Assessing genome conservation on pangenome graphs with PanSel

Matthias Zytnicki

April 2024

1 Abstract

Motivation: With more and more telomere-to-telomere genomes assembled, pangenomes make it possible to capture the genomic diversity of a species. Because they introduce less biases, pangenomes, represented as graphs, tend to supplant the usual linear representation of a reference genome, augmented with variations. However, this major change requires new tools adapted to this data structure. Among the numerous questions that can be addressed to a pangenome graph is the search for conserved regions, *i.e.* genomic that are likely retained during evolution.
Results: In this article, we present a new tool, named PanSel, which finds genomic regions that are significantly conserved, or divergent.
Availability: PanSel, written in C++11 with no dependency, is available at `https://github.com/mzytnicki/pansel`.
Contact: matthias.zytnicki@inrae.fr

2 Introduction

Genome diversity has long been studied for genetic diseases, genome improvement of species of agricultural interest, among others. It is possible to discover, and characterize, this diversity using a reference genome, and the sequencing of other individuals, usually with short reads. While the method provides excellent results with short variations, such as single nucleotide variations, they hardly can detect longer variations, such as structural variants.

With the advent of new, long, almost errorless, reads, it is now possible to produce several telomere-to-telomere haplotypes for complex eukaryotes. These haplotypes, if they include the diversity of the population of interest (usually, a species), may be compared, leading to pangenomic studies [1]. A pangenome can be efficiently stored into a dedicated data structure, which is usually called a variation graph [6]. This graph contain so-called *segments*, which are sequence chunks found in at least one haplotype, and *paths*, which are lists of segments (possibly in reverse-complement) representing haplotypes. A given segment, if

it belongs to several paths, represents a homologous sequence, shared by the haplotypes.

Although pangenomic studies have been published for almost two decades in prokaryotes, they are relatively new for multi-cellular eukaryotes, and changing the framework from a genome supplemented with variations to a variation graph is no simple task. Notably, it requires adapting all the tools which were developed for the previous framework, and they are numerous. Mapping reads to a genome, for instance, is the first step of many analyses, which include genome-wide association studies, transcriptome quantification, or protein binding assessment. Read mapping methods have been extended, so that is now possible to map reads directly to graphs with Giraffe [13] for short reads, or GraphAligner [11] for long reads, for instance. Reports consistently show that these new methods are less biased towards the reference genome.

Other methods include the conservation of genomic sequence, *i.e.* sequences that evolve slower (or faster) than the expected natural drift. Such methods, made for linear genomes, include phastCons [12], GERP and GERP++ [2], and phyloP [10]. These methods rely on two types of information: a multiple sequence alignment, and a phylogeny. Briefly, they estimate the length of the branches of a phylogenetic tree at each position of the genome alignment. The metrics then can be use to detect conserved, or rapidly evolving, regions.

Here, we would like to extend these ideas to the variation graph. However, it requires adaptation, at several levels. First, the phylogenetic tree is usually not known, or even not defined, between individuals of the same species, because inbreeding may be frequent. Second, the base-level multiple alignment —which tentatively puts in the same column the nucleotides that evolved from the common, ancestor nucleotide— is not modeled in the variation graph; finding a conservation value for each nucleotide is thus not directly applicable.

In PanSel, we adapt the idea of the quantification of the genomic conservation to variation graphs, using simple ideas. First, we use sliding windows of fixed size on the reference haplotype. Second, we estimate the average number of edit distance between each haplotype on each window. Last, significantly under- and over-conserved windows are detected by fitting a mixture model.

3 Implementation

PanSel in a C++ tool that reads a GFA file representing one chromosome. Given a window size s, it tries to detect conserved segments (hereafter name *boundary segments*) distant by s nucleotides on the reference path. For each pair of boundary segment, it extracts the sub-path for each haplotype. It estimates the average edit distance between each pair of sub-paths. Considering two paths, we approximate the edit distance by the sum of the sizes of the segments which belong in one path, and not the other.

The distribution of the edit distance is then fitted with a simple mixture model. The most conserved part follows a negative binomial (fitted using the R package fitdistrplus [3]), which can be considered as an extension of the Pois-

son distribution with an over-dispersion, which is common in genome biology. In the conserved regions, we suppose that the number of editions is roughly uniform throughout the genome, and each window can be considered as an observation of a Poisson distribution. However, we observe a heavy tail on the right-hand side of the distribution, which represents divergent sequences, and is not compatible with a negative binomial distribution. We modeled it using a log-normal distribution, and finally fitted the whole distribution with the average of the two previously found distributions. A more detailed description of the implementation, and the fit, is given in Supplementary Section 1.

In our computer (Debian x86_64 6.1.76-1), PanSel took less than 25 minutes, and not more than 20GB RAM, for each chromosome.

4 Results

We applied PanSel with three sliding window size (1000, 10,000, and 100,000) on the Human Pangenome [9]. We took the 2022-03-11 release, computed by MiniGraph-Cactus [7].

4.1 Comparison with other annotations

First, we discarded the regions that overlapped with known gaps (see Supplementary Section). We then divided the regions into 8 categories, with the most conserved regions being in the first bin, and the most divergent regions being in the last bin. For the sake of space, we will only present results for a window size of 1kb, but results for other windows sizes follow the same trend, and are presented in Supplementary Section 3.1.

We compared each bin with several available annotations. First, we counted the number of structural variants that overlapped with each bin. As seen in Figure 1 (first row), the number of variants increases with the bin number, and the final bin overlaps with significantly more known variants. This confirms that our method correlates with known data.

We computed the average conservation score, provided by PhyloP [10] using 100 vertebrates, on each bin. The second row of Figure 1 show that the conservation decreases with the bin number. This confirms that genomic regions that are more conserved in human are also more conserved in vertebrates.

We compared the bins with the location of known protein coding exons, provided by GenCode [5], and computed the number of exonic nucleotides found in each bin. As seen in the third row of Figure 1, conserved regions contain more coding exons, as expected.

