

Valection: Design Optimization for Validation and Verification Studies

Christopher I. Cooper^{1,+}, Delia Yao^{1,+}, Dorota H. Sendorek^{1,+}, Takafumi N. Yamaguchi¹, Christine P'ng¹, Kathleen E. Houlahan^{1,2}, Cristian Caloian¹, Michael Fraser³, SMC-DNA Challenge Participants, Kyle Ellrott^{4,5,6}, Adam A. Margolin^{4,5,7}, Robert G. Bristow^{2,3}, Joshua M. Stuart⁶, Paul C. Boutros^{1,2,8,*}

¹ Ontario Institute for Cancer Research, Toronto, Canada
² Department of Medical Biophysics, University of Toronto, Toronto, Canada
³ Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
⁴ Computational Biology Program, Oregon Health & Science University, Portland, OR, USA
⁵ Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA
⁶ Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA
⁷ Sage Bionetworks, Seattle, WA, USA
⁸ Department of Pharmacology & Toxicology, University of Toronto, Toronto, Canada
⁺ Equal first authors
^{*} Corresponding author: Paul.Boutros@oicr.on.ca; 661 University Avenue, Suite 510; Toronto, Ontario, Canada; M5G 0A3; 416-673-8564

Author Emails:

Christopher I. Cooper	chriscooper1991@gmail.com
Delia Yao	delia.yao@gmail.com
Dorota H. Sendorek	dorota.sendorek@oicr.on.ca
Takafumi N. Yamaguchi	takafumi.yamaguchi@oicr.on.ca
Christine P'ng	chspng@gmail.com
Kathleen E. Houlahan	katie.houlahan@oicr.on.ca
Cristian Caloian	cristian.caloian@gmail.com
Michael Fraser	michael.fraser@oicr.on.ca
Kyle Ellrott	ellrott@ohsu.edu
Adam A. Margolin	margolin@ohsu.edu
Robert G. Bristow	robert.bristow@manchester.ac.uk
Joshua M. Stuart	jstuart@ucsc.edu
Paul C. Boutros	paul.boutros@oicr.on.ca

41 **Keywords:** verification, validation, candidate-selection, next-generation
42 sequencing

43 **Abstract**

44 **Background**

45 Platform-specific error profiles necessitate confirmatory studies where
46 predictions made on data generated using one technology are additionally
47 verified by processing the same samples on an orthogonal technology. In
48 disciplines that rely heavily on high-throughput data generation, such as
49 genomics, reducing the impact of false positive and false negative rates in
50 results is a top priority. However, verifying all predictions can be costly and
51 redundant, and testing a subset of findings is often used to estimate the true
52 error profile. To determine how to create subsets of predictions for validation
53 that maximize inference of global error profiles, we developed Valection, a
54 software program that implements multiple strategies for the selection of
55 verification candidates.

56 **Results**

57 To evaluate these selection strategies, we obtained 261 sets of somatic
58 mutation calls from a single-nucleotide variant caller benchmarking challenge
59 where 21 teams competed on whole-genome sequencing datasets of three
60 computationally-simulated tumours. By using synthetic data, we had complete
61 ground truth of the tumours' mutations and, therefore, we were able to
62 accurately determine how estimates from the selected subset of verification
63 candidates compared to the complete prediction set. We found that selection
64 strategy performance depends on several verification study characteristics. In
65 particular the verification budget of the experiment (*i.e.* how many candidates

66 can be selected) is shown to influence estimates.

67 **Conclusions**

68 The Valection framework is flexible, allowing for the implementation of
69 additional selection algorithms in the future. Its applicability extends to any
70 discipline that relies on experimental verification and will benefit from the
71 optimization of verification candidate selection.

72 **Background**

73 High-throughput genomics studies often exhibit error profiles that are biased
74 towards certain data characteristics. For example, predictions of single-
75 nucleotide variants (SNVs) from DNA sequencing data have error profiles
76 biased by local sequence context [1-2], mappability of the region [3] and many
77 other factors [4-5]. The false positive rate for individual predictions in high-
78 throughput studies is frequently high [6-7], while the false negative rate is
79 difficult to estimate and rarely known. Critically, error rates can vary
80 significantly between studies because of tissue-specific characteristics, such
81 as DNA quality and sample purity, and differences in data processing
82 pipelines and analytical tools. In cancer studies, variations in normal tissue
83 contamination can further confound genomic and transcriptomic analyses [8-
84 10].

85 Taken together, these factors have necessitated the wide-spread use of
86 studies with orthogonal technologies, both to verify key hits of interest and to
87 quantify the global error rate of specific pipelines. In contrast to a *validation*
88 *study*, which typically approaches the same biological question using an
89 independent set of samples (e.g. like a test dataset in a machine learning
90 exercise), we define a *verification study* as interrogating the same sample-set

91 with an independent method (*i.e.* a method that generates analogous data
92 using a distinct chemistry). The underlying concept is that if the second
93 technique has separate error profiles from the first, a comparative analysis
94 can readily identify false positives (*e.g.* in inconsistent, low quality calls) and
95 even begin to elucidate the false negative rate (*e.g.* from discordant, high
96 quality calls).

97 The choice of verification platform is critical as it determines both the tissue
98 and financial resources required. There is typically a wide range of potential
99 verification technologies for any given study. While confirmation of DNA-
100 sequencing results traditionally involves gold-standard Sanger sequencing
101 [11-12], the drawbacks of this approach (*e.g.* high financial and resource
102 costs) and advancements in newer sequencing techniques have shifted the
103 burden of variant verification to other technologies [13-15]. For example, a
104 typical Illumina-based next-generation sequencing (NGS) whole-genome or
105 whole-exome experiment may be verified by sequencing a separate library on
106 a different but similar machine [16]. This offers the advantages of high-
107 throughput, low cost and the opportunity to interrogate inter-library differences
108 [17]. Other groups have applied mass-spectrometric based corroboration of
109 individual variants, which has the benefit of technological independence [18-
110 19].

