Industrial Training Report

Muhammad Farhan bin Mohd Tahir

3 January 2017

Contents

Company

Company Background

- Novocraft Technologies SDN BHD

NOVOCRAFT
TECHNOLOGIES SDN BHD

- Culmination of over 8 years of experience in software development and performance tuning, bioinformatics research

Company

Company Background

- Mision
  - To be a globally recognized provider of innovative and accurate bioinformatics tools and services for discovery, primarily in the next-generation genomics (NGS) space.
- Vision
  - Intelligent software that automates the complex data processing pipelines required to conduct genomics analysis.
  - Ensuring better predictive power from NGS data in personalized genomics & therapies.
  - Intelligent software that automates the complex data processing pipelines required to conduct genomics analysis.

Company

Company Organization's Chart

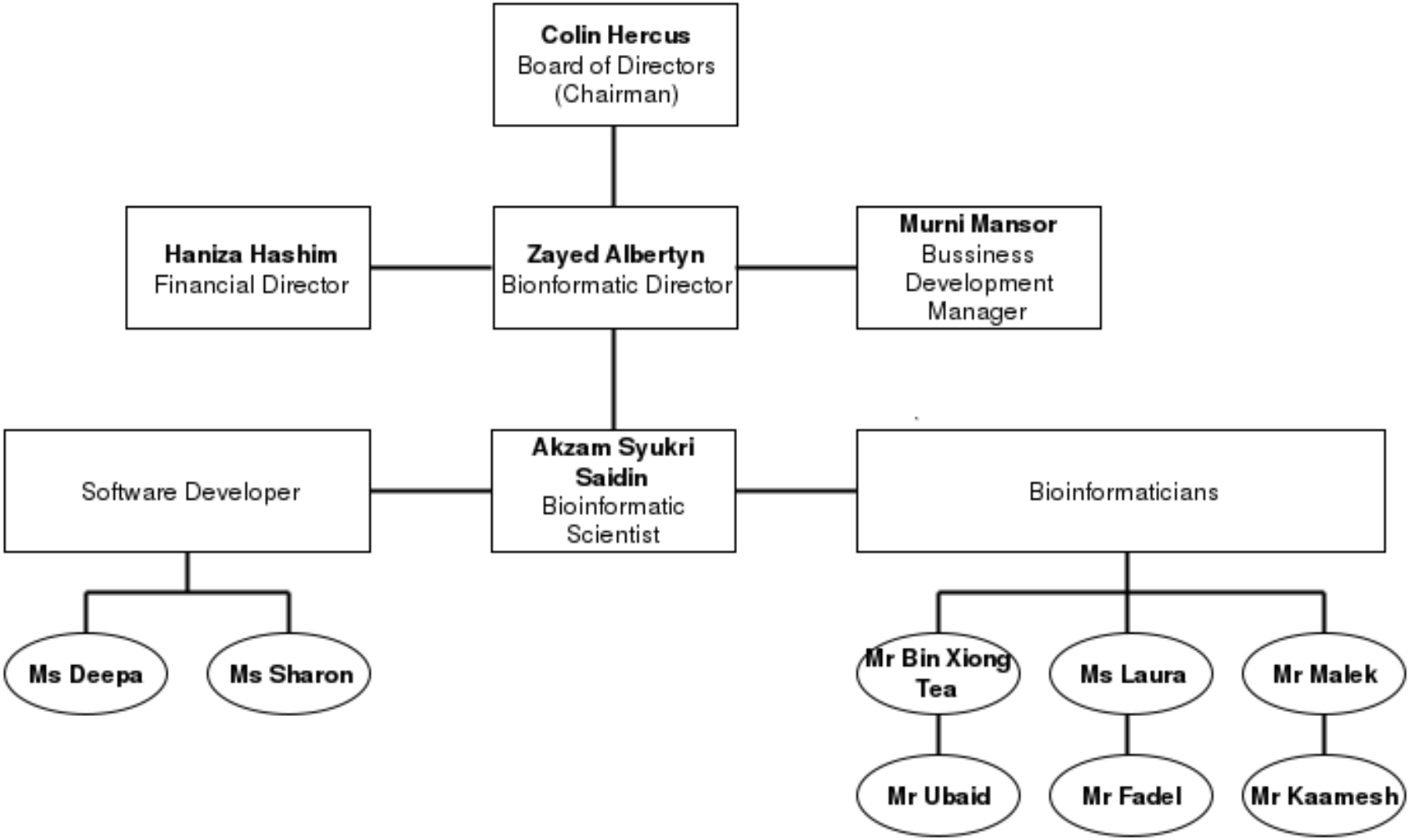


Figure 1:

Company

Company's Services and Product

- Services:
  - Bioinformatics Consultancy
  - Pipeline and Software Development Services
  - Bioinformatics Contract Research Services
- Products:
  - NovoSort
  - NovoAlignCS
  - NovoAlign
  - NovoWorx (Latest released tools)

Project's Background

- Problem Statement:
  - Illumina Sequencing technology is a tool that used by the biologist/biological researcher to extract the genome sequence from any biological sample. However, the output from Illumina sequencing may contain sequences that are not originally found in the sample, but chemically synthesized, such as, the Adapter Sequence which is used to link the ends of two other DNA molecules. In order to get a clean genomic sequence, the adapter sequence needs to be identified and removed. A program has previously been developed in Novocraft to overcome this issue, Nevetherless, there are room for improvement in terms of the program's result accuracy and executional speed.
- Project's Assigned:
  - To develop a program to find an adapter sequence from two paired-end FASTQ file (input file from user).
- Program's Name:
  - PEAdapterFinder program (**Pair Ends Adapter Finder**)
- Project's Initiator:

Project's Aim and Objectives

- Project's Aim:
  - to develop program to find an adapter sequence from two paired-end FASTQ file (input file from user).
- Project's Objective:
  - To implement the code that can make the program's result more accurate than before.
