Differences in software performance for analyses of

High Throughput Ion Torrent DNA barcode data

Bio-informatics, Year 3 2012-2013, 1st and 2nd term, BPDSA, 8EC

Authors: Alex Hoogkamer, Thomas Bolderink and Roeben Vink

Supervisors: Dr. Sigrid Beiboer, Dr. Barbara Gravendeel and Dr. Rutger Vos

Content

[Introduction 2](#_Toc341204170)

[Research questions 3](#_Toc341204171)

[Problem 1 3](#_Toc341204172)

[Hypothesis 1 3](#_Toc341204173)

[Problem 2 3](#_Toc341204174)

[Hypothesis 2 3](#_Toc341204175)

[Goals 4](#_Toc341204176)

[Products 4](#_Toc341204177)

[Project boundaries 4](#_Toc341204178)

[Quality 5](#_Toc341204179)

[project organization 5](#_Toc341204180)

[Project agreements 6](#_Toc341204181)

[Planning 7](#_Toc341204182)

[Software Testing Flowchart 8](#_Toc341204183)

[Project flowchart 8](#_Toc341204184)

[Risk analysis 9](#_Toc341204185)

[Results 10](#_Toc341204186)

[Discussion 12](#_Toc341204187)

[Problem 1 12](#_Toc341204188)

[Hypothesis 1 12](#_Toc341204189)

[Answer 1 12](#_Toc341204190)

# Introduction

In recent years, Traditional Chinese Medicines (TCMs) have become increasingly popular in western countries. These TCMs make use of endangered species of plants or animals and may contain potentially dangerous components. This is a serious danger to the continued existence of certain plants and animals in the wild. In order to better protect these organisms and to prevent harm to humans, it is necessary to identify which species and components are present in TCMs. Since TCMs contain a mix of species they are currently analyzed with DNA barcoding using a high‐throughput technique: Ion Torrent sequencing. This saves a lot of time and money.

Careful information of the DNA barcoding regions present in these medicines is necessary.

However, without good software, it might prove very hard to make correct observations of the DNA barcodes found in TCM samples. To do this, three different packages will be compared: Geneious, CLC Bio and Life Technologies Ion Torrent software(Torrent Suite). This will be done by comparing a dataset from a single TCM sample on all three packages, so that decent assumptions can be made about each package's unique set of features and methods in relation to their function of properly identifying (illegally‐used) plant species.  
 Important functions to compare include the way data is imported, the workload necessary  
to gain proper results, how much options there are in modifying BLAST settings and how accurate the results are. Other important factors are user friendliness of the program and visual representation of the results. A program with clear, color‐coded results will rank higher than a program that only displays names in a single list of names and scores. Additionally, this project will include the creation of a pipeline, the steps of which will initially be decided upon via an ad‐hoc basis. Following review of these steps, a streamlined version will be decided upon and created.

# Research questions

## Problem 1

Why do different software packages give different blast results?

## Hypothesis 1

Different programs use different settings, perform more or less accurate BLASTs, trim the reads differently or not at all, and/or do not use the same reference databases.

## Problem 2

What would be the added value of a newly designed pipeline next to already existing software?

## Hypothesis 2

The newly designed pipeline would be for free (licenses for other software are expensive), offers more flexibility for data handling, and/or produces more accurate results.

# Goals

This project will have several goals. Because this is a educational project most of the goals will be about educating ourselves and gaining new skills.

* Writing reports in English.
* Gain experience in a 20 weeks project.
* Examining software.
* Learn how to recommend software based on own research.
* Programming a pipeline.
* Learning to work with Microsoft Word.
* Communication with teachers and clients.
* Project management.
* Learning to write a Plan of Action showing insight in assignment.

# Products

At the end of the project the following products will we handed over:

* Preliminary report consisting of our findings concerning the examined software packages and an advice on the program that performed the best during the testing.
* A pipeline that can automatically trim the Ion Torrent dataset, BLAST (P) the trimmed dataset, reliably identify organisms in the dataset and show the results in a clear and simple format.
* Manual of the pipeline, explaining how to use the program and containing further information (e.g. formats, options etc).
* Final report, this will give a detailed explanation about the methods, findings of the project and the code of the pipeline program (well documented for further use).

