QnGene: A scalable query engine optimized for analysis of genomic data

[Extended Abstract] *

David Lauzon, Simon Grondin,
Alain April
École de Technologie Supérieure (ÉTS)
1100, rue Notre-Dame Ouest,
Montréal, QC, Canada
david.lauzon.2@ens.etsmtl.ca
simon.grondin.1@ens.etsmtl.ca
alain.april@etsmtl.ca

Keywords

ADAM; BigData; Bioinformatics; Data management; Digital Health; Genomics; GWAS; Software Engineering

1. KEY CHALLENGES

1.1 Big data

Handling large data sets is becoming quite a challenge in the field of bioinformatics. Bioinformatics softwares such as plink[9] are not designed to scale automatically to multiple computing nodes [5] and to process rapidly very large datasets, on demand.

1.2 Multiple softwares and file schemas

Implementing genome-wide association studies (GWAS) requires the intricate knowledge of multiple genomic softwares [7]. The know-how to query, filter, extract, aggregate and perform analysis is specific to each software. As they also use their own set of unique file schemas, this results in ”the parsing and manipulation of data the most time consuming and error prone part of a study” [7].

1.3 Multiple data sources

To raise another challenge, researchers want to integrate data from multiple sources such as dbSNP[10], Cartagene[1], genotypes, clinical databases, and GWAS results, to cite but a few.

1.4 Fast insight

Researchers frequently request urgent ad hoc queries, which are slowed down because of the many manual operations that must be done: data extraction, data conversion, joining, aggregates, etc. The majority of bioinformatics softwares are driven by command line interfaces (CLI) instead of graphical user interfaces (GUI).

1.5 Extensibility

When a use case cannot be fulfilled completely by installed softwares, the bio-informaticians are stucked with a though choice: 1) find, learn, install, and configure a new software (if it exists); or 2) write custom code from scratch. Depending on the complexity of the need, the latter can be a tedious task since most GWAS softwares do not have an application programming interface (API).

2. CURRENT SOLUTIONS AND SHORTFALLS

2.1 OLTP Relational databases

A popular solution to the analysis of genomic data is to export the data into an OLTP relational database, and use structured query language (SQL) commands to manipulate and query the data. SQL is generally easier than the query language of a specific individual bioinformatics software such as plink[9] and allows to join multiple data sources with little effort. The main drawbacks are that 1) modelling the database correctly is hard and takes expertise, 2) traditional OLTP databases do not scale easily to handle very large amount of data (i.e. which is now more common), and 3) newly generated data must be constantly integrated back into the database.

2.2 Workflow platforms

Workflow platforms for genomic analysis such as Galaxy[2, 3, 4] helps to transition from a software to another. Their current strategy proposed is to develop a GUI for each supported software. This approach has drawbacks such as: 1) it lags behind the underlying softwares in terms of functionality; 2) not all software are supported and it's tedious to integrate a new software; and 3) sometimes plain SQL is just more efficient to use and less cumbersome for power users.

2.3 ADAM

These solutions do not scale well when BigData size has to be processed and they are not optimized for sharing genomic experiments between researchers.

ACM DH '16 April 11–13, 2016, Montreal, QC, CANADA

ACM ISBN T.B.A..
DOI: T.B.A.

ADAM[8], an open source project developed by UC Berkeley's Amplab and supported by the National Institute of Health provides both an application programming interface (API) and a command line interface (CLI) to manipulate sequencing data at very large scale. ADAM is leveraging the most advanced BigData technology and thus scales well with very large genomic datasets, but does not yet have support for genomic analytic data structures.

3. PROPOSED SOLUTION

3.1 Open source GWAS data warehouse

We are currently developing QnGENE, an open source software that we believe could lay the ground work for a generic platform to store and analyze the genomic data of clinical databases, such as Cartagene[1].

QnGene will provide a central data store to integrate all the datasets generated during GWAS experiments.

3.2 Generic schemas

First, we will design generic data structures for genomic analysis that we'll contribute to the ADAM schemas, which already support sequencing (fasta, BAM, SAM, VCF) and dbSNP data. The plan is to add support for patients, phenotypes, GWAS results, and clinical data into ADAM.

Different but related schemas (e.g. plink's 20+ gwas file schemas) will be merged into a single standard schema to simplify the data manipulations and minimize manipulation errors.

3.3 Scalable

The data structures discussed in 3.2 will be pre-joined which will speed up both writing the query and the processing of the query. Akin to ADAM, storage and processing of this data can be done locally or efficiently distributed on a cluster. This will avoid the burden to manually split the data and merge it afterward (handled by QnGene).

3.4 Compatibility

We'll support importing from and exporting to the current most popular GWAS softwares (exact list to be determined) such as plink[9] and snptest[6]. This will eliminate the need for manual conversions and daily struggles with their numerous file schemas.

3.5 Simple generic unified data access

Then, we will offer both CLI and SQL interfaces over files stored using our extended ADAM schemas. Common data management operations such as querying, filtering, sorting, extracting will be standardized across all supported schemas.

Therefore, it won't be necessary to learn how to access the data from each individual software. Also, QnGENE will allow to access and join multiple files - a currently non-trivial task in genomic analysis.

3.6 Extensible

QnGENE will expose its internal building blocks via an API that allows a programmer to easily extend its initial capabilities and benefit from its scalability.

Therefore minimal effort will be required to extend QnGENE with a custom algorithm for which a mathematical formula is known.

3.7 Analysis features

At first, analysis will still need to be performed with existing bioinformatics softwares. But the extensible nature of QnGENE will make it easy to progressively incorporate the analysis algorithms that need better performance, whether they are developed internally or contributed by the community.

We're currently looking to improve the scalability of the associations and dosage modules of plink[9], and integrate these algorithms into QnGene.

4. REFERENCES

[1] P. Awadalla, C. Boileau, Y. Payette, Y. Idaghdour, J.-P. Goulet, B. Knoppers, P. Hamet, and C. Laberge. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *International Journal of Epidemiology*, 42(5):1285–1299, 2013.

[2] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor. *Galaxy: a web-based genome analysis tool for experimentalists.*, volume Chapter 19. 2010.

[3] B. Giardine, C. Riemer, and R. Hardison. Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, 15:1451–1455, 2005.

[4] J. Goecks, A. Nekrutenko, and J. Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, jan 2010.

[5] H. Huang, S. Tata, and R. J. Prill. Bluesnp: R package for highly scalable genome-wide association studies using hadoop clusters. *Bioinformatics*, 29(1):135–136, 2013.

[6] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

[7] F. Muñiz-Fernandez, A. Carreño, C. Morcillo-Suarez, A. Navarro, et al. Genome-wide association studies pipeline (gwaspi): a desktop application for genome-wide snp analysis and management. *Bioinformatics*, 27(13):1871–1872, 2011.

[8] F. A. Nothaft, M. Massie, T. Danford, Z. Zhang, U. Laserson, C. Yeksigian, J. Kottalam, A. Ahuja, J. Hammerbacher, M. Linderman, M. J. Franklin, A. D. Joseph, and D. A. Patterson. Rethinking Data-Intensive Science Using Scalable Analytics Systems Categories and Subject Descriptors. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015.

[9] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[10] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.