

Germline Contamination and Leakage in Whole Genome Somatic Single Nucleotide Variant Detection

Dorota H. Sendorek^{1,*}, Cristian Caloian^{1,*}, Kyle Ellrott^{3,4}, J. Christopher Bare², Takafumi N. Yamaguchi¹, Adam D. Ewing^{3,5}, Kathleen E. Houlahan¹, Thea C. Norman², Adam A. Margolin^{2,4,6}, Joshua M. Stuart³, Paul C. Boutros^{1,7,8,**}

¹ Informatics & Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Canada

² Sage Bionetworks, Seattle, WA, USA

³ Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA

⁴ Computational Biology Program, Oregon Health & Science University, Portland, OR, USA

⁵ Mater Research Institute, University of Queensland, Woolloongabba, Queensland, Australia

⁶ Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA

⁷ Department of Medical Biophysics, University of Toronto, Toronto, Canada

⁸ Department of Pharmacology & Toxicology, University of Toronto, Toronto, Canada

* Equal Contributions

** Corresponding Author: paul.boutros@oicr.on.ca; 661 University Avenue, Suite 510; Toronto, Ontario, Canada; M5G 0A3; 416-673-8564

Author Emails:

Dorota H. Sendorek	dorota.sendorek@oicr.on.ca
Cristian Caloian	cristian.caloian@gmail.com
Kyle Ellrott	ellrott@ohsu.edu
J. Christopher Bare	christopherbare@gmail.com
Takafumi N. Yamaguchi	takafumi.yamaguchi@oicr.on.ca
Adam D. Ewing	adam.ewing@mater.uq.edu.au
Kathleen E. Houlahan	Katie.Houlahan@oicr.on.ca
Thea C. Norman	thea.norman@sagebase.org
Adam A. Margolin	margolin@ohsu.edu
Joshua M. Stuart	jstuart@ucsc.edu
Paul C. Boutros	paul.boutros@oicr.on.ca

Keywords: cancer genomics, next-generation sequencing, mutation calling, germline contamination, germline leakage, patient identifiability, single nucleotide variant, SNV

Abstract

Background

The clinical sequencing of cancer genomes to personalize therapy is becoming routine across the world. However, concerns over patient re-identification from these data lead to questions about how tightly access should be controlled. It is not thought to be possible to re-identify patients from somatic variant data. However, somatic variant detection pipelines can mistakenly identify germline variants as somatic ones, a process called “germline leakage”. The rate of germline leakage across different somatic variant detection pipelines is not well-understood, and it is uncertain whether or not somatic variant calls should be considered re-identifiable. To fill this gap, we quantified germline leakage across 259 sets of whole-genome somatic single nucleotide variant (SNVs) predictions made by 21 teams as part of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge.

Results

The median somatic SNV prediction set contained 4,325 somatic SNVs and leaked one germline polymorphism. The level of germline leakage was inversely correlated with somatic SNV prediction accuracy and positively correlated with the amount of infiltrating normal cells. The specific germline variants leaked differed by tumour and algorithm. To aid in quantitation and correction of leakage, we created a tool, called GermlineFilter, for use in public-facing somatic SNV databases.

Conclusions

The potential for patient re-identification from leaked germline variants in somatic SNV predictions has led to divergent open data access policies, based on different assessments of the risks. Indeed, a single, well-publicized re-identification event could reshape public perceptions of the values of genomic data sharing. We find that modern somatic SNV prediction pipelines have low germline-leakage rates, which can be further reduced, especially for cloud-sharing, using pre-filtering software.

57

58 **Background**

59 The appropriate limits on data sharing remains a contentious issue throughout biomedical research, as
 60 shown by recent controversies [1]. Studies such as the Personal Genome Project (PGP) have pioneered
 61 open sharing of patient data for biomedical research, while ensuring that enrolled patients consent to
 62 risks of identification [2]. In fact, analysis of PGP data has showed that a majority of participants can
 63 be linked to a specific named individual [3]. Identifiability is greatly facilitated when researchers
 64 release all generated data online – as is standard in some fields [4]. This public, barrier-free release has
 65 numerous advantages. It can minimize storage costs, increase data redundancy to reduce the risk of
 66 data-loss and maximize data availability and re-use. As a result, it is argued that barrier-free deposition
 67 of genomic data in public repositories like GEO [5-6] or dbGaP [7-8] promotes collaborative work and
 68 maximizes the value of already-funded research [9]. Further, many researchers believe they have an
 69 ethical duty to release all data [10].

