# Numbered Heading 1

Every new chapter should start on an odd page.

## numbered Heading 2

### numbered Heading 3

#### numbered Heading 4

Body text in font Garamond 12pt

<https://www.duo.uio.no/bitstream/handle/10852/37680/Bragstad.pdf?sequence=5>

<https://www.duo.uio.no/bitstream/handle/10852/37431/Furenes_Master.pdf?sequence=1>

Tom side

Title for master thesis

Sabba Ifzal

May 1, 2014

Tom side

Acknowledgements

Takke folk som har hjulpet til

Veiledere – Torbjørn, Lex, Geir Kjetil og Ksenia

Andre folk i 10 – Sveinung, Kai, Jon

Ruth for sensur

Ghada for samtaler

etc

Abstract

Kort oppsummert hele oppgaven

Why you do this

How you do this

The results

Conclusion

Preface

Hvem er oppgaven ment for? Andre bioinformatikkstudenter

Contents

[Chapter 1 Numbered Heading 1 1](#_Toc377385015)

[1.1 numbered Heading 2 1](#_Toc377385016)

[1.1.1 numbered Heading 3 1](#_Toc377385017)

[Acknowledgements 5](#_Toc377385018)

[Abstract 6](#_Toc377385019)

[Preface 8](#_Toc377385020)

[Contents 10](#_Toc377385021)

[List of figures 12](#_Toc377385022)

[List of tables 13](#_Toc377385023)

[Chapter 2 Introduction 14](#_Toc377385024)

[2.1 Motivation 14](#_Toc377385025)

[2.2 Hypothesis? 14](#_Toc377385026)

[Chapter 3 Background 15](#_Toc377385027)

[3.1 DNA and RNA 15](#_Toc377385028)

[3.2 Sequencing 15](#_Toc377385029)

[3.2.1 De novo sequencing 15](#_Toc377385030)

[3.2.2 Next generation sequencing 15](#_Toc377385031)

[3.3 Genome Assembly 15](#_Toc377385032)

[3.3.1 Quality measures 15](#_Toc377385033)

[3.3.2 Quast 15](#_Toc377385034)

[3.4 Bioinformatics 15](#_Toc377385035)

[3.4.1 Benchmarking 16](#_Toc377385036)

[3.5 Problem areas 16](#_Toc377385037)

[3.5.1 Assembly 16](#_Toc377385038)

[3.5.2 GAGE-B 16](#_Toc377385039)

[3.5.3 Galaxy framework 16](#_Toc377385040)

[3.6 Technology and formats 16](#_Toc377385041)

[3.6.1 Python 16](#_Toc377385042)

[3.6.2 Bash 16](#_Toc377385043)

[3.6.3 Galaxy 16](#_Toc377385044)

[3.6.4 All assembler 16](#_Toc377385045)

[3.6.5 Fasta and fastq 16](#_Toc377385046)

[3.7 Evaluation methods 16](#_Toc377385047)

[Chapter 4 Material 17](#_Toc377385048)

[4.1 Dataset 17](#_Toc377385049)

[4.2 Reference genome 17](#_Toc377385050)

[4.3 Assemblers? 17](#_Toc377385051)

[Chapter 5 Methods 18](#_Toc377385052)

[5.1 Description 18](#_Toc377385053)

[5.2 Implementation 18](#_Toc377385054)

[Chapter 6 Results 19](#_Toc377385055)

[Chapter 7 Discussion 20](#_Toc377385056)

[7.1 Interpreting the results 20](#_Toc377385057)

[7.2 Analysis of galaxy tool 20](#_Toc377385058)

[7.2.1 Performance 20](#_Toc377385059)

[7.2.2 Potential use 20](#_Toc377385060)

[7.2.3 Weakness 20](#_Toc377385061)

[7.3 Further work 20](#_Toc377385062)

[7.4 Conclusion 20](#_Toc377385063)

[Appendix a (Plot/Figure) 21](#_Toc377385064)

[Appendix B (Tables) 22](#_Toc377385065)

List of figures

[figure 1 21](#_Toc377132394)

[Table 7-1 Bla bla bla 21](#_Toc377132395)

List of tables

[figure 1 21](#_Toc377132379)

[Table 7-1 Bla bla bla 21](#_Toc377132380)

# Introduction

## Background

Why decide to have this project…

Why should the project be done, why use time

## PROBLEM STATEMENT / AIMs

What you actually doing in the project.

Visualize quastresult and add information quast-like, that can be useful for determining which assembly is the best.

Same principle can be added.

Organize a automatized principle. Make something to compare.

## Literature review

Quast, gage assemblathon

Hvordan andre har funnet den beste løsningen

## Problem solution

Briefly about the solution. How you solved it.

