PENGEMBANGAN APLIKASI IDENTIFIKASI *SINGLE NUCLEOTIDE POLYMORPHISM* PADA GENOM KEDELAI MENGGUNAKAN *RULE BASED CLASSIFIER*

CITRA PUSPITA RAHMAN

DEPARTEMEN ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2015

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi berjudul Pengembangan Aplikasi Identifikasi *Single Nucleotide Polymorphism* pada Genom Kedelai Menggunakan *Rule Based Classifier* adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Desember 2015

Citra Puspita Rahman
NIM G64110059

ABSTRAK

CITRA PUSPITA RAHMAN. Pengembangan Aplikasi Identifikasi *Single Nucleotide Polymorphism* pada Genom Kedelai Menggunakan *Rule Based Classifier*. Dibimbing oleh WISNU ANANTA KUSUMA dan MUHAMMAD ABRAR ISTIADI.

Rendahnya produksi kedelai di Indonesia dapat disebabkan oleh beberapa faktor, salah satunya adalah kualitas varietas kedelai lokal yang yang kurang baik. Perakitan varietas unggul baru dapat dimaksimalkan dengan pemuliaan tanaman dan dapat dibantu oleh teknologi *Next-Generation Sequencing* (NGS). Namun, aplikasi-aplikasi NGS yang tersedia selama ini masih memiliki akurasi rendah. Tujuan penelitian ini adalah mengembangkan aplikasi identifikasi SNP (bernama GPSNP) pada genom kedelai dengan menggunakan *rule* hasil optimasi *Genetic Programming* (GP) sebagai *classifier*. Evaluasi aplikasi dilakukan dengan cara membandingkan SNP yang telah diperoleh dengan dua aplikasi identifikasi SNP lainnya yaitu Freebayes dan SAMtools. Pada salah satu data (BAM C01), GPSNP, Freebayes, dan SAMtools masing-masing menghasilkan 2 397 923 SNP, 737 945 SNP, dan 67 401 SNP. Kemudian *reliable true* SNP, yakni SNP yang muncul di posisi yang sama pada ketiga aplikasi, berjumlah 57 590 SNP. Hasil ini menunjukkan aplikasi GPSNP dapat mengidentifikasi sebagian besar *true* SNP, tetapi masih terdapat banyak *false positive* SNP yang teridentifikasi.

Kata kunci: identifikasi SNP, kedelai, *next-generation sequencing*

ABSTRACT

CITRA PUSPITA RAHMAN. Development of Application for Single Nucleotide Polymorphism Identification in the Soybean Genome Using Rule-Based Classifier. Supervised by WISNU ANANTA KUSUMA and MUHAMMAD ABRAR ISTIADI.

Low soybean production in Indonesia can be caused by several factors, one of them is low quality of soybean varieties. Development of new soybean variety can be maximized by plant breeding and can be aided by Next-Generation Sequencing Technology (NGS). Unfortunately, most of NGS applications still have low accuracy. The aim of this study is to develop an application for SNP identification (named GPSNP) in soybean genome with rules obtained from Genetic Programming (GP) as a rule-based classifier. The evalution was conducted by comparing the SNP results with two other applications (Freebayes and SAMtools). From one of research data (BAM C01), GPSNP, Freebayes, and SAMtools obtained 2 397 923 SNPs, 737 945 SNPs, and 67 401 SNPs respectively. The reliable true SNPs (SNPs appearing on the same position in those three applications) were 57 590 SNPs. This result indicate that GPSNP is able to identify most of true SNPs, although there are still many false positive SNPs identified.

Keywords: next-generation sequencing, SNP identification, soybean

PENGEMBANGAN APLIKASI IDENTIFIKASI *SINGLE NUCLEOTIDE POLYMORPHISM* PADA GENOM KEDELAI MENGGUNAKAN *RULE BASED CLASSIFIER*

CITRA PUSPITA RAHMAN

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer
pada
Departemen Ilmu Komputer

DEPARTEMEN ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2015

Penguji: Toto Haryanto, SKom MSi

Judul Skripsi: Pengembangan Aplikasi Identifikasi *Single Nucleotide Polymorphism* pada Genom Kedelai Menggunakan *Rule Based Classifier*
Nama : Citra Puspita Rahman
NIM : G64110059

Disetujui oleh

DrEng Wisnu Ananta Kusuma, ST MT
Pembimbing I

Muhammad Abrar Istiadi, SKomp MKom
Pembimbing II

Diketahui oleh

Dr Ir Agus Buono, MSi MKom
Ketua Departemen

Tanggal Lulus: 03 DEC 2015

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan.

Penulis mengucapkan terima kasih kepada seluruh pihak yang terlibat dalam penelitian ini yaitu:

1 Bapak Yuzarsil Rahman, ibu Ermita dan kakak Ayu Vista Rahman atas kasih sayang, doa, semangat dan dorongan kepada penulis hingga dapat menyelesaikan penelitian ini.
2 Bapak DrEng Wisnu Ananta Kusuma, ST MT dan Bapak Muhammad Abrar Istiadi, SKomp MKom selaku dosen pembimbing yang telah memberikan arahan, masukan, ide, bantuan, serta dukungan dalam penyelesaian penelitian ini.
3 Farino, Fadhlan Rizal, Erwansyah Adriantama, Riko Ahmad Maulana, Muhammad Abruri yang telah membantu mengatasi kesulitan pemrograman dan pengunduhan data yang penulis hadapi.
4 Teman terdekat Fitri, Ira Hastuti, Ulfa Nikmatiya, Weni Handayani, Hanif Bagus Guritno, Fikry Destian, Muhammad Fhadil serta teman sekamar Khariza Dwi Sepriani atas dukungan, motivasi serta bersedia menjadi tempat curhat penulis.
5 Rekan-rekan bimbingan Dezika Geniya, Ikhsan Wisudhandi Wibawa, Meylinda Nur Puspita yang telah banyak membantu penulis.
6 Rekan-rekan seperjuangan Ilmu Komputer 48 atas segala kenangan, suka duka, kebersamaan, bantuan serta dukungannya.
7 Rekan-rekan sekelas R-SBI SMA Negeri 3 Bukittinggi yang menjadi motivasi penulis untuk tetap bertahan melanjutkan kuliah. Semoga kelak kita bisa berjumpa lagi sebagai orang-orang sukses.

Bogor, Desember 2015

Citra Puspita Rahman

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PENDAHULUAN

Latar Belakang

Tanaman pangan yang sering dikonsumsi masyarakat Indonesia setelah padi dan jagung adalah kedelai. Konsumsi masyarakat terhadap kedelai yang cenderung meningkat setiap tahunnya tidak diiringi dengan produksi kedelai di dalam negeri. Terbukti dengan konsumsi kedelai rata-rata masyarakat Indonesia adalah 2.2 juta ton pertahun, sedangkan produksi kedelai pada tahun 2009 adalah 0.97 juta ton dan menurun menjadi 0.78 juta ton pada tahun 2013 (BPS 2014). Rendahnya produksi kedelai dalam negeri dapat disebabkan oleh beberapa faktor antara lain: adanya berbagai penyakit dan hama yang menyerang kedelai, luas tanam, teknik budi daya, dan varietas kedelai yang digunakan (Marwoto 2011). Mahalnya benih dan buruknya kualitas kedelai lokal membuat petani memilih untuk tidak menanam kedelai atau lebih memilih menanam benih impor yang harganya lebih murah. Penyediaan varietas lokal yang bermutu (varietas unggul) merupakan salah satu strategi penting yang dapat dilakukan untuk peningkatan produksi kedelai (Atman 2009).

