SUPPLEMENTARY INFORMATION

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

Pau Puigdevall Robert Castelo*

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

* To whom correspondence should be addressed (robert.castelo@upf.edu).

This is the supplementary information of the article on the `GenomicScores` Bioconductor software package version 1.3.24. For documentation on specific funcionality and working examples of use of the package please consult its reference manual and vignette, both available from the package landing web page at:

`https://bioconductor.org/packages/GenomicScores`

for Bioconductor release version 3.7 or higher.

Supplementary Note

To have a sense of the extent of the trade-off between precision and compression in a specific case, we compare here original *phastCons* scores with the ones obtained by rounding their precision to one decimal place, and stored in the annotation package `phastCons100way.UCSC.hg19`. Because *phastCons* scores measure conservation, we sampled uniformly at random one thousand *phastCons* scores from differently conserved regions, concretely CDS and 3'UTR. These sampled scores are included in the `GenomicScores` package, to illustrate this comparison in the acompanying vignette. Interestingly, among the *phastCons* scores sampled from 1000 CDS positions, there are only 198 different values despite the apparently very high precision of some of them.

```
> origpcscoCDS <- readRDS(system.file("extdata", "origphastCons100wayhg19CDS.rds",
                                      package="GenomicScores"))
> origpcscoCDS
GRanges object with 1000 ranges and 1 metadata column:
         seqnames    ranges strand |               score
            <Rle> <IRanges>  <Rle> |           <numeric>
     [1]     chr1    976248      * |   0.763999998569489
     [2]     chr1   1886625      * | 0.00300000002607703
     [3]     chr1   3751636      * |                   1
     [4]     chr1   5950945      * |   0.236000001430511
     [5]     chr1   6638784      * |    0.99099999666214
     ...      ...       ...    ... .                 ...
   [996]     chrX 154113587      * |                   1
   [997]     chrX 154261706      * |                   1
   [998]     chrX 154305499      * |                   1
   [999]     chrY  15481226      * |                   1
  [1000]     chrY  21871402      * |   0.981000006198883
  -------
  seqinfo: 25 sequences from an unspecified genome
> length(unique(origpcscoCDS$score))
[1] 198
```

Similarly, in 3'UTR regions, only 209 unique *phastCons* scores are observed.

```
> origpcsco3UTRs <- readRDS(system.file("extdata", "origphastCons100wayhg193UTR.rds",
                                        package="GenomicScores"))
> origpcsco3UTRs
GRanges object with 1000 ranges and 1 metadata column:
         seqnames    ranges strand |               score
            <Rle> <IRanges>  <Rle> |           <numeric>
     [1]     chr1   1189930      * |                   0
     [2]     chr1   1227810      * |                   0
     [3]     chr1   1595390      * |  0.0149999996647239
     [4]     chr1   1595685      * |                   0
     [5]     chr1   2336044      * |                   0
     ...      ...       ...    ... .                 ...
   [996]     chrX 125684491      * |                   0
   [997]     chrX 129190576      * |                   1
   [998]     chrX 135299066      * |   0.196999996900558
   [999]     chrX 148560948      * | 0.00300000002607703
  [1000]     chrX 153289202      * |                   0
  -------
  seqinfo: 25 sequences from an unspecified genome
> length(table(origpcsco3UTRs$score))
[1] 209
```

In Supplementary Figure S1 we show a visual comparison between original and rounded *phastCons* scores. The two panels on top compare the whole range of scores observed in CDS (left) and 3'UTR (right) regions. The rounding effect can be better observed in the cumulative distributions shown in the panels at the bottom, again for CDS (left) and 3'UTR (right) regions.