Last, we extracted the annotation computed by ChromHMM [4]. Briefly, ChromHMM splits the genome into 100 categories [14], based on ChIP-Seq data. These categories are associated with different states of the chromatin (*e.g.* "actively transcribed", or "Tx"). Each state may have different levels (*e.g.* there are eight "Tx" states), and we merged these levels in this study. Here, we focused on the "Tx" and the "HET" (for heterochromatin) states, but

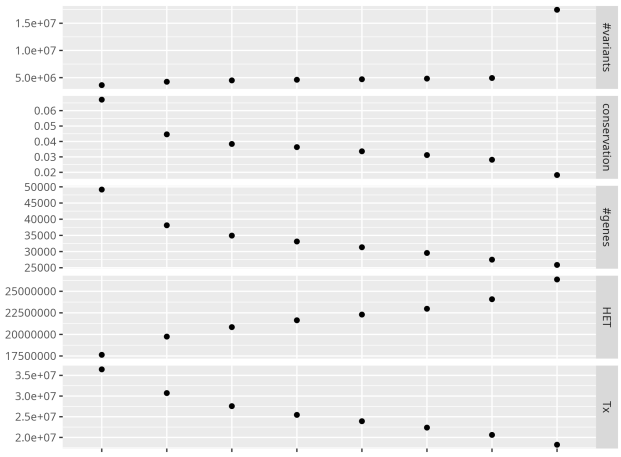


Figure 1: Each column represents a bin. The bins on the left include the most conserved regions, whereas the ones on the right include the most divergent regions. For each bin, we computed the number of overlapping structural variants (first row), the average vertebrate conservation (second row), the exonic coverage (third row), the predicted heterochromatin coverage (fourth row), and the predicted actively transcribed coverage (last row).

all the results can by found in Supplementary Section 3.2. As seen in fourth row of Figure 1, heterochromatin-rich regions are more divergent. The last row shows that actively transcribed regions are more conserved, as expected.

4.2 Gene set enrichment

We selected the regions that were significantly conserved and divergent (p-value of 5%) for a window size of 10,000 (a window size of 1,000 is usually too small to capture the full extent of a gene). We found about 150Mbp and 470Mbp of conserved and divergent regions respectively. We then extracted the overlapping genes (6666 and 9503 genes in conserved and divergent regions respectively), and looked for enriched gene ontologies using gProfiler [8].

Full results are provided in Supplementary Section 3.3. Briefly, conserved regions are related to very general molecular functions (*e.g.* protein binding), biological processes (*e.g.* multicellular organism development) or cellular compartment (*e.g.* cytoplasm).

Divergent regions are related to more specific molecular functions (*e.g.* antigen binding), biological processes (*e.g.* adaptive immune response) or cellular compartment (*e.g.* immunoglobulin complex), which are related to immunity.

This would suggest that basic molecular functions are located in conserved regions, whereas immunity related genes are located in divergent regions, which is expected.

5 Competing interests

No competing interest is declared.

6 Data availability

The pipe-line used to produce the results can be retrieved from `https://github.com/mzytnicki/pansel_paper`.

References

[1] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 10 2016.

[2] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Computational Biology*, 6(12):e1001025, December 2010.

[3] Marie Laure Delignette-Muller and Christophe Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4), 2015.

[4] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, February 2012.

[5] Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Carles Boix, Silvia Carbonell Sala, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Kevin L Howe, Toby Hunt, Osagie G Izuogu, Rory Johnson, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Ferriol Calvet Riera, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Maxim Y Wolf, Jinuri Xu, Yucheng T Yang, Andrew Yates, Daniel Zerbino, Yan Zhang, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Michael L Tress, and Paul Flicek. Gencode 2021. *Nucleic Acids Research*, 49(D1):D916–D923, December 2020.

[6] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello,

Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, October 2018.

[7] Glenn Hickey, Jean Monlong, Jana Ebler, Adam M. Novak, Jordan M. Eizenga, Yan Gao, Haley J. Abel, Lucinda L. Antonacci-Fulton, Mobin Asri, Gunjan Baid, Carl A. Baker, Anastasiya Belyaeva, Konstantinos Billis, Guillaume Bourque, Silvia Buonaiuto, Andrew Carroll, Mark J. P. Chaisson, Pi-Chuan Chang, Xian H. Chang, Haoyu Cheng, Justin Chu, Sarah Cody, Vincenza Colonna, Daniel E. Cook, Robert M. Cook-Deegan, Omar E. Cornejo, Mark Diekhans, Daniel Doerr, Peter Ebert, Jana Ebler, Evan E. Eichler, Susan Fairley, Olivier Fedrigo, Adam L. Felsenfeld, Xiaowen Feng, Christian Fischer, Paul Flicek, Giulio Formenti, Adam Frankish, Robert S. Fulton, Shilpa Garg, Erik Garrison, Nanibaa' A. Garrison, Carlos Garcia Giron, Richard E. Green, Cristian Groza, Andrea Guarracino, Leanne Haggerty, Ira M. Hall, William T. Harvey, Marina Haukness, David Haussler, Simon Heumos, Kendra Hoekzema, Thibaut Hourlier, Kerstin Howe, Miten Jain, Erich D. Jarvis, Hanlee P. Ji, Eimear E. Kenny, Barbara A. Koenig, Alexey Kolesnikov, Jan O. Korbel, Jennifer Kordosky, Sergey Koren, HoJoon Lee, Alexandra P. Lewis, Wen-Wei Liao, Shuangjia Lu, Tsung-Yu Lu, Julian K. Lucas, Hugo Magalhães, Santiago Marco-Sola, Pierre Marijon, Charles Markello, Tobias Marschall, Fergal J. Martin, Ann McCartney, Jennifer McDaniel, Karen H. Miga, Matthew W. Mitchell, Jacquelyn Mountcastle, Katherine M. Munson, Moses Njagi Mwaniki, Maria Nattestad, Sergey Nurk, Hugh E. Olsen, Nathan D. Olson, Trevor Pesout, Adam M. Phillippy, Alice B. Popejoy, David Porubsky, Pjotr Prins, Daniela Puiu, Mikko Rautiainen, Allison A. Regier, Arang Rhie, Samuel Sacco, Ashley D. Sanders, Valerie A. Schneider, Baergen I. Schultz, Kishwar Shafin, Jonas A. Sibbesen, Jouni Sirén, Michael W. Smith, Heidi J. Sofia, Ahmad N. Abou Tayoun, Françoise Thibaud-Nissen, Chad Tomlinson, Francesca Floriana Tricomi, Flavia Villani, Mitchell R. Vollger, Justin Wagner, Brian Walenz, Ting Wang, Jonathan M. D. Wood, Aleksey V. Zimin, Justin M. Zook, Tobias Marschall, Heng Li, and Benedict Paten. Pangenome graph construction from genome alignments with minigraph-cactus. *Nature Biotechnology*, May 2023.

