On the many advantages of using the VariantExperiment class to store, exchange and analyze SARS-CoV-2 genomic data and associated metadata

Jérôme Ambroise[1,*], Laurent Gatto[2,*], Julie Hurel[1], Bertrand Bearzatto[1], Jean-Luc Gala[1]

*The first two authors contributed equally to this work

1: Center for Applied Molecular Technologies, Institute of Clinical and Experimental Research, Université Catholique de Louvain (UCLouvain), 1200 Brussels, Belgium

2: Computational Biology and Bioinformatics Unit, de Duve Institute, Université Catholique de Louvain (UCLouvain), 1200 Brussels, Belgium

Abstract

On Friday, 19 March 2021, WHO organized a virtual global workshop highlighting the need for a globally coordinated plan to increase SARS-CoV-2 genetic sequencing capacities to detect SARS-CoV-2 mutations and variants, and to monitor virus genomic evolution worldwide. One week later, in another virtual meeting, it focused on sero epidemiology for SARS-CoV-2 variants of concern and variants of interest. Efficient monitoring of the virus relies on the storage, handling and sharing of the genomic data and the associated metadata. In this manuscript, we demonstrate how the Bioconductor VariantExperiment class addresses these needs, offering a robust and efficient solution to the requirements laid out by the WHO.

Background

The first COVID-19 case was reported in Wuhan city, Hubei province of China in December 2019. The pathogen causing the disease was soon identified as a novel coronavirus, closely related to severe acute respiratory syndrome coronavirus (SARS-CoV), and renamed novel coronavirus SARS-CoV-2. In light of the growing number of cases at Chinese and around the world, the World Health Organization (WHO) Emergency Committee declared a global health emergency on 30 January 2020.

On Friday, 19 March 2021, WHO organized a virtual global workshop on enhancing sequencing for SARS-CoV-2 in the context of the elaboration of a globally coordinated plan to increase SARS-CoV-2 genetic sequencing capacities to detect SARS-CoV-2 mutations and variants, and to monitor virus genomic evolution worldwide. WHO is indeed working with Member States and partners to increase SARS-CoV-2 sequencing capacities and encourage timely sharing of geographically representative sequences and supporting data. One of the main objectives of the coordinated plan is to promote an efficient discovery and reporting of new Variants of Concern (VOC) among the many existing and upcoming genetic variants within the SARS-CoV-2 genome. During this workshop, the importance of metadata and quality metrics of SARS-CoV-2 genomic data was emphasized. In particular, the lack of current minimum standards of metadata and the need for a system enabling to keep all components well synchronized was emphasized. Being able to store and analyze SARS-COV2 genomic data and associated high-quality metadata will favor the efficient discovery of VOC, Variant of Interest (VOI), or Variant of High Consequences in the upcoming months [1]. Of course, this will be enhanced by sharing of all genomics and associated metadata between countries and continents.

Bioconductor is an open-source, open-development software project for the analysis and comprehension of high-throughput data in genomics and molecular biology. The Bioconductor project aims to enable inter-disciplinary research, collaboration and rapid development of scientific software [2]. Among the packages developed and maintained by the core developers of the project, VariantAnnotation and VariantExperiment can be used to import data from VCF and GDS files and store them into structured data model objects [3,4].

The aim of this paper is to demonstrate the many advantages of using these Bioconductor packages for storing, sharing, and analyzing SARS-CoV-2 genomic data and associated metadata. In order to illustrate this purpose, an object of the VariantExperiment class was created with publicly available SARS-COV-2 genomic (Whole Genome Sequencing WGS) data and used in a R markdown as a demonstration tool.

Methods

In order to illustrate the advantages of using the VariantExperiment data model, we demonstrate how it can be used to create and manipulate genomics SARS-CoV2 data. To this end, 10 fastq raw data files generated by the MinION Oxford Nanopore Technology (ONT) were downloaded from the European Nucleotide Archive (ENA) repository. Data were mapped to the reference genome MN908947.3 with minimap2 2.17 [5] and SAM files were processed using samtools 1.9. [6] The calling of the mutations was performed using freebayes 1.3.2 [7]. After quality filtering with vcffilter, the resulting VCF file and a dummy metadata (including dummy patients characteristics as well as preanalytical, analytical and bioinformatics parameters) file were processed with the VariantExperiment and the VariantAnnotation packages in order to create the desired object of the VariantExperiment class. Finally, an R markdown report and its html rendering[1] were created to illustrate the structure and benefit of the object [8]**.**

Results

The Rmarkdown which illustrates the structure and the manipulation of the VariantExperiment object created from SARS-CoV-2 genomic data is available on github[2]. The structure of the VariantExperiment class is illustrated in Figure 1. The advantages of such Bioconductor objects in the context of SARS-COV2 genomic data analysis are described below.