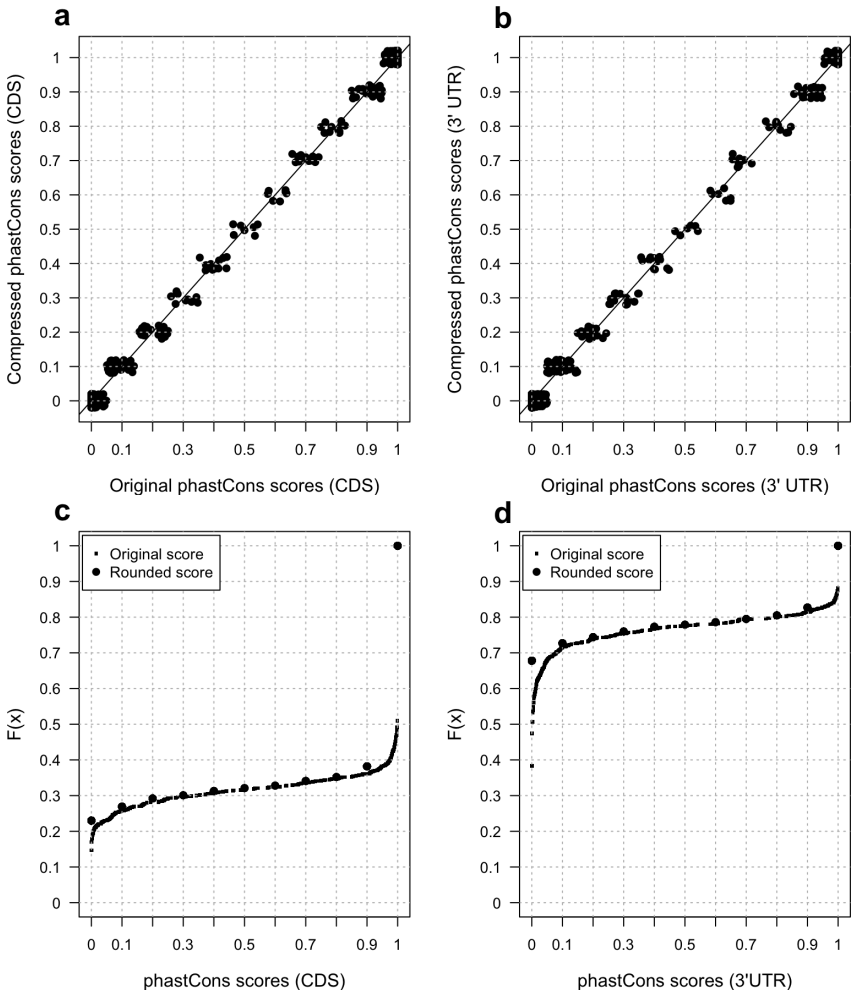

Figure S1: Original and lossy-compressed *phastCons* scores. Top panels (a, b): comparison of the distribution of values. Bottom panels (c, d): comparison of the cumulative distribution.

In these bottom panels, *phastcons* scores in CDS and 3'UTR regions display very different cumulative distributions. In CDS regions, most of the genomic scores ($> 60\%$) are found between the values of 0.9 and 1.0, while around 25% of the scores are found below 0.1. Indeed, these are the range of values where lossy compression loses more precison. The cumulative distribution of 3'UTR shows the same critical points, with the difference that most of scores are found below 0.1 ($> 70\%$).

The bottom plots in Supplementary Figure S1 also reveal that when the cumulative distribution is of interest, such as in the context of filtering genetic variants above or below certain threshold of conservation, the quantization of *phastCons* scores to one decimal place provides sufficient precision for a wide range of conservation values.

Table S1: Bioconductor annotation packages storing genomic scores and supported by the `GenomicScores` package. See `https://bioconductor.org/install` for instructions about installing Bioconductor packages. The column *Precision* indicates the number of decimal places (DP) or significant figures (SF) to which original genomic scores have been rounded.

