**#=== 01-readme.docx for Zoonomia downloads**

# pf sullivan, UNC/KI, 07/2022

# contents :: file descriptions & methods

# always check the md5sum

**#=== file location, view access**

https://drive.google.com/drive/folders/1d9u7VBa-bDdMNXzV5a1gEIKTtiDcAYnc?usp=sharing

**#=== files**

# all coordinates are hg38 / GRCh38, reference species is *Homo sapiens*

# bed files are a UCSC format for a genomic interval of the form chr \t start \t end

# where chr is chr1-chr22 chrX chrY, start is 0-based and end is 1-based.

# some files have a header row that are comments explaining file content

**### all\_mammals.genome.coverage.hg38.bed**

bed file of genomic intervals covered by Zoonomia alignment (i.e., constraint score was emitted)

chr1 10074 136896 # first 5 rows, no header

chr1 138896 208744

chr1 257607 298981

chr1 348433 370613

chr1 370640 445520

## compute total bases covered

library(data.table)

library(tidyverse)

fread("all\_mammals.genome.coverage.hg38.bed") %>%

rename(chr=V1, start0=V2, end=V3) %>%

summarise(sum(end-start0))

## 2,852,623,265 bases

**### phyloPam.fdr05.merge.hg38.bed.gz**

bed file of genomic intervals constrained in 240 mammals (all mammals phyloP score ≥ 2.270, 3.53% of the genome). md5sum 97ae6e71955a114d9d8810f17a3d3c50

chr1 11708 11709 # first 5 rows, no header

chr1 11930 11931

chr1 11935 11936

chr1 11940 11941

chr1 11975 11976

## compute total bases constrained in 240 mammals

fread("phyloPam.fdr05.merge.hg38.bed.gz") %>%

rename(chr=V1, start0=V2, end=V3) %>%

summarise(sum(end-start0))

## 100,651,377 bases

**### phastConsPr.high.merge.hg38.bed.gz**

bed file of genomic intervals constrained in 43 primates (phastCons primate score ≥ 0.961). md5sum 25fb159bb49cd9352063fdbfc4f515a4

chr1 12006 12038 # first 5 rows, no header

chr1 12050 12074

chr1 12179 12234

chr1 12291 12297

chr1 12392 12407

## compute total bases constrained in 43 primates

fread("phastConsPr.high.merge.hg38.bed.gz") %>%

rename(chr=V1, start0=V2, end=V3) %>%

summarise(sum(end-start0))

## 101,134,907 bases

**### the folder called: chr\_.data.v2.bed.gz/**

This folder contains data files containing per-base constraint information for each human chromosome (chr1-chr22, chrX, chrY). File sizes range from 260 mb to 2.6gb, total size of 24 chr\* files is 33 gb.

Each file has 2 header rows beginning with “##”. Columns 1-3 are hg38 position (chr, start0, end). Column 4 is the hg19 position (using UCSC hg38ToHg19 liftOver); if missing, liftOver failed or mapped to a different chromosome. Columns 5-7 are constraint scores: phyloP.am=phyloP for all 240 mammals, phastCons.am=phastCons score for all 240 mammals, phastCons.pr=phaseCons score for 43 primates. Nspecies are the number of species aligning to this position.