[1] https://uclouvain-cbio.github.io/VariantExperiment-COVID19-UseCase/

[2] https://github.com/UCLouvain-CBIO/VariantExperiment-COVID19-UseCase

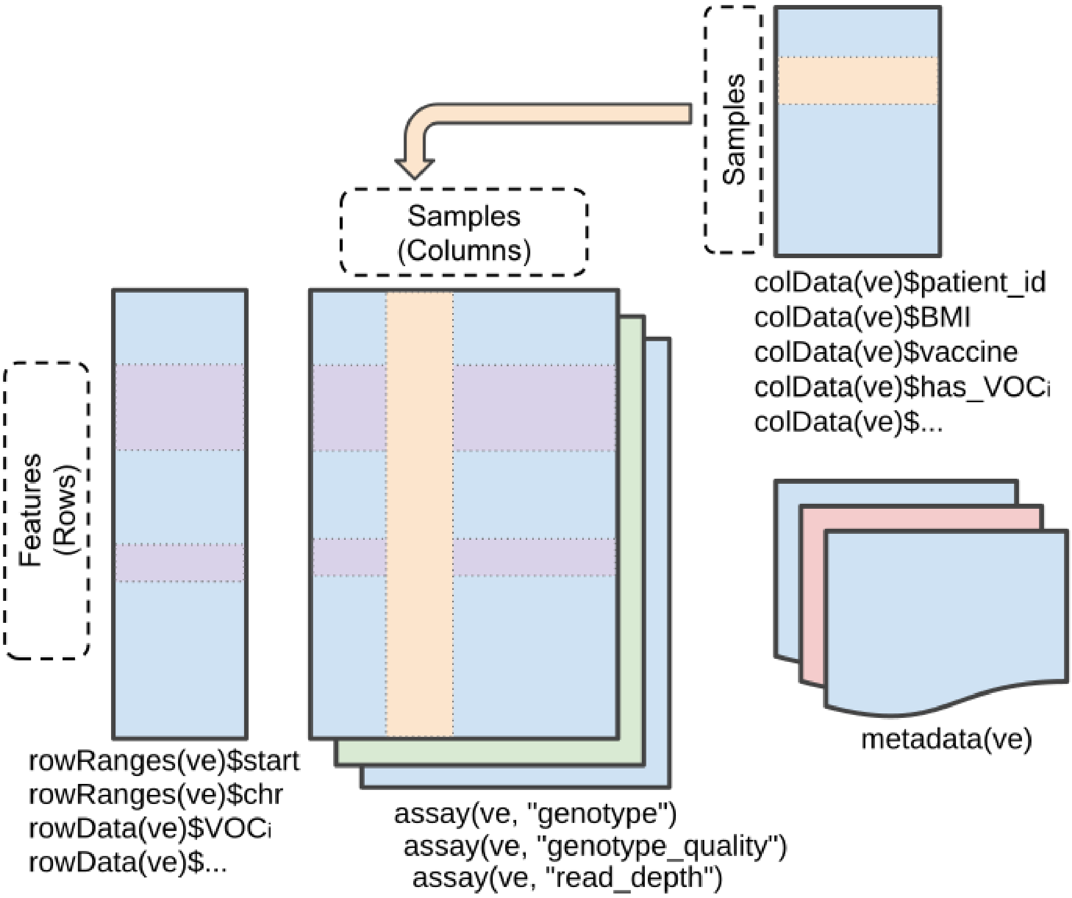


Figure 1: Illustration of a VariantExperiment object for handling SARS-COV2 genomic data and associated metadata. The latter should be interpreted in the broad sense of the term and includes sample characteristics, pre-analytical, analytical and data processing steps (stored in the colData component) as well as mutation and variant characterisation (stored in the rowData component). The generation and manipulation of such an object is demonstrated in our accompanying vignette [8]. The figure is adapted from the SummarizedExperiment package with permission of the author, Jim Hester.