70 Nevertheless, there are at least four counter-arguments in favour of a conservative approach to data
 71 protection. First, the groups generating the data have uniquely intimate knowledge of it and studies
 72 done without their participation can be more prone to errors, although improved documentation of the
 73 research process can mitigate this effect [1]. Second, the desire to immediately release data may oppose
 74 the desire to explore complex inter-linked questions. The initial report of a dataset may not fully reflect
 75 the magnitude of work that goes into generating it, particularly for clinical trials. With immediate data
 76 release, the data collectors may find themselves under time constraints, unable to comprehensively
 77 exploit the data they produced without competition from subsequent researchers who are able to use the
 78 data freely. This effectively disincentivizes the challenging work of dataset creation, producing a

situation akin to a tragedy of the commons. Third, the inherent value in large datasets may enable data producers to seek commercialization opportunities by keeping data resources private. Fourth, many studies involve data derived from human subjects that contain revealing and personal information, which is under legal protection [11]. Legislation designed to protect patient privacy, such as the Health Insurance Portability and Accountability Act (HIPAA) [12], the Common Rule [13] and the European Union's General Data Protection Regulation [14] impose harsh financial and professional penalties for violations. As genomic data becomes widely available and techniques for interpreting them improve, de-identification grows increasingly difficult, challenging implementation of barrier-free access that upholds ethical considerations. We focus here on this fourth challenge, or re-identifiability.

Earlier studies have quantified how much DNA information is required to identify individuals. One suggests that as few as 30-80 statistically-independent single nucleotide polymorphisms (SNPs) suffice [15]. Under certain circumstances, small segments of DNA can even be used to recover participants' names by accessing publicly available, commercial genealogy websites [16]. These problems are compounded by deficiencies in techniques used to prevent re-identification: for example, pooling DNA samples does not prevent detection of any individual sequence [17]. More recently, research into information leakage demonstrated how easily patients can be linked back to data from which they previously had been disassociated by correlating seemingly disparate features, namely from phenotypic and genotypic datasets, in what is referred to as a 'linking attack' [18-19].

In cancer research, many studies concentrate on identifying somatic mutations that are induced in the process of tumourigenesis and tumour evolution. Identifying these causative mutations can lead to discovery of novel biomarkers and potential therapeutic targets, making public data release critical for accelerating research. Because these mutations are found in the tumour and not in an individual's germline genome, they do not, by themselves, provide identifying information. Barrier-free release of

somatic mutational data can, in theory, occur without compromising patient privacy.

However, tools used to distinguish somatic mutations from germline are imperfect, and sometimes the predicted somatic mutations are in fact germline genetic variants. This “germline leakage” can occur in several ways. Most next-generation sequencing (NGS) base calling algorithms have low error rates [20], including both undetected true variants (false negatives) while some non-existent variants get reported (false positives). These false positives can occur for several reasons, including low coverage (number of reads aligning to a specific position in the genome), which reduces statistical confidence [21]. Even datasets with high total coverage have variable coverage across the genome with particular regions getting sampled at lower rates either through stochastic or structurally biased factors. As a result, sets of somatic variant predictions can be contaminated with germline variants, particularly in the case of single nucleotide variants (SNVs). To account for these errors, some groups filter out any variant seen in a germline database like dbSNP, while others allow only release of mutations in the exome [22]. Still others allow public release of somatic variant predictions from the whole genome [23]. These variations reflect differing views on the likelihoods and risks of germline leakage, and many groups have not yet developed or articulated specific policies.

To help improve our understanding of the magnitude of germline leakage, we analyzed a set of 259 somatic mutation predictions made by 21 groups from around the world on three synthetic tumours during the ICGC-TCGA DREAM Somatic Mutation Calling-DNA (SMC-DNA) Challenge [24]. We developed a software tool, called GermlineFilter, which can help to quantify and mitigate the risks of germline leakage for publicly available somatic SNV data.

Results

Gold Standards of Germline Leakage

We sought to evaluate the extent of germline contamination in contemporary cancer whole-genome sequencing (WGS) datasets, particularly those comprising somatic SNV predictions across the entire genome. To do so, we exploited the synthetic data from the ICGC-TCGA DREAM SMC-DNA Challenge [24-25], which benchmarked somatic SNV predictions using synthetic and real tumour-normal whole-genome pairs/. The generation of the synthetic tumours and their properties are fully detailed in Ewing *et al.* [25]. Briefly, high coverage binary alignment map (BAM) files were obtained from cell lines HCC1143 and HCC1954 [26]. BAMSurgeon [25] was used to randomly 'spike-in' germline mutations into the BAM files. Each file was then split into two: one file representing a synthetic tumour and the other file representing the matched normal. The tumour BAM file was finalized by adding somatic mutations: both SNVs and structural variants. This methodology allows for the creation of a “gold standard” dataset in which the precise locations of germline and somatic variants are known, enabling comprehensive assessment of leaked germline mutations. We focused on the first three synthetic tumours from SMC-DNA, referred to as IS1, IS2 and IS3. These tumours vary in the number of mutations, normal contamination and subclonal complexity (**Additional file 1: Supplementary Table 1**) [25]. The synthetic tumours have been available to the public for several years and have thus accumulated a large number of somatic mutation calling results from various submitted methods. Additionally, the organizers ran several widely used algorithms with default settings as a baseline [25]. In total, we evaluated 5,792,868 somatic mutations that included 259 analyses by 21 teams across the three tumours ($n_{IS1} = 120$; $n_{IS2} = 71$; $n_{IS3} = 68$).

