BIOKIMIA DI ERA *BIG DATA* GENOMIK: TANTANGAN, APLIKASI DAN PELUANG INOVASI

FREDY Z. SAUDALE*

Program Studi Kimia FST Undana, Jl. Adi Sucipto Penfui Kupang Indonesia

Article Received: 23 November 2020

Article Accepted: 01 December 2020

Abstract

The completion of human genome project at beginning of 21st century with the advancement of computer technology has transformed Biochemistry into a genomic era. Further, it is accelerated by parallel and massive genome sequencing technology known as next generation sequencing (NGS) that enhances the identification of genetic variants associated with complex diseases such as cancer, diabetes and Alzheimer. Currently, this knowledge has been driving the development of precision and personalized medicine. Wisely applied, it is believed that the explosion of genomic big data can be of great use in advancing the diagnosis, therapy and drug discovery to combat complex diseases.

Keywords: Biochemistry, big data, genomics, cloud computing

Abstrak

Selesainya pemetaan DNA manusia pada awal abad ke-21 bersamaan dengan kemajuan komputer telah mentransformasi Biokimia ke dalam era genomik. Ini dipercepat dengan kemajuan teknologi pengurutan DNA secara paralel dan masif yang dikenal dengan *next generation sequencing* (NGS) yang telah menghasilkan ledakan maha data yang memampukan ilmuwan untuk mengidentifikasi varian-varian genetika penting yang berkaitan dengan perkembangan penyakit kompleks seperti kanker, diabetes maupun Alzheimer. Diterapkan dengan bijak, sinergi antara Biokimia dan maha data dapat memacu inovasi dalam diagnosis, terapi, desain dan penemuan obat menuju pengobatan yang lebih terpresisi dan personal berdasarkan informasi genom masing-masing individu.

Kata kunci: Biokimia, maha data, genomik, komputasi awan

Pendahuluan

Revolusi sains di awal abad 21 melalui selesainya proyek pemetaan DNA manusia (*human genome project*) disertai perkembangan komputer telah mentransformasi Biokimia ke dalam visi dan era genomik. Ini semakin diakselerasi melalui kemajuan teknologi sekuensing generasi baru secara parallel dan masif atau yang dikenal dengan *next generation sequencing* (NGS) yang tidak hanya mampu mengidentifikasi mutasi penyebab varian pada satu gen tapi juga banyak gen yang terkait penyakit secara bersamaan, lebih mendalam, cepat dan murah [1,2].

**Corresponding Author: Jl. Adisucipto-Penfui Kupang 85110 telp.(+62380)8037977, e-mail: fredy_saudale@staf.undana.ac.id*

Saat ini NGS telah dan sedang merevolusi bidang kedokteran klinis dalam hal diagnosis dan monitoring untuk pemberian terapi yang lebih tepat terhadap penyakit-penyakit seperti infeksi[3], kanker[4], diabetes[5], dan neurodegeneratif[6]. Namun demikian, NGS telah menghasilkan ledakan mahadata genomik (*genomic big data*) yang belum pernah ada sebelumnya yang memberikan tantangan khusus[7]. Ketersediaan teknologi komputasi yang dapat dimanfaatkan untuk menyimpan, mengolah dan kolaborasi data besar tersebut sangat penting diperlukan untuk kepentingan riset[8]. Tantangan lainnya adalah studi prediksi dan korelasi dalam penelitian menggunakan mahadata genomik NGS dengan menerapkan metodologi statistika dan matematika tidak serta merta memberikan pengetahuan tentang hubungan kausalitas atau sebab akibat antara varian dan penyakit. Tambahan pula banyak varian-varian yang sebelumnya dianggap berkorelasi dengan penyakit tertentu ternyata tidak menunjukkan pengaruhnya. Akibatnya identifikasi dan validasi varian patogenik secara akurat masih menjadi permasalahan utama.

Tinjauan ilmiah ini bertujuan memberikan gambaran terkini sinergi antara bidang ilmu Biokimia dengan Genetika dan Biologi Molekuler yang menjadi dasar munculnya era genomik, juga tantangan yang diberikan melalui kelimpahan maha data genomik hasil NGS dan bagaimana teknologi komputasi awan (*Cloud*) saat ini sedang dimanfaatkan untuk manajemen, penyimpanan dan kolaborasi data. Di bagian akhir akan dibukakan 4 tren aplikasi dimana Biokimia melalui kolaborasinya dengan keilmuan terkait seperti Genetika, Biologi Molekuler, Bioinformatika, Biofisik, Kimia Komputasi, Farmakogenomik dan Biologi Komputasi dapat memainkan peranan yang signifikan dalam pemanfaatan mahadata genomik.

Melalui telaah ilmiah ini penulis berpendapat bahwa era data besar genomik, ditengah tantangan yang masih ada, memberikan peluang kepada ilmu Biokimia dalam penerapan teknik dan metodologinya juga perluasan kolaborasinya dengan keilmuan yang lain dalam upaya identifikasi dan validasi mekanistik terhadap varian-varian patogenik dengan mengintegrasikan pendekatan komputasional (*in silico*), modeling, simulasi maupun eksperimental. Dengan identifikasi dan validasi yang *solid* ini maka mahadata genomik dapat dimanfaatkan sebaik-baiknya untuk pengembangan diagnosis dan terapi medis yang lebih akurat, serta desain obat yang selektif terhadap penyakit-penyakit yang masih eksis sehingga bisa membawa manfaat bagi masyarakat luas.

Dari Enzim Ke Gen

Di awal abad ke-20, Biokimia dan Genetika merupakan keilmuan yang masih berdiri sendiri. Studi yang mendominasi Biokimia pada saat itu salah satunya berpusat pada pencarian enzim yang mentransformasi perubahan substrat menjadi produk yang dapat diidentifikasi[9,10].

Sementara itu, Genetika klasik fokus mempelajari hereditas yang ditentukan oleh faktor keturunan yang ditemukan Gregor Mendel di tahun 1865[11]. Adalah Archibald Garrod, seorang dokter dan biokimiawan asal Inggris pada tahun 1902, yang pertama kali menggambarkan adanya hubungan antara faktor keturunan (yang kemudian diketahui sebagai gen) dengan enzim yang mengkatalisis proses biokimiawi tubuh[12]. Pada masa itu, sifat dan karakteristik tentang apa itu faktor keturunan atau gen belum sepenuhnya dipahami. Garrod menangani pasien penderita penyakit metabolisme alkaptonuria yang merupakan gangguan non-fatal di mana urin berubah warna menjadi hitam karena ketidakmampuan tubuh memecah molekul yang disebut alkapton (yang pada orang normal dipecah menjadi molekul lain yang tidak berwarna). Garrod berpendapat bahwa penyakit alkaptonuria disebabkan oleh cacat metabolisme bawaan (*inborn errors of metabolism*) yang disebabkan oleh bentuk resesif dari salah satu faktor keturunan yang digagas Mendel.

Pada tahun 1941, Biokimia dan Genetika akhirnya dipersatukan melalui hipotesis yang dikenal dengan “satu gen-satu enzim” (*one gene-one enzyme hypothesis*) yang menyatakan bahwa setiap gen mengkode satu enzim yang bertanggungjawab dalam mengkatalisis satu reaksi metabolisme[13]. Hipotesis ini didasarkan atas penelitian George Bradle dan Edward Tatum menggunakan jamur kapang *Neurospora crassa*[14]. Bradle dan Tatum menunjukkan bahwa mutan jamur kapang hasil iradiasi tidak bisa tumbuh di dalam medium dengan kondisi nutrisi minimal yang didesain cukup untuk jamur bisa berkembang[15]. Penambahan asam amino arginin pada media nutrien menyebabkan mutan jamur kapang dapat kembali bertumbuh yang mengindikasikan mutasi terjadi pada gen dari enzim yang berperan dalam biosintesis arginin. Penemuan Bradle dan Tatum ini memperkuat hipotesis Garrod bahwa proses metabolisme dikontrol oleh enzim yang dikodekan oleh satu gen[16].

