Mail Brigitte :

> *Bonjour Gaëtan,*
> *Nous sommes à la dernière phase de notre clonage positionnel d'un QTL sur le chromosome 9.*
> *En principe, on n'a plus que 10 gènes mais par sécurité, on va prendre un peu plus large.*
> *Soit les positions sur Nipponbare: Os09 : 18855737-19122396*
> *Ce qu'on voudrait, ce sont les séquences d'Azucena et d'IR64 correspondant à cette séquence chez Nipponbare. Avec un alignement et une annotation éventuellement.*
> *Je passe te voir*
> *Brigitte*

Objective: Find all genes on IR64 corresponding to the region identified on Nipponbare

Context :

Schatz et al (2014), New whole genome de novo assemblies of three divergent strains of rice (O. sativa) documents novel gene space of aus and indica.

http://biorxiv.org/content/biorxiv/early/2014/04/02/003764.full.pdf

Howto:

1. **Set the environnment**

Create a directory

```
mkdir bioperl
```

Create an alias

```
export FORMATION=$PWD/bioperl
ls $FORMATION
```

Move into this directory

```
cd $FORMATION
mkdir $FORMATION/bbmh
```

Create symbolic link with Nipponbare (MSU7) and IR64 fasta file

```
ln -s /bank/oryza_sativa_japonica/MSU7_pep_20111031
ln -s /home/droc/OryGenesDB/IR64/IR64_prot.faa
ln -s /home/droc/OryGenesDB/IR64/os.ir64.cshl.draft.1.0.scaffold.fa
```

Copy perl script

```
mkdir $FORMATION/bin
export PATH=$PATH:$FORMATION/bin
cp /home/droc/OryGenesDB/IR64/script/* $FORMATION/bin
```

Now, all the scripts include in the bin directory can be directly access.

For example:

```
length.pl
```

How many scaffolds were produces on this assembly?

```
grep –c ">"os.ir64.cshl.draft.1.0.scaffold.fa
```

2. **Comparison of both proteome with Blast**

If necessary, you need to format your fasta before running a blast

```
formatdb -i IR64_prot.faa
```

Run reciprocal blast between Nipponbare and IR64

```
blast_cluster.pl --help
```

Do not run the following command, it take too many time

- IR64 vs Nipponbare

```
nohup blast_cluster.pl -input IR64_prot.faa --program blastp --database MSU7_pep_20111031 --output IR64_vs_MSU --directory $PWD/bbmh/ --num_seq_by_batch 1000 --evalue 1e-10 –format 6&
```

The option --format 6 is very important to get the output as tabulated file.

- Nipponbare vs IR64

```
nohup blast_cluster.pl -input MSU7_pep_20111031 --program blastp --database IR64_prot.faa --output MSU_vs_IR64 --directory $PWD/bbmh/ --num_seq_by_batch 1000 --evalue 1e-10 –format 6&
```

Get result for reciprocal blast between IR64 and MSU proteome

```
cd bbmh
ln -s /home/droc/OryGenesDB/IR64/bbmh/IR64_vs_MSU
ln -s /home/droc/OryGenesDB/IR64/bbmh/MSU_vs_IR64
```

Find putative ortholog

```
cd $FORMATION
findBBMH.pl bbmh/
```

"_vs_" this separator is required if you used findBBMH.pl

If you want to have a look onto this program

```
which findBBMH.pl
```

How many relations were identify?

```
wc –l MSU.BBMH
```

Now we want to have for each relation the location on the assembly. This information are integrated on the GFF3 files.

```
ln -s /home/droc/OryGenesDB/os_japonica/gff3/version7.0/MSU_v7.gff3
ln -s /home/droc/OryGenesDB/IR64/IR64.gff3
```

3. Analyse the result

parse_gff.pl: Extract for a GFF3 file, the primary_id, the reference sequence (chromosome, or contig name), and the location (start, end, strand).

```
parse_gff.pl -g IR64.gff3 -o locus_ir64.txt
```

-t, --tag by default is "gene" , this correspond to the third column of the GFF3 file.

