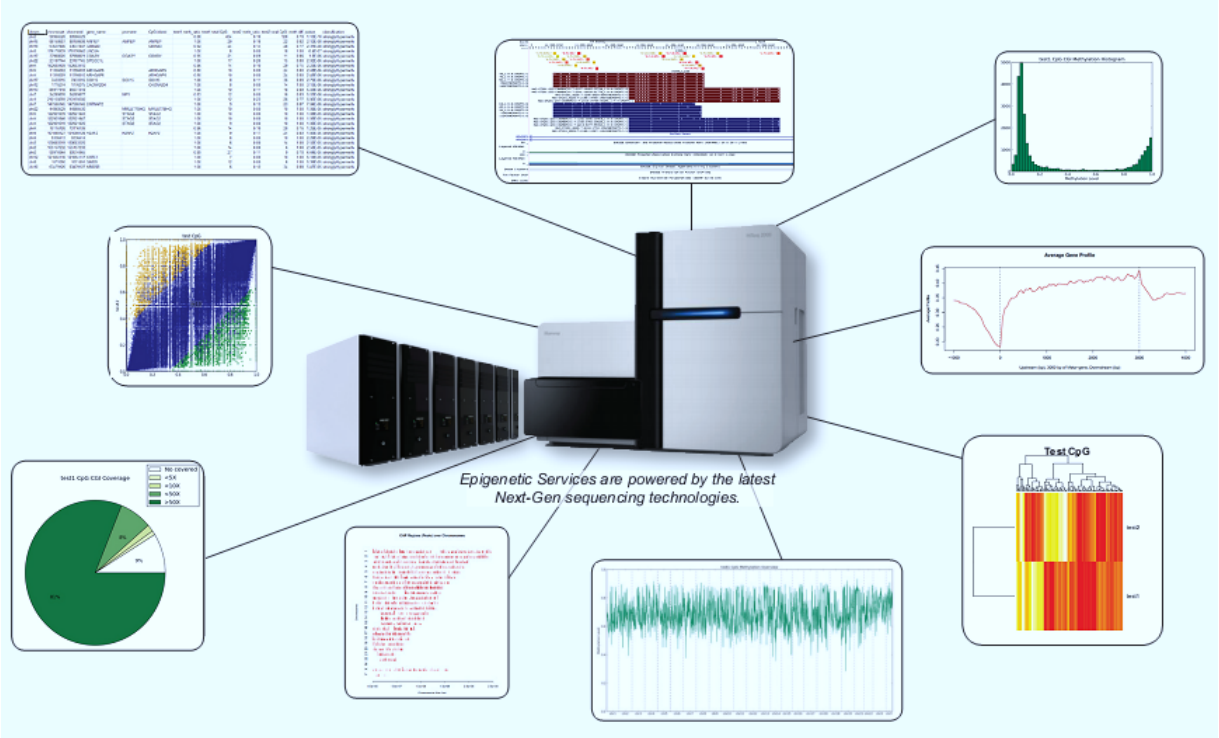
Bioinformatika pada teknologi sekuensing generasi baru

*Pada masa sekarang ini hingga satu dekade ke depan, para peneliti akan dihadapkan pada tantangan penyimpanan, pengolahan dan analisis data besar (*big data*) yang dihasilkan dari teknologi* Next-Generation Sequencing *(NGS). Analisis sekuen genom menggunakan pendekatan NGS merupakan proses konversi dari materi biologis belum bermakna (sampel DNA atau RNA) menjadi kode informasi (kode biner dan kode basa nukleotida) dan diterjemahkan menjadi informasi biologis bermakna. Bioinformatika terlibat secara nyata dan memiliki peran sangat penting dalam rangka mempermudah setiap tahapan analisis NGS.*



Ilustrasi pemanfaatan teknologi NGS dalam analisis ekspresi gen, kuantifikasi metilasi genom hingga pemetaan kromosom. Sumber gambar: http://www.epibeat.com.

