


GenomicScores

–latest developments and future challenges–

Robert Castelo

robert.castelo@upf.edu @robertclab

Dept. of Experimental and Health Sciences
Universitat Pompeu Fabra
Barcelona, Spain

European Bioconductor Conference
Brussels, Belgium
December 9th, 2019



Genome analysis

GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor

Pau Puigdevall and Robert Castelo*

Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona 08003, Spain

*To whom correspondence should be addressed.

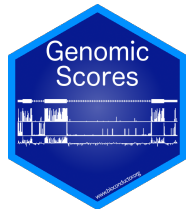
Associate Editor: John Hancock

Received on January 8, 2018; revised on April 12, 2018; editorial decision on April 13, 2018; accepted on April 25, 2018

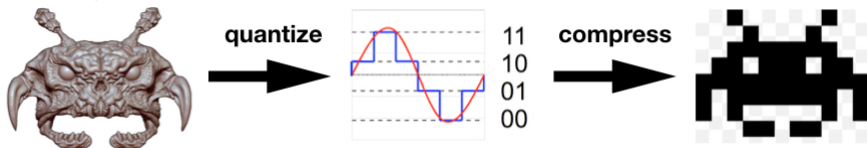
Abstract

Summary: Genomewide position-specific scores, such as those estimating conservation, constraint, fitness or mutation tolerance, are ubiquitous in current genome analyses. The diversity of sources and formats of these scores, as well as their size, increase the burden to use them. We present *GenomicScores*, a Bioconductor package that provides efficient storage and seamless access of genomewide position-specific scores from R, facilitating their use in genome analysis workflows.

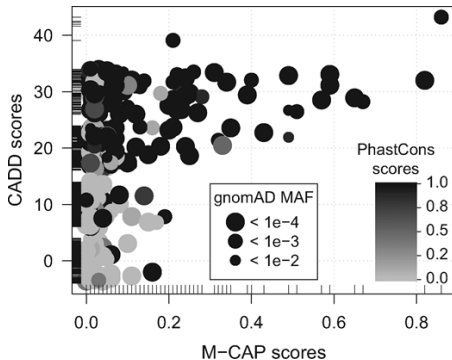
<https://doi.org/10.1093/bioinformatics/bty311>



Approach: lossy compression



Different types of scores may be quantized differently



GScores objects through annotation packages

```
> library(MafDb.gnomAD.r3.0.GRCh38) ## minor allele frequencies from gnomAD 3.0
> mafdb <- MafDb.gnomAD.r3.0.GRCh38
> mafdb
```

GScores object

```
# organism: Homo sapiens (UCSC, hg38)
# provider: BroadInstitute
# provider version: r3.0
# download date: Oct 18, 2019
# loaded sequences (SNRs): chrY, chr15
# loaded sequences (nonSNRs): chr3
# loaded populations (SNRs): AF, AF_afr
# loaded populations (nonSNRs): none
# default scores population: AF
# number of sites: 595 millions
# maximum abs. error (def. pop.): 0.00273
# use 'citation()' to cite these data in publications
```

```
> citation(mafdb)
```

Konrad J Karczewski, et al. (2019). "Genetic Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes." *bioRxiv*, 531210. doi: 10.1101/531210 (URL: <https://doi.org/10.1101/531210>).

```
> ## CCR5-delta32 reported to be protective against HIV-1 infection
> gscores(mafdb, GRanges("3:46373452-46373484"), type="nonsnrs")
```

GRanges object with 1 range and 1 metadata column:

	seqnames	ranges	strand	AF
	<Rle>	<IRanges>	<Rle>	<numeric>
[1]	chr3	46373452-46373484	*	0.07

seqinfo: 1 sequence from hg38 genome; no seqlengths

Latest developments: individual allele frequencies

chr start

chr end

...	0.12	...	2·10e-5	...	7·10e-4	...
-----	------	-----	---------	-----	---------	-----

no values MAF no values MAF no values MAF no values

raw-Rle of length 57227415 with 593692 runs

```
Lengths: 2781621      1      20      1 ...      1      28      1 339684
Values :      00      7f      00      80 ...      71      00      72      00
```