Assessment of Germline Leakage

To quantify germline leakage in submissions to the SMC-DNA tumours, we created a Python program called GermlineFilter, which simultaneously evaluates germline leakage in somatic SNV predictions and filters them in real-time to allow barrier-free access to the final results. The overall process has two steps (**Figure 1**). During the initial preprocessing step, a germline caller is run on paired tumour and normal BAM files to generate the germline variant calls. Current germline callers have high accuracy rates which can be attributed to diploidy-based assumptions of normal human tissue, assumptions that do not hold for somatic variants due to a host of issues (*e.g.* intra-tumour heterogeneity, tissue cellularity, genomic instability). In the following step, each germline SNP is compared against the somatic SNV predictions to be filtered, provided in standard variant call format (VCF), and the matches are identified. Finally, somatic SNV calls can now be filtered, either by rejecting entire submissions that exceed an acceptable level of leakage or by simply removing the calls that match a germline variant.

Germline Contamination Reduces Somatic SNV Prediction Accuracy

The 259 somatic call VCFs submitted during the IS1, IS2 and IS3 phases of the SMC-DNA challenge contained a median of 4,325 SNV calls (averaging 22,366 SNV calls). Each of these was run through GermlineFilter to quantify germline leakage in terms of the number of true germline SNPs misidentified as somatic SNVs. Prediction accuracy for each submission was measured using the F_1 -score (*i.e.* the harmonic mean of precision and recall) in keeping with the metrics used in the DREAM SMC-DNA challenge.

Germline leakage was highly variable across submissions, ranging from 0 to 45,300, with a median of 1 per submission. The median leakage rate across tumours ranged from 0 (IS3), to 2 (IS1) and went up

as high as 6 (IS2). IS2 contained the highest normal contamination (20%), suggesting that even low normal contamination can increase germline leakage. For each tumour, we compared germline count to the previously reported F_1 -scores (**Figure 2a**) and found a highly significant negative correlation in each of the three tumours (Spearman's $\rho_{IS1} = -0.557$, $\rho_{IS2} = -0.477$, $\rho_{IS3} = -0.410$, **Additional file 1: Supplementary Table 1**). For a number of algorithms, the germline variants make up a substantial fraction of the total calls, showing an association with the number of false positive calls (**Figure 2b**). Thus germline leakage is, as expected, associated with reduced overall accuracy of mutation calling.

Quantifying Germline Leakage across Tumours and Between Algorithms

Submissions were further analyzed to determine recurrence of individual germline contaminants across the mutation calling algorithms. For these purposes, only the highest F_1 -score submission from each team was selected, as in the primary report of the somatic SNV data [25]. This was done separately for each tumour, resulting in 15 submissions for IS1, 12 for IS2 and 11 for IS3. A plurality of submissions harboured no germline variants (IS1 = 40.0%; IS2 = 41.7%; IS3 = 45.5%), but there was substantial variability, with one submission containing 43 germline SNPs (**Additional file 2: Supplementary Table 2**).

Individual leaked germline variants varied significantly across algorithms (**Figure 3**). Of the 85 germline variants leaked in the 12 IS2 submissions (all with an $F_1 > 0.863$), only five were identified more than once. Similarly, of the 23 germline variants leaked in the 11 IS3 submissions, only two were identified more than once. Leaked variants were distributed uniformly across chromosomes. These data suggest that in modern pipelines, germline leakage rates are low and different variants are leaked by different pipelines.

In addition to the spiked-in mutations, common SNP sites were also analyzed. The Exome Aggregation

Consortium (ExAC) has produced a library of variant sites seen across 60,706 individuals [27]. These sites represent locations where samples commonly deviate from the reference. Due to the very large number of individuals represented, this set of SNP sites is often used as a filter of possible germline variant sites. ExAC provides ~9.3 million potential common SNP sites, much more than the thousands of spiked-in mutations. The number of false positive calls using ExAC as a filter remained very low (medians: IS1 = 2; IS2 = 3; IS3 = 1.5). As these sites are publicly available and known to be common for SNPs, most modern somatic calling pipelines can directly incorporate this information into their filtering strategy.