111 Apart from choice of technology, all groups must make decisions regarding
112 the *scope* of their verification work. For example when considering genome-
113 wide discovery, it may be appropriate to verify only known candidate drug
114 target mutations or unexpected novel functional aberrations. However, in
115 many contexts having an unbiased estimate of the global error rate is critical.
116 This is particularly true when benchmarking different data-generating methods

117 or when looking at genome-wide trends. It remains unclear how best to select
118 targets for verification studies, particularly in the context of fairly comparing
119 multiple methods and providing unbiased performance metric estimates. To
120 address this problem, we created Valection, a software tool that implements a
121 series of diverse variable selection strategies, thereby providing the first
122 framework for guiding optimal selection of verification candidates. To
123 benchmark different strategies, we exploit data from the ICGC-TCGA DREAM
124 Somatic Mutation Calling Challenge (SMC-DNA), where we have a total of
125 2,051,714 predictions of somatic SNVs made by 21 teams through 261
126 analyses [20, 4]. We show that the optimal strategy changes in a predictable
127 way based on characteristics of the verification experiments.

128 **Results**

129 We began by developing six separate strategies for selecting candidates for
130 verification (**Figure 1**). The first is a naïve approach that samples each
131 mutation with equal probability, independent of whether a mutation is
132 predicted by multiple algorithms or of how many calls a given algorithm has
133 made ('random rows'). Two simple approaches follow that divide mutations
134 either by recurrence ('equal per overlap') or by which algorithm made the call
135 ('equal per caller'). Finally, we created three approaches that account for both
136 factors: 'increasing per overlap' (where the probability of selection increases
137 with call recurrence), 'decreasing per overlap' (where the probability of
138 selection decreases with call recurrence) and 'directed-sampling' (where the
139 probability of selection increases with call recurrence while ensuring an equal
140 proportion of targets is selected from each caller). All methods have
141 programmatic bindings in four separate open-source languages (C, R, Perl
142 and Python) and are accessible through a systematic API through the

143 Valection software package. Valection thus becomes a test-bed for groups to
 144 try new ways of optimizing verification candidate-selection strategies.

145 To compare the six methods outlined above, we used data from tumour-
 146 normal whole-genome sequencing pairs from the ICGC-TCGA DREAM
 147 Somatic Mutation Calling Challenge [20, 4]. These tumours differ in major
 148 characteristics such as normal contamination, sub-clonality and mutation rate.
 149 We chose to work with simulated tumours because we know the ground truth
 150 of their mutational profiles, allowing a precise evaluation of the effectiveness
 151 of different selection schemes in estimating the true underlying error rates.
 152 Altogether, there are results available from 261 SNV calling analyses
 153 performed by 21 teams. We designed a rigorous parameter-sweeping
 154 strategy, considering different numbers of SNV calling algorithms and different
 155 quantities of verification candidate targets. The experimental design is
 156 outlined in **Figure 2**.

157 We assessed the performance of the candidate-selection strategies in two
 158 ways. First, we considered how close the predicted F_1 score from a simulated
 159 verification experiment is to that from the overall study. We calculated
 160 precision in two modes: 'default' (as described in Methods) and 'weighted'
 161 (where precision scores were modified so that unique calls carried more
 162 weight than calls predicted by multiple callers). Second, we assessed the
 163 variability in this result across 10 replicate runs of each strategy, allowing us
 164 to gauge how much random chance elements of variant-selection perturb the
 165 results of a given method (*i.e.* a stability analysis).

166 Overall, across all simulations, the 'equal per caller' approach performs best,
 167 showing a negligible mean difference between subset and total F_1 scores
 168 while, additionally, displaying low variability (*i.e.* small spread) in F_1 score

169 differences across all runs (**Figure 3**). Both the number of algorithms tested
 170 and the verification budget size (*i.e.* the number of candidates being selected)
 171 factor into which strategy performs optimally. Specifically, when there are
 172 large numbers of algorithms or the number of possible verification targets is
 173 low, the 'equal per caller' method does extremely well ($n_{\text{targets}} = 100$;
 174 **Supplementary Figure 1**). By contrast, when the number of verification
 175 targets is substantially larger (*i.e.* a considerable proportion of all predictions
 176 will be tested), the 'random rows' method shows similar performance levels
 177 ($n_{\text{targets}} = 1000$ and $n_{\text{targets}} = 2500$; **Supplementary Figures 2 and 3**,
 178 **respectively**). However, the 'random rows' method performs poorly when
 179 prediction set sizes are highly variable (*i.e.* a small number of callers has a
 180 large fraction of the total calls), resulting in some callers with no calls by which
 181 to estimate performance. This was the case for runs with verification budgets
 182 of $n_{\text{targets}} = 250$ (**Supplementary Figure 4**), $n_{\text{targets}} = 500$ (**Supplementary**
 183 **Figure 5**) and, in particular, $n_{\text{targets}} = 100$ (**Supplementary Figure 1**). Missing
 184 scores were treated as missing data.

185 However, the effects of the verification experiment characteristics described
 186 above alone do not account for all the variability observed across the
 187 simulations. Comparing runs of matching parameter combinations across the
 188 three synthetic tumours reveals some inter-tumour differences. Unlike with
 189 tumours IS1 (**Supplementary Figure 6**) and IS2 (**Supplementary Figure 7**),
 190 the 'random rows' method performs best on tumour IS3 suggesting tumour
 191 characteristics may have an impact on target selection strategy performance
 192 (**Supplementary Figure 8**). The 'equal per caller' method is only the second
 193 best selection strategy for the IS3 dataset.

194 We further assessed variability in the results of the selection strategies by

195 running 10 replicate runs of each. The results in **Figure 4** show that the
 196 consistency of performance across simulations trends with the overall
 197 performance of the selection strategy. An overall positive effect of the
 198 adjustment step ('weighted mode') on the selection strategies is also visible
 199 with the exception of the 'random rows' method, on which the weighted
 200 precision calculation appears to have no effect. A closer look at the recall and
 201 precision scores reveals that the approach with the poorest recall score,
 202 'decreasing with overlap' (**Supplementary Figure 9a**), also shows the most
 203 sensitivity to the weighted adjustment step in precision calculations
 204 (**Supplementary Figure 9b**). Altogether, across methods, recall scores tend
 205 to mirror F_1 scores in both magnitude and amount of spread, which is lower in
 206 approaches with higher recall. In contrast, precision scores are highly variable
 207 across most selection approaches, regardless of their overall performance.