Perakitan kultivar unggul baru yang berdaya hasil dan berkualitas tinggi dapat dimaksimalkan dengan pemuliaan tanaman. Teknologi pemuliaan konvensional selama ini telah terbukti berhasil meningkatkan produksi kedelai. Namun, teknologi ini masih memiliki kekurangan seperti mahalnya biaya dan lamanya waktu operasional yang dibutuhkan. Kekurangan dari pemuliaan konvensional tersebut mulai teratasi dengan ditemukannya marka molekuler (Azrai 2005). *Single Nucleotide Polymorphism* (SNP) merupakan salah satu marka molekuler yang banyak digunakan saat ini dan dapat merepresentasikan perbedaan basa di antara dua individu. Pengidentifikasian SNP dapat dibantu oleh teknologi *next-generation sequencing* (NGS) untuk membaca data sekuens DNA yang diteliti. Teknologi NGS memiliki kelebihan yaitu mampu menghasilkan data yang sangat besar dengan waktu yang singkat dan membuat proses sekuensing DNA menjadi lebih efisien dan murah dibandingkan dengan metode konvensional (Metzker 2010).

Terdapat banyak *tools* atau aplikasi NGS yang dapat digunakan untuk mengidentifikasi SNP pada genom kedelai. Aplikasi NGS yang tersedia antara lain: SAMtools, GATK, Freebayes, dan SoapSNP. Namun, aplikasi-aplikasi tersebut masih memiliki akurasi yang rendah. Rendahnya akurasi disebabkan oleh penggunaan model probabilistik sebagai dasar pengklasifikasian SNP pada kebanyakan aplikasi NGS tersebut. Banyaknya *false SNP* (SNP bukan sebenarnya) yang dihasilkan karena akurasi yang buruk dapat mempengaruhi secara signifikan hasil analisis SNP. Metode *machine learning* baru-baru ini banyak digunakan untuk meningkatkan akurasi pada pengidentifikasian SNP (O'Fallon *et al.* 2013).

Beberapa penelitian telah menggunakan metode *machine learning* terkait dengan pengidentifikasian SNP pada genom kedelai. Pada penelitian sebelumnya Istiadi *et al.* (2014) menggunakan algoritme C4.5 dan fitur statistik DNA untuk mengklasifikasikan SNP pada data sekuens DNA genom kedelai budi daya hasil penelitian Lam *et al.* (2010). Pada percobaan tersebut diperoleh akurasi yang

cukup baik yaitu 93%, namun masih memiliki *sensitivity* dan *specificity* yang rendah yaitu 56.7% dan 3.1%. Perbedaan yang signifikan antara *sensitivity* dan *specificity* tersebut menggambarkan adanya ketidakseimbangan distribusi pada kedua kelas *true* dan *false* SNP pada data. Untuk mengatasi masalah tersebut, Istiadi *et al.* (2014) kemudian menerapkan metode *random undersampling* pada data set dan diperoleh *sensitivity* dan *specificity* yang lebih baik yaitu 92.8% dan 13.8%. Hasil tersebut menunjukkan penerapan metode *random undersampling* pada data set menghasilkan akurasi yang lebih baik dibandingkan tanpa penerapan metode *random undersampling*. Data set yang digunakan oleh Istiadi *et al.* (2014) juga digunakan oleh Hasibuan *et al.* (2014) dengan metode *Support Vector Machine* (SVM) dan juga menerapkan metode *random undersampling* untuk memperbaiki masalah distribusi kelas. Hasibuan *et al.* (2014) memperoleh hasil yang lebih baik yaitu *sensitivity* 88% dan *specificity* 15%, sedangkan data *imbalanced* menghasilkan *sensitivity* dan *specificity* yang lebih rendah yaitu 51% dan 2%.

Pada penelitian Istiadi *et al.* (2014), algoritme C4.5 selain dapat menghasilkan akurasi yang cukup baik juga menghasilkan *decision tree* yang dapat diolah menjadi sekumpulan *rule.* Namun, *tree* yang dihasilkan berukuran besar dan jika diolah menjadi *rule set* akan terbentuk 94 *rule*. Hasibuan *et al.* (2014) tidak menghasilkan aturan atau *rule* yang dapat dijadikan sebagai *classifier* karena metode SVM berbasiskan *black-box* yang artinya model tidak dapat diinterpretasi karena berupa deretan angka yang menyatakan *support vector*.

Pada penelitian ketiga, Istiadi *et al.* (2015) melakukan pengklasifikasian SNP berdasarkan fitur statistik sekuens DNA dengan menggunakan metode *Genetic Programming* (GP). Kelebihan dari metode *genetic programming* adalah *rule set* yang terbentuk relatif sedikit, sederhana, dan dapat diinterpretasi dengan mudah. *Rule set* yang terbentuk dari hasil optimasi GP tersebut dapat diterapkan ke dalam aplikasi sebagai *classifier* sehingga diharapkan dapat meningkatkan akurasi pada pengidentifikasian SNP. Penelitian ini akan mengembangkan aplikasi identifikasi SNP yang bernama GPSNP dan menerapkan *rule* hasil optimasi GP ke dalam aplikasi sebagai *classifier*.

Perumusan Masalah

Bagaimana mengembangkan aplikasi identifikasi SNP berdasarkan *rule based classifier* yang mampu membedakan *true* dan *false* SNP secara akurat?

Tujuan Penelitian

Tujuan penelitian ini adalah mengembangkan suatu aplikasi identifikasi SNP pada genom kedelai yang berbasiskan *rule* (*rule based classifier*).

Manfaat Penelitian

Penelitian ini diharapkan dapat memudahkan pengguna terutama peneliti pemuliaan tanaman dan bioinformatik untuk dapat mengidentifikasi SNP pada genom kedelai secara lebih akurat.

Ruang Lingkup Penelitian

Batasan penelitian ini mencakup:

1 Genom rujukan yang digunakan adalah sekuens genom kedelai budi daya Williams 82 versi rilis v1 dengan 8x *coverage*.
2 Data rujukan yang digunakan untuk data masukan hanya terbatas pada data Gm01 (kromosom 1 pada genom kedelai).
3 *Rule* yang diterapkan adalah *rule* paling optimal yaitu *rule* FP3 yang diperoleh dari hasil penelitian Istiadi *et al.* (2015).
4 Proses identifikasi SNP dibatasi sampai SNP putatif (belum divalidasi secara biologi) dan tidak mencakup identifikasi indel (*insertion-deletion*).

TINJAUAN PUSTAKA

Kedelai

Kedudukan tanaman kedelai budi daya dalam sistematik tumbuhan (taksonomi) diklasifikasikan sebagai berikut: (1) kingdom *Plantae*; (2) divisi *Spermatophyta*; (3) sub-divisi *Angiospermae*; (4) kelas *Dicotyledonae*; (5) ordo *Polypetales*; (6) famili *Leguminosae*; (7) sub-famili *Papilionoideae*; (8) genus *Glycine*; (9) spesies *Glycine L Max*. Spesies yang paling dekat dengan kedelai budi daya (*G. max*) adalah kedelai liar *G. soja*, *G. clandestina* dan *G.usuriennsis*. Tanaman kedelai pada umumnya mempunyai susunan genom diploid dengan 20 pasang kromosom atau 2n=40 (Rukmana danYuniarsih 1996).

