Bpsda:

Plan van Aanpak

Groep 3: Rick de Graaf, Stephen Pieterman en Carla Stegehuis

## 

## Inhoudsopgave

[1.Introduction 3](#_Toc369263751)

[2.Objective 5](#_Toc369263755)

[3.Products 6](#_Toc369263756)

[4. Project organisatie 8](#_Toc369263757)

[5. Planning project 10](#_Toc369263763)

[6. Risico analyse 11](#_Toc369263764)

[7.Projectgrenzen 12](#_Toc369263815)

[8.Eindverantwoordelijke planning 13](#_Toc369263816)

[9.Ondertekening 14](#_Toc369263817)

[10.Referentielijst 15](#_Toc369263818)

# 1.Introduction

**Next-generation sequencing**

Next-generation sequencing (NGS) is the term used for a number of high-throughput DNA sequencing techniques that generate results more rapidly, at higher volume and lower cost than the heretofore most commonly used technique, Sanger sequencing. A wide variety of next-generation platforms exists, each with their own approach to sequencing. Each platform differs in read length, read quality and read costs. The platform used for this project is the Illumina platform. Illumina delivers reads with approximately 2 x 100 base pairs in length, which is relatively short in contrary to other platforms. Illumina uses a technique called “paired-end” sequencing, which uses ligate adapters that bind at both ends of the DNA fragment1. These adapters are capable of binding onto the flow cell causing the formation of a bridge (called bridge PCR). The flow cell is subequently washed with a variation of ddNTP's. This constitutes a PCR whose product will eventually form clusters on the flow cell. After every wash a photo is taken to determine what ddNTP was bound.

**De novo assembly**

As next-generation sequencing is rapidly being adopted, sequence read databases are overflowing with data. The analysis of this continues growing data flow called assembly can be aided by a reference genome. At times when no reference genome is present another technique will be applied called de novo.

**Why is de novo assembly hard?**

Missing a reference causes a lack of indication of the correctness of a performed assembly. Especially in case of two species in need of research in terms of differences. Locating mutations in DNA with uncertainty about the correctness of the assembly can be misleading during research2 . To do this as right as possible there are different software tools that can be used. Each tools has its own algorithm and parameters.

**What algorithms are used?**

There are two kinds of algorithms, Overlap-algorithm and the de-bruijn algorithm. The overlap-algorithm is generally used for sanger sequencing and the de-bruijn algorithm is more often used for next-generation sequenced data. De-bruijn’s algorithm uses less computational memory which is preferred when working with big data sets3 . There are different kinds of software tools that can be used, in tabel 1 an overview of different traits of the assemblers. It shows whether it is open-source or not, which algorithm and the “+” if the requirements are low.

Tabel 1: List with several assembly tools with URL and algorithm and requirements

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | URL | License | Algorithm | Requirements |
| Velvet | <http://www.ebi.ac.uk/~zerbino/velvet/> | Open-source | De-bruijn |  |
| Ray | <http://denovoassembler.sourceforge.net/> | Open-source | De-bruijn |  |
| Abyss | <http://www.bcgsc.ca/platform/bioinfo/software/abyss> | Open-source | De-bruijn |  |
| SOAPdenovo2 | <http://soap.genomics.org.cn/soapdenovo.html> | Open-source | De-bruijn | + |
| CLC | [http://www.clcbio.com/products/clc-assembly-cell](http://www.clcbio.com/products/clc-assembly-cell/)/ | Commercial | De-bruijn |  |
| ALLPATHS-LG | <http://www.broadinstitute.org/software/allpaths-lg/blog/> | Open-source | De-bruijn |  |
| PASHA | <http://sourceforge.net/projects/pasha/> | Open-source | De-bruijn | + |

# 

# The genomes

# In this project there will be worked with the genome of two species, the Rhagoletis cerasi and the Gonioctena quinquepunctata. These are two insects and therefore have larger genomes than bacteria. The data consists of paired-end illumina reads, these are very short and are more intensive for the computer4. These larger genomes are harder to assemble and cost more time and computational memory and cpu. The two species go by the name of the cherry fruit fly (Rhagoletis cerasi) and some kind of leaf beetle (Gonioctena quinquepunctata). They both live on a host plant and cause trouble on the agriculture.