Bioinformatika merupakan bagian yang tidak terpisahkan dari teknologi sekuensing generasi baru atau yang sering disebut *Next-Generation Sequencing* (NGS). Pada masa sekarang ini hingga satu dekade ke depan, para peneliti akan dihadapkan pada tantangan penyimpanan, pengolahan dan analisis data besar (*big data*) yang dihasilkan dari teknologi NGS [1]. Setidaknya, empat tahapan wajib dilakukan dalam analisis sekuen nukleotida menggunakan *platform* NGS yaitu (1) pemanggilan basa (*base calling*) pasca sekuensing, (2) penjajaran dan penggabungan sekuen (*assembly*), (3) anotasi sekuen (*annotation*), dan (4) integrasi data bioinformatika menjadi data biologis (*data integration*).

Pertama, analisis pemanggilan basa (*base calling*) pada potongan pendek sekuen (100-500 pasangan basa) yang disebut *reads* pasca prosesi sekuensing genom atau transkrip dilakukan menggunakan piranti lunak berbasis modul perintah (*Command Line Interface*, CLI). **Kedua**, *contig* dan *scaffold* disusun menggunakan penjajaran dan penggabungan (*assembly*) sekuen nukleotida pendek tersebut. *Contig* dan *scaffold* merupakan bentuk gabungan *reads* yang umumnya memiliki panjang ratusan hingga ratusan ribu pasang basa. **Ketiga**, anotasi dan visualisasi sekuen *contig* dan *scaffold* kemudian dilakukan menggunakan piranti lunak

berbasis grafis (*Graphical Processing Unit, GPU*) seperti BLAST2GO. **Keempat**, keseluruhan analisis bioinformatika dari *platform* NGS seperti GO (*Gene Ontology*), KEGG (*Kyoto Encyclopedia of Genes and Genomes*) hingga DGE (*Differential Gene Expression*) diintegrasi menjadi satu luaran informasi yang koheren serta memberikan makna biologis [2]. Saat ini, beberapa *tool* bioinformatika untuk analisis data NGS tersedia secara daring dan tidak dipungut biaya, seperti *Galaxy* (https://usegalaxy.org/) dan *Genomic tools* (http://molbiol-tools.ca/Genomics.htm) [3].

Analisis bioinformatika terlibat bahkan sejak sekuen nukleotida mentah (*raw nucleotide sequence*) dihasilkan dari mesin sekuensing seperti Illumina, SFF, HDF5, CG atau SOLID). Proses sekuensing sendiri mendeteksi basa-basa nukleotida dan mengubahnya secara komputasi menjadi data *reads*. Sistem format FASTQ digunakan untuk mengukur kualitas dari sekuen *reads* yang dihasilkan. Pengukuran tersebut pada dasarnya adalah memberikan penilaian apakah basa yang terbaca akurat atau tidak. Data dalam bentuk FASTQ sukar untuk digunakan pada peneliti di laboratorium karena data berukuran besar dan masih berbentuk kode angka dan karakter, bukan dalam bentuk kode basa nukleotida. Oleh sebab itu, data FASTQ pada umumnya dikonversi menjadi bentuk kompak yang disebut SAM (*Sequence Alignment Map*) dan kemudian lebih dikompres menjadi BAM (*Binary Alignment Map*) [2].

Penjajaran sekuen *reads* untuk membentuknya menjadi sekuen yang lebih panjang berupa *contig* dan/atau *scaffold* dapat dilakukan dengan dua pendekatan berbasis bioinformatika yaitu (1) pemetaan komparatif (*comparative mapping*) dimana sekuen *reads* disejajarkan dengan sekuen genom referensi (*reference genome*) dan (2) penggabungan sekuen dengan memanfaatkan sekuen *reads* yang saling tumpang tindih (*overlapping reads*). Pendekatan kedua tersebut lebih dikenal sebagai *de novo assembly* [4]. Saat ini, berbagai genom referensi telah tersedia untuk spesies baik hewan dan tanaman. Umumnya, para peneliti yang melakukan pendekatan *de novo assembly* akan tetap membandingkan dan mengkonfirmasi sekuen mereka menggunakan referensi genom yang telah ada baik pada spesies yang sama maupun pada spesies terdekat. Dalam mengkonfirmasi hasil *assembly* secara manual, para peneliti menggunakan piranti lunak seperti Tablet® untuk memvisualisasikan sekuen *contig* atau *scaffold* mereka [5].