  - To implement the code that can increase the speed of program's execution time.

Term's Definition

- FASTQ File
  - Line 1: Begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
  - Line 2: Raw DNA sequence.
  - Line 3: Begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
  - Line 4: Encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

Term's Definition

FASTQ File

- 4-Line FASTQ file example

```
@HWI-EAS179_0001:5:1:7:1350#0/1
CGTGATTTTTGTATGTAAGTTTTTTCTTTTGAAGTNNTTTGTTCAGTGAAACAANAATTGNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1350#0/1
HHDH@HHHHGHHHHHFF>FFEFFFFFFFFFEE@E>##A?AFA:GC15C##########################
@HWI-EAS179_0001:5:1:7:72#0/1
GGTGATCCTCAAAGTCAAACTCGACTAGTTCGGAANNGGTCTTATTCTGTCAATNTCATANNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:72#0/1
HHHEEDGHHHCAGE4HGHHHGHAGE8:HEDFBF3<##A7.58=,1BECD/F0.B######################
@HWI-EAS179_0001:5:1:7:746#0/1
CGTGATCTTATATCATGGAGTTGGGCAGTGCATACNNTTCTTCGTGTATTTATTNATTTCNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:746#0/1
HHGHHHGHHCDAD;5DGGADGDGGGFAFFFC>CC9##=<2CCCC<D>7D<<5@@######################
@HWI-EAS179_0001:5:1:7:1657#0/1
CCTGATACGCAGACAATTTAATGCTAGAAATTTTCNNTAATTGGTCCTTGATCANTGTCGNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1657#0/1
HHDEHHFHHHHHHHH5GGG<DFFFF=D;>ACACCC##@6:>@@C:FFCFF9@@?#@>;A@################
@HWI-EAS179_0001:5:1:7:1002#0/1
CGTGATGATCAAACAATTCAGCATGACTCTTTGGANNTCCATATCCAGTCCCTTNCGAATNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1002#0/1
CGTTATHHBHFA>G;ADDFDGGGEE.G8GD=9:C5##<=><9?<CFFF?FC7?######################
@HWI-EAS179_0001:5:1:7:14#0/1
AATTATTGGACTTTGTGTGTAATTTATAACAAGTCNNCCATTTTCCTGTAGGGTNTATTCNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:14#0/1
GFGAGA>DDDGGGDGD?FFF.7<???478;155+5##=08@=94+CEC6A<C@@#A3?##################
@HWI-EAS179_0001:5:1:7:1350#0/1
CGTGATTTTTGTATGTAAGTTTTTTCTTTTGAAGTNNTTTGTTCAGTGAAACAANAATTGNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:1350#0/1
HHDH@HHHHGHHHHHFF>FFEFFFFFFFFFEE@E>##A?AFA:GC15C##########################
@HWI-EAS179_0001:5:1:7:72#0/1
GGTGATCCTCAAAGTCAAACTCGACTAGTTCGGAANNGGTCTTATTCTGTCAATNTCATANNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:72#0/1
HHHEEDGHHHCAGE4HGHHHGHAGE8:HEDFBF3<##A7.58=,1BECD/F0.B######################
@HWI-EAS179_0001:5:1:7:746#0/1
CGTGATCTTATATCATGGAGTTGGGCAGTGCATACNNTTCTTCGTGTATTTATTNATTTCNNNNNNNNNNNNNNNN
+HWI-EAS179_0001:5:1:7:746#0/1
HHGHHHGHHCDAD;5DGGADGDGGGFAFFFC>CC9##=<2CCCC<D>7D<<5@@######################
```