Discussion

Barrier-free access to genomic data can expand its utility, maximizing investments in research funding, enabling citizen-scientists and facilitating collaboration. Strong barriers to access can limit these positive consequences of large investments in dataset generation. Indeed, even when data is made available through protected databases, the processes to gain access can be time-consuming, advantaging labs or institutions that have resources dedicated to gaining and maintaining data-access authorizations. Accessibility can be skewed by variability in the standards, knowledge and impartiality of data access committees that authorize use of controlled data [28-29].

We quantified the amount of leakage in three comprehensively studied tumours used in a crowd-sourced prediction benchmarking challenge. While some submissions showed large amounts of germline leakage, the median submission leaked only one germline SNP, and indeed the top three teams for each tumour leaked none. Given that the SMC-DNA Challenge was run in 2014-2015 and that detection pipelines and the quality of genomic data have improved further since, it appears that

modern optimized variant-calling pipelines leak an insignificant number of germline variants on many tumours, well below the 30-80 independent SNP range needed for re-identification [15].

However, several caveats must be evaluated when considering barrier-free access to whole-genome somatic SNV predictions. First, the data we evaluated only included three tumours, and further evaluations on larger numbers with a range of cellularities will be critical to generalize these conclusions. Additionally, while we considered the amount of germline leakage in tumors with different subclonal complexities, we did not investigate whether germline leakage is more likely in genomic regions with specific tumour characteristics (*e.g.* mutational hotspots, trinucleotide context, subclonality, copy number alterations, loss of heterozygosity, *etc.*). On-going work from the ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG) may provide the data necessary to address this. Second, genomic alterations other than nuclear SNVs (*e.g.* germline copy number variants and mitochondrial polymorphisms) may provide information contributing to identifiability. Third, while most individual pipelines leaked few variants, aggregating multiple pipelines could increase the information content: the union of variants across all 12 pipelines from IS2 contain 85 leaked SNPs, potentially providing sufficient information for re-identification [15]. Since ensemble calling generally adopts a 'majority rules' approach [30], which would remove most germline variants due to low recurrence, this is most relevant in cases of malicious intent.

Taken together, our findings suggest that germline contamination in somatic SNV calling is relatively rare, and supports additional consideration of barrier-free access to these data. Re-identification risks can be substantially reduced by incorporating automated checks into the data release process, designed to identify germline leakage and remove these prior to data release. GermlineFilter provides a convenient and secure way to monitor leakage by individual algorithms, and may be useful as a front-end to cloud-based SNV databases to quantify and minimize risk in real-time.

233

234

Methods

235

Software

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

GermlineFilter works in an encrypted fashion, allowing its use on a public server. It is executed in two steps (**Figure 1**). For the first step, performed offline, a VCF file containing germline calls is generated using paired tumour and normal BAM files. The user can additionally specify alternative germline call sets, such as a list of all dbSNP entries, although these would elevate the false-negative rate by removing true somatic mutations. For each germline SNP in the VCF file, we extract the chromosome, position, reference base and alternate base. This information is hashed and written to a file that is then encrypted. It is this file of hashes rather than the actual variants that is then transferred to the server. Next, online somatic VCF filtering is performed. At runtime, the truth germline VCF is decrypted in memory and the somatic VCF undergoes preprocessing and hashing. Finally, an in-memory comparison of hashes is done and the number of matches is returned. GermlineFilter can spawn multiple instances to process multiple germline VCFs for different tumours or multiple somatic VCFs for a single tumour. The user chooses the encryption and hashing protocols, with strong default settings in place to help minimize risks such as hash collisions. An additional feature for local use allows the user to obtain a list of the actual positions of the germline leaks within the somatic VCF. This list can be used to filter out the germline mutations in preparation for publication.

The GermlineFilter software package was written in Python 2.7 and it is supported for Unix and Linux platforms. The encryption and hashing is done using the *PyCrypto* v2.6.1 Python module. The tool currently supports two encryption protocols -- *AES* (default) and *Blowfish*, as well as two hashing protocols -- *SHA512* (default) and *md5*, selected for their security and broad usage. GermlineFilter v1.2

is the stable version and it is available for download at: <https://pypi.python.org/pypi/GermlineFilter>.
Alternatively, it can be installed via pip install GermlineFilter.

Data

The analysis data was taken from Ewing *et al.* [25] and it consists of the first three publicly available *in silico* datasets from the ICGC-TCGA DREAM Somatic Mutation Calling Challenge and their corresponding SNV submissions from the challenge participants. The truth germline calls were generated using *GATK HaplotypeCaller* v3.3. A description of the synthetic tumour data and a summary of participating teams and their submissions can be found in **Additional file 1: Supplementary Table 1**. All challenge submissions and their scores are listed in **Additional file 2: Supplementary Table 2**.