[8] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, May 2023.

[9] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, Kerstin Howe, Miten Jain, Tsung-Yu Lu, Charles

Markello, Fergal J. Martin, Matthew W. Mitchell, Katherine M. Munson, Moses Njagi Mwaniki, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, David Porubsky, Pjotr Prins, Jonas A. Sibbesen, Jouni Sirén, Chad Tomlinson, Flavia Villani, Mitchell R. Vollger, Lucinda L. Antonacci-Fulton, Gunjan Baid, Carl A. Baker, Anastasiya Belyaeva, Konstantinos Billis, Andrew Carroll, Pi-Chuan Chang, Sarah Cody, Daniel E. Cook, Robert M. Cook-Deegan, Omar E. Cornejo, Mark Diekhans, Peter Ebert, Susan Fairley, Olivier Fedrigo, Adam L. Felsenfeld, Giulio Formenti, Adam Frankish, Yan Gao, Nanibaa' A. Garrison, Carlos Garcia Giron, Richard E. Green, Leanne Haggerty, Kendra Hoekzema, Thibaut Hourlier, Hanlee P. Ji, Eimear E. Kenny, Barbara A. Koenig, Alexey Kolesnikov, Jan O. Korbel, Jennifer Kordosky, Sergey Koren, HoJoon Lee, Alexandra P. Lewis, Hugo Magalhães, Santiago Marco-Sola, Pierre Marijon, Ann McCartney, Jennifer McDaniel, Jacquelyn Mountcastle, Maria Nattestad, Sergey Nurk, Nathan D. Olson, Alice B. Popejoy, Daniela Puiu, Mikko Rautiainen, Allison A. Regier, Arang Rhie, Samuel Sacco, Ashley D. Sanders, Valerie A. Schneider, Baergen I. Schultz, Kishwar Shafin, Michael W. Smith, Heidi J. Sofia, Ahmad N. Abou Tayoun, Françoise Thibaud-Nissen, Francesca Floriana Tricomi, Justin Wagner, Brian Walenz, Jonathan M. D. Wood, Aleksey V. Zimin, Guillaume Bourque, Mark J. P. Chaisson, Paul Flicek, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, David Haussler, Ting Wang, Erich D. Jarvis, Karen H. Miga, Erik Garrison, Tobias Marschall, Ira M. Hall, Heng Li, and Benedict Paten. A draft human pangenome reference. *Nature*, 617(7960):312–324, May 2023.

[10] Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, October 2009.

[11] Mikko Rautiainen and Tobias Marschall. Graphaligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1), September 2020.

[12] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, July 2005.

[13] Jouni Sirén, Jean Monlong, Xian Chang, Adam M. Novak, Jordan M. Eizenga, Charles Markello, Jonas A. Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, Namrata Gupta, Stacey Gabriel, Thomas W. Blackwell, Aakrosh Ratan, Kent D. Taylor, Stephen S. Rich, Jerome I. Rotter, David Haussler, Erik Garrison, and Benedict Paten. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574), December 2021.

[14] Ha Vu and Jason Ernst. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biology*, 23(1), January 2022.

A Description of the method

A.1 GFA parsing

PanSel reads the `S` and `P` lines of the GFA file. All the other lines are skipped. It stores the lengths of the segments (but not the nucleotidic sequences), and the paths (as lists of segments).

A.2 Conserved segments

In this step, PanSel finds all the segments that included in almost all paths. These segments represented conserved sequences.

Ideally, these segments should be used by all paths. However, in some GFA files, variations graphs have been enriched with additional variations (using long or short reads). These variations can be stored using `P` lines, but do not represent haplotypes.

The number of haplotypes can be provided by the user.

Alternatively, PanSel can detect it. For each segment, PanSel counts the number of paths that include it. It computes a distribution, with the number of paths passing through a node on the x-axis, and the number of such nodes on the y-axis.

In our experiments, we found that the right-most local maximum of the previous distribution accurately finds the number of haplotypes. We compute this local maximum, m.

The *conserved* segments that are included in at least m path are then stored.

A.3 Sliding windows

PanSel then follows the path of the reference haplotype (provided by the user). It scans each segment, which can be represented as a genomic interval $[a, b]$ on the reference haplotype. It first stops at the conserved segment, A representing interval $[a, b]$. This is the first boundary segment. From there, it tries to find the next boundary segment, B, such that the distance between A and B is approximately w, the size of the user-defined sliding window. More formally, we are looking for, B, representing interval $[c, d]$, such that $c - b \leq w \leq d - a$.

There may be no such boundary segment: the position of the ending sliding window may contain variations. In this case, we look for the next boundary segment, and record that the size of the sliding window will be larger than expected.

A.4 Computing the edit distance

We extract the sub-paths of each haplotype between each pair of selected, consecutive boundary segment. The names of the segments are then stored for each sub-path.

Here, we store only once each segment for each sub-path. If there is a loop, it is thus discarded.

The segments are then sorted (by order of apparition in the GFA file) for each sub-path. We thus also loose any structure in the paths, and the structural variations may be lost. However, we can then consider a sub-path as a set of segments. This simplifies the next step.

We then compute, for each pair of sub-paths, an approximation of the edit distance. To do so, we find the symmetric difference of the two sub-paths, which is a set of segments. The edit distance is the sum of the sizes of these segments.

Although an approximation, it works well in practice, because loops and inversions are rare in the genome.

