Binf 6110 Assignment 1: Analysis of the D. discoideum reference genome

Jason Moggridge 1159229

23/01/2021

Overview of *Dictyostelium discoideum* genome

I chose to examine the genome of the social amoeba *Dictyostelium discoideum*, which was published by Eichinger *et al.* in 2005. *D. discoideum* is a interesting model organism for many areas of research in molecular biology, particularly because of their ability to cooperatively form a multicellular super-organism for the purpose of sporulation under starvation conditions (Williams, 2010). The genome was created by whole chromosome shotgun sequencing using the Sanger method. For this, chromosomes were separated using PFGE. Isolated chromosomes were fragmented and then cloned using plasmids and yeast artificial chromosomes. As this was among the first protozoan genome projects, there was no draft genome to guide contig assembly. As such, data from *HAPPY* mapping, previously mapped genes, and YACs were used in creating the assembly. *HAPPY* mapping is akin to linkage mapping but uses DNA fragmentation and PCR to get distance information for sequence-tagged sites. This information was used to assign reads from each library to chromosome-specific bins through `BLAST` or `Atlas Overlapper`. Binned reads were then joined into contigs by either `GAP4` or `PHRED/PHRAM/CONSED` software. Then read-pair data and `BLAST` searches of all sequence data were used to extend contigs and create scaffolds. Gap were closed by various strategies, depending on whether sequencing was challenged by repetitive or A+T rich regions. The authors presented an analysis of the genome's nucleotide composition, repeats, transposons, tRNAs, telomeric and centromeric regions. In addition, the genome was used for gene prediction and analysis of the predicted proteome; predicted peptide domains were used to create a phylogeny to place *D. discoideum* among other eukaryotic phyla.

Links - (Browse genome record on NCBI) - (Zipped FASTA file through ftp)

Analysis of genome

Lines starting with '$' are unix commands and '#' are comments, otherwise lines are R code

I obtained the *Dictyostelium discoideum* reference genome from NCBI Refseq as follows:

```
$ rsync rsync://ftp.ncbi.nih.gov/genomes/refseq/protozoa/Dictyostelium_discoideum/
  all_assembly_versions/GCF_000004695.1_dicty_2.7/
  GCF_000004695.1_dicty_2.7_genomic.fna.gz  ./
$ gunzip GCF_000004695.1_dicty_2.7_genomic.fna.gz
$ mv GCF_000004695.1_dicty_2.7_genomic.fna ./dicty_genome.fna
```

Genome assembly size

```
# remove lines with headers (>), strip newline characters (\n), count characters
$ grep -v ’>’ dicty_genome.fna | tr -d ’\n’ | wc -c
```

The *D. discoideum* assembly has a size of 34.2 Mbp.

Number of chromosomes/scaffolds

```
# count scaffold headings
$ grep ’>’ dicty_genome.fna | wc -l
```

There are 41 scaffolds in the assembly; these include 6 genomic chromosomes, the mitochondrial chromosome and a plasmid chromosome, as well as 33 unplaced scaffolds.

Scaffold lengths

```
# create a table with headings: Scaffold, Length
$ touch seqlens.tsv
$ echo -e ’Scaffold\tLength’ > seqlens.tsv
# loop over headings, get length of following sequence for each
$ awk ’/^>/ {if (seqlen){print seqlen}; printf $0"\t";seqlen=0;next; }
{seqlen += length($0)}END{print seqlen}’ dicty_genome.fna >> seqlens.tsv
```

```
library(tidyverse)
seqlen <- read_delim(’data/seqlens.tsv’, ’\t’)
seqlen <- seqlen %>%
  arrange(desc(Length)) %>%
  mutate(Rank = row_number(),
         Type = case_when(
           str_detect(Scaffold, ’chrUn_’) ~ ’Unplaced’,
           str_detect(Scaffold, ’plasmid’) ~ ’Plasmid’,
           str_detect(Scaffold, ’mitochond’) ~ ’Mitochondrial’,
           TRUE ~ ’Nuclear chr.’)
  )
plt.title <- expression(paste(’Fig. 1: ’, italic("D. discoideum"), " scaffold lengths"))

ggplot(seqlen) +
  geom_point(aes(x = Rank, y = Length, colour = Type)) +
  scale_y_log10() +
  labs(x=’Scaffold’,
       y = ’Scaffold length (bp)’,
       subtitle = plt.title) +
  theme(legend.position = ’right’,
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 8))
```

