Establishing reference samples for detection of somatic mutations and germline variants with NGS technologies

Li Tai Fang[1*], Bin Zhu[2*], Yongmei Zhao[3*], Wanqiu Chen[4], Zhaowei Yang[4,30], Liz Kerrigan[5], Kurt Langenbach[5], Maryellen de Mars[5], Charles Lu[6], Kenneth Idler[6], Howard Jacob[6], Ying Yu[7], Luyao Ren[7], Yuanting Zheng[7], Erich Jaeger[8], Gary Schroth[8], Ogan D. Abaan[8], Justin Lack[3], Tsai-Wei Shen[3], Keyur Talsania[3], Zhong Chen[4], Seta Stanbouly[4], Jyoti Shetty[9], Bao Tran[9], Daoud Meerzaman[10], Cu Nguyen[10], Virginie Petitjean[11], Marc Sultan[11], Margaret Cam[12], Tiffany Hung[13], Eric Peters[13], Rasika Kalamegham[13], Sayed Mohammad Ebrahim Sahraeian[1], Marghoob Mohiyuddin[1], Yunfei Guo[1], Lijing Yao[1], Lei Song[2], Hugo YK Lam[1], Jiri Drabek[14,15], Roberta Maestro[15,16], Daniela Gasparotto[15,16], Sulev Kõks[15,17,18], Ene Reimann[15,18], Andreas Scherer[19,15], Jessica Nordlund[20,15], Ulrika Liljedahl[20,15], Roderick V Jensen[21], Mehdi Pirooznia[22], Zhipan Li[23], Chunlin Xiao[24], Stephen Sherry[24], Rebecca Kusko[25], Malcolm Moos[26], Eric Donaldson[27], Zivana Tezak[28], Baitang Ning[29], Jing Li[30], Penelope Duerken-Hughes[31], Huixiao Hong[29#], Leming Shi[7#], Charles Wang[4,31#], Wenming Xiao[28,29#], and The Somatic Working Group of SEQC-II Consortium

[1]Bioinformatics Research & Early Development, Roche Sequencing Solutions Inc., 1301 Shoreway Road, Suite #300, Belmont, CA 94002; [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, 9609 Medical Center Drive, Bethesda, Maryland 20892, USA.; [3]Advanced Biomedical and Computational Sciences, Biomedical Informatics and Data Science Directorate, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD21701; [4]Center for Genomics, Loma Linda University School of Medicine, 11021 Campus St., Loma Linda, CA 92350; [5]ATCC (American Type Culture Collection), 10801 University Blvd, Manassas, VA 20110; [6]Computational Genomics, Genomics Research Center (GRC), AbbVie, 1 North Waukegan Road, North Chicago, IL 60064; [7]State Key Laboratory of Genetic Engineering, School of Life Sciences and Shanghai Cancer Center, Fudan University, 2005 SongHu Road, Shanghai, China 200438; [8]Core applications group, Product development, Illumina Inc , 200 Lincoln Centre Dr. Foster City, CA 94404; [9]Genomics Laboratory, Cancer Research Technology Program, Frederick National Laboratory for Cancer Research, 8560 Progress Drive, Frederick, MD21701; [10]Computational Genomics Research, Center for Biomedical Informatics and Information Technology (CBIIT), National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850; [11]Biomarker development, Novartis Institutes for Biomedical Research, Fabrikstrasse 10, CH-4056 Basel, Switzerland; [12]CCR Collaborative Bioinformatics Resource (CCBR), Office of Science and Technology Resources, Center for Cancer Research, NCI, Bldg 37, Rm 3041C, 37 Convent Drive, Bethesda, MD 20892; [13]Companion Diagnostics Development, Oncology Biomarker Development, Genentech, 1 DNA Way, South San Francisco, CA 94080; [14]IMTM, Faculty of Medicine and Dentistry, Palacky University Olomouc, Hnevotinska 5, 77900 Olomouc, the Czech Republic; [15]member of EATRIS ERIC- European Infrastructure for Translational Medicine; [16]Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, National Cancer Institute, Unit of Oncogenetics and Functional Oncogenomics, Via Gallini 2, 33081 Aviano (PN), Italy; [17]Perron Institute for Neurological and Translational Science, Verdun St, Nedlands, Western Australia, 6009, Australia; [18] the Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Murdoch, 6150, Western Australia; [19]Institute for Molecular Medicine Finland (FIMM),