For each sliding window, we compute the average edit distance between all pairs of sub-paths.

A.5 Fitting the average edit distance

We first compute the distribution of the average number of editions per region. We compute the mode m, and extract the region corresponding to $x \in [0, 2m]$. We censor the rest of the distribution, and fit a negative negative binomial using the fitdistcens function. We also fit the whole distribution with a log-normal, and plot the average of the two distributions.

In order to find the significantly conserved (resp. divergent) regions, we take the extreme 5% of the negative binomial (resp. log-normal) distribution. The threshold can be modified by the user.

The fit are plotted in figures 2, 3, and 4.

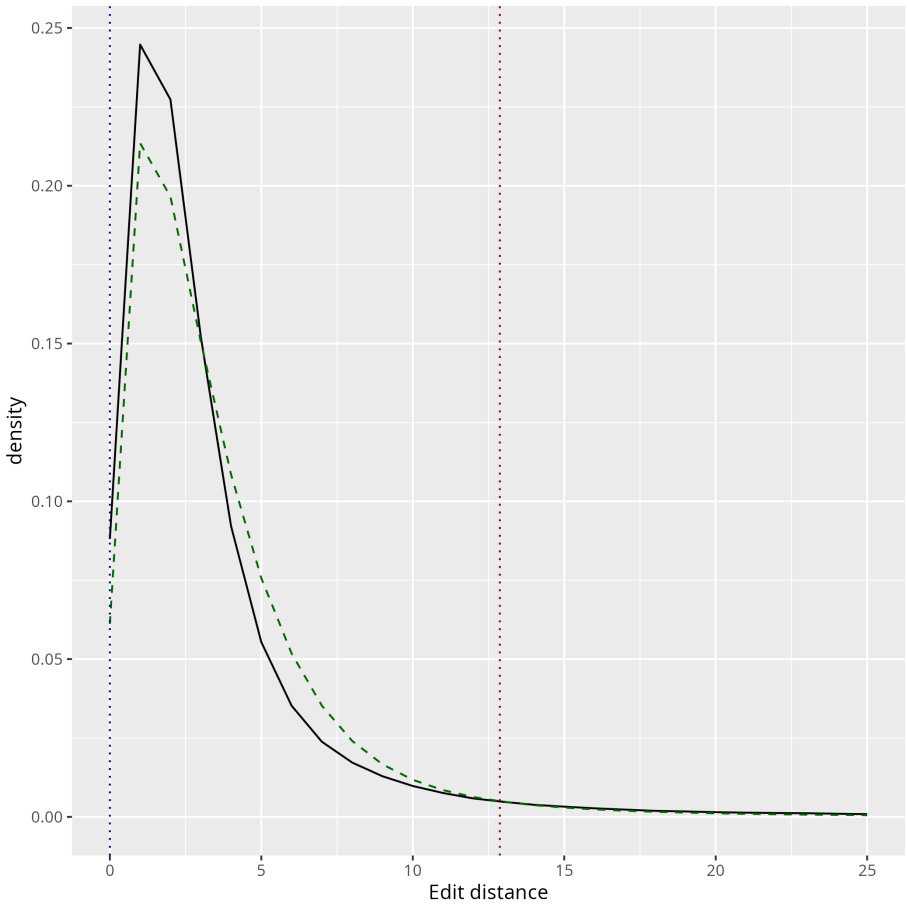


Figure 2: Distribution of the average edit distance, in solid black line, for a window size of 1000. The fit is in dashed green, the threshold for significantly conserved (resp. divergent) regions is in dotted blue (resp. red).

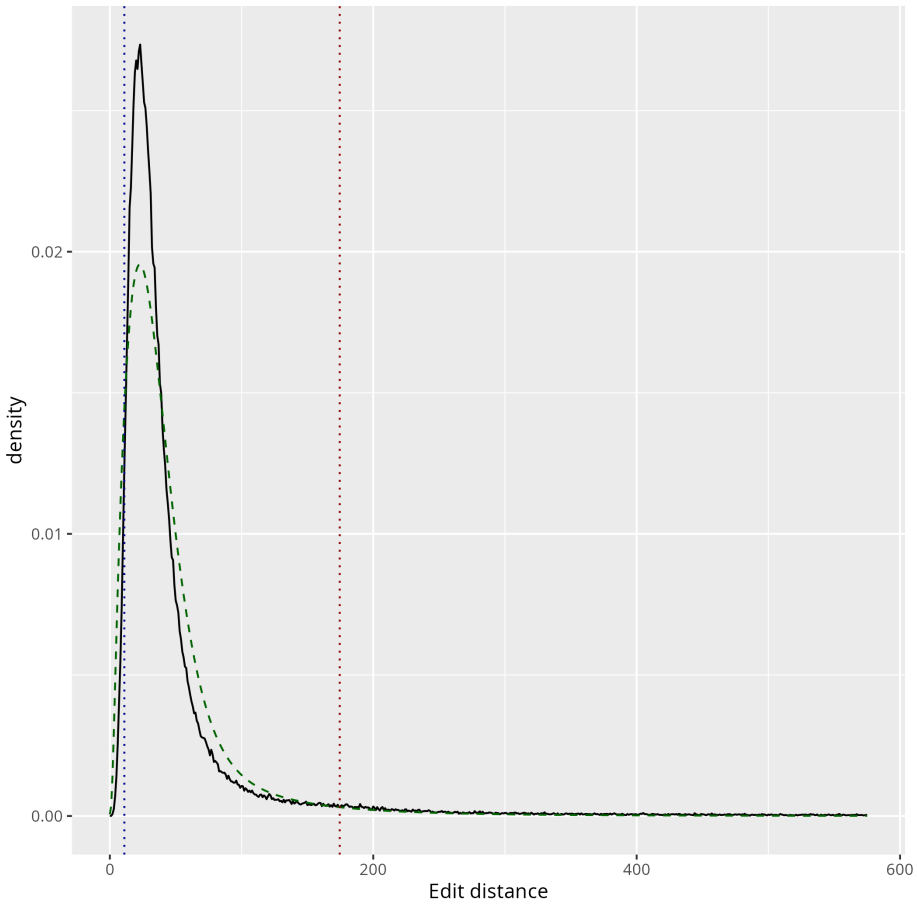


Figure 3: Distribution of the average edit distance, in solid black line, for a window size of 10,000.

