

**IMPLEMENTASI *PACKAGE BIOCONDUCTOR* PADA *SOFTWARE R*
UNTUK OLAH DATA HASIL *GENOTYPING BY SEQUENCING*
PADA TANAMAN CABAI (*Capsicum annuum*)**

TUGAS AKHIR

**Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana
Jurusan Statistika**



Hawila Sonya Savitri

14 611 176

**JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018**

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul : Implementasi *Package Bioconductor* pada
Software R untuk Olah Data Hasil
Genotyping by Sequencing pada Tanaman
Cabai (*Capsicum annuum*).

Nama Mahasiswa : Hawila Sonya Savitri

Nomor Mahasiswa : 14 611 176

**TUGAS AKHIR TELAH DIPERIKSA DAN DISETUJUI UNTUK
DIUJIKAN
Yogyakarta, 1 Agustus 2018**

Pembimbing I



Dr. techn. Rohmatul Fajriyah, S.Si., M.Si.

Pembimbing II



Muryanto, S.P., M.Si.

HALAMAN PENGESAHAN

TUGAS AKHIR

**IMPLEMENTASI *PACKAGE BIOCONDUCTOR* PADA *SOFTWARE R*
UNTUK OLAH DATA HASIL *GENOTYPING BY SEQUENCING*
PADA TANAMAN CABAI (*Capsicum annuum*)**

Nama Mahasiswa : Hawila Sonya Savitri

Nomor Mahasiswa : 14 611 176

**TUGAS AKHIR INI TELAH DIAJUKAN
PADA TANGGAL 1 AGUSTUS 2018**

Nama Penguji

Tanda Tangan

1. Husna Nugrahapraja, Ph.D.

: 

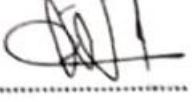
2. M. Hasan Sidiq Kurniawan, S.Si., M.Sc.

: 

3. Muryanto, S.P., M.Si

: 

4. Dr. techn. Rohmatul Fajriyah, S.Si., M.Si.

: 

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam



KATA PENGANTAR



Assalaamu'alaikum warahmatullaahi wabarakaatuh

Alhamdulillahirobbil alamin, puji syukur senantiasa penulis panjatkan atas kehadiran Allah SWT yang telah memberikan rahmat, karunia, dan hidayah-Nya sehingga penulis dapat menyelesaikan tugas akhir yang berjudul “Implementasi *Package Bioconductor* pada *Software R* untuk Olah Data Hasil *Genotyping by Sequencing* pada Tanaman Cabai (*Capsicum annuum*)”. Shalawat serta salam tak lupa penulis haturkan kepada junjungan Nabi Besar Muhammad SAW, karena berkat perjuangan beliau kita dapat merasakan nikmat Islam yang begitu luar biasa serta indahny ilmu pengetahuan.

Penelitian ini dilakukan sebagai salah satu persyaratan yang harus dipenuhi untuk mencapai gelar sarjana statistika strata satu di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Penulis menyadari bahwa dalam penyusunan tugas akhir ini tidak lepas dari bimbingan, dorongan, dan bantuan dari berbagai pihak, sehingga kegiatan yang telah direncanakan dapat terlaksana dengan baik dan terselesaikan pada waktu yang tepat. Oleh karena itu, perkenankan penulis menghaturkan terima kasih kepada :

1. Mama Nurul tercinta sebagai penyemangat terbesar bagi penulis yang luar biasa memberikan kasih sayang serta mendoakan yang terbaik untuk penulis dan bekerja keras tanpa kenal lelah demi kelancaran studi penulis. Bude Aniek, Bude Inien, Om Yoyok dan seluruh keluarga penulis yang telah memberikan semangat kepada penulis, memberikan dukungan baik moril maupun materil.
2. Bapak Fathul Wahid, S.T., M.Sc., Phd, selaku Rektor Universitas Islam Indonesia.
3. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia, Yogyakarta.