Identifikasi SNP (*SNP Calling*)

Pengidentifikasian SNP dari teknologi NGS membutuhkan banyak proses di dalamnya. Identifikasi SNP atau SNP *calling* juga akan menghasilkan data sementara (*temporary data*) yang berukuran besar hasil dari penjajaran data rujukan dan data *read*. Masing-masing dari data tersebut memiliki format, struktur, dan *code* yang berbeda. Berikut penjelasan masing-masing format yang digunakan dan dihasilkan dalam proses SNP *calling* menggunakan teknologi NGS:

1 FASTA

Format FASTA digunakan untuk menyimpan informasi setiap basa pada genom rujukan (Gambar 1). Baris pertama pada data FASTA berisikan informasi kromosom dan daerah pemetaan gen, sedangkan baris ke 2 sampai 860947 (pada data rujukan Gm01) berisikan basa pada sekuens rujukan (Altmann *et al.* 2012).

```
>Gm01:1..51656713
GGTTTGGTGTTTGGGTTTTAGGTTTTAGGTTTTAGGTTTTACGGTTTAGGGTTTATG
TATGGTTTAGGGTTTAGGGTTAGGAAATAATTTGGGTCTTTCATCTTTCAACAAAAA
```

Gambar 1 Contoh data berformatkan FASTA

2 FASTQ

Hampir semua *platform* NGS menggunakan format FASTQ untuk menyimpan data *read* (Gambar 2). Data FASTQ terdiri atas 4 baris yaitu: (1) *title line* berisi informasi tentang ID dari sekuens *read*; (2) *sequence line* berisi sekuens DNA dari data *read*; (3) karakter ”+” merupakan pemisah antara baris sekuens dan kualitas *string*; (4) *quality line* berisi kualitas dari sekuens dalam karakter ASCII (Cock *et al*. 2010).

```
@FC42G5WAAXX:5:1:6:810#0/1
GAAGAGCTTGCGCATAAAGANATCGGCTACNNNNNNNNNNNTTA
+
abb[]`bab]a`bbaaba TD\`BBBBBBBBBBBBBBBBBBBBB
```

Gambar 2 Contoh data berformat FASTQ

3 SAM

Format *Sequence Alignment/MAP* (SAM) digunakan untuk menyimpan semua informasi penting tentang data NGS (Gambar 3). Baris pada data SAM berisikan informasi tentang: (1) sekuens DNA yang digunakan (SQ); (2) aplikasi penjajaran (PQ); dan (3) kualitas dari penjajaran (Li *et al.* 2009).

```
@SQ    SN:Gm01:1..51656713     LN: 51656713
@PQ    ID: bwa   PN: bwa   VN:0.6.2-r126
FC42G5WAAXX:5:1:6:810#0 77  *    0     0     *     *     0
```

Gambar 3 Contoh data berformat SAM

4 BAM

Format BAM merupakan versi biner dari SAM dan menyimpan informasi yang sama dengan SAM (Li *et al.* 2009). Berikut visualisasi dari data berformat BAM (Gambar 4):

```
1         11        21        31        41        51
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
           TTTAGGGTTTATAKTTTAGGKTTTASGGTTTAGGGTTTAGGGTTTAG
           TTTAGGGTTTATAGTTTAGGGTTTAC           gtttagggt
```

Gambar 4 Contoh data berformat BAM

METODE

Data Penelitian

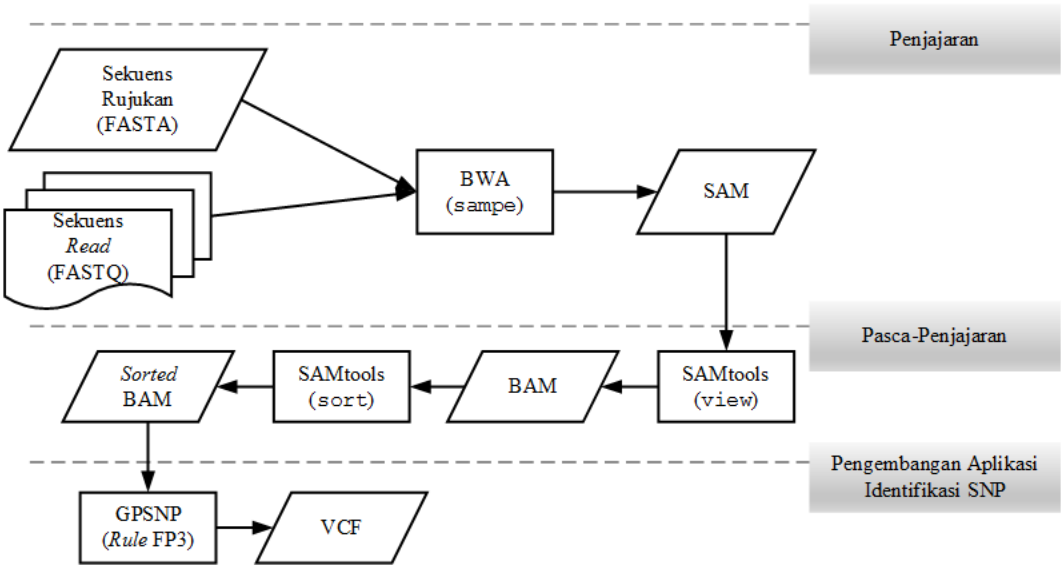
Penelitian ini menggunakan dua data yaitu data rujukan dan data *read*. Data rujukan dan *read* yang digunakan sama dengan yang digunakan oleh Istiadi *et al.* (2015). Data rujukan adalah data kedelai budi daya (*Glycine max*) varietas williams 82 versi rilis v1 dengan 8x *coverage* data. Data dapat diperoleh melalui

alamat http://www.phytozome.net/soybean.php. Data rujukan yang disediakan berformatkan FASTA dan dipetakan menjadi 20 kromosom kedelai (Gm01 sampai Gm20).

Data sekuens *reads* yang digunakan berformatkan FASTQ dan terdiri atas dua data (*paired-end reads*). Data *read* tersebut adalah data hasil dari penelitian Lam *et al*. (2010) dan dapat diunduh dari http://public.genomics.org.cn/BGI/soybean_resequencing. Data yang digunakan berjumlah 14 data aksesi kedelai budi daya dengan panjang 75 bp (*base pair*) dengan kode C01, C02, C08, C14, C16, C17, C19, C24, C27, C30, C33, C34, dan C35.

Tahapan Penelitian

Tahapan penelitian dilakukan sesuai tahapan *SNP calling pipeline* yang ada pada Altmann *et al.* (2012). SNP *calling pipeline* terdiri atas 3 tahapan utama yaitu proses penjajaran, pasca-penjajaran, dan pengembangan aplikasi identifikasi SNP (*lihat* Gambar 5). Data hasil olahan pada proses penjajaran dan pasca-penjajaran akan digunakan sebagai masukan pada aplikasi GPSNP yang akan dikembangkan.



Gambar 5 Diagram alir SNP *calling pipeline*

Tahap Pertama: Penjajaran

Sebelum dilakukan penjajaran data rujukan (*reference*) yang berformatkan FASTA dengan data *read* yang berformatkan FASTQ dilakukan terlebih dahulu proses kontrol kualitas. Proses kontrol kualitas yaitu berupa pemeriksaan kualitas data *read* dengan menggunakan aplikasi FastQC dan pemotongan kualitas data *read* yang bernilai kurang dari 20 pada kedua ujung *read* menggunakan aplikasi PRINSEQ. Setelah dilakukan proses kontrol kualitas, proses penjajaran data sekuens dapat dilakukan.

Proses penjajaran (*alignment*) data rujukan dan *read* berguna untuk menemukan basa *read* yang berbeda dengan basa pada data rujukan. Proses penjajaran dilakukan secara terpisah antara kedua data *paired-end reads*. Aplikasi yang digunakan untuk menjajarkan adalah aplikasi BWA (*Burrows-Wheeler*

Aligner). Penggunaan *options* `aln` pada aplikasi BWA akan menghasilkan data berformatkan SAI. Kemudian dengan menggunakan *options* `sampe`, kedua data dari hasil penjajaran akan digabung dan menghasilkan data berformatkan SAM. Penggunaan aplikasi BWA sebagai aplikasi untuk penjajaran dikarenakan aplikasi BWA dikenal cepat dan efisien dalam penggunaan memori (Li dan Durbin 2009).