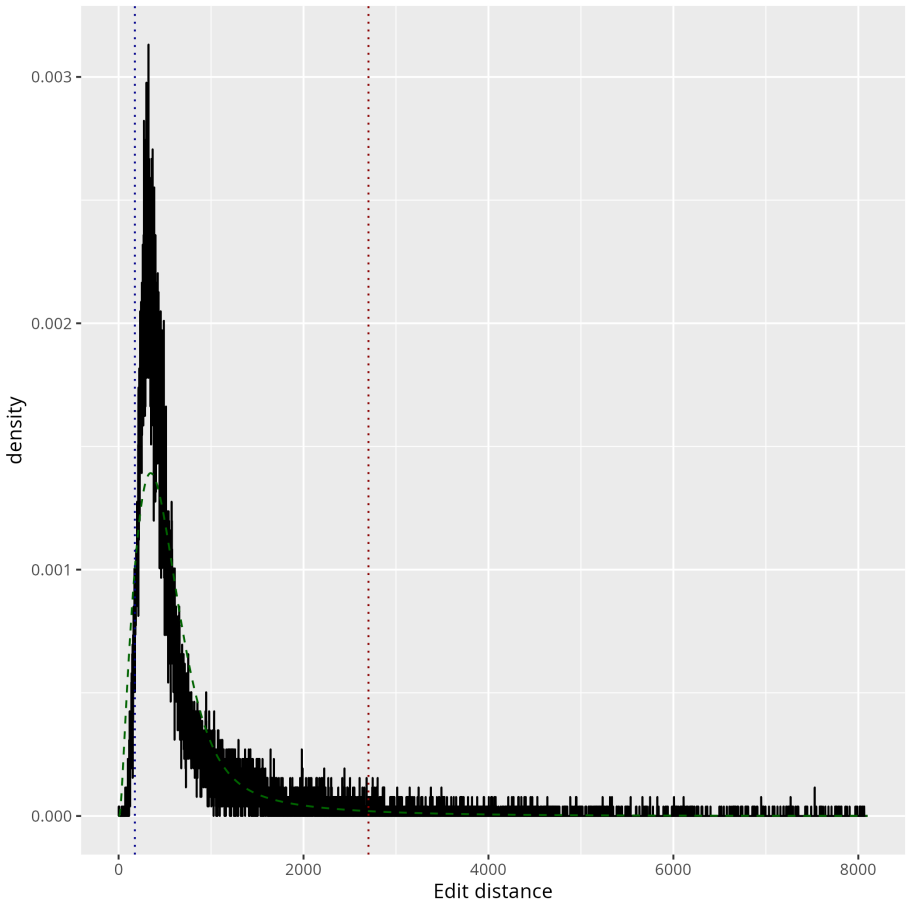


Figure 4: Distribution of the average edit distance, in solid black line, for a window size of 100,000.

B Data used

B.1 Pangenome graphs

The pangenome graphs were downloaded from the Human Pangenome initiative:
`https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/scratch/`

We chose the `2022_03_11_minigraph_cactus` release, and downloaded one `vg` file per chromosome.

Each file was then converted to GFA format using `vg view`.

We chose GRCh38 as a reference, since many annotation are available.

B.2 Genome annotations

- Assembly gaps:
  `https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/latest/hg38.agp.gz`
- List of structural variants from dbVar:
  `https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_assembly/GRCh38/vcf/GRCh38.variant_call.all.vcf.gz`
- 100 vertebrate conservation using PhyloP:
  `https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phyloP100way/hg38.phyloP100way.bw`
- Genome annotation:
  `https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_44/gencode.v44.annotation.gtf.gz`
- ChromHMM annotation:
  `https://public.hoffman2.idre.ucla.edu/ernst/2K9RS/full_stack/full_stack_annotation_public_release/hg38/hg38_genome_100_segments.bed.gz`

C Results

C.1 Scores for the other window sizes

This part gives the same results as Figure 1 in the main document, with different window sizes.

C.1.1 Window size of 10k

For 10k window size, results are very similar to 1k size.

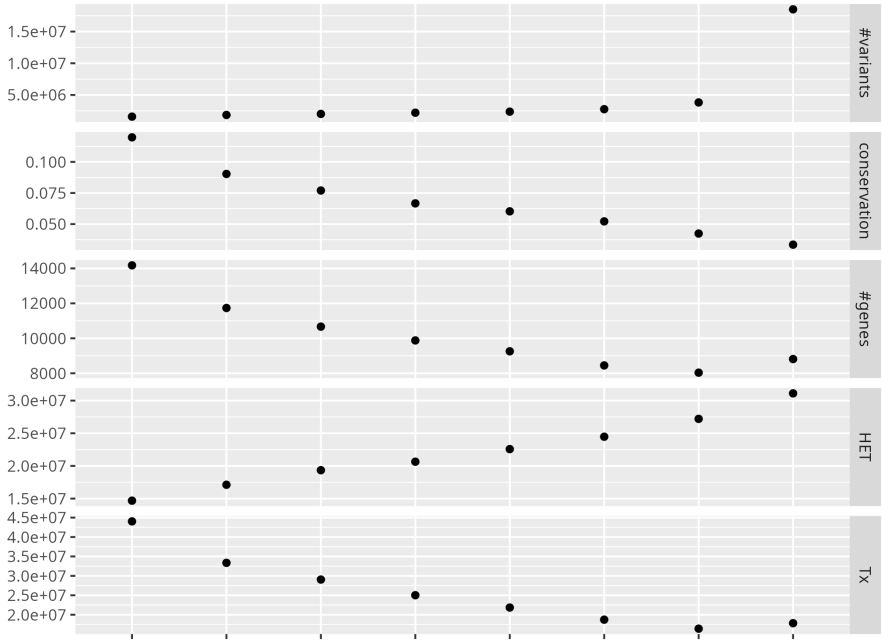


Figure 5: Results for window size of 10k. See Figure 1 in the main document for a detailed legend.