# To compare the different sizes of genome from multiple organisms:

Tabel 2: Genome differences between several organismen

|  |  |  |
| --- | --- | --- |
| Scientific name | Common name | Size |
| *Escherichia coli* | coli bacteria | 4.6 MB |
| *Saccharomyces cervisiae* | yeast | 12.1 MB |
| *Drosophila melanogaster* | fruit fly | 139.5 MB |
| *Caenorhabditis elegans* | roundworm | 100.2 MB |
| *Homo sapiens* | human | 3 GB |
| *Sus scrofa* | pig | 2.7 GB |

# 2.Objective

**Genome sequencing at Naturalis Biodiversity Center**

The research facility of the Naturalis Biodiversity Centre is trying to understand life and in addition to that they are searching for social and economic uses to improve life by using nature and its products. Alongside this applied research they’re also working on more fundamental research, in particular speciation5,6.

Speciation is an evolutionary process of which new biological species arise. During this project the focus is on allopatric speciation, which is a more specific form of speciation. When populations of the same species become isolated from each other due to geographical changes or emigration, environmental characteristics can change. For instance the habitat has a different climate or the food sources are different of what they normally eat. This causes a genetic interchange between the two populations. When reproduction takes place, the new genetic information is passed through to the new generation and a species with a slightly different genetic information arises. They call this a ‘daughter species’.

This is what happens with allopatric speciation. Daughter species arise from a so called ‘parent species’ when geological change or emigration appears which causes a division between the parent species and part of the original population.5 6

Ecosystems supply many products and services which are indispensable for humans. Most common known products are the air we breathe, the water we drink, the meat and fish but also fruits and vegetables we get from breeding and cultivation. Though genes in former known products can indicate what will make the current product better.

Gained knowledge of the behaviour of speciation and the functioning of the ecosystems, together with the newest techniques make that scientists are continuously searching for the new ways to use the power of nature.5 6

Our contribution to the research projects of Naturalis will mainly be to the genome sequencing projects.

**The projects objective**

The main goal of the project is researching different kinds of assembly tools and comparing these to each other, in order to perform a high quality assembly on two insect species. The main focus will be on assembly tools that can run on the HPC infra of the Biodiversity Center. It is important to choose the right tools for the assembly. The second objective is to make a user guide for biologists who are not well known with data assembly. As written in the introduction, the choice of assembly tool can make the difference between a high quality assembled dataset or a low quality assembled dataset. Therefore it is important that we weigh out different assembly tools and parameters to find the one with the highest quality for the same type of data. This all adjusted to HPC of the Biodiversity Center.

We want to make sure that we find tools that will be of use for the biodiversity center the next couple of years. Other goals are to put a good result on the table with respect to the client and we will also strive to a good cooperation between the members of the group and between the group and the client.

# 

# 3.Products

There are many software tools to use, as explained in the introduction some are open-source or commercial and others are more computer friendly. Though the most have in common that they use de-bruijn algorithm, they have different parameters. Also some tools are more geared towards the method the data has been sequenced with. If we look at SOAPdenovo2 for instance, this tool is geared for the short reads of the illumina machine7. Therefore way more attractive to use than a regular tool to assemble.

PASHA another tool claims to be a fast and more lightweight assembler8, this can be quite useful when the computer you are using is not capable or is having trouble with other assemblers. But the speed may cut down the quality of the assembled data. Then there is ALLPATHS-LG which handles smaller and larger amount of data and is often used by big institutes9. Velvet is much seen but not updated since 2011 and the last real update since 201010, therefore may be outdated and compared to other assemblers not as good. Despite all this it is still a widely used assembler and along with SOAPdenovo2 it is one of the open-source assemblers which assembles with less computational memory.

