Conceptual models and databases for searching the genome

Anna Bernasconi
Dept. Electronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
anna.bernasconi@polimi.it

Pietro Pinoli
Dept. Electronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
pietro.pinoli@polimi.it

ABSTRACT

Genomics is an extremely complex domain, in terms of concepts, their relations, and their representations in data. This tutorial introduces the use of ER models in the context of genomic systems: conceptual models are of great help for simplifying this domain and making it actionable. We carry out a review of successful models presented in the literature for representing biologically-relevant entities and grounding them in databases. We draw a difference between conceptual models that aim to explain the domain and conceptual models that aim to support database design and heterogeneous data integration. Genomic experiments and/or sequences are described by several metadata, specifying information on the sampled organism, the used technology, and the organizational process behind the experiment. Instead, we call data the actual regions of the genome that have been read by sequencing technologies and encoded into a machine-readable representation. First, we show how data and metadata can be modeled, then we exploit the proposed models for designing search systems, visualizers, and analysis environments. Both domains of human genomics and viral genomics are addressed, surveying several use cases and applications of broader public interest. The tutorial is relevant to the EDBT community because it demonstrates the usefulness of conceptual models' principles within very current domains; in addition, it offers a concrete example of conceptual models' use, setting the premises for interdisciplinary collaboration with a greater public (possibly including life science researchers).

1 INTRODUCTION

Genomics has to date become a big data generation domain [42]. Since 2008 the costs to sequence a single human genome have experienced a significant drop [33]; consequently, a growing number of experimental samples has been deposited in public archives. Unfortunately, this has not been matched by contextual data curation and integration. Especially, models in the domain are overly complex and do not allow practical use or guidance for database design. Moreover, experiments descriptions are very heterogeneous and lack standards, while semantic integration can only be achieved by cumbersome linking to specialized ontologies [3]. The recent COVID-19 pandemic has brought general attention also to genomics of infectious diseases and microbial research, including viral typing. Laboratories around the world started sequencing samples extracted from patients with COVID-19, harboring SARS-CoV-2 viral bio-material, leading to the collection of several million sequences [31].

Conceptual models can bring useful support in this context, especially by providing a shared clarification of this domain that drives data integration solutions [4, 6]. These, in turn, allow for building data management systems, which empower effective search over the genome.

This tutorial carries out a review of classic successful models presented in literature for representing biologically-relevant entities [26, 34, 36, 38] (i.e., *explanatory models*) and grounding them on databases [44**?**] (i.e., *data-design models*). We draw a difference between conceptual models and databases that aim to explain, unfold and query the domain knowledge and those that are functional to design databases and to heterogeneous data integration directed to the deployment of analytical and research-oriented services. Two exemplary and current domains are considered, with a focus on the need for data-driven approaches: i) human genomics (including expression of genes, somatic and inherited mutations) and ii) viral sequence genomics (including the ones of SARS-CoV-2, the virus responsible for COVID-19).

2 HUMAN GENOMICS

Big genomic datasets are organized as collections of samples. Samples are the basic unit of information, containing experimental data that corresponds to a given individual and preparation (e.g., cell line and antibody used) that first undergoes Next Generation Sequencing [40] (producing "raw data"), then alignment and calling processes (producing "processed data"). Each sample includes DNA segments or regions (possibly the whole genome) – called *region data* in the following – and it is associated with information about the performed experiment, i.e., *metadata* of the sample.

We introduce the Genomic Conceptual Model (GCM, [9]), for describing metadata and high-level properties of regions (see Figure 1). The schema was built through a top-down method, by abstraction of important conceptual properties of genomic sources. The schema is centered on the notion of *experimental item*, typically a file containing genomic regions and their properties, which is analyzed from four points of view: 1) the biological process observed in the experiment (with the biosample being sequenced, derived from a tissue or a cell culture) and donor; 2) the management of the experiment, describing the organization behind it; 3) the technological process used for the production of the item; 4) The parameters used for internal organization.

An integration pipeline called META-BASE [5] can be followed to build a repository based on the GCM: data are downloaded from the original sources (with heterogeneous formats), transformed into a key-value format with rules or prediction models [16], and cleaned to reduce redundancy. At the schema level, sources' information is mapped within a database based on the GCM; at the value level, this information is normalized using biomedical ontologies [3]. The repository can then be queried by means of GenoSurf [13], which allows searching experimental data within a database using semantically enriched queries.

