Jurnal Agro*Biogen* 11(1):7–16

Pembentukan Pustaka Genom, Resekuensing, dan Identifikasi SNP Berdasarkan Sekuen Genom Total Genotipe Kedelai Indonesia

(Genomic Library Construction, Resequencing, and SNP Identification Based on Whole-Genome Sequences of Indonesian Soybean Genotypes)

I Made Tasma*, Dani Satyawan, dan Habib Rijzaani

Balai Besar Penelitian dan Pengembangan Bioteknologi dan Sumber Daya Genetik Pertanian, Jl. Tentara Pelajar 3A, Bogor 16111 Indonesia
Telp. (0251) 8337975; Faks. (0251) 8338820; *E-mail: imade.tasma@gmail.com

Diajukan: 24 November 2014; Direvisi: 19 Desember 2014; Diterima: 23 Maret 2015

ABSTRACT

Resequencing of the soybean genome facilitates SNP marker discoveries useful for supporting the national soybean breeding programs. The objectives of the present study were to construct soybean genomic libraries, to resequence the whole genome of five Indonesian soybean genotypes, and to identify SNPs based on the resequence data. The studies consisted of genomic library construction and quality analysis, resequencing the whole-genome of five soybean genotypes, and genome-wide SNP identification based on alignment of the resequence data with reference sequence, Williams 82. The five Indonesian soybean genotypes were Tambora, Grobogan, B3293, Malabar, and Davros. The results showed that soybean genomic library was successfully constructed having the size of 400 bp with library concentrations range from 21.2–64.5 ng/μl. Resequencing of the libraries resulted in 50.1 x 10^9 bp total genomic sequence. The quality of genomic library and sequence data resulted from this study was high as indicated by *Q score* of 88.6% with low sequencing error of only 0.97%. Bioinformatic analysis resulted in a total of 2,597,286 SNPs, 257,598 insertions, and 202,157 deletions. Of the total SNPs identified, only 95,207 SNPs (2.15%) were located within exons. Among those, 49,926 SNPs caused missense mutation and 1,535 SNPs caused nonsense mutation. SNPs resulted from this study upon verification will be very useful for genome-wide SNP chip development of the soybean genome to accelerate breeding program of the soybean.

Keywords: Soybean, genomic library, NGS, SNP, genome variation.

ABSTRAK

Rekuensing genom total kedelai memfasilitasi penemuan marka *single nucleotide polymorphism* (SNP), insersi, dan delesi. Tujuan penelitian ini adalah mengonstruksi pustaka genom, meresekuen genom total, dan mengidentifikasi SNP genom genotipe kedelai Indonesia. Penelitian terdiri atas konstruksi dan analisis kualitas pustaka genom, resekuensing genom total menggunakan *next generation sequencing* (NGS), dan identifikasi SNP dengan penjajaran (*alignment*) data resekuen dengan sekuen rujukan varietas Williams 82. Materi genetik yang digunakan, yaitu lima genotipe kedelai Indonesia (Tambora, Grobogan, B3293, Malabar, dan Davros). Hasil penelitian menunjukkan bahwa pustaka genom total yang berhasil dikonstruksi berukuran 400 bp dengan konsentrasi berkisar antara 21,2–64,5 ng/μl. Resekuensing pustaka genom ini menghasilkan data sekuen sebanyak 50,1 x 10^9 bp. Kualitas pustaka genom dan sekuen yang dihasilkan sangat tinggi dengan nilai *Q score* 88,6% dan tingkat kesalahan pembacaan basa yang rendah (0,97%). Semua fitur menunjukkan kualitas pustaka dan sekuen pada penelitian ini termasuk kategori ideal. Analisis bioinformatika menghasilkan variasi SNP sebanyak 2.597.286, variasi insersi sebanyak 257.598 dan variasi delesi 202.157. Dari total SNP yang diidentifikasi hanya 95.207 SNP (2,15%) berada di ekson (pengode mRNA). Dari jumlah SNP yang ada di ekson, 49.892 SNP menyebabkan *missense mutation* (mutasi DNA yang mengubah susunan asam amino) dan 1.497 SNP menyebabkan *nonsense mutation* (mutasi DNA yang menghasilkan stop kodon). Setelah diverifikasi, SNP hasil penelitian ini dapat digunakan untuk mendesain *chip* SNP genom total kedelai untuk mendukung program percepatan pemuliaan kedelai.

Kata kunci: Kedelai, pustaka genom, NGS, SNP, variasi genom.

PENDAHULUAN

Kendala utama budi daya kedelai di Indonesia adalah produktivitasnya yang masih rendah dengan rerata nasional sekitar 1,3 t/ha. Untuk mengatasi kendala produksi tersebut diperlukan terobosan teknologi pemuliaan dengan memanfaatan informasi genomika kedelai untuk percepatan program pemuliaan kedelai nasional.

Genomika modern menyediakan sumber daya pemuliaan tanaman, seperti berbagai tipe marka DNA (*single nucleotide polymorphism* [SNP], insersi dan delesi [*insertion and deletion*/indel], dan *simple sequence repeat* [SSR]), gen dan *quantitative trait loci* (QTL) unggul serta sistem deteksinya menggunakan teknik berkapasitas tinggi. Deteksi alel, gen, dan QTL unggul pada koleksi plasma nutfah dapat dipercepat dengan penggunaan teknologi sekuensing kapasitas tinggi (*high throughput sequencing*) (Schuster, 2008; Zhang *et al.,* 2011) dan teknologi genotiping kapasitas tinggi (*high throughput genotyping*).

Salah satu teknologi sekuensing kapasitas tinggi untuk sekuensing genom total adalah *next generation sequencing* (NGS) platform yang menyediakan miliaran basa (sekitar 300–600 Gbp) informasi genetik dalam satu kali menjalankan alat (Pattersson *et al*., 2009). Teknik sekuensing ini juga menghasilkan data sekuen yang cukup akurat dengan akurasi data sekuen yang dihasilkan dapat mencapai 99,99% (Zhang *et al.,* 2011). Berbeda dengan teknologi Sanger, sekuensing dengan NGS menghasilkan data sekuen yang berukuran relatif pendek (50–150 bp), namun kuantitas data sekuen yang dihasilkan sangat besar.