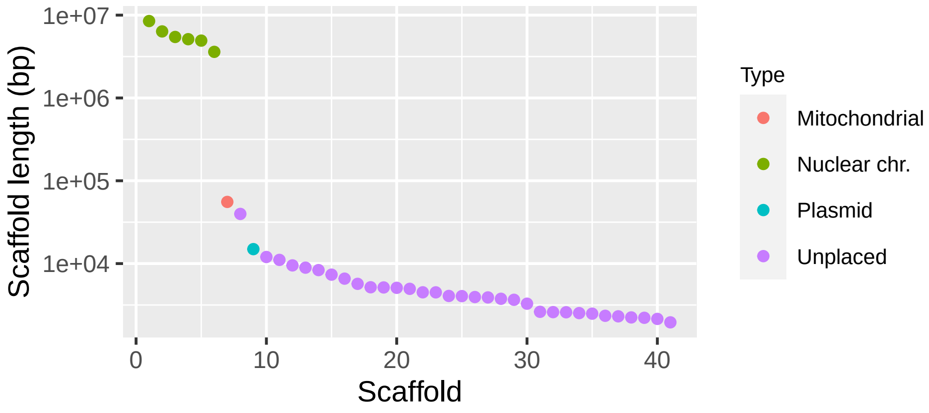

Fig. 1: *D. discoideum* scaffold lengths

The mean scaffold length is 834.3 kbp. Scaffold lengths are shown in fig. 1.

N50 & L50

I was able to recreate the N50 (5,450,249 bp) and the L50 (3) statistics reported on NCBI as follows:

```
N50 <- function(scaffold_lengths){
  lengths <- sort(scaffold_lengths, decreasing = T)
  total <- sum(lengths); cumul <- 0
  for (i in seq_along(lengths)){
    cumul <- cumul + lengths[i]
    if (cumul >= total/2) return(lengths[i])
  }
}
dicty.N50 <- N50(seqlen$Length)
dicty.L50 <- which(sort(seqlen$Length, decreasing = T) == dicty.N50)
```

Ambiguous bases (Ns)

```
# Ncontent
grep -v ’^>’ dicty_genome.fna | tr -cd ’Nn’ | wc -c
# 23142

# Total
grep -v ’>’ dicty_genome.fna | tr -d ’\n’ | wc -c
# 34204973
```

```
N_percent <- 23142 * 100 / 34204973
```

Only a very small portion (0.068 %) of the assembly are ambiguous bases (N).

Ratio of repetitive to unique sequence

```
# repetitive regions bases count
```

```
grep -v ’>’ dicty_genome.fna | tr -d ’\n’ | tr -cd acgt | wc -c
# 18093064

# unique regions bases count
grep -v ’>’ dicty_genome.fna | tr -d ’\n’ | tr -cd ACGT | wc -c
# 16088767
```

The ratio of repetitive to unique sequence is 1.12.

What proportion of the genome is unplaced scaffolds?

```
seqlen %>%
  group_by(Type) %>%
  summarize(‘Percent of Genome‘ = round(sum(Length)/sum(seqlen$Length)*100, 2))
```

Type	Percent of Genome
Mitochondrial	0.16
Nuclear chr.	99.23
Plasmid	0.04
Unplaced	0.56

Discussion

The *D. discoideum* reference genome is relatively small (34.2 Mbp) and the assembly is of remarkably high quality and completeness. The assembly has nearly complete scaffolds for the six nuclear chromosomes, with very few unplaced scaffolds (33; fig. 1). The percentage of the assembly in these unplaced scaffolds is only 0.56 %. Correspondingly, the assembly has a very large N50 (5.45 Mbp) and a very small L50 (3 scaffolds). The sequence data is of high quality, with only 0.07 % ambiguous bases (N). The genome contains a relatively even split between repetitive sequence and unique sequence (ratio = 1.12).

The quality of the genome is surprisingly good, given that the effort was reliant on cloning and Sanger sequencing technology of the time, and dealt with complex repetitive regions that are difficult to assemble, as well as A+T-rich tracts that can be difficult to clone and sequence (Eichinger *et al.*, 2011). To create this near-perfect *de-novo* genome assembly, the authors applied data from contiguity methods that were commonly used at the time: genetic mapping (HAPPY maps), physical mapping (with YACs, *etc.*), and paired reads. Overall, this reference genome should prove to be an invaluable resource for studying this intriguing model amoeba.

References

- Eichinger L., *et al.*. 2005. The genome of the social amoeba Dictyostelium discoideum. *Nature.* May 5; 435(7038): 43–57.
- Williams, JG. 2010. Dictyostelium finds new roles to model. *Genetics*, 185(3):717–726.