Annotation Package	Description	Source	Precision
`fitCons.UCSC.hg19`	fitCons scores for the human genome (hg19).	(Gulko *et al.*, 2015)	2 DP
`MafDb.1Kgenomes.phase1.hs37d5` `MafDb.1Kgenomes.phase1.GRCh38`	MAF data from the 1000 Genomes Project Phase 1 for the human genome (GRCh37 & GRCh38).	(The 1000 Genomes Project Consortium, 2012)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`MafDb.1Kgenomes.phase3.hs37d5` `MafDb.1Kgenomes.phase3.GRCh38`	MAF data from the 1000 Genomes Project Phase 3 for the human genome (GRCh37 & GRCh38).	(The 1000 Genomes Project Consortium, 2015)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`MafDb.ESP6500SI.V2.SSA137.hs37d5` `MafDb.ESP6500SI.V2.SSA137.GRCh38`	MAF data from NHLBI ESP 6500 exomes for the human genome (GRCh37 & GRCh38).	(Tennessen *et al.*, 2012)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`MafDb.ExAC.r1.0.hs37d5` `MafDb.ExAC.r1.0.GRCh38`	MAF data from ExAC 60,706 exomes for the human genome (GRCh37 & GRCh38).	(Lek *et al.*, 2016)	MAF>0.1, 2 SF MAF≤ 0.1,1 SF
`MafDb.ExAC.r1.0.nonTCGA.hs37d5` `MafDb.ExAC.r1.0.nonTCGA.GRCh38`	MAF data from ExAC 53,105 nonTCGA exomes for the human genome (GRCh37 & GRCh38).	(Lek *et al.*, 2016)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`MafDb.gnomAD.r2.0.1.hs37d5` `MafDb.gnomAD.r2.0.1.GRCh38`	MAF data from gnomAD 15,496 genomes for the human genome (GRCh37 & GRCh38).	(Lek *et al.*, 2016)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`MafDb.gnomADex.r2.0.1.hs37d5` `MafDb.gnomADex.r2.0.1.GRCh38`	MAF data from gnomAD 123,136 exomes for the human genome (GRCh37 & GRCh38).	(Lek *et al.*, 2016)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`MafDb.TOPMed.freeze5.hg19` `MafDb.TOPMed.freeze5.hg38`	MAF data from NHLBI TOPMed 62,784 genomes for the human genome (hg19 & hg38).	(TOPMed Consortium, 2017)	MAF>0.1, 2 SF MAF≤ 0.1, 1 SF
`phastCons100way.UCSC.hg19`	phastCons scores derived from the alignment of the human genome (hg19) to other 99 vertebrate species.	(Siepel *et al.*, 2005)	1 DP (default) 2DP
`phastCons100way.UCSC.hg38`	phastCons scores derived from the alignment of the human genome (hg38) to other 99 vertebrate species.	(Siepel *et al.*, 2005)	1 DP (default) 2DP
`phastCons7way.UCSC.hg38`	phastCons scores derived from the alignment of the human genome (hg38) to other 6 vertebrate species.	(Siepel *et al.*, 2005)	1 DP (default) 2DP

Table S2: Bioconductor `AnnotationHub` resources storing genomic scores and supported by the `GenomicScores` package. See functions `availableGScores()` and `getGScores()` from the package to download these genomic score sets. The column *Precision* indicates the number of decimal places (DP) or significant figures (SF) to which original genomic scores have been rounded.