Furthermore, several presentations will be given. For school two presentations will be held and a third (final) presentation will be held for Naturalis.

# Project boundaries

For the first part of the project the main focus lies on discovering why three software-packages get different results while the input remains the same.  
 The second part of the project consists of building a pipeline to make analyzing the data generated by the sequencer easier and more user friendly (only display less complicated options, fixed parameters, etc).  
 It is important not to stray too far from these set tasks, if such a thing might occur we will plan a additional meeting to discuss this problem.

The duration of the project according to the modulewijzer for each member will be 224 hours distributed over a time period of 20 weeks.

# Quality

In order to maintain a high quality report the following points will be monitored:

* Correct grammar.
* Substantive level of information.
* Style.
* Lay-out.
* Correct Image subscripts.
* Relevance.
* The source of the information.

The final report is read by all members prior to submission and checked again for any mistakes.

# Project organization

Chair:

Alex Hoogkamer  
Email: [aqhoogkamer@outlook.com](mailto:aqhoogkamer@outlook.com)

Email: [S1047388@student.hsleiden.nl](mailto:S1047388@student.hsleiden.nl)   
Tel: 06-54797996

Secretaries:

Roeben Vink

Email: [roeben@zonnet.nl](mailto:roeben@zonnet.nl)  
Email: [S1047911@student.hsleiden.nl](mailto:S1047911@student.hsleiden.nl)  
Tel: 06 55533002

Thomas Bolderink

Email: [Thomas\_bolderink@hotmail.com](mailto:Thomas_bolderink@hotmail.com)  
Email: [s1047798@student.hsleiden.nl](mailto:s1047798@student.hsleiden.nl)   
Tel: 06 11283106

Project supervisors:

Sigrid Beiboer

Email: [beiboer.s@hsleiden.nl](mailto:beiboer.s@hsleiden.nl)

Dr. Barbara Gravendeel

Email: [barbara.gravendeel@naturalis.nl](mailto:barbara.gravendeel@naturalis.nl)

Dr. Rutger Vos

Email: [Rutgeraldo@gmail.com](mailto:Rutheraldo@gmail.com)

Clients:

Dr. Barbara Gravendeel

Email: [barbara.gravendeel@naturalis.nl](mailto:barbara.gravendeel@naturalis.nl)

Dr. Rutger Vos

Email:[Rutgeraldo@gmail.com](mailto:Rutgeraldo@gmail.com)

# Project agreements

The role of secretary is exchanged weekly between Roeben and the others. The presidency is fixed and thus will not be exchanged by the members. During the meetings, the progress of the overall planning is reviewed and modified if necessary. Any irregularities and/or problems of the previous week are discussed. The action points of the meeting will be mailed within 24 hours by the secretary to the project members and supervisors.  
 The project supervisor and clients will be available via email and personal contact for questions and/or uncertainties. Each week there will be a meeting with the project supervisor as scheduled on the school planning. The client will be kept informed via a weekly email containing results and what progress is made. It is expected that mails are answered as soon as possible and that appropriate help will be offered if needed.  
 For this project we will use Github. All members of the project group will have access to the Github folder used for the project. Everyone will put data, script components and parts of the Plan of Action and report in the shared Github folder when done. Every time a member makes changes to a document a version number is added to the document name. Each member will have his own part to work on the script so comments in the script will be added containing the date and time and the changes made/work done. A new document will be made containing the version number and the changes made in that version.   
 In the occasion that any member is not able to join the weekly meetings he is expected to inform the other members of his absence as soon as possible. Both the module supervisor and clients will receive a copy of the final products by e-mail.