Pada tahun 1953, melalui penemuan struktur DNA oleh James Watson dan Francis Crick, Biokimia kemudian dipersatukan dengan Biologi Molekuler yang mempelajari bagaimana informasi genetika tersimpan di dalam struktur biomolekul yaitu DNA yang kemudian di transfer ke RNA lalu Protein atau yang dikenal dengan Dogma Sentral[17,18]. Penemuan kode genetik oleh Francis Crick, Leslie Barnett, Sydney Brenner, Richard Watts-Tobin pada tahun 1961 dan Marshall Nirenberg pada tahun 1966 memotivasi ilmuwan lain untuk menyelidiki efek perubahan susunan basa-basa DNA terhadap urutan asam amno yang dapat merubah fungsi protein menjadi bersifat patogenik[19]. Pada tahun 1957, Vernon Ingram adalah ilmuwan yang pertama kali membandingkan hemoglobin dari sel normal dengan sel sabit anemia (*anemia sickle cells*) dan menemukan bahwa perbedaannya hanyalah pada perubahan satu asam amino[20]. Sel sabit anemia terbentuk oleh karena mutasi nukelotida tunggal pada kode DNA yang mengubah asam glutamat (GAG) menjadi valin (GUG). Individu yang memiliki 2 kopi gen sel sabit anemia tidak bisa memproduksi hemoglobin secara normal yang berakibat menghambat sel darah merah yang membawa oksigen

dari paru-paru ke seluruh bagian tubuh. Penemuan Ingram menunjukkan bagaimana mutasi pada DNA dapat mengubah asam amino tunggal yang menyebabkan protein menjadi patogenik[21]. Di tahun-tahun berikutnya, pengetahuan ini kemudian mengakselerasi pencarian penyakit metabolik lainnya yang disebabkan oleh mutasi pada satu gen (*mendelian monogenetic disease*) seperti talasemia dan sistik fibrosis[22]. Namun demikian fakta lain yang ditemukan adalah banyak penyakit kompleks muncul akibat interaksi maupun mutasi beberapa gen (*polygenic complex disease*) yang juga melibatkan faktor nongenetik lingkungan[23]. Sehingga untuk mengurai kompleksitas seperti itu dibutuhkan peta lengkap tentang variasi genetik dari keseluruhan DNA manusia (*genome*). Hal inilah yang kemudian menggerakan usaha untuk memetakan keseluruhan DNA manusia yang dimulai pada tahun 1990 yang kemudian membawa Biokimia bersama Genetika dan Biologi Molekuler masuk ke era baru genomik.

Era Genomik

Selesainya Proyek Genom Manusia (*Human Genome Project*) di tahun 2003 menandakan dimulainya era genomik[24]. Proyek genom manusia berhasil memetakan sebanyak 3,2 miliar (3,2 x 10^9) pasangan basa-basa kimia yaitu guanin (G), sitosin (C), adenin (A), dan timin (T) yang membentuk sekuen DNA di dalam 23 pasang kromosom dari inti sel manusia[25]. Sekitar 20.000-30.000 sekuen DNA atau gen mengkode pembentukan protein fungsional pada organisme hidup[26].

Proyek genom manusia membutuhkan waktu 13 tahun dengan memakan anggaran sebesar $ 2.7 milyar melalui kolaborasi 200 institut penelitian di Amerika Serikat dan juga kontribusi dari 18 negara diantaranya Inggris, Jerman, Jepang dan Cina [27,28]. Presiden Bill Clinton menyambut selesainya proyek pemetaan DNA manusia dengan pernyataan, “Tidak diragukan, ini adalah peta yang paling penting dan menakjubkan yang pernah dihasilkan oleh manusia. Hari ini kita mempelajari bahasa di mana Tuhan menciptakan kehidupan.”[29]. Namun demikian, memetakan genom manusia barulah langkah awal dalam upaya memahami instruksi biologis tentang kehidupan yang dikodekan dalam DNA [30].

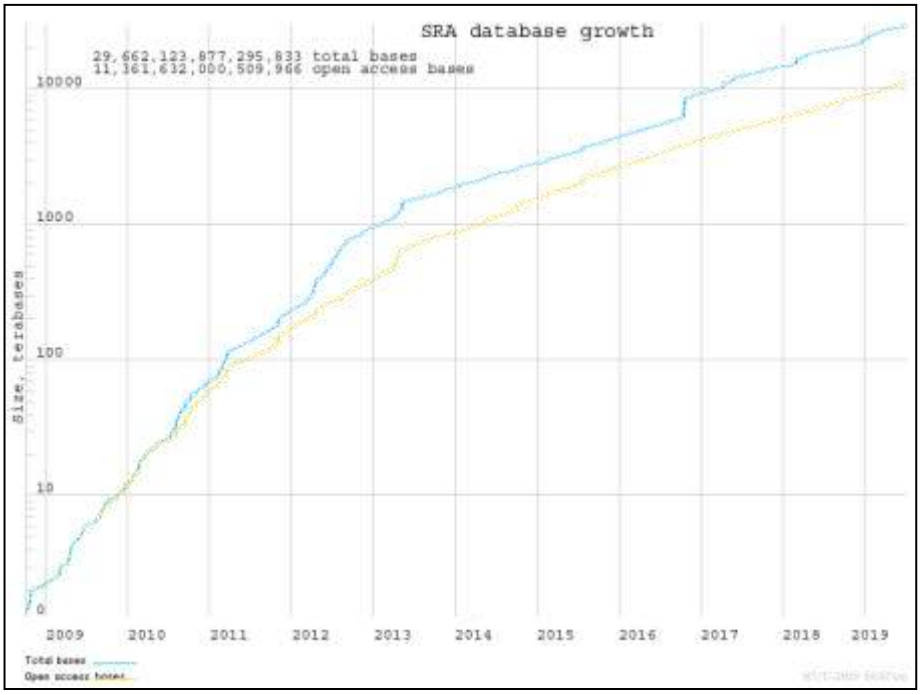
Jika pemetaan genom manusia adalah revolusi sains terbesar dalam 20 tahun terakhir, maka pencapaian terbesar sains dalam 20 tahun ke depan adalah bagaimana mendapatkan pengetahuan yang bermakna dari data genom manusia tersebut [25,31–33]. Pengetahuan tersebut dipercaya akan merevolusi bidang kesehatan melalui inovasi pengobatan yang lebih terpresisi dan personal dalam penanganan penyakit yang lebih kompleks[34–36]. Dalam visi pengobatan presisi dan personal, profil genom atau DNA pasien, disamping informasi klinis, menjadi standar medis yang memandu dalam upaya diagnosis, monitoring dan pemberian terapi penyakit seperti infeksi, kanker, diabetes dan neurodegeneratif[37,38]. Peta lengkap dan detail dari genom manusia dipercaya

akan sangat berguna dalam menjelaskan korelasi antara variasi genetik dengan resiko terhadap penyakit dan respon terhadap obat-obatan[39,40].

Tantangan Mahadata Genomik

Untuk mencapai tujuan tersebut jutaan bahkan milyaran data genom telah berhasil dipetakan menggunakan teknologi sekuensing generasi baru atau *Next Generation Sequencing* (NGS)[41]. Melalui NGS pemetaan genom bisa dilakukan dalam hitungan hari dengan jangkauan pengurutan basa-basa dari keseluruhan gen atau DNA organisme (*genome*) secara lebih luas, dalam dan detail atau yang dikenal dengan istilah *Whole Genome Sequencing* (WGS)[42,43]. Teknologi NGS diprediksi akan menurunkan biaya pemetaan semua gen atau DNA manusia menjadi $ 1000 per genom[44–46]. Selain itu NGS juga mampu mendapatkan data biologis yang melimpah dari hasil pengurutan keseluruhan segmen gen yang mengkode protein atau *Whole Exome Sequencing* (WES) dan juga urutan sekuen RNA (RNA seq) [47,48].

Namun demikian, sebagai konsekuensinya, kecepatan dan terjangkaunya biaya NGS telah meningkatkan produksi data sekuen mentah (*raw sequencing data*) dalam jumlah yang belum pernah ada sebelumnya. Sebanyak 29,6 Peta (29,6 x 10^{15}) total basa telah dihasilkan dari 5.331.832 sampel biologis melalui 204.429 studi penelitian dalam 10 tahun terakhir (Gambar 1).