Tahap Kedua: Pasca-Penjajaran

Data SAM dikompres menjadi data berformatkan BAM dengan menggunakan *options* `view` pada aplikasi SAMtools. Data BAM kemudian diurutkan berdasarkan posisi kromosom dari terendah sampai tertinggi dengan menggunakan *options* `sort` pada aplikasi SAMtools. Kompresi data SAM menjadi BAM bertujuan memperkecil ukuran data pada tahap SNP *Calling*.

Tahap Ketiga: Pengembangan Aplikasi Identifikasi SNP

Aplikasi identifikasi SNP GPSNP dikembangkan menggunakan metode *software development life cycle* (SDLC) dengan model *incremental* (Gambar 6). Model *incremental* pada masing-masing *build* terdiri atas tahapan analisis kebutuhan, desain alur kerja sistem, implementasi, dan evaluasi (Pressman 2001). Berikut penjelasan dari tahapan pengembangan sistem dengan menggunakan metode SDLC *incremental*:

1 Analisis Kebutuhan

Pada tahap analisis kebutuhan sistem dilakukan proses analisis kebutuhan atau fungsi yang diperlukan oleh aplikasi GPSNP. Tahapan ini dimulai dengan mengidentifikasi sistem yang telah dikembangkan sebelumnya kemudian menemukan kebutuhan selanjutnya dari sistem.

2 Desain Alur Kerja Sistem

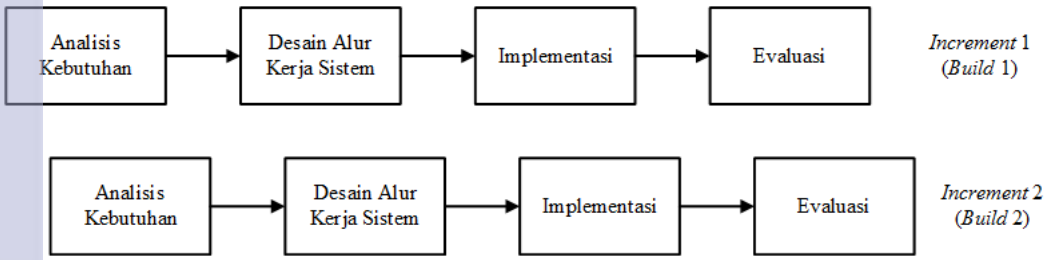
Pada tahapan ini dilakukan perancangan alur kerja sistem yang akan dikembangkan. Alur kerja sistem dirancang sesuai dengan kebutuhan dari sistem yang telah diperoleh dari tahapan analisis kebutuhan.

3 Implementasi

Sistem yang telah dirancang kemudian akan diimplementasikan dengan menggunakan bahasa pemrograman Java dengan *library* SAMtools.

4 Evaluasi

Pada tahap evaluasi dilakukan pengukuran kinerja dari aplikasi GPSNP yang telah dikembangkan.



Gambar 6 Tahapan SDLC dengan model *incremental*

Lingkungan Pengembangan

Pengembangan aplikasi dilakukan pada komputer dengan spesifikasi prosesor 2x Intel(R) Pentium(R) @ 2.30GHz dan RAM 4.0 GiB. Perangkat lunak sistem operasi yang digunakan adalah Debian GNU/Linux 7.8. Bahasa pemrograman Java dengan *library* SAMtools dan IDE IntellijIDEA *community edition* 14.0.

HASIL DAN PEMBAHASAN

Penjajaran

Pada proses *quality control*, kualitas data telah diperiksa dengan menggunakan aplikasi FastQC. Hasil yang diperoleh adalah semua data *read* yang digunakan memiliki kualitas yang baik atau *poor quality read* bernilai 0 (Gambar 7). Pemotongan *read* tidak perlu lagi dilakukan karena data *read* memiliki kualitas yang baik.

Measure	Value
Filename	c351.fq.gz
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	18464291
Sequences flagged as poor quality	0
Sequence length	75
%GC	35

Gambar 7 Contoh hasil analisis *quality control* menggunakan aplikasi FastQC

Pada proses berikutnya yaitu proses penjajaran, sudah dilakukan penjajaran data sekuens *read* sebanyak 14 data *paired-end reads* masing-masing dengan panjang 75 bp (*base pair*) dengan 1 data rujukan Gm01. Penjajaran data rujukan dan *read* membutuhkan waktu eksekusi rata-rata 1 jam untuk 1 data *paired-end reads*. Waktu eksekusi yang cukup lama tersebut disebabkan oleh ukuran data genom kedelai yang besar yaitu rata-rata berukuran 1 GB. Penjajaran data rujukan dan *reads* menghasilkan data sekuens DNA berformatkan SAM.

Pasca-Penjajaran

Pada proses pasca-penjajaran (*post-processing alignment)*, telah dilakukan pengkompresian data SAM menjadi data BAM. Pengkompresian data SAM menjadi data BAM dapat menghemat penyimpanan *file* rata-rata hingga mencapai 35% dari keseluruhan data SAM.

Pengembangan Aplikasi Identifikasi SNP

Analisis Kebutuhan

Pada tahap ini dilakukan pengidentifikasian sistem pengklasifikasian yang telah dikembangkan oleh Istiadi *et al.* (2015). Pada *build* 1 yaitu pada penelitian Istiadi *et al.* (2015) telah dilakukan pengklasifikasian SNP dengan menggunakan beberapa fitur statistik pada sekuens DNA. Fitur yang digunakan dapat dilihat pada Tabel 1 dan penjelasan dari masing-masing fitur dapat dilihat di Istiadi *et al.* (2015). Dari fitur-fitur yang telah didapatkan, dibangun suatu *classifier* berbasis *rule* yang dioptimasi dengan GP. Pembangkitan *rule* klasifikasi dilakukan dengan menggunakan tiga algoritme yaitu Bojarczuk, De Falco, dan Tan.

Tabel 1 Fitur statistik pada DNA

No	Fitur	*Code*
1	Tipe variasi	*ts.tv*
2	Maksimum kualitas alel mayor dan minor	*max.qual.major, max.qual.minor*
3	Rata-rata kualitas alel mayor dan minor	*mean.qual.major,mean.qual.minor*
4	Jarak relatif dengan ujung fragmen	*rel.dist*
5	Kedalaman	*total.depth*
6	Kualitas *alignment*	*mean.mapping.qual*
7	Jarak kandidat SNP terdekat	*nearest.flank*
8	Peluang *error*	*error prob*
9	Banyaknya perulangan dinukleotida	*dinuc.repeat*
10	*Strand bias*	*strand.bias*
11	Total *mismatch* area	*area.mismatch*
12	Panjang homopolimer	*homopolymer.length*
13	Keragaman nukleotida	*nuc.diversity*
14	Banyaknya *mismatch* pada *read*	*mismatch.alt*
15	Keseimbangan alel	*allele.balance*
16	Kualitas basa pengapit	*mean.nearby.qual*