Dengan tersedianya sekuen genom rujukan (*reference genome sequence*) berbagai spesies tanaman penting, teknologi NGS ini menjadi sangat handal (*powerful*) dalam mengidentifikasi variasi genom suatu spesies tanaman melalui penelitian resekuensing berbagai genotipe anggota spesies dalam rangka penemuan gen-gen unggul dan berbagai marka DNA dalam jumlah besar yang dapat digunakan dalam program pemuliaan tanaman. Penjajaran (*alignment*) data resekuen dengan sekuen genom rujukan menghasilkan jutaan variasi genom, seperti SNP, indel, dan SSR (Tasma, 2014).

Peningkatan secara signifikan jumlah basa yang dihasilkan NGS dan biaya sekuensing per basa yang semakin menurun, membuat kemajuan yang signifikan dalam penelitian di bidang genomika (Metzker, 2010; Voelkerding *et al.,* 2009). Hal ini didukung oleh perkembangan teknik analisis bioinformatika modern yang telah mampu memanfaatkan data sekuen pendek yang dihasilkan NGS yang mencakup data genom secara komprehensif dalam jumlah besar tersebut untuk mempercepat penemuan variasi-variasi dalam genom berbagai individu anggota spesies untuk penemuan gen dan QTL unggul mendukung program pemuliaan tanaman dengan lebih cepat.

Dengan teknologi NGS, karakterisasi genotipe koleksi plasma nutfah tanaman dapat dilakukan secara lebih komprehensif pada level genom sehingga penemuan dan pelabelan gen-gen unggul lebih efisien, akurat, dan dapat dilakukan lebih cepat. Penanganan materi genetik tanaman yang dikoleksi di bank gen juga menjadi lebih efisien karena yang dikoleksi hanya materi genetik yang berbeda (*distinct*) secara genetik.

Keberhasilan penelitian sekuensing menggunakan platform NGS sangat bergantung pada kualitas pustaka genom yang digunakan (Ansorge, 2009; Metzker, 2010). Pustaka genom merupakan kumpulan sekuen atau urutan DNA suatu organisme (Wahyudi, 2001; Wulandari, 2009). Pustaka genom mengandung semua untaian DNA suatu genom. Pada pustaka genom tersimpan gen (*coding regions*) total yang dipunyai oleh suatu organisme dan daerah bukan pengode protein (*noncoding regions*) (Pattersson *et al*., 2009; Suharsono, 2002). Fragmen DNA (pustaka genom) yang disekuen menggunakan NGS disiapkan dengan menempelkan adaptor pada ujung fragmen. Adaptor tersebut pada tahapan selanjutnya (klasterisasi pustaka genom) akan menempel pada sebuah *flow cell* sebelum dilakukan sekuensing (Commins *et al.,* 2011; Mardis, 2008).

Pengurutan basa individu anggota spesies lebih mudah dilakukan dengan tersedianya peta genom rujukan (*reference sequence*) spesies tersebut. Tersedianya peta rujukan mempercepat analisis data resekuen berbagai individu anggota spesies. Penjajaran data resekuen dengan peta sekuen genom rujukan menghasilkan berbagai variasi genom, seperti SNP, indel, dan SSR yang merupakan sumber utama marka DNA untuk tujuan pemuliaan tanaman. Resekuensing genom total dilakukan untuk mengidentifikasi variasi genetik pada genom secara menyeluruh (Tasma, 2014; Zhang *et al.,* 2011). Untuk organisme yang telah ada sekuen rujukannya, dengan teknologi sekuensing NGS, resekuensing genom total dapat dilakukan dengan cepat dan variasi genetik dapat diidentifikasi dengan cepat. Dengan demikian, teknologi ini mempercepat baik identifikasi maupun pelabelan gen unggul untuk menunjang program pemuliaan jangka panjang.

Peta sekuen rujukan genom kedelai (varietas Williams 82) telah tersedia sejak lima tahun terakhir dan dapat diakses oleh publik (Schmutz *et al*., 2010). Genom kedelai yang berukuran 1,1 miliar basa (Gb) diprediksi memiliki 46.430 gen yang mengode protein.

Kandungan gen tersebut sekitar 70% lebih banyak dibanding dengan kandungan gen tanaman model *Arabidopsis*. Tersedianya sekuen genom rujukan kedelai ini memfasilitasi percepatan identifikasi basis genetik banyak karakter penting kedelai melalui penelitian resekuensing berbagai genotipe kedelai untuk mendukung percepatan pembentukan varietas unggul baru kedelai nasional.

Salah satu marka DNA yang akhir-akhir ini sangat populer dan banyak digunakan untuk genotiping kapasitas tinggi adalah SNP (Tasma, 2014). SNP ialah variasi urutan DNA yang terjadi ketika satu nukleotida (A, T, C, atau G) dalam urutan genom diubah. SNP merupakan marka DNA yang paling umum karena jumlahnya hampir tidak terbatas dalam genom suatu organisme. Marka SNP bialelik, jumlahnya hampir tidak terbatas pada genom, mudah diotomatisasi, dan data SNP dari satu laboratorium dengan laboratorium lainnya sangat mudah digabungkan yang membuat marka SNP lebih praktis penerapannya antar laboratorium (Leonforte *et al*., 2013; Tasma, 2014). Marka lainnya adalah insersi, yaitu penambahan satu basa pada lokasi spesifik, dan delesi, yaitu pengurangan satu basa pada lokasi spesifik dalam genom (Väli *et al*., 2008). Gabungan kedua marka disebut indel. Namun, frekuensi indel pada genom jauh lebih rendah dibanding dengan SNP. Koleksi SNP dan indel merupakan sumber daya marka DNA utama sebagai materi dasar pembentukan *chip* SNP untuk genotiping kapasitas tinggi (*high throughput genotyping*) untuk deteksi dan pelabelan gen-gen dan QTL yang berasosisasi dengan karakter ekonomis penting pada kedelai seperti produktivitas, ukuran dan kualitas biji, ketahanan terhadap cekaman biotik (hama dan penyakit), dan toleransi terhadap cekaman lingkungan suboptimal (toleran keracunan Al, salinitas, dan kekeringan).

Tujuan penelitian ini adalah mengonstruksi pustaka genom, meresekuen, dan mengidentifikasi variasi SNP dan indel yang dihasilkan dari data sekuen genom total varietas kedelai Indonesia. SNP dan indel yang dihasilkan digunakan untuk mengonstruksi *chip* SNP dan sebagai sumber marka DNA komprehesif untuk mendukung program pemuliaan kedelai nasional.