For each of the 259 submissions we calculated: precision (the fraction of submitted calls that are true somatic SNVs), recall (the fraction of true somatic SNVs that are identified by the caller) and the F_1 -score (the harmonic mean of precision and recall), as previously reported [25]. The F_1 -score was selected to be the accuracy metric as it does not rely on true negative information which, given the nature of somatic variant calling on whole genome sequencing data, would overwhelm alternative scoring metrics such as specificity (the fraction of non-SNV bases that are correctly identified as such by the caller).

Each tumour's germline calls were encrypted separately using default methods: AES for encryption and SHA512 for hashing. Somatic calls from all challenge submissions were filtered against their corresponding tumour's encrypted germline calls. For a somatic SNV call to be designated a germline leak, it exactly matched a germline variant at the chromosome, position, reference allele and alternate allele.

277 The resulting germline leak counts were compared to F_1 -scores using Spearman correlation. The best
 278 team submissions per tumour were selected to look at leaked germline variant recurrence across
 279 tumours and mutation callers. Best submissions were defined as having the highest F_1 -score.

280 **Visualization**

281 All data figures were created using custom R scripts executed in the R statistical environment (v3.2.3)
 282 using the *BPG* (v5.6.8) package [31].

283 **List of abbreviations used**

284 BAM: binary alignment map

285 DREAM: Dialogue on Reverse-Engineering Assessment and Methods

286 GATK: Genome Analysis Toolkit

287 HIPAA: Health Information Portability and Accountability Act

288 ICGC: International Cancer Genome Consortium

289 NGS: next-generation sequencing

290 PGP: Personal Genome Project

291 SMC: Somatic Mutation Calling

292 SNP: single nucleotide polymorphism

293 SNV: single nucleotide variant

294 TCGA: The Cancer Genome Atlas

295 VCF: variant call format

296

297 **Declarations**

298 **Ethics approval and consent to participate**

299 Not applicable

300 **Consent for publication**

301 Not applicable

302 **Availability of data and materials**

303 The datasets supporting the conclusions of this article are available on Synapse (syn312572) at:

304 <https://www.synapse.org/#!Synapse:syn312572/wiki/61509>

305 And in the Supplementary of Ewing *et al.* [25]. The main GermlineFilter project page is at:

306 <https://labs.oicr.on.ca/boutros-lab/software/germlinefilter>

307 And the source-code is freely available at:

308 <https://pypi.python.org/pypi/GermlineFilter/1.2>

309 **Competing interests**

310 All authors declare that they have no competing interests.

311 **Funding**

312 This study was conducted with the support of the Ontario Institute for Cancer Research to PCB through
 313 funding provided by the Government of Ontario. This work was supported by Prostate Cancer Canada
 314 and is proudly funded by the Movember Foundation - Grant #RS2014-01. This project was supported
 315 by Genome Canada through a Large-Scale Applied Project contract to P.C.B., S.P. Shah and R.D.
 316 Morin. This work was supported by the Discovery Frontiers: Advancing Big Data Science in Genomics
 317 Research program, which is jointly funded by the Natural Sciences and Engineering Research Council

(NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), Genome Canada, and the Canada Foundation for Innovation (CFI). P.C.B. was supported by a Terry Fox Research Institute New Investigator Award and a CIHR New Investigator Award. The following NIH grants supported this work: R01-CA180778 (J.M.S.), U24-CA143858 (J.M.S.), and U54-HG007990 (A.A.M.). The authors thank Google Inc. (in particular N. Deflaux) and Annai Biosystems (in particular D. Maltbie and F. De La Vega) for their ongoing support of the ICGC-TCGA DREAM Somatic Mutation Calling Challenge.

Authors' contributions

ADE, AAM, JMS and PCB initiated the project. CC created GermlineFilter and performed validation studies. KE, CC, JCB, TCN, AAM, JMS and PCB created the ICGC-TCGA DREAM Somatic Mutation Calling Challenge. DHS, CC, TNY, KEH and ADE created datasets and analyzed submission data. Research was supervised by AAM, JMS and PCB. The first draft of the manuscript was written by DHS, and approved by all authors.

Acknowledgements

The authors thank all members of the Boutros lab and all ICGC-TCGA DREAM Somatic Mutation Calling Challenge Participants for their support and thoughtful commentary.