208 **Discussion**

209 Assessing and comparing the quality of new prediction tools is an important
 210 step in their adoption and the truth of their results is arguably the most
 211 important component of their quality. When the resources required to
 212 independently verify results are substantial, it is vital to choose an unbiased
 213 but maximally informative set of results. This is naturally true not just for
 214 somatic SNVs, but other predictions like structural variants, fusion proteins,
 215 alternative splicing events and epigenetic phenomena, e.g. methylation and
 216 histone marks. Ongoing research into the error profiles of various data types
 217 increases our understanding of what factors influence verification rates [21].
 218 This information helps in distinguishing high- from low-quality calls and goes
 219 towards minimizing the amount of prediction verification required. However,
 220 with the continuous emergence of new data-generating technologies, e.g.

221 third generation sequencing [22], benchmarking studies assessing false
 222 positive and false negative rates are likely to remain a fundamental
 223 component of computational biological research well into the foreseeable
 224 future. Having standardized methods for comparing workflows in contexts
 225 such as these will ease the uptake of new techniques more confidently.
 226 Valtion is a first step towards standardizing and optimizing verification
 227 candidate selection.

228 Evaluation of the target candidate selection approaches presented in this
 229 study provides an in-depth view of the effects of call recurrence and algorithm
 230 representation on a verification candidate set. Nonetheless, this is by no
 231 means an exhaustive set of selection strategies. Although, our findings
 232 suggest that the most straightforward approaches (e.g. 'random rows') are
 233 often the most effective, future implementations of more complex strategies
 234 may highlight additional factors important to target candidate selection.

235 The need for informative verification target selections also highlights the
 236 importance of simulators for experimental biology, since the best suited
 237 method may vary from dataset to dataset. Indeed, as our findings here
 238 suggest, optimal candidate-selection strategies for somatic SNV calls may
 239 even be affected by various tumour data characteristics. A complete
 240 assessment of error profiles is impossible without access to multifarious
 241 datasets with an established ground truth. As such, there is a need for reliable
 242 simulators in biology to create and analyze gold-standard synthetic datasets
 243 to help guide top empirical research. For some time computationally-
 244 simulated data has been used to circumvent the difficulties that arise when
 245 working with real data [23]. The production of varied synthetic data is
 246 comparatively cheap and efficient, restricted only by the computational power

247 and storage space required to generate and hold it. With complete control
248 over data feature profiles, researchers are able to query numerous biological
249 questions simultaneously. As demonstrated here, and specific to cancer
250 genomics, synthetic tumour data can expedite accurate estimation of false
251 negative rates which are difficult to determine in genome-wide mutation
252 calling, thus mitigating the need for large-scale wet lab validation of non-
253 variants. It is important to note, however, that the utility of synthetic data is
254 limited to non-exploratory research. Biological processes or data features that
255 are unknown or poorly understood cannot be adequately simulated, leading to
256 a lack of 'real-world' complexity. Therefore, the interplay between
257 experimental and simulated data is critical to the advancement of 'big data'
258 disciplines such as genomics. As such, subsequent assessment using
259 comprehensively-characterized real data will be vital to further optimizing
260 candidate-selection strategy.

261 **Conclusions**

262 Verification of somatic SNV calls made on NGS tumour data is critical due to
263 the high numbers of false positive and false negative calls. However, a
264 thorough search to identify all erroneous calls is a cumbersome and
265 expensive task. Our findings suggest that it may also be an avoidable one.
266 Fewer verification targets may be sufficient to characterize global error rates
267 in data, provided that there is proper optimization of the target candidate
268 selection process. We find that this optimization must factor in not just the
269 scope of the verification study but, conceivably, the characteristics of the
270 dataset itself. To date, few studies have assessed candidate-selection
271 methods for verification purposes. Here, we begin to explore the alternatives
272 available to big data analysts performing confirmatory studies that are both

273 efficient and thorough. By releasing our Valection software publicly, we
274 encourage groups across the wider research community to continue this work.
275 With a straightforward implementation and easy application, Valection has the
276 potential for maximal impact across a wide range of disciplines that rely on
277 verification studies.

278 **Methods**

279 **Selection Strategies & Software**

280 The **random rows** selection strategy (**Figure 1b**) samples calls at random
281 without replacement from the entire set of calls, and continues until the
282 verification budget has been reached, or there are no more calls left.

283 The **directed-sampling** selection strategy (**Figure 1c**) begins by constructing
284 a matrix. Row 1 contains all the calls made only by individual callers, row 2
285 contains the calls made by exactly 2 callers, all the way to row N, which
286 contains the calls that were made by all of the N callers. Each column, j, of the
287 matrix contains only the calls made the j^{th} caller. Note that this means in all
288 rows past 1, calls appear in multiple cells on the same row. Any given cell
289 holds zero or more calls. To select calls, the following procedure is followed
290 for each row, from N to 1, and for each cell in that row, ordered by ascending
291 number of calls:

- 292 • Calculate the cell budget as the total remaining verification budget
293 divided among the yet unexamined cells in the rest of the matrix.
- 294 • Select calls without replacement from the cell in question up to the cell
295 budget (these calls become invalid selections for future cells). Each call
296 selected reduces the total remaining verification budget.
- 297 • If any budget remains once all cells have been selected from, the

298 process is repeated.

299 The **equal per caller** selection strategy (**Figure 1d**) divides the verification
300 budget equally among all callers. The set of calls that each individual caller
301 made is sampled from without replacement up to that caller's portion of the
302 total budget. A call selected by one caller becomes an invalid choice for all
303 other callers. If a single caller does not have enough available calls (calls not
304 yet selected in another caller's budget), its remaining budget is distributed
305 equally to the other callers.

306 The **equal per overlap** selection strategy (**Figure 1e**) is based around the
307 number of times each call was made. With N callers, the verification budget is
308 divided N ways. Out of the set of calls made only once (all the calls unique to
309 any caller), calls are selected without replacement up to the sub-budget. This
310 is repeated for all the calls made by exactly two callers, and so on up every
311 level of overlap. If a single level of overlap does not have enough available
312 calls (calls not yet selected in another overlap level's budget), its remaining
313 budget is distributed equally to the other levels.

314 The **increasing with overlap** selection strategy (**Figure 1f**) is similar to equal
315 per overlap, but instead of selecting an equal number of calls at every level of
316 overlap, it selects a number from each level of overlap proportional to the
317 level of overlap.