HiLIFE, P.O.Box 20, FI-00014 University of Helsinki, Finland; [20]Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Molekylär Medicin, Box 1432, BMC, Uppsala, 75144 Sweden; [21]Department of Biological Sciences, Virginia Tech, Life Sciences 1, 970 Washington St., Blacksburg, VA 24061; [22]Bioinformatics and Computational Biology Core, National Heart Lung and Blood Institute, National Institutes of Health, 12 SOUTH DR. Bethesda MD 20892; [23]Sentieon Inc., 465 Fairchild Drive, Suite 135, Mountain View CA 94043; [24]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,45 Center Drive, Bethesda, Maryland 20894; [25]Immuneering Corporation, One Broadway 14th Fl Cambridge MA 02142 USA; [26]The Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, FDA, Silver Spring, Maryland; [27]Division of Antiviral Products, Office of Antimicrobial, Center for Drug Evaluation and Research, FDA, Silver Spring, Maryland; [28]The Center for Devices and Radiological Health, U.S. Food and Drug Administration, FDA, Silver Spring, Maryland; [29]Bioinformatics branch, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079; [30]Departemnt of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, 510182, P. R. China; [31]Department of Basic Science, Loma Linda University School of Medicine, Loma Linda, CA 92350

* Contributed equally

To whome correspondence should be addressed to: Wenming.Xiao@fda.hhs.gov, chwang@llu.edu, lemingshi@fudan.edu.cn and Huixiao.Hong@fda.hhs.gov

Abstract

We characterized two reference samples for NGS technologies: a human triple-negative breast cancer cell line and a matched normal cell line. Leveraging several whole-genome sequencing (WGS) platforms, multiple sequencing replicates, and orthogonal mutation detection bioinformatics pipelines, we minimized the potential biases from sequencing technologies, assays, and informatics. Thus, our "truth sets" were defined using evidence from 21 repeats of WGS runs with coverages ranging from 50X to 100X (a total of 140 billion reads). These "truth sets" present many relevant variants/mutations including 193 COSMIC mutations and 9,016 germline variants from the ClinVar database, nonsense mutations in *BRCA1/2* and missense mutations in *TP53* and *FGFR1.* Independent validation in three orthogonal experiments demonstrated a successful stress test of the truth set. We expect these reference materials and "truth sets" to facilitate assay development, qualification, validation, and proficiency testing. In addition, our methods can be extended to establish new fully characterized reference samples for the community.

Introduction

In oncology, accurate somatic mutation detection is essential to diagnose cancer, pinpoint

87 targeted therapies, predict survival, and identify resistance mutations. Despite the recent
88 explosion of technological advancements, many studies have reported difficulties in obtaining
89 consistent and concordant somatic mutation calls from individual platforms or pipelines[1–3], which
90 hampers clinical validation and advancement of these biomarkers.
91
92 As more sequencing technologies can detect clinically actionable somatic mutations for
93 oncology, the need grows stronger for benchmark samples with known "ground-truth" variants.
94 Such a publicly available sample set would allow platform and pipeline developers to quantify
95 accuracy of somatic mutation calls, study reproducibility across platforms or pipelines, perform
96 validation usig orthogonal techniques, and calibrate best practices of protocols and methods. The
97 FDA has released a guidance on the use of NGS technologies for *in vitro* diagnosis of suspected
98 germline diseases[4], in which well-characterized reference materials are recommended to
99 establish NGS test performance.
100
101 In the absence of well-characterized samples with somatic mutations, normal samples
102 such as the Platinum Genome[5], HapMap[6] cell lines, or Genome in a Bottle (GiaB) consortium
103 materials[7,8] are often used in clinical test development and validation of somatic applications.
104 Also there are some gene-specific reference samples available, such as KRAS in the WHO 1st
105 International Reference Panel[9], or from synthetic materials[10]. Such samples do not adequately
106 address cancer-specific quality metrics such as somatic mutation variant allele frequency (VAF),
107 heterogeneity, tumor mutation burden (TMB), etc. Therefore, cancer reference samples with an
108 abundance of well-defined genetic alterations characterized across the whole genome are highly
109 desirable and urgent needed.
110
111 Previous attempt has characterized a cancer cell line (from metastatic melanoma) that
112 inquired somatic mutations (SNV/indels) in exon regions only. Germline variants and somatic
113 mutations across the rest of the genome were not defined[11]. In addition, this dataset is
114 distributed under dbGAP-controlled access, limiting its accesibility and utility. In fact, a recent
115 landscape analysis of currently available somatic variant reference samples published by the
116 Medical Devices Innovation Consortium (MDIC) did not identify any reference mutation sets that
117 can be used to evaluate the somatic mutation calling accuracy on a whole-genome basis[12].
118
119 To fulfill this unmet need, we chose a pair of cell lines, HCC1395 (triple-negative breast
120 cancer) and HCC395BL (B lymphocytes) from the same donor, supplied by the American Type
121 Culture Collection (ATCC). These two specific cell lines were chosen because they are rich in
122 testable features (CNVs, SNVs, indels, SVs, and genome rearrangements[13]), and may have a
123 potential to serve as a long-term, publicly available, and renewable reference samples with
124 appropriate consent from donor. Using multiple next generation sequencing (NGS) platforms,
125 sequencing centers, and various bioinformatics analysis pipelines we profiled these tumor-
126 normal matching cell lines. Thus, we minimized biases that were specific to any platform,
127 sequencing center, or bioinformatic algorithm, to create a list of high-confidence mutation calls
128 across the whole genome, here called the "truth set." A subset of these calls was further
129 confirmed with orthogonal targeted sequencing and Whole Exome Sequencing (WES). We also
130 sequenced a series of titrations between HCC1395 and HCC1395BL genomic DNA (gDNA) to