BAHAN DAN METODE

Bahan Tanaman

Bahan tanaman yang digunakan adalah lima genotipe kedelai Indonesia, yaitu Grobogan, Tambora, Davros, Malabar, dan B3293. Grobogan dan Tambora adalah varietas unggul dengan produktivitas tinggi dan ukuran biji besar. B3293 adalah galur yang memiliki karakter terkait toleransi terhadap keracunan aluminium. Tambora adalah varietas unggul berbiji besar introduksi dari Filipina. Davros adalah varietas unggul produktivitas tinggi dan kebanyakan varietas unggul Indonesia memiliki latar belakang genom Davros. Malabar adalah salah satu varietas kedelai Indonesia toleran naungan. Hasil analisis filogenetik menunjukkan bahwa genotipe yang dipilih mempunyai kekerabatan jauh, khususnya Tambora yang unik dan membentuk klaster sendiri dalam pohon filogenetik (Santoso *et al*., 2006; Satyawan *et al*., 2014; Tasma *et al*., 2008) sehingga sangat baik digunakan untuk deteksi variasi genom, seperti SNP dan indel.

Isolasi, Uji Kualitas, dan Uji Kuantitas DNA sebagai Materi untuk Konstruksi Pustaka Genom

DNA genomik diisolasi dari tepung daun muda dari lima genotipe kedelai (Grobogan, Tambora, B3293, Malabar, dan Davros) menggunakan bufer CTAB dengan mengikuti metode Michiels *et al*. (2003) yang dimodifikasi (Satyawan dan Tasma, 2011). DNA dilarutkan dalam 50 μl bufer TE (Tris 10 mM pH 8, EDTA 1 mM pH 7,5), kemudian konsentrasinya diukur dengan spektrofotometer Nano Drop™ (Thermo Scientific, USA). Pita DNA genomik yang dihasilkan dielektroforesis menggunakan gel agarosa 1%.

Fragmentasi DNA Genomik dengan Teknik Nebulisasi

Fragmentasi DNA genomik kedelai dengan teknik nebulisasi (Surzcky, 1990). Kit dan reagen untuk nebulisasi DNA genomik disediakan oleh Illumina Inc. (USA). DNA genomik diberikan perlakuan gas dengan tekanan 40 psi selama 6 menit. DNA hasil nebulisasi dipurifikasi dengan metode pengendapan DNA, sentrifugasi, dan pencucian pelet DNA. Cara kerja detil nebulisasi telah dilaporkan sebelumnya (Kosasih, 2012). DNA hasil fragmentasi dielektroforesis dan diukur konsentrasinya. Fragmen DNA berukuran 400 bp dipilih dan dipotong dari gel untuk digunakan sebagai materi dasar penyiapan pustaka genom.

Konstruksi Pustaka Genom Total Kedelai

Konstruksi pustaka genom total kedelai dilakukan dengan menggunakan protokol *TruSeq LTLibrary Preparation* dari Illumina Inc. (USA). Konstruksi pustaka dilakukan dengan urutan kerja sebagai berikut (Kosasih, 2012): (1) modifikasi ujung fragmen menggunakan *Insert Modification Plate* (IMP) dan *AMPure XP Beads*; (2) adenilasi ujung 3' DNA dengan penempelan adaptor, purifikasi DNA hasil ligasi, dan amplifikasi PCR; (3) pemurnian pustaka genom; (4) validasi pustaka genom. Pada setiap langkah pem-

buatan pustaka genom ini dilakukan pengukuran kuantitas dan kualitas DNA.

Sekuensing Pustaka Genom Total Kedelai

Proses sekuensing genom total terdiri atas klasterisasi pustaka dan sekuensing klaster DNA pustaka genom. Klasterisasi DNA pustaka genom menggunakan *cBot Cluster Generation* mengikuti protokol dari Illumina Inc. (USA). Sekuensing pustaka genom dilakukan menggunakan NGS HiSeq 2000 (Illumina®). Reagen yang digunakan dan tahapan sekuensing dilakukan dengan mengikuti protokol sekuensing dari Illumina Inc. (USA). Sekuensing dilakukan dua arah (*paired-end reads*) dengan panjang pembacaan DNA 2 x 100 siklus.

Parameter yang diamati pada penelitian sekuensing meliputi hasil sekuen total, densitas klaster pustaka genom per lajur pada *flow cell*, persentase klaster *passing filter* (PF), persentase *phasing*, persentase *prephasing*, jumlah bacaan basa per lajur, jumlah bacaan basa PF per lajur, persentase hasil sekuen dengan hanya satu kesalahan pembacaan per 1.000 basa (%> = 30), dan tingkat kesalahan bacaan pada lajur control (PhiX).

Deteksi SNP dan Indel

Data sekuen final dari kelima genotipe kedelai dijajarkan (*aligned*) dengan data sekuen genom rujukan kedelai varietas Williams 82 (Schmutz *et al*., 2010) menggunakan perangkat lunak (*software*) Bowtie2 (Langmead dan Salzberg, 2012) diikuti dengan identifikasi SNP menggunakan *mpileup* dalam SAMtools (Li *et al*., 2009). Anotasi lokasi dan prediksi efek dari SNP dilakukan menggunakan perangkat lunak snpEff (Cingolani *et al*., 2012) dan analisis diversitas genetik kelima genotipe kedelai dengan sekuen genom rujukan kedelai dilakukan menggunakan perangkat lunak DarWin (Perrier dan Jacquemoud-Collet, 2006).

HASIL DAN PEMBAHASAN

Kualitas dan Kuantitas DNA Genomik sebagai Materi Dasar untuk Konstruksi Pustaka Genom Total Kedelai

Protokol isolasi DNA genomik kedelai menghasilkan DNA dengan kualitas dan kuantitas tinggi. Contoh DNA genomik (dari tiga varietas yang diteliti) yang dihasilkan pada penelitian ini utuh dengan pita yang besar dan terang dengan rasio A_{260}/A_{280} dan A_{260}/A_{230} berturut-turut berkisar antara 2,03–2,06 dan 1,90–2,03 (Tabel 1). Kisaran standar mutu DNA yang baik adalah 1,8–2,0. Ukuran pita terlihat sangat mirip antara DNA satu genotipe dan genotipe lainnya. Hasil uji kuantitatif menunjukkan bahwa konsentrasi DNA genom total ketiga genotipe sangat tinggi, yaitu 902,4 ng/μl (Tambora), 1485,7 ng/μl (B3293), dan 1401,5 ng/μl (Grobogan) (Tabel 1). Kuantitas dan kualitas DNA yang tinggi ini sangat mendukung sebagai penyedia materi dasar untuk konstruksi pustaka genom total kedelai karena pembentukan pustaka genom membutuhkan DNA dengan kuantitas dan kualitas tinggi sehingga semua reaksi selama proses pembentukan pustaka genom dapat berjalan dengan baik.