318 The **decreasing with overlap** selection strategy (**Figure 1g**) is identical to
319 increasing with overlap, but the number of calls selected at each level is
320 inversely proportional to the level of overlap.

321 All of these methods are available through four commonly used programming
322 languages C, Perl, Python and R. The implementations have robust user-level
323 documentation and are openly available at both their appropriate public

324 repositories (*i.e.* CPAN, PyPI and CRAN) and on our website at:
 325 labs.oicr.on.ca/boutros-lab/software/valection.

326 The selection strategy algorithms were implemented in C, and compiled using
 327 the GNU Compiler Collection (v4.8.1). The implementations also made use of
 328 GLib (v 2.44.0). The R statistical environment (v3.1.3) was used for statistical
 329 analysis and data subsetting. Perl (v5.18.2) was used to coordinate the
 330 simulations. All plots were generated with the same version of R using the
 331 “BPG” (v5.2.8) [24], “lattice” (v0.20-31) and “latticeExtra” (v0.6-26) packages.
 332 The analysis scripts are also available at [http://labs.oicr.on.ca/boutros-](http://labs.oicr.on.ca/boutros-lab/software/valection)
 333 [lab/software/valection](http://labs.oicr.on.ca/boutros-lab/software/valection).

334 **Simulated Data**

335 To test the accuracy of these different approaches empirically, we applied
 336 them to gold-standard data from the ICGC-TCGA DREAM Somatic Mutation
 337 Calling Challenge [20]. This is a global crowd-sourced benchmarking
 338 competition aiming to define the optimal methods for the detection of somatic
 339 mutations from NGS-based whole-genome sequencing. The challenge has
 340 two components, one using simulated data created using BAMSurgeon
 341 software [4] and the other using experimentally-verified analyses of primary
 342 tumours. To test the accuracy of our approaches on representation
 343 algorithms, we exploited the SNV data from the first three *in silico* tumours.
 344 This dataset comprises 261 genome-wide prediction sets made by 21 teams
 345 and there are no access restrictions. The raw BAM files are available at SRA
 346 with IDs SRX570726, SRX1025978 and SRX1026041. Truth files are
 347 available as VCFs at <https://www.synapse.org/#!/Synapse:syn2177211>.
 348 Prediction-by-submission matrices for all submissions are provided in
 349 Supplementary Tables 1-3, as well as the best submissions from each team in

350 Supplementary Table 4, truth calls in Supplementary Tables 5-7 and a
 351 confusion matrix in Supplementary Table 8.

352 To probe a range of possible verification studies, we ran a very broad set of
 353 simulations. For each run, we pre-specified a tumour, a number of algorithms
 354 and a number of mutations to be selected for verification, and ran each of the
 355 candidate-selection strategies listed above. We then calculated the F_1 score
 356 (along with precision and recall) based on the verification study, assuming
 357 verification results are ground truth. Finally, we compared the true F_1 for a
 358 given algorithm on a given tumour across all mutations to the one inferred
 359 from the verification experiment.

360 We used three separate tumours with diverse characteristics
 361 (<https://www.synapse.org/#!Synapse:syn312572/wiki/62018>), including a
 362 range of tumour cellularities and the presence or absence of sub-clonal
 363 populations. We selected subsets of algorithms for benchmarking in four
 364 different ways:

- 365 i) the complete dataset (X)
- 366 ii) the single best submission from each team (X-best)
- 367 iii) three randomly selected entries from X-best (repeated 10 times)
- 368 iv) 25 randomly selected entries from X (repeated 10 times)

369 Lastly, we considered verification experiment sizes of 100, 250, 500, 1000
 370 and 2500 candidates per tumour. Thus, in total, we analyzed each of the
 371 candidate-selection algorithms in 22 datasets for 3 tumours and 5 verification
 372 sizes, for 330 total comparisons.

373 **Statistical Analyses**

374 The precision, recall and F_1 score of each caller were calculated as follows,

375 from the caller's true positive (TP), false positive (FP) and false negative (FN)
 376 values, as estimated by the selection strategy. Here, FNs are true calls
 377 sampled by the selection strategy that were not made by the caller in question
 378 (i.e. another caller made it).

$$precision = \frac{TP}{TP + FP} \quad (1)$$

379

$$recall = \frac{TP}{TP + FN} \quad (2)$$

380

$$F_1 score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (3)$$

381

382 When no calls were selected to calculate a value for a caller, scores were
 383 given values of N/A. This happened primarily with the 'random rows' method.
 384 Additionally, each precision score was calculated in an adjusted and
 385 unadjusted manner. A caller's precision in the unadjusted form was calculated
 386 exactly as described above, using all the calls made by the caller and
 387 selected for verification as the TPs and FPs. In the adjusted form, the
 388 selected calls were first divided into groups, according to how many callers
 389 made the call. Then, the precision was calculated separately using the calls
 390 from each group. The final precision was calculated as a weighted average of
 391 the precision of each group of calls, with weights equal to the total number of
 392 calls (verified and unverified) that caller made at that overlap level. Thus, in a
 393 two-caller example, a caller that made 100 unique calls and 50 calls shared

394 with the other caller would count its precision from unique calls twice as
395 strongly as its precision from shared calls.

396 **List of abbreviations**

397 SNV: single-nucleotide variant

398 NGS: next-generation sequencing

399 ICGC: International Cancer Genome Consortium

400 TCGA: The Cancer Genome Atlas

401 DREAM: Dialogue for Reverse Engineering Assessments and Methods

402 SMC-DNA: Somatic Mutation Calling DNA Challenge

403 TP: true positive

404 FP: false positive

405 FN: false negative

406 **Declarations**

407 **Ethics approval and consent to participate**

408 Not applicable

409 **Consent for publication**

410 Not applicable

411 **Availability of data and material**

412 The datasets supporting the conclusions of this article are included in its
413 additional files and in the Supplementary of Ewing *et al.* [4]. The main
414 Valection project page is at:

415 <http://labs.oicr.on.ca/boutros-lab/software/valection>

416 Programmatic bindings for the source-code are additionally available at:

417 <https://pypi.python.org/pypi/valelection/1.0.1>

418 <http://search.cpan.org/dist/Bio-Sampling-Valelection/>

419 **Competing interests**

420 All authors declare that they have no competing interests.

421 **Funding**

422 This study was conducted with the support of the Ontario Institute for Cancer
423 Research to PCB through funding provided by the Government of Ontario.