Several features and phenomena can be measured on the genomes of humans (not only mutations); these can be represented using an interval-based paradigm, resulting in the Genomic Data Model (GDM) [28], using spatial regions (with a start,

10.48786/edbt.2022.57

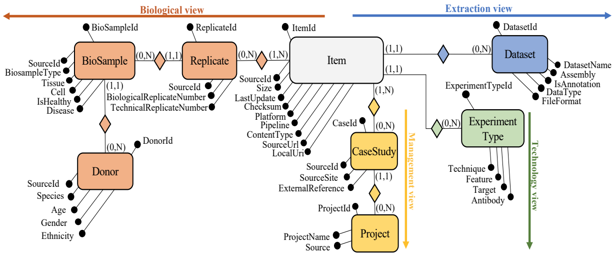

Figure 1: The Genomic Conceptual Model [9].

a stop, and an arbitrary set of features) to answer complex biological queries in a data-driven fashion, leveraging big-data and cloud computing optimizations.

This data-driven paradigm is demonstrated to participants by the GenoMetric Query Language (GMQL) [29] a formal language that combines relational algebra and domain-specific operators to effectively query GDM interval-based data. GMQL is compiled down to an intermediate representation (IR). Each instance of IR is a direct acyclic graph, where nodes are atomic operations and connections represent the flow of the information, which goes from the input data to the result of the query. The IR is further split into two sub-graphs: one elaborates the portion of the query related to the regions and the other manipulates the metadata. Having an intermediate representation allows to implement classical relational databases optimizations (e.g., avoiding loading unused tables, performing selection before the join), while the peculiarity of the presence of two sub-graphs leads to specialized optimizations, such as the meta-first [35]. In the meta-first optimization, the nodes of the IR are rearranged in such a way that all the operations on the metadata are executed before even loading the region data; this allows to filter the region data (which are usually much bigger than the corresponding metadata) directly at the reading stage, thus reducing significantly the consumption of both memory and execution time. The optimized IR is then interpreted by an execution engine. The main implementation uses Apache Spark [27], but other engines based on Apache Flink, SciDB (an array-based DBMS for data-intensive applications) and PostgreSQL has been evaluated [17, 18, 23, 25]. One of the most challenging aspects in the development of GMQL using scalable engines such as Apache Spark, lies in the implementation of theta-joins. We demonstrate to participants an effective strategy based on binning the genome in non-overlapping portions and assigning each region to each bin it overlaps. The computations of the join are executed within each bin, and a filtering strategy of the results avoids the generations of duplicates in the output [12]. The selection of the best binning size is a critical point, as large bins create fewer duplicates on the input data while smaller ones lead to a faster *intra-bin* computation [22].

Similarly, the IR allows reusing the same execution engines to build new frameworks. We show in particular pyGMQL [32], a Python wrapper of GMQL with additional perks, including the ability to convert instances of GDM in Pandas DataFrames for visualization and statistical analysis. The result can be then re-converted to GDM and elaborated with GMQL.

The cloud-based implementation has driven the design of Federated GMQL [15], a large infrastructure based on communication protocols and federated query execution mechanisms and policies, to connect multiple geographically separated instances of the GMQL system, in order to share data and evaluate queries in a federated and privacy-performing fashion.

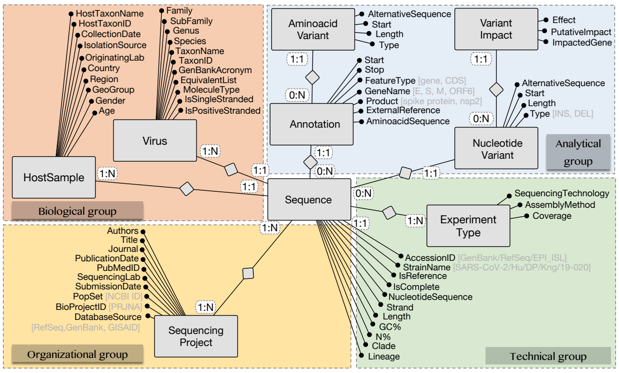

Figure 2: The Viral Conceptual Model [7].

Finally, building upon the repository of big datasets and the computational engine, another system has been realized to overcome technological barriers for biologists: GeCoAgent is a conversational agent [20] where users may explore and analyze genomic datasets using a natural language interface that maps into concepts and actions.

To summarize, through several examples, we demonstrate to participants how well-designed data models can be leveraged to develop applications that are both accessible to the final user and with high performance in terms of execution time and memory consumption.