# Planning

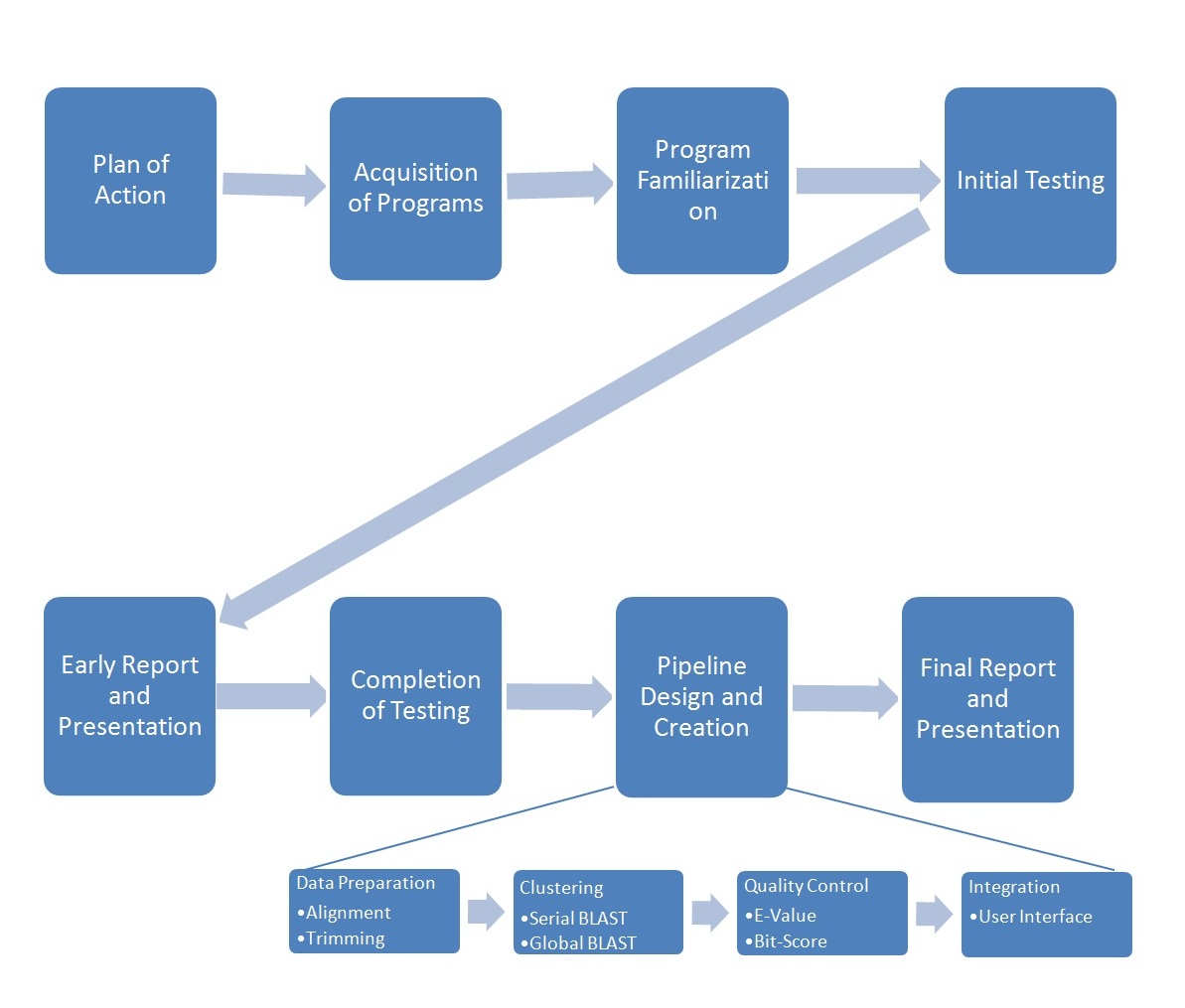
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **What** | **Person** | **Who** | **Start** | **End** | **Deliverables** |
| Plan of action | Alex | Alex,Roeben,Thomas | 24-9-2012 | 1-10-2012 | Plan of Action .docx/.pdf |
| Software Testing (data input, time, number of actions) | Roeben |  | 28-9-2012 | 9-10-2012 | Put data on Github |
| Software Testing (output, settings, validity) | Thomas |  | 9-10-2012 | 17-10-2012 | Put results on Github |
| Software Testing with incense data | Alex |  | 17-10-2012 | 30-10-2012 | Put results on Github |
| Clc Bio |  | Roeben |  |  |  |
| Geneious |  | Thomas |  |  |  |
| Torrent Suite |  | Alex |  |  |  |
| Preliminary report | Thomas | Alex,Roeben,Thomas | 2-11-2012 | 9-11-2012 | Report |
| Presentation 1 | All | Alex,Roeben,Thomas | 22-11-2012 | 29-11-2012 | PowerPoint |
| Pipeline Script (research) | Thomas | Alex,Roeben,Thomas | 31-10-2012 | 7-11-2012 | Put challenge on Github |
| Pipeline Script (programming) | Roeben | Alex,Roeben,Thomas | 7-11-2012 | 27-12-2012 | Put scripts on Github |
| Pipeline Script (validation) | Alex | Alex,Roeben,Thomas | 18-12-2012 | 11-1-2013 | Script, validation report |
| Presentation 2 | All | Alex,Roeben,Thomas | 16-1-2013 | 21-1-2013 | PowerPoint |
| Final report | Alex | Alex,Roeben,Thomas | 14-1-2013 | 28-1-2013 | Report |

