|  |
| --- |
| **NGS and its clinical application** |

1 Assignment

* 1. Background

*Clinical application in the microbiology*

Crucial steps in the world of bacterial pathogen studies and the subsequent infections are the identification of a pathogen, to determine its pathogenic potential, to test its susceptibility to antimicrobial drugs and, for public health purposes, knowledge on the relatedness of the pathogen to other strains. This in order to investigate transmission routes and to allow recognition of outbreaks. Each of these steps depends on many different specialized, species-specific methods which require knowledge of clinical microbiology. These labour-intense, complex and often slow techniques will yield the relevant information necessary. Preferably, all of the information necessary to determine these important factors would be gained in a single step. The genome sequence of an isolate contains all, or almost all of the information required to identify the species, reveal its properties (virulence and resistance) and monitor the (clonal) spread and emergence of these pathogens. Using a next-generation sequencing technique, the full sequence of a bacteria can be determined.

*Clinical application in the oncology*

Cancer genome sequencing studies most of the time have one to four aims; (1) To determine which somatic mutations are likely to contribute to the cancer phenotype (2) Discovering driving mutations. This will lead to improved understanding of basic cancer biology and consequently treatment discovery and development. (3) Characterizing clonal evolution (important regards to cancer treatment). This characterization can be achieved at the nucleotide level using sequencing. (4) Usage for advanced personalized medicine. Cancer genomes can be sequenced to reduce toxicity and increase efficacy trough selecting the correct treatment for the correct patient (dose and time).

1.2 Learning goals

Using this assignment you will learn;

* The clinical application of NGS in the microbiology
* The clinical application of NGS in cancer research

1.3 Literature

*1, Application of next generation sequencing in clinical microbiology and infection prevention.* [Deurenberg RH](https://www.ncbi.nlm.nih.gov/pubmed/?term=Deurenberg%20RH%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Bathoorn E](https://www.ncbi.nlm.nih.gov/pubmed/?term=Bathoorn%20E%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Chlebowicz MA](https://www.ncbi.nlm.nih.gov/pubmed/?term=Chlebowicz%20MA%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Couto N](https://www.ncbi.nlm.nih.gov/pubmed/?term=Couto%20N%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Ferdous M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ferdous%20M%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [García-Cobos S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Garc%C3%ADa-Cobos%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Kooistra-Smid AM](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kooistra-Smid%20AM%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)2, [Raangs EC](https://www.ncbi.nlm.nih.gov/pubmed/?term=Raangs%20EC%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Rosema S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rosema%20S%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Veloo AC](https://www.ncbi.nlm.nih.gov/pubmed/?term=Veloo%20AC%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Zhou K](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhou%20K%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)3, [Friedrich AW](https://www.ncbi.nlm.nih.gov/pubmed/?term=Friedrich%20AW%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)1, [Rossen JW](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rossen%20JW%5BAuthor%5D&cauthor=true&cauthor_uid=28042011)4.

*2, Applications of next-generation sequencing.* <https://www.nature.com/collections/jmgqdxpvsk/content/content>

3, The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide. <https://apps.who.int/iris/bitstream/handle/10665/274443/WHO-CDS-TB-2018.19-eng.pdf>

4, Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification,

<https://www.ncbi.nlm.nih.gov/pubmed/28146134>

5, Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3219767/>

6, Unifying cancer and normal RNA sequencing data from different sources .

[https://www.nature.com/articles/sdata201861](https://www.nature.com/articles/sdata201861 7)

[7](https://www.nature.com/articles/sdata201861 7), Tumor exome sequencing and copy number alterations reveal potential predictors of intrinsic resistance to multi-targeted tyrosine kinase inhibitors, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5777758/>

1.4 Other information

The assignment will be handed in on BB at March 20 23.59 (week 7). Make the assignment alone. If the assignment is insufficient, you will need to do the alternative assignment in this file. The due date will be 24:00 Friday week 10.