Genomic data

The VariantExperiment object model enables to store .vcf and .gds data into a structured data model, namely RangedSummarizedExperiment object [9]. These objects contain one or more assays, each represented by a matrix-like object. In the context of the SARS-CoV-2 genomic data, the first assay should therefore contain the called genotype (Figure 2, left). Additional assay can be used to store quality metrics associated with the genotype calling (e.g. genotype quality calculated as a Phred-scaled probability of the called genotype, read depth on Figure 2, right). It is worth noting that large assays can also be stored on disk (rather than in memory), thus enabling the handling and processing of large datasets (i.e., hundreds of variants and thousands of patients) that would not fit in memory.

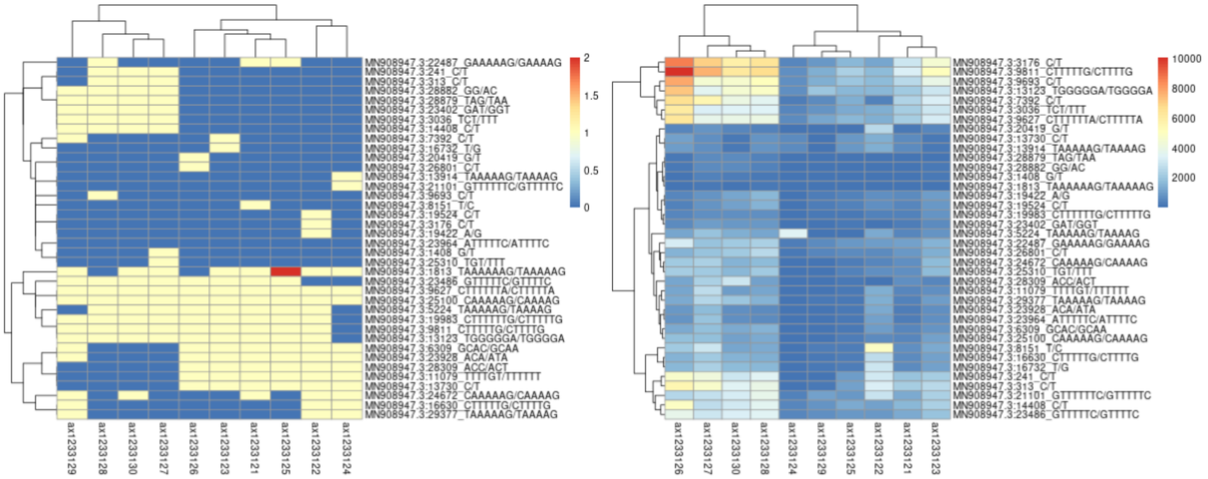


Figure 2: Visualisation of the two assay slots stored in our example data. On the left, the genotypes, encoded as 0s for the reference allele, and 1, 2, .. for alternative references. On the right, we see the sequencing read depth and how some samples have systematic low read depth (blue).

Metadata

The metadata terms includes a wide range of data annotations, including sample characteristics, as well as variant and mutation characterisation. Linking genomic data and their associated metadata is crucial in omics sciences. Accordingly, international collaboration should be put in place in order to standardize the metadata ontology.