4. Dr. RB. Fajriya Hakim, S.Si., M.Si., selaku Ketua Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia beserta jajarannya, terima kasih atas kerja keras bapak ibu dalam memberikan ilmu yang manfaat.
5. Ibu Dr. techn. Rohmatul Fajriyah, S.Si., M.Si., selaku dosen pembimbing pertama yang telah membimbing penulis, meluangkan waktu dan tenaga, memberikan keyakinan kepada penulis, memberikan semangat, arahan, saran serta ilmu yang bermanfaat sehingga tugas akhir ini dapat terselesaikan dengan baik.
6. Bapak Muryanto, S.P., M.Si. selaku pembimbing kedua, Bapak Boy Valenza Damiri, Mas Sandi Arya Rumintang, Ibu Anisa beserta jajaran staf divisi Bioteknologi PT East West Seed Indonesia yang telah mengenalkan ilmu pengetahuan baru dibidang bioinformatika, memberikan saran maupun nasihat, memberikan semangat, serta membimbing penulis dalam memahami setiap detail ilmu bioinformatika.
7. Kawan-kawan seperjuangan *Bioinformatics* Ika, Nurlina, Umami, Nanda, Husain, Ari, Leni, Himelda, Maidah, Shodiq, Gilang dan Aat yang sudah banyak memberikan kenangan dan tak lupa saling mendukung serta menguatkan satu sama lain karena kita luar biasa.
8. Teman-teman, kakak-kakak dan adik-adik Statistika UII yang sudah memberikan dukungan, bantuan, serta berbagi ilmu dan pengalaman yang berharga kepada penulis selama menjalani kuliah di Universitas Islam Indonesia.
9. Semua pihak yang telah memberikan semangat, dukungan, doa, dan bantuan kepada penulis baik secara langsung maupun tidak langsung yang tidak dapat penulis sebutkan satu-persatu.

Semoga bantuan dan dukungan yang telah diberikan mendapat balasan dari Allah SWT. Penulis menyadari sepenuhnya bahwa dalam penyusunan laporan penelitian ini masih banyak kekurangan, oleh sebab itu penulis mengharap kritik dan saran yang bersifat membangun dalam pengembangan di masa mendatang

dan bermanfaat bagi yang membaca serta penulis khususnya. Semoga Allah SWT selalu melimpahkan rahmat dan hidayah-Nya kepada kita semua, Aamiin.

Wassalamualaikum warahmatullaahi wabarakatuh

Yogyakarta, Juli 2018

Penulis

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PERSETUJUAN PEMBIMBING	iii
HALAMAN PENGESAHAN	iv
KATA PENGANTAR	v
DAFTAR ISI	viii
DAFTAR GAMBAR	x
DAFTAR TABEL	xi
DAFTAR LAMPRAN	xii
PERNYATAAN	xiv
INTISARI	xv
ABSTRACT	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Batasan Masalah	4
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
BAB II TINJAUAN PUSTAKA	6
2.1 Penelitian tentang Cabai	6
2.2 Penelitian GBS dengan Memanfaatkan <i>Next-Generation Sequencing</i>	6
2.3 Penelitian GBS untuk Identifikasi SNPs	7
BAB III LANDASAN TEORI	9
3.1 Cabai	9
3.1.1 Taksonomi Cabai	9
3.2 Pemuliaan Tanaman	10
3.3 <i>Genotyping by Sequencing</i> (GBS)	10
3.4 <i>Single Nucleotide Polymorphisms</i> (SNP)	11

3.5 Bioconductor	12
3.5.1 <i>ShortRead Package</i>	13
3.5.2 <i>Rqc Package</i>	13
3.5.3 <i>Rbowtie Package</i>	14
3.5.4 <i>Rsamtools Package</i>	14
BAB IV METODOLOGI PENELITIAN.....	16
4.1 Populasi dan Sampel Penelitian.....	16
4.2 Variabel dan Definisi Operasional Penelitian	16
4.3 Jenis dan Sumber Data	17
4.4 Tempat dan Waktu Penelitian	17
4.5 Metode dan Analisis	17
4.6 Tahapan Penelitian	18
BAB V HASIL DAN PEMBAHASAN.....	19
5.1 Deskriptif Data <i>Fastq</i>	19
5.1.1 Basa Sekuens.....	19
5.1.2 Sekuens, Panjang <i>Read</i> , Rangkaian Basa dan Kualitas Basa	22
5.2 Kualitas Sekuens	26
5.3 GC Content.....	28
5.4 Preparasi <i>SNP Calling</i>	30
5.4.1 <i>Alignment</i> Menjadi <i>SAM</i> Format.....	30
5.4.2 Merubah <i>SAM</i> ke <i>BAM</i> Format.....	28
BAB VI KESIMPULAN DAN SARAN.....	33
6.1 Kesimpulan.....	33
6.2 Saran	33
DAFTAR PUSTAKA	35
LAMPIRAN.....	39

DAFTAR GAMBAR

Gambar 3.1	<i>Single Nucleotide Polymorphisms</i>	11
Gambar 4.1	Tahap analisis data	18
Gambar 5.1	Menampilkan <i>file directory</i> atau lokasi penyimpanan data.....	19
Gambar 5.2	Plot kualitas <i>read file</i> data A1_142804-1_1 dan A1_142804-1_2.....	27
Gambar 5.3	Plot rata-rata kualitas <i>read</i> per basa data A1_142804-1_1 dan A1_142804-1_2.....	28
Gambar 5.4	Plot <i>GC content</i> data A1_142804_1_1.....	29