C.1.2 Window size of 100k

At 100k, the signal is less clear, probably because windows are too large, and include very diverse types of regions.

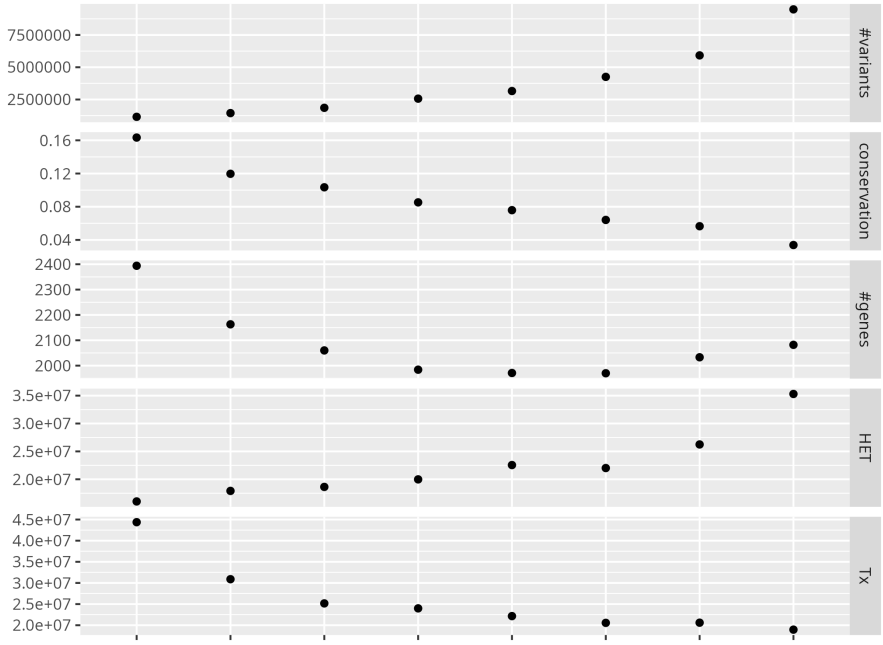


Figure 6: Results for window size of 100k. See Figure 1 in the main document for a detailed legend.

C.2 ChromHMM full results

This part gives the coverage for all ChromHMM categories. The categories are:

- Acet: acetylations
- BivProm: bivalent promoters
- DNAse: DNase I hypersensitivity
- EnhA: active enhancers
- EnhWk: weak enhancers
- GapArt: assembly gaps and alignment artifacts
- HET: heterochromatin
- PromF: Flanking promoters
- Quies: repressive or inactive
- ReprPC: polycomb repressed
- TSS: transcription start site
- Tx: strong transcription
- TxEnh: transcribed enhancers
- TxEx: transcript exons
- TxWk: weak transcription
- ZNF: zinc finger

C.2.1 Window size of 1k

Figure 7 clearly shows that repressed states, including HET, Quies, and ReprPC, tend to co-localize with divergent regions. It is also the case for Acet states, which is a weak mark of promoter or enhancers.

On the other hand, active states, including BivProm, EnhWk, PromF, TSS, TxEnh, TxEx, TxWe, and ZNF, tend to co-localize with conserved regions.

The DNase state does not seem to follow any clear trend.

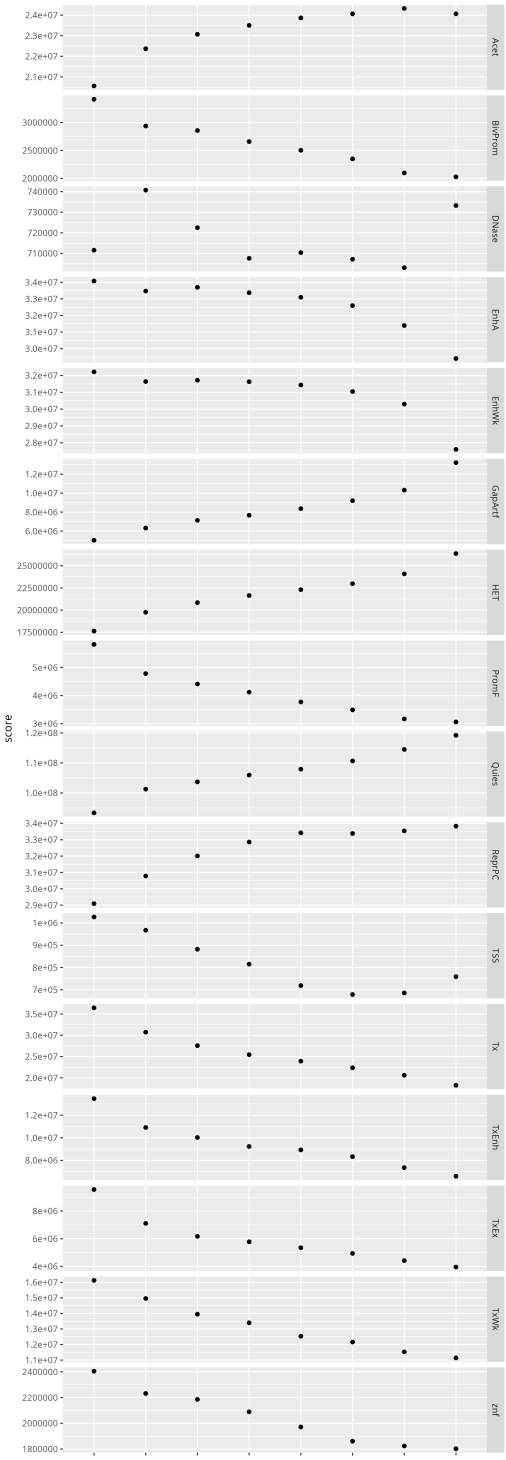


Figure 7: Results for window size of 10k.

C.2.2 Window size of 10k

Results for a window size 10k are similar to the previous ones. The correlation, however, seems slightly less clear, probably because larger windows include different types of regions.

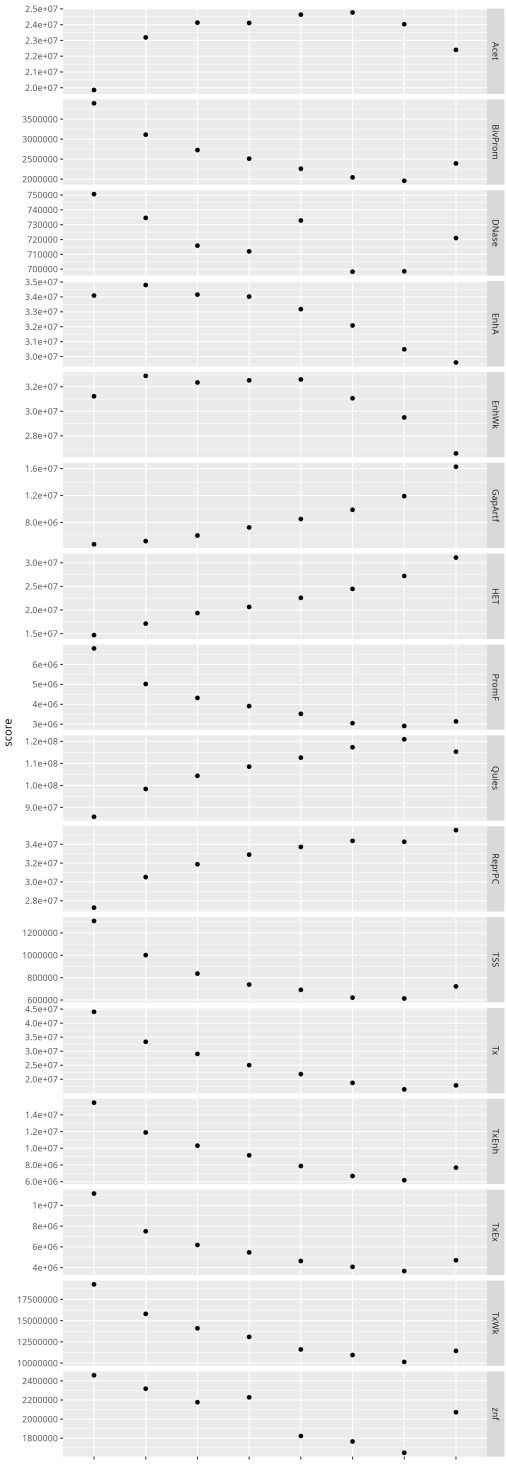


Figure 8: Results for window size of 10k.

C.2.3 Window size of 100k

Results for a window size 100k are similar to the previous ones, but even less clear.

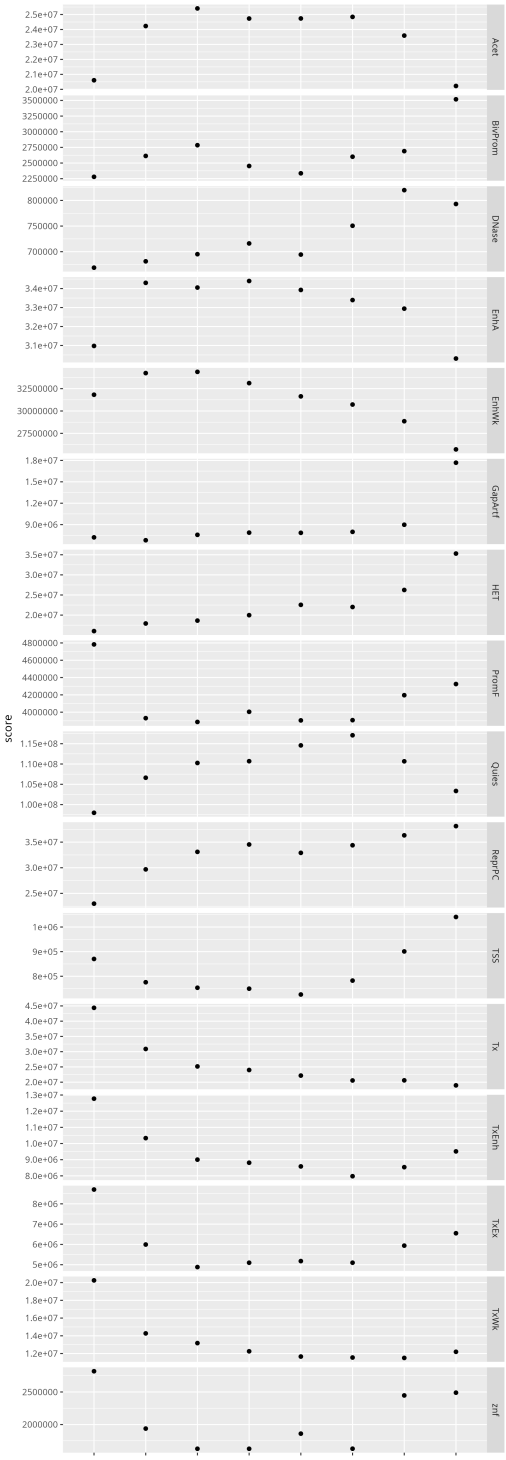
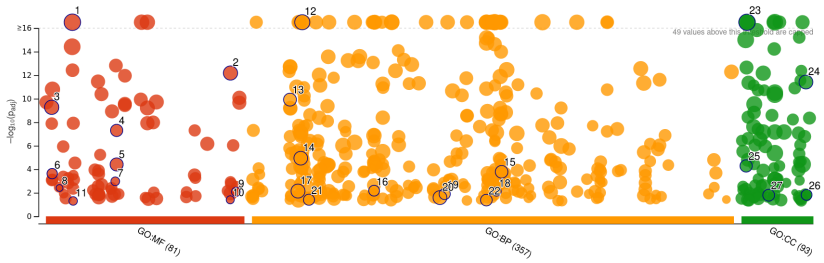


Figure 9: Results for window size of 100k.

C.3 Gene set enrichment

First gene set enrichment results for conserved regions are given in Figure 10. The first figure shows significantly enriched gene sets in the molecular function (MF), biological process (BP), and cellular compartment (CC) ontologies. The y-axis give the p-value.