There are also commercial tools on the market (CLC), the pro is that these are mostly up-to-date and have a professional interface. CLC is also known for high quality results using less memory11. They can offer more help, whereas open-source tools may require more Internet searching. To reduce the RAM usage ABySS is also a competitor as an assembler and can handle larger genomes12.

It can be stated that numerous things are important when looking for the best assembly tools, important it is to look at the speed of an assembler. If an assembler is slow it can take over a many hours before an assembly is done and there is not always enough time. Some assemblers may have a high reputation of making high quality assembly but then may use a lot memory and cpu. The computer which is at our disposal has a 192GB memory, which may not cut it. This can be a problem if most of the assembly tools need more.

The data that is sequences is paired-end, this means that it is more computationally intensive for the assemblers. The most important thing to concentrate on maybe the quality of the assembly in the end. There are different things that can be of influence on the data, like the assembly size and the amount and length of contigs.

Because we are working with eukaryotic data there are a lot of transposons in the data and can be of negative influence on the amount of contigs. Also different assemblers give different lengths with N50, if the lengths are very short and not above the N50 the data isn’t very useful12. Then it can happen that important piece of data may be thrown away because it was too short and couldn’t not be assembled right.

To prevent these things it may be better to choose assembly tools that take more larger reads, around ~100 bp, such as SOAPdenovo and ALLPATH-LG. These take larger reads using less RAM7,9.

To use all the parameters of the assemblers, it is useful to make shell scripts in order to see what parameters are used in the command line. With this it should be easier for the biologist to know what the parameters mean and how they should use it for the best results. The genomes that will be assembled are *de novo* which means that one of the products are two whole new genomes of two insects.

The products of this project will consist of different assemblies and one definitive, of the definitive one a user guide will be made. Furthermore a paper must be written with all the steps of the project, the methods and probably some programming scripts. Of this all there will be two presentations, one of the progress and a final presentation. All of this work must also be written in an logbook.

# 

**Final products**

At the end the projectgroup will deliver a few products these are;

Tabel 3: summary of the products

|  |
| --- |
| PVA |
| Progress presentation (after the first period) |
| Project report |
| (Bash) scripts and other relevant scripts in Github |
| Config files from assembly programs in Github |
| Assembly’s (output from the tool(s)) |
| Userguide on how to use the assembly tools |
| Presentation |
| Log with amounts of hours spent on this project (individual) |

**Description of the data**

This is a short description of the data that will be used during this project.

Tabel 4: list with several facts about the data which will be used during this project

|  |  |
| --- | --- |
| Description | Size or length |
| Number of files | 29 files |
| Total size  (packed gz format) | 81,8 GB |
| Total size (unpacked fastq format) estimation\* | 245,4 GB |
| Average size (packed) | 2,82 GB |
| Smallest file | 1,3 GB |
| Largest file | 7,0 GB |
| Read length | 100 basepairs long |

**Workstation specificaties**

Following some specifications of the workstation:

Tabel 5: system specifications of the workstation  
which is available during this project

|  |  |  |
| --- | --- | --- |
| Besturingsysteem | type | Linux |
| CPU | Aantal aanwezig | 8 |
|  | Snelheid | 2.40 GHz |
|  | Aantal cores | 4 |
|  | Cache grootte | 10.240 KB |
| RAM - geheugen | Grootte | 198,2 GB |

# 

# 4. Project organisatie

**Samenstelling projectgroep**

De projectgroep bestaat uit vier project leden. Binnen de groep zullen de taken van voorzitter en notulist worden verdeeld. De taak van voorzitter zal door Rick de Graaf worden uitgevoerd. De taak van notulist zal rouleren tussen de overige drie projectleden.