For Nipponbare, we need to change the tag value to "mRNA" to match with MSU.BBMH output file

```
parse_gff.pl  -g MSU_v7.gff3  -t mRNA  -o locus_nipponbare.txt
```

```
[droc@marmadais bioperl]$ head -7 MSU.BBMH
Os02g35140.1    maker-scaffold_700-snap-gene-0.34
Os05g40850.1    maker-scaffold_222-snap-gene-2.92
Os02g33400.1    maker-scaffold_7-pred_gff_Fgenesh-gene-3.1
Os07g10630.1    maker-scaffold_1207-snap-gene-0.20
Os04g44420.1    snap-scaffold_423-processed-gene-1.54
Os02g51490.1    maker-scaffold_5-snap-gene-9.62
Os01g01010.1    maker-scaffold_0-pred_gff_Fgenesh-gene-0.1
```

```
[droc@marmadais formation bioperl]$ more IR64.gff3
##gff-version 3
##sequence-region scaffold_0 1 2857538
##gff-version 3
scaffold_0      IR64    contig  1       2857538 .       .       1       ID=scaffold_0
scaffold_0      maker   gene    7330    7659    .       -       .       ID=maker-
scaffold_0-pred_gff_Fgenesh-gene-0.1;Name=maker-scaffold_0-pred_gff_Fgenesh-gene-0.1
scaffold_0      maker   mRNA    7330    7659    .       -       .       ID=maker-
scaffold_0-pred_gff_Fgenesh-gene-0.1-mRNA-1;Parent=maker-scaffold_0-pred_gff_Fgenesh-gene-
0.1;Name=scaffold_0_FG000002-mRNA.1;_QI=0|-1|0|1|-1|1|1|0|109
scaffold_0      maker   CDS     7330    7659    .       -       0       Parent=maker-
scaffold_0-pred_gff_Fgenesh-gene-0.1-mRNA-1
scaffold_0      maker   exon    7330    7659    .       -       .       Parent=maker-
scaffold_0-pred_gff_Fgenesh-gene-0.1-mRNA-1
```

```
[droc@marmadais formation bioperl]$ more MSU_v7.gff3
##gff-version 3
Os01    MSU     gene    2903    10817   .       +       .
ID=Os01g01010;Name=Os01g01010
Os01    MSU     mRNA    2903    10817   .       +       .
ID=Os01g01010.1;Parent=Os01g01010
Os01    MSU     polypeptide     3449    10297   .       +       1       ID=Os01g01010.1-
pep;Derives_from=Os01g01010.1
Os01    MSU     CDS     3449    3616    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     4357    4455    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     5457    5560    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     7136    7944    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     8028    8150    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     8232    8320    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     8408    8608    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     9210    9617    .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     10104   10187   .       +       .       Parent=Os01g01010.1
Os01    MSU     CDS     10274   10297   .       +       .       Parent=Os01g01010.1
```

Get putative ortholog for a specific region

```
get_location.pl -b MSU.BBMH -n locus_nipponbare.txt -i locus_ir64.txt -e 19200000 -c Os09 -s
18800000 -o qtl.txt
```

Sort data

```
sort -k 3 qtl.txt
```

Q1: How many scaffolds on IR64 cover the QTL region?

Q2: How many genes are predicted on each scaffold?

Index fasta file

```
fasta-make-index os.ir64.cshl.draft.1.0.scaffold.fa
```

Get sequence for a specific ID

```
fasta-fetch os.ir64.cshl.draft.1.0.scaffold.fa scaffold_829 > scaffold_829.fna
fasta-fetch os.ir64.cshl.draft.1.0.scaffold.fa scaffold_534 > scaffold_534.fna
fasta-fetch os.ir64.cshl.draft.1.0.scaffold.fa scaffold_134 > scaffold_134.fna

length.pl scaffold_829.fna
length.pl scaffold_534.fna
length.pl scaffold_134.fna
```

Q3: Does all the scaffolds are complete on Nipponbare ?

Q4: Does all the genes are present?

```
grep scaffold_829 locus_ir64.txt  | cut -f 1 > scaffold_829_gene.txt
grep scaffold_534 locus_ir64.txt  | cut -f 1 > scaffold_534_gene.txt
grep "scaffold_134-" locus_ir64.txt  | cut -f 1 > scaffold_134_gene.txt
```

```
fasta-make-index IR64_prot.faa
fasta-fetch IR64_prot.faa -f scaffold_829_gene.txt  > scaffold_829_gene.faa
fasta-fetch IR64_prot.faa -f scaffold_534_gene.txt > scaffold_534_gene.faa
fasta-fetch IR64_prot.faa -f scaffold_134_gene.txt > scaffold_134_gene.faa
```

Run a Blastp against SwissProt database

```
blast_cluster.pl  -input scaffold_829_gene.faa -directory $PWD -database
/bank/uniprot/taxo33090_review_yes.faa-program blastp -output scaffold_829_gene.out -evalue
1e-20
```

Parse BLAST result

```
parse_blast_result.pl -f scaffold_829_gene.out -n 1
```