Pustaka Genom Genotipe-genotipe Kedelai Indonesia

Fragmentasi DNA genomik kedelai dengan teknik nebulisasi menghasilkan pita DNA dengan ukuran berkisar antara 100–800 bp dengan intensitas DNA paling tinggi berada pada ukuran pita sekitar 500 bp. Secara visual, DNA genomik setelah fragmentasi terlihat seperti *smear*. Hal tersebut menunjukkan bahwa DNA genom telah berhasil difragmentasi dan menghasilkan ukuran fragmen sesuai yang diharapkan, yaitu antara 100–800 bp. Hasil tersebut sesuai dengan ukuran fragmen DNA yang dibutuhkan dalam pembuatan pustaka genom untuk sekuensing genom total dengan NGS.

Kemurnian dan konsentrasi DNA setelah fragmentasi masih cukup baik dengan rasio A_{260}/A_{280} berkisar antara 2,01–2,09 yang mendekati nilai ideal kualitas DNA rasio A_{260}/A_{280} sekitar 2,0. Fragmen DNA yang berukuran 400 bp dipilih dan dipurifikasi dari gel dan DNA hasil purifikasi digunakan sebagai cetakan dalam reaksi PCR. DNA hasil amplifikasi PCR ini adalah pustaka genom kedelai yang diharapkan.

Pustaka genom yang diperoleh divalidasi untuk memastikan kelayakannya untuk sekuensing. Elektroforegram hasil elektroforesis menunjukkan bahwa pustaka genom dari tiga genotipe kedelai telah berhasil dikonstruksi yang ditandai dengan munculnya pita DNA pada ukuran 400 bp (Gambar 1). Hasil tersebut sesuai dengan ukuran fragmen DNA yang dipilih

Tabel 1. Kuantitas dan kemurnian DNA genom total tiga genotipe kedelai.

Genotipe	A_{260}	Konsentrasi DNA (ng/µl)	$A_{260}/_{280}$	$A_{260}/_{230}$
Tambora	18,05	902,40	2,05	1,93
B3293	29,71	1485,70	2,06	2,09
Grobogan	28,03	1401,50	2,03	1,90

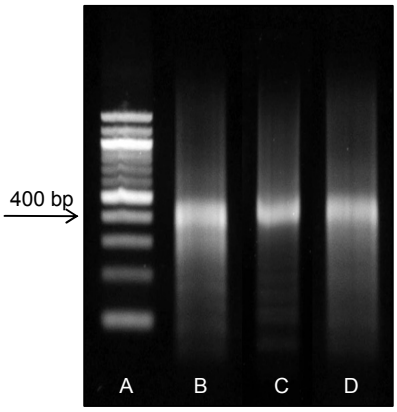
pada saat tahap purifikasi fragmen DNA hasil ligasi. Konsentrasi ketiga pustaka genom kedelai yang telah dikonstruksi berkisar antara 21,2–64,5 ng/μl (Tabel 2). Konsentrasi ini jauh lebih tinggi dari 2 ng/μl, yaitu konsentrasi yang diperlukan untuk sekuensing menggunakan NGS HiSeq 2000 (Illumina®).

Pustaka genom dua genotipe kedelai lainnya (Davros dan Malabar) dikonstruksi dengan cara yang sama dengan yang dilakukan pada konstruksi pustaka genom total genotipe Tambora, B3293, dan Grobogan. Hasil pustaka genom Davros dan Grobogan hampir sama dengan hasil pustaka genom Tambora, B3293, dan Grobogan (data tidak ditunjukkan).

Data Resekuen Genom Total Lima Genotipe Kedelai Indonesia

Jumlah basa total yang dihasilkan pada sekuensing pustaka genom kedelai sebanyak 50,1 x 10^9 bp atau 50,1 Gbp (Tabel 3). *Yield perfect* yang diperoleh pada penelitian sekuensing ini sebesar 6,3 Gbp (Tabel 3). Nilai tersebut menunjukkan jumlah basa yang secara akurat melewati *passing filter* (PF). PF adalah sejenis algoritma yang dipakai oleh komputer untuk menentukan keakuratan pembacaan basa saat sekuensing. PF yang digunakan berdasarkan data sekuen PhiX yang berperan sebagai kontrol pada tahapan sekuensing menggunakan NGS HiSeq 2000 (Illumina®). Hasil ini juga terkait dengan *percent perfect* sebesar 65,6% untuk data kontrol pada lajur 4 menggunakan DNA PhiX. Semua data di atas menunjukkan bahwa pustaka DNA dari kelima genotipe kedelai yang dikonstruksi sangat ideal digunakan untuk penelitian sekuensing menggunakan NGS HiSeq 2000 (Illumina®).

Hasil sekuensing menunjukkan bahwa densitas klaster pustaka genom yang dihasilkan berkisar antara 199–447 K/mm^2 (Tabel 4). Dari keenam lajur *flow cell* yang digunakan, hanya ada dua lajur *flow cell* yang termasuk kategori ideal, yaitu lajur B3293 7 pM dan B3293 13 pM. Densitas klaster pustaka genom termasuk kategori ideal jika klaster pustaka genom memiliki densitas berkisar antara 400–600 K/mm^2 (Illumina, 2009). Densitas klaster pustaka genom yang rendah dapat disebabkan oleh konsentrasi pustaka genom yang digunakan dalam proses klasterisasi



Gambar 1. Elektroforegram DNA pustaka genom tiga genotipe kedelai. DNA pustaka genom hasil amplifikasi PCR berukuran 400 bp sesuai dengan ukuran fragmen yang dipilih untuk konstruksi pustaka genom. A = penanda DNA, B = Tambora, C = B3293, D = Grobogan.

Tabel 2. Kuantitas dan kemurnian DNA pustaka genom tiga genotipe kedelai.