Alle documenten zullen zowel op Google Docs als in de gedeelde Dropbox map worden bewaard. Daarnaast hebben alle leden van de groep een Github account. Dit account zal worden gebruikt om geschreven programma’s onderling te delen. De opdrachtgever heeft ons toegang verleend tot een workstation op het naturalis. Hierop staat alle data waarmee gewerkt zal gaan worden. Uiteindelijk zullen alle assembly tools ook hierop gebruikt gaan worden.

De groep heeft twee begeleiders. De opdrachtgever en tevens begeleider vanuit het Naturalis is Rutger Vos en Jeroen Pijpe zal ons vanuit de hogeschool begeleiden tijdens dit project. Met zowel de opdrachtgever als met de begeleider van uit de hogeschool, zal elke week een vergadering ingeplant worden. De notulen hiervan zullen ook naar de projectbegeleider verstuurd worden.

**Contactgegevens**

Tabel 6: contactgegevens van de projectleden

|  |  |
| --- | --- |
| Project leden  Rick de Graaf  [rickdegraaf@gmail.com](mailto:rickdegraaf@gmail.com)  0619916176  Stephen Pieterman  [sjpieterman@gmail.com](mailto:sjpieterman@gmail.com)  0654322215  Carla Stegehuis  [stegehuis.c@gmail.com](mailto:stegehuis.c@gmail.com)  0623334404 | Opdrachtgever  Rutger Vos  [Rutger.Vos@naturalis.nl](mailto:Rutger.Vos@naturalis.nl)  Projectbegeleider  Jeroen Pijpe  [pijpe.j@hsleiden.nl](mailto:pijpe.j@hsleiden.nl) |

**Verantwoordelijkheden**

**Afspraken**

1. Alle groepsleden moeten altijd aanwezig zijn bij de vergaderingen, ingeroosterde projecturen en andere gemaakte afspraken.
2. Als een groepslid niet aanwezig kan zijn moet hij/zij hiervoor een geldige reden hebben en dit moet je minimaal 1 uur van te voren melden door middel van een sms of bellen.
3. Afspraken en deadlines moeten altijd worden nagekomen.
4. Als je vastloopt meld dit gelijk aan je groepsgenoten, deze zullen je dan zo veel mogelijk proberen te helpen.
5. Bij problemen met iemand of irritaties moet dit ook zo snel mogelijk aan de andere groepsleden worden verteld.
6. Er wordt verwacht dat elk groepslid 100% voor het project gaat.
7. Groepsleden moeten bereikbaar op mobiel zijn tussen 09:00 en 19:00 uur doordeweeks en tussen 12:00 en 17:00 uur in het weekend.
8. Binnen 24 uur moeten groepsleden reageren op mails of telefoontjes van groepsleden.

**Consequenties**

Bij lichte overtreding (bijv. 1 regel overtreden) wordt er eerst een waarschuwing gegeven.

Bij de volgende overtreding worden in overleg met de begeleider de maatregelen getroffen.

Bij ernstige overtreding (naar mening van de andere groepsleden en de begeleider) kan een groepslid uit de groep worden verwijderd

# 

# 5. Planning project

Project activities/week (The number stands for the amount of people who are working on a particular task.)

Tabel 7: first table, 2 sep until 21 okt. The left column has the activities, the remaining columns have the numbers. The number stands for the amount of people who are working on a particular task

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weeknumber | Week1 | Week2 | Week3 | Week4 | Week5 | Week6 | Week7 | Week8 |
| Start date | 2-sep | 9-sep | 16-sep | 23-sep | 30-sep | 7-okt | 14-okt | 21-okt |
| Reading project description and pdf file | 4 | 4 | 4 |  |  |  |  |  |
| Working on PVA parts |  | 4 | 4 |  |  |  |  |  |
| Deadline Pva |  |  | 4 |  |  |  |  |  |
| Researching De-novo assembly |  |  |  | 2 | 2 |  |  |  |
| Searching for assembly tools |  |  |  | 4 | 4 | 4 |  |  |
| Analyzing data on workstation |  |  |  | 2 |  |  |  |  |
| Testing assembly tools |  |  |  |  |  |  | 2 | 2 |
| Pooling the datasets |  |  |  |  |  |  | 2 | 2 |

