Performance evaluation of variant calling methods with or without pedigree information

Kaname Kojima
kojima@megabank.tohoku.ac.jp
Naoki Nariai
nariai@megabank.tohoku.ac.jp
Takahiro Mimori
mimori@megabank.tohoku.ac.jp
Mamoru Takahashi
takahashi@megabank.tohoku.ac.jp
Yumi Yamaguchi-Kabata
yamaguchi@megabank.tohoku.ac.jp
Yukuto Sato
yuksato@megabank.tohoku.ac.jp
Masao Nagasaki
nagasaki@megabank.tohoku.ac.jp

Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-machi, Aoba-ku, Sendai-shi, Miyagi 980-8573, Japan

Keywords: next generation sequencing, variant call, phase informative read, pedigree information

Variant detection from genome-wide sequencing data is essential for the analysis of disease-causing mutations and elucidation of disease mechanisms. However, variant calling in low coverage regions is difficult due to sequence read errors and mapping errors. Hence, variant calling approaches that are robust to low coverage data are demanded. To address such an issue, several methods considering pedigree information have been proposed. Here, we compare the performance of variant calling methods considering pedigree information and methods without it. For variant calling methods with pedigree information, we use Pedigree Caller [3], Trio Caller [1], and PolyMutt [4]: Pedigree Caller considers pedigree information and phase-informative reads, TrioCaller considers pedigree information and linkage disequilibrium between variants, and PolyMutt considers pedigree information and also estimates *de novo* mutations. For variant calling methods without pedigree information, we use GATK Unified Genotyper [2] and BCFtools for SAMtools output [5]. In performance evaluation, we use two types of sequencing data for parent-offspring trios with several read coverages: one is synthetically generated sequencing data for a parent-offspring trio based on variant calling results in the 1000 Genomes Project [6] and the other is real sequencing data for a HapMap parent-offspring trio.

References

[1] Chen, W., Li, B., Zeng, Z., Sanna, S., Sidore, C., Busonero, F., Kang, H. M., Li, Y., and Abecasis, G. R., Genotype calling and haplotyping in parent-offspring trios, *Genome Research*, 23(1):142–151, 2013.

[2] DePristo, M. A., Banks, E., Poplin, R. Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A, del Angel, G. Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and, Daly, M. J., A framework for variation discovery and genotyping using next generation DNA sequencing data, *Nature Genetics*, 43:491–498, 2011.

[3] Kojima, K., Nariai, N., Mimori, T., Takahashi, M., Yamaguchi-Kabata, Y., Sato, Y., and Nagasaki, M., A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads, *Bioinformatics*, accepted, 2013.

[4] Li, B., Chen, W., Zhan, X., Busonero, F., and Sanna, S., Sidore, C., Cucca, F., Kang, M. H., and Abecasis, G. R., A likelihood based framework for variant calling and de novo mutation detection in families, *PLoS Genetics*, 8(10), 2012.

[5] Li, H., Ruan, J., and Durbin, R., Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome Research*, 18(11):1851–1858, 2008.

[6] 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., A map of human genome variation from population-scale sequencing, *Nature*, 467(7319):1061-1073, 2010.