Genotipe	A_{260}	Konsentrasi DNA (ng/µl)	$A_{260}/_{280}$	$A_{260}/_{230}$
Tambora	0,46	23,20	1,92	2,12
B3293	1,28	64,50	1,99	2,37
Grobogan	0,42	21,20	1,84	1,80

Tabel 3. Jumlah basa dan kualitas klaster pustaka genom hasil sekuensing genom total lima genotipe kedelai menggunakan NGS HiSeq 2000 (Illumina®).

Yield total (Gbp)*	*Projected total yield* (Gbp)	*Yield perfect* (Gbp)**	*Yield <=3 errors* (Gbp)**	*% Perfect [number of cycles]***	*%<=3 errors [number of cycles]***
50,1	50,1	6,3	9,1	65,6 [99]	94,6 [99]

*Data sekuen seluruh lajur dari *flow cell*. **Data PhiX, yang terletak pada lajur 4 pada *flow cell* yang berfungsi sebagai kontrol internal selama proses sekuensing.

lebih rendah daripada konsentrasi yang seharusnya (7 pM atau 13 pM). Hal tersebut dapat terjadi antara lain karena kesalahan dalam pengukuran konsentrasi pustaka genom sebelum klasterisasi (Quail, 2008).

Selain densitasnya, pada klaster pustaka genom dihitung juga persentase klaster yang melewati PF. Persentase klaster PF yang dihasilkan selama proses sekuensing memiliki nilai lebih dari 93% (Tabel 4). Hasil tersebut menunjukkan bahwa nilai persentase klaster PF termasuk ke dalam kategori ideal. Nilai persentase klaster PF yang ideal memiliki nilai lebih dari 70% (Illumina, 2009). Rendahnya nilai persentase klaster PF dapat disebabkan oleh beberapa hal, di antaranya densitas klaster yang terlalu tinggi (lebih dari 600 K/mm^2), ukuran klaster pustaka genom yang terlalu panjang, dan tingginya nilai persentase *dephasing* (*phasing* dan *prephasing*) selama proses sekuensing. Nilai persentase *phasing* dan *prephasing* merupakan salah satu indikator kualitas selama proses sekuensing. *Phasing* dan *prephasing* menggambarkan hilangnya sinkronisasi dalam pembacaan basa dari urutan salinan klaster. Hal tersebut mengakibatkan dalam pembacaan basa hasil sintesis selama proses sekuensing ada yang dibaca terlalu cepat (*prephasing*) dan ada yang terlalu lambat (*phasing*).

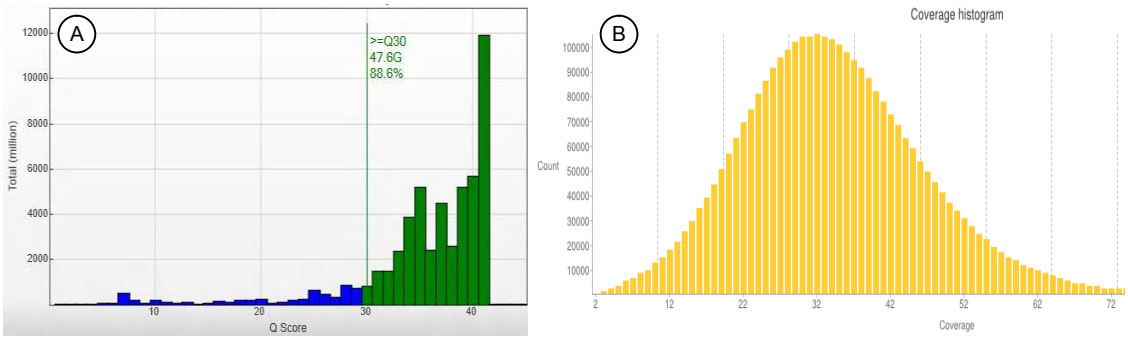
Kualitas urutan basa hasil sekuensing dapat diketahui melalui nilai *Q score*. *Q score* yang dihasilkan pada penelitian ini menunjukkan nilai di atas 30 dengan nilai persentase di atas 90% di setiap lajurnya (Tabel 4). Nilai *Q score* 30 (Q30) berarti dari 1.000 basa yang dibaca hanya satu basa yang salah pembacaan (Ewing dan Green, 1998; Voelkerding *et al.*, 2009). Secara keseluruhan, nilai Q30 yang dihasilkan selama proses sekuensing adalah 88,6% dengan jumlah basa sebanyak 47,6 x 10^9 bp (Gambar 2A).

Nilai *Q score* didefinisikan sebagai hubungan logaritmik antara proses *base calling* selama sekuensing dengan kemungkinan tingkat kesalahan yang dihasilkan dalam proses tersebut (Ewing dan Green, 1998). Voelkerding *et al.* (2009) menyatakan bahwa *base calling* diartikan sebagai proses konversi data hasil pencitraan pendaran fluoresen selama proses sekuensing menjadi basa-basa yang berurutan. Selama proses sekuensing, sekuen yang dihasilkan di lajur kontrol (PhiX) dibanding dengan sekuen PhiX yang telah tersedia di basis data (database). Hal tersebut bertujuan mengetahui tingkat kesamaan antara sekuen yang dihasilkan pada lajur kontrol dengan sekuen PhiX pada basis data. Nilai kesamaan (*aligned*) yang diperoleh selama proses sekuensing sebesar 95,3% (Tabel 5).

Tingkat kesalahan selama proses sekuensing juga dapat diketahui dari nilai *error rate* (%). Nilai *error rate* menunjukkan persentase perbandingan kesalahan pembacaan basa antara pustaka genom dengan kontrol yang dibaca pada beberapa siklus sekuensing, yaitu pada siklus 35, 75, dan 100. Secara keseluruhan, nilai *error rate* yang dihasilkan selama sekuensing sangat rendah, yaitu sebesar 0,97% (Tabel

Tabel 4. Densitas dan kualitas pustaka genom hasil sekuensing lima genotipe kedelai Indonesia menggunakan NGS HiSeq 2000 (Illumina®).