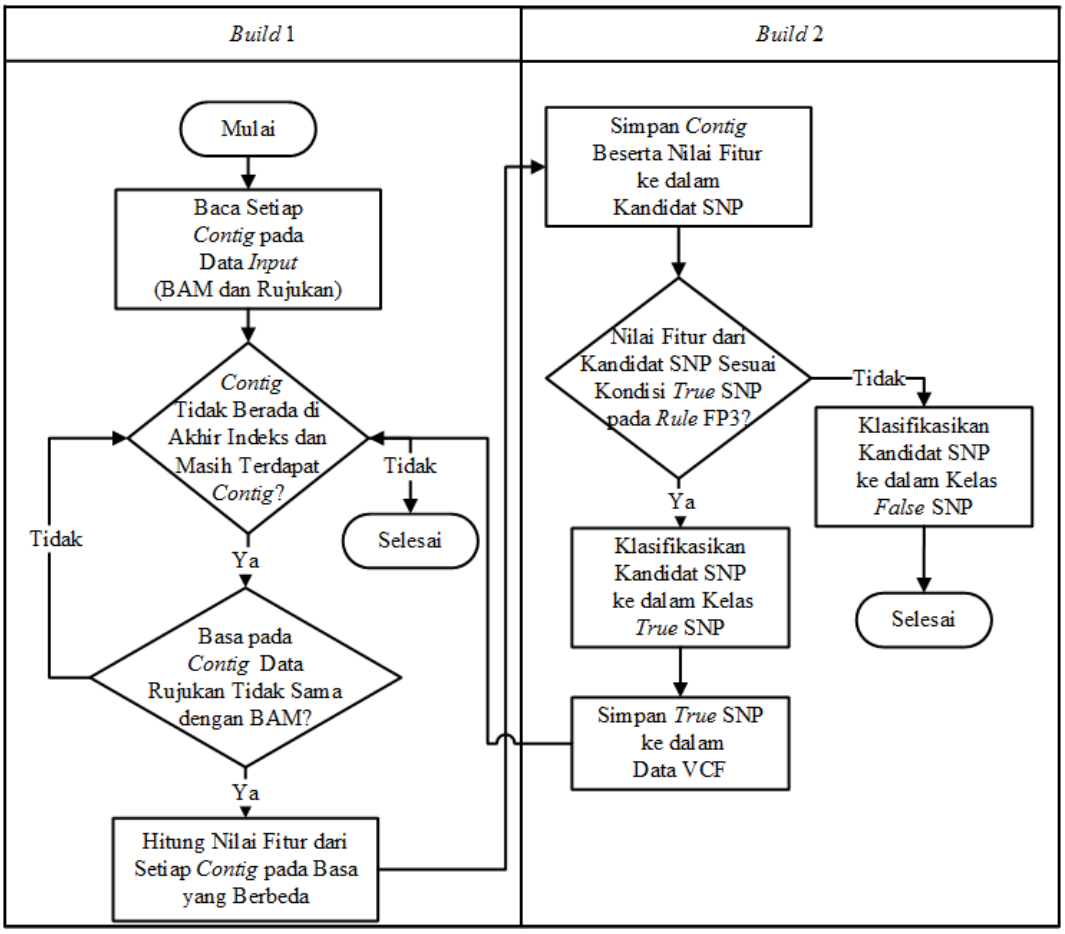
Masing-masing algoritme digunakan dalam 9 jenis percobaan yaitu (1) B1-B9 pada algoritme Bojarczuk; (2) F1-F9 pada algoritme De Falco; dan (3) T1-T9 pada algoritme Tan. Kemudian *rule* hasil dari ketiga algoritme tersebut dibandingkan dan dicari *rule* yang paling optimal untuk dijadikan sebagai *classifier*. Dari hasil perbandingan diperoleh, algoritme De Falco menghasilkan *rule* yang tidak terlalu kompleks seperti algoritme Tan dan tidak terlalu sederhana seperti algoritme Bojarczuk. Namun demikian, fungsi *fitness*-nya membuat *classifier* yang dihasilkannya hanya baik dalam mengidentifikasi kelas *false*. Oleh karena itu, fungsi *fitness* algoritme De Falco dimodifikasi agar menghasilkan *sensitivity* dan *specificity* yang baik. Dari fungsi *fitness* yang telah dimodifikasi diperoleh *rule* paling optimal yaitu *rule* FP3 dengan akurasi 93.99%.

Pada penelitian ini (*build* 2), *rule* FP3 yang telah diperoleh dari penelitian Istiadi *et al.* (2015) akan dijadikan sebagai *classifier* yang mampu membedakan kandidat SNP sebagai *true* dan *false* SNP dengan baik. Kemudian data hasil

pengidentifikasian SNP tersebut akan disimpan ke dalam keluaran (*output*) yang berformatkan VCF (*Variant Calling Format*). Penyimpanan data SNP ke dalam format VCF dilakukan karena VCF telah menjadi standar penyimpanan data SNP pada kebanyakan aplikasi NGS (Altmann *et al.* 2012).

Desain Alur Kerja Sistem

Tahap ini mendesain bentuk umum tahapan kerja aplikasi GPSNP (Gambar 8). Pada *build* 1 yaitu pada penelitian Istiadi *et al.* (2015), aplikasi GPSNP bekerja dengan cara membaca masukan (*input*). Kemudian (1) data BAM yang telah di urutkan; (2) data indeks BAM; (3) data genom rujukan berformatkan FASTA; dan (4) data indeks rujukan, keempat data tersebut dijadikan sebagai masukan pada aplikasi GPSNP. Setiap baris *contig* dibaca dan dicari basa pada data BAM yang tidak sama dengan basa pada rujukan. Selama *contig* tidak berada diakhir indeks, dilakukan penghitungan setiap nilai fitur statistik dari basa pada setiap baris *contig* yang ada.



Gambar 8 Gambaran umum desain alur kerja sistem

Pada *build* 2, setiap basa pada *contig* beserta nilai fitur statistiknya akan disimpan ke dalam kandidat SNP. R*ule* FP3 yang telah diterapkan kemudian akan digunakan sebagai *classifier* yang mampu membedakan *true* dan *false* SNP (Gambar 9). Jika nilai fitur pada *contig* yang terdapat pada kandidat SNP sesuai dengan kondisi *true* SNP pada *rule* FP3, maka kandidat SNP tersebut akan

diklasifikasikan ke dalam kelas *true* SNP begitu juga sebaliknya. *Contig-contig* yang telah diklasifikasikan ke dalam kelas *true* SNP oleh *rule* FP3, kemudian disimpan ke dalam keluaran (*output*) yang berformatkan VCF. *Sequence* dan class *diagram* yang menggambarkan alur kerja dan *class-class* dari aplikasi GPSNP dapat dilihat pada Lampiran 1 dan 2.

```
IF((max.qual.minor greather than 59.699) AND
   (total.depth smaller or equal than 82.149) AND
   (allele.balance greater or equal than 0.067))
then class is True SNP

ELSE IF((max.qual.minor smaller or equal than 60.101)OR
        (total.depth greater than 60.973) OR
        (((allele.balance is not between range 0.048
           until 0.930)))
then class is False  SNP

ELSE class is False SNP
```

Gambar 9 *Pseudocode rule* FP3

Implementasi

Pada tahap ini, aplikasi GPSNP telah berhasil dikembangkan dan diimplementasikan di dalam lingkungan sistem operasi Linux dan bahasa pemrograman Java dengan *library* SAMtools. Data VCF merupakan data akhir yang dihasilkan dari proses SNP *calling* (Gambar 10). VCF format terdiri atas meta informasi, *header*, dan data (SAMtools 2013). Bagian pertama yaitu meta informasi, menjelaskan tentang *code-code* yang terdapat pada kolom FORMAT dan INFO, beserta posisi *contig* dalam kromosom dan alamat data rujukan yang digunakan. Bagian kedua yaitu *header* berisi nama-nama dari kolom. Bagian ketiga yaitu data, berisi informasi yang dibutuhkan dalam proses analisis SNP. Berikut penjelasan pada masing-masing kolom pada bagian data pada data VCF:

meta informasi

```
##fileformat=VCFv4.1
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered reads used for calling)">
##INFO=<ID=VD,Number=A,Type=Integer,Description="Number of reads containing alt allele">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at site">
##contig=<ID=Gm01:1..51656713,length=51656713>
##reference=file:///home/csi2/Downloads/c01/gm01.fasta
```

header

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	c01-sorted

data

Gm01:1..51656713	12	.	T	G	64.0	.	DP=2,VD=1	GT:PL	0/0:3,20,17
Gm01:1..51656713	13	.	G	A	62.0	.	DP=2,VD=1	GT:PL	0/0:3,20,17
Gm01:1..51656713	16	.	T	G	65.0	.	DP=2,VD=1	GT:PL	0/0:3,20,17
Gm01:1..51656713	20	.	A	G	62.0	.	DP=2,VD=1	GT:PL	0/0:3,20,17

Gambar 10 Hasil keluaran aplikasi GPSNP

1 #CHROM

Kolom #CHROM berisikan informasi nama dari kromosom beserta *contig*.