424 This work was supported by Prostate Cancer Canada and is proudly funded
425 by the Movember Foundation - Grant #RS2014-01. Dr. Boutros was
426 supported by a Terry Fox Research Institute New Investigator Award and by a
427 CIHR New Investigator Award. This project was supported by Genome
428 Canada through a Large-Scale Applied Project contract to PCB and Drs.
429 Sohrab Shah and Ryan Morin. This work was supported by the Discovery
430 Frontiers: Advancing Big Data Science in Genomics Research program,
431 which is jointly funded by the Natural Sciences and Engineering Research
432 Council (NSERC) of Canada, the Canadian Institutes of Health Research
433 (CIHR), Genome Canada and the Canada Foundation for Innovation (CFI).
434 This work was supported by the National Cancer Institute of the US National
435 Institutes of Health under award number R01CA180778 (J.M.S., K.E.).

436 **Authors' contributions**

437 Initiated the project: CIC, JMS, PCB

438 Data preparation: TNY, CC, KEH

439 Generated tools and reagents: CIC, DY, TNY, CP, KE

440 Performed statistical and bioinformatics analyses: CIC, DY, DHS, TNY, KEH

441 Supervised research: MF, KE, AAM, RGB, JMS, PCB

442 Wrote the first draft of the manuscript: DHS, PCB

443 Approved the manuscript: all authors

444 **Acknowledgements**

445 The authors thank Dr. Jared Simpson and Dr. John McPherson for thoughtful
446 discussions and all members of the Boutros lab and the SMC-DNA team for
447 helpful suggestions.

448 **References**

1. Abnikova I, Leonard S, Skelly T, Brown A, Jackson D, Gourtovaia M, Qi G, Te Boekhorst R, Faruque N, Lewis K, Cox T. Analysis of context-dependent errors for Illumina sequencing. *J Bioinform Comput Biol*. 2012 Apr;10(2):1241005.
2. Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L. Identification and correction of systemic error in high-throughput sequence data. *BMC Bioinformatics*. 2011 Nov 21;12:451.
3. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P. Fast computation and applications of genome mappability. *PLoS One*. 2012;7(1):e30377.
4. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY, ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Kellen MR, Norman TC, Haussler D, Friend SH, Stolovitzky G, Margolin AA, Stuart JM, Boutros PC. Combining tumor genome simulation with crowdsourcing

- to benchmark somatic single-nucleotide-variant detection. *Nat Methods*. 2015 Jul;12(7):623-30.
5. Hofmann AL, Behr J, Singer J, Kuipers J, Beisel C, Schraml P, Moch H, Beerenwinkel N. Detailed simulation of cancer exome sequencing data reveals differences common limitations of variant callers. *BMC Bioinformatics*. 2017 Jan 3;18(1):8. doi: 10.1186/s12859-016-1417-7.
 6. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M. Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol*. 2012 Jan;30(1):78-82.
 7. Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res*. 2014 Nov;24(11):1734-9.
 8. Quon G, Haider S, Deshwar AG, Cui A, Boutros PC, Morris Q. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med*. 2013 Mar 28;5(3):29.
 9. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol*. 2012 May;30(5):413-21.
 10. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Cote A, Shah SP. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014 Apr;11(4):396-8.

11. Lee LG, Connell CR, Woo SL, Cheng RD, McArdle BF, Fuller CW, Halloran ND, Wilson RK. DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. *Nucleic Acids Res.* 1992 May 25;20(10):2471-83.
12. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E, Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Committee. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med.* 2013 Sep;15(9):733-47.
13. Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, van Spaendonck-Zwarts KY, van Tintelen JP, Sijmons RH, Jongbloed JD, Sinke RJ. Targeted next-generation sequencing can replace Sanger in clinical diagnostics. *Hum Mutat.* 2013 Jul;34(7):1035-42.
14. Nelson AC, Bower M, Baughn LB, Henzler C, Onsongo G, Silverstein KAT, Schomaker M, Deshpande A, Beckman KB, Yohe S, Thyagarajan B. Criteria for clinical reporting of variants from a broad target capture NGS assay without Sanger verification. *JSM Biomark.* 2015;2(1):1005.
15. Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, Deignan JL. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. *Genet Med.* 2014 Jul;16(7):510-5.
16. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature.* 2015 Jan 29;517(7536):576-82.

17. Chong LC, Albuquerque MA, Harding NJ, Caloian C, Chan-Seng-Yue M, de Borja R, Fraser M, Denroche RE, Beck TA, van der Kwast T, Bristow RG, McPherson JD, Boutros PC. SeqControl: process control for DNA sequencing. *Nat Methods*. 2014 Oct;11(10):1071-5.
18. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*. 2010 May 27;465(7297):473-7.
19. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, Schuster SC. Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One*. 2013 Feb 6;8(2):e55089. doi: 10.1371/journal.pone.0055089.
20. Boutros PC, Margolin AA, Stuart JM, Califano A, Stolovitzky G. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biol*. 2014 Sep 17;15(9):462. doi: 10.1186/s13059-014-0462-7.
- 449 21. Park M-H, Rhee H, Park JH, Woo HM, Choi BO, Kim BY, Chung KW,
450 Cho YB, Kim HJ, Jung JW, Koo SK. Comprehensive analysis to
451 improve the validation rate for single nucleotide variants detected by
452 next-generation sequencing. *PLoS One*. 2014 Jan 29;9(1):e86664.
- 453 22. Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S, Goodwin S,
454 McCombie WR, Schatz M. Third-generation sequencing and the future
455 of genomes. *bioRxiv* 048603; doi: <https://doi.org/10.1101/048603>.

- 456 23. Escalona M, Rocha S, Posada D. A comparison of tools for the
457 simulation of genomic next-generation sequencing data. Nat Rev
458 Genet. 2016 Aug;17(8):459-69.
- 459 24. P'ng C, Green J, Chong LC, Waggott D, Prokopec SD, Shamsi M,
460 Nguyen F, Mak DYF, Lam F, Albuquerque MA, Wu Y, Jung EH,
461 Starmans MHW, Chan-Seng-Yue MA, Yao CQ, Liang B, Lalonde E,
462 Haider S, Simone NA, Sendorek D, Chu KC, Moon NC, Fox NS,
463 Grzadkowski MR, Harding NJ, Fung C, Murdoch AR, Houlahan KE,
464 Wang J, Garcia DR, de Borja R, Sun RX, Lin X, Chen GM, Lu A, Shiah
465 Y-J, Zia A, Kearns R, Boutros P. BPG: seamless, automated and
466 interactive visualization of scientific data. bioRxiv 156067; doi:
467 <https://doi.org/10.1101/156067>.

