A Review of Scaling Genome Sequencing Data Anonymisation

Nikolai J. Podlesny and Anne V.D.M. Kayem and Christoph Meinel

Abstract Sequencing genomes and analysing their variations can make an essential contribution to healthcare research on drug discovery and advancing clinical care, for instance. Genome sequencing data, however, presents a special case of highly sparsely populated, multi-attribute, high-dimensional data, in which each record (tuple) can be associated with more than tens of thousands of attributes on average. Since anonymising genome sequencing data is a necessary pre-processing step for privacy-preserving genomic data analysis for personalised care, discovering all the quasi-identifier combinations required to preserve anonymity is essential; This requires verifying an exponential number of quasi-identifier candidates to identify and remove all unique data values, an NP-hard problem for larger datasets. Furthermore, recent work classifies this problem to be at the very least W[2]-complete and not a fixed-parameter tractable problem. Thus, achieving efficient and scalable anonymisation of genome sequence data is a challenging problem. In this paper, we summarise the uniqueness of ensuring privacy in the context of (whole) genome sequencing. Further, we show and compare the latest trends to discover quasi-identifiers (QID) in large-scale genome data and concepts to counter the exponential runtime growth during QID candidate processing in this field. Finally, we present an architecture incorporating previous enhancements to enable near real-time QID discovery in high-dimensional genome data based on vectorised GPU-acceleration. Achieving anonymisation processing in our experiments in just a few seconds, which corresponds to speedups by factor 100, can be essential in life-or-death situations like triage.

1 Introduction

Sequencing genomes and analysing patients' genomic variations contribute to personalised medicine and research on vaccine discovery, triage, and advancing clinical

Nikolai J. Podlesny and Anne V.D.M. Kayem and Christoph Meinel
Hasso-Plattner-Institute, Potsdam, Germany,
e-mail: {Nikolai.Podlesny,Anne.Kayem,Christoph.Meinel}@hpi.de

care. The technical achievement of (whole) genome sequencing was denominated repetitively as an enormous promise for the greater public good [1] and proved its advancements during the past global pandemic of COVID-19 [2]. Simultaneously, the gained insights in individuals' genomics currently come with fear and cost of its patient's privacy. In contrast to standard high-dimensional datasets from commercial and web data, typically used for experimentation and testing anonymisation algorithms, genome sequence data presents a worst-case scenario of high-dimensional data. Data gathering techniques often produce highly sparsely populated, multi-attribute, high-dimensional whole-genome sequence data that can have up to thousands of describing attributes per record [3]. Protecting this information from private data exposure when sharing for research, triage, or other means is fundamental and an integral part of our highest ethical standards. Yet, established strategies and algorithmic approaches for privacy-preserving data processing and publishing have been summarised as impractical to ensure privacy for these special case data settings [1]. This classification comes back to the NP-hard nature of syntactic data anonymisation techniques and the necessity to verify the exponential number of quasi-identifier candidates. The situation that tens of thousands of attributes need to be checked links back to the algorithmic enumeration problem, yet necessary to find quasi-identifiers that endanger patient privacy. Achieving efficient and scalable anonymisation of genome sequence data is thus a challenging problem.

Contributions. We pick up the open problem of privacy-preserving publishing of patients genome data and make two contributions in this paper:

- summarise, show, and compare latest trends to discover quasi-identifiers (QID) in large-scale genome data and concepts to counter the exponential runtime growth during QID candidate processing in this field
- present an architecture incorporating previous enhancements to enable near real-time QID discovery in high-dimensional genome data based on vectorised GPU-acceleration

Outline of the paper. The rest of the paper is structured as follows. We discuss related work in Section 2. In Section 3, we offer characteristics of genome sequence data and discuss the matter of quasi-identifiers (QID) in genome sequence data. Section 4 addresses scaling patterns for conducting QID discovery on a large-scale and presents the latest concept of accelerating the QID search. We present and compare results from different QID search implementations to enable genome sequence data anonymisation in Section . Section 6 summarises the contributions of this paper and discusses avenues for future research in genome data anonymisation.