Tabel 8: second table, 28 okt until 16 dec

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Weeknumber | Week9 | Week10 | Week11 | Week12 | Week13 | Week14 | Week15 | Week16 |
| Start date | 28-okt | 4-nov | 11-nov | 18-nov | 25-nov | 2-dec | 9-dec | 16-dec |
| Using the tool(s) on the data | 3 | 3 | 3 |  |  |  |  |  |
| Analyzing results from assembly |  |  | 2 | 3 | 3 |  |  |  |
| Locating loci-positions |  |  |  |  | 2 | 2 |  |  |
| Progress presentation preparations |  |  | 2 | 2 |  |  |  |  |
| Progress presentation (No exact day yet) |  |  |  | 2 | 2 | 2 |  |  |
| Writing documentation |  |  |  |  | 3 | 3 | 3 |  |
| Spellcheck on documentation |  |  |  |  |  |  | 2 | 2 |
| Writing project report |  |  |  |  |  |  | 3 | 3 |

Tabel 9: third and final table, 23 dec until 27 jan

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Weeknumber | Week17 | Week18 | Week19 | Week20 | Week21 | Week22 |
| Startdate | 23-dec | 30-dec | 6-jan | 13-jan | 20-jan | 27-jan |
| Writing project report | 3 | 3 |  |  |  |  |
| Preparations for the final presentation |  | 2 | 2 |  |  |  |
| Final touch on the userguide and the report |  |  |  | 3 | 3 |  |
| Deadline project 27 jan |  |  |  |  |  | 1 |
| Presentation project period 2 week 9 |  |  |  | 2 |  |  |

# 

# 6. Risico analyse

Enkel risico’s met een waarde hoger of gelijk aan 20 zijn ernstig genoeg om een voorzorgsmaatregel voor te nemen.

Tabel 10: de risico analyse van het project met waardes, als deze waarde hoger is dan 20 is het ernstig genoeg om een voorzorgsmaatregel te nemen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Omschrijving risico** | **Kans op optreden1** | **Mate van impact1** | **Kans op (tijdig) ontdekken2** | **Risico**  (product) | **Voorzorgsmaatregelen** |
| Tools zijn te zwaar voor de workstation | 3 | 5 | 1 | 15 |  |
| Het lukt niet om een de-novo assembly correct uit te voeren | 2 | 5 | 1 | 10 |  |
| Loci-posities niet te lokaliseren | 2 | 4 | 1 | 8 |  |
| **Projectmatige risico’s** |  |  |  |  |  |
| Het project loop vast (bijv. doordat student of opdrachtgever afhaakt) en wordt niet voltooid | 2 | 5 | 2 | 20 | Om niet vast te lopen zal er frequent samenkomsten plaatsvinden tussen opdrachtgever, studenten en de begeleider. |
| Zelf gestelde deadlines niet haalbaar | 3 | 2 | 1 | 6 |  |
| Product deadlines niet haalbaar | 2 | 5 | 1 | 10 |  |
| Voortgang verloren door bijv. computer crash | 2 | 5 | 1 | 10 | Tijdens het werken aan producten wordt gebruik gemaakt van Google-docs en Dropbox. |
| Het aantal voorgeschreven studie-uren wordt overschreden | 1 | 3 | 2 | 6 |  |

1) Schaal van 1 tot en met 5, des te lager des te lager de kans op optreden en lagere de mate van impact.

2) Schaal van 1 tot en met 5, des te lager des te hoger de kans op ontdekken.