2 POS

Kolom POS berisi informasi tentang posisi dari *contig* yang mengandung SNP. Posisi diurutkan secara numerik dari urutan terendah sampai tertinggi.

3 ID

Kolom ID berisi tentang informasi ID dari SNP pada *contig*. Aplikasi SAMtools telah terhubung dengan *database* variasi genetik atau dbSNP di National Center for Biotechnology Information (NCBI). dbSNP menyimpan SNP yang telah didaftarkan dan dievaluasi kemudian SNP yang telah didaftarkan tersebut akan memiliki ID tersendiri misalnya rs144773400 (SAMtools 2013). VCF akan menampilkan ID dari SNP jika memiliki kesamaan posisi *contig* pada genom yang sama di *database* dbSNP. Namun, dikarenakan aplikasi GPSNP tidak terhubung dengan dbSNP, kolom ID hanya menampilkan karakter '**.**'.

4 REF

Kolom REF (*reference*) menampilkan basa yang ada pada data rujukan. Basa yang ditampilkan terbatas pada basa A, C, G, dan T, sedangkan untuk basa N (basa yang tidak diketahui jenisnya) tidak akan ditampilkan pada VCF.

5 ALT

Kolom ALT (*alternative*) berisi basa yang diidentifikasi sebagai *true* SNP pada posisi *contig* tersebut.

6 QUAL

Kolom QUAL berisikan kualitas dari tiap basa pada *contig*. Nilai kualitas yang ditampikan adalah nilai kualitas rata-rata pada alel minor atau alel kedua yang sering muncul setelah alel mayor.

7 *FILTER*

Kolom FILTER akan menampilkan apakah *contig* telah melewati *filter* atau tidak. Aplikasi SAMtools menggunakan *filter* tersendiri yang dapat membuang *false* SNP dan akan menampilkan kata "*PASS*" jika SNP telah melewati *filter* (SAMtools 2013). Aplikasi GPSNP menyaring *false* SNP berdasarkan *rule* yang telah diterapkan. Oleh karena itu, pada kolom *filter* pada keluaran VCF aplikasi GPSNP hanya akan ditampilkan karakter "**.**".

8 INFO

Kolom info berisikan informasi tentang *total depth* (DP) dan *variant depth* (VD). *Total depth* adalah jumlah keseluruhan *reads* yang dijajarkan pada posisi adanya variasi, sedangkan *variant depth* adalah banyaknya variasi pada posisi tersebut (Istiadi *et al.* 2015).

9 FORMAT

Kolom format berisi informasi terkait dengan *genotype* dari data *sample*. Informasi yang ditampilkan terdiri atas dua yaitu *genotype* (GT) dan *Phred-Scale Genotype Likelihood* (PL) yang dipisahkan oleh karakter "**:**" (SAMtools 2013). Informasi GT menampilkan peluang dari suatu alel merupakan *heterozygot* atau *homozygot*. Terdapat tiga *code* yang ditampilkan yaitu 0/0, 0/1, dan 1/1. *Code* 0/0

akan ditampilkan jika basa pada *contig* tersebut berpeluang merupakan *homozygot* pada rujukan (*homozygous reference*). *Code* 0/1 untuk basa yang berpeluang merupakan *heterozygous,* dan *code* 1/1 jika basa tersebut berpeluang *homozygot* pada SNP (*homozygous alternative*). Informasi kedua yaitu PL menampilkan nilai peluang dari *heterozygous*, *homozygous reference*, dan *homozygous alternative* dari setiap *contig*.

10 *Sample name*

Kolom *sample name* menampilkan nama dari sampel (data BAM) yang digunakan.

Evaluasi

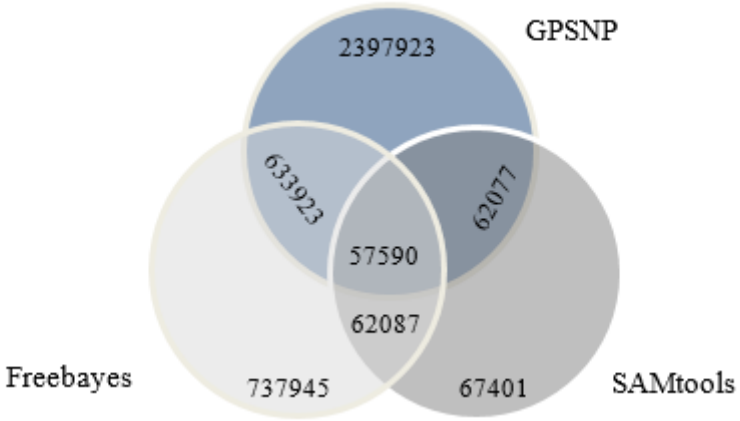
Dikarenakan tidak ada data biologi yang dapat dijadikan sebagai pembanding, pengujian aplikasi GPSNP dilakukan dengan cara membandingkan keluaran GPSNP dengan dua aplikasi lainnya yaitu SAMtools dan Freebayes. *True positive* SNP atau SNP sebenarnya adalah SNP yang hadir pada posisi yang sama di ketiga aplikasi. Penghitungan *true positive* SNP dilakukan dengan cara membandingkan posisi SNP pada VCF, kemudian dihitung SNP yang muncul pada posisi yang sama pada ketiga aplikasi. Keseluruhan hasil perbandingan jumlah SNP putatif yang ditemukan pada aplikasi GPSNP, SAMtools, dan Freebayes dapat dilihat pada Tabel 2.