3 VIRAL GENOMICS

The COVID-19 pandemic has attracted incredible attention to the mechanisms of birth, spread, and evolution of viruses. Extensive sequencing has been performed since January 2020, reaching today about 10 million sequences deposited in open databases [30].

Similarly to what was presented in the human genomics case, also in this domain, we can adopt a methodology that includes a *modeling* phase, analyzing the peculiarity of data and proposing a conceptual model to unify relevant sources; an *integration* phase, in which relevant databases are selected and integrated by means of pipelines that feed a large repository; and a *search* phase, when methods for querying data are built to respond to scientists' needs.

Following this paradigm, the Viral Conceptual Model (VCM [7], see Figure 2) has been designed to describe genomic samples' metadata, their sequencing, their several functional parts, their mutations (i.e., deviation) with respect to an expected reference, and the presence of such mutations on parts of the virus that are critical for vaccine and serological assays design. In this tutorial, we focus on the design of this conceptual model, as it can be exploited to implement highly performing data-intensive applications. It allows the development of methods and tools to efficiently answer complex research queries, able to replicate the scientific results of recent articles, hence demonstrating considerable potential in supporting virology research.

The ViruSurf [14] database has been developed based on the VCM, fed by a high-performance pipeline of sequence data extraction and processing [2]. Such pipeline extracts information from data archives where worldwide laboratories deposit SARS-CoV-2 sequences (namely, GenBank [39], COG-UK [43], and GISAID [41]), applies content curation (including standardizazion, cleaning, computation of metrics and variant calling), and database content optimization. On top of this database, a powerful search system of viral sequences allows building queries using metadata or mutations patterns as filters. In parallel, the same database supports Episurf [10], a system to query and analyze epitopes, which are portions of viruses recognized by the host immune system and thus fundamental to the development of vaccines. Both systems feed the VirusViz visualization engine [11], which receives the results of queries and provides metadata summarization, variant descriptions, and variant visualization. VirusViz offers rich customization and the possibility of grouping and comparatively analyzing populations of interest (allowing, for instance, to trace new variants since their starting dates in a precise location). The system was employed to study the evolution of SARS-CoV-2 variants only based on their mutational patterns, providing interesting insights about variant emergence [37, 45]. In a complementary way, ViruClust [19] is a tool for fine-tuning clusters of sequences, used for fine-grain surveillance.

The VCM can be further abstracted if one aims to represent both the data and the external knowledge that is being collected about SARS-CoV-2, such as notions on variants with: 1) their effects (in terms of disease severity, transmissibility, vaccine escape, etc.); 2) their composition (in terms of sets of mutations); 3) their characteristic mutations. Mutations are distinct due to their original and alternative nucleotide or amino acid residues and their location, e.g., within particular regions of the genome with given functions [1].

4 JOINT DIRECTIONS

The elements described during the tutorial are part of a broad vision: availability of conceptual models, related databases, and search systems for both humans and viruses' genomics will provide important opportunities for genomic and clinical research, especially if the sequences of viruses (or of other pathogens) can be connected to genotype and phenotype information regarding its host, i.e., the human organism. In this direction, we show a unifying model that interlinks a viral genome to the genomic features of the human being who has been infected [6]. This vision has been embraced by the COVID-19 Host Genetics Initiative, which aims at gathering an open community of thousands of researchers who produce, share, and analyze data to learn the genetic determinants of COVID-19 susceptibility, severity, and outcomes [24]. Within this international group, we engaged in the design, structuring, and harmonization of a comprehensive data dictionary to help with the submission of individual-level data. The phenotype refers to severe patients who were hospitalized; it has about 200 clinical variables that have been progressively consolidated and annotated, describing demographics, exposure, risk factors, co-morbidities, hospitalization admission and course, and longitudinal encounters with symptoms, treatments, and lab data. The data dictionary [8] can be used to format clinical phenotype data that are currently being collected and hosted by EGA, the European Genome-Phenome Archive of EMBL-EBI [21]. The initiative already collected a considerable amount of results, currently reaching 9.4 K critically ill cases, 25 K hospitalized cases, and 125 K reported cases of SARS-CoV-2 infection with almost 3M controls.

5 LEARNING GOALS

Participants of the tutorial can learn how conceptual models are used to represent genomic entities and how they can be employed as a basis for building usable databases and systems. They learn basic notions of viral and human genomes and how to pose meaningful queries to a number of different search systems for genomic research. The gap between a very complex topic and its understanding is reduced by means of a conceptual modeling approach. The proposed paradigm may be adapted to other highly specialized domains, both within life sciences and within data science in the broader sense. The tutorial also provides a series of practical tasks, ordered by increasing complexity for both human and viral genomics.