References

1. Longo DL, Drazen JM. Data Sharing. N Engl J Med. 2016;374:276-7.
2. Personal Genome Project. Harvard Medical School, Boston. 2017.
<http://www.personalgenomes.org>. Accessed 12 Oct 2017.
3. Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name (a

- re-identification experiment). CoRR. 2013;abs/1304.7605:<http://arxiv.org/abs/1304.7605>.
4. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, Austin CP, Berglund L, Bobrow M, Bountra C, Brookes AJ, Cambon-Thomsen A, Carter NP, Chisholm RL, Contreras JL, Cooke RM, Crosby WL, Dewar K, Durbin R, Dyke SO, Ecker JR, El Emam K, Feuk L, Gabriel SB, Gallacher J, Gelbart WM, Granell A, Guarner F, Hubbard T, Jackson SA, Jennings JL, Joly Y, Jones SM, Kaye J, Kennedy KL, Knoppers BM, Kyrpides NC, Lowrance WW, Luo J, MacKay JJ, Martín-Rivera L, McCombie WR, McPherson JD, Miller L, Miller W, Moerman D, Mooser V, Morton CC, Ostell JM, Ouellette BF, Parkhill J, Raina PS, Rawlings C, Scherer SE, Scherer SW, Schofield PN, Sensen CW, Stodden VC, Sussman MR, Tanaka T, Thornton J, Tsunoda T, Valle D, Vuorio EI, Walker NM, Wallace S, Weinstock G, Whitman WB, Worley KC, Wu C, Wu J, Yu J. Prepublication data sharing. *Nature*. 2009;461:168-70.
5. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207-10.
6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41:D991-5.
7. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. The NCBI dbGaP Database of Genotypes and Phenotypes. *Nat Genet*. 2007;39:1181-6.

8. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* 2014;42:D975-9.
9. Rodriguez LL, Brooks DB, Greenberg JH, Green ED. Research ethics. The complexities of genomic identifiability. *Science.* 2013;339:275-6.
10. Lolkema MP, Gadellaa-van Hooijdonk CG, Bredenoord AL, Kapitein P, Roach N, Cuppen E, Knoers NV, Voest EE. Ethical, legal, and counseling challenges surrounding the return of genetic results in oncology. *J Clin Oncol.* 2013;31:1842-8.
11. Lowrance WW, Collins FS. Ethics. Identifiability in genomic research. *Science.* 2007;317:600-2.
12. U.S. Department of Health & Human Services: Health Information Privacy. <http://www.hhs.gov/hipaa/>.
13. U.S. Department of Health & Human Services: Federal Policy for the Protection of Human Subjects ('Common Rule'). <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/>.
14. European Commission: Justice: Protection of personal data. <http://ec.europa.eu/justice/data-protection/>.
15. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science.* 2004;305:183.
16. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science.* 2013;399:321-4.
17. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly

- complex mixtures using high-density SNP genotyping microarrays. *PloS Genet.* 2008;4:e1000167
18. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat Methods.* 2016;13:251-6.
19. Craig DW. Understanding the links between privacy and public data sharing. *Nat Methods.* 2016;13:211-2.
20. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* 2009;10:R32.
21. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet.* 2010;11:685-96.
22. The Cancer Genome Atlas Data Portal. Data Levels and Data Types: DNA Sequencing. <https://tcga-data.nci.nih.gov/tcga/tcgaDataType.jsp>. Accessed 29 Jan 2016.
23. International Cancer Genome Consortium. Goals, structure, policies & guidelines. 2008. https://icgc.org/files/icgc/ICGC_April_29_2008_en.pdf. Accessed 01 Feb 2016.
24. Boutros PC, Ewing AD, Ellrott K, Norman TC, Dang KK, Hu Y, Kellen MR, Suver C, Bare JC, Stein LD, Spellman PT, Stolovitzky G, Friend SH, Margolin AA, Stuart JM. Global optimization of somatic variant identification in cancer genomes with a global community challenge. *Nat Genet.* 2014;46:318-9.
25. 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;12:623-30.
26. Gazdar AF, Kurvari V, Virmani A, Gollahon L, Sakaguchi M, Westerfield M, Kodagoda D, Stasny V, Cunningham HT, Wistuba II, Tomlinson G, Tonk V, Ashfaq R, Leitch AM, Minna JD, Shay JW. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int J Cancer*. 1998;78:766-774.
27. Exome Aggregation Consortium, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* 2015; doi: <http://dx.doi.org/10.1101/030338>.
28. Shabani M, Dyke SOM, Joly Y, Borry P. Controlled access under review: improving the governance of genomic data access. *PloS Biol*. 2015;13:e1002339.
29. Joly Y, de Vries-Seguin E, Chalmers D, Ouellette BFF, Yamada J, Bobrow M, Knoppers BM for the ICGC Data Access Compliance Office and the ICGC International Data Access Committee. Analysis of five years of controlled access and data sharing compliance at the International Cancer Genome Consortium. *Nat Genet*. 2016;48:224-5.

30. Sage Bionetworks. TCGA Unified Ensemble “MC3” Call Set. 2016.

<https://www.synapse.org/#!Synapse:syn7214402/wiki/>. Accessed 11 Oct 2017.

