BIOS 426/826-001

Homework Assignment 2

Due on 11:59PM, September 16, 2015

We want to sequence a species of algae, which is close to *Chlorella variabilis* NC64A.

Question 1) What is GeneBank Accession number for the genome sequence of *Chlorella variabilis* NC64A? We need to know this, because we want to use it as a reference for our *de novo* assembly.

The GenBank assecssion number for the genome is

GenBank: ADIC00000000.1

<http://www.ncbi.nlm.nih.gov/nuccore/ADIC00000000.1>

GCF\_000147415.1 is accession number for the assembly

Question 2) Which published article describe the sequencing and assembly of *Chlorella variabilis* NC64A genome? Please show me the citation.

Blanc G, [Duncan G](http://www.ncbi.nlm.nih.gov/pubmed/?term=Duncan%20G%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Agarkova I](http://www.ncbi.nlm.nih.gov/pubmed/?term=Agarkova%20I%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Borodovsky M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Borodovsky%20M%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Gurnon J](http://www.ncbi.nlm.nih.gov/pubmed/?term=Gurnon%20J%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Kuo A](http://www.ncbi.nlm.nih.gov/pubmed/?term=Kuo%20A%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Lindquist E](http://www.ncbi.nlm.nih.gov/pubmed/?term=Lindquist%20E%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Lucas S](http://www.ncbi.nlm.nih.gov/pubmed/?term=Lucas%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Pangilinan J](http://www.ncbi.nlm.nih.gov/pubmed/?term=Pangilinan%20J%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Polle J](http://www.ncbi.nlm.nih.gov/pubmed/?term=Polle%20J%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Salamov A](http://www.ncbi.nlm.nih.gov/pubmed/?term=Salamov%20A%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Terry A](http://www.ncbi.nlm.nih.gov/pubmed/?term=Terry%20A%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Yamada T](http://www.ncbi.nlm.nih.gov/pubmed/?term=Yamada%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Dunigan DD](http://www.ncbi.nlm.nih.gov/pubmed/?term=Dunigan%20DD%5BAuthor%5D&cauthor=true&cauthor_uid=20852019),[Grigoriev IV](http://www.ncbi.nlm.nih.gov/pubmed/?term=Grigoriev%20IV%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), [Claverie JM](http://www.ncbi.nlm.nih.gov/pubmed/?term=Claverie%20JM%5BAuthor%5D&cauthor=true&cauthor_uid=20852019), Van Etten JL. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. [Plant Cell.](http://www.ncbi.nlm.nih.gov/pubmed/20852019) 2010 Sep;22(9):2943-55.

# This work was achieved by a UNL group, lead by Dr. Van Etten.

Question 3) What is the genome size of *Chlorella variabilis* NC64A. How many chromosomes does it have? How many genes does it have?

Genome size is 46.16 Mbp

Number of Chromosome = 12

Number of Gene = 9780

Question 4) Based on the knowledge of the *Chlorella variabilis* NC64A genome, can you estimate the genome size of the algae species that we want to sequence? What is it?

Around 50 Mbp

We got quotations from two companies for sequencing and assembly. Please see the following.

**Company #1:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Insert Size** | **# of Library** | **Library** | **# of reads** | **Read length** | **Effective Coverage** | **Type of Sequencing** |
| 500bp | 1 | PE | 55.6M | 90 | 100x | HiSeq 2000 |
| 6kb | 1 | ME | 16.6M | 90 | 30x | HiSeq 2000 |
| 10kb | 1 | ME | 11.1M | 90 | 20x | HiSeq 2000 |
| Total | 3 |  |  |  | 150x | -- |

Quotation from the Company #1: “A package price for library+sequencing+ Bioinforamtics(survey+assembly+annotation+evolution): $19,200.  With an additional 10% discount of this project, the final price is down to $17,280.”

**Company #2:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Insert Size** | **# of Library** | **Library** | **# of reads** | **Read length** | **Effective Coverage** | **Type of Sequencing** |
| 600bp | 1 | PE | 66.6M | 150bp | 200x | HiSeq 2500 |
| 5kb | 1 | ME | 66.6M | 75bp | 100x | MiSeq |
| Total | 2 | ME |  |  | 300x | -- |

Quotation from the Company #2: “DNA extraction (optional) ($150)+ NGS library prep ($600)+ Sequencing ($3,000)+ Bioinformatics ($7,500)=#11,100.”

Question 5) Please finished the above to tables. For the column of “Library”, if you think that library is a standard paired-end library, please fill it with “PE”. If you think that library is a Mate pair library, please fill it with “MP”. You need to estimate the total number of reads for each library based on the length of read, the expected coverage, and your estimated genome size.

Coverage=(Read Length x Number of reads)/Genome size

Question 6) If you are leading and supervising this sequencing project, which company are you going to use based on their quotations, and why? Or, you can compare the advantages and disadvantages of two companies’ designs.

For Company #1, the length of short reads is short and the coverage is lower than that for the Company #2.

For Company #2, it will be better if they can make an additional longer ME library, such as 10K ME. The quotation of Company #2 is significantly cheaper than that of Company #1.