Benefit of system. Quast need install etc, but not this.output structure is rather good so we reuse code for our output.

# Background

Since the thesis will involve discussion of problem areas in bioinformatics that require some biological knowledge, this chapter will provide the basics of genome assembly, benchmarking in bioinformatics and the Galaxy Project.

## Genome Assembly

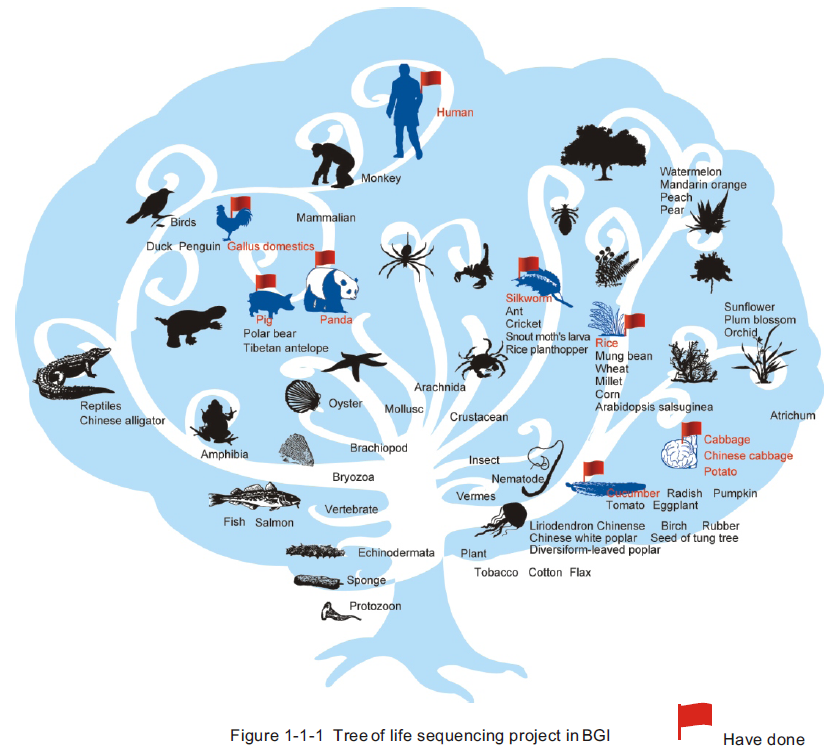


Figure 4‑1 – The tree of life sequencing project in beijing Genomics Institute  
Source - <http://www.hitseq.com/images/BGI-seq-denovo.png>

Genome assembly is the process of trying to reconstruct a sequence from smaller sequences called reads. This process could be trivial if each read had a unique placement, but it doesn’t, thus making assembly quite a difficult process. There are a lot of written work describing different assembly methods, and how the assembly process works (which we’ll come back to later), but it is difficult to find written work describing how or why assembly is a difficult process and challenges regarding it. One paper that try to discuss this is *Genome assembly reborn: recent computational challenges* by Mihai Pop,[[1](#_ENREF_1)] where Pop use solving a jigsaw puzzle as a metaphor to an assembly process. Another complementary paper used for this section is *Genome assembly forensics: finding the elusive mis-assembly* by Phillippy, Scatz and Pop[[2](#_ENREF_2)]

One of the problems regarding assembly process is about genome repeats which can be compared to large stretches of sky in a jigsaw puzzle. Repeats tend to confuse the assembly process, because they seem identical to the assembler. They also make it difficult to distinguish between sequencing error and polymorphism among near-identical repeats. There are also the tandem repeats making it even more difficult. Many of the problems concerning repeats can be generalized in the following two categories:

**1. Repeat collapse and expansion**

The assembly can incorrectly gauge the number of repeats, thus including too few or too many copies, which can result in phenotypic differences, such as Huntigton’s disease.

The assembler can also mis-join reads originating from distinct repeat copies into one unit or include extra copies of repeat, both which can be noted as a higher or lower density of reads.

**2. Sequence rearrangement and inversion**

The assembler can shuffle the order of multiple repeat copies, which could be misinterpreted as a biological rearrangement event, meaning that one could draw wrong conclusions depending on the rearranged sequence.

During both repeat collapse and rearrangement, reads may get placed in a wrong copy of a repeat, therefore SNP could be a useful indicator of such a mis-assembly.

Another problem is looking at the complexity of an assembly, which depends on the number of reads being assembled. An assembly becomes more complex the more reads it assembles, making it quite complex, considering the development of shorter reads and sequencing tools that generates several million reads in a reasonable time and at a low cost.