Gambar 1. Pertumbuhan total basa yang berhasil dipetakan (kurva garis warna biru) dan yang bebas diakses publik (kurva garis warna kuning) dan yang tersimpan di database *Sequence Read Archive* (SRA). (Sumber: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/, di akses 5 Juli 2019)

Data ini tersimpan dan bisa diakses oleh publik melalui pangkalan data *Sequence Read Archive* (SRA) yang dikelola Pusat Nasional Informasi Bioteknologi (NCBI) di Washington. SRA menyimpan data pengurutan mentah dan informasi penjajaran dari platform atau instrumen NGS seperti Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, dan Pacific®. Jumlah data ini diprediksi akan terus bertambah dua kali lipat setiap 10-20 bulan menandakan hadirnya era maha data biologis *(Biological Big Data)* [49,50].

Peningkatan jumlah dan kompleksitas maha data biologis secara eksponensial telah membawa tantangan besar dalam hal manajemen, penyimpanan, akses dan analisis data[7]. Hal ini karena analisis mahadata genomik membutuhkan proses komputasi yang sangat intensif, ruang penyimpanan data yang besar dan juga investasi infrastruktur komputer yang tidak sedikit[51,52]. Analisis dan transfer data digital genomik dengan ukuran *Gigabyte*, *Terabyte* bahkan *Petabyte* tampak mustahil dilakukan dengan model komputasi tradisional melalui seperangkat *desktop* dan *software* yang terinstalasi secara lokal. Tantangan ini telah mendorong kolaborasi dengan ilmu komputer dan IT melalui pengadopsian teknologi komputasi awan atau *cloud* [8,53].

Bantuan Komputasi Awan

Komputasi awan atau *Cloud* menawarkan banyak keuntungan dan fasilitas komputasi yang dibutuhkan dalam riset genomik maupun biomolokuler di era data besar[54]. *Cloud* menyediakan perangkat komputasi virtual (*virtual machine*) dengan prosesor, memori, kapasitas penyimpanan skala besar yang dapat diakses sesuai kebutuhan (*on demand*) melalui internet dimana saja[55]. Dengan ini institusi penelitian tidak perlu lagi melakukan investasi infrastruktur yang besar untuk membangun dan memelihara server komputasi seperti superkomputer maupun *High Performace Computing*[56]. Dengan *cloud*, beban biaya modal bisa dialihkan ke beban biaya operasional[57]. Pengguna *cloud* pada dasarnya hanya membayar sewa perangkat komputasi virtual berdasarkan waktu dan jenis layanan yang digunakan (*pay-as-you-go*) dengan harga yang relatif terjangkau[58]. Dengan *cloud* juga pengguna bisa mengatur kapasitas perangkat komputasi virtual yang diperlukan secara proporsional sesuai jumlah data yang dianalisis untuk meningkatkan daya analisis dan ruang penyimpanan data[59]. Fleksibilitas dari *cloud* ini mengatasi kekurangan kluster komputer lokal dengan memberikan sumber daya perangkat virtual tak terbatas dan elastis. *Cloud* juga memberikan jaminan keamanan data dengan memberikan pengguna kontrol penuh terhadap akses masuk (*log in*), penggunaan enkripsi data, *firewall*, perlindungan kata sandi dan transfer data secara aman.

Beberapa vendor komersial penyedia layanan komputasi awan diantaranya adalah Amazon Web Services – Elastic Cloud Computing (AWS EC2), Google App Engine, Microsoft

Windows Azure, dan Alibaba Cloud. Dengan melakukan pendaftaran ke dalam salah satu penyedia layanan komputasi awan tersebut pengguna bisa memilih kapasitas mesin komputasi virtual seperti sistem operasi, jumlah prosesor, ukuran memori, dan kapasitas *hard disk* sesuai kebutuhan dan anggaran. Sebagai gambaran AWS EC2 mengenakan biaya sebesar $ 0.723 per jam untuk pengoperasian mesin komputasi berkapasitas 16 CPU virtual dan 64 GB RAM dan $ 0.0021 per 50 GB per bulan untuk penyimpanan data[60]. Melalui penyediaan infrastruktur virtual dengan penyimpanan tak terbatas dan pemrosesan data berkinerja tinggi, *cloud* saat ini telah dimanfaatkan dalam mempercepat penelitian juga inovasi di bidang kesehatan dan desain pengembangan obat-obatan baru di era data besar biologis.

Aplikasi

Dibawah ini akan dijelaskan 4 area penerapan yang sedang menjadi tren dimana Biokimia—dalam kolaborasinya dengan Genetika, Biologi Molekuler, Bioinformatika, Kimia Fisika, Biologi/Kimia Komputasi dan Genomik—dapat berkontribusi dalam pemanfaatan maha data genomik (Tabel 1).

Pertama, identifikasi sekuen varian yang berdampak fungsional *in silico*. Jenis variasi DNA yang paling umum adalah variasi nukleotida tunggal atau *single nucleotide variation* (SNV), misalkan A menjadi C atau T digantikan G[61]. Secara umum SNV terjadi secara alami dalam genom manusia dimana terdapat sekitar 4-5 juta di dalam manusia[62]. SNV memberikan dasar genetik terhadap keunikan fisik yang terlihat seperti warna kulit, rambut, atau tinggi badan[63]. SNV juga menjelaskan mengapa seseorang lebih rentan terhadap penyakit tertentu seperti kanker atau diabetes dibandingkan yang lain[64,65]. Hal ini karena SNV non-sinonim (nsSNV) atau yang terjadi pada basa-basa DNA yang mengkode protein fungsional akan merubah asam amino menjadi berbeda secara karakteristik sehingga bisa berdampak negatif pada aktivitas biologisnya yang berkaitan dengan perkembangan penyakit. Saat ini identifikasi SNV pembawa resiko penyakit dalam data besar biologis sudah beberapa dilakukan dengan memanfaatkan platform *Cloud* [66–68]. Sebagai contoh, data digital genom sebesar 2.5 Petabyte (PB) dari sebanyak 33 tipe kanker dari sekitar 11.000 pasien yang dihasilkan proyek ambisius pemetaan genom sel kanker atau *The Cancer Genome Atlas* (TCGA) tersimpan dalam Google Cloud yang bisa diakses peneliti diseluruh dunia [69]. Serta data variasi DNA sebesar 7,3 Terabyte (TB) dari 2504 individu dalam 26 populasi di 5 benua yang dihasilkan Proyek 1000 Genom Manusia (kelanjutan dari proyek Pemetaan Genom Manusia) tersedia melalui Amazon Cloud [97].

Tabel 1. Aplikasi Biokimia dan kolaborasinya dengan bidang ilmu terkait dalam identifikasi dan validasi varian juga pemanfaatan data besar genomik untuk desain dan pengembangan obat

No	Topik	Kolaborasi	Metode	Penjelasan	Referensi
1.	Identifikasi varian berdampak fungsional *in silico*	Bioinformatika, Genomik dan Biologi Komputasi	SIFT (*Sorting Intolerant From Tolerant*)	Menggunakan konservasi asam amino dari protein homolog untuk memprediksi dampak fungsional pergantian asam amino non-sinonim dari varian	[70]
			Polyphen-2	Menerapkan probabilitas pembelajaran mesin (*Machine Learning*) menggunakan klasifikasi *Naïve Bayes* terhadap substitusi asam amino protein homolog dengan melihat juga aspek fisikokimia.	[71]
			Provean	Tidak hanya memprediksi dampak fungsional varian, tetapi juga insersi dan delesi (indel) yang merubah susunan asam amino yang penting secara fungsional	[72]
			MutantAsessor	Memprediksi dampak fungsional dari substitusi asam amino dalam protein berdasarkan pola konservasi evolusi dari asam amino yang terpengaruh dalam homolog protein.	[73]
2.	Validasi dampak varian terhadap struktur 3D, dinamika dan interaksi protein fungsional	Kimia Fisik, Biofisika, Kimia Komputasi, Bioinformatika	Homologi komparatif (MODELLER, SWISS-MODEL)	Membangun model 3D protein yang belum diketahui strukturnya menggunakan struktur *template* protein homolog hasil eksperimen	[74–76]
			Threading (PHYRE2, HHPRED)	Membangun strukur 3D protein berdasarkan	[77,78]

