

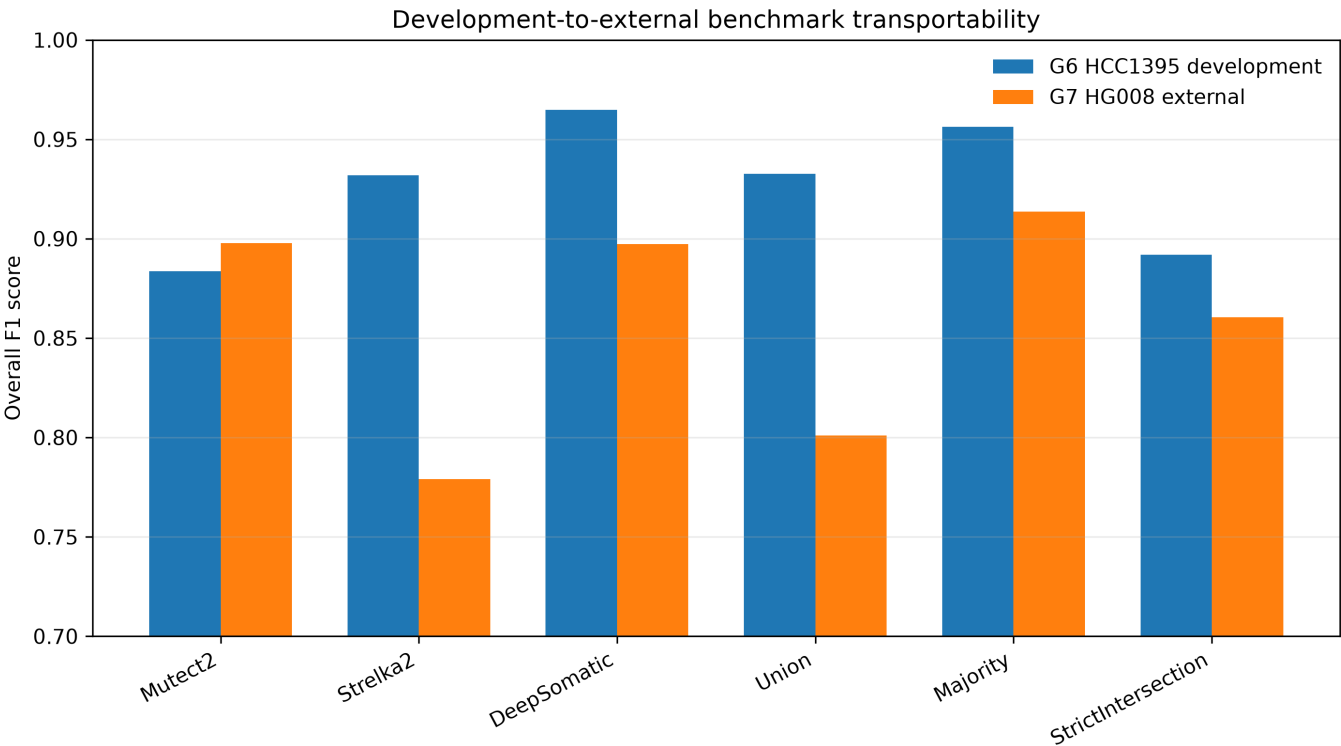
# Supplementary Information

## Prospectively frozen external benchmarking reveals benchmark-context dependence in somatic variant calling

Md Mahmudul Hasan, PhD | AmbioSpire LLC, Ohio, USA

This supplement contains only results and records already approved in the locked P003/T001 scientific publication release v1.0.0. The bounded caller-support, INDEL-length, germline-context, within50bp and all-region-bootstrap analyses were prespecified before HG008 access and executed after completion of the primary external evaluation. The within50bp analysis remains prespecified exploratory. The HCC1395-to-HG008 comparison remains post-hoc descriptive because the benchmark systems are nonexchangeable.

