CAPTUREC ANALYSER 2 – USER MANUAL AND HELP

1) USER MANUAL
2) HELP / USAGE

1) USER MANUAL

This manual is also available with command : perl CCanalyser2.pl --man

NAME

CCanalyser.pl

SYNOPSIS

This script uses a sam file as input. This needs to have been generated as follows:

1. Perform adaptor trimming of the raw fastq files if this has not been performed by the sequencer. I tend to use trim_galore to do this with the paired settings. For example: nohup trim_galore --paired Paired_end_1.fastq Paired_end_2.fastq&

2. If you are using 150 bp paired end reads from the Miseq then the overlapping reads need to be merged. I use FLASH to do this. For example
   module load flash
   nohup flash --interleaved-output Paired_end_1.fq Paired_end_2.fq&
   Then concatenate the separate files of merged and unmerged reads:
   cat out.notCombined.fastq out.extendedFrags.fastq -> Combined_reads.fastq

3. If you sequence the data with reads that are shorter than the fragments then you may not need to merge the reads with flash but you will need to merge the reads from the PE1 and PE2 by interleaving them making sure that the order of the reads is maintained.
   I have written a

4. I normally align the fastq files with bowtie using one processor only because it is crucial that the reads are kept in strict order otherwise the script will not function properly.
   If you use more than one processor then it is crucial the reads are sorted so that they are ordered by name.

The script only parses the first digit in the cigar string, which should be fine with bowtie but be careful with other aligners (it will only take the first digit if the cigar is 20M29M

I tend to use m2, best and strata Ð but this is partly because m2 is needed otherwise all of the reads for alpha globin will be thrown (because of the gene duplication) so you may well want to use your own setting.

I would use 'best' so that each read can only align once otherwise each read could be counted many times.

For example:

```
nohup bowtie -p 1 -m 2 --best --strata --sam --chunkmb 256 --sam /databank/indices/bowtie/mm9/mm9 Sample_REdig.fastq Sample_REdig.sam &
```

The script needs to have two other input files:

1. A file of all of the restriction enzyme fragments in the genome. This is best made using the script dpngenome.pl.
   The coordinates are of the middle of the restriction fragment in the format chr:start-stop

2. A file of the input coordinates of the capture oligonucleotides it needs to in the following 9 column tab separated format.
   Be careful to ensure the /n new line is used rather than /r (as can be inserted by excel for example).
   1. Name of capture (avoid spaces in the names please)
   2. Chromosome of capture fragment
   3. Start coordinate of capture fragment
   4. End of capture fragment
   5. Chromosome of proximity exclusion
   6. Start coordinate of proximity exclusion
   7. End coordinated of proximity exclusion
   8. Position of SNP
   9. Base of SNP

The script requires the perl modules Data::Dumper; Getopt::Long and Pod::Usage and it needs wigToBigWig.

Depending on your system setup you may need to load wigToBigWig BEFORE running this script with the command module load ucsctools

This script will create a subdirectory (of the directory the script is in) named after the sample and it will put into this wig, windowed wig and sam files of all of the reads that are to be reported

It also outputs a sam file of all of the reads mapping to the capture region.

In addition it outputs 2 report files. One containing the statistics and the other containing the data relating to exclusion of duplicates.

The script will convert the wig files to bigwig, copy them to your public folder and generate a track hub so that all you need to do is paste the url for the track hub into the UCSC genome browser to see the data.

EXAMPLE

```
    CCanalyser.pl -f input_sam_file.sam -o input_oligo_file.txt -r input_restriction_enzyme_coordinates_file.txt -s short_sample_name -pf public_folder -pu
public_url
```

OPTIONS

```
    -f          Input filename
    -r          Restriction coordinates filename
    -o          Oligonucleotide position filename
    -pf         Your public folder (e.g. /hts/data0/public/username)
    -pu         Your public url (e.g. sara.molbiol.ox.ac.uk/public/username)
    -s          Sample name (and the name of the folder it goes into)
    -w          Window size (default = 2kb)
    -i          Window increment (default = 200bp)
    -dump       Print file of unaligned reads (sam format)
    -snp        Force all capture points to contain a particular SNP
    -limit      Limit the analysis to the first n reads of the file
    -genome     Specify the genome (mm9 / hg18)
    -globin     Combines the two captures from the gene duplicates (HbA1 and HbA2)
```