31. P'ng C, Green J, Chong LC, Waggott D, Prokopec SD, Shamsi M, Nguyen F, Mak DYF, Lam F, Albuquerque MA, Wu Y, Jung EH, Starmans MHW, Chan-Seng-Yue MA, Yao CQ, Liang B, Lalonde E, Haider S, Simone NA, Sendorek D, Chu KC, Moon NC, Fox NS, Grzadkowski MR, Harding NJ, Fung C, Murdoch AR, Houlahan KE, Wang J, Garcia DR, de Borja R, Sun RX, Lin X, Chen GM, Lu A, Shiah Y-J, Zia A, Kearns R, Boutros P. BPG: Seamless, Automated and Interactive Visualization of Scientific Data. bioRxiv 2017; doi: <https://doi.org/10.1101/156067>.

Figure Legends

Figure 1: GermlineFilter Workflow

Locally, tumour-normal BAM files are submitted to a germline caller (*e.g.* GATK) to create a germline SNV call VCF file, which is later hashed and encrypted. The encrypted, hashed germline calls can now be moved to any server and used to filter for germline leakage in somatic SNV call VCF files. The output is the germline count found in the somatic calls.

Figure 2: Assessment of somatic SNV prediction accuracy against germline leakage

(a) F_1 -scores for each submission are plotted against the germline count (as determined by GermlineFilter). Submissions for different tumours are colour-coded (IS1 = orange, IS2 = green, IS3 = purple). The grey area represents 30-80 counts: the minimum number of independent SNPs required to correctly identify a subject, according to Lin *et al.* [15]. (b) Proportions of germline calls as found in total submission calls (upper panel) and in false positive submission calls (lower panel) per tumour. The horizontal red lines indicate the 30 count mark (the lower bound of the 30-80 SNP range

453 mentioned above).

454 **Figure 3: Germline leakage across all tumours (IS1, IS2, IS3) and SNV-calling algorithms.**

455 Teams are consistently colour-coded across multiple tumours. Barplots show F_1 -scores from each
 456 team's top-scoring submission. Leaked variants are displayed below with their corresponding
 457 chromosomes. Variant bars that overlap horizontally represent recurrent germline leaks.

458
 459 **Additional files**

460 **Additional file 1: Supplementary Table 1.**

461 XLS 12.3 kb

462 Tumour information from each tumour challenge (IS1, IS2, IS3). This includes information on *in silico*
 463 tumour construction, composition, and a summary of participating teams and their challenge
 464 submissions.

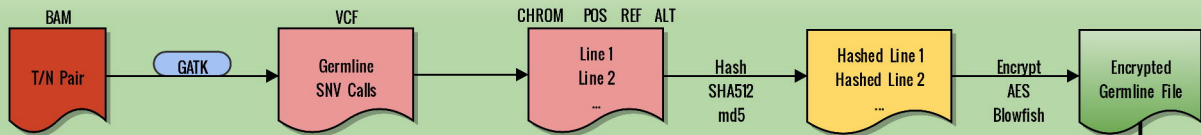
465 **Additional file 2: Supplementary Table 2.**

466 XLS 36.4 kb

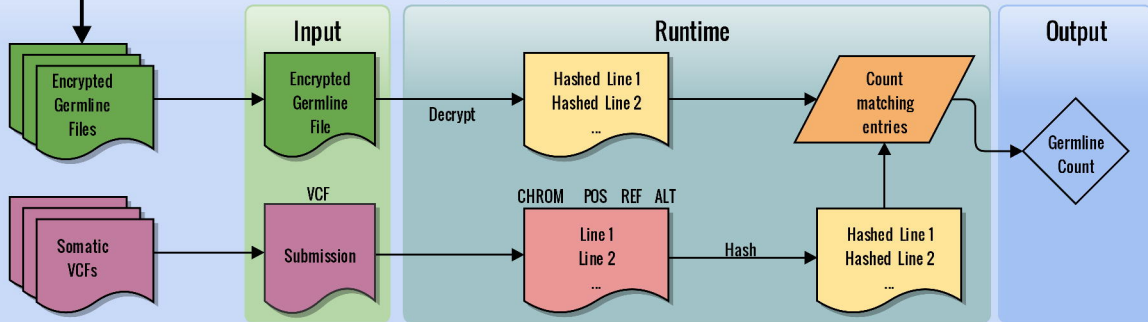
467 Contains the following information for every challenge submission: tumour, submission ID, precision,
 468 recall, F_1 -score, and the number of germline variants leaked.