confirm candidate somatic SNV/indels.

We defined truth sets containing somatic mutations and germline variants in a paired cell lines, HCC1395/HCC1395BL, with methods that minimized potential bias from library preparation, sequencing center, or bioinformatics pipeline. While the "truth set" germline variants in HCC1395BL can be used for benchmarking germline variant detection, the "truth set" somatic mutations in HCC1395 can be used for benchmarking cancer mutation detection with VAF as low as 5%. Many of variants and mutations have clinical implications. In the coding regions, a total of 193 somatic mutations are documented in the COSMIC database and 8 germline variants are annotated as pathogenic in the ClinVar database. Interestingly, there is a nonsense somatic mutation in the *BRCA2* gene and a nonsense germline variant in the *BRAC1* gene. Other hotspot somatic mutations are also observed in the *TP53* and *FGFR1* genes. Thus, we believe these paired cell lines may be highly valuable for those looking for reference samples to benchmark products in detection of mutations in these four genes.

Results

Massive data generated to characterize the reference samples

To provide reference samples for the community well into the future, a matched pair, HCC1395 and HCC1395BL was selected for profiling[14]. Previous studies of this triple negative breast cancer cell line have revealed the existence of many somatic structural and ploidy changes[13], which are confirmed by our cell karyotype and cytogenetic analysis (Suppl. Fig S1, S2). Several attempts have been made to identify SNVs and small indels[15–17]. Given that appropriate consent from the donor has been obtained for tumor HCC1395 and normal HCC1395BL for the purposes of genomic research, we sought to characterize this pair of cell lines as publicly available reference samples for the NGS community. In this manuscript, we focued our efforts on germline and somatic SNVs and indels. By performing numerous sequencing experiments with multiple platforms at different sequencing centers, we obtained high-confidence call sets of both somatic and germline SNVs and indels (Table 1). Larger structural variants and copy number analysis will be included in a separate manuscript that will discuss these fundings in greater detail.

Initial Determination of Somatic Mutation Call Set

High-confidence somatic SNVs and indels were obtained based primarily on 21 pairs of tumor-normal Whole Genome Sequencing (WGS) replicates from six sequencing centers; sequencing depth ranged from 50X to 100X (see manuscript DOI: 10.1101/626440). Each of the 21 tumor-normal sequencing replicates was aligned with BWA MEM[18], Bowtie2[19], and NovoAlign[20] to create 63 pairs of tumor-normal Binary Sequence Alignment/Map (BAM) files. Six mutation callers (MuTect2[21], SomaticSniper[22], VarDict[23], MuSE[24], Strelka2[25], and TNscope[26]) were applied to discover somatic mutation candidates for each pair of tumor-normal BAM files (Fig. 1). SomaticSeq[27] was then utilized to combine the call sets and classify the candidate mutation calls into "PASS", "REJECT", or "LowQual". Four confidence levels (HighConf, MedConf, LowConf, and Unclassified) were determined based on the cross-aligner and cross-sequencing center

reproducibility of each mutation call. HighConf and MedConf calls were grouped together as the "truth set" (also known as high-confidence somatic mutations). The call set in its entirety is referred to as the "super set" which includes low-confidence (LowConf) and likely false positive (Unclassified) calls. For low-VAF (Variant Allele Frequency) calls, a HiSeq data set with 300× coverage and a NovaSeq data set with 380× coverage were employed to rescue initial LowConf and Unclassified calls into the truth set. The details are described in the Methods.