6 TARGET AUDIENCE, PREREQUISITE KNOWLEDGE

The tutorial targets researchers that are curious about the application of conceptual modeling and database theory to the complex and current domain of genomics. The tutorial aims to explore the strength of conceptual models' and databases' principles within very practical applied scenarios. Basic knowledge of ER models syntax is required. No previous knowledge of biology and genomics is requested, as we cover all the basic ingredients that are necessary to understand the workflow and the interactive session's queries.

7 EARLIER VERSIONS OF THE TUTORIAL

The tutorial has been presented to the International Conference on Conceptual Modeling in 2021 (https://er2021.org/papers.html), focused on modeling aspects rather than on their applications to databases. In this version, we originally provide insights on the technological challenges brought by genomic (and in general biological) data that can be addressed practically. Additionally, we share results of recent database-driven research on SARS-CoV-2 evolution and platforms that support current studies.

8 ABOUT THE AUTHORS

Anna Bernasconi Anna Bernasconi obtained her PhD cum laude at Politecnico di Milano, within the "Data-driven Genomic Computing" ERC Awarded project (2016-2021), under the supervision of Professor Stefano Ceri. She is now a postdoctoral researcher in the Department of Electronics, Information, and Bioengineering at Politecnico di Milano and a visiting researcher at Universitat Politècnica de València. Her research focuses on conceptual modeling, data integration, semantic web, and biological data analysis. Since the COVID-19 pandemic, her research has focused on viral genomics, by building models, databases, and Web search systems for viral sequences and their variants. She is active in the conceptual modeling community, with the presentation of one tutorial, several papers, and the organization of two workshops on conceptual models and web applications for life sciences (ER and ICWE conferences).

Pietro Pinoli works as a Researcher Fellow and lecturer at the Department of Electronics, Information, and Bioengineering at the Politecnico di Milano (Italy). He received his PhD cum laude in 2017, with a thesis titled "Modeling and Querying Genomic Data" where he proposed and benchmarked data structures and algorithms to manage, search and elaborate huge collections of genomic datasets, by means of cloud and distributed technologies. He has been visiting PhD student at Harvard University (Cambridge, MA, US). His research interests include bioinformatics and computational biology, databases and data management, big data technology and algorithms, machine learning and natural language processing, and drug repurposing. He participated in the Italian PRIN GenData, ERC GeCo, and EIT VirusLab projects.

9 ACKNOWLEDGEMENTS

This work was funded by the ERC Advanced Grant 693174 GeCo (data-driven Genomic Computing) and by the EIT Digital innovation activity 20663 "DATA against COVID-19". The authors would like to acknowledge the P.I. of the projects Prof. Stefano Ceri for his continuous support and inspirational discussions.

REFERENCES

[1] Ruba Al Khalaf, Tommaso Alfonsi, Stefano Ceri, and Anna Bernasconi. 2021. CoV2K: A Knowledge Base of SARS-CoV-2 Variant Impacts. In *Research Challenges in Information Science*, Samira Cherfi, Anna Perini, and Selmin Nurcan (Eds.). Springer, 274–282.

[2] Tommaso Alfonsi, Pietro Pinoli, and Arif Canakoglu. 2022. High Performance Integration Pipeline for Viral and Epitope Sequences. *BioTech* 11, 1 (2022), 7.

[3] Anna Bernasconi, Arif Canakoglu, Andrea Colombo, and Stefano Ceri. 2018. Ontology-Driven Metadata Enrichment for Genomic Datasets. In *International Conference on Semantic Web Applications and Tools for Life Sciences (CEUR Workshop Proceedings)*, Christopher J. O. Baker, Andra Waagmeester, Andrea Splendiani, Oya Deniz Beyan, and M. Scott Marshall (Eds.), Vol. 2275.

[4] Anna Bernasconi, Arif Canakoglu, Marco Masseroli, and Stefano Ceri. 2021. The road towards data integration in human genomics: players, steps and interactions. *Briefings in Bioinformatics* 22, 1 (2021), 30–44.

[5] Anna Bernasconi, Arif Canakoglu, Marco Masseroli, and Stefano Ceri. 2022. META-BASE: A Novel Architecture for Large-Scale Genomic Metadata Integration. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 1 (2022), 543–557.