Line 1: identifier

Line 2: Raw DNA sequence

Line 3: separator '+'

Line 4: Qualities Value

Term's Definition

FASTQ File

- Multi-line FASTQ file example

```
@HWI-1KL125:208:H84HDADXX:2:1101:1550:2180 1:N:0:CTCTCTACCTCTCTAT
AGGCAGGAGGGCTCACCTGGCACTGGCACGTGTAGCAGGGGTCTGGTGGGGCTAGCTCAGATCCCAGCAGGCCCTCTGAACAGTTACT
CAAGGCCCTGTC
+
?=?DDBDDDDDHHIIFH>ABEFGF@FGGIB8DGGEDF?BDG7<=9@C@FBHEBEBB;66;;6>@CCC@C<A=?@AA?AAC@@<:ACA:
@:A@ABBB8A@C
@HWI-1KL125:208:H84HDADXX:2:1101:1712:2201 1:N:0:CTCTCTACCTCTCTAT
GCCCATGTCACCCCATTTGGGCAAAGAGCCCCACTTTGTTTCACCTCATGGTGTCCTACTCCAAGCCACACATGTCAAAAGAAATATATCAGCACATTCA
+
+1?DDABACFDFBH@??CFCGIECGI@;F;EHEGIIGIHHIIFEEGEHGII)8C>=@CEGEIIC@D-7AAB?>;;6;AC>9?B?CCDDDDDC>>@?ACCC
@HWI-1KL125:208:H84HDADXX:2:1101:1613:2213 1:N:0:CTCTCTACCTCTCTAT
GCTATGTGCTGAAGCCAGACTTCCTGCGTGATATCCAGAGTTCTTTCCACCCTGAGAAGCCCATCAGCCCTTTCAAAGCCCAGACTCTCTTAATCCAGGT
+
==?B?DDBBHHHF@EGH@HG<FHHHEBFF:GBCHIHIIGCBBGH9BEGBFIIDDHDHH@@GEC3CDCGGEE33;;?DDECCCA?@CCCCCC:>3>;CC9:
@HWI-1KL125:208:H84HDADXX:2:1101:1728:2228 1:N:0:CTCTCTACCTCTCTAT
TGCCAGAACAGCTTAAGAACATACCAGCAAAAGGAAGCAGTCTGTGCCACATGGTTCTGGTTCAATCTTGATTTCTTTATTCTTGCGTTTATATACATTA
+
@@CFFFFFHHHHHJJIIIJJJJIJJJJJJIJJJJJIIJGHHAGFEGIIIIEHGJIJIGHGIIGIIGIJII>HHHHHHHHHFFFFFFDDDDDDEEFEDEFD
@HWI-1KL125:208:H84HDADXX:2:1101:1784:2100 1:N:0:CTCTCTACCTCTCTAT
GTTTTTATCCCCATTTTATAGCTAAAGGGACAGAGGCTTAGAGAAATCATGAAATTCGCCTTTGTCATACATCTTGTACATTGCGGAC
CTGGGAATCTGC
+
@@@FFFFFHHGHDFEEHGGIGI<CHHGIEHIJC@CCHIGEACDHIGEHIIGHDGDHEFGBAHGE@EHIGCECHH=>;?@DBD@CDD??
@CD9@9>A>AA@
@HWI-1KL125:208:H84HDADXX:2:1101:1962:2139 1:N:0:CTCTCTACCTCTCTAT
ACTTTTGAGCTTCCTGAGGAATTTAGATTTTTGCTTGTTTCTTTTGTTTTTCAGACTCATGAAATGGCCACAGATGATAAGACGTCCC
CAACCTGTCTCT
+
@@CFFDEFHHFGHJIJHJJGHGHIIIGIIJIJIIIJJGIIJIIIJJIJFGEHGIGECHGHGIEHEHIIJJJJHHEHGHFFEFFCCDDD
DDDBDDDDDCCD
@HWI-1KL125:208:H84HDADXX:2:1101:1908:2143 1:N:0:CTCTCTACCTCTCTAT
GGCCTGGGGCTCCACCCTAAGGCCTATGGGCTGCCAGTCCTCTCCCTTGCAGTTACCTGGGATGGCGCCACCACATCATCCTCGCTGGACTCTGCTGCGG
+
C@CFFFFDFHDHHGIIJHHIGGEHGIIEIGIGGIJEGGGHIJGHJIGIJJIJAHIIIGIIIIEEACBFCDDDDDDDCDDDCDDDDDDBBBDDDDDCDABB
```