A breakdown of the four confidence levels is displayed in Fig. 2a. In the truth set, HighConf calls consist of 94% of the SNVs and 79% of the indels. LowConf calls typically do not have enough "PASS" classifications across the 63 data sets to be included in the truth set. Variants calls labeled as Unclassified are not reproducible and likely false positives, with more "REJECT" classifications than "PASS". The vast majority of the calls in the super set are either HighConf or Unclassified. In other words, super set calls tend to be either highly reproducible or not at all reproducible.

In general, HighConf calls were classified as "PASS" in the vast majority of the data sets, with no variant read in the matched normal and high mapping quality scores. MedConf calls tended to be low-VAF (VAF $\lesssim$ 0.10) variants. Due to stochastic sampling of low frequency variants, MedConf calls were not reproduced as highly across different sequencing replicates as HighConf calls. LowConf calls (not a part of truth set) tended to have VAF near or below our detection limits (VAF $\lesssim$ 0.05). Distinguishing the LowConf calls with sequencing noise is challenging because they were not reproduced enough to be high-confidence calls (Fig. 2b).

Independent AmpliSeq confirmation of Call Set

We randomly selected 450 SNV and 21 indel calls of different confidence levels from the super set and performed PCR-based AmpliSeq with approximately 2000× depth for tumor and normal cells on an Illumina MiSeq sequencer. As we treated the AmpliSeq data set as a confirmatory experiment, simple rules were devised to determine whether a variant call was deemed positively confirmed, not confirmed, or uninterpretable based on the presence or absense of somatic mutation evidence in the AmpliSeq data. Overall, positively confirmed calls had at least 100 variant-supporting reads in the tumor but had no variant read in the normal sample, despite sequencing depths of 600× or more in the normal. Not confirmed calls either had no more variant-supporting read than the expected from base call errors, and/or had VAF≥10% in the normal cells. Uninterpretable calls did not satisfy the criteria for either positive or no validation, either because they did not have enough read depth (<50) or had fewer than 10 variant-supporting reads. (See Methods for details).

Both HighConf and MedConf SNV calls had very high coverage in validation and thus had impressive validation rates (99% and 92%) (Table 2). There were only three HighConf SNV calls that were not confirmed by AmpliSeq. Two of them had germline signals below the detection limit of 50× in the WGS, and the third one was likely an actual somatic mutation missed by AmpliSeq. There were only seven "positively confirmed" Unclassified SNV calls. Four of those seven were either a part of di-nucleotide change or had deletions within 1 bp of the call. The other three had low mapping quality scores (MQ), which drove the categorization of

"Unclassified". This result suggests that some of the "positively confirmed" Unclassified calls might be false positives after all, but it also exposes the limitations of our truth set with regard to complex variants and low mappability regions. LowConf and Unclassified calls (not part of the truth set) also had higher fractions of uninterpretable calls, which consist of low-coverage genomic positions or ambiguous variant signals. In addition, there were also 17 HighConf, 2 MedConf, 1 LowConf, and 1 Unclassified indel calls re-sequenced by AmpliSeq. The only not confirmed HighConf indel call was caused by a germline signal. The lone Unclassified indel call was not confirmed (we expect Unclassified calls to be not confirmed). For the inquisitive reader, these discrepant calls (i.e., not confirmed HighConf calls and confirmed Unclassified calls) are discussed in greater detail in the Supplementary Material.

The VAF calculated from the truth set correlated highly with the VAF calculated from AmpliSeq data set, especially for HighConf calls (Fig. 2c). On the other hand, almost all the data points at the bottom of the graph (i.e., VAF = 0 by AmpliSeq) are Unclassified calls (red). It suggests that despite high VAFs (from 21 WGS replicates) for some of the calls, they were categorized correctly as Unclassified (implying likely false positives). In addition, a large number of uninterpretable Unclassified calls (red X's) lying at the bottom suggest those were correctly labeled as Unclassified in addition to the not confirmed ones (open red circles). Moreover, some of the seven "positively confirmed" Unclassified calls had dubious supporting evidence. Taken together, these results suggest that the actual true positive rate for the Unclassified calls may be even lower than the validation rate (11%) we reported here. The indel equivalent is portrayed in Suppl. Fig S7a.