Lajur	Densitas (K/mm^2)	Klaster PF (%)	*Phasing* (%)	*Prephasing* (%)	*Reads* (Mbp)	*Read PF* (Mbp)	*%> = 30*
B3293 7 pM	416	94,12	0,274	0,343	76,63	72,06	91,6
Grobogan 7 pM	242	95,58	0,282	0,384	44,61	42,64	93,1
Tambora 7 pM	199	95,49	0,291	0,429	36,68	35,02	91,9
Kontrol (PhiX)	603	91,15	0,297	0,336	11,19	10,34	84,0
B3293 13 pM	447	93,85	0,273	0,380	82,45	77,38	91,9
Grobogan 13 pM	302	94,90	0,280	0,399	55,70	52,86	92,5
Tambora 13 pM	264	94,94	0,282	0,387	48,63	46,17	92,8



Gambar 2. Profil DNA hasil sekuensing lima genotipe kedelai menggunakan NGS HiSeq 2000 (Illumina®). A = nilai *Q score* 30 (terdapat hanya satu kesalahan dalam setiap 1.000 basa) hasil sekuensing lima genotipe kedelai, B = *sequence coverage* dari penelitian resekuensing di atas.

Tabel 5. Tingkat kesalahan (*error rate*), persentase basa sama (*aligned*), dan intensitas basa hasil sekuensing tiga genotipe kedelai pada mesin NGS HiSeq 2000 (Illumina®).

Lane	*Aligned (%)*	*Error rate (%)*	*Error rate 35 cycles (%)*	*Error rate 75 cycles (%)*	*Error rate 100 cycles (%)*	*Intensity cycle 1*	*% Intensity cycle 20*
B3293 7 pM	0	0	0	0	0	4669	80,2
Grobogan 7 pM	0	0	0	0	0	4900	80,9
Tambora 7 pM	0	0	0	0	0	4586	80,0
Kontrol (PhiX)	95,3	0,97	0,20	0,59	0	3864	80,5
B3293 13 pM	0	0	0	0	0	4557	80,1
Grobogan 13 pM	0	0	0	0	0	4673	79,7
Tambora 13 pM	0	0	0	0	0	4832	79,4

5). Hasil tersebut menunjukkan bahwa sekuensing yang telah dilakukan berlangsung dengan baik dengan tingkat kesalahan yang sangat kecil.

SNP dan Indel yang Teridentifikasi dari Data Sekuen Lima Genotipe Kedelai Indonesia

Sequencing coverage

Sequencing coverage atau *sequencing depth* ialah banyaknya pengulangan sekuen yang diperoleh pada setiap segmen DNA tertentu dalam genom organisme yang disekuen. *Sequencing coverage* 40 kali artinya rata-rata setiap segmen DNA pada bagian genom disekuen sebanyak 40 kali. Pada penelitian ini, rata-rata setiap bagian genom tersekuensing 34 kali (Gambar 2B). Gambar 2B juga menunjukkan bahwa 95% bagian genom tersekuensing minimal 10 kali. Pengulangan sekuensing ini sangat penting untuk mendapatkan tingkat akurasi data SNP yang tinggi. Hal ini penting untuk mencegah atau memperkecil peluang deteksi SNP yang salah sebagai akibat kesalahan sekuensing. Dengan pengulangan sekuen pada bagian genom yang sama minimal sebanyak 10 kali, peluang kesalahan deteksi SNP atau indel karena kesalahan sekuensing sangat kecil. Dengan demikian, SNP yang dideteksi dengan pengulangan sekuensing (*sequence coverage*) yang tinggi memiliki tingkat akurasi yang lebih tinggi dibanding dengan data sekuen dengan *sequence coverage* yang rendah. Resekuensing lima genotipe kedelai Indonesia menggunakan sekuensing dua arah (*paired-end sequencing*) menghasilkan data sekuen seperti terlihat pada Gambar 2B.

Variasi SNP dan indel yang terdeteksi pada genom kedelai Indonesia

Analisis bioinformatika melalui penjajaran data resekuen lima genotipe kedelai Indonesia dengan data sekuen rujukan genom kedelai varietas Williams 82 (Schmultz *et al*., 2010) menghasilkan variasi DNA sebanyak 3.150.869 variasi (Tabel 6). Dari jumlah tersebut, 2.597.286 variasi SNP, 257.598 variasi insersi, dan sebanyak 202.157 variasi delesi. SNP ialah variasi DNA yang terjadi karena terjadinya perubahan satu basa yang berubah menjadi basa lainnya. Insersi ialah variasi DNA pada posisi tertentu pada genom berupa terjadinya penambahan basa pada genotipe yang sekuennya diamati jika dibanding dengan sekuen genom rujukan. Delesi ialah variasi DNA berupa terjadinya pengurangan atau penghilangan basa pada posisi tertentu pada genom kedelai yang diamati data sekuennya jika dibanding dengan sekuen genom rujukan. Perubahan variasi DNA bervariasi pada bagian genom yang berbeda. Tabel 7 menunjukkan data sebaran perubahan tersebut.

Sebagian besar perubahan terjadi pada bagian area genom di luar ekson, yaitu area antargen (*intergenic region*) (31,59%), bagian hulu gen (*upstream region*) (29,76%), bagian hilir gen (*downstream region*) (26,64%), dan intron (8,65%). Variasi yang ditemukan pada gen (ekson, intron, dan *untranslated region*/UTR) hanya 11,81%. Dengan demikian, variasi genom yang diamati sebagian besar terjadi di luar gen (88,19%). Hal ini disebabkan antara lain oleh bagian genom di luar sekuen gen umumnya mengandung banyak sekuen berulang (*repeated sequence*) yang frekuensi perubahan basa dari basa satu ke basa lainnya biasanya tinggi yang mengakibatkan tingkat perubahan (*change rate*) lebih tinggi pada area di luar gen dibanding dengan area gen.

Di antara variasi DNA yang diamati, hanya sebagian kecil variasi tersebut, yaitu sebanyak 95.154 variasi (2,15%) ada pada ekson (*protein coding region*) yang memengaruhi sifat-sifat fisiologis tanaman, termasuk karakter-karakter unggul tertentu pada kedelai. Dari analisis lanjutan terhadap variasi yang terdapat pada ekson tersebut, 49.892 variasi dapat menyebabkan mutasi DNA yang mengubah jenis asam amino (*missense mutation*) pada protein yang dihasilkan oleh gen tersebut dan sebanyak 1.497 variasi dapat menyebabkan mutasi DNA yang menghasilkan stop kodon (*nonsense mutation*) pada gen yang mengandung variasi DNA tersebut.