Tabel 2 Jumlah SNP putatif yang ditemukan pada seluruh data *read*

Read	GPSNP	SAMtools	Freebayes	GPSNP dan SAMtools	GPSNP dan Freebayes	SAMtools dan Freebayes	GPSNP dan Freebayes dan SAMtools
C01	2397923	67401	737945	62077	633923	62087	57590
C02	2531680	36396	652218	34605	581413	33872	32336
C08	493598	649	274102	568	105485	562	495
C12	51316230	67935	749134	62951	651086	62898	58604
C14	51598617	67093	745463	61968	644161	61937	57550
C16	1828781	47287	567567	43199	481474	43975	40472
C17	2335471	19208	737843	17503	628724	17797	16345
C19	2669759	73513	621754	68760	543015	48536	45835
C24	2237079	6876	664712	6439	585591	6448	6078
C27	2172106	61631	695949	56375	586538	56742	52356
C30	2529677	37835	103860	36449	91967	7233	6880
C33	173149	1899	35653	1702	29760	1826	1648
C34	785346	1	413730	1	208168	1	1
C35	2604853	79165	822999	73412	711669	72903	68068

Berdasarkan hasil penelitian, pada salah satu data (BAM C01) diperoleh 2 397 923 SNP pada aplikasi GPSNP, 737 945 SNP pada aplikasi Freebayes, dan 67 401 SNP pada aplikasi SAMtools (Gambar 11). Kemudian *reliable true* SNP (SNP yang dapat dipercaya) yaitu SNP yang muncul di posisi yang sama pada ketiga aplikasi berjumlah 57 590 SNP. Hasil ini menunjukkan aplikasi GPSNP

dapat mengidentifikasi sebagian besar *true* SNP, namun masih terdapat banyak *false positive* SNP yang teridentifikasi. Hal ini sesuai dengan yang disampaikan pada Istiadi *et al.* (2015), *rule* FP3 memiliki *sensitivity* sebesar 92.39%, *specificity* 86.63%, tetapi masih memiliki *precision* yang rendah yaitu 30.14%. Hasil tersebut menggambarkan *rule* FP3 dapat mendeteksi sebagian besar *true* SNP dan *false* SNP, namun masih mendeteksi banyak *false positive* SNP dikarenakan adanya ketidakseimbangan distribusi kelas. Ketidakseimbangan distribusi kelas ini dapat diatasi dengan menerapkan metode *random undersampling* seperti yang dilakukan oleh Istiadi *et al.* (2014) dan Hasibuan *et al.* (2014).



Gambar 11 Hasil keluaran aplikasi GPSNP, Freebayes, dan SAMtools

SIMPULAN DAN SARAN

Simpulan

Pada penelitian ini telah dikembangkan aplikasi yang dapat mengidentifikasi SNP pada genom kedelai dengan menggunakan *rule* FP3 dari hasil optimasi *Genetic Programming* sebagai *classifier*. Hasil keluaran dari aplikasi yaitu berupa data berformatkan VCF yang berisikan data SNP yang telah teridentifikasi. Berdasarkan hasil penelitian, pada data BAM C01 diperoleh 2 397 923 SNP pada aplikasi GPSNP, 737 945 SNP pada aplikasi Freebayes, dan 67401 SNP pada aplikasi SAMtools. *Reliable true* SNP yang hadir di ketiga aplikasi tersebut yaitu berjumlah 57 590 SNP. Hasil ini menunjukkan aplikasi GPSNP dapat mengidentifikasi sebagian besar *true* SNP namun masih terdapat banyak *false positive* SNP yang teridentifikasi.

Saran

Saran-saran untuk penelitian selanjutnya:

1 Penghitungan nilai fitur pada aplikasi GPSNP dilakukan pada setiap fitur statistik pada *contig* yang ada. Namun, dari *rule* FP3 diketahui bahwa untuk mengklasifikasikan SNP hanya dibutuhkan tiga fitur statistik yaitu kedalaman

read, keseimbangan alel, dan kualitas alel minor. Oleh karena itu, pada penelitian selanjutnya tidak perlu dilakukan penghitungan fitur statitistik selain ketiga statistik yang telah disebutkan sebelumnya agar sistem dapat bekerja lebih efisien.

2 *Rule* yang diterapkan pada penelitian ini berupa *rule* statis yaitu *rule* yang memiliki nilai tetap yang diperoleh dari hasil optimasi GP. Diharapkan pada penelitian selanjutnya, dapat diterapkan *rule* dinamis yaitu *rule* yang nilai fitur statistiknya dapat berubah sesuai dengan hasil pelatihan pada data *training*.

DAFTAR PUSTAKA

Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B. 2012. A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum Genet*. 131(10): 1541-54.

Atman. 2009. Strategi peningkatan produksi kedelai di Indonesia. *J Ilmiah Tambua*. 8(1): 39-45.

Azrai M. 2005. Ulasan pemanfaatan markah molekuler dalam proses seleksi pemuliaan tanaman. *J AgroBiogen* 1(1):26-37.

[BPS] Badan Pusat Statistik. 2014. Statistik Tanaman Pangan. [diunduh Februari 2015]. Tersedia di http://www.bps.go.id/tnmn_pgn.php.

Cock PJA , Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleid Acids Research*. 38(6): 1767-1771.

Hasibuan LS, Kusuma WA, Suwarno WB. 2014. Identification single nucleotide polymorphism using support vector machine on imbalanced data. Di dalam: *Proceedings of International Conference on Advance Computer Science and Information Systems*, *ICACSIS* 2014; 2014 Okt 18; Jakarta, Indonesia. Jakarta (ID): IEEE.

Istiadi MA, Kusuma WA, Tasma IM. 2014. Application of decision tree classifier for single nucleotide polymorphism discovery from next-generation sequencing data. Di dalam: *Proceedings of International Conference on Advance Computer Science and information Systems*, *ICACSIS* 2014; 2014 Okt 18-19; Jakarta, Indonesia. Jakarta (ID): IEEE. hlm 85-89.

Istiadi MA, Kusuma WA, Tasma IM. 2015. Identifikasi *Single Nucleotide Polymorphism* pada Genom Kedelai Menggunakan *Genetic programming* [tesis]. Bogor (ID): Institut Pertanian Bogor.

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS, Zhang G. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*. 42(12): 1053-1059.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 25(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennel T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The

sequence alignment/map format and SAMtools. *Bioinformatics*. 25(16): 2078-2079

Marwoto, Hardaningsih S, Taufiq A. 2011. *Masalah Hama, Penyakit dan Hara pada Tanaman Kedelai*. Bogor (ID): Puslitbangtan.

Metzker ML. 2010. Sequencing technologies-the next generation. *Nat Rev Genet*. 11(1):31-46.

O'Fallon BD, Wooderchack-Donahue W, Crocket DK. 2013. A support vector machine for identification of single-nucleotide polymorphism from next generation sequencing data. *Bioinformatics*. 29(11): 1361-6.