...	TRUE	...	FALSE	...	TRUE	...
-----	------	-----	-------	-----	------	-----

no values isREF no values isREF no values isREF no values

raw-Rle of length 57227415 with 27 runs

```
Lengths: 7592292      1 466749      1 ...      5      1 543803
Values :      00      01      00      01 ...      00      01      00
```

```
> gscores(mafdb, GRanges("15:28111713"), pop=c("AF", "AF_afr"), ref="C", alt="T")
```

GRanges object with 1 range and 4 metadata columns:

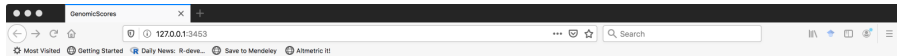
```
  seqnames   ranges strand |   AF_REF AF_afr_REF   AF_ALT AF_afr_ALT
    <Rle> <IRanges> <Rle> | <numeric> <numeric> <numeric> <numeric>
[1] chr15 28111713 * |      0.5      0.87      0.5      0.13
-----
```

seqinfo: 1 sequence from hg38 genome; no seqlengths



Latest developments: shiny web app

- First version of a shiny web app (developed by Pablo Rodriguez).
- Choose among installed annotation packages with *GScores* objects.
- Enter genomic coordinates manually or uploading a BED file.
- Browse results or download them as a BED or CSV file.



GenomicScores WebApp

Select a GScore object

MafDb.gnomAD.r3.0.GRCh38

Select an available population

AF AF_afr AF_amr AF_fin

Input genomic coordinates

Manually Uploading BED file

Chr name Start End

15 28111713 28111713

Output type

Genomic range Individual positions

GScores

About Session Info

```
GScores object
# organism: Homo sapiens (UCSC, hg38)
# provider: BroadInstitute
# provider version: r3.0
# download date: Oct 18, 2019
# loaded sequences (SNRs): chrY
# loaded sequences (nonSNRs): none
# default scores population: AF
# number of sites: 595 millions
# maximum abs. error (def. pop.): 0.00273
# use 'citation()' to cite these data in publications
```

Konrad J Karczewski, et al. (2019). "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes." *_bioRxiv_*, 531210. doi: 10.1101/531210 (URL: <https://doi.org/10.1101/531210>).

Show 10 entries

Search:

	seqnames	start	end	width	strand	AF	AF_afr	AF_amr	AF_fin
1	chr15	28111713	28111713	1	*	0.5	0.13	0.25	0.13

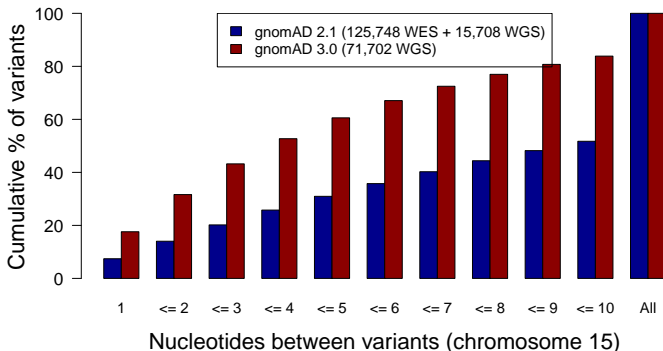
Showing 1 to 1 of 1 entries

Previous 1 Next

[Download BED](#) [Download CSV](#)

Future challenges

- Individual allele frequencies are currently correct for biallelic variants only. How to efficiently store them for multiallelic variants?
- Current growth of gnomAD catalogs makes *Rle* vectors increasingly less efficient. What could be an alternative?



Suggestions, bugs, feature requests and acknowledgments

- Suggestions to robert.castelo@upf.edu (everything) and pablosebastian.rodriguez@upf.edu (shiny app); stickers available!
- Bugs and feature requests through <https://github.com/rcastelo/GenomicScores/issues>
- User questions through <https://support.bioconductor.org>
- Acknowledgments to:
 - The Bioconductor core team for their continuous advice and support in developing GenomicScores.
 - Funding: TIN2015-71079-P (MINECO/FEDER, UE).