# 7.Projectgrenzen

Het aantal te behalen studiepunten voor dit project bedraagt acht punten. Dit staat gelijk aan 224 studie-uren. Tijdens dit project zal zo goed mogelijk als kan aan het aantal voorgeschreven studie-uren gehouden worden. Dit zodat de student geen achterstand op loopt bij andere studie gerelateerde activiteiten. Met een totaal van 224 uur verspreid over 20 weken maakt dit 11,2 uur per week per projectlid. Uitkomend op een totaal van 44,8 uur per week voor de gehele projectgroep. Dit dient als richtlijn gehanteerd te worden. Met de mogelijkheid minder drukke periodes te compenseren met de drukkere periodes. Iedere student houdt een logboek bij waarin terug is te zien hoeveel uur besteed is aan het project.

Op het gebied van assemblies zal er beperkt worden tot de data die in eerste instantie geleverd is, afkomstig van de boorvlieg en het bladhaantje.

Binnen het bestek van het zelf schrijven of aanpassen van bestaande algoritmes gelden geen beperkingen zolang dit past binnen het aantal voorgeschreven studie-uren.

# 

# 8.Eindverantwoordelijke planning

Tabel 11: eindverantwoordelijke tabel, de linker rij is de activiteit en de rechterrij is de verantwoordelijke persoon

|  |  |
| --- | --- |
| Activities | Responsible projectmember |
| Reading projectdescription and pdf file | Rick |
| Working on PVA parts | Rick |
| Deadline Pva | Rick |
| Researching De-novo assembly | Carla |
| Searching for assembly tools | Benjamin |
| Analyzing data on workstation | Stef |
| Testing assembly tools | Stef |
| Pooling the datasets | Carla |
| Using the tool(s) on the data | Stef |
| Analyzing results from assembly | Carla |
| Locating loci-positions | Carla |
| Progress presentation preparations | Rick |
| Progress presentation (No exact day yet) | Rick |
| Writing documentation | Carla |
| Spellcheck on documentation | Carla |
| Writing project report | Stef |
| Preparations for the final presentation | Rick |
| Last checks on the userguide and report | Rick |
| Deadline project 27 jan | Carla |
| Presentation project periode 2 week 9 | Stef |

# 

# 9.Ondertekening

**Handtekeningen mits akkoord**

Jeroen Pijpe(projectbegeleider) Rutger Vos(opdrachtgever)

Rick de Graaf Stephen Pieterman

Carla Stegehuis

# 10.Referentielijst

1. “An Introduction to Next-GenerationSequencing Technology” illumina.com | http://res.illumina.com/documents/products/illumina\_sequencing\_introduction.pdf
2. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies(2011), Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, Bairong Shen\* Center for Systems Biology, Soochow University, Suzhou, Jiangsu, China. 1
3. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies(2011), Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, Bairong Shen\* Center for Systems Biology, Soochow University, Suzhou, Jiangsu, China. 5-6
4. “Software packages for next gen sequence analysis “ seqanswers | http://seqanswers.com/forums/showthread.php?t=43
5. Naturalis |www.naturalis.nl
6. “Het mysterie der mysteriën, over evolutie en soortvorming” Menno Schilthuizen, juni 2002
7. SOAPdenovo | http://soap.genomics.org.cn/soapdenovo.html
8. “PASHA: Parallelized Short Read Assembly” PASHA | http://sourceforge.net/projects/pasha/
9. ALLPATHS-LG | http://www.broadinstitute.org/software/allpaths-lg/blog/
10. “Velvet, Sequence assembler for very short reads” EMBL-EBI | http://www.ebi.ac.uk/~zerbino/velvet/
11. “CLC Assembly Cell” CLCbio | http://www.clcbio.com/products/clc-assembly-cell/
12. “ABySS” BC Cancer Agency | http://www.bcgsc.ca/platform/bioinfo/software/abyss
13. PVA [Eisen vernieuwd PvA 1.3.doc](http://blackboard.hsleiden.nl/@@BE9D7CCBF9596952EF0DC524D55052B8/courses/1/bpr4/content/_845175_1/Eisen%20vernieuwd%20PvA%201.3.doc) geraadpleegd op 11 september 2013