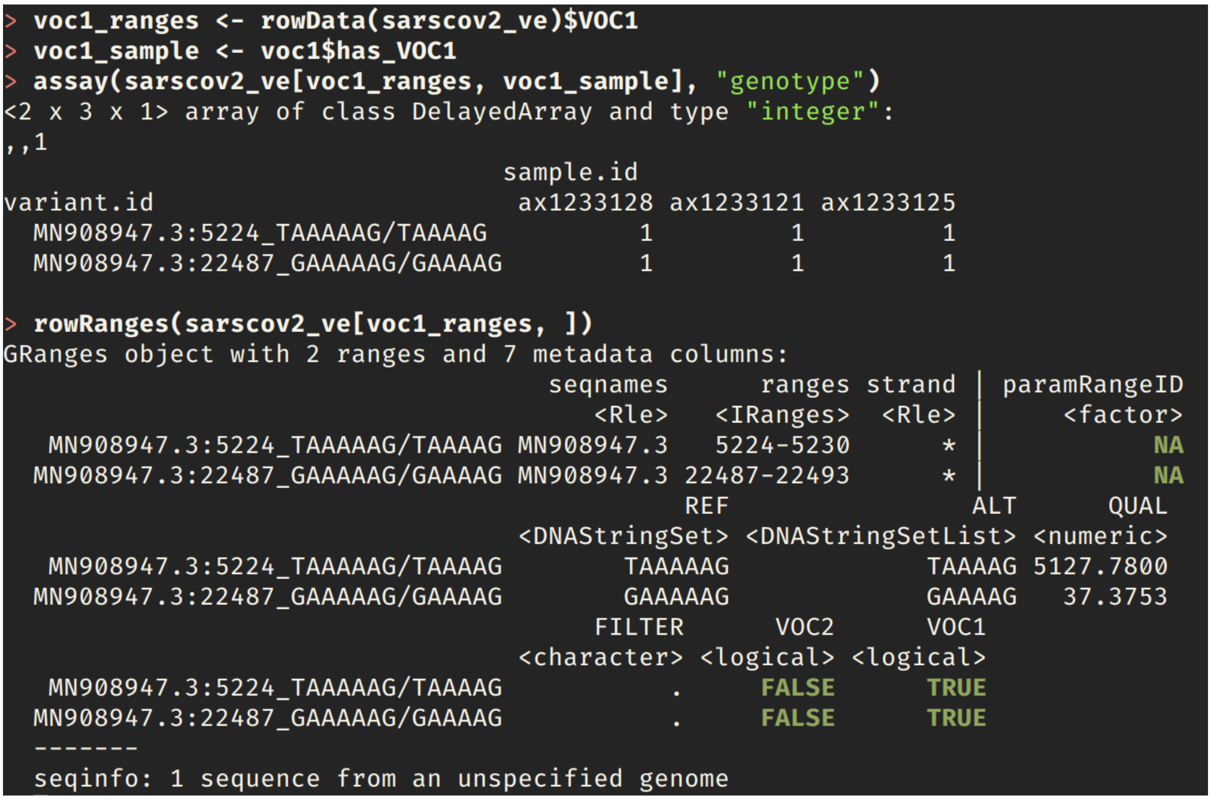
When using the VariantExperiment infrastructure, sample and feature metadata are saved as tabular data (DataFrame or DelayedDataFrame, when the data is large and requires on-disk access). The sample metadata, referred to specifically as *colData* (column/sample metadata) offers a structured way to describe sample characteristics (such as patient identifiers, age, any clinical variables such as BMI, or any comorbidities, whether patients were vaccinated, with which vaccine and when), pre-analytical and analytical parameters (such as library preparation or the sequencing platform that was used to generate the data), as well as indications about the specific bioinformatics pipelines (algorithms and software versions) that were used for batches of samples of various origins. The individual rows (features) are also annotated using exact genomics coordinates [10], protein-level alteration, as well as variants specification (e.g.VOC; see below) to cite a few. These annotations are referred to as *rowData* or *rowRanges* (when referring to genomic coordinates). As for the genomic data, that rapid accumulation of annotations will become a computational bottleneck, the on-disk representation of the metadata can be used to reduce its memory footprint, enabling the processing of many large files.

Another key aspect of this data model is the coordination of the assays and their row and columns metadata during all data manipulation, subsetting and processing. Accordingly, any subsetting along the assay rows (features) or columns (samples) are propagated along the colData and rowData/rowRanges metadata in a coordinated fashion. This property therefore ensures to keep a consistent synchronization between genomic data and all associated metadata.

Mutation and variant characterization

As illustrated in Figure 1, the description of each called mutation is saved and annotated in the rowData and rowRanges metadata slots. In the context of the analysis of SARS-COV2 genomic data, this component should include several columns including the mutation position along the virus's genomic coordinates as well as nucleic acid and amino acid changes. Moreover, an indicator may be inserted to reflect if a mutation is part of VOC or VOI signature, and should regularly be updated according to the evolution of scientific knowledge. One such indicator column should be added for each VOC and VOI (Figure 3).