				identifikasi fragmen lipatan (*fold recognition*) yang sama yang didapat dari protein yang tersimpan di database yang memiliki kekerabatan dekat maupun jauh secara evolusi	
			Ab initio (ROBETTA, I-TASSER)	Melakukan pemodelan 3D protein target hanya berdasarkan asam amino penyusunnya	79–81
			Dinamika Molekuler (NAMD, GROMACS, AMBER, CHARMM)	Memprediksi dampak varian pada kestabilan dan pergerakan konformasi 3D protein dengan menerapkan medan gaya (*force field*)	82,83
			Molecular Docking (DOCK, Auto Dock4,AutoDock Vina, GLIDE, GOLD)	Memprediksi dampak varian terhadap interaksi protein dgn ligan (substrat, protein, molekul, obat)	84
			Kristalografi sinar-X, NMR, Cryo-EM	Elusidasi strukur 3D protein melalui eksperimen	85–87
3.	Validasi varian melalui pengujian fungsional *in vitro, in vivo*, dan *ex vivo*	Genetika, Biologi Molekuler, Genomik	*In vitro Splicing assay*	Menguji dampak varian terhadap penjalinan segmen mRNA yang mengkode protein fungsional	88
			Deep Mutational Scanning	Metode mutagenesis di mana ekspresi protein dan pemilihan mutan digabungkan dengan NGS untuk menentukan berbagai dampak varian	89
			Mutagenesis yang di mediasi CRISPR/Cas-9	Pengujian dampak varian pada segmen gen pengkode maupun pengatur (*regulatory or coding genes*)	90
4.	Desain obat	Kimia Fisik,	SBDD (*Structure-*	Desain	91

berbasis data besar genomik dengan bantuan komputer	Biofisika, Kimia Komputasi, Bioinformatika, Farmakogenomik	*based drug design*)	pengembangan obat berdasarkan ketersediaan struktur 3D protein target	
		LBDD (*Ligand-based drug design*)	Desain pengembangan obat berdasarkan ketersediaan struktur 3D senyawa kimia aktif atau ligan terhadap protein target	[92]
		Similarity Searching	Pencarian senyawa kimia baru yang memiliki kemiripan dengan senyawa obat yang sudah untuk tujuan optimalisasi	[93]
		Pharmacophore modeling	Pemodelan senyawa kimia ligan berpotensi obat berdasarkan fitur-fitur yang berperan untuk peningkatan profil farmakologinya	[94]
		QSAR (CoMFA, CoMSIA)	Studi hubungan struktur ligan berpotensi obat dengan aktivitas farmakologisnya	[95]
		Virtual Screening	Penyeleksian ribuan hingga jutaan senyawa kimia berpotensi sebagai obat di database dengan menggunakan komputer	[96]

Demokratisasi data besar genomik melalui *Cloud* memudahkan akses dan mempercepat pemrosesan data[98]. Peneliti hanya membutuhkan koneksi internet dan melakukan keseluruhan analisis, penyimpanan, transfer data di dalam *Cloud*. Analisis ulang data besar genomik di dalam *Cloud* telah menolong peneliti dalam mengidentifikasi tidak hanya variasi-variasi DNA yang umum (*common variants*) tapi juga yang baru dan langka (*rare variants*) yang selama ini tidak terdeteksi yang berperan dalam meningkatkan resiko penyakit kanker [99].

Beberapa metode komputasional (*in silico*) juga telah banyak dikembangkan untuk mengidentifikasi sekuen varian yang membawa dampak fungsional. Diantaranya yang paling

popular adalah SIFT[70], Polyphen-2[71], Provean[72] dan MutantAsessor[73]. Pada prinsipnya semua program tersebut memprediksi frekuensi perubahan asam amino non-sinonim pada varian dan dibandingkan dengan urutan asam-asam amino yang tidak berubah atau konservatif dari protein homolog menggunakan pendekatan statistik probabilitas. Asam-asam amino konservatif pada protein homolog umumnya tidak banyak mengalami perubahan karena telah diseleksi alam melalui evolusi untuk menjalankan fungsi yang penting bagi organisme. Akibatnya frekuensi perubahan non-sinonim pada asam amino konservatif ini akan mengganggu aktivitas dan fungsi dari protein terkait. Beberapa metode komputasional lain yang telah dikembangkan untuk mengidentifikasi dan menginterpretasi dampak varian berdasarkan sekuen genom serta penerapannya pada penyakit kanker dapat dilihat pada tinjauan literatur dari referensi[100–102]

Kedua, validasi dampak varian terhadap struktur, dinamika dan interaksi protein fungsional. Variasi pada urutan basa-basa DNA yang mengkode protein fungsional dapat membawa efek negatif bagi tubuh dan meningkatkan resiko terhadap penyakit (nsSNV)[103,104]. Hal ini disebabkan karena nsSNV dalam sekuens DNA yang mengkode protein fungsional akan merubah konformasi struktur dan aktivitas biologisnya[105]. Variasi genetik karena mutasi DNA yang mengkode protein BRCA2 dan APOE misalkan berkorelasi dengan peningkatan faktor resiko terhadap kanker payudara dan prostat[106,107]. Studi Biofisika dan Biologi Struktural telah digunakan untuk mempelajari varian genomik terhadap struktur 3D protein dan interaksinya dengan ligan dengan menggunakan eksperimen kristalografi sinar-X, NMR dan Cryo-EM[108,109]. Sampai bulan Juni 2019 sebanyak 153.328 struktur 3D protein hasil eksperimen telah berhasil dielusidasi dan tersimpan di PDB (*Protein Data Bank*). Namun demikian kekurangan penentuan struktur 3D protein melalui eksperimen adalah biayanya yang mahal dan memakan waktu lama, melibatkan *trial-and error* dengan tingkat kegagalan yang tinggi dan tidak jarang juga memerlukan faktor keberuntungan untuk mendapatkan padatan kristal protein dengan kemurnian tinggi[110]. Untuk mengatasi dua tantangan tersebut diatas, pemodelan molekuler menggunakan pendekatan komputasional secara popular telah banyak diterapkan untuk memprediksi struktur 3D protein target yang belum bisa didapatkan melalui eksperimen (lihat Tabel 1) [74,111]. Pemodelan molekuler, dinamika dan *docking* secara komputasional telah banyak diterapkan untuk memvalidasi efek nsSNV terhadap struktur, gerakan, interaksi dan energi internal protein dalam ruang tiga dimensi (3D)[82,83,112]. Namun demikian kekurangannya adalah studi simulasi dan dinamika protein dalam ruang 3D memerlukan proses komputasi yang sangat intensif dan memakan waktu karena dilakukan dengan menganalisis pergerakan setiap atom-atom penyusun protein yang berjumlah puluhan hingga ratusan ribu. *Cloud* menyediakan platform komputasi berkinerja tinggi yang memfasilitasi simulasi dan dinamika protein dengan lebih cepat[57]. Mengetahui perubahan energi internal antara protein normal dan protein mutan dapat memberikan validasi seberapa besar

pengaruh nsSNV terhadap struktur dan fungsi protein. Informasi ini juga sangat berguna dalam mengetahui efek nsSNV terhadap interaksi protein dengan obat yang menjelaskan kenapa beberapa individu memberikan resistansi terhadap pemberian obat-obatan tertentu.