2 Related Work

Whole-genome sequencing is the process of determining the order of nucleotides, and their four bases (adenine, guanine, cytosine, and thymine) in a given organism [4]. After comparing individuals' sequence against a standard, variations – known as single-nucleotide polymorphism (SNPs) – are marked to determine the delta. Such

processing results in detailed information on the organism's DNA, which can serve as an essential basis for health-related research in diagnostics, prevention, and exploration. However, as McGuire et al. [5] have pointed out, whole-genome sequencing research raises several ethical considerations such as genetic discrimination, psychological impacts due to inadvertent information exposure, and loss of anonymity. Especially in the health sector and on a genome level, privacy compromises might have dramatic consequences for patients like the de-anonymisation of US Governor William Weld's medical information [6], the exposure of tens of thousands of private health data that included patient names, dates of birth, social security numbers, lab results, and diagnostics data [7].

Already after the completion of the first whole-genome sequences, McGuire et al. raise privacy concerns, discuss complexities of informed consent, and outline the need for an empirical study on the effects of data sharing [5]. Naveed et al. offered survey insights on genome data privacy with biomedical specialists, characterise field-specific privacy problems and provide an enumeration of key privacy challenges in genomics like large scale datasets [8]. Based on a case study on Alzheimer's disease, Wagner substantiates the uniqueness of individual genomic specifics, their highly sensitive information and further offers measures for genomic privacy, metric selection, interpretation, and visualisation options [9]. Humbert et al. present reconstruction attacks based on statistical relationships between the genomic variants through graphical models and belief propagation to expose familial correlations [10]. The "US Presidential Commission for the study of bio-ethical issues" identified genome sequencing as clinical care advancement and enormous promise for the greater public good. Simultaneously, they address data privacy challenges and conclude impracticability for absolute privacy in whole-genome sequencing (WGS) [1].

Guaranteeing privacy technically is challenging since anonymising datasets is known to be NP-hard [11], W[2]-complete and not fixed-parameter tractable (FPT) [12] due to the underlying enumerative combinatorial nature [13]. Whole-genome sequencing (WGS) often produces thousands of attributes per record. Because the attribute values differ from per to person, it is possible to obtain highly sparsely populated, multi-attribute, high-dimensional datasets for genomic analysis with at least tens of thousands of describing attributes (columns) [3]. To avoid private data exposure like in the instance of US Governor William Weld's medical information [6], unique attribute values must be removed as part of the anonymisation processing. These unique attribute values are known as quasi-identifiers (QID) serving in combination with auxiliary data attackers as a link to draw a conclusion and derive private information [14].There is not much research addressing the search for quasi-identifier in the environment of large-scale genome sequencing to the best of our knowledge. Malin et al. presented a method to protect genomic sequences data through generalisation lattices for smaller datasets with less than 400 sequences [15]. Chen et al. offer a scalable approach similar to the "seed-and-extend" method to outsource genome sequence mapping on low-cost cloud platforms [16]. While targeting a secure computation setup based on hashing and fingerprinting for data linkage, the variation mapping outcome may still serve as a quasi-identifier and be used for private data exposure. In the context of genome-wide association studies, Johnson et al. [17]

offer privacy-preserving data mining algorithms introducing non-trivial amounts of random noise, which promise more accurate results on correlation analysis (e.g., SNP disease coherence). Yet, open questions remain towards scalability and practicability of sparsely populated, multi-attribute datasets. Kushida et al. offer a systematic literature review of 1798 prospective citations and conclude that current de-identification strategies have their limitations, full anonymisation is challenging, and further work is needed notably to protect genetic information [18].

As part of this work, we will pick up the previously raised concerns and stated problems with scalability and privacy, particularly data anonymisation for large scale genome sequencing data. To this end, we will provide some background on genome sequencing and the requirements for anonymising genome sequencing data successfully in the next section.

3 Genome Sequence Data Anonymisation

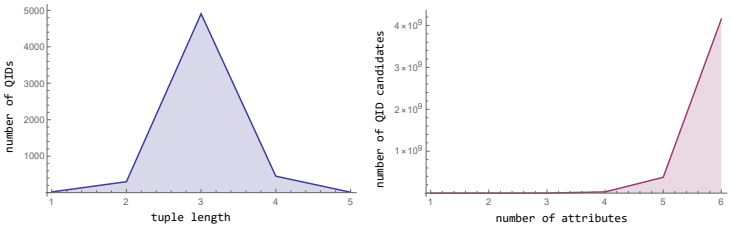
In recent years, however, the increased need to share genome sequence data with third-party data analytics service providers, for instance, fortify the issue of privacy. While whole-genome sequencing (WGS) and data anonymisation are well-explored research areas, their combination lacks practical insights.