Table 1: Planning.

Incense data: Naturalis is currently working on an new Ion Torrent data set acquired from incense with fully known ingredients. When the data set is ready it will be passed to us. Earlier testing will be done with data provided of TCMs with unknown ingredients and contamination.

## Software Testing Flowchart

## Project Flowchart



# Risk analysis

Main probable causes of project failure include the unavailability of licenses or the sudden incorporation of a large update halfway through the project, which would in theory require a complete redo of that part of the research process.  
 Smaller concerns include hardware failure (e.g. some computers do not have enough memory for some of the tasks of the software) and other mechanical failures. Personal injury during the project might also result in loss of knowledge and results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Risk | Description | Chance of appearance | Impact | Chance of discovery | Risk factor |
| 1 | Inadequate programming experience | 3 | 2 | 1 | 6 |
| 2 | Problems due to hardware failure | 3 | 3 | 1 | 9 |
| 3 | Bad communication/communication error | 3 | 3 | 2 | 18 |
| 4 | Not following up on agreements | 2 | 3 | 2 | 12 |
| 5 | Missing deadline | 3 | 5 | 4 | 60 |
| 6 | Project members quitting the project | 2 | 4 | 2 | 16 |
| 7 | Low quality products | 2 | 4 | 2 | 16 |

Table 2: Risk analysis.

For all risk factors 1 is low and 5 is high. Three factors are taken into account: chance of appearance , impact and chance of discovery. The risk factor is calculated by multiplying the three individual values . Anything under 10 constitutes a minor risk. A score between 11 and 29 is considered a risk that warrants small preventive measures. Scores above 30 are serious risks that need to be contained as much as possible (see Appendix 1).

# Materials and Methods

We compared three software packages with a first IonTorrent dataset provided by Naturalis by

running a default analysis and scoring the identifications obtained.

|  |  |  |  |
| --- | --- | --- | --- |
| Name Program | Developer | Version | Platform |
| Clc Bio Workbench | Clc Bio. | 6.7.1 | Linux/Windows |
| Ion Torrent suite | Life Technologies. |  | Linux |
| Geneious | Biomatters Ltd. | 6.0.3 | Windows |

We compared user friendliness, program functionality, integration with online applications and databases and output provided in terms of E-value and similarity.

**About ION Torrent**

The Torrent suite does not have options or functions to trim or BLAST sequences. There are plug-ins available yet none of them will add the ability to trim. Torrent suite has a different focus than the other programs and requires a different level of user skill. It's main purpose is to process the data an ION Torrent cycle delivers. Its main features include searching for gen variations and assembly.