## Zoonomia constraint, pfs 12-2021, hg38

## chr start0 end hg19bp phyloP.am phastCons.am phastCons.pr Nspecies

chr9 10059 10060 10060 0.053 0.004 0.004 2

chr9 10060 10061 10061 0.064 0.007 0.007 2

chr9 10061 10062 10062 0.064 0.01 0.01 2

chr9 10062 10063 10063 0.064 0.012 0.012 2

### md5sum

3a40d0d6740ebb3521bc5fa977b81d27 chr1.data.v2.bed.gz

6753f2f59d9427aebae558eddd579a64 chr10.data.v2.bed.gz

539477de96217e5fdd4cef5faa87be80 chr11.data.v2.bed.gz

1fca23b80e548fd7706dcfdb48eb45ef chr12.data.v2.bed.gz

dfd454b0b5e91d0e263954922914ec47 chr13.data.v2.bed.gz

0c2179ff48b6579c994bc0bc7a9f9065 chr14.data.v2.bed.gz

32af7439114dc40ebf3e7f8cc156d7fd chr15.data.v2.bed.gz

840b754fff54b6585b77b142cb7a4dbb chr16.data.v2.bed.gz

473e00833f7685ea4a06dd0f8e7ebbe5 chr17.data.v2.bed.gz

f7aae39f0a3abee09366e1dd28d51083 chr18.data.v2.bed.gz

d683006063335bf867632a652bbb2bd4 chr19.data.v2.bed.gz

b1492a4e87262212a8b3d5b82156fb2b chr2.data.v2.bed.gz

b9dd6de78fac8ce494f2bdabbec1a082 chr20.data.v2.bed.gz

c13d57a11d28c83ed4ee99ddeab41eee chr21.data.v2.bed.gz

98440de15a1cb928a73f3f717fb153cd chr22.data.v2.bed.gz

a90bf6f6a68a9bafa5fcc3c98e909a81 chr3.data.v2.bed.gz

57af4665b6d317bb5edf64247abb0c6d chr4.data.v2.bed.gz

eddeaad8b88bf43922be71bc2ee3f946 chr5.data.v2.bed.gz

b69f28c80f91dc4730a6eb6fbf8a9d93 chr6.data.v2.bed.gz

1ee627f7fab2954669ca8c4ae04401b8 chr7.data.v2.bed.gz

2fa6a512a30074fe26a80736430a473c chr8.data.v2.bed.gz

9cb5647386c4d32ca817a018e20266f0 chr9.data.v2.bed.gz

cb14735ed1d26eeddaf901a273cdd7b3 chrX.data.v2.bed.gz

d61b596228e54d3ceed2046fb53aa87e chrY.data.v2.bed.gz

## ***Section 3: Deriving Mammalian and Primate Constraint Measures for the Human Genome***

By: Michael Dong, Ola Wallerman, Xue Li, Matthew J Christmas, Voichita D. Marinescu, Jennifer R. S. Meadows

In this section, we describe the generation of the key constraint metrics. *PhyloP scores* were computed for each human position to reflect constraint over mammalian evolution (240 mammal species), and *PhastCons scores* reflect constraint over the evolution of 43 primate species.

*Notes.* Full details of the rationale behind Zoonomia, justification of the species selected, the specific species, the sources of whole genome sequences, the builds used, and the reference-free alignment approach are given in a prior paper [(*11*)](https://paperpile.com/c/W0mHww/TXhY). As these data are extensive, they are not duplicated here. The Zoonomia alignment has 241 genome sequences from 240 mammalian species (*Canis lupus familiaris* was included twice, and handled as described below). In this section, “MAF” refers to Multiple Alignment Format. Human gene models were based on *GENCODE v36 (11/2020)* [(*12*)](https://paperpile.com/c/W0mHww/kzAR), and the genome build was *hg38/GRCh38* (unless otherwise noted).