AnnotationHub Resource	Description	Source	Precision
`cadd.v1.3.hg19`	Combined Annotation Dependent Depletion, CADD scores: deleteriousness of single nucleotide variants in the human genome (hg19)	(Kircher *et al.*, 2014)	PHRED$\leq$ 50, 1 SF PHRED$>$ 50, set to 50
`fitCons.UCSC.hg19`	fitCons scores for the human genome (hg19).	(Gulko *et al.*, 2015)	2 DP
`linsight.UCSC.hg19`	linsight scores for the human genome (hg19).	(Huang *et al.*, 2017)	2 DP
`mcap.v1.0.hg19`	Mendelian Clinically Applicable Pathogenicity, M-CAP scores: pathogenicity classifier for rare missense variants in the human genome (hg19).	(Jagadeesh *et al.*, 2016)	2 DP
`phastCons100way.UCSC.hg19`	phastCons scores derived from the alignment of the human genome (hg19) to other 99 vertebrate species.	(Siepel *et al.*, 2005)	1 DP
`phastCons100way.UCSC.hg38`	phastCons scores derived from the alignment of the human genome (hg38) to other 99 vertebrate species.	(Siepel *et al.*, 2005)	1 DP
`phastCons7way.UCSC.hg38`	phastCons scores derived from the alignment of the human genome (hg38) to other 6 vertebrate species.	(Siepel *et al.*, 2005)	1 DP
`phastCons27way.UCSC.dm6`	phastCons scores derived from the alignment of the Drosophila melanogaster genome (dm6) to other 26 insect species.	(Siepel *et al.*, 2005)	1 DP
`phastCons60way.UCSC.mm10`	phastCons scores derived from the alignment of the mouse genome (mm10) to other 59 vertebrate species.	(Siepel *et al.*, 2005)	1 DP
`phyloP60way.UCSC.mm10`	Phylogenetic p-values (phyloP) conservation scores of the mouse genome (mm10) to other 59 vertebrate species.	(Pollard *et al.*, 2010)	PHRED$<$ 2, 1 DP to closest 0.5 PHRED$\geq$ 2, 1 DP PHRED$>$ 10 are set to 10
`phyloP100way.UCSC.hg19`	Phylogenetic p-values (phyloP) conservation scores of the human genome (hg19) to other 99 vertebrate species.	(Pollard *et al.*, 2010)	PHRED$<$ 2, 1 DP to closest 0.5 PHRED$\geq$ 2, 1 DP PHRED$>$ 10 are set to 10
`phyloP100way.UCSC.hg38`	Phylogenetic p-values (phyloP) conservation scores of the human genome (hg38) to other 99 vertebrate species.	(Pollard *et al.*, 2010)	PHRED$<$ 2, 1 DP to closest 0.5 PHRED$\geq$2, 1 DP PHRED$>$ 10 are set to 10

Table S3: Compression ratios for some representative genomic score sets. The source files `GenomicScores/inst/scripts/make-data_*.R` contain all the R instructions that transform the original genomic scores into lossy-compressed ones. The methods `qfun()` and `dqfun()` on a `GScores` object provide the quantization and dequantization functions; see reference manual and package vignette.

Score set and source URL	Original	Compressed	Ratio

`fitCons.UCSC.hg19`
`http://compgen.cshl.edu/`
`fitCons/0downloads/tracks/`
`V1.01/i6/scores/fc-i6-0.bw`
	76 Mb	25 Mb	≈ 3

`phyloP100way.UCSC.hg19`
`http://hgdownload.cse.ucsc.edu/`
`goldenPath/hg19/phyloP100way/`
`hg19.100way.phyloP100way`
	5.1 Gb	1.2 Gb	≈ 4

`phastCons100way.UCSC.hg19`
`http://hgdownload.cse.ucsc.edu/`
`goldenPath/hg19/phastCons100way/`
`hg19.100way.phastCons`
	2.5 Gb	233 Mb (1DP)
		774 Mb (2DP)
			$\approx$ 3-10

`mcap.v1.0.hg19`
`http://bejerano.stanford.edu/`
`MCAP/downloads/dat/mcap_v1_0.txt.gz`
	729 Mb	61 Mb	≈ 12

`cadd.v1.3.hg19`
`http://krishna.gs.washington.edu/`
`download/CADD/v1.3/whole_genome_SNVs.tsv.gz`
	80 Gb	716 Mb	≈ 114

References

Gulko, B. *et al.* (2015). Probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, **47**, 276–283.

Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**(4), 618.

Jagadeesh, K. A. *et al.* (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.*, **48**(12), 1581–1586.

Kircher, M. *et al.* (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**(3), 310–315.

Lek, M. *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**(7616), 285–291.

Pollard, K. S. *et al.* (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.

Siepel, A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Tennessen, J. A. *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, **526**, 68–74.

TOPMed Consortium (2017). Trans-omics for precision medicine (topmed). Allele frequency data accessed on Sep. 2017.