Ketiga, pengujian fungsional varian patogenik *in vitro, in vivo*, dan *ex vivo*. Terlepas dari kemajuan luar biasa yang telah diberikan NGS dalam menghasilkan data genomik yang besar, juga ilmu bioinformatika melalui pengembangan algoritma komputer untuk identifikasi varian-varian patogenik maupun ilmu komputasi melalui pemodelan dalam beberapa tahun terakhir, harus perlu dicatat bahwa semua itu belumlah memberikan validasi yang paripurna. Pengujian fungsional *in vitro*, *in vivo* maupun *ex vivo* di dalam laboratorium adalah metode mekanistik yang paling *solid* untuk menunjukkan patogenisitas dari varian-varian yang berhasil diidentifikasi dan diprediksi secara komputasional. Oleh sebab itu kedepannya validasi fungsional di laboratorium akan menjadi pelengkap penting dari analisis pendahuluan data besar genomik NGS yang bersifat prediktif dan korelatif dari varian-varian menggunakan bioinformatika, juga analisis struktural pendahuluan menggunalan simulasi maupun pemodelan komputasi. *Deep mutational scanning* (DMS) adalah metode pengujian fungsional yang memanfaatkan NGS untuk menganalisis dalam satu percobaan aktivitas dari banyak varian unik protein [113]. Karena kedalaman cakupan analisis mutasi ini, DMS menyediakan data informasi yang dapat dianalisis untuk mengungkapkan banyak dampak pada karakteristik dan kestabilan protein oleh karena variasi gen. Pendekatan DMS sejauh ini banyak dilakukan secara *in vivo* menggunakan ragi *Saccharomyces cerevisiae*, mengingat reagen dan teknologi yang tersedia luas untuk jenis organisme ini. Variasi pada basa-basa DNA juga bisa menyebabkan perubahan proses penjalinan transkrip mRNA (*mRNA splicing*) yang digunakan sebagai cetakan translasi protein. Akibatnya varian tersebut bisa menghasilkan fragmen protein yang terpotong dan tidak utuh yang mempengaruhi fungsi biologisnya. Teknik *in vitro splicing assay* mulai banyak digunakan untuk menguji dampak varian terhadap penjalinan segmen mRNA yang mengkode protein fungsional[88]. Selain itu teknik pengeditan genom menggunakan CRISPR/Cas9 juga sedang mulai diterapkan juga untuk menguji efek varian tidak hanya pada segmen DNA pengkode protein tapi juga non-pengkode yang umumnya digunakan untuk menghasilkan transkrip RNA yang berperan penting dalam regulasi jalur informasi genetika [90,114].

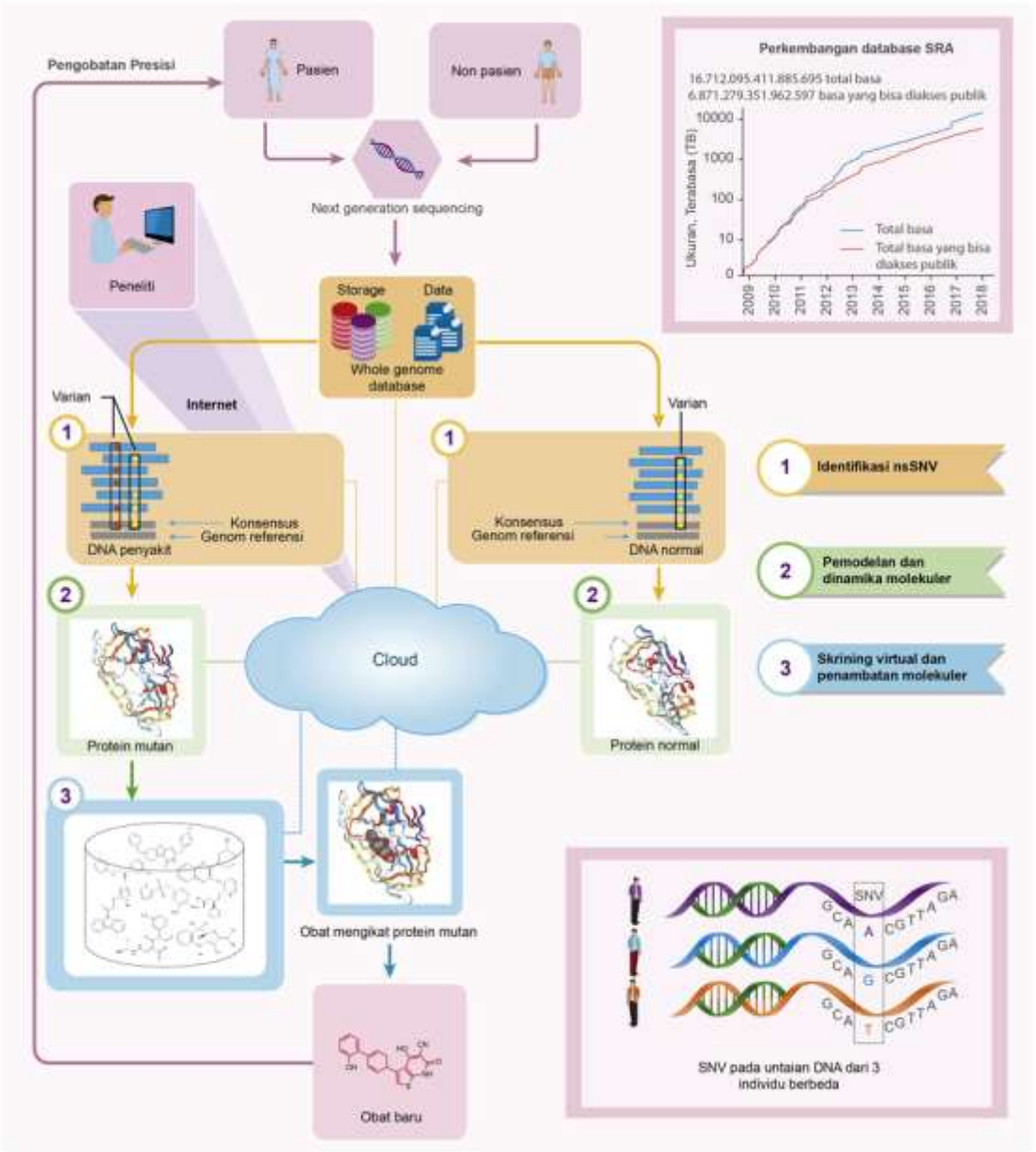
Keempat, desain dan pengembangan obat berbasis genomik dengan bantuan komputer. Terdapat dua metode komputasional yang dikenal dalam tahapan awal (*early stage*) riset desain dan pengembangan obat; pertama desain berdasarkan ketersediaan struktur 3D protein target **(***structure-based drug design* atau *SBDD*)[115] dan kedua berdasarkan ketersediaan struktur 3D senyawa kimia aktif atau ligan terhadap protein target (*ligand-based drug design* atau *LBDD*)[116]. Integrasi SBDD dan LBDD dikenal sebagai teknologi desain obat dengan bantuan

komputer/komputasi (*Computer-aided drug design* atau CADD)[117]. CADD melalui SBDD dan LBDD telah menjadi strategi yang sangat penting yang diterapkan secara luas dalam industri farmasi, maupun akademia untuk tujuan skrining virtual (*virtual screening*) komponen-komponen kimia aktif yang terseleksi (*compound selection*) yang selanjutnya dapat dioptimalisasi (*lead optimization*) untuk penemuan obat baru (*new drug discovery*)[118]**.** LBDD didasarkan atas informasi senyawa kimia aktif atau ligan yang telah diketahui secara eksperimen berinteraksi dan memodulasi protein target[119]. LBDD menggunakan pendekatan statistik dan pemodelan matematika dalam mempelajari hubungan antara struktur dan aktivitas senyawa kimia, baik yang tersedia di dalam database publik mauapun yang didapat dari hasil skrining virtual pada SBDD, dengan menerapkan QSAR (*Three-Dimensional Quantitative Structure-Activity Relationship*)[120]. Dua pendekatan QSAR yang sangat umum digunakan adalah CoMFA (*Comparative Molecular Field Analysis*) dan CoMSIA (*Comparative Molecular Simalrity Indices Analysis*)[121]. CoMFA didasarkan atas konsep bahwa aktivitas biologis dari senyawa kimia bergantung pada medan molekular yang mengelilinginya seperti efek sterik dan elektrostatik. Sementara CoMSIA adalah pengembangan lebih lanjut dari CoMFA dimana selain efek sterik dan elektrostatik, efek hidrofobik, efek donor ikatan hidrogen dan akseptor ikatan hidrogen juga mempengaruhi aktivitas biologis suatu molekul. Hasil analisis dapat digunakan untuk memprediksi senyawa-senyawa kimia baru turunan yang mempunyai kemiripan struktur dan sifat fisika kimianya menggunakan teknik *similarity searching* dan *pharmacopore modeling* untuk mengoptimalkan aktivitas biologisnya[122,123].