Genome Sequence Data: A Characterisation. As part of genome sequencing, the order of nucleotides is determined. For this purpose, the whole genome is itemised in its four bases (adenine, guanine, cytosine, and thymine) [4]. Given this classification, a single whole-genome dataset can be coded as a tuple (POSITION, VALUE) for

$$0 < POSITION < 250M\ (INT),\ VALUE \in \{A, C, G, T, a, c, g, t, *\} \qquad (1)$$

with CHAR(1). This may be represented in a traditional relational data schema combined with an arbitrary patient identifier. When processing nucleotides' order, the idea is to compare the actual genomic sequence to a reference case to identify substitutions for a single nucleotide at a specific position within the genome. Those substitutions or mutations are known as *single-nucleotide polymorphism (SNP)*. Current research suggests the existence of roughly 4 to 5 million SNPs in a person's whole genome making its combination, and especially its mutation composition unique [19, 9]. Even the combination of four or fewer SNPs is often unique and can serve as a quasi-identifier (see Figure 1a). Such uniqueness combined with auxiliary data might result in individual de-anonymisation, draw conclusions, and derive private information [6, 14].

Anonymity Processing of Genomes Sequence Data. The combination of (rare) describing attributes like SNPs, disease history, intake adherence or drug subscriptions can form unique patterns. While anonymity is known as the quality of lacking the characteristic of recognisability or distinction, those unique patterns can be abused to re-identify data records to their original owner. Discovering and dissolving unique patterns through attribute combinations, also known as quasi-identifier, is one of the core principles in data anonymisation. In genomics, one of the most common



(a) Growth of QIDs and candidates in genomic datasets

(b) Growth of QID candidates over increasing attributes

Fig. 1: Genome sequencing data characteristics

mutations is SNP *C677T* within the *MTHFR* enzyme. *C677T* represents a single value mutation, where C (cytosine) is replaced by T (thymidine) at position 677 with the effect of decreasing the enzyme activity by up to 70%. While this particular SNP is relatively commonly represented in the population, combining this insight with several other non-related mutations makes the combination unique. Therefore, individuals' anonymity in a dataset shall be granted if no quasi-identifier remains.

Definition 1. Quasi-identifier
Let $F = \{f_1, .., f_n\}$ be a set of all attribute values and $B := \mathscr{P}(F) = \{B_1, .., B_k\}$ its power set, i.e. the set of all possible attribute value combinations. A set of selected attribute values $B_i \in B$, is called a quasi-identifier, if B_i identifies at least one entity uniquely and all attribute values $f_j \in B_i$ are not standalone identifiers.

During the search for quasi-identifiers (QID) all attribute value combinations as QID candidates need to be assessed which sums up to: $C_2(n) = \sum_{r=1}^{n} \binom{n}{r} = \sum_{r=1}^{n} \frac{n!}{(r!(n-r)!)} = 2^n - 1$ where n is the population of attributes (SNPs) and r the subset of n, while r must equal all potential lengths of subsets of attributes. Figure 1b delineates this exponential candidate growth. A proven and subtle strategy to discover quasi-identifiers is by iterating over all QID candidates, grouping by their attribute values set and counting each group size [20]. Standard aggregate functions like SQL92 can accomplish this. As soon as multiple grouped attributes have a group count, it serves as a quasi-identifier.

Hereinafter, architectural considerations required for realising whole genome sequencing will be considered, quasi-identifiers discovery in practice and on a large scale studied and subsequently data anonymisation activities assessed.

4 Scaling QID search for Genome Sequencing in Practice

Distributing and scaling enumeration problems, like the discovery for quasi-identifiers in large-scale genome sequencing data, is reducible to high-performance computing. Typically, a search scheme can be scaled horizontal or vertical depending on several factors like response time, pricing, scalability demands and network I/O. For

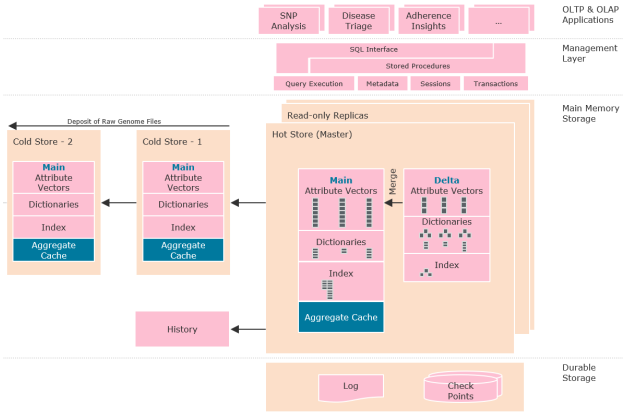


Fig. 2: Vertical Scaling Architecture

the alignment process of genome sequencing data to find mutations, a distributed horizontal architecture approach like Map-Reduce has been implemented in the real-world before [21].

While Map-Reduce is a well-established framework for cheap and easy scaling of computing resources [22], vertical scaling techniques combined with the latest in-memory concepts like dictionary encoding, reverse indices and optimises on L1-L3 cache hits for lightning-fast calculations and reduction of CPU cycles [23, 24] proved their applicability in the past as well [25, 26]. Figure 2 illustrates such architecture.

As mentioned before in Section 3, the QID search can be implemented relatively efficiently through grouping with standard aggregate functions in SQL92 considering a record *count* c of $c = 1$. As a result of this, any column combination which is unique for at least a single row serves as a quasi-identifier. If a group of attributes has less than two and more than none data records, this attribute combination serves as a quasi-identifier, respectively. As hundreds of millions of attribute combinations and therefore aggregation statements needed to be processed, there are options for optimising its execution to achieve the best query runtime possible. Grouping and

Table 1: Cache Hierarchy & CPU Cycle Latency [27]

Data Source	**Cycle Latency (approx.)**
L1 CACHE hit	4 cycles
L2 CACHE hit	10 cycles
L3 CACHE hit	line unshared 40 cycles
L3 CACHE hit	shared line in another core 65 cycles
L3 CACHE hit	modified in another core 75 cycles
remote L3 CACHE	100-300 cycles
Local DRAM	60 ns

aggregation statements are extremely efficient in column-wise storage settings with reversed indices in place [23], as not the entire dataset needs to be loaded, iterated

and parsed. Instead, the courser can jump directly to the corresponding row in the virtual-file, appreciating a reverse-index hash table and allocating just the row's content representing the column values. Such mechanic reflects mainly in vertical scaling settings [24] on the query performance, as parallel aggregation, dictionary encoding and particularly the L1-L3 cache hierarchy significantly accelerates cache hits, reducing the cycle latency (see Table 1). These optimisation features not only fall in place for a single aggregation function but also for executing a variety of grouping statements with overlapping input variables. Here, the same dictionary encoding and intelligent sorting of aggregating tuples may significantly increase L1-L3 cache hits minimising CPU cycles for the entire query runtime. The dictionary itself already contains the count in its index table in the best case. Separately, parallel aggregation can also be applied for multiple grouping operations sharing the same (partial) memory, which again improves cache hits, making the main memory I/O the slowest bottleneck in the hardware chain. Similar observations have been explored by Kessler et al. while implementing *k*-anonymity, and local differential privacy as part of an enterprise database management system view principle [28].

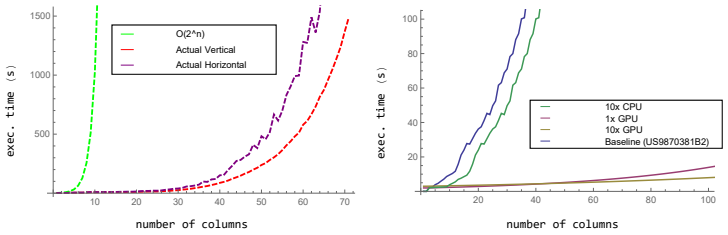
Recent technological enhancements offered the new perspective of leveraging GPU hardware to accelerate computationally intensive tasks through massive-parallelisation. GPUs offer a high compute density, with massive computations per memory access, high throughput, and increased latency tolerance. While GPUs memory capacities are still limited, it also lacks a similar large L1-L3 cache as with CPU architecture. Individual CPU cores are faster and smarter than a single GPU core instruction set. The sheer number of GPU cores and the massive parallelism they offer make up the single-core clock speed difference limited instruction sets. For this purpose, Braghin et al. work [29] can be extended to or rendered as vector operation using standard SQL-like aggregation with out-of-the-box libraries like the open data science framework cuDF, a RAPIDS Nvidia initiative that enables GPU acceleration. This way, given a single Tesla V100, the QID search can be parallelised on 5120 CUDA cores promising a massive runtime acceleration.

These trends and concepts will assess different scaling options in the following section.

5 Experiments

Our empirical model aims to compare QID discovery performance on different scaling and optimisation schemes for large high-dimensional genome datasets. To this end, we mimic real-world environments by employing the following hardware for our experiments.

Hardware. Our experiments were conducted on a GPU-accelerated high-performance compute cluster, housing 160 CPU cores (Xeon Gold 6140), 1TB RAM, and 10x Tesla V100 with a combined Tensor performance of 1120 TFlops. The execution environment for GPU related experiments will be restricted to 10x dedicated CPU core and a single, dedicated Tesla V100 GPU.



(a) Comparison of scaling patterns for QID discovery

(b) Vertical runtime comparison of CPU vs GPU

Fig. 3: Query runtime complexity for quasi-identifier (QID) discovery

Experiments. The evaluation is based on a synthetic genome dataset publicly available at github.com [30]. A variety of work exists exploring cache optimisation options [31] also particularly for aggregation functions on massively parallel processing systems [32, 33]. We are particularly interested in the runtime performance uplift of interacting aggregation functions executed during the QID search scheme. Acknowledging the theoretical time complexity of $O(2^n)$ [13], Figure 3a depicts the outcome of algorithmic runtimes with different scaling patterns using similar emulated environments. A tiny improvement can be derived from Figure 3a in favour of vertical scaling. This is not unexpected as with vertical scaling, the network I/O is reduced and the aggregation based quasi-identifier (QID) search scheme benefits (see Section 4). A much larger acceleration can be seen in Figure 3b, where based on the vertical scaling pattern GPU has been employed to execute the same search scheme in a vectorised manner. Using the available GPU resources, a massive decrease in computation is apparent. While the time complexity remains for the enumeration problem, parallelisation alleviates its runtime effects quite massively by factors of 100x in the given scenario.

Subsuming, we observe a light performance uplift as anticipated with vertical scaling patterns over horizontal ones for the same underlying dataset and similar hardware environment. We expect the query runtime uplift to be fortified with an increasing number of describing attributes. Simultaneously, main memory capacity is a natural limitation towards this approach's practicability, since as soon as swapping takes the place of cache performance, uplifts will collapse. Simultaneously, we observe a massive decrease in runtime for GPU accelerated clusters promising near real-time results to discover privacy endangering quasi-identifiers in high-dimensional datasets.

6 Conclusion & Future Work

Understanding the human genome through its sequencing is often referred to as the most valuable insight possible [1]. Doing it on a large-scale is a huge milestone for personalised care and digital health. In this work, the open problem of privacy-

preserving publishing of patients' genome data has been considered. We presented and compared the latest trends to discover quasi-identifiers (QID) in large-scale genome data and discussed optimisation concepts to counter the exponential runtime growth during QID candidate processing in this field. Further, we present an architecture incorporating previous enhancements, including vectorised GPU-acceleration and showed that it enables near real-time QID discovery in highly sparsely populated, multi-attribute, high-dimensional genome datasets. The experiments confirm that implementing quasi-identifier discovery via standard aggregate functions in SQL92 is a practical solution for large-scale genome sequencing. Combined with GPU hardware, it permits query runtime improvements by more than 100x. As future work, conducting more extensive experiments would help identify re-identification risks concerning auxiliary data. In particular, real-world evidence on data anonymisation's side-effects on the conducted genome-disease correlation analysis would be exciting.

References

1. A Gutmann, J Wagner, Y Ali, AL Allen, JD Arras, BF Atkinson, NA Farahany, AG Garza, C Grady, SL Hauser, et al. Privacy and progress in whole genome sequencing. *Presidential Committee for the Study of Bioethical*, (2012), 2012.
2. Clinton R Paden, Ying Tao, Krista Queen, Jing Zhang, Yan Li, Anna Uehara, and Suxiang Tong. Rapid, sensitive, full-genome sequencing of severe acute respiratory syndrome coronavirus 2. *Emerging infectious diseases*, 26(10):2401, 2020.
3. Ivo Sbalzarini. The Algorithms of Life - Scientific Computing for Systems Biology, June 2019. Keynote talk at ISC High Performance.
4. International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860, 2001.
5. Amy L McGuire, Timothy Caulfield, and Mildred K Cho. Research ethics and the challenge of whole-genome sequencing. *Nature Reviews Genetics*, 9(2):152, 2008.
6. Daniel Barth-Jones. The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. *Then and Now (July 2012)*, 2012.
7. Jessica Davis. Health data, medical documents exposed by labcorp website error, Jan 2020.
8. Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):1–44, 2015.
9. Isabel Wagner. Evaluating the strength of genomic privacy metrics. *ACM Transactions on Privacy and Security (TOPS)*, 20(1):1–34, 2017.
10. Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. Quantifying interdependent risks in genomic privacy. *ACM Transactions on Privacy and Security (TOPS)*, 20(1):1–31, 2017.
11. Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.
12. Thomas Bläsius, Tobias Friedrich, and Martin Schirneck. The Parameterized Complexity of Dependency Detection in Relational Databases. In Jiong Guo and Danny Hermelin, editors, *11th International Symposium on Parameterized and Exact Computation (IPEC 2016)*, volume 63 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 6:1–6:13, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