[6] Anna Bernasconi, Arif Canakoglu, Marco Masseroli, Pietro Pinoli, and Stefano Ceri. 2021. A review on viral data sources and search systems for perspective mitigation of COVID-19. *Briefings in Bioinformatics* 22, 2 (2021), 664–675.

[7] Anna Bernasconi, Arif Canakoglu, Pietro Pinoli, and Stefano Ceri. 2020. Empowering Virus Sequence Research Through Conceptual Modeling. In *Conceptual Modeling*, Gillian Dobbie, Ulrich Frank, Gerti Kappel, Stephen W. Liddle, and Heinrich C. Mayr (Eds.). Springer, 388–402.

[8] Anna Bernasconi and Stefano Ceri. 2022. Interoperability of COVID-19 Clinical Phenotype Data with Host and Viral Genetics Data. *BioMed* 2, 1 (2022), 69–81.

[9] Anna Bernasconi, Stefano Ceri, Alessandro Campi, and Marco Masseroli. 2017. Conceptual modeling for genomics: building an integrated repository of open data. In *International Conference on Conceptual Modeling*. Springer, 325–339.

[10] Anna Bernasconi, Luca Cilibrasi, Ruba Al Khalaf, Tommaso Alfonsi, Stefano Ceri, Pietro Pinoli, and Arif Canakoglu. 2021. EpiSurf: metadata-driven search server for analyzing amino acid changes within epitopes of SARS-CoV-2 and other viral species. *Database* 2021 (2021).

[11] Anna Bernasconi, Andrea Gulino, Tommaso Alfonsi, Arif Canakoglu, Pietro Pinoli, Anna Sandionigi, and Stefano Ceri. 2021. VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants. *Nucleic Acids Research* 49, 15 (2021), e90.

[12] Michele Bertoni, Stefano Ceri, Abdulrahman Kaitoua, and Pietro Pinoli. 2015. Evaluating cloud frameworks on genomic applications. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 193–202.

[13] Arif Canakoglu, Anna Bernasconi, Andrea Colombo, Marco Masseroli, and Stefano Ceri. 2019. GenoSurf: metadata driven semantic search system for integrated genomic datasets. *Database* 2019 (2019).

[14] Arif Canakoglu, Pietro Pinoli, Anna Bernasconi, Tommaso Alfonsi, Damianos P Melidis, and Stefano Ceri. 2021. ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Research* 49, D1 (2021), D817–D824.

[15] Arif Canakoglu, Pietro Pinoli, Andrea Gulino, Luca Nanni, Marco Masseroli, and Stefano Ceri. 2021. Federated sharing and processing of genomic datasets for tertiary data analysis. *Briefings in Bioinformatics* 22, 3 (2021).

[16] Giuseppe Cannizzaro, Michele Leone, Anna Bernasconi, Arif Canakoglu, and Mark J Carman. 2020. Automated integration of genomic metadata with sequence-to-sequence models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 187–203.

[17] Simone Cattani, Stefano Ceri, Abdulrahman Kaitoua, and Pietro Pinoli. 2017. Bi-dimensional binning for big genomic datasets. In *Proc. of the 4th ACM SIGMOD Workshop on Algorithms and Systems for MapReduce and Beyond*.

[18] Simone Cattani, Stefano Ceri, Abdulrahman Kaitoua, and Pietro Pinoli. 2017. Evaluating big data genomic applications on SciDB and Spark. In *International Conference on Web Engineering*, Vol. 10360. Springer, 482–493.

[19] Luca Cilibrasi, Pietro Pinoli, Anna Bernasconi, Arif Canakoglu, Matteo Chiara, and Stefano Ceri. 2022. ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time. *Bioinformatics* 38, 7 (2022), 1988–1994.

[20] Pietro Crovari, Sara Pidò, Pietro Pinoli, Anna Bernasconi, Arif Canakoglu, Franca Garzotto, and Stefano Ceri. 2021. GeCoAgent: a conversational agent for empowering genomic data extraction and analysis. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–29.

[21] P Flicek and E Birney. 2007 March 30. The European Genotype Archive: Background and Implementation [White paper]. https://www.ebi.ac.uk/ega/sites/ebi.ac.uk.ega/files/documents/ega_whitepaper.pdf

[22] Andrea Gulino, Abdulrahman Kaitoua, and Stefano Ceri. 2018. Optimal binning for genomics. *IEEE Trans. Comput.* 68, 1 (2018), 125–138.