Data lokasi SNP di atas memberi keleluasaan memilih SNP untuk digunakan dalam mendesain *low*

Tabel 6. Variasi genom yang terdeteksi pada setiap kromosom kedelai hasil analisis penjajaran sekuen genom lima varietas kedelai Indonesia dengan sekuen genom rujukan kedelai varietas Williams 82.

Kromosom	Panjang sekuen (bp)	Jumlah basa berubah (bp)	Laju perubahan (*change rate*)
Gm01	55.915.595	164.050	340
Gm02	51.656.713	140.745	367
Gm03	47.781.076	182.784	261
Gm04	49.243.852	122.041	403
Gm05	41.396.504	102.278	410
Gm06	50.722.821	207.015	245
Gm07	44.683.157	137.522	324
Gm08	46.995.532	156.487	300
Gm09	46.843.750	152.358	307
Gm10	50.959.635	91.056	559
Gm11	39.172.790	115.306	339
Gm12	40.113.140	99.504	403
Gm13	44.408.971	168.210	264
Gm14	49.711.204	169.969	292
Gm15	50.939.160	218.143	233
Gm16	37.397.385	194.427	192
Gm17	41.906.774	164.865	254
Gm18	62.308.140	280.791	221
Gm19	50.589.441	148.520	340
Gm20	46.773.167	111.122	420
Total/rataan	972.068.482	3.150.869	312*

Gm = *Glycine max* (L.) Merr., bp = *base pair*/pasang basa.

*Laju perubahan rerata = 312, artinya rata-rata ditemukan satu variasi DNA pada setiap 312 basa pada genom kedelai.

Tabel 7. Lokasi terjadinya perubahan variasi DNA pada genom kedelai.

Tipe perubahan	Jumlah perubahan (bp)	Frekuensi perubahan (%)
Downstream	1.174.917	26,637
Exon	95.154	2,159
Intergenic	1.393.216	31,586
Intron	381.701	8,645
Splice site acceptor	537	0,012
Splice site donor	606	0,014
Upstream	1.312.546	29,757
UTR 3 prime	36.219	0,821
UTR 5 prime	15.985	0,362

UTR = *untranslated region*.

and high density chip SNP kedelai. Marka SNP yang berada di luar gen dipilih yang menyebar di seluruh genom kedelai yang mewakili bagian-bagian kromosom nongen. Biasanya sekitar 50% SNP untuk *chip* SNP dipilih dari area nongen ini untuk mewakili genom yang menyebar pada berbagai kromosom. Lima puluh persen sisanya dipilih dari area gen untuk mewakili area genom yang mengodekan protein penentu berbagai fenotipe tanaman kedelai. Kombinasi marka SNP dari dua area genom ini menghasilkan *chip* SNP yang bermanfaat untuk melabel karakter-karakter penting kedelai selain untuk uji kekerabatan dan pengelompokan genotipe kedelai. *Chip* SNP mempercepat penemuan gen, QTL, dan marka karakter penting untuk digunakan dalam program pemuliaan berbasis genom menggunakan teknologi *genome selection* dan *marker-assisted selection* (MAS).

Tipe variasi pada ekson (SNP atau indel), yang ada pada ekson yang dapat mengubah komposisi asam amino produk gen tersebut, perlu dipelajari lebih jauh melalui penelitian genomika fungsional (*functional genomics*) untuk mengisolasi gen-gen potensial, di antaranya gen-gen yang mengode adaptasi di lingkungan tropis, seperti insensitivitas terhadap panjang hari, ketahanan terhadap hama dan penyakit yang beradaptasi di daerah tropis, dan karakter penting lainnya.

Berdasarkan data genomik dua puluh kromosom kedelai, variasi DNA yang diperoleh dari penelitian ini bervariasi antar kromosom kedelai. Dari 972.068.482 pasang basa genom kedelai yang diamati, terdapat total perubahan (variasi DNA) sebanyak 3.150.869 basa atau rata-rata ditemukan satu variasi DNA dari setiap 312 basa dari genom kedelai (Tabel 6).

Perubahan tersebut dapat berupa SNP, insersi, atau delesi. Perubahan yang diamati bervariasi bergantung pada jenis dan ukuran kromosom. Jumlah variasi yang diamati umumnya proporsional dengan ukuran kromosom. Secara umum, semakin besar ukuran kromosom, semakin besar variasi DNA yang ditemukan. Variasi DNA terbanyak ditemukan pada kromosom 18 (Gm18) sebanyak 280.791 dan paling sedikit ditemukan pada kromosom Gm12 sebanyak 99.504 variasi (Tabel 6). Namun, tingkat perubahan variasi DNA (*change rate*) tidak sama antar kromosom. Tingkat perubahan terbesar ditemukan pada kromosom Gm18 (ditemukan satu perubahan pada setiap 221 basa DNA) dan terkecil ditemukan pada kromosom Gm10 (ditemukan satu perubahan pada setiap 559 basa DNA) (Tabel 6). Ini menunjukkan bahwa komposisi DNA setiap kromosom tidak sama. Tingkat perubahan yang besar terjadi pada hulu (*upstream*) dan hilir (*downstream*) sekuen gen (Tabel 7). Segmen-segmen DNA tersebut di antaranya merupakan sekuen berulang (*repeated sequences*) yang tingkat mutasinya pada level nukleotida biasanya tinggi. Semakin tinggi suatu kromosom mengandung sekuen berulang, semakin besar peluang perubahan variasi DNA yang diamati pada kromosom tersebut.

Penelitian ini menghasilkan lebih dari 2,5 juta SNP dan sekitar 460 ribu indel. Jumlah ini cukup besar untuk tanaman menyerbuk sendiri seperti kedelai. Besarnya jumlah variasi yang ditemukan karena materi genetik yang digunakan pada penelitian ini kekerabatannya jauh seperti ditunjukkan oleh Satyawan *et al*. (2014) yang membandingkan kelima genotipe kedelai ini dengan genotipe kedelai asal Tiongkok yang menunjukkan bahwa aksesi kelima genotipe berkerabat relatif jauh. Kekerabatan yang jauh menghasilkan variasi SNP dan indel yang lebih banyak dibanding dengan genotipe yang berkerabat dekat.