Raw DNA line

Qualities
Value line

Term's Definition

- **Adapter Sequence**
  - Adapter Sequence is a short, chemically synthesized, double stranded DNA molecule. which is used to link the ends of two other DNA molecules.
- **Needleman-Wunsch Algorithm (NW)**
  - NW is an algorithm used in bioinformatics to align two DNA sequences.
- **Consensus Sequence (CS)**
  - CS is used to find the highest frequency of DNA base(A/T/G/C) that need to be assign at certain place.

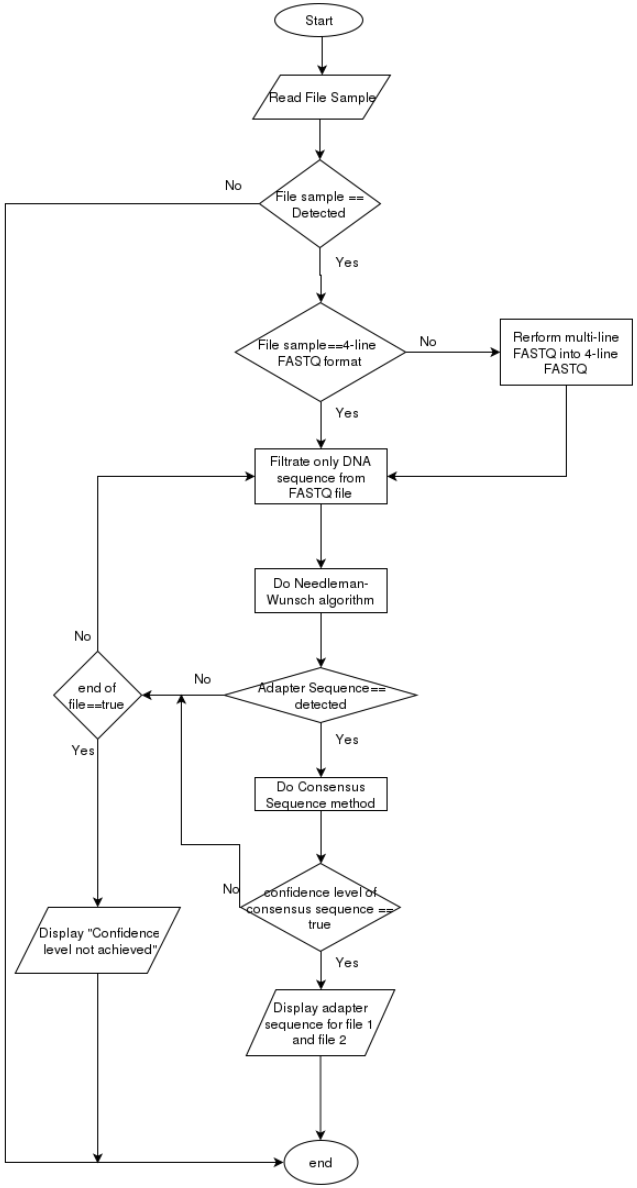
Programming Language and Tools Used

- Programming Language
  - C++ language
- Tools

Tools	Description
NetBeans IDE -version 8.1	Provide framework to develop C++ program.
GPERFTOOLS (PPROF) -version 2.0	Provide tools that can help which line in the code that consume the most time during executional of the program.
GNU gdb -version 7.4.1-debian	Tools that can help in trace if there any error during executional of program such as *Segmentation Fault*

Program's Flowchart

- Flowchart of PEAdapterFinder in finding adapter sequence.



Method Implemented

Improved Code

- Accuracy

Original Code	**Function**	Improved code
There is no code to convert multi-line FASTQ file into 4-line FASTQ file. So when the input file is multi-line FASTQ file, the output may incorrect.	**Convert/Reform multi-line FASTQ file into 4-line FASTQ file**	Code to convert from multi-line FASTQ file into 4-line FASTQ file is developed. So then, when the program want to filtrate only DNA sequence from FASTQ file, it just choose the string from line 2, 6, 10, …, (N_{BEFORE}+4)
Original code compare each character in each line whether the line only had bases (A/T/G/C/N) to be assign as DNA sequence. But if in case when the input file is multi-line FASTQ file, for example, there is two line which represent only ONE DNA sequence, the program will assume it is two different	**Filtration only DNA sequence from 4 line FASTQ format**	Improved code use line count to filtrate the only DNA sequence line from 4-line FASTQ file. This may be accurate than original code because the program will assign the string at line 2, 6, 10, …, (N_{BEFORE}+4) as DNA sequence which it is confirmed as DNA sequence's line.