[23] Olha Horlova, Abdulrahman Kaitoua, and Stefano Ceri. 2020. Array-based data management for genomics. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 109–120.

[24] COVID-19 Host Genetics Initiative et al. 2020. The COVID-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *European Journal of Human Genetics* 28, 6 (2020), 715.

[25] Abdulrahman Kaitoua, Pietro Pinoli, Michele Bertoni, and Stefano Ceri. 2016. Framework for supporting genomic operations. *IEEE Trans. Comput.* 66, 3 (2016), 443–457.

[26] C. Maria Keet. 2003. Biological data and conceptual modelling methods. *Journal of Conceptual Modeling* 29, 1 (2003), 1–14.

[27] Marco Masseroli, Arif Canakoglu, Pietro Pinoli, Abdulrahman Kaitoua, Andrea Gulino, Olha Horlova, Luca Nanni, Anna Bernasconi, Stefano Perna, Eirini Stamoulakatou, et al. 2019. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. *Bioinformatics* 35, 5 (2019), 729–736.

[28] Marco Masseroli, Abdulrahman Kaitoua, Pietro Pinoli, and Stefano Ceri. 2016. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods* 111 (2016), 3–11.

[29] Marco Masseroli, Pietro Pinoli, Francesco Venco, Abdulrahman Kaitoua, Vahid Jalili, Fernando Palluzzi, Heiko Muller, and Stefano Ceri. 2015. GenoMetric Query Language: a novel approach to large-scale genomic data management. *Bioinformatics* 31, 12 (2015), 1881–1888.

[30] Amy Maxmen. 2021. Omicron blindspots: why it's hard to track coronavirus variants. *Nature* (2021), 579–579.

[31] Amy Maxmen. 2021. One million coronavirus sequences: popular genome site hits mega milestone. *Nature* 593, 7857 (2021), 21–21.

[32] Luca Nanni, Pietro Pinoli, Arif Canakoglu, and Stefano Ceri. 2019. PyGMQL: scalable data extraction and analysis for heterogeneous genomic datasets. *BMC bioinformatics* 20, 1 (2019), 1–11.

[33] National Human Genome Research Institute. [n.d.]. The cost of sequencing a human genome. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/. Accessed: 2022-04-06.

[34] Norman W Paton, Shakeel A Khan, Andrew Hayes, Fouzia Moussouni, Andy Brass, Karen Eilbeck, Carole A Goble, Simon J Hubbard, and Stephen G Oliver. 2000. Conceptual modelling of genomic information. *Bioinformatics* 16, 6 (2000), 548–557.

[35] Pietro Pinoli, Stefano Ceri, Davide Martinenghi, and Luca Nanni. 2019. Metadata management for scientific databases. *Information Systems* 81 (2019), 1–20.

[36] Sudha Ram and Wei Wei. 2004. Modeling the semantics of 3D protein structures. In *International Conference on Conceptual Modeling*. Springer, 696–708.

[37] Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, and Erik on behalf of COVID-19 Genomics Consortium UK (CoG-UK) Volz. [n.d.]. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. Accessed: 2022-04-06.

[38] José F Reyes Román, Óscar Pastor, Juan Carlos Casamayor, and Francisco Valverde. 2016. Applying conceptual modeling to better understand the human genome. In *International Conference on Conceptual Modeling*. Springer, 404–412.

[39] Eric W Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi. 2019. GenBank. *Nucleic Acids Research* 47, D1 (2019), D94–D99.

[40] Stephan C Schuster. 2008. Next-generation sequencing transforms today's biology. *Nature methods* 5, 1 (2008), 16–18.

[41] Yuelong Shu and John McCauley. 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* 22, 13 (2017).

[42] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. Big Data: Astronomical or Genomical? *PLOS Biology* 13 (07 2015).

[43] The COVID-19 Genomics UK (COG-UK) consortium. 2020. An integrated national scale SARS-CoV-2 genomic surveillance network. *The Lancet Microbe* (2020).

[44] Liangjiang Wang, Aidong Zhang, and Murali Ramanathan. 2005. BioStar models of clinical and genomic data for biomedical data warehouse design. *International Journal of Bioinformatics Research and Applications* 1, 1 (2005), 63–80.

[45] Wenjuan Zhang, Brian D Davis, Stephanie S Chen, Jorge M Sincuir Martinez, Jasmine T Plummer, and Eric Vail. 2021. Emergence of a novel SARS-CoV-2 variant in Southern California. *Jama* 325, 13 (2021), 1324–1326.