*Clinical application in the microbiology*

Next-generation sequencing technologies and their impact on microbial genomics, <https://academic.oup.com/bfg/article/12/5/440/206814>

Background

A good infection prevention policy is aimed, among other things, at preventing the spread of particularly resistant micro-organisms. When a resistant bacteria (eg an ESBL-producing Gram-negative bacterium) is found in a patient, isolation measures are taken to prevent this bacterium from spreading to other patients. In many hospitals, contact research is done when a resistant bacterium is found in a patient. For this, a culture is taken in all patients who have been in contact with this patient (eg roommates) to see if there has been any dissemination.

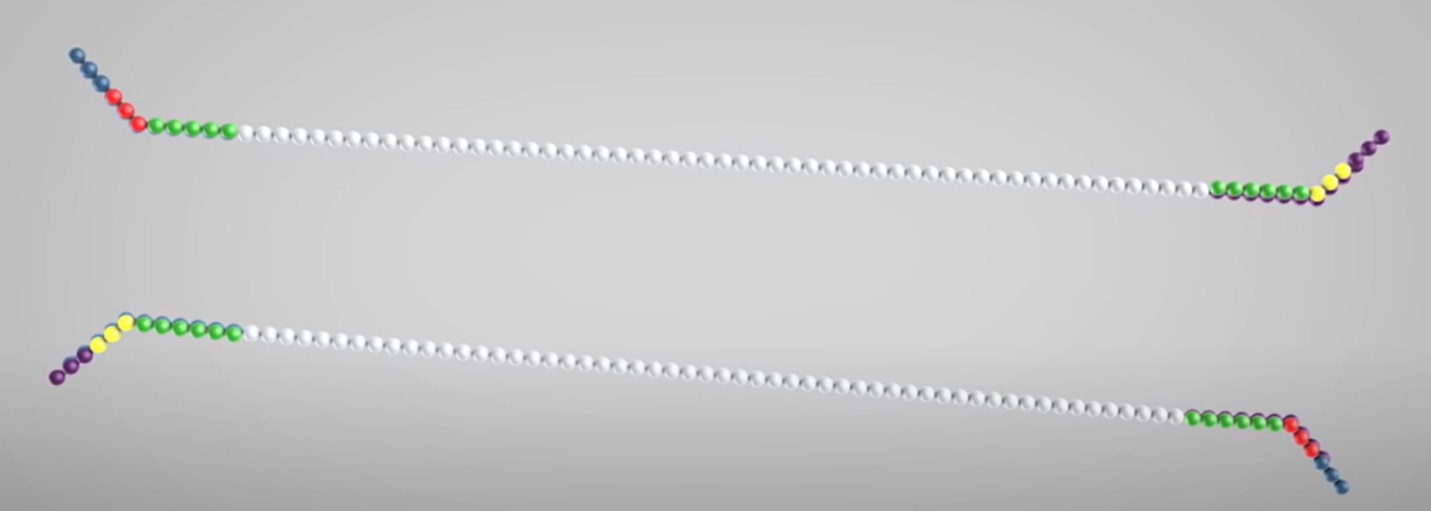
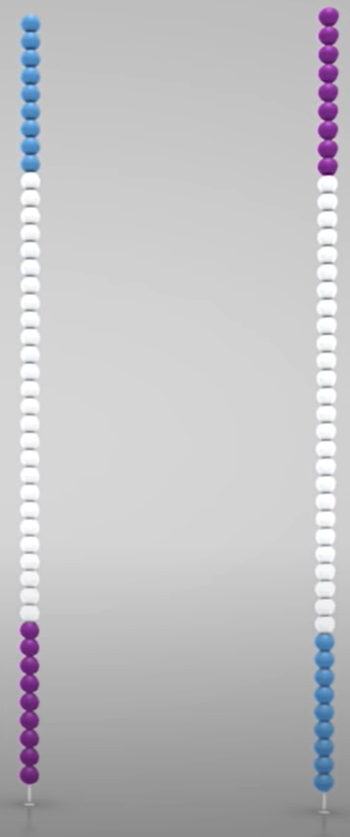
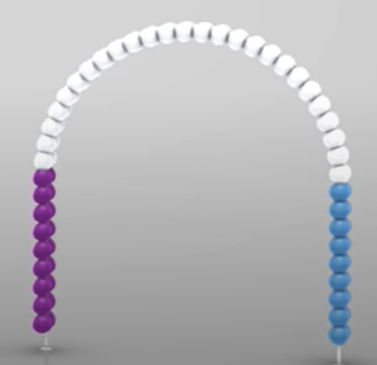
Situation