Method Implemented

Improved Code

- Speed

Original Code	**Function**	Improved Code
Original code comparing **each** character in **each** line for **both** file to assign whether the line is DNA sequence or not. This may increase in program's executional time since the program need to do comparison two times for each line (because have 2 file input): First comparison: **Each** character in **each** line in file 1. Second comparison: **Each** character in **each** line in file 2.	**Filtration only DNA sequence from 4-line FASTQ file**	Improved code just choose only line 2, 6, 10, …, (N_{BEFORE} +4) as DNA sequence at both file. *(If the input file is multi-line FASTQ file, the program will reformat it into 4-line FASTQ file. Then the program use above method in order to filtrate DNA sequence from FASTQ file)* This may reduce program's execution time since the program do not need to compare each character for each string for both file.

Method Implemented

Improved Code

- Speed

Original code implementing matrices and switch statement in order to determine the match/mismatch between 2 bases from file 1 and file 2 respectively. Since matrices need the usage of an array, this may increase the program's executional time.	**Determination of match or mismatch between two bases**	Improved code implementing if-else statement rather than switch statement. This code just store the result of match/mismatch in a single variable rather than use array and matrices. This may reduce program's executional time.
In original code, there is the function (user-define function) that code by the developer (Rayan) to find the maximum score between 3 score. User-define function may increase the program's executional time.	**Finding maximum score to be filled in score-matrix table**	In improved code, built-in function are implemented rather than using user-define function to find which score is higher between 3 score. Built-in function: *Max(max(SCORE1, SCORE2), SCORE3);* This may reduce program's