Pada tahapan berikutnya, bioinformatika dibutuhkan lebih spesifik dalam menganotasi, mengubah dan menerjemahkan sekuen nuleotida menjadi informasi genomika tingkat tinggi seperti menentukan daerah penyandi protein (*coding sequence,* CDS), daerah yang tidak menyandi protein (*non coding*), bentuk isoform sekuen mRNA, sinyal peptida, dan elemen repetitif (*repeat elements*) [6]. Pada organisme eukariotik yang memiliki struktur genom lebih kompleks dibanding organisme prokatiotik, anotasi genom menjadi lebih sulit dan menantang. Analisis pada genom yang lebih kompleks umumnya dilakukan menggunakan rangkaian proses anotasi yang disebut GAP (*genome annotation pipeline*). Pada basis data publik, *The NCBI Eukaryotic Genome Annotation Pipeline* tersedia untuk genom eukariotik (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/). GAP tersebut terdiri dari langkah pertama berupa identifikasi dan mengeliminasi elemen genom repetitif (mikrosatelit, retrotransposon dan transposon) menggunakan *RepeatMasker*, *Censor* atau *WindowMasker* [7]. Langkah pertama tersebut penting untuk menyaring sekuen repetitif yang dapat mengganggu analisis BLAST dalam anotasi sekuen penyandi protein. Langkah berikutnya dari GAP adalah anotasi transkrip, penjajaran protein/domain, prediksi model gen, penamaan gen dan lokus, dan pemberian GeneID [8].

Analisis bioinformatika tingkat lanjut pada NGS harus menghasilkan makna biologis dari sekuen genom. Salah satu analisis tingkat lanjut tersebut adalah GO (*Gene Ontology*). GO menyediakan terminologi dari gen, interaksi protein-protein yang terlibat sebagai komponen seluler, fungsi molekuler dan proses biologis terkait. Dalam sejarahnya, analisis GO dimulai hanya dari basis data dari tiga organisme model yaitu *FlyBase* (*Drosophila*), the *Saccharomyces Genome Database* (SGD) and the *Mouse Genome Database* (MGD). Saat ini, Konsorsium Kontributor GO telah terbentuk dan terus menambahkan basis data baru dari organisme yang genomnya baru disekuen. Daftar contributor GO tersebut dapat diakses melalui tautan: http://geneontology.org/page/go-consortium-contributors-list. Salah satu konsorsium pada tanaman adalah *The Plant Ontology* yang direpresentasikan oleh *Phytozome* (https://phytozome.jgi.doe.gov/pz/portal.html) [9]. Selain GO, analisis ontologi serupa juga disediakan oleh beberapa provider yang berbeda seperti: *the Open Biological and Biomedical Ontologies* (OBBO), *Reactome*, *DAVID*, and *the KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database*.

Pada akhirnya, analisis sekuen genom menggunakan pendekatan NGS merupakan proses konversi dari materi biologis belum bermakna (sampel DNA atau RNA) menjadi kode informasi (kode biner dan kode basa nukleotida) dan diterjemahkan menjadi informasi biologis bermakna. Bioinformatika terlibat secara nyata dan memiliki peran sangat penting dalam rangka mempermudah setiap tahapan analisis NGS. Bioinformatika terintegrasi dalam NGS telah menjadi bagian tidak terpisahkan dalam rangka memberikan makna biologis pada sekuen.

Referensi

1. Horner DS, Pavesi G, Castrignanò T, De Meo PDO, Liuni S, Sammeth M, et al. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Briefings in Bioinformatics. 2010;11(2):181-97.
2. Kulski JK. Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. In: Kulski DJ, editor. Next Generation Sequencing - Advances, Applications and Challenges: InTech; 2016.
3. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 2016;44(W1):W3-W10.
4. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671-82.
5. Milne I, Bayer M, Stephen G, Cardle L, Marshall D. Tablet: Visualizing Next-Generation Sequence Assemblies and Mappings. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 253-68.
6. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(5):329-42.
7. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006;22(2):134-41.

8. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Research. 2012;40(Database issue):D130-D5.
9. Cooper L, Jaiswal P. The Plant Ontology: A Tool for Plant Genomics. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 89-114.