In patient A an urine culture is taken on 1 August 2016 in connection with suspicion of urinary tract infection. On August 4, 2016, an ESBL-producing E. coli appears to grow in the culture. The patient is placed in (contact) isolation on a single room. Because the patient was in a multiple-person room at the time of the removal, a rectum culture was taken from the five roommates (patients B, C, D, E and F) on 4 August 2016 in order to rule out the spread of the ESBL-producing bacterium to one or more of the roommates. An ESBL-producing bacterium is also detected for three of the roommates (patient B, C and D)

Questions

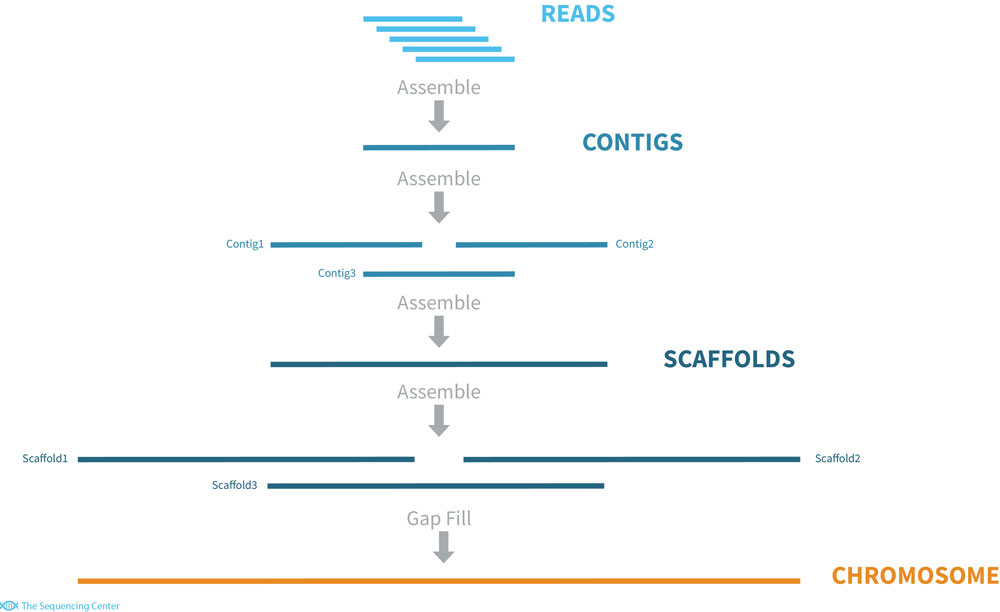
1. One of the most used methods to determine the whole genome sequence of a bacteria is Illumina. Explain how this method is working

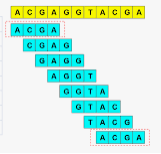
Illumina is a next-generation sequencing technology, responsible for generating more than 90% of the world’s sequencing data. The Illumina sequencing workflow is composed of 4 steps:

* **Sample prep**: adapters are added to the ends of the DNA fragments, during the amplifications motifs are introduced, like the sequencing binding site (DNA-binding proteins can bind to it) indices and regions complimentary to the cell oligos (short piece of single stranded DNA). 
* **Cluster generation**: During clustering each fragment molecule is amplified, with the use of 2 types of oligos. By the first oligo hybridization enabled, so this oligo of the fragment strand will bind to other oligo. Then a polymerase creates a complement of the hybridized fragment, now there is a double stranded string, this is denatured and the original template is washed away. Now the string can bend over to the other oligo on the outer end of the strand. Again a polymerase can form a double stranded string. Now we have 2 single stranded copies of the molecule. With these single strings the process is repeated over and over. So all the fragments will be cloned. The reversed strands will be washed of, so only the forward strands will remain. For unwanted priming the 3’ends are blocked.
* **Sequencing**: to produce the first read the extension can begin. Fluorescently tagged nucleotides will be added, when a nucleotide binds to the template string, a light source is created. So each nucleotide has it’s own color so you can determine which nucleotide is bonded to the string. After the completion of the first read, the product is washed away. The template binds to the second oligo, the beginning of a new strand is created with the fluorescently nucleotides, polymerases extends this, now we have two single strands again. Now the forward strand is removed. Now the whole process starts over again.
* **Data analysis**: this process produces millions of reads, that will represent all the fragments. Sequences from pulled samples are separated, based on the unique indices. For each sample reads with similar regions are clustered. Also the forward and reverse reads are paired, these sequences are aligned to the reference genome for the identification.

1. To investigate this outbreak, two different approaches can be used. We will continue with option 1: All ESBL-strains are sequenced using the Illumina method
   1. Explain an approach to perform a *denovo* assembly on these sequences (in detail!) Use the terms and explanation of: k-mer, contig, scaffolds

*De-novo* assembly, is a method for constructing genomes from a large DNA fragment, but in this method there is no template of the genome to compare with. A single contig is constructed from two or more overlapping reads (DNA fragments). A scaffold is a set of oriented contigs, the scaffolds may have to be reversed to get the right orientation. Some scaffolds can fill in gaps between other scaffolds, when all these scaffolds are joined together, a single chromosome is constructed. The image below shows schematically how the *De-novo* assembly works.



Counting k-mers is an essential component of many methods in bioinformatics, such as genome assembly, or the correction of sequence reads, A k-mer is a string of any length ‘k’, on the image to the right we can see a DNA string (yellow) with underneath al the k-mers (blue), this k-mer has the length of four base pairs, so we call this a 4-mer. So this DNA string consists of 11 bases, but 8 4-mers.

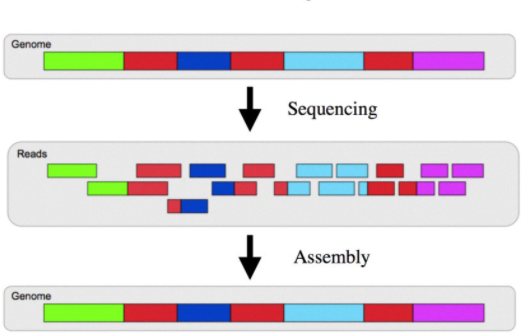
So with the *De-novo* assembly we can check if the DNA of patients B, C and D contains any similarity with the DNA of patient A.

* 1. A criteria to investigate your *De-novo* assembly is the N50; what is it and should this be high or low?

N50 is the measurement to the describe the quality of the assembled genomes that are fragmented in contigs of different lengths. So N50 needs to be the shortest contig length to cover 50% of the whole genome. Which means that half of the genome sequence is covered by contigs who are larger or equal to the N50 contig size. To calculate the N50, you need to sort all the contigs by length, go to the base in the center of the genome, get the size of this contig and so you have the N50. It should be the half of the length of the genome.

* 1. Which method are you going to use to investigate if these strains are identical and explain how this method is working (not based on SNPs).

To investigate if these strains are identical I am going to use the Cancer Cell Line Encyclopedia (CCLE), this SNP data base is generated to match the SNPs in the cancer cells against the genome.

We can use the string of DNA from patient A as reference for all the other DNA strigs. So wit whole genome assembly we can check if the DNA from the other patients are the same as the reference. Genome assembly is the process of putting multiple pieces of one genome back together in the right order to get a whole string, but then you know what the string will look like. The image below shows schematically how the genome assembly works. This is almost the same as *De-novo* assembly but here we have the reference genome, to check if the DNA is the same.