Achievements

Result's Accuracy

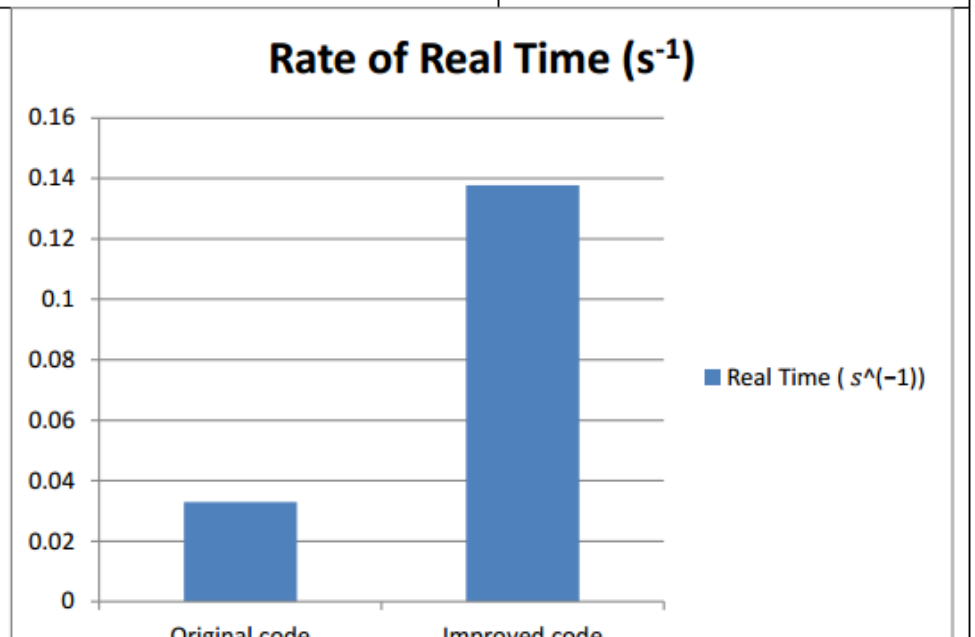
- File Sample
  - Four-line FASTQ file: fourLine1.fastq & fourLine2.fastq
  - Multi-line FASTQ file: multiLine1.fastq & multiLine2.fastq

Characteristics	Original Code	Improved Code
4-line FASTQ file Expected Output: Adapter 1: ATTACGAAATAATCGGATTC Adapter2: ATTACGAAATAATCGGATTC	Output: Adapter1: ATTACGAAATAATCGGATTC Adapter2: ATTACGAAATAATCGGATTC • Percentage of Accuracy: 100 %	Output: Adapter1: ATTACGAAATAATCGGATTC Adapter2: ATTACGAAATAATCGGATTC • Percentage of Accuracy: 100%
Multi-line FASTQ file Expected output: Adapter Sequence from file 1: ATTACGAAATAATCGGATTC	Output: Adapter1: AGCCTAATCGGGATCTCATC Adapter2: CAGTAATCTTACTGTGATCC *Base with line means the base is different from expected base	Output: Adapter1: ATTACGAAATAATCGGATTC Adapter2: ATTACGAAATAATCGGATTC • Percentage of