Peluang Inovasi

Penerapan riset CADD dengan memanfaatkan maha data genomik dan bantuan komputasi awan memberikan peluang inovatif baik melalui pengembangan metodologi, algoritma komputasi maupun desain dan penemuan obat baru di era pengobatan personal dan presisi saat ini. Penulis mencoba mengusulkan dan menggambarkan ide tersebut dalam bentuk bagan pada Gambar 2. Pertama-tama, keseluruhan genom (*whole genome*) dari pasien dipetakan menggunakan NGS dengan genom non-pasien (orang sehat) sebagai kontrol atau referensi. Data besar genomik yang dihasilkan kemudian disimpan di pangkalan data di *Cloud* yang bisa diakses oleh peneliti melalui internet di mana saja. Data besar ini kemudian digunakan sebagai *input* untuk 3 tahapan riset berikutnya; (1) identifikasi varian DNA yang mempengaruhi pengkodean protein fungsional normal menjadi mutan yang berbeda secara karakteristik fisikokimia atau non-sinonim (nsSNV). (2) pemodelan dan dinamika molekuler untuk menginvestigasi efek varian terhadap struktur 3D, dinamika dan kestabilan protein target yang sedang dipelajari dan (3) penerapan teknik *Docking* dan *virtual screening*.

Dengan teknik penambatan molekuler (*molecular docking*) akan dapat diketahui pengaruh variasi genetik terhadap interaksi protein dengan obat. Variasi DNA pada tempat pengikatan protein dengan obat dapat mempengaruhi proses interaksi di antara mereka. Ini bisa menyebabkan resistensi terhadap obat. Validasi ini akan menolong peneliti dalam memprediksi dan mendesain struktur obat baru yang bisa mengikat dan berinteraksi dengan protein mutan dan menghambat fungsinya yang membawa efek buruk bagi kesehatan. Kemudian *virtual screening* dilakukan untuk skrining senyawa kimia berpotensi obat yang diperoleh dari database ZINC[124], DrugBank[125] maupun ChEMBL[126] dengan protein target yang diprediksi berkorelasi dengan perkembangan penyakit[96,127]. Senyawa kimia yang didapatkan selanjutnya dimodifikasi, diuji secara pra klinis dan klinis sebelum dipasarkan menjadi obat baru yang dapat diterapkan kembali kepada pasien.



Gambar 2. Peluang inovasi memanfaatkan integrasi mahadata genomik dan komputasi awan menuju pengobatan personal dan presisi (*precision and personal medicine*). Keterangan: Sisipan kanan atas: Pertumbuhan data basa-basa nukleotida yang berhasil dipetakan menggunakan NGS di SRA (data dari Mei 2018). Sisipan kanan bawah: ilustrasi SNV (*single nucleotide Variation*) pada 3 individu berbeda.

Kesimpulan

Era *big data* genomik memberikan perspektif dan teknologi baru dalam upaya untuk menjawab permasalahan kesehatan yang semakin kompleks di abad 21 ini. Biokimia dalam sinerginya dengan bidang keilmuan Genetika, Biologi Molekuler, Bioinformatika, Biologi Komputasi, Kimia Komputasi, Biofisika dan Farmakogenomik dapat berkontribusi dalam upaya identifikasi dan validasi yang akurat terhadap varian-varian yang diduga berkaitan dengan perkembangan penyakit. Identifikasi sekuen varian DNA yang bersifat patogenik, validasi dampak varian tersebut terhadap struktur, dinamika dan interaksi protein, juga aplikasi dalam desain pengembangan obat adalah area-area riset yang membawa peluang inovasi dimana Biokimia akan tetap terus relevan dalam sinerginya dengan bidang keilmuan yang terkait tersebut. Jika hal ini bisa dimanfaatkan dengan baik maka ledakan *big data* genomik bernilai sangat besar untuk pengembangan diagnosis dan terapi yang lebih akurat serta obat selektif untuk memerangi penyakit menular dan kompleks yang masih eksis di abad 21 ini seperti kanker, obesitas, diabetes dan Alzheimer.

Daftar Pustaka

1 C. Di Resta et al., “Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities,” *EJIFCC*, vol. 29, no. 1, pp. 4–14, 2018.

2 S. S. Jamuar and E.C. Tan, “Clinical application of next-generation sequencing for Mendelian diseases,” *Hum Genomics*, vol. 9, no. 10, pp. 1-6, 2015.

3 M. Gwinn, D et al., “Next-Generation Sequencing of Infectious Pathogens,” *JAMA*, vol. 321, no. 9, pp. 893–894, 2019.

4 R. Kamps et al., “Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification,” *Int J Mol Sci*, vol. 18, no. 2, pp. 1-57, 2017.

5 Ellard, S.; Franco, E. D. Next-Generation Sequencing for the Diagnosis of Monogenic Diabetes and Discovery of Novel Aetiologies. *Genet. Diabetes* **2014**, *23*, 71–86.

6 V. V. Giau et al., “Genetic analyses of early-onset Alzheimer’s disease using next generation sequencing,” *Sci. Rep.*, vol. 9, no. 1, p. 1-10, 2019.

7 V. Marx, *Biology: The big challenges of big data*. *Nature*, vol. 498, no. 7453, pp. 255-260, 2013.

8 A. O'Driscoll et al, "'Big data', Hadoop and cloud computing in genomics," *J. Biomed. Inf*, vol. 46, no. 5, pp. 774–781, 2013.

9 D. E. Cane, "Back to Basics: Assigning Biochemical Function in the Post-Genomic Era," *Chem. Biol.*, vol. 11, no. 6, pp. 741–743, 2004.

10 Quastel, J. H. The Development of Biochemistry in the 20th Century. *Mol. Cell. Biochem.* **1985**, *69* (1), 17–26.

11 J. Gayon, "From Mendel to epigenetics: History of genetics," *C. R. Biol.*, vol. 339, no. 7–8, pp. 225–230, 2016.

12 Prasad, C.; Galbraith, P. A. Sir Archibald Garrod and Alkaptonuria–'Story of Metabolic Genetics.' *Clin. Genet.* **2005**, *68* (3), 199–203.

13 Horowitz, N. H. One-Gene-One-Enzyme: Remembering Biochemical Genetics. *Protein Sci.* **1995**, *4* (5), 1017–1019.

14 B. S. Strauss, "Biochemical Genetics and Molecular Biology: The Contributions of George Beadle and Edward Tatum," *Genetics*, vol. 203, no. 1, pp. 13–20, 2016.

15 G. W. Beadle and E. L. Tatum, "Genetic control of biochemical reactions in Neurospora," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 27, no. 11, p. 499-506, 1941.

16 Beadle, G. W. Genes and Chemical Reactions in Neurospora. *Science* **1959**, *129* (3365), 1715–1719.

17 Cobb, M. 1953: When Genes Became "Information." *Cell* **2013**, *153* (3), 503–506.

18 J. D. Watson and F. H. C. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, no. 4356, p. 737-738, 1953.

19 F. H. C. Crick et al., "General Nature of the Genetic Code for Proteins," *Nature*, vol. 192, no. 4809, p. 1227-1232, 1961.

20 Ingram, V. M. Gene Mutations in Human Haemoglobin: The Chemical Difference between Normal and Sickle Cell Haemoglobin. *Nature* **1957**, *180* (4581), 326–328.

21 V. M. Ingram, "Abnormal human haemoglobins: I. The comparison of normal human and sickle-cell haemoglobins by 'fingerprinting,'" *Biochim. Biophys. Acta.*, vol. 28, pp. 539–545, 1958.

22 S. E. Antonarakis and J. S. Beckmann, "Mendelian disorders deserve more attention," *Nat. Rev. Genet.*, vol. 7, no. 4, p. 277-282, 2006.

23 E. Duncan, M. Brown, and E. M. Shore, "The Revolution in Human Monogenic Disease Mapping," *Genes (Basel)*, vol. 5, no. 3, pp. 792–803, 2014.

24 A. E. Guttmacher and F. S. Collins, Welcome to the genomic era. *N. Engl. J. Med*, vol. 349, no. 10, pp. 996-998 2003.

25 Collins, F. S.; Morgan, M.; Patrinos, A. The Human Genome Project: Lessons from Large-

Scale Biology. *Science* **2003**, *300* (5617), 286–290.

26 I. H. G. S. Consortium, “Finishing the euchromatic sequence of the human genome,” *Nature*, vol. 431, no. 7011, p. 931-945, 2004.

27 Chial, H. DNA Sequencing Technologies Key to the Human Genome Project. *Nat. Educ.* **2008**, *1* (1), 219.

28 The Cost of Sequencing a Human Genome https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (accessed Jul 19, 2019).