### Supplementary Figure 1



The descriptive HCC1395-to-HG008 comparison is supplied as Supplementary\_Figure\_1.png. It does not estimate a population-level transportability effect and does not change the frozen HG008 primary endpoint.

# Supplementary Table 1

Supplementary\_Table\_1\_within50bp.tsv contains the full prespecified exploratory nogermmlinewithin50bp results for the six frozen procedures.

| method              | partition | tp    | fp   | fn   | precision          | recall             | f1                 | n_truth | n_called | un |
|---------------------|-----------|-------|------|------|--------------------|--------------------|--------------------|---------|----------|----|
| Mutect2             | overall   | 10068 | 312  | 954  | 0.9699421965317919 | 0.9134458356015243 | 0.9408466498458088 | 11022   | 10380    | 0  |
| Mutect2             | snv       | 7101  | 84   | 446  | 0.9883089770354906 | 0.9409036703325825 | 0.9640238935650285 | 7547    | 7185     | 0  |
| Mutect2             | indel     | 2967  | 75   | 508  | 0.97534516765286   | 0.8538129496402878 | 0.9105416602731318 | 3475    | 3042     | 0  |
| Strelka2            | overall   | 9859  | 902  | 1163 | 0.9161787937923985 | 0.8944837597532208 | 0.9052013037689942 | 11022   | 10761    | 0  |
| Strelka2            | snv       | 7483  | 833  | 64   | 0.8998316498316499 | 0.9915198091957069 | 0.943453319044317  | 7547    | 8316     | 0  |
| Strelka2            | indel     | 2376  | 69   | 1099 | 0.9717791411042945 | 0.6837410071942446 | 0.8027027027027027 | 3475    | 2445     | 0  |
| DeepSomatic         | overall   | 9721  | 48   | 1301 | 0.9950864981062545 | 0.8819633460352023 | 0.9351161560290511 | 11022   | 9769     | 0  |
| DeepSomatic         | snv       | 7507  | 39   | 40   | 0.9948316989133316 | 0.9946998807473169 | 0.9947657854634598 | 7547    | 7546     | 0  |
| DeepSomatic         | indel     | 2214  | 9    | 1261 | 0.9959514170040485 | 0.6371223021582734 | 0.7771147771147772 | 3475    | 2223     | 0  |
| union               | overall   | 10646 | 1199 | 376  | 0.898775854791051  | 0.9658864090001814 | 0.9311234530108891 | 11022   | 11845    | 0  |
| union               | snv       | 7515  | 915  | 32   | 0.891459074733096  | 0.9957599045978535 | 0.940727295487263  | 7547    | 8430     | 0  |
| union               | indel     | 3131  | 131  | 344  | 0.9598405885959534 | 0.9010071942446043 | 0.9294938399881253 | 3475    | 3262     | 0  |
| majority            | overall   | 9999  | 58   | 1023 | 0.9942328726260317 | 0.907185628742515  | 0.9487167322928033 | 11022   | 10057    | 0  |
| majority            | snv       | 7492  | 38   | 55   | 0.9949535192563081 | 0.9927123360275606 | 0.9938316641241627 | 7547    | 7530     | 0  |
| majority            | indel     | 2507  | 20   | 968  | 0.9920854768500198 | 0.7214388489208633 | 0.8353882039320226 | 3475    | 2527     | 0  |
| strict_intersection | overall   | 9003  | 5    | 2019 | 0.9994449378330373 | 0.8168209036472509 | 0.8989515726410384 | 11022   | 9008     | 0  |
| strict_intersection | snv       | 7084  | 3    | 463  | 0.9995766897135601 | 0.9386511196501921 | 0.968156348230149  | 7547    | 7087     | 0  |
| strict_intersection | indel     | 1919  | 2    | 1556 | 0.9989588755856325 | 0.5522302158273381 | 0.7112676056338029 | 3475    | 1921     | 0  |