DAFTAR TABEL

Tabel 3.1	<i>Install ShortRead package</i>	13
Tabel 3.2	<i>Install Rqc package</i>	13
Tabel 3.3	<i>Install Rbowtie package</i>	14
Tabel 3.4	<i>Install Rsamtools package</i>	15
Tabel 4.1	Definisi operasional penelitian	16
Tabel 5.1	<i>Command</i> untuk menginstall <i>package</i>	19
Tabel 5.2	Banyaknya basa A, C, G, dan T pada setiap <i>file</i> data	20
Tabel 5.3	<i>Base calling accuracy</i>	21
Tabel 5.4	Banyaknya <i>read</i> pada setiap <i>file</i> data	21
Tabel 5.5	<i>Syntax</i> menampilkan sekuens, panjang <i>read</i> dan rangkaian basa	22
Tabel 5.6	Sekuens, panjang <i>read</i> dan rangkaian basa A1_142804-1_1	23
Tabel 5.7	Kualitas setiap basa A1_142804-1_1	23
Tabel 5.8	Nilai kualitas	24
Tabel 5.9	Rangkaian basa tiap sekuens A1_142804-1_2	25
Tabel 5.10	Kualitas basa tiap sekuens A1_142804-1_2	25
Tabel 5.11	<i>Phred quality score</i>	26
Tabel 5.12	<i>Syntax</i> untuk menginstall <i>package Rqc</i>	27
Tabel 5.13	<i>Syntax</i> memunculkan plot <i>GC content</i>	29
Tabel 5.14	<i>Syntax</i> untuk menginstall <i>package Rsamtools</i>	31
Tabel 5.15	Hasil <i>alignment</i> A1_142804-1_1	31

DAFTAR LAMPIRAN

Lampiran 1	<i>Script quality assesment</i>	39
Lampiran 2	Hasil “baseCalls”	39
Lampiran 3	Hasil "readCounts"	39
Lampiran 4	<i>Script</i> menampilkan data sekuens DNA.....	40
Lampiran 5	<i>American Standard Code for Information Interchange</i> (ASCII) ...	42
Lampiran 6	<i>Output sequence reads</i> dan kualitas basa data A1_142804-1_1	43
Lampiran 7	<i>Output sequence reads</i> dan kualitas basa data A1_142804-1_2	43
Lampiran 8	<i>Output sequence reads</i> dan kualitas basa data A2_142804-1_1	44
Lampiran 9	<i>Output sequence reads</i> dan kualitas basa data A2_142804-1_2	44
Lampiran 10	<i>Output sequence reads</i> dan kualitas basa data A3_142804-1_1	45
Lampiran 11	<i>Output sequence reads</i> dan kualitas basa data A3_142804-1_2	45
Lampiran 12	<i>Output sequence reads</i> dan kualitas basa data B1_84860-1_1	46
Lampiran 13	<i>Output sequence reads</i> dan kualitas basa data B1_84860-1_2	46
Lampiran 14	<i>Output sequence reads</i> dan kualitas basa data B2_84860-1_1	47
Lampiran 15	<i>Output sequence reads</i> dan kualitas basa data B2_84860-1_2	47
Lampiran 16	<i>Output sequence reads</i> dan kualitas basa data B3_84860-1_1	48
Lampiran 17	<i>Output sequence reads</i> dan kualitas basa data B3_84860-1_2	48
Lampiran 18	<i>Script</i> kualitas sekuens <i>reads</i>	49
Lampiran 19	Plot kualitas <i>read file</i> data A2_142804-1_1& A2_142804-1_2.....	50
Lampiran 20	Plot kualitas <i>read file</i> data A3_142804-1_1 & A3_142804-1_2.....	50
Lampiran 21	Plot kualitas <i>read file</i> data B1_84860-1_1 & B1_84860-1_2.....	51
Lampiran 22	Plot kualitas <i>read file</i> data B2_84860-1_1 & B2_84860-1_2.....	51
Lampiran 23	Plot kualitas <i>read file</i> data B3_84860-1_1 & B3_84860-1_2.....	52
Lampiran 24	Plot rata-rata kualitas <i>read</i> per basa data A2_142804-1_1 & A2_142804-1_2.....	52
Lampiran 25	Plot rata-rata kualitas <i>read</i> per basa data A3_142804-1_1 & A3_142804-1_2.....	53