**Neutral Model.** Three human neutral models (autosomes, chrX, and chrY) were generated using the HAL-format (HAL Tools v2.1) [(*13*)](https://paperpile.com/c/W0mHww/tXlr) of the Zoonomia alignment of 241 mammalian sequences [(*11*)](https://paperpile.com/c/W0mHww/TXhY). Ancestral repeats were identified by applying RepeatMasker Open-4.0. 2013-2015 (<http://www.repeatmasker.org>) to the ancestral sequence of the mammal alignment. Sequence from the penultimate ancestral branch (fullTreeAnc238) was used instead of the most ancestral sequence (fullTreeAnc239), as interspersed repeats on fullTreeAnc238 had better reconstruction of the eutherian ancestral form than fullTreeAnc239 (A. Smit, personal communication, 2 Feb 2021). Repeat coordinates were converted to fullTreeAnc239 using halLiftOver [(*13*)](https://paperpile.com/c/W0mHww/tXlr). Ancestral sequence repeat sets were filtered to exclude: non-mammalian repeats shared with birds and reptiles; repeat sequences annotated as structural RNA copies, satellites, tandem repeats, and low complexity annotations; and regions where synteny was not present across the four major branches of the mammalian tree (Xenarthran, Afrotherian, Laurasiatherian and Euarchontoglires). A representative random set of ancestral repeat positions (i.e., 100Kb) from the above list was the input for the neutral evolution model calculation. An ancestor-referenced multiple alignment format alignment was extracted (hal2maf). PhyloFit (with default parameters, --subst-mod REV --EM) was applied to the corrected tree to estimate branch lengths, fixing the tree topology. The resulting alignment was processed (mafDuplicateFilter) to remove sequences that aligned multiple times to the same region in order to avoid biases due to species over-alignment. These steps was repeated for the sex-specific models but using repeat positions from chrX or chrY of the human-referenced alignment (i.e., 100Kb). Primate neutral models were constructed in the same fashion, with the ancestral branch reconstruction based on the 43 primates present in the alignment. Synteny with five of the reconstructed branches in the primates tree were checked.

**PhyloP and phastCons constraint score calculation.** The alignment was preprocessed with mafDuplicateFilter (<https://github.com/dentearl/mafTools>), in order to keep one sequence per species (i.e., the best match) in each alignment block compared to the sequence to the consensus for that block [(*14*)](https://paperpile.com/c/W0mHww/oTiOV). For the primate-specific data, non-primate species were filtered out from the alignment using the mafSpeciesSubset command from mafTools. Alignment depth, ACGT ratios, and lists of species across the alignment were collected using the Bio.Align package from BioPython (<https://biopython.org/wiki/Multiple_Alignment_Format>) in a custom script. Total branch lengths for each alignment block in the multiple alignment format files were calculated using the tree\_doctor command with the branch length (--branchlen) option from the PHAST software package. Two constraint scores, phyloP and phastCons, were calculated using the PHAST package, (<https://github.com/CshlSiepelLab/phast>). PhyloP (v1.5) was used to calculate per-base constraint and acceleration p-values [(*15*)](https://paperpile.com/c/W0mHww/K3ME8). Scores were presented as –log p-values under null hypothesis of neutral evolution, where computation involved performing a likelihood ratio test at each alignment column (--method LRT) and with the output of constraint and acceleration scores (--mode CONACC). PhyloP scores calculated on the human-referenced, MAF-formatted, duplicate-filtered alignment range from -20 to 9.28. Scores from the human-referenced, primates-only alignment (43-way) range from -20 to 1.26. Higher scores reflect greater deviation from the neutral models and hence greater constraint.

Scores calculated on the primates have lower ranges, as the total branch lengths of clade-specific trees are relatively lower compared to the entire mammalian tree. PhastCons (Phast v1.5), implemented a phylogenetic hidden Markov model (phylo-HMM) to identify evolutionarily constrained elements [(*16*, *17*)](https://paperpile.com/c/W0mHww/waaPe+79LvZ). In contrast to phyloP single base constraint, phastCons metrics incorporate the columns of flanking bases. Model parameters were selected to match those used to generate the phastCons output for the UCSC 100 Species Vertebrate Multiz Alignment (expected-length=45, target-coverage=0.3, and rho=0.31). The phastCons, a per-base constraint score, was outputted.

**Constraint score thresholds.** We applied two metrics of constraint for 2.85 billion bases in the human genome, phyloP base scores across the evolutionary history of mammals (240 species, range -20 to 9.28, higher scores indicating greater constraint) and windowed phastCons base scores across the evolutionary history of primates (43 species, range 0 to 1, higher scores indicating greater constraint). Both metrics compared the observed phylogeny to a neutral model. The phastCons metric focuses on bases which may have diverged over mammalian history but remained under constraint over the shorter primate branch lengths.