Orthogonal Confirmation of Call Set with WES on Ion Torrent

We have also sequenced the tumor-normal pair with Whole Exome Sequencing (WES) on the Ion Torrent S5 XL sequencer with the Agilent SureSelect All Exon + UTR v6 hybrid capture. The sequencing depths for the HCC1395 and HCC1395BL were 34× and 47×, respectively. Results from this Ion Torrent sequencing were leveraged to evaluate high-VAF SNV calls (Table 1 and 2). HighConf and MedConf SNV calls had high positive validation rates (99% and 89%). However, because the Ion Torrent sequencing was performed at much lower depth, nearly 50% of the calls were deemed uninterpretable (compared with 16% for AmpliSeq, despite having AmpliSeq custom target enriched for low-confidence calls vs. WES). The trend of higher uninterpretable fraction with lower confidence level calls was even more pronounced in this data set because the coverage was too low to confirm or invalidate many low-VAF calls. The validation rate for MedConf calls (predominantly low-VAF calls) may have suffered due to low coverage.

The VAF correlation between truth set and Ion Torrent WES (R=0.928) is lower than that between truth set and AmpliSeq (R=0.958), although the vast majority of the HighConf SNV calls in Ion Torrent data still stay within the 95% confidence interval area (Fig. 2d).

There are uninterpretable Unclassified calls (red X's) at the bottom for high-VAF calls, which is again highly suggestive that the true positive rate for Unclassified calls may be lower than the reported validation rate (25%) for Ion Torrent data as well. The indel equivalent is

263 included in Suppl. Fig. S7b.
264
265 **Independent Confirmation of Call Set with WES on HiSeq**
266
267 We used 14 HiSeq WES replicates from six sequencing centers to evaluate the
268 concordance between these data sets and the WGS data sets employed to construct the truth
269 set. While the WES data sets were not sequenced from orthogonal platforms, they provide
270 insights in terms of the reproducibility of our call sets in different library preparations. The scatter
271 plot between the super set derived VAF and medium HiSeq WES-derived VAF is presented in Fig.
272 2e. Almost all truth set (HighConf and MedConf calls) variants had consistent VAFs calculated
273 from both sources.
274
275 Again, simple rules were implemented for validation with the WES data as well (Table 2).
276 The validation rate for HighConf, MedConf, LowConf, and Unclassified SNV calls by WES were
277 100%, 98.4%, 93.1%, and 42.4%. These validation rates are higher than other methods because
278 these WES data were sequenced on the same platform and sequencing centers as those used to
279 build the truth set. Thus, the truth set variant calls are reproducible in WES, though these data
280 sets do not eliminate sequencing center or platform specific artifacts that may exist in both WGS
281 and WES data sets. The indel equivalent is the subject of Suppl. Fig. S7c.
282
283 **Validation with tumor content titration series**
284
285 To evaluate the effects of tumor purity, we pooled HCC1395 DNA with HCC1395BL DNA
286 at different ratios to create a range of admixtures representing tumor purity levels of 100%, 75%,
287 50%, 20%, 10%, 5%, and 0%. For each tumor DNA dilution point, we performed WGS on a HiSeq
288 4000 with 300 × total coverage by combining three repeated runs (manuscript DOI:
289 10.1101/626440). We plotted the VAF fitting score between the expected values based on the
290 super set vs. the observed values at each tumor fraction (Fig. 2f). For real somatic mutations,
291 their observed VAF should scale linearly with tumor fraction in the tumor-normal titration series.
292 In contrast, the observed VAF for sequencing artifacts or germline variants will not scale in this
293 fashion. Fig. 2f shows that the fitting scores for HighConf and MedConf calls are much higher than
294 LowConf and Unclassified calls across all VAF brackets, indicating that the HighConf and MedConf
295 calls contain far more real somatic mutations than LowConf and Unclassified calls. The formula
296 [Eq. 2] for the fitting score is described in the Methods.
297
298 **Definition and Confirmation of Germline SNVs/Indels in matched normal**
299
300 For the 21 WGS sequencing replicates of HCC1395BL (aligned with BWA MEM, Bowtie2,
301 and NovoAlign to create 63 BAM files) we employed four germline variant callers, i.e.,
302 FreeBayes[28], Real Time Genomics (RTG)[29], DeepVariant[30], and HaplotypeCaller[31], to discover
303 germline variants (SNV/indels). To consolidate all the calls, a generalized linear mixed model
304 (GLMM) was fit for each set of SNV calls which are sequenced at different centers on various
305 replicates, aligned by the three aligners, and discovered by the four callers. We estimated the
306 SNVs/indel call probability (SCP) averaged across four factors (sequencing center, sequencing