Data SNP hasil penelitian ini merupakan sumber daya marka DNA dalam jumlah yang besar sebagai bahan dasar untuk pembentukan *chip* SNP, baik yang densitasnya rendah (*low density SNP chip*) maupun yang densitasnya tinggi (*high density SNP chip*). *Chip* SNP digunakan untuk melabel gen dan QTL karakter penting dengan cepat melalui analisis asosiasi menggunakan *high throughput genotyping system* pada kedelai. SNP dan indel yang ada pada ekson (*protein coding region*) perlu diteliti lebih jauh untuk penemuan gen-gen potensial dalam rangka perbaikan varietas kedelai nasional. Marka dan gen tersebut dapat digunakan dalam pemuliaan kedelai dengan MAS dan *genomic selection system* untuk mendukung percepatan pemuliaan kedelai nasional.

KESIMPULAN

Pustaka genom genotipe kedelai lokal Indonesia yang berukuran 400 bp telah dikonstruksi dengan konsentrasi berkisar antara 21,2 ng/μl dan 64,5 ng/μl yang sangat ideal untuk penelitian sekuensing genom total menggunakan NGS HiSeq. Jumlah basa yang dihasilkan selama proses sekuensing sebanyak 50,1 x 10^9 bp dengan kualitas dan karakteristik sekuen kualitas tinggi dan tingkat kesalahan pembacaan basa yang sangat rendah (0,97%). Penjajaran data resekuen lima genotipe kedelai Indonesia dengan sekuen genom rujukan kedelai varietas Williams 82 menghasilkan variasi DNA sebanyak 3.150.869. Variasi DNA tersebut terdiri atas 2.597.286 variasi SNP, 257.598 variasi insersi, dan 202.157 variasi DNA delesi. Hanya sebagian kecil variasi tersebut (2,15%) berlokasi pada ekson (*protein coding region*) yang dapat mengubah asam amino protein atau menghasilkan stop kodon.

UCAPAN TERIMA KASIH

Penelitian ini dibiayai dari APBN BB Biogen TA 2011–2012. Tim penulis mengucapkan banyak terima kasih kepada Sdr. Andi Kosasih dan Ratna Utari atas keterlibatannya dalam penelitian ini.

DAFTAR PUSTAKA

Ansorge, W.J. 2009. Next generation DNA sequencing techniques. Nat. Biotechnol. 25:195–203.

Cingolani, P., A. Platts, and M. Coon. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6:80–92.

Commins, J., C. Toft, and M.A. Fares. 2011. Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insect. Biol. Proced. Online 11:52–78.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186–194.

Illumina. 2009. Sequencing analysis software: User guide. Illumina Inc., San Diego.

Langmead, B. and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Kosasih, A. 2012. Konstruksi dan analisis kualitas pustaka genom kedelai (*Glycine max* [L.] Merr.) untuk sekuensing genom total. Skripsi S1, Institut Pertanian Bogor, Bogor.

Leonforte, A., S. Sudheesh, N.O.I. Cogan, P.A. Salisbury, M.E. Nicolas, M. Materne, J.W. Forster, and S. Kaur. 2013. SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity

tolerance in field pea (*Pisum sativum* L.). BMC Plant Biol. 13:161–174.

Li, H., B. Handsaker, and A. Wysoker. 2009. The sequence alignment/map format and SAM tools. Bioinformatics 25:2078–2079.

Mardis, E.R. 2008. Next generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9:387–402.

Metzker, M.L. 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. 11:31–46.

Michiels, A., W. Van den Ende, M. Tucker, L. Van Riet, and A. Van Laere. 2003. Extraction of high quality genomic DNA from latex containing plants. Anal. Biochem. 315:85–89.

Pattersson, E., J. Lundeberg, and A. Ahmadian. 2009. Generations of sequencing technologies. Genomics 93:105–111.

Perrier, X. and J. Jacquemoud-Collet. 2006. DARwin software. http://darwin.cirad.fr/ (diakses 3 Jan. 2012).

Quail, S. 2008. A large genome center's improvements to the Illumina sequencing system. Nat. Methods 5:1005–1010.

Santoso, T.J., D.W. Utami, dan E.M. Septiningsih. 2006. Analisis sidik jari DNA plasma nutfah kedelai menggunakan markah SSR. J. AgroBiogen 2(1):1–7.

Satyawan, D. and I.M. Tasma. 2011. Genetic diversity analysis of *Jatropha curcas* provenances using randomly amplified polymorphic DNA markers. J. AgroBiogen 7(2):47–55.

Satyawan, D., H. Rijzaani, and I.M. Tasma. 2014. Characterization of genomic variation in Indonesian soybean (*Glycine max*) varieties using next-generation sequencing. Plant Genet. Resour. 12:S109–S113.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G.D. May, Y. Yu, T. Sakurai, T. Umezawa, M.K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. Zhang, K. Shinozaki, H.T. Nguyen, R.A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R.C. Shoemaker, and S.A. Jackson. 2010. Genome sequence of the paleopolyploid soybean. Nature 463:178–183.

Schuster, S.C. 2008. Next generation sequencing transform today's biology. Nat. Methods 5:16–18.

Suharsono. 2002. Konstruksi pustaka genom kedelai kultivar Slamet. Hayati 9:67–70.

Surzcky, S. 1990. A Fast method to prepare random fragment sequencing libraries using a new procedure of DNA shearing by nebulization and electroporation. The International Conference on the Status and Future of Research on the Human Genome. Human Genome II. San Diego.

Tasma, I.M. 2014. *Single nucleotide polymorphism* (SNP) sebagai marka DNA masa depan. Warta Biogen 10(3):7–10.

Tasma, I.M., A. Warsun, and Asadi. 2008. Development and characterization of F_2 population for molecular mapping of aluminum-toxicity tolerant QTL in soybean. J. AgroBiogen 4(1):1–8.

Väli, Ü., M. Brandström, M. Johansson, and H. Ellegren. 2008. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet. 9:8. doi:10.1186/1471-2156-9–8.

Voelkerding, K.V., S.A. Dames, and J.D. Durtschi. 2009. Next generation sequencing: From basic research to diagnostics. Clin.Chem. 55(4):641–658.

Wahyudi, A.T. 2001. Perpustakaan gen: Bagaimana mengontruksinya? Hayati 8:27–30.

Wulandari, Y.R.E. 2009. Konstruksi pustaka genom *Tibouchina langsdorffiana* Baill. Tesis S2, Institut Pertanian Bogor, Bogor.

Zhang, J., R. Chiodini, A. Badr, and G. Zhang. 2011. The impact of next generation sequencing on genomics. J. Genet. Genomics 38:95–109.