468 Figure Legends

469 Figure 1: Valection Candidate-Selection Strategies

- 470 a) A hypothetical scenario where we have results from three callers available.
471 Each call is represented using a dot. SNV calls that are shared by multiple
472 callers are represented with matching dot colours. b) The 'random rows'
473 method where all unique calls across all callers are sampled from with equal
474 probability. c) The 'directed-sampling' method where a 'call overlap-by-caller'
475 matrix is constructed and the selection budget is distributed equally across all
476 cells. d) The 'equal per caller' method where the selection budget is
477 distributed evenly across all callers. e) The 'equal per overlap' method where
478 the selection budget is distributed evenly across all levels of overlap (*i.e.* call
479 recurrence across callers). f) The 'increasing with overlap' method where the
480 selection budget is distributed across overlap levels in proportion to the level

481 of overlap. g) The 'decreasing with overlap' method where the selection
482 budget is distributed across overlap levels in inverse proportion to the level of
483 overlap.

484 **Figure 2: Verification Selection Experimental Design**

485 Verification candidates were selected from somatic mutation calling results of
486 multiple algorithms run on three *in silico* tumours (IS1, IS2, and IS3).
487 Candidate selection was performed separately on each tumour's set of results
488 using all combinations of five different verification budgets (*i.e.* number of calls
489 selected) and six different selection strategies. F_1 scores were calculated for
490 each set of selected calls and compared to F_1 scores calculated from the full
491 prediction set. To compare the effect of the numbers of algorithms used,
492 datasets were further subset using four different metrics.

493 **Figure 3: All Simulation Results for Selection Strategy Parameter** 494 **Combinations**

495 Overall, the best results are obtained using the 'equal per caller' method. The
496 'random rows' approach scores comparably except in cases where there is
497 high variability in prediction set sizes across callers. Calls from low-call callers
498 are less likely to be sampled at random and, in cases where none are
499 sampled, it is not possible to get performance estimates for those callers.
500 Failed estimate runs are displayed in grey.

501 **Figure 4: F_1 Scores Across Replicate Runs.**

502 Top selection strategies perform consistently across replicate runs. Strategies
503 are ordered by median scores. The adjustment step in precision calculations
504 improves the 'equal per caller' method, but shows little effect on 'random
505 rows'.

506 **Additional files**

507 **Additional file 1: Supplementary Figure 1.**

508 TIFF 9.4 Mb

509 Simulations with 100 verification targets, across all tumours. The 'equal per
510 caller' method (weighted mode) performs optimally as the 'random rows'
511 method generates N/As.

512 **Additional file 2: Supplementary Figure 2.**

513 TIFF 9.2 Mb

514 All simulations with 1000 verification targets, across all tumours. The best
515 results come from the 'random rows' and the 'equal per caller' (weighted
516 mode) methods.

517 **Additional file 3: Supplementary Figure 3.**

518 TIFF 8.9 Mb

519 All simulations with 2500 verification targets, across all tumours. The best
520 results come from the 'random rows' and the 'equal per caller' (weighted
521 mode) methods.

522 **Additional file 4: Supplementary Figure 4.**

523 TIFF 9.9 Mb

524 All simulations with 250 verification targets, across all tumours. The 'equal per
525 caller' method (weighted mode) performs optimally as the 'random rows'
526 method generates N/As.

527 **Additional file 5: Supplementary Figure 5.**

528 TIFF 9.6 Mb

529 All simulations with 500 verification targets, across all tumours. The 'equal per
530 caller' method (weighted mode) performs optimally as the 'random rows'
531 method generates N/As.

532 **Additional file 6: Supplementary Figure 6.**

533 TIFF 18 Mb

534 All simulations for tumour IS1. Optimal results are achieved with the 'equal per
535 caller' method (weighted mode).

536 **Additional file 7: Supplementary Figure 7.**

537 TIFF 12 Mb

538 All simulations for tumour IS2. Optimal results are achieved with the 'equal per
539 caller', 'increasing per overlap' and 'equal per overlap' methods (weighted
540 mode).

541 **Additional file 8: Supplementary Figure 8.**

542 TIFF 14 Mb

543 All simulations for tumour IS3. Optimal results are achieved with the 'random
544 rows' method, regardless of how precision is calculated.

545 **Additional file 9: Supplementary Figure 9.**

546 TIFF 4.1 Mb

547 a) Recall scores from all runs, displayed per candidate-selection strategy. b)
548 Precision scores from all runs, calculated with and without a weight
549 adjustment step (default mode and weighted mode, respectively) and
550 displayed per candidate-selection strategy.

551 **Additional file 10: Supplementary Table 1.**

552 CSV 57 Mb

553 A prediction-by-submission matrix of all SNV call submissions for tumour IS1

554 where SNV predictions are annotated with chromosome ("CHROM") and

555 position ("END").

556 **Additional file 11: Supplementary Table 2.**

557 CSV 29 Mb

558 A prediction-by-submission matrix of all SNV call submissions for tumour IS2

559 where SNV predictions are annotated with chromosome ("CHROM") and

560 position ("END").

561 **Additional file 12: Supplementary Table 3.**

562 CSV 3.6 Mb

563 A prediction-by-submission matrix of all SNV call submissions for tumour IS3

564 where SNV predictions are annotated with chromosome ("CHROM") and

565 position ("END").

566 **Additional file 13: Supplementary Table 4.**

567 CSV 3.3 kb

568 A summary table of the top team submissions for each tumour, includes

569 submission ID, team alias, the number of true positives, true negatives, false

570 positives and false negatives, as well as the precision, recall and F₁ scores.

571 **Additional file 14: Supplementary Table 5.**

572 CSV 3.1 Mb

573 A table of all predicted SNVs for tumour IS1, annotated by chromosome

574 (“chrom”) and position (“pos”), and a “truth” column for whether the call is a
575 true positive (1) or not (0).