Enriched gene sets are very vague, and shared in most species.



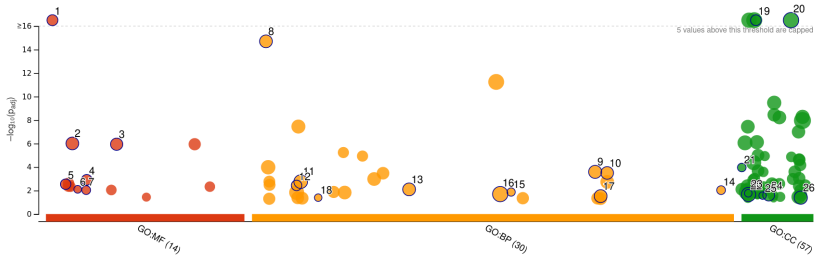
ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005515	protein binding	9.395×10^{-57}
2	GO:MF	GO:0140096	catalytic activity, acting on a protein	6.609×10^{-13}
3	GO:MF	GO:0003677	DNA binding	5.409×10^{-10}
4	GO:MF	GO:0030695	GTPase regulator activity	4.989×10^{-8}
5	GO:MF	GO:0030674	protein-macromolecule adaptor activity	3.903×10^{-5}
6	GO:MF	GO:0003730	mRNA 3'-UTR binding	2.355×10^{-4}
7	GO:MF	GO:0030371	translation repressor activity	1.057×10^{-3}
8	GO:MF	GO:0004415	hyaluronoglucosaminidase activity	3.916×10^{-3}
9	GO:MF	GO:0140938	histone H3 methyltransferase activity	8.347×10^{-3}
10	GO:MF	GO:0140036	ubiquitin-dependent protein binding	3.964×10^{-2}
11	GO:MF	GO:0008066	glutamate receptor activity	4.958×10^{-2}
12	GO:BP	GO:0007275	multicellular organism development	8.575×10^{-91}
13	GO:BP	GO:0006325	chromatin organization	1.204×10^{-10}
14	GO:BP	GO:0007155	cell adhesion	1.148×10^{-5}
15	GO:BP	GO:0051301	cell division	1.646×10^{-4}
16	GO:BP	GO:0030010	establishment of cell polarity	6.634×10^{-2}
17	GO:BP	GO:0006915	apoptotic process	7.149×10^{-3}
18	GO:BP	GO:0050919	negative chemotaxis	7.792×10^{-3}
19	GO:BP	GO:0043087	regulation of GTPase activity	1.270×10^{-2}
20	GO:BP	GO:0042592	homeostatic process	2.577×10^{-2}
21	GO:BP	GO:0008361	regulation of cell size	3.900×10^{-2}
22	GO:BP	GO:0048511	rhythmic process	4.207×10^{-2}
23	GO:CC	GO:0005737	cytoplasm	1.876×10^{-53}
24	GO:CC	GO:1902494	catalytic complex	3.854×10^{-12}
25	GO:CC	GO:0005667	transcription regulator complex	5.225×10^{-5}
26	GO:CC	GO:1902911	protein kinase complex	1.507×10^{-2}
27	GO:CC	GO:0034702	monoatomic ion channel complex	1.596×10^{-2}

version	e111_eg58_p18_30541362
date	3/22/2024, 6:10:00 PM
organism	hsapiens

g:Profiler

Figure 10: Results for window size of 100k.

First gene set enrichment results for divergent regions are given in Figure 10. Here, some genes related to immunity can be found in all three ontologies.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0003823	antigen binding	3.516×10^{-26}
2	GO:MF	GO:0005509	calcium ion binding	9.764×10^{-7}
3	GO:MF	GO:0030695	GTPase regulator activity	1.139×10^{-5}
4	GO:MF	GO:0015085	calcium ion transmembrane transporter activity	1.228×10^{-3}
5	GO:MF	GO:0004843	cysteine-type deubiquitinase activity	2.877×10^{-3}
6	GO:MF	GO:0008559	ABC-type xenobiotic transporter activity	7.878×10^{-3}
7	GO:MF	GO:0015020	glucuronosyltransferase activity	9.168×10^{-3}
8	GO:BP	GO:0002250	adaptive immune response	1.989×10^{-15}
9	GO:BP	GO:0098657	import into cell	2.569×10^{-4}
10	GO:BP	GO:0120035	regulation of plasma membrane bounded cell pr…	3.155×10^{-4}
11	GO:BP	GO:0007155	cell adhesion	1.739×10^{-3}
12	GO:BP	GO:0006805	xenobiotic metabolic process	3.661×10^{-3}
13	GO:BP	GO:0034330	cell junction organization	7.939×10^{-3}
14	GO:BP	GO:2000463	positive regulation of excitatory postsynaptic po…	9.254×10^{-3}
15	GO:BP	GO:0052695	cellular glucuronidation	1.343×10^{-2}
16	GO:BP	GO:0051179	localization	2.003×10^{-2}
17	GO:BP	GO:0099536	synaptic signaling	2.969×10^{-2}
18	GO:BP	GO:0009812	flavonoid metabolic process	4.015×10^{-2}
19	GO:CC	GO:0019814	immunoglobulin complex	1.535×10^{-50}
20	GO:CC	GO:0071944	cell periphery	5.387×10^{-38}
21	GO:CC	GO:0000137	Golgi cis cisterna	1.093×10^{-4}
22	GO:CC	GO:0005833	hemoglobin complex	1.696×10^{-4}
23	GO:CC	GO:0005856	cytoskeleton	2.168×10^{-2}
24	GO:CC	GO:0034702	monoatomic ion channel complex	2.439×10^{-2}
25	GO:CC	GO:0031838	haptoglobin-hemoglobin complex	2.706×10^{-2}
26	GO:CC	GO:0099081	supramolecular polymer	4.028×10^{-2}

version	e111_eg58_p18_30541362
date	3/22/2024, 6:10:28 PM
organism	hsapiens

g:Profiler

Figure 11: Results for window size of 100k.