A mammalian phyloP threshold was determined by converting absolute phyloP scores into p-values and then to q-values using a false discovery rate (FDR) correction [(*18*)](https://paperpile.com/c/W0mHww/vrzCl) (R function *qvalue*). Outputs were classified as constrained or accelerated based on the sign of the score, any column with a resulting q-value ≤ 0.05 was deemed significantly constrained or accelerated (5% FDR 240 mammal phyloP constraint score ≥ 2.27, 3.53% of the human genome). We took a different approach for primate constraint given the fewer species (43) and lesser branch lengths available; a phyloP score across primates species would be underpowered to discriminate highly constrained bases from the background. We identified the phastCons threshold that yielded a similar fraction of the genome under constraint as for all mammals (i.e., phastCons base score ≥ 0.96, 3.54% of the human genome). Given that the 43 primates are a subset of 240 mammals, the locations of significantly constrained bases in mammals and primates overlap.

We note that we validated the mammalian phyloP threshold, the use of different scores to measure constraint in mammals and primates, and the base pair resolution of mammalian phyloP scores by using heritability analyses of human diseases and complex traits.

**Fraction constrained of human genome.** We estimated the lower bound for the fraction of sites under purifying selection across the genome (πhat) by comparing the empirical cumulative distribution function (ECDF) of phyloP scores across the genome to the ECDF of ancestral repeats, following the same method detailed in [(*19*)](https://paperpile.com/c/W0mHww/9mCj) where:

πhat = 1 - mins F(s)/G(s)

Here, F is the ECDF for all sites across the genome (a function of scores s) and G is the ECDF for ancestral repeats (i.e., neutrally evolving sites). We extracted coordinates of ancestral repeats from the UCSC repeatmasker track and filtered these to retain a set of ancestral repeats present in four clades of the mammalian tree (Xenarthran, Afrotherian, Laurasiatherian, and Euarchontoglires). The ECDFs of phyloP values for all positions across the genome, as well as those only in the set of ancestral repeats, were calculated using the ecdf function in R v.4.0.4. PhyloP values below -1.5 were excluded so that the extreme left tail of the ECDFs did not heavily influence estimates of constraint. We estimated πhat for human- and primate-centered phyloP sets.

***References:***

11.  [Zoonomia Consortium, A comparative genomics multitool for scientific discovery and conservation. *Nature*. **587**, 240–245 (2020).](http://paperpile.com/b/W0mHww/TXhY)

12.  [A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, P. Flicek, GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).](http://paperpile.com/b/W0mHww/kzAR)

13.  [G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*. **29**, 1341–1342 (2013).](http://paperpile.com/b/W0mHww/tXlr)

14.  [D. Earl, N. Nguyen, G. Hickey, R. S. Harris, S. Fitzgerald, K. Beal, I. Seledtsov, V. Molodtsov, B. J. Raney, H. Clawson, J. Kim, C. Kemena, J.-M. Chang, I. Erb, A. Poliakov, M. Hou, J. Herrero, W. J. Kent, V. Solovyev, A. E. Darling, J. Ma, C. Notredame, M. Brudno, I. Dubchak, D. Haussler, B. Paten, Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.* **24**, 2077–2089 (2014).](http://paperpile.com/b/W0mHww/oTiOV)

15.  [A. Siepel, K. S. Pollard, D. Haussler, in *Research in Computational Molecular Biology* (Springer, Berlin, Heidelberg, 2006), pp. 190–205.](http://paperpile.com/b/W0mHww/K3ME8)

16.  [J. Felsenstein, G. A. Churchill, A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**, 93–104 (1996).](http://paperpile.com/b/W0mHww/waaPe)

17.  [A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).](http://paperpile.com/b/W0mHww/79LvZ)

18.  [J. D. Storey, A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).](http://paperpile.com/b/W0mHww/vrzCl)

19.  [K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).](http://paperpile.com/b/W0mHww/9mCj)