1. Another option to investigate this outbreak is to make a good reference strain of the outbreak strain and compare the others with the reference strain
   1. Mention another next-generation sequencing method which can be used to determine the sequence of an isolate which produces long reads with a high accuracy, and explain how this method works

SMRT (Single-Molecule sequencing in Real Time) is a system that involves a single stranded DNA molecule to attach to DNA polymerase. Then this DNA is sequences as the polymerase adds fluorescent complementary bases to this strand. This fluorescent is cut of and then measured or recorded. SMRT can generate reads from 10-15 kilobases long. It is easier and cheaper to assemble long reads .

* 1. Another sequencing platform, the oxford nanopore technology, also produces long reads. However, the accuracy is under debate. Explain how this method is working.

A nanopore is a hole, Oxford nanopore passes DNA or RNA through this pore and measures the changes in the current. This changes in current can identify this molecule that is passed through. The single stranded DNA is pulled through the nanopore

* 1. The outbreak strain has been sequenced using a platform which generates long reads and an assembly has been performed. Explain your workflow how you will compare the other isolates to this outbreak strain. Now a SNP-analysis is allowed.

*Clinical application in the oncology*

Exome sequencing explained: a practical guide to its clinical application [https://academic.oup.com/bfg/article/15/5/374/2240049](https://academic.oup.com/bfg/article/15/5/374/2240049?searchresult=1)

The average person inherits 3 to 4 million single-nucleotide polymorphisms (SNPs). Compared to these millions of inherited polymorphisms, there are relatively few (specifically, thousands to tens of thousands) candidate somatic single-nucleotide variants (SNVs) in the cancer genomes of adults. To have confidence that these candidates are real somatic SNVs, research­ers must have identified the large majority of inherited SNPs in the matched normal genome.

In diagnostic settings, an assortment of genes is investigated for these SNPs; the exome. This is called Whole exome sequencing (WES).

Questions

1. What is difference between whole genome sequencing and whole exome sequencing? Can you think of an advantage of using WGS above WES and vice versa?

Whole genome sequencing (WGS) is the process of determining nearly the whole genome of an organism at a single time. So also the DNA that is present in the mitochondria, or the chloroplast of the plant cells. Whole exome sequencing (WES) is a technique for sequencing only the protein-coding regions of the genes in a genome (so the exome), this is only about 2% of the whole genome. There are only 2 steps in this process, the selection of the DNA that encodes proteins, and after this to sequence this DNA. WES will cost much less than WGS and is faster. WGS is effective in rare diseases studies, because this technique is an efficient way to identify genetic variants of genes.

1. SNP calling has several sources of false positives and false negatives. Can you think of two artefacts which cause false positive SNPs?

False positives can cause major problems for the downstream processes, like the generation of genotypic platforms. The two artefacts that causes false positive SNPs are using poorly assembled fragmented reference sequences, this will lead to a increase in the number of false positive SNPs. And also longer reads cause more damage, when mismapped, so there are more mismatches, which led to an increased number of false positives.

1. Investigating cancer exome is looking for a needle in the haystack. Approximately >20.000 coding variants are identified in each individual that differs from the wildtype sequence. What kind of information do you need to investigate further?

Maybe you need to know the genome of an other type of cancer so you can check the similarities between the genomes.

1. Using NGS, the quality is important of a specific nucleotide. A Phred score is used to explain this. What is Phred score?

A Phred quality score is a measurement for the quality of the nucleotides generated by DNA sequencing. The FASTQ format encodes the Phred scores as ASCII characters alongside the read sequences. So if a base has a Phred score of 30, the chances that this base is called incorrectly are 1 in 1000.

1. Describe 2 different laboratory methods that can be used to do exon sequencing.

So for the exon sequencing you can use the Illumina DNA prep with Exome Enrichment Kit, or the exome sequencing data analysis. The Illumina is a fast, integrated sequencing library preparation workflow for a lot of applications. And the exome sequencing data provides accurate, ultra-rapid analysis of exome sequencing data and other NGS data.