307 replicate, aligner, and caller), and examined the variance of SCP across these factors. The SNV
308 candidates considered were called at least four times (out of a maximum of 21x3x4=252 times)
309 by various combination of the four factors. The frequency histogram of the averaged SCPs
310 demonstrates a bimodal pattern (Fig. 3a). The vast majority of SNV calls (97%) had SCPs either
311 below 0.1 (57%) or above 0.9 (40%). Only a small minority of calls (3%) lie between 0.1 and 0.9.
312 This indicates when SNVs were repeatedly sequenced and called, only a small proportion of them
313 would be recurrently called as SNVs, and those recurrent calls were in fact highly recurrent.
314
315 Each of our germline SNV or indel calls had annotated SCP. See the Methods and Eq. 2 for
316 details. Suppl. Table S7 demonstrates that, of the highest-confidence calls (SCP=1, i.e, they were
317 called everywhere), the validation rates were approximately 99% for SNV and 98% for indels by
318 Illumina MiSeq, and 98% and 97% for Ion Torrent. Of the 11 SNV with SCP below 0.5, all were not
319 confirmed by MiSeq. Other calls had intermediate validation rates.
320
321 Figs 3b and 3c display that the vast majority of confirmed germline VAF was around 50%
322 and 100%. A considerable number of lower-confidence germline SNV calls clustered around 20%
323 VAF in non-exonic regions (Fig. 3b), with a large proportion of them being uninterpretable during
324 validation. Scatter plots for indels are qualitatively similar (Suppl. Fig. S13).
325
326 **SNV Functional Relevance and TMB Benchmarks**
327
328 Among the truth set somatic mutations, 186 COSMIC SNVs and 7 COSMIC indels are in
329 the coding region. One hotspot somatic mutation of particular biological significance is a *TP53*
330 *c.128G>A* (COSMIC99023, chr17:7675088 C>T, VAF>99%), which causes an amino acid change
331 *p.Arg43His* that leads to the inactivation of *TP53* tumor suppressive function[32]. In addition, there
332 is also a stop gain mutation in *BRCA2 c.4777G>T* (COSMIC13843, chr13:32339132 G>A), which
333 causes a nonsense at *p.Glu1593**, though it is only a heterozygous variant with VAF of 37.5%.
334 Furthermore, there is a missense mutation in *FGFR1 c.473C>T* (COSM1456963, chr8: 38428420
335 G>A, VAF>99%).
336
337 Of the over 3.5 million high-confidence germline variants discovered in HCC1395BL, 9,016
338 of them are in the ClinVar database. Most of them were annotated as “benign” or “like benign”;
339 however, 8 SNVs were annotated as “pathogenic” (Suppl. Table S9). One germline variant likely
340 to substantially increase the risk of an affected patient to develop breast cancer is a premature
341 stop gain in *BRCA1* (chr17:43057078, *c.5251C>A*, *p.Arg1751**, ClinVar #55480, OMIM Entry
342 #604370). The lifetime risk of breast cancer for carriers of this variant is 80 to 90%[33]. The
343 premature stop codon deactivates *BRCA1*’s function to repair DNA double-strand breaks. It is one
344 of the most common germline variants among breast cancer patients. HCC1395 has both *BRCA1*
345 and *TP53* completely inactivated, one from germline and one acquired somatically. The loss of
346 two critical tumor suppressor genes likely contributed to tumorigenesis. A full list of COSMIC
347 somatic mutations and ClinVar germline variants in the coding region is provided in Supplemental
348 File 2.
349
350 Tumor mutational burden (TMB) is defined as the number of non-synonymous somatic