2) USAGE INSTRUCTIONS / HELP

This help is also available with command : perl CCanalyser2.pl --help or   perl CCanalyser2.pl -h
Usage:

This script uses a sam file as input. This needs to have been generated as follows:

1. Perform adaptor trimming of the raw fastq files if this has not been performed by the sequencer. I tend to use trim_galore to do this with the paired settings. For example: nohup trim_galore --paired Paired_end_1.fastq Paired_end_2.fastq&

2. If you are using 150 bp paired end reads from the Miseq then the overlapping reads need to be merged. I use FLASH to do this. For example

```
module load flash
nohup flash --interleaved-output Paired_end_1.fq Paired_end_2.fq&
```

Then concatenate the separate files of merged and unmerged reads:

```
cat out.notCombined.fastq out.extendedFrags.fastq -> Combined_reads.fastq
```

3. If you sequence the data with reads that are shorter than the fragments then you may not need to merge the reads with flash but you will need to merge the reads from the PE1 and PE2 by interleaving them making sure that the order of the reads is maintained.

I have written a

4. I normally align the fastq files with bowtie using one processor only because it is crucial that the reads are kept in strict order otherwise the script will not function properly.

If you use more than one processor then it is crucial the reads are sorted so that they are ordered by name.

The script only parses the first digit in the cigar string, which should be fine with bowtie but be careful with other aligners (it will only take the first digit if the cigar is 20M29M

I tend to use m2, best and strata Ð but this is partly because m2 is needed otherwise all of the reads for alpha globin will be thrown (because of the gene duplication) so you may well want to use your own setting.

I would use 'best' so that each read can only align once otherwise each read could be counted many times.

For example:

```
nohup bowtie -p 1 -m 2 --best --strata --sam --chunkmb 256 --sam /databank/indices/bowtie/mm9/mm9 Sample_REdig.fastq Sample_REdig.sam &
```

The script needs to have two other input files:

1. A file of all of the restriction enzyme fragments in the genome. This is best made using the script dpngenome.pl.

The coordinates are of the middle of the restriction fragment in the format chr:start-stop

2. A file of the input coordinates of the capture oligonucleotides it needs to in the following 9 column tab separated format.

Be careful to ensure the /n new line is used rather than /r (as can be inserted by excel for example).

1. Name of capture (avoid spaces in the names please)

2. Chromosome of capture fragment
3. Start coordinate of capture fragment
4. End of capture fragment
5. Chromosome of proximity exclusion
6. Start coordinate of proximity exclusion
7. End coordinated of proximity exclusion
8. Position of SNP
9. Base of SNP

The script requires the perl modules Data::Dumper; Getopt::Long and Pod::Usage and it needs wigToBigWig.
Depending on your system setup you may need to load wigToBigWig BEFORE running this script with the command module load ucsctools

This script will create a subdirectory (of the directory the script is in) named after the sample and it will put into this wig, windowed wig and sam files of all of the reads that are to be reported
It also outputs a sam file of all of the reads mapping to the capture region.
In addition it outputs 2 report files. One containing the statistics and the other containing the data relating to exclusion of duplicates.
The script will convert the wig files to bigwig, copy them to your public folder and generate a track hub so that all you need to do is paste the url
for the track hub into the UCSC genome browser to see the data.

Options:

-f	Input filename
-r	Restriction coordinates filename
-o	Oligonucleotide position filename
-pf	Your public folder (e.g. /hts/data0/public/username)
-pu	Your public url (e.g. sara.molbiol.ox.ac.uk/public/username)
-s	Sample name (and the name of the folder it goes into)
-w	Window size (default = 2kb)
-i	Window increment (default = 200bp)
-dump	Print file of unaligned reads (sam format)
-snp	Force all capture points to contain a particular SNP
-limit	Limit the analysis to the first n reads of the file

-genome Specify the genome (mm9 / hg18)

-globin Combines the two captures from the gene duplicates (HbA1 and HbA2)