Achievements

Program's Speed

- File Sample: sample1.fastq & sample2.fastq

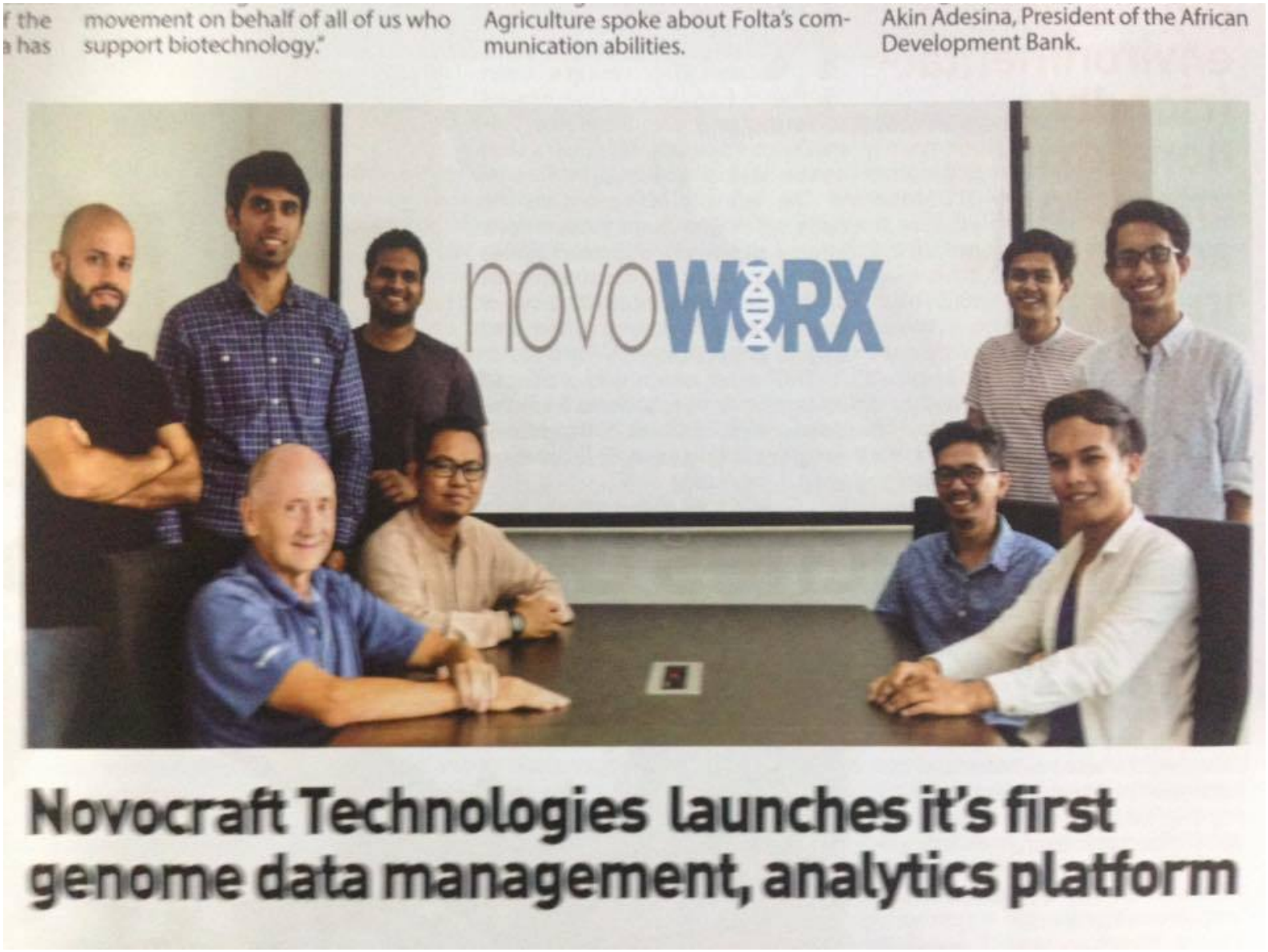
Characteristics	Original Code	Improved Code
Real time of program's executional (s)	30.315	7.261
Rate of Real Time of program's executional (s^{-1})	0.032987	0.1377221
Graph of code pattern vs rate of real time of program's executional		

Demonstration

Conclusion

- Aim of this project had been achieved
- Objective of this project had been achieved

Conclusion



f the movement on behalf of all of us who Agriculture spoke about Folta's com- Akin Adesina, President of the African
a has support biotechnology." munication abilities. Development Bank.

Novocraft Technologies launches it's first genome data management, analytics platform

Figure 2:

Conclusion

Figure 3:

Conclusion

Figure 4: