# SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data

Zhi Wei[1,*], Wei Wang[1], Pingzhao Hu[2], Gholson J. Lyon[3], Hakon Hakonarson[3]

[1]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
[2]The Centre for Applied Genomics (TCAG) and Program in Genetics and Genome Biology, the Hospital for Sick Children, Toronto, ON, Canada
[3]Center for Applied Genomics, the Children's Hospital of Philadelphia, and Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, USA

*To whom correspondence should be addressed. Tel: +1 973 642 4497; Fax: +1 973 596 5777; Email: zhiwei@njit.edu

## Supplementary Material

Sup Table 1 FDR for SNVer and the Fisher's exact test employed by CRISP.

| Nominal FDR level = 0.1 | | | | | | | Nominal FDR level = 0.05 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAF: 0.05 - 0.5 | | MAF: 0.01 - 0.05 | | MAF: 0.001 - 0.01 | | | MAF: 0.05 - 0.5 | | MAF: 0.01 - 0.05 | | MAF: 0.001 - 0.01 | |
| N | SNVer | CRISP | SNVer | CRISP | SNVer | CRISP | N | SNVer | CRISP | SNVer | CRISP | SNVer | CRISP |
| 50 | 0.013 | 0.400 | 0.013 | 0.242 | 0.035 | 0.142 | 50 | 0.007 | 0.395 | 0.010 | 0.230 | 0.015 | 0.120 |
| 100 | 0.010 | 0.453 | 0.016 | 0.312 | 0.032 | 0.146 | 100 | 0.005 | 0.451 | 0.014 | 0.302 | 0.019 | 0.127 |
| 150 | 0.010 | 0.469 | 0.013 | 0.353 | 0.024 | 0.159 | 150 | 0.005 | 0.467 | 0.009 | 0.343 | 0.012 | 0.139 |
| 200 | 0.009 | 0.476 | 0.014 | 0.382 | 0.022 | 0.159 | 200 | 0.005 | 0.475 | 0.005 | 0.373 | 0.003 | 0.137 |
| 250 | 0.008 | 0.480 | 0.010 | 0.400 | 0.020 | 0.168 | 250 | 0.003 | 0.478 | 0.005 | 0.391 | 0.006 | 0.145 |
| 375 | 0.006 | 0.484 | 0.011 | 0.424 | 0.012 | 0.183 | 375 | 0.002 | 0.482 | 0.006 | 0.415 | 0.004 | 0.161 |
| 500 | 0.007 | 0.486 | 0.011 | 0.437 | 0.016 | 0.198 | 500 | 0.004 | 0.483 | 0.004 | 0.424 | 0.007 | 0.173 |
| 750 | 0.006 | 0.487 | 0.010 | 0.444 | 0.019 | 0.212 | 750 | 0.002 | 0.484 | 0.005 | 0.431 | 0.008 | 0.179 |
| 1000 | 0.004 | 0.488 | 0.008 | 0.446 | 0.014 | 0.220 | 1000 | 0.002 | 0.485 | 0.004 | 0.432 | 0.007 | 0.180 |
| 1500 | 0.004 | 0.488 | 0.007 | 0.448 | 0.015 | 0.220 | 1500 | 0.001 | 0.485 | 0.002 | 0.433 | 0.007 | 0.179 |
| 2000 | 0.004 | 0.488 | 0.006 | 0.450 | 0.015 | 0.228 | 2000 | 0.002 | 0.484 | 0.004 | 0.437 | 0.007 | 0.178 |

GATK:

Based on the latest recommendations from the authors of GATK (http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v2), we removed variant calls based on having any of the following criteria: 1) SNVs within clusters (3 SNVs within 10 bp of each other); (2) more than four reads with mapping quality of zero (MQ0) and more than 10% of reads with mapping quality of zero; (3) strand bias (SB) higher than or equal to -1.0; (4) SNV quality score less than 30; (5) quality-by-depth (QD) score less than 5.0; (6) largest Contiguous Homopolymer Run of Variant Allele (HRun) more than 5; or (7) SNVs around a potential indel. Finally, we removed all variants with depth coverage less than 6 so that we have comparable numbers of variant as SNVer.