29 Nerlich, B.; Dingwall, R.; Clarke, D. D. The Book of Life: How the Completion of the Human Genome Project Was Revealed to the Public. *Health (N. Y.)* **2002**, *6* (4), 445–469.

30 L. Hood and D. Galas, “The digital code of DNA,” *Nature*, vol. 421, no. 6921, p. 444-448, 2003.

31 E. Birney, “The making of ENCODE: lessons for big-data projects,” *Nature*, vol. 489, no. 7414, p. 49-51, 2012.

32 E. P. Consortium, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, p. 57-74, 2012.

33 E. D. Green et al, “Human Genome Project: Twenty-five years of big biology,” *Nat. News*, vol. 526, no. 7571, p. 29-31, 2015.

34 Bayer, R.; Galea, S. Public Health in the Precision-Medicine Era. *N. Engl. J. Med.* **2015**, *373* (6), 499–501.

35 Collins, F. S.; Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **2015**, *372* (9), 793–795.

36 Hamburg, M. A.; Collins, F. S. The Path to Personalized Medicine. *N. Engl. J. Med.* **2010**, *363* (4), 301–304.

37 Langreth, R.; Waldholz, M. New Era of Personalized Medicine Targeting Drugs for Each Unique Genetic Profile. *The oncologist* **1999**, *4* (5), 426–427.

38 K. Offit, “Genomic profiles for disease risk: predictive or premature?,” *JAMA*, vol. 299, no. 11, pp. 1353–1355, 2008.

39 F. S. Collins et al, “A vision for the future of genomics research,” *Nature*, vol. 422, no. 6934, p. 835-847, 2003.

40 F. S. Collins and V. A. McKusick, “Implications of the Human Genome Project for medical science,” *JAMA*, vol. 285, no. 5, pp. 540–544, 2001.

41 M. L. Metzker, “Sequencing technologies—the next generation,” *Nat. Rev. Genet.*, vol. 11, no. 1, p. 31-46, 2010.

42 S. C. Schuster, “Next-generation sequencing transforms today’s biology,” *Nat. Methods*, vol. 5, no. 1, p. 16-18, 2007.

43 J. Shendure and H. Ji, “Next-generation DNA sequencing,” *Nat. Biotechnol.*, vol. 26, no. 10, p. 1135-1145, 2008.

44 S. T. Bennett et al, “Toward the $1000 human genome,” *Pharmacogenomics*, vol. 6, no. 4, pp. 373-382, 2005.

45 E. Check Hayden, “Technology: the $1,000 genome,” *Nat. News*, vol. 507, no. 7492, p. 294-295, 2014.

46 Wolinsky, H. The Thousand-Dollar Genome. *EMBO Rep.* **2007**, *8* (10), 900–903.

47 F. Ozsolak and P. M. Milos, “RNA sequencing: advances, challenges and opportunities,” *Nat. Rev. Genet.*, vol. 12, no. 2, p. 87-98, 2011.

48 B. Rabbani, M. Tekin, and N. Mahdieh, “The promise of whole-exome sequencing in medical genetics,” *J. Hum. Genet.*, vol. 59, no. 1, p. 5-15, 2014.

49 Schatz, M. C. Biological Data Sciences in Genome Research. *Genome Res.* **2015**, *25* (10), 1417–1422.

50 Z. D. Stephens et al., “Big data: astronomical or genomical?,” *PLoS Biol.*, vol. 13, no. 7, p. e1002195, 2015.

51 Fan, J.; Han, F.; Liu, H. Challenges of Big Data Analysis. *Natl. Sci. Rev.* **2014**, *1* (2), 293–314.

52 Y. Qin et al, “The current status and challenges in computational analysis of genomic big data,” *Big Data Res.*, vol. 2, no. 1, pp. 12–18, 2015.

53 B. Langmead and A. Nellore, “Cloud computing for genomic data analysis and collaboration,” *Nat. Rev. Genet.*, vol. 19, no. 4, p. 208-2019, 2018.

54 C. Yang et al, “Big Data and cloud computing: innovation opportunities and challenges,” *Int. J. Digit. Earth*, vol. 10, no. 1, pp. 13–53, 2017.

55 Avram, M.-G. Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective. *Procedia Technol.* **2014**, *12*, 529–534.

56 Calabrese, B.; Cannataro, M. Cloud Computing in Healthcare and Biomedicine. *Scalable Comput. Pract. Exp.* **2015**, *16* (1), 1–18.

57 J.-P. Ebejer et al, “The emerging role of cloud computing in molecular modelling,” *J. Mol. Graph. Model.*, vol. 44, pp. 177–187, 2013.

58 Pallis, G. Cloud Computing: The New Frontier of Internet Computing. *IEEE Internet Comput.* **2010**, *14* (5), 70–73.

59 Varghese, B.; Buyya, R. Next Generation Cloud Computing: New Trends and Research Directions. *Future Gener. Comput. Syst.* **2018**, *79*, 849–861.

60 AWS. Amazon EC2 Pricing - Amazon Web Services https://aws.amazon.com/ec2/pricing/ (accessed Jul 5, 2019).

61 Katsonis, P.; Koire, A.; Wilson, S. J.; Hsu, T.-K.; Lua, R. C.; Wilkins, A. D.; Lichtarge, O. Single Nucleotide Variations: Biological Impact and Theoretical Interpretation. *Protein Sci.* **2014**, *23* (12), 1650–1666.

62 NIH. What are single nucleotide polymorphisms (SNPs)? https://ghr.nlm.nih.gov/primer/genomicresearch/snp (accessed Jul 5, 2019).

63 D. B. Goldstein, "Common genetic variation and human traits," *N. Engl. J. Med.*, vol. 360, no. 17, p. 1696-1698, 2009.

64 N. Deng et al, "Single nucleotide polymorphisms and cancer susceptibility," *Oncotarget*, vol. 8, no. 66, p. 110635-110649, 2017.

65 Mathers, J. C.; Hesketh, J. E. The Biological Revolution: Understanding the Impact of SNPs on Diet-Cancer Interrelationships. *J. Nutr.* **2007**, *137* (1), 253S–258S.

66 M. Cao et al, "A fast and accurate SNP detection method on the cloud platform," in *2015 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2015, pp. 2186–2191.

67 R. J. Mashl et al., "GenomeVIP: a cloud platform for genomic variant discovery and interpretation," *Genome Res.*, vol. 27, no. 8, pp. 1450–1459, 2017.

68 G. Minevich et al, "CloudMap: a cloud-based pipeline for analysis of mutant genome sequences," *Genetics*, vol. 192, no. 4, pp. 1249–1269, 2012.

69 Google Inc. TCGA Cancer Genomics Data in the Cloud — Google Genomics v1 documentation https://googlegenomics.readthedocs.io/en/latest/use_cases/discover_public_data/isb_cgc_data.html (accessed Jul 5, 2019).

70 Ng, P. C.; Henikoff, S. SIFT: Predicting Amino Acid Changes That Affect Protein Function. *Nucleic Acids Res.* **2003**, *31* (13), 3812–3814.

71 I. A. Adzhubei et al., "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, p. 248-249, 2010.

72 Choi, Y.; Chan, A. P. PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics* **2015**, *31* (16), 2745–2747.

73 Reva, B.; Antipin, Y.; Sander, C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Res.* **2011**, *39* (17), e118–e118.

74 Ginalski, K. Comparative Modeling for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2006**, *16* (2), 172–177.

75 Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: An Automated Protein Homology-Modeling Server. *Nucleic Acids Res.* **2003**, *31* (13), 3381–3385.

76 Webb, B.; Sali, A. Protein Structure Modeling with MODELLER. In *Functional Genomics*;

Springer, 2017; pp 39–54.

77 A. Hildebrand et al, “Fast and accurate automatic structure prediction with HHpred,” *Proteins Struct. Funct. Bioinforma.*, vol. 77, no. S9, pp. 128–132, 2009.

78 Kelley, L. A. Fold Recognition. In *From Protein Structure to Function with Bioinformatics*; Springer, 2017; pp 59–90.

79 D. E. Kim et al, “Protein structure prediction and analysis using the Robetta server,” *Nucleic Acids Res.*, vol. 32, no. suppl_2, pp. W526–W531, 2004.

80 J. Lee et al, “Ab initio protein structure prediction,” in *From protein structure to function with bioinformatics*, Springer, 2017, pp. 3–35.