351 variants per unit area of the genome, i.e., typically the number of non-synonymous mutations
352 per Mbps[34]. Recent literature increasingly has reported correlations between TMB and response
353 to anti-PD(L)-1 immunotherapy treatment[35]. The "gold standard" to measure TMB is to perform
354 tumor-normal WES and find the total number of non-synonymous mutations (all in the coding
355 regions). Due to the high cost and time required for WES, researchers are trying to infer TMB
356 with much smaller and less expensive targeted oncology panels. One way to increase the
357 statistical power of a much smaller panel is to measure all somatic mutations, including
358 synonymous mutations, which is expected to correlate highly with frequency of non-synonymous
359 mutations if we believe most somatic mutations, especially in high-TMB patients, occur more-or-
360 less randomly. We inferred TMB with various commercially available target panels. The
361 uncertainties of mutation rate (calculated as the 95% binomial confidence interval) inferred by
362 smaller oncology panels are quite large, so we advise caution when attempting to infer TMB from
363 targeted oncology panels (Suppl. Table S10).
364
365 **Defining Genome Callable Regions**
366
367 Accurate variant calling requires an abundance of high-quality reads aligned accurately to
368 the genomic coordinates in question. False positives are overwhelmingly enriched in genomic
369 regions where the alignments are challenging, base call qualities are low, and/or reported
370 coverage is far from the mean or median[36]. There are parts of the human genome that cannot be
371 covered by current technologies (Fig. 4a). To obtain the callable regions, we ran GATK CallableLoci
372 on each of the 63 HCC1395 and HCC1395BL BAM files to identify regions of low coverage (<10),
373 ultra-high coverage ($8\times$ the mean coverage of the sample), difficult to map ($MQ<20$), poor
374 reads (Base Quality Score $BQ<20$), or with N in the reference genome. We then created
375 consensus callable regions that we deemed callable for our truth set. A limitation of our callable
376 regions and our truth set is that they were defined and relied on short-read sequencing
377 technologies (i.e., Illumina sequencers), because currently only high-accuracy short-read
378 technologies are fit for somatic variant detection due to their low VAF. Variant calls outside the
379 consensus callable regions were labeled NonCallable in the super set and truth set to warn users
380 of these potential problems (details in Methods). NonCallable regions consisted of approximately
381 8% of the genome but contained over 34% of all Unclassified calls and 23% of all LowConf calls in
382 the super set (Suppl. Table S6).
383
384 The consensus callable regions consist of a total of 2.73 billion bps (Fig. 4b). In comparison
385 with GiaB NA12878 genome's more strictly defined high-confidence (HC) regions[7], 88% of our
386 consensus callable regions are in common with GiaB's HC regions. On the other hand, 98% of
387 GiaB's HC regions are a part of our conensus callable regions. Unlike GiaB's HC which exclude
388 regions with structural variations as well as regions where variant calls are inconsistent with
389 pidigree or regions with unexplained pipeline inconsnstencies, when there were disagreements
390 in a variant call from various sequencing data, we did not exclude the region. Instead, we
391 attempted to resolve these discrepancies. When there were nearby structural changes, we relied
392 on machine learning algorithms to resolve these challenging events. As a result, our conensus
393 callable regions included some difficult genomic regions, such as human leukocyte antigen (HLA)
394 and olfactory receptor genes which contain high homologus sequences. The confidence (or the

395 lack thereof) we hold for each variant call is annotated on a per call basis. We have demonstrated
396 some benchmarking results with different regions in the Supplementary, section 1.10.
397
398
399 ## Discussion
400
401 Through a community effort, we generated a high confidence somatic mutation call set
402 with limit of detection (LOD) at 5% VAF (Fig. 2b). To employ as an accuracy benchmark, we
403 recommend considering the variant calls labeled with both HighConf and MedConf as true
404 positives. These true positive variants can be used to assess sensitivity, i.e., the fraction of those
405 variants detected by a pipeline. On the other hand, variant calls labeled as Unclassified plus any
406 unspecified genomic coordinates are likely false positives. LowConf calls could not be confidently
407 determined here and should be blacklisted for current accuracy evaluation. LowConf calls had
408 validation rates around 50%, and often had VAF below our 50× depth detection limit. They
409 represent opportunities for future work to ascertain their actual somatic status.
410
411 The confidence level of each variant call was determined by the "PASS" classifications
412 provided by SomaticSeq across different sequencing centers with different aligners (see
413 Methods). If a variant was not detected by any caller in a data set, it was considered "Missing" in
414 that data set, which is common for low-VAF calls due to stochastic sampling. For most calls,
415 however, they either had "PASS" classifications or "REJECT or Missing" classifications, but not
416 both. Few variant candidates had a large number of "PASS *and* REJECT" classifications (Suppl. Fig.
417 S6a). HighConf calls had many "PASS" classifications, very few "REJECT" classifications, and a full
418 range of VAFs. MedConf calls had fewer "PASS" calls (still high), still very few "REJECT"
419 classifications, but were mostly low-VAF, which explains the lower number of "PASS" calls.
420 LowConf calls had even fewer "PASS" calls than MedConf though they overlaped significantly,
421 and also a low number of "REJECT" classifications. LowConf calls tended to have even lower VAF
422 than MedConf, around or below our detection limit (Fig. 2b). Only Unclassified calls suffered a
423 significant number of "REJECT" classifications, and they also displayed a full range of VAF. The
424 performance of Unclassified calls indicated that SomaticSeq labeled them "REJECT" due to poor
425 mapping, poor alignment, germline risk, or causes other than lack of variant reads. HighConf and
426 Unclassified calls are far apart in all of the metrics describedabove.
427
428 Variant re-sequencing with AmpliSeq (Suppl. Fig. S6c) pointed to a high validation rate for
429 HighConf and MedConf calls. Suppl. Fig. S6c also contains a cluster of Unclassified and LowConf
430 calls in the middle of the XY plane, representing calls with some conflicts (i.e., large number of
431 "PASS" and "REJECT" calls).
432
433 Each time a human cell divides, somatic mutations could be introduced by replication
434 errors. Somatic mutations can occur much more frequently in cancer cells with malfunctioning
435 DNA repair systems. It is not feasible to detect extremely low-VAF somatic mutations because
436 they may appear in few tumor cells. Our "truth set" for somatic mutation was built upon WGS
437 with 50×-100X coverages, and thus it was designed to detect somatic mutations limited to 5% of
438 VAF. Variants with low-VAF (≤12%) were cross-referenced with two data sets with depths over