576 **Additional file 15: Supplementary Table 6.**

577 CSV 2.5 Mb

578 A table of all predicted SNVs for tumour IS2, annotated by chromosome
579 (“chrom”) and position (“pos”), and a “truth” column for whether the call is a
580 true positive (1) or not (0).

581 **Additional file 16: Supplementary Table 7.**

582 CSV 329 kb
























583 A table of all predicted SNVs for tumour IS3, annotated by chromosome
584 (“chrom”) and position (“pos”), and a “truth” column for whether the call is a
585 true positive (1) or not (0).

586 **Additional file 17: Supplementary Table 8.**

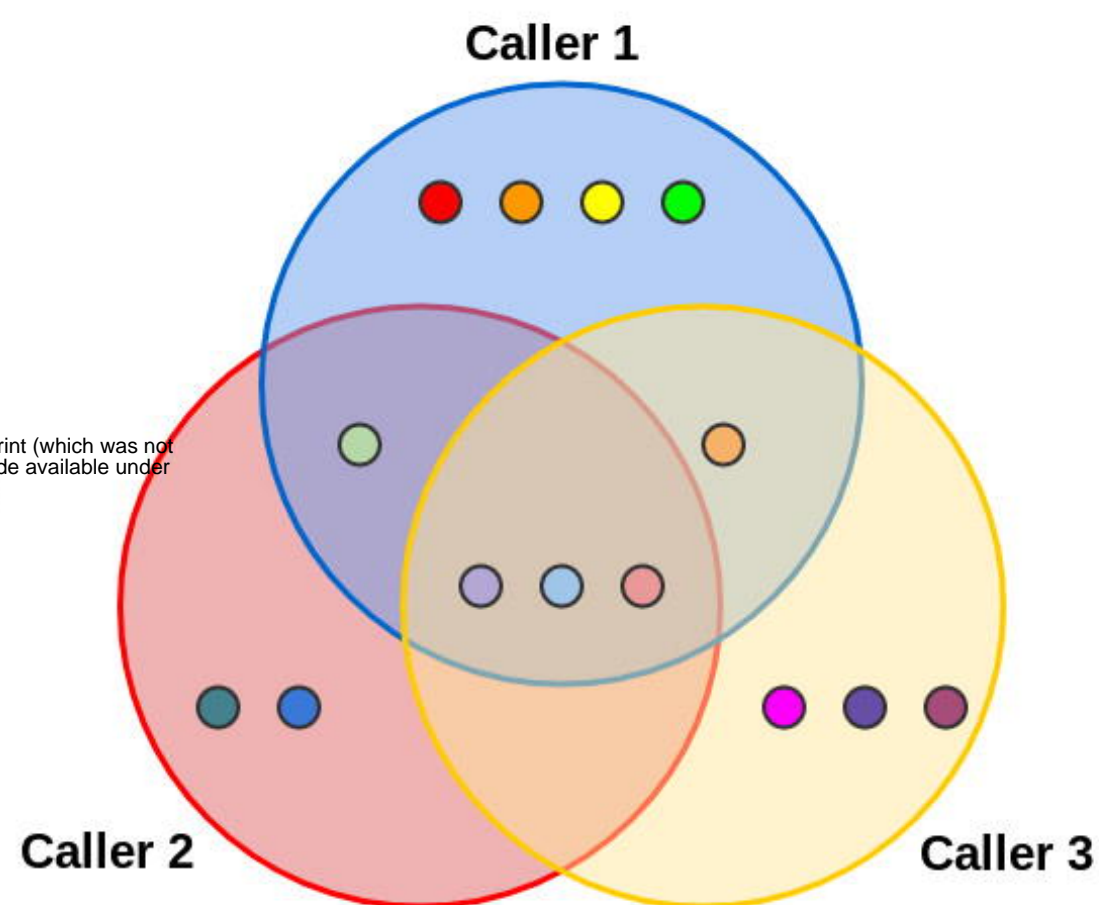
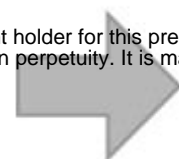
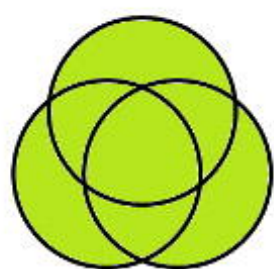
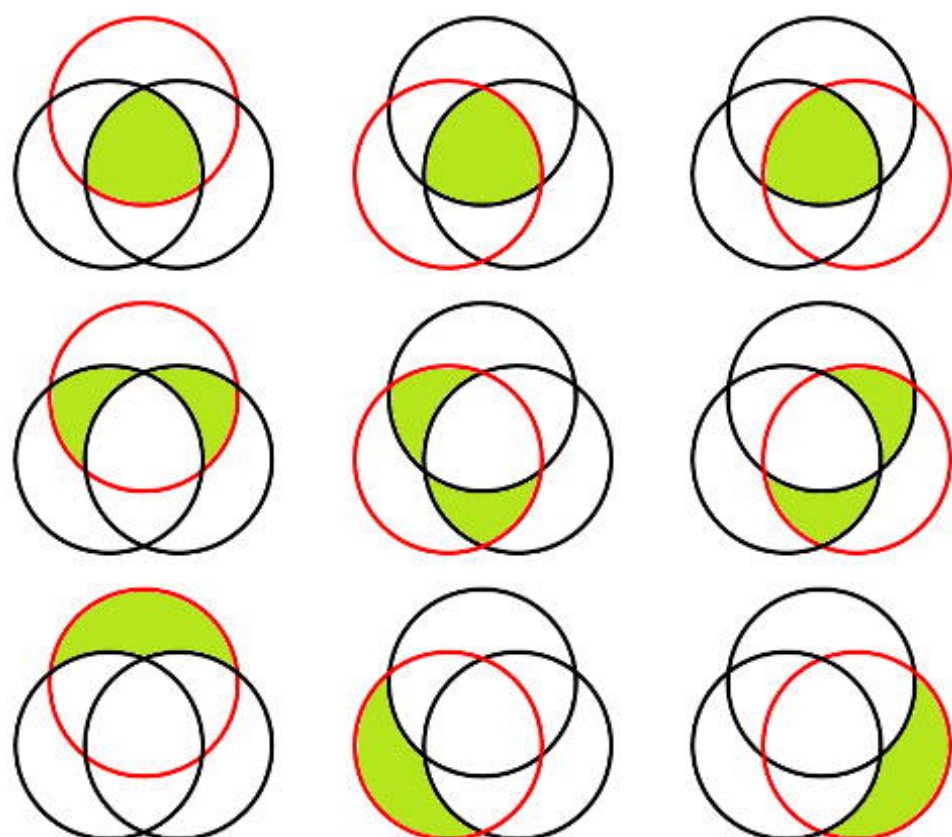
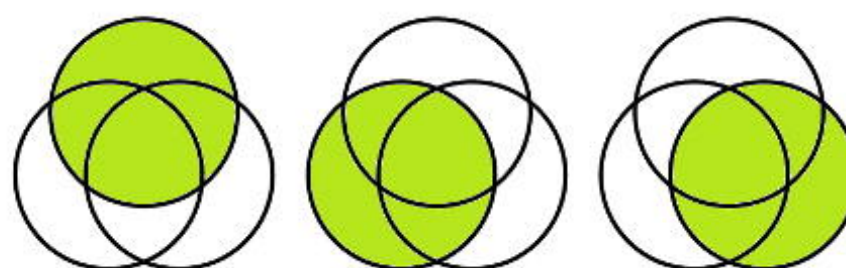
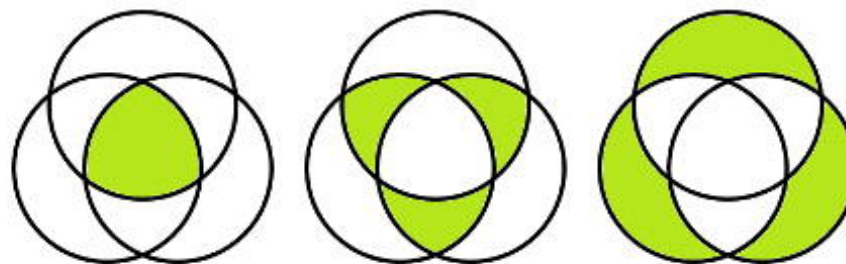
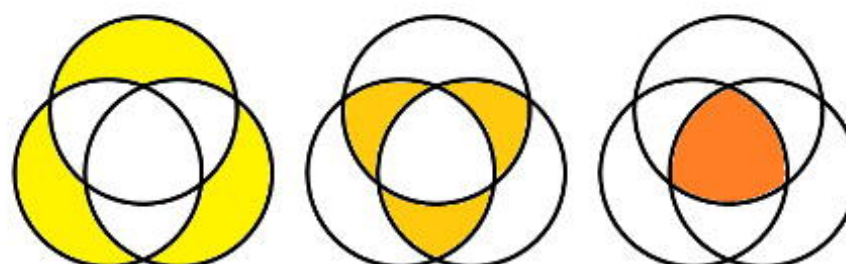
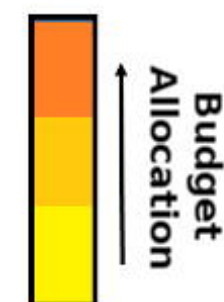
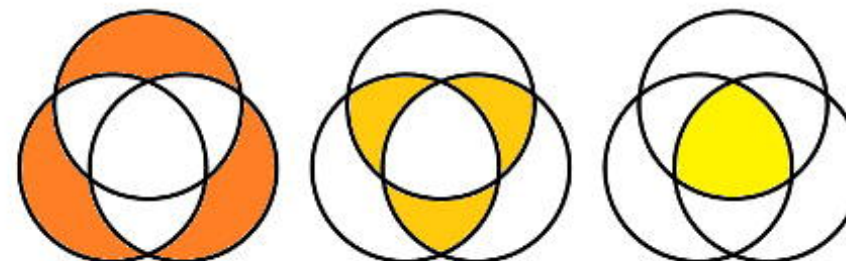
587 CSV 20 kb

588 A summary table of all submissions from across all tumours, includes
589 submission ID, the number of true positives, true negatives, false positives
590 and false negatives, as well as the precision, recall and F_1 scores.

a

Caller 1	Caller 2	Caller 3
		
		
		
		
		
		
		
		
		

bioRxiv preprint doi: <https://doi.org/10.1101/254839>; this version posted January 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**b****c****d****e****f****g**

Verification Selection Experimental Design

bioRxiv preprint doi: <https://doi.org/10.1101/254493>; this version posted January 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Experiment Condition Parameters

Tumour	Data subset method	Verification Budget
IS1	All data	100
IS2	Best from each team	250
IS3	Best from 3 random teams	500
	Best from 3 random teams	1000
	25 random submissions	2500

All combinations

Experiment Conditions

Each condition

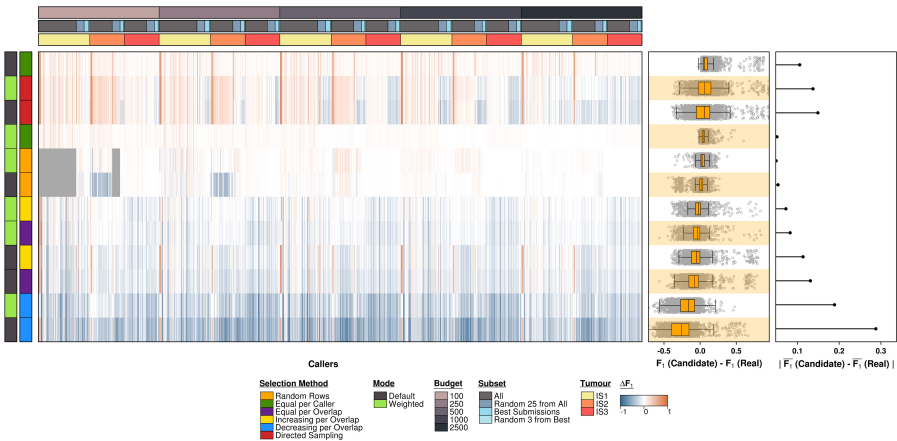
Condition

verification selection strategies

selected calls

calculate F1 scores

Plot and analyze all results



Selection Method

- Random Rows
- Equal per Caller
- Equal per Overlap
- Increasing per Overlap
- Decreasing per Overlap
- Directed Sampling

Mode

- Default
- Weighted

