs into a file
for i in *.paf; do awk -F ' ' 'NR==1{print $1 " " $11/$2 " " $10/$11}' $i | awk '{if ($2 > 0.99 && $3 >
# to count the contigs for given tresholds
for i in *.paf; do awk -F ' ' 'NR==1{print $1 " " $11/$2 " " $10/$11}' $i | awk '{if ($2 > 0.99 && $3 >
```

tres len	tres qual	n° ctgs
0.99	0.99	5
0.99	0.98	11
0.98	0.99	5
0.98	0.98	11
0.97	0.97	15
0.96	0.96	20
0.95	0.95	23

The 12 Contigs which most probably are duplicates are the following :

contig	super-scaffold
Peex113Ctg02834_obj	Super-Scaffold_14460
Peex113Ctg13959_obj	Super-Scaffold_5422
Peex113Ctg07317_obj	Super-Scaffold_2567
Peex113Ctg02434_obj	Super-Scaffold_14390
Peex113Ctg13182_obj	Super-Scaffold_1611
Peex113Ctg07641_obj	Super-Scaffold_14840
Peex113Ctg17963_subseq_1:69396_obj	Super-Scaffold_11267
Peex113Ctg01240_obj	Super-Scaffold_14928
Peex113Ctg03546_obj	Super-Scaffold_4410
Peex113Ctg05113_obj	Super-Scaffold_163

PAF fileformat

PAF is a text format describing the approximate mapping positions between two set of sequences. If PAF is generated from an alignment, column 10 equals the number of sequence matches, and column 11 equals the total number of sequence matches, mismatches, insertions and deletions in the alignment.