300× to ascertain their presence. While we do not expect our truth set to be 100% accurate or 100% comprehensive, the AmpliSeq and Ion Torrent data sets demonstrated combined 99% and 91% validation rates for HighConf and MedConf SNV calls, respectively. AmpliSeq also showed a 94% validation rate for HighConf indel calls. VAF of 5% represents the lower detection limit of the first release of the somatic mutation truth set, even though there are many true mutations with VAF under that threshold. We recommend that if using this truth set as a benchmark, novel variant calls (i.e., variants calls not present in our super set) with VAF<5% should be blacklisted from the accuracy calculations because we cannot confidently determine their status. Due to losses of chr6p, chr16q, and chrX in HCC1395BL (Suppl. Fig S1, S2), somatic mutations in these regions were excluded.

For the first time, tumor-normal paired "reference samples" with a whole-genome characterized somatic mutation and germline "truth sets" are available to the community. Our samples, data sets, and the list of known somatic mutations can serve as a public resource for evaluating NGS platforms and pipelines. The massive and diverse amount of sequencing data generated from multiple platforms at multiple sequencing centers can help tool developers to create and validate new algorithms and to build more accurate artificial intelligence (AI) models for somatic mutation detection. The reference samples and call set presented here can help in assay development, qualification, validation, and proficiency testing. Such community defined tumor-normal paired reference samples can be helpful in quality assessment by clinical laboratories engaged in NGS, data exchange between laboratories, characterization of gene therapy products, and premarket review of NGS-based products. Furthermore, the methodology used in this study can be extended to establish truth sets for additional cancer reference samples. Other reference sample efforts may be able to build on the data sets we established or consider using these samples as a genomic background for other reference samples.

Methods

See Online Methods

Acknowledgements

We thank Justin Zook of the National Institute of Standard Technology for advice in establishing reference samples and truth set; Sivakumar Gowrisankar of Novartis, Susan Chacko of the Center for Information Technology, the National Institute of Health for their assistance with data transfer; Dr. Jun Ye of Sentieon for providing the Sentieon software package. We also appreciate Dr. Laufey Amundadottir of the Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), for the sponsorship and the usage of the NIH Biowulf cluster; Drs Reena Phillip, Yun-Fu Hu, Sharon Liang, and You Li of the Center for Devices and Radiological Health, U.S. Food and Drug Administration, for their advices on study design and manuscript writing; and Seven Bridges for providing storage and computational support on the Cancer Genomic Cloud (CGC). The CGC has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No.

HHSN261201500003I. Chunlin Xiao and Steve Sherry were supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. This work also used the computational resources of the NIH Biowulf cluster (http://hpc.nih.gov). Original data was also backed up on the servers provided by Center for Biomedical Informatics and Information Technology (CBIIT), NCI. We would also like to thank the partially support from the Ardmore Institute of Health (AIH) grant (2150141) and Dr. Charles A. Sims' gift to Loma Linda University (LLU) Center for Genomics. The LLU Center for Genomics is partially supported by AIH grant (2150141) and Charles A. Sims' gift.

Disclaimer

This is a research study, not intended to guide clinical applications. The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as endorsement.

Data availability

All raw data (FASTQ files) are available on NCBI's SRA database (SRP162370). The truth set for somatic mutations in HCC1395, VCF files derived from individual WES and WGS runs, and source codes are available on NCBI's ftp site (ftp://ftp-trace.ncbi.nlm.nih.gov/seqc/ftp/Somatic_Mutation_WG/). Some alignment files (BAM) are also available on Seven Bