BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	ISAnalytics: a new bioinformatic R package for integrated analysis of viral integration sites in gene therapy applications
All Authors	Pais G(1), Spinozzi G(1), Montini E(1), Calabria A(1)

Affiliation

(1) San Raffaele Telethon Institute for Gene Therapy, Ospedale San Raffaele

Motivation

Gene therapy is an emerging field for the treatment of genetic diseases. Viral vectors are commonly used as delivery tool of the therapeutic transgene. Monitoring safety and long-term efficacy of the treatment exploits viral integration site (IS) analysis allowing the tracking of clones and their progeny over time, tissues and lineages. To retrieve and analyze the whole repertoire of IS in in-vivo patients, as well as in animal models, genomics technologies combined with high-throughput platforms are used, generating increasing amount of large data to be integrated and processed as a whole, leading to a rapid scale-up of the overall volume of data. This data growth and accumulation is not paired by hardware improvements, thus posing computational limitations for the analysis of biosafety and long-term efficacy. Such limitations acted as the driving force for the development of an integrated scalable tool, enabling clonal studies in gene therapy applications. Several requirements and computational aspects deeply impacted on the design and development process of the solution, such as the structure of input and output data, the ease of use, the scalability in terms of memory footprint and computational time, and finally the reproducibility of results, along with the production of in-depth documentation with reproducible examples. Since bioinformatics tools for IS retrieval already exist, such as VISPA2, we designed a downstream application that can process IS datasets both from the output of VISPA2 and from other custom sources.

Methods

Data structure was one of the most crucial aspects that was taken into account in the designing phase, since it impacts on performance both in terms of time and space. Usually, IS data are released in the form of a sparse matrix, where rows are IS (cases) and columns are samples (variables), with the exception of the first 3-5 columns, which are always genomics features. In contrast with the current data structure, which contains a very high density of non-significant values, an alternative data representation was proposed: this structure, commonly referred to as “tidy”, is characterized by a percentage of non-significant values equal to zero, which leads to lower memory impact and consequently a significant decrease in computational time for recurring operations. Another key aspect under investigation was the performance bottleneck constituted by the import of data files into main memory to allow manipulation and analysis: since the root of the problem is a single reading operation of a large sized file, we proposed a “divide-et-impera” solution with parallelism exploitation to decrease computational time and flatten memory peaks. Both our proposed solutions were tested against the currently used procedures using bootstrap simulations and benchmarking to assess the potential gain/loss in terms of time and memory consumed: 151 random samples were generated by applying a series of modifiers with bootstrapping approach respectively 5%, 20%, 50% and 100%, on both rows and columns. Benchmarks for the data structure comparison were performed by applying two versions of the same function which produces aggregated data: the first works on the original sparse-matrix-like data structure, while the second one works on the new data structure.

Results

The results of the preliminary benchmarks confirmed our initial hypotheses both on data structure and data import. For data structure comparisons, benchmarks showed, on average, a fold-change of 104 to 106 between the classical approach and the new approach in terms of computational time elapsed (measured in milliseconds). For the import strategies comparison, only a brief exploratory analysis was performed prior the actual package development since a decrease in computational time was expected, due to the intrinsic nature of parallel computation algorithms, but it was later integrated by a use-case scenario performed on real clinical trial data post package development: again, we observed a fold-change of 104 to 106 between the classical approach and the “divide-et-impera” approach in terms of computational time elapsed, and an estimated decrease in RAM peaks (measured in MB) of 106-fold.

Besides improved performance, ISAnalytics provides a metadata-driven operation workflow with an automated and detailed report system, which is adaptable to many other application fields other than gene therapy.

ISAnalytics provides a system of integrated automated reports for the majority of functionalities: reports are saved in HTML format and often include interactive widgets, plots and various amounts of information which is crucial in large and complex scenarios to ensure the reproducibility of results and easier monitoring for potential problems from the very beginning.

Info	
-	
Figure	
-	
Availability	https://doi.org/doi:10.18129/B9.bioc.ISAnalytics

Corresponding Author	
Name, Surname	Giulia, Pais
Email	pais.giulia@hsr.it
Submitted on	29.04.2021