81 A. Roy et al, “I-TASSER: a unified platform for automated protein structure and function prediction,” *Nat. Protoc.*, vol. 5, no. 4, p. 725-738, 2010.

82 F. Fratev et al, “Combination of genetic screening and molecular dynamics as a useful tool for identification of disease-related mutations: ZASP PDZ domain G54S mutation case,” *J. Chem. Inf. Model.*, vol. 54, no. 5, pp. 1524–1536, 2014.

83 B. Kamaraj and R. Purohit, “In silico screening and molecular dynamics simulation of disease-associated nsSNP in TYRP1 gene and its structural consequences in OCA3,” *BioMed Res. Int.*, vol. 2013, pp. 1-13, 2013.

84 Lybrand, T. P. Ligand—Protein Docking and Rational Drug Design. *Curr. Opin. Struct. Biol.* **1995**, *5* (2), 224–228.

85 Hillisch, A.; Hilgenfeld, R. The Role of Protein 3D-Structures in the Drug Discovery Process. In *Modern methods of drug discovery*; Springer, 2003; pp 157–181.

86 Murata, K.; Wolf, M. Cryo-Electron Microscopy for Structural Analysis of Dynamic Biological Macromolecules. *Biochim. Biophys. Acta BBA-Gen. Subj.* **2018**, *1862* (2), 324–334.

87 Shi, Y. A Glimpse of Structural Biology through X-Ray Crystallography. *Cell* **2014**, *159* (5), 995–1014.

88 A. Y. Filatova et al., “Functional reassessment of PAX6 single nucleotide variants by in vitro splicing assay,” *Eur. J. Hum. Genet.*, vol. 27, no. 3, p. 488-493, 2019.

89 D. M. Fowler and S. Fields, “Deep mutational scanning: a new style of protein science,” *Nat. Methods*, vol. 11, no. 8, p. 801-807, 2014.

90 P. Mali et al., “RNA-guided human genome engineering via Cas9,” *Science*, vol. 339, no. 6121, pp. 823–826, 2013.

91 Kalyaanamoorthy, S.; Chen, Y.-P. P. Structure-Based Drug Design to Augment Hit Discovery. *Drug Discov. Today* **2011**, *16* (17–18), 831–839.

92 C. Acharya et al, “Recent advances in ligand-based drug design: relevance and utility of the

conformationally sampled pharmacophore approach," *Curr. Comput. Aided Drug Des.*, vol. 7, no. 1, pp. 10–22, 2011.

93 Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discov. Today* **2002**, *7* (17), 903–911.

94 J. S. Mason et al, "3-D pharmacophores in drug discovery," *Curr. Pharm. Des.*, vol. 7, no. 7, pp. 567–597, 2001.

95 J. Verma et al, "3D-QSAR in drug design-a review," *Curr. Top. Med. Chem.*, vol. 10, no. 1, pp. 95–115, 2010.

96 A. Olğaç et al, "Cloud-Based High Throughput Virtual Screening in Novel Drug Discovery," in *High-Performance Modelling and Simulation for Big Data Applications*, Springer, 2019, pp. 250–278.

97 AWS. 1000 Genomes Project and AWS https://aws.amazon.com/1000genomes/ (accessed Jul 5, 2019).

98 Richards, D. The democratization of data in the cloud https://www.infoworld.com/article/3090084/the-democratization-of-data-in-the-cloud.html (accessed Jul 5, 2019).

99 J. R. Huyghe et al., "Discovery of common and rare genetic risk variants for colorectal cancer," *Nat. Genet.*, vol. 51, no. 1, p. 76-87, 2019.

100 B. J. Raphael et al, "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine," *Genome Med.*, vol. 6, no. 1, p. 1-17, 2014.

101 R. Tian et al, "Computational methods and resources for the interpretation of genomic variants in cancer," *BMC Genomics*, vol. 16, no. 8, p. S7, 2015.

102 J. Zhang et al, "Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing," *Brief. Bioinform.*, vol. 15, no. 2, pp. 244–255, 2013.

103 E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," *Nat. Rev. Genet.*, vol. 11, no. 6, p. 415-425, 2010.

104 J. G. Taylor et al, "Using genetic variation to study human disease," *Trends Mol. Med.*, vol. 7, no. 11, pp. 507–512, 2001.

105 R. Bhattacharya et al, "Impact of genetic variation on three dimensional structure and function of proteins," *PloS One*, vol. 12, no. 3, p. e0171355, 2017.

106 D. F. Easton et al., "A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer–predisposition genes," *Am. J. Hum. Genet.*, vol. 81, no. 5, pp. 873–883, 2007.

107 F. Yencilek et al., “Apolipoprotein E Genotypes in Patients with Prostate Cancer,” *Anticancer Res.*, vol. 36, no. 2, pp. 707–711, 2016.

108 M. Hicks et al, “Functional characterization of 3D protein structures informed by human genetic diversity,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 18, pp. 8960–8965, 2019.

109 A. Stein et al, “Biophysical and mechanistic models for disease-causing protein variants,” *Trends Biochem. Sci.*, vol. 44, no. 7, pp. 575-588, 2019.

110 A. M. Davis et al, “Application and limitations of X-ray crystallographic data in structure-based ligand and drug design,” *Angew. Chem. Int. Ed.*, vol. 42, no. 24, pp. 2718–2736, 2003.

111 T. L. Nero et al, “Protein structure and computational drug discovery,” *Biochem. Soc. Trans.*, vol. 46, no. 5, pp. 1367–1379, 2018.

112 L. Ponzoni and I. Bahar, “Structural dynamics is a determinant of the functional significance of missense variants,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 16, pp. 4164–4169, 2018.

113 Starita, L. M.; Fields, S. Deep Mutational Scanning: A Highly Parallel Method to Measure the Effects of Mutation on Protein Function. *Cold Spring Harb. Protoc.* **2015**, *2015* (8),

114 M. F. Sentmanat et al, “A survey of validation strategies for CRISPR-Cas9 editing,” *Sci. Rep.*, vol. 8, no. 1, p. 888, 2018.

115 Saudale, F.Z., "Pemodelan Homologi Komparatif Struktur 3D Protein dalam Desain dan Pengembangan Obat," *Al-Kimia*., vol. 8, no.1, p.93, 2020

116 T. Wang et al., “Advances in computational structure-based drug design and application in drug discovery,” *Curr. Top. Med. Chem.*, vol. 16, no. 9, pp. 901–916, 2016.

117 B. Bordás et al, “Ligand-based computer-aided pesticide design. A review of applications of the CoMFA and CoMSIA methodologies,” *Pest Manag. Sci.*, vol. 59, no. 4, pp. 393–400, 2003.

118 G. Sliwoski et al, “Computational methods in drug discovery,” *Pharmacol. Rev.*, vol. 66, no. 1, pp. 334–395, 2014.

119 S. J. Y. Macalino et al, “Role of computer-aided drug design in modern drug discovery,” *Arch. Pharm. Res.*, vol. 38, no. 9, pp. 1686–1701, 2015.

120 C. Acharya et al, “Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach,” *Curr. Comput. Aided. Drug. Des*, Vol. 7, no. 1, pp. 10-22, 2011.

121 J. Verma et al, “3D-QSAR in Drug Design - A Review,” *Curr. Top. Med. Chem*, vol. 10, no. 1, pp. 95-115, 2010.

122 D. Stumpfe and J. Bajorath, “Similarity searching,” *Wiley Interdiscip. Rev.: Comput. Mol.*

Sci., vol. 1, no. 2, pp. 260–282, 2011.

123 S.-Y. Yang, “Pharmacophore modeling and applications in drug discovery: challenges and recent advances,” *Drug Discovery Today*, vol. 15, no. 11, pp. 444–450, 2010.

124 Irwin, J. J.; Shoichet, B. K. ZINC- a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.

125 D. S. Wishart et al., “DrugBank: a comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Res.*, vol. 34, no. suppl_1, pp. D668–D672, 2006.

126 A. Gaulton et al., “ChEMBL: a large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1100–D1107, 2011.

127 M. Capuccini et al, “Large-scale virtual screening on public cloud resources with Apache Spark,” *J. Cheminformatics*, vol. 9, no. 15, p. 1-6, 2017.