13. Nikolai J Podlesny, Anne VDM Kayem, and Christoph Meinel. Attribute compartmentation and greedy ucc discovery for high-dimensional data anonymization. In *Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy*, pages 109–119. ACM, 2019.
14. Raymond Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang, and Jian Pei. Minimality attack in privacy preserving data publishing. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB ?07, page 543?554. VLDB Endowment, 2007.
15. Bradley A Malin. Protecting genomic sequence anonymity with generalization lattices. *Methods of information in medicine*, 44(05):687–692, 2005.
16. Yangyi Chen, Bo Peng, XiaoFeng Wang, and Haixu Tang. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. In *NDSS*, 2012.
17. Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087, 2013.
18. Clete A Kushida, Deborah A Nichols, Rik Jadrnicek, Ric Miller, James K Walsh, and Kara Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82, 2012.
19. Lister Hill Center for Biomedical Communications. Genomic research, 2019.
20. Nikolai J Podlesny, Anne VDM Kayem, Stephan von Schorlemer, and Matthias Uflacker. Minimising information loss on anonymised high dimensional data with greedy in-memory processing. In *International Conference on Database and Expert Systems Applications*, pages 85–100. Springer, 2018.
21. Cathrine Jespersgaard, Ali Syed, Piotr Chmura, and Peter Løngreen. Supercomputing and secure cloud infrastructures in biology and medicine. *Annual Review of Biomedical Data Science*, 3:391–410, 2020.
22. Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.
23. Hasso Plattner and Alexander Zeier. *In-memory data management: technology and applications*. Springer Science & Business Media, 2012.
24. Franz Färber, Norman May, Wolfgang Lehner, Philipp Große, Ingo Müller, Hannes Rauhe, and Jonathan Dees. The sap hana database–an architecture overview. *IEEE Data Eng. Bull.*, 35(1):28–33, 2012.
25. Matthieu-P Schapranow, Franziska Häger, and Hasso Plattner. High-performance in-memory genome project: A platform for integrated real-time genome data analysis. In *Proceedings of the 2nd International Conference on Global Health Challenges*, pages 5–10, 2013.
26. Matthieu-P Schapranow, Hasso Plattner, and Christoph Meinel. Applied in-memory technology for high-throughput genome data processing and real-time analysis. *Proceedings of the XXI Winter Course of the Centro Avanzado Tecnológico de Análisis de Imagen*, pages 35–42, 2013.
27. David Levinthal. Performance analysis guide for intel® core™ i7 processor and intel® xeon™ 5500 processors, 2009.
28. Stephan Kessler, Jens Hoff, and Johann-Christoph Freytag. Sap hana goes private: from privacy research to privacy aware enterprise analytics. *Proceedings of the VLDB Endowment*, 12(12):1998–2009, 2019.
29. Stefano Braghin, Aris Gkoulalas-Divanis, and Michael Wurst. Detecting quasi-identifiers in datasets, January 16 2018. US Patent 9,870,381.
30. Nikolai J. Podlesny. Synthetic genome data, 2021.
31. David Michael Pullen and Michael Antony Sieweke. Optimizing cache efficiency within application software, October 17 2006. US Patent 7,124,276.
32. Bhashyam Ramesh, Timothy Brent Kraus, and Todd Allan Walter. Optimization of sql queries involving aggregate expressions using a plurality of local and global aggregation operations, March 16 1999. US Patent 5,884,299.
33. Hasso Plattner, Stephan Mueller, Jens Krueger, Juergen Mueller, and Christian Schwarz. Aggregate query-caching in databases architectures with a differential buffer and a main store, August 22 2017. US Patent 9,740,741.