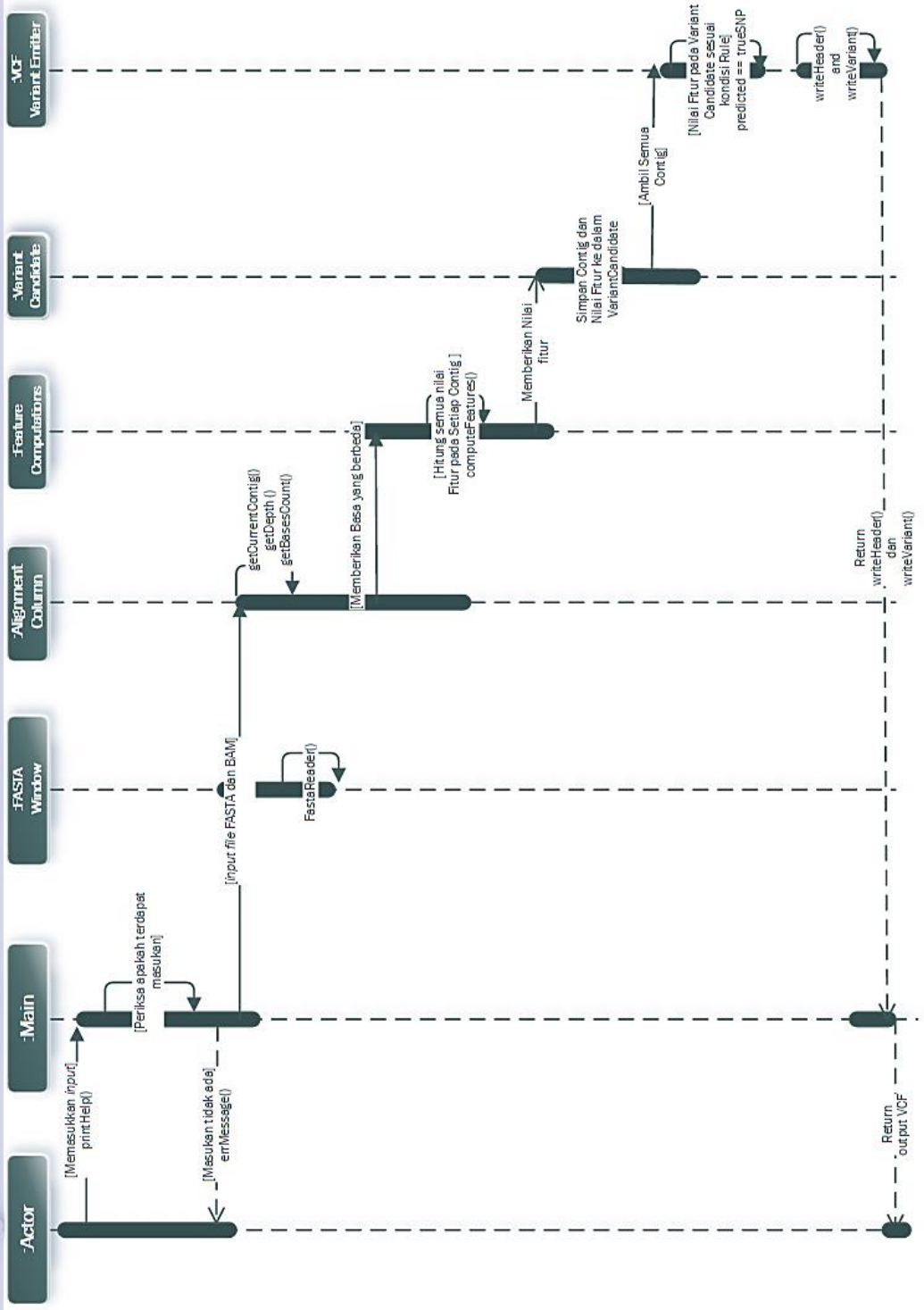
Pressman RS. 2001. *Software Engineering a Practitioner's Approach*, Ed ke-5. Boston (US): McGraw-Hill.

Rukmana R, Yuniarsih Y. 1996. *Kedelai Budi daya dan Pasca Panen*. Yogyakarta (ID): Kanisius.

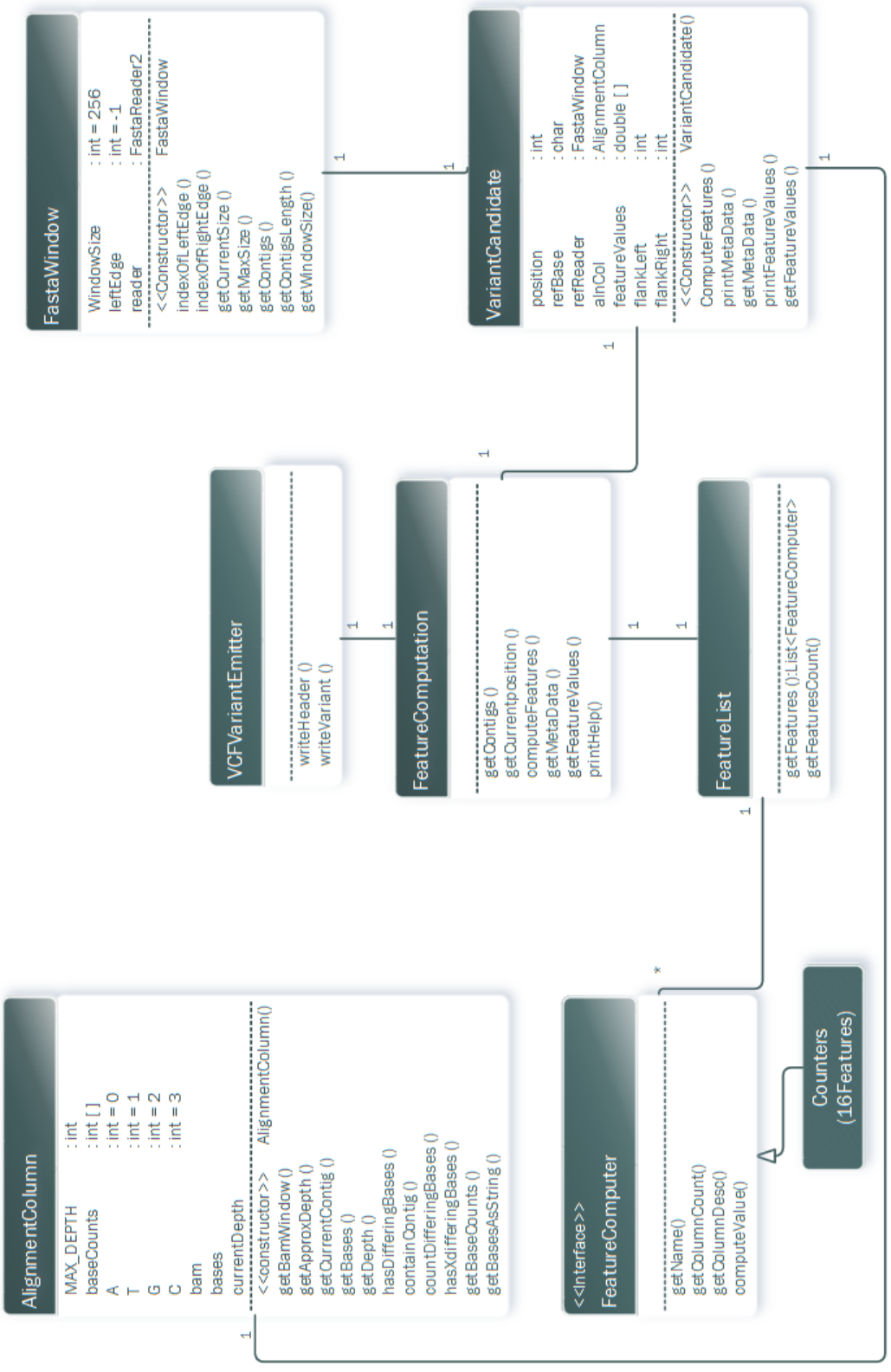
[SAMtools]. 2013. The variant call format (VCF) version 4.1 specification. [diunduh Agustus 2015]. Tersedia di https://samtools.github.io/hts-specs/VCFv4.1.pdf

LAMPIRAN

Lampiran 1 *Sequence diagram*



Lampiran 2 *Class diagram*



RIWAYAT HIDUP

Penulis dilahirkan di Bukittinggi pada tanggal 20 mei 1993 dari ayah bernama Yuzarsil Rahman dan ibu bernama Ermita. Penulis merupakan anak kedua dari empat bersaudara. Pada tahun 2011, penulis menamatkan pendidikan SMA di SMA Negeri 3 Bukittinggi dan pada tahun yang sama diterima menjadi mahasiswa departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor (IPB). Selama aktif menjadi mahasiswa, penulis pernah menjadi salah satu pengurus Organisasi Mahasiswa Daerah IPMM (Ikatan pelajar Mahasiswa Minang) pada tahun 2011 dan Himpunan Mahasiswa Ilmu Komputer pada tahun 2012. Penulis juga mengikuti beberapa kepanitian seperti IT-Today, IDEA (IPB's *Dedication for Education*) dan beberapa kepanitian lainnya. Penulis juga pernah melaksanakan Praktik Kerja Lapangan di kantor Direktorat Jenderal Perdagangan Dalam Negeri, Kementerian Perdagangan pada tahun 2014.