The software itself is at face value simple to use but can be a bit tricky at times. It comes as a virtual machine which needs to be set-up before it can be used. While someone with experience in Linux and virtual machines in general will find it easy to do, it can be daunting for someone with no experience.

**About Geneious**

Installing Geneious is a comforting experience that does not require much experience on the part of the user. It is easy to work with and the menus are usually clean and simple. Tutorials are readily available and are one of the first things one sees when they boot up the program. These tutorials are partitioned in segments with exercises with varying difficulty, ranging from a Beginner tutorial to ones useful to any Advanced user. It can be concluded is not hard to get used to working with Geneious as long as one follows these tutorials. They are available online as well.

**About CLC Genomics Workbench**

CLC Genomics Workbench, hereafter referred to as Clc Bio, is extremely easy to use and definitely seems to be designed by a team of professionals. It breezes through its tasks, and equally positive statements can be made about its stability.

Unlike Geneious Clc does not grant you a tutorial when one first starts it. However, the argument can be made that Clc Bio is essentially the non-freeware version of Geneious, as the two programs are strikingly similar down to the interface and options menus. If one is familiar with the one, they can nearly be guaranteed to be somewhat familiar with the other.

When using BLASTn on Clc Bio for the first time, the default settings will be equal to the ones you find on the NCBI website, with one important difference. The score for a mismatch is -3 instead of -2. This might cause differentiating results.

**About the pipeline program**

This project included designing a pipeline that will try to provide an enhanced, customized solution for the type of problem that is generally encountered when analyzing Ion Torrent data from TCMs, i.e. finding different results when using different commercially available software, paying a lot of money for a license and still not knowing if CITES species are present in a dataset without doing additional analyses. By creating the pipeline ourselves we could eliminate the need to pay for a 10,000 dollar license and allow an untrained user to make use of the same functions and gain accurate results similar to those found in commercially available programs.

Specifically, this pipeline must be able to automatically trim the Ion Torrent dataset(s), Blast these and then reliably identify organisms in the dataset and identify these as CITES species. The results are to be shown in a clear, concise and simple format. The user gives the program the location of a fastq file. After specifying all desired parameters, the sequences are trimmed, and run through the BLAST database. After that, a quality check is done that filters hits with above 97% coverage and 95% identity.

# Results

Below the presence/absence of the three packages compared is provided for the different aspects mentioned above (Table 3).

|  |  |  |  |
| --- | --- | --- | --- |
|  | CLC BIO | Geneious | Torrent Suite |
|  |  |  |  |
| In-Program download function | + | + | - |
| In-Program sequence searcher | + | + | + |
| De Novo Assembly | + | + | +- |
| Epigenomics | + | + | - |
| Transscriptomics | + | + | +- |
| Resquencing analysis | + | + | + |
| Tracking Tools | + | + | - |
| NGS Core Tools | + | + | - |
| BLAST | + | + | - |
| Molecular Biology tools | + | + | - |
| Trim Primers | + | + | - |
| Trim on Phred quality | + | + | - |
| Classical Sequence | + | + | + |
| Aligning to reference genome | + | + | + |
| Output in FASTQ | + | + | + |
| Output in SAM/BAM | + | + | + |

Table 3: Comparison of functionality between CLC Bio, Geneious and Torrent Suite. Plus (+) means the function is present in the software, minus (-) means it is not present and plus minus (+-) means it is not present by default but can be activated by way of plug in.

For a fair comparison we first looked up the default settings for each program (Table 4).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CLC BIO | Geneious | Torrent Suite | NCBI BLAST |
| Match score | 1 | 1 | N/A | 1 |
| Mismatch score | -3 | -2 | N/A | -2 |
| Read size | 11 | 28 | N/A | 28 |
| Gap existence | 5 | Linear | N/A | Linear |
| Gap extension | 2 | Linear | N/A | Linear |
| Filter low complexity | + | + | N/A | + |
| Algorithm | BLASTn | megablast | N/A | megablast |
| Database | NCBI Nucleotide Collection(nr) | NCBI Nucleotide Collection(nr) | N/A | NCBI Nucleotide Collection(nr) |

Table 4: Comparison of the default setting of the programs used for BLAST. Note that Torrent Suite is empty because it lacks the ability to use BLAST.