## Supplementary Table 2

Supplementary\_Table\_2\_all\_region\_bootstrap.tsv contains the prespecified all-region sensitivity uncertainty results using 5-Mb blocks, 10,000 replicates, seed 3003, linear percentile intervals and minimum defined fraction 0.95.

| method              | partition | tp    | fp   | fn   | point_f1           | ci_low              | ci_high           |
|---------------------|-----------|-------|------|------|--------------------|---------------------|-------------------|
| Mutect2             | overall   | 12469 | 822  | 5738 | 0.7917328084322814 | 0.7853633500570103  | 0.797991485619482 |
| Mutect2             | snv       | 8456  | 296  | 998  | 0.9289245303746018 | 0.9231756016714462  | 0.934452338859964 |
| Mutect2             | indel     | 4013  | 311  | 4740 | 0.6137493308862889 | 0.6026268719841767  | 0.624720893304730 |
| Strelka2            | overall   | 12149 | 4475 | 6058 | 0.6975969682179668 | 0.6691324835678832  | 0.720501995317457 |
| Strelka2            | snv       | 9013  | 3952 | 441  | 0.8040501360453187 | 0.7582916100273988  | 0.841199730774238 |
| Strelka2            | indel     | 3136  | 523  | 5617 | 0.5053174347405737 | 0.49333153861085777 | 0.517189937013353 |
| DeepSomatic         | overall   | 11618 | 133  | 6589 | 0.7756192002136324 | 0.7689285305921033  | 0.782157375009817 |
| DeepSomatic         | snv       | 8948  | 99   | 506  | 0.96729906491541   | 0.9632053315451884  | 0.971021159473427 |
| DeepSomatic         | indel     | 2670  | 34   | 6083 | 0.4660905996334119 | 0.454225981982979   | 0.478038191857156 |
| union               | overall   | 13500 | 5297 | 4707 | 0.7296508485569128 | 0.7017857510454688  | 0.752096681474794 |
| union               | snv       | 9107  | 4255 | 347  | 0.7982994389901823 | 0.7531795600206782  | 0.835078827681338 |
| union               | indel     | 4393  | 827  | 4360 | 0.6287840835897803 | 0.6177713619511712  | 0.63939177433758  |
| majority            | overall   | 12069 | 126  | 6138 | 0.7939609236234458 | 0.7875414663804782  | 0.800292885159443 |
| majority            | snv       | 8952  | 89   | 502  | 0.9680454176804542 | 0.9638515129305675  | 0.971858415934950 |
| majority            | indel     | 3117  | 37   | 5636 | 0.5235575711766188 | 0.5117251666059565  | 0.535281471293880 |
| strict_intersection | overall   | 10667 | 7    | 7540 | 0.7386863335757072 | 0.731559778923804   | 0.745648704159754 |
| strict_intersection | snv       | 8358  | 3    | 1096 | 0.9383104125736739 | 0.9331210433943543  | 0.943301701978649 |
| strict_intersection | indel     | 2309  | 4    | 6444 | 0.4173142960419302 | 0.40478110843335996 | 0.429706962091964 |

## Supplementary Table 3

Supplementary\_Table\_3\_development\_external\_descriptive.tsv contains the post-hoc descriptive development-to-external comparison. The two benchmark systems differ in samples, data generation, truth construction and evaluated regions.

| Method             | G6_HCC1395_F1 | G7_HG008_F1 | Delta_G7_m |
|--------------------|---------------|-------------|------------|
| Mutect2            | 0.883600      | 0.897772    | +0.014172  |
| Strelka2           | 0.932000      | 0.778996    | -0.153004  |
| DeepSomatic        | 0.964800      | 0.897278    | -0.067522  |
| Union              | 0.932700      | 0.801075    | -0.131625  |
| Majority           | 0.956200      | 0.913635    | -0.042565  |
| StrictIntersection | 0.891900      | 0.860568    | -0.031332  |