ZfaParallel: an improved R package to locate the optimal testing region for rare variant association tests using shared-memory parallel computing

Yexian Zhang[1], Chaorong Chen[2], Maggie Haitian Wang[3,4], Shuai Liu[1], Meiyu Duan[1], Lan Huang[1], and Fengfeng Zhou[1,*].

[1]BioKnow Health Informatics Lab, College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012,

[2]BioKnow Health Informatics Lab, College of Software, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012,

[3]Division of Biostatistics and Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, N.T, Hong Kong SAR and

[4]CUHK Shenzhen Research Institute, Shenzhen, China.

*To whom correspondence should be addressed.

Supplementary materials

An illustrative dataset to evaluate ZfaParallel

The exome-wide case-control genotyping dataset of the Han Chinese population of psoriasis was chosen to evaluate the ZfaParallel (Yang, et al., 2019). This dataset with the accession number GSE131670 was downloaded from the Gene Expression Omnibus (GEO) dataset. Its sampled cohort consisted of 12,398 Han Chinese participants, among which there were 6,489 cases and 5,909 controls. Each sample was hybridized using the microarray platform (Illumina Human Exome Fine BeadChip, GPL26699) and has 270,241 genomic variants. This study excluded those variants with minor allele frequency (MAF)>1% or call rate <95% from further analysis (Wang, et al., 2017). After the quality control step, there were 68,208 variants remaining for further analysis.

Case study

ZfaParallel with W-test was applied on all the human autosomal chromosomes with the initial window size 128. A genomic region was considered as being psoriasis-associated if its combined *p*-value is smaller than 0.01 after the Bonferroni correction. The rare variants were annotated with the software ANNOVAR (Wang, et al., 2010) based on the database RefSeq (Pruitt, et al., 2014). ZfaParallel detected a genomic region of two rare variants (exm-rs1269854 and exm534463) and its psoriasis-association statistical significance *p*-value=3.01e-18. These two variants were within the gene Tenascin XB (*TNXB*), and the genomic region of gene *TNXB* on chromosome 6 carried 25 variants with gene-based psoriasis-associated significance *p-value*=3.13e-3.

The gene *TNXB* plays important roles in organizing and maintaining structures of tissues that support skin, muscles, and regulating production of certain types of collagen (https://ghr.nlm.nih.gov/gene/TNXB#resources). Mutations in *TNXB* is causal to a recessive form of the Ehlers-Danlos syndrome, a type of skin disease characterized by collagen deficiency (Demirdas, et al., 2017; Schalkwijk, et al., 2001). Through the proposed algorithm, we identified a novel gene that is highly susceptible to psoriasis with relevant functions in the physiology of skin disorder. The two biomarker variants may be worthy of further experimental validations.

Conclusions

We implemented a parallel version ZfaParallel of the R package ZFA, which was designed for detecting a genomic region with multiple rare variants and a significant phenotype-association. The efficient C++ implementation and the shared-memory parallel computing of ZfaParallel improved the running speed of ZFA by over 37 times on the human autosomal chromosomes. The heuristic rules of ZfaParallel to further refine the rare variants in a genomic region detected rare variant combinations with much better phenotype-association significances.

References

Demirdas, S., *et al.* Recognizing the tenascin-X deficient type of Ehlers-Danlos syndrome: a cross-sectional study in 17 patients. *Clin Genet* 2017;91(3):411-425.

Pruitt, K.D., *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;42(Database issue):D756-763.

Schalkwijk, J., *et al.* A recessive form of the Ehlers-Danlos syndrome caused by tenascin-X deficiency. *N Engl J Med* 2001;345(16):1167-1175.

Wang, K., Li, M. and Hakonarson, H. ANNOVAR: functional annotation of genetic

variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38(16):e164.

Wang, M.H., *et al.* A Zoom-Focus algorithm (ZFA) to locate the optimal testing region for rare variant association tests. *Bioinformatics* 2017;33(15):2330-2336.

Yang, C., *et al.* Exome-Wide Rare Loss-of-Function Variant Enrichment Study of 21,347 Han Chinese Individuals Identifies Four Susceptibility Genes for Psoriasis. *J Invest Dermatol* 2019.