Lampiran 26 Plot rata-rata kualitas <i>read</i> per basa data B1_84860-1_1 & B1_84860-1_2	53
Lampiran 27 Plot rata-rata kualitas <i>read</i> per basa data B2_84860-1_1 & B2_84860-1_2	54
Lampiran 28 Plot rata-rata kualitas <i>read</i> per basa data B3_84860-1_1 & B3_84860-1_2	54
Lampiran 29 Menampilkan plot <i>GC content</i>	55
Lampiran 30 Plot <i>GC content</i> data A2_142804-1_1	56
Lampiran 31 Plot <i>GC content</i> data A2_142804-1_2	57
Lampiran 32 Plot <i>GC content</i> data A3_142804-1_1	57
Lampiran 33 Plot <i>GC content</i> data A3_142804-1_2	57
Lampiran 34 Plot <i>GC content</i> data B1_84860-1_1	58
Lampiran 35 Plot <i>GC content</i> data B1_84860-1_2	58
Lampiran 36 Plot <i>GC content</i> data B2_84860-1_1	58
Lampiran 37 Plot <i>GC content</i> data B3_84860-1_1	59
Lampiran 38 Plot <i>GC content</i> data B3_84860-1_2	59
Lampiran 39 <i>Alignmnet</i>	59
Lampiran 40 Konversi <i>sam</i> ke <i>bam</i>	66
Lampiran 41 Membaca <i>file bam</i>	67
Lampiran 42 Data <i>bam</i> A1_142804-1_1	69
Lampiran 43 Data <i>bam</i> A1_142804-1_2	69
Lampiran 44 Data <i>bam</i> A2_142804-1_1	70
Lampiran 45 Data <i>bam</i> A2_142804-1_2	70
Lampiran 46 Data <i>bam</i> A3_142804-1_1	70
Lampiran 47 Data <i>bam</i> A3_142804-1_2	70
Lampiran 48 Data <i>bam</i> B1_84860-1_1	71
Lampiran 49 Data <i>bam</i> B1_84860-1_2	71
Lampiran 50 Data <i>bam</i> B2_84860-1_1	71
Lampiran 51 Data <i>bam</i> B2_84860-1_2	71
Lampiran 52 Data <i>bam</i> B3_84860-1_1	72
Lampiran 53 Data <i>bam</i> B3_84860-1_2	72

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 31 Juli 2018



Penulis

**IMPLEMENTASI PACKAGE BIOCONDUCTOR PADA SOFTWARE R
UNTUK OLAH DATA HASIL GENOTYPING BY SEQUENCING
PADA TANAMAN CABAI (*Capsicum annuum*)**

Oleh: Hawila Sonya Savitri

Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

INTISARI

Penelitian ini menerapkan *package* dari *Bioconductor* pada *software* R dalam menangani data sekuens DNA cabai yang disimpan dengan format *fastq*. Data tersebut akan dicari lokasi SNPs dengan cara *alignment* data *fastq* dengan referensi genom cabai Zunla yang disimpan dalam format *fasta*. Sebelumnya terlebih dahulu digunakan *package ShortRead* dan *Rqc* untuk melihat kualitas data sekuens DNA yang dihasilkan oleh mesin sekuensing. Hasil sekuens dapat dikatakan sempurna karena nilainya mendekati 100%. Selanjutnya dilakukan *alignment* data *fastq* dan *fasta* yang hasilnya disimpan dalam *sam file* dengan bantuan *package Rbowtie*. Kemudian data dalam format *sam* dikonversi menjadi *bam* dengan menggunakan *package Rsamtools*. Hasilnya yaitu dari dua belas data yang dimiliki, data *bam* A2_142804-1_1 memiliki jumlah lokasi sekuens DNA terbanyak yang didalamnya mengandung SNPs, yaitu sebanyak 95.349 *ranges*.

Kata Kunci: *Alignment*, Cabai, *Fastq*, Referensi Genom, Sekuens DNA

**IMPLEMENTATION OF BIOCONDUCTOR'S PACKAGE IN R SOFTWARE
FOR DATA PROCESSING OF GENOTYPING BY SEQUENCING
ON HOT PEPPER (*Capsicum annuum*)**

Hawila Sonya Savitri

Department of Statistics, Faculty of Matematics and Natural Sciences

Islamic University of Indonesia

ABSTRACT

*This study applied package from Bioconductor in software R to handling data of hot pepper (*Capsicum annuum*) DNA sequence which stored with fastq format. These data will be searched the location of SNPs by alignment the fastq file with genom reference of pepper Zunla (stored in fasta format). Previously we used ShortRead and Rqc packages to quality assessment of the DNA sequence data generated by the sequencing machine. The DNA sequence can be considered perfect because the value is close to 100%. The next step is alignment of fastq and fasta file and the result is stored in sam file by using Rbowtie package. Then, sam data is converted into bam using Rsamtools package. From twelve fastq datas, A2_142804-1_1 has the most number location of DNA sequence which contain SNPs. There are 95,349 ranges.*

Keyword: *Alignment, Hot pepper, Fastq, Genome Reference, DNA Sequence*