Analysis_report

Jennifer Semple

10/29/2019

Custom plots for nanoHIC

Results of mapping are read in from hdf5 file in frag_files folder using the script js_plotHDF5. Many stats along the way are saved into the important_stats list. This document describes them a bit.

```
sampleName<-"20190501_HIC6_7_barcode08_pass"
importantStats<-readRDS(paste0("./rds/importantStats_",sampleName,".rds"))
```

Total read and fragment numbers

TotalReadsSequenced and *TotalFragsSequenced* refer to the highest read and fragment number that were assigned by th MC-HiC-guppy pipeline when processing the fastq reads and giving the sequential numbers as IDs. After mapping the total number of reads/frags goes down considerably.

TotalMappedReads and *TotalMappedFragments* refer to the number of reads and fragments in the hdf5 file. all these reads have a MQ>0 (i.e. they are mapped).

```
format(importantStats$TotalReadsSequenced,big.mark=",")
```

```
## [1] "296,505"
```

```
format(importantStats$TotalFragsSequenced,big.mark=",")
```

```
## [1] "930,041"
```

```
format(importantStats$TotalMappedReads,big.mark=",")
```

```
## [1] "218,290"
```

```
format(importantStats$TotalMappedFrags,big.mark=",")
```

```
## [1] "435,830"
```

Filtering of reads

The hdf5 table contains also reads mapped to multiple locations.

uniquelyMappedFrags refers to the number of fragments mapped uniquely in the genome (occur only once in the hdf5 file). These reads were given a uniqueMapping score of 1.

multiMappedFrags refers to the number of fragments that have 2 or more occurrences in the genome. The multiMappedFrags were given a uniqueMapping score of 0. They were filtered to keep the one with the best mapping score for any given fragment.

ambigMappedFrags counts the number of fragments where more than one alignment had the same top mapping score, these were considered ambiguous mappings (many of these had a low MQ, but some had a MQ as high as 250). Finally the reads were filtered to remove any mtDNA fragments, as there cannot be real physical interactions.

bestUniqMappedFrag gives the count of all singly mapped reads and multimapped reads for which a top-mapping-score fragment could be identified. Histogram and plots of the MQ of these fragments and reads are in **mappingQuality_bestUniqMapped.pdf**

trueHops exclude all fragments that are in tandem and closer than 500bp (or whatever threshold the pipeline puts).

trueHops_MQgt20 include only true hops that also have a mapping quality 20 or higher. These are the fragments that are used for generating the HiC matrix in the MC-HiC pipeline. These are saved in the ./rds/bestUniqMapped_sampleName.rds file.

```
format(importantStats$uniquelyMappedFrags,big.mark=",")
```

```
## [1] "353,654"
```

```
format(importantStats$multiMappedFrags,big.mark=",")
```

```
## [1] "82,176"
```

```
format(importantStats$ambigMappedFrags,big.mark=",")
```

```
## [1] "620"
```

```
format(importantStats$bestUniqMappedFrags,big.mark=",")
```

```
## [1] "391,135"
```

```
format(importantStats$trueHops,big.mark=",")
```

```
## [1] "350,124"
```

```
format(importantStats$trueHops_MQgt20,big.mark=",")
```

```
## [1] "296,123"
```

Minimum, maximum and median read and fragment lengths are also recorded for bestUniqMapped

```
importantStats["medianReadLength"]
```

```
## $medianReadLength
## [1] 682
```

```
importantStats["medianFragLength"]
```

```
## $medianFragLength
## [1] 249
```

```
importantStats["minReadLength"]
```

```
## $minReadLength
## [1] 74
```

```
importantStats["minFragLength"]
```

```
## $minFragLength
## [1] 16
```

```
importantStats["maxReadLength"]
```

```
## $maxReadLength
## [1] 19971
```

```
importantStats["maxFragLength"]
```

```
## $maxFragLength
## [1] 5000
```

Summarising data per read

Data is grouped by read ID and then number of fragments and number/fraction of multi fragment reads are calculated. This is performed for all reads (bestUniqMapped), for all fragments with MQ>=20, and for all reads with MQ>=20 that are also true hops.

```
importantStats["maxFragPerRead_bestUniqMapped"]
```

```
## $maxFragPerRead_bestUniqMapped
## [1] 43
```

```
importantStats["FractionMultiFragReads_bestUniqMapped"]
```

```
## $FractionMultiFragReads_bestUniqMapped
## [1] 0.41
```

```
importantStats["maxFragPerRead_MQ>=20"]
```

```
## $`maxFragPerRead_MQ>=20`
## [1] 37
```

```
importantStats["FractionMultiFragReads_MQ>=20"]
```

```
## $`FractionMultiFragReads_MQ>=20`
## [1] 0.38
```

```
importantStats["maxFragPerRead_TrueHop"]
```

```
## $maxFragPerRead_TrueHop
## [1] 16
```

```
importantStats["FractionMultiFragReads_TrueHop"]
```

```
## $FractionMultiFragReads_TrueHop
## [1] 0.33
```

Summarising data about mtDNA

Mitochondrial DNA should have no 3D contact with nuclear DNA, therefore it serves as an internal control.

Are mtDNA over or under respresented?

Ideally there would be no mtDNA fragments in the read. But one certainly should hope that they are under-represnted compared to nuclear fragments. To calculate that we can compare to the relative size of the mtDNA (13794bp) to the whole genome (100Mb)

```
importantStats$numFragsMtDNA
```

```
## [1] 491
```

```
importantStats$percentFragsMtDNA
```

```
## [1] 0.1253747
```

```
# Note: MtDNA size is 13794 bp
importantStats$percentGenomeMtDNA
```

```
## [1] 0.013794
```

We can do the same calculation at the read level

```
importantStats$numAllReads
```

```
## [1] 218038
```

```
importantStats$numReadsMtDNA
```

```
## [1] 378
```

```
importantStats$percentReadsWithMtDNA
```

```
## [1] 0.17
```

But actually what most worries use are reads that have fragments from both nuclear and mitochondrial genomes:

```
importantStats$numMixedReadsMtDNA
```

```
## [1] 190
```

```
importantStats$percentMixedReadsWithMtDNA
```

```
## [1] 0.09
```