We then performed a global BLAST on all 22 contigs in the Ion Torrent sample provided (Table 5; Appendix 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MID 6 contig | Length of consensus sequence | Accession | Description | Bit score | Query coverage | E-value | Max identity |
|  |  |  |  |  |  |  |  |
| 1 | 212 | [JX270570](http://www.ncbi.nlm.nih.gov/nucleotide/411031104?report=genbank&log$=nucltop&blast_rank=1&RID=APZ5EJCN015) | *Debaryomyces* sp. | [387](http://blast.ncbi.nlm.nih.gov/Blast.cgi#411031104) | 100% | 2,00E-104 | 99 |
| 2 | 191 | [JF907990](http://www.ncbi.nlm.nih.gov/nucleotide/344332949?report=genbank&log$=nucltop&blast_rank=1&RID=APZ5EJCN015) | *Entoloma madidum* | [353](http://blast.ncbi.nlm.nih.gov/Blast.cgi#344332949) | 100% | 2,00E-94 | 100 |
| 3 | 160 | [GU370763](http://www.ncbi.nlm.nih.gov/nucleotide/289064262?report=genbank&log$=nucltop&blast_rank=1&RID=APZ5EJCN015) | Uncultured fungus clone | [296](http://blast.ncbi.nlm.nih.gov/Blast.cgi#289064262) | 100% | 4,00E-77 | 100 |
| 4 | 151 | [JQ700297](http://www.ncbi.nlm.nih.gov/nucleotide/378942770?report=genbank&log$=nucltop&blast_rank=1&RID=APZ5EJCN015) | *Piptoporus betulinus* | [279](http://blast.ncbi.nlm.nih.gov/Blast.cgi#378942770) | 100% | 4,00E-72 | 100 |
| 5 | 215 | [JX171169](http://www.ncbi.nlm.nih.gov/nucleotide/401878784?report=genbank&log$=nucltop&blast_rank=1&RID=APZ5EJCN015) | *Cryptococcus magnus* | [396](http://blast.ncbi.nlm.nih.gov/Blast.cgi#401878784) | 99% | 4,00E-107 | 100 |

Table 5: Selection of the first 5 contigs on which a BLAST search was performed. The search was done with the BLAST implementation on the NCBI website using default values. For all results see Appendix 2.

We then performed a global BLAST on all 22 contigs in the Ion Torrent sample provided with Clc Bio (Table6; Appendix 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MID 6 contig | Length of consensus sequence | Accession | Description | Bit score | Query coverage | E-value | Max identity |
| 1 | 208 | HF545662 | *Debaryomyses hansenii* | 412,82 | 97 | 4,34E-112 | 100 |
| 2 | 191 | JF907990 | *Entoloma madidum* | 379,12 | 100 | 6,05E-102 | 100 |
| 3 | 160 | GU370763 | Uncultured fungus | 270,09 | 100 | 4,01E-069 | 100 |
| 4 | 151 | JQ700297 | *Piptoporus betulinus* | 299,83 | 100 | 4,49E-078 | 100 |
| 5 | 214 | JX171169 | *Cryptococcus magnus* | 424,72 | 98,6 | 1,14E-115 | 100 |

Table 6: Selection of the first 5 contigs on which a BLAST search was performed. The search was done with the BLAST implementation of CLC Bio using default values. For all results see Appendix 2.