One of the most time consuming task is probably the computation of overlaps. This task can have assembly errors which can occur due to limitations of the assembly algorithm, or by providing incorrect or incomplete assembly-parameters. It can be difficult to see where there are indels (an insertion or deletion of bases), mis-join, or find the exact placement of reads, and the detection of these errors are what scientists try to improve.

### Sequencing

When looking at sequencing in a biological context, it is usually referred to as a process (a method or technology) that is used to determine the order of nucleotides in a DNA molecule. This process can be used to determine the order of genes, full chromosomes or entire genomes. This is important because the sequence in which the nucleotides appears in gives scientists valuable information about that part of the DNA which can, for instance, be used to look for disease-causing mutations in genes.

Reads

Coverage

Contig

Scaffolds

#### De novo sequencing

De novo sequencing (from latin as “from the beginning”, “afresh” or “anew”) is a collective term used for methods that sequence a DNA sequence for the first time (meaning that it has no previously known sequence), known genomes where significant structural variation is expected, or for microbial sequencing that includes experimental strains and genomes with high plasticity. This kind of sequencing use starting materials such as bacterial, viral, phage, fungus, eukaryote genomic DNA, fragmented DNA or BACs and fosmids. It can use human genome, but substantial challenges exist for its application to this.

#### Next generation sequencing

Next-Generation sequencing (NGS) technologies are usually applied as a general term for sequencing platforms that use post-Sanger technology to sequence DNA fragments in parallel. They were developed to reduce weaknesses in de novo sequencing. As a result, this technology had a major increase in efficiency as well as a decrease in cost per base and could produce a lot of data cheaply compared to de novo sequencing. The “down side” was that the read length was shorter, thus creating an assembly challenge [ ] as it made for example sequencing repeated elements difficult and could result in mis-assembled genomes as Figure X [ ] shows below.

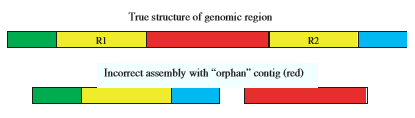


Figure 4‑2 – The challenge when sequencing repeated elements  
SOURCE: STEVEN L. SALZBERG AND JAMES A. YORKE   
 BEWARE OF MIS-ASSEMBLED GENOMES FIG.1

### Assembly

#### Assembly algorithms

#### Assembly challenges

### Quality measures

#### Quast

## Bioinformatics

### Benchmarking

## Galaxy

Introduction to galaxy

# Material

All dataset used in the application here

## Datasets

## Reference genomes

## Software

# Methods

This chapter covers a description of the tool and implementation.

## Reproducing the GAGE-B results

## Galaxy

# Results

Samme som ch5

# Discussion

## Interpreting the GAGE-B results

## Analysis of galaxy tool

### Performance

### Potential use

### Weakness

## Further work

## Conclusion

Referanseliste!

Glossary =)

Glossary

|  |  |
| --- | --- |
|  |  |
|  |  |
| BAC | Abbreviation of *bacterial artificial chromosome* |
| Bacterial artificial chromosome (BAC) | An artificially constructed segment of nucleic acid used for transforming and cloning in bacteria, usually E. coli. |
| bacteriophage | A virus that parasitizes a bacterial cell |
| chromosome | A threadlike, gene-carrying structure found in the nucleus. Each chromosome consists of one very long DNA molecule and associated proteins |
| de novo sequencing |  |
| deoxyribonucleic acid (DNA) | A double-stranded, helical nucleic acid molecule capable of replicating and determining the inherited structure of a cell's proteins |
| DNA | Abbreviation of *deoxyribonucleic acid* |
| eukaryote | An organism whose cells contain membraine-bound organelles and whose DNA is enclosed in a cell nucleus and is associated with proteins |
| fungus | About 80,000 known species of organisms of the kingdom Fungi, which includes the yeasts, rusts, smuts, mildews, molds, mushrooms, and toadstools |
| genome | The complete complement of an organism's genes; an organism's genetic material |
| nucleus | The chromosome-containing organelle of a eukaryotic cell |
| phage | A virus that infects bacteria; also called a bacteriophage |
| protein | A three-dimensional biological polymer constructed from a set of 20 different monomers called amino acids |
| viral | A biological virus |
| virus | A submicroscopic, noncellular particle composed of a nucleic acid core and a protein coat (capsid); parasitic; reproduces only within a host cell. |
|  |  |

Appendix a  
(Plot/Figure)



figure 8‑1

Appendix B  
(Tables)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Table 7-1 Bla bla bla

1. Pop, M.,

*Genome assembly reborn: recent computational challenges.*

Briefings in bioinformatics, 2009. **10**(4): p. 354-66.

2. Phillippy, A.M., M.C. Schatz, and M. Pop,

*Genome assembly forensics: finding the elusive mis-assembly.*

Genome Biology, 2008. **9**(3).