On Friday 26th March WHO organized a meeting focusing on sero epidemiology for SARS-CoV-2 VOC. During this meeting WHO already reported 3 and 6 confirmed VOC and VOI, respectively. However, 16 other signals of potential VOIs/VOCs are in assessment. Accordingly, the number of columns in the rowData metadata slots is expected to grow in the following months/years. Conversely, samples can be annotated as having been infected by a particular VOC or VOI in the colData component (Figure 3). Examples of such variant and infection characterisation are documented in our companion vignette.



```
> voc1_ranges <- rowData(sarscov2_ve)$VOC1
> voc1_sample <- voc1$has_VOC1
> assay(sarscov2_ve[voc1_ranges, voc1_sample], "genotype")
<2 x 3 x 1> array of class DelayedArray and type "integer":
,,1
                                 sample.id
variant.id                        ax1233128 ax1233121 ax1233125
  MN908947.3:5224_TAAAAAG/TAAAAG          1         1         1
  MN908947.3:22487_GAAAAAG/GAAAAG         1         1         1

> rowRanges(sarscov2_ve[voc1_ranges, ])
GRanges object with 2 ranges and 7 metadata columns:
                                    seqnames      ranges strand | paramRangeID
                                       <Rle>   <IRanges>  <Rle> |     <factor>
   MN908947.3:5224_TAAAAAG/TAAAAG MN908947.3   5224-5230      * |           NA
  MN908947.3:22487_GAAAAAG/GAAAAG MN908947.3 22487-22493      * |           NA
                                             REF               ALT      QUAL
                                  <DNAStringSet> <DNAStringSetList> <numeric>
   MN908947.3:5224_TAAAAAG/TAAAAG        TAAAAAG            TAAAAG 5127.7800
  MN908947.3:22487_GAAAAAG/GAAAAG        GAAAAAG            GAAAAG   37.3753
                                       FILTER      VOC2      VOC1
                                  <character> <logical> <logical>
   MN908947.3:5224_TAAAAAG/TAAAAG           .     FALSE      TRUE
  MN908947.3:22487_GAAAAAG/GAAAAG           .     FALSE      TRUE
  -------
  seqinfo: 1 sequence from an unspecified genome
```

Figure 3: Illustration of the definition of a variant of interest (VOC1) and how the VOC1 indicator is used to extract the alleles that define it (`voc1_ranges` variable). The `voc1_sample` variable selects the samples that have been annotated as being infected by that variant. We then show how the 3 samples have the alternative allele (a deletion) in the 2 sites at positions 5224-5230 and 22487-22493 respectively.

Conclusions

In this paper, we illustrate how the VariantExperiment Bioconductor package can be used to store, share and analyze SARS-COV2 genomic data and their associated metadata. For many years, the Bioconductor infrastructure has proved to match the requirements for handling omics data (e.g. bulk or single-cell transcriptomics, genomics, proteomics, ….). To this end, Bioconductor employs a flexible object-oriented paradigm that enables to encapsulate multiple object components into a single instance and preserve the relations between primary (i.e. genomic, transcriptomic, ..) data and metadata. As proposed here, the integration of SARS-CoV-2 genomic data and associated metadata in the VariantExperiment object model stenghtens the discovery and annotation of new VOCs and VOIs among the fast expanding list of newly emerging SARS-CoV-2 variants. Moreover, the VariantExperiment object may include other complementary data such as quality metrics and technical parameters, among which pre-analytical (i.e library preparation kit), analytical (sequencing technology) and bioinformatics (software version and parameters) data. Integrating the latter into the same object would substantially broaden the field of genomic data exploitation, e.g. enable stakeholders to answer queries that address the impact of technical parameters on data quality.

Funding

Julie Hurel is supported by the PANDEM-2 european project (Grant Agreement 883285). The PANDEM-2 project implements and demonstrates the most important novel concepts and IT systems to improve the capacity of European pandemic detection, monitoring and response. The funding body did not play any roles in the design of the study and in writing the manuscript.

References

1. CDC. Cases, Data, and Surveillance. 24 Mar 2021 [cited 31 Mar 2021]. Available: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html

2. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12: 115–121.

3. Liu Q, Pagès H, Morgan M. VariantExperiment: A RangedSummarizedExperiment Container for VCF/GDS Data with GDS Backend. 2020. Available: https://github.com/Bioconductor/VariantExperiment

4. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014;30: 2076–2078.

5. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100.

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

7. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv [q-bio.GN]. 2012. Available: http://arxiv.org/abs/1207.3907

8. Gatto L, Ambroise J. Using a VariantExperiment data object to manage COVID19 data. 2021. doi:10.5281/zenodo.4663114

9. Morgan M, Obenchain V, Hester J, Pagès H. SummarizedExperiment: SummarizedExperiment container. 2020. Available: https://bioconductor.org/packages/SummarizedExperiment

10. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9: e1003118.