Remus: a web application for prioritization of regulatory variants in monogenic diseases

Paweł Sztromwasser[1*], Damian Skrzypczak[1,3], Arkadiusz Michalak[1], Wojciech Fendler[1,2]

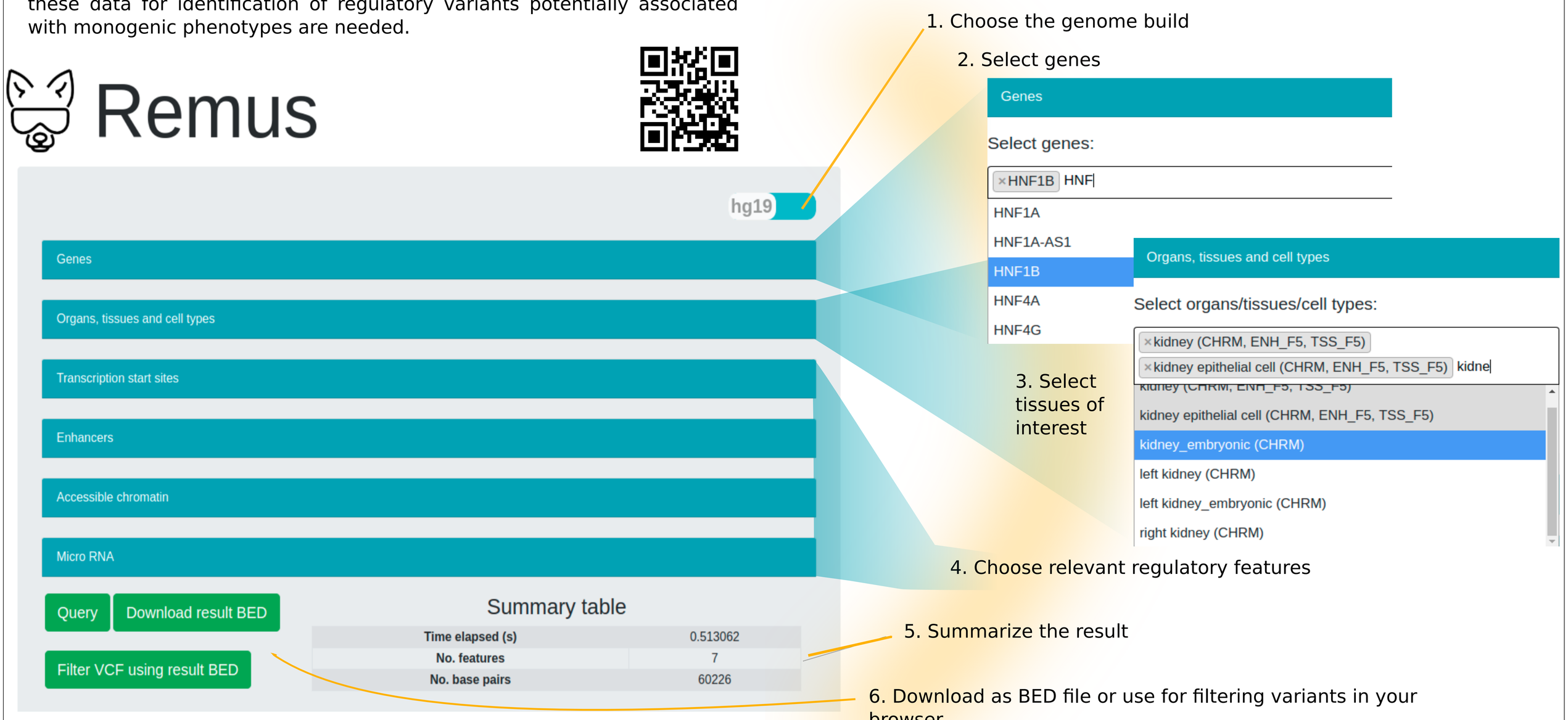
1. Department of Biostatistics and Translational Medicine, Medical University of Lodz, Łódź, Poland
2. Department of Radiation Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA
3. Wroclaw University of Enviromental and Life Sciences, Wrocław, Poland
(*) pawel.sztromwasser@umed.lodz.pl

Background

Whole-genome testing offers unparalleled diagnostic rates for monogenic disorders[1,2], at the same time providing an opportunity to interrogate variants in regulatory elements[3]. However, due to the vast size of the regulome, great number of variants, and difficulty in predicting their phenotypic impact, searching pathogenic variants in the regulatory regions remains challenging. Rich regulatory data produced by global initiatives such as ENCODE[4] and FANTOM[5] allowed to map tissue-specific regulatory features onto the human genome, opening new opportunities for studying the regulome, also in monogenic diseases. New tools facilitating use of these data for identification of regulatory variants potentially associated with monogenic phenotypes are needed.

Results

Remus is a Web application that facilitates identification of regulatory regions potentially associated with the expression of input genes. Starting from a small set of input genes implicated in pathogenesis of a disease, Remus allows creating a list of regulatory features active in a set of tissues of interest. Customizable search and step-by-step process allows for iterative building of a list of coordinates representing genomic locations of elements that likely play a role in regulating expression of the input genes in the user-selected tissues. The coordinates can subsequently be used for in-browser filtering of variants, making it suitable for sensitive data. Remus is available at http://remus.btm.umed.pl



	Cell-types (mean size [Mb])	Tissues (mean size [Mb])
Enhancers	81 (4.6)	61 (15.3)
Transcription start sites	92 (0.76)	53 (0.76)
Accessible chromatin	59 (69.9)	100 (87.9)

Coordinate liftover of datasets available only for single genome build, provided extra 17.2% enhancer regions, and 31.9% more open chromatin regions, on average per tissue. The open chromatin regions allow to restrict 450,000 experimentaly verified miRNA-gene interactions to average 7.5 and 9% interactions relevant to user-selected cell-type and tissue, respectively.

To evaluate Remus in searching pathogenic mutations, we used a set of distal regulatory variants reported causative for 23 distinct monogenic disorders[3], and a manually curated list of tissues affected by these disorders. 19 out of 23 enhancers and promoters located 100bp-120kb away from the regulated genes appeared among the candidate regions identified by Remus. Using an example of a correctly identified *BLK* enhancer located 19.7kb upstream of the gene[9], tissue-agnostic screening +/- 50kb of the gene's transcript resulted in 43.6kb enhancer regions, in contrast with 5.8kb (13.4%) yielded by Remus as active in the pancreas. Genome-wide, analysis focused on a single tissue (or cell-type) confines considered enhancer regions to average 14.8 (or 6.9%) of the total.

Methods

Coordinates of tissue-specific regulatory regions were extracted from ENCODE[4] and FANTOM5[6] repositories. Regions originating from the same tissues and cell-types were merged and collapsed. Datasets available in hg19 and GRCh38 coordinate system were converted to GRCh38 and hg19 respectively, and merged-in. Regions were categorized into promoters, enhancers, and accessible chromatin, and provided for user operations in the application. Information about accessible chromatin loci is additionally used to filter non-tissue-specific interactions, such as microRNA - transcript interactions from miRTarBase[7] and miRWalk[8]. The connection between genes and regulatory features is made based on experimental evidence of interaction or user-tunable distance in the genome. Variant filtering is implemented in JavaScript and happens entirely in user's browser.

Conclusion

Remus facilitates identification of regulatory regions potentially associated with a monogenic disease and can supplement analysis of coding variation with an aim of improving diagnostic yield of whole-genome sequencing. Although the analysis is currently limited to phenotypes affecting available tissues and celltypes, with time, continously expanding repertoire of tissue-specific regulatory features will enable studying a wider spectrum of rare disorders.

References

1. Gilissen, C. et al. Nature, vol. 511, no. 7509, pp. 344–347, 2014
2. Lionel, A.C. et al. Genet. Med., vol. 20, no. 4, pp. 435–443, 2018
3. Smedley D. et al. AJHG, vol. 99, no. 3, pp. 595–606, 2016
4. The ENCODE Project Consortium, Nature, vol. 489, pp. 57–74, 2012
5. The FANTOM Consortium, Nature, vol. 507, pp. 462–70, 2014
6. Ienasescu, H. et al. Database, vol. 2016, p. baw144, 2016
7. Chou, C.H. et al. Nucl. Acids Res., vol. 46, pp. D296–D302, 2018
8. Sticht, C. et al. PLoS One, vol. 13, no. 10, p. e0206239, 2018
9. Borowiec, M. et al. PNAS, vol. 106, no. 34, pp. 14460–5, Aug. 2009

Acknowledgements

Project funded by NCN Polonez grant no 2016/23/P/NZ2/04251. The project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665778.