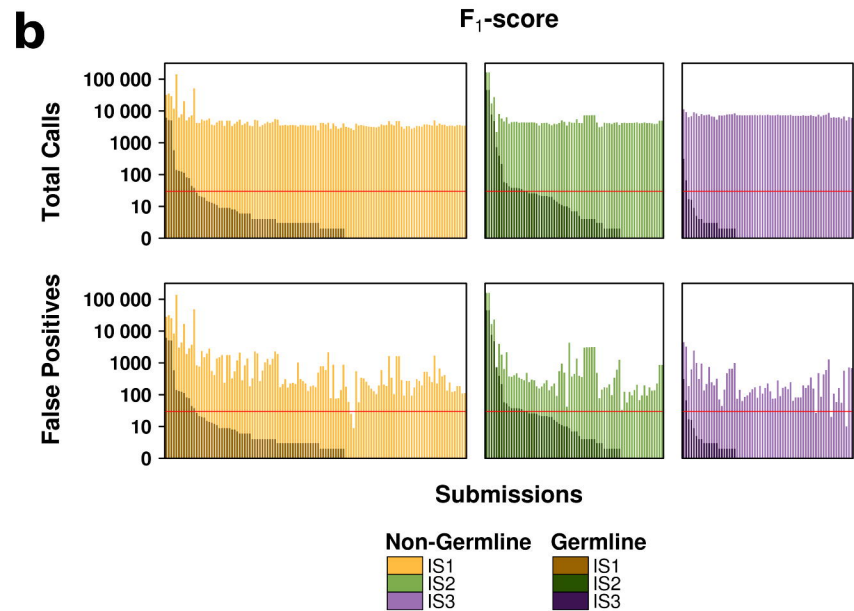
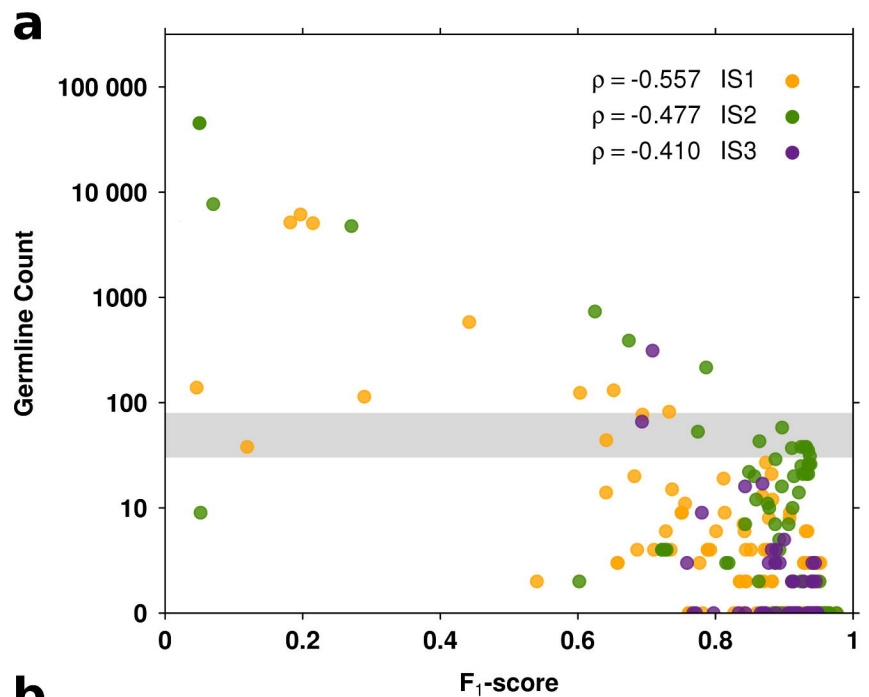
Germline Filter Workflow

In-House Preprocessing Of Truth Germline VCF

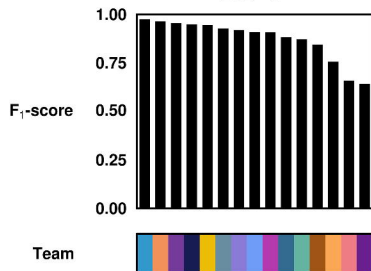


Filtering Somatic VCFs On Server

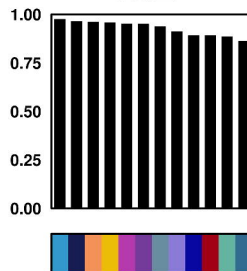




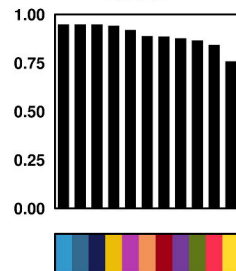
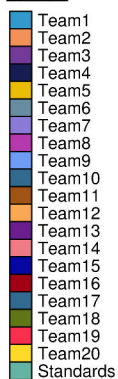
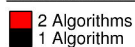
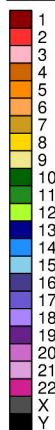
IS1



IS2



IS3

**Teams****Recurrent Variant****Chromosomes**

Position