We then performed a global BLAST on all 22 contigs in the Ion Torrent sample provided with Geneious (Table7; Appendix 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MID 6 contig | Length of consensus sequence | Accession | Description | bit-score | Query coverage | E-value | Max identity |
| 1 | - | - | - | - | - | - | - |
| 2 | 191 | JF907990 | *Entoloma madidum* | 353,83 | 1 | 2,36E-94 | 1 |
| 3 | 160 | GU370763 | Uncultured fungus | 296,584 | 1 | 4,04E-77 | 1 |
| 4 | 151 | AY966448 | *Piptoporus betulinus* | 279,964 | 1 | 4,07E-72 | 1 |
| 5 | 214 | FN400938 | *Cryptococcus magnus* | 396,303 | 1 | 3,87E-107 | 1 |

Table 7: Selection of the first 5 contigs on which a BLAST search was performed. The search was done with the BLAST implementation of Geneious using default values. For all results see Appendix 2.

Clc Bio does not have standard functionality for megablast, so we decided to modifty the NCBI website to Clc Bio’s settings. This was not a problem because we already had a Geneious comparison that used megablast in reserve. Thus we performed a global BLAST on all 22 contigs in the Ion Torrent sample with NCBI BLAST, equaling the settings to ClC Bio, a 1,-3 match/mismatch score.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MID 6 contig | Bit score | Accession | Description | query coverage | E-value | Max identity |
| 1 | 412 | JX270570.1 | *Debaryomyces* | 0,98 | 4E-112 | 1 |
| 2 | 379 | JF907990.1 | *Entoloma madidum* | 1 | 6E-102 | 1 |
| 3 | 317 | GU370763.1 | Uncultured fungus clone | 1 | 2E-83 | 1 |
| 4 | 299 | JQ700297.1 | *Piptoporus betulinus* | 1 | 4E-78 | 1 |
| 5 | 424 | JX171169.1 | *Cryptococcus magnus* | 0,99 | 1E-115 | 1 |

Table 8: Results after adjusting default settings of NCBI BLAST set to the settings of Clc Bio It is notable this table is similar to table 6.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MID 6 contig | bit-score | Accession | Description | Query coverage | E-value | Max identity |
| 1 | 378.19 | JN942657 | Debaryomyces prosopidis | 100% | 1,08E-101 | 99,50% |
| 2 | 345,73 | JF907990 | Entoloma madidum | 100% | 6,37E-92 | 100% |
| 3 | 289,825 | GU370763 | Uncultured fungus clone | 100% | 4,29E-75 | 100% |
| 4 | 273,595 | AY966448 | Piptoporus betulinus | 100% | 3,30E-70 | 100% |
| 5 | 387,207 | JX164076 | Cryptococcus sp. | 99,,53% | 2,08E-104 | 100% |

Table 9: Results after adjusting default settings of Geneious to the settings of Clc Bio. This table is also notably similar to table 6.

We compared the Phred scores of the different programs.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clc Bio | Geneious | Torrent Suite |
| Quality | 0.05 | 0.05 | N/A |

Table 10: Phred scores used as trimming settings.

# Discussion

Several differences were found by us between the different software packages compared. First of all not all programs supported the same features. Clc Bio and Geneious were much more comprehensive as compared with Torrent Suite. Secondly, we found Clc Bio to be deviant in default settings, leading to a significantly different result with the global BLASTs. We experimentally tested this by adjusting the default settings of Clc Bio with those of Geneious, which resulted in similar, but not exact results. Thirdly, with Torrent Suite one cannot trim reads or access the internet.

Interesting to note is that Geneious’ results differ not because the output is different, but because of the way the output is listed. It is tough to sort the results in a way similar to NCBI and Clc Bio. Sometimes the results have multiple organisms with exactly the same scores. These are then listed alphabetically on NCBI BLAST and Clc Bio, Geneious sorts on bit score, but doesn’t do a secondary sorting on alphabetical order, shuffling the results to appear different.

# Conclusions

## Problem 1

Why do different software packages give different results?

## Hypothesis 1

Different programs use different settings, perform more or less accurate BLASTs, trim the reads differently or not at all, and/or do not use the same reference databases.

## Answer 1

The reason why different software gives different results is because the settings for BLAST in Clc Bio and Geneious are different. Ion Torrent does not give results at all since it does not have functionality capable of performing BLASTs and cannot trim. The programs that are capable of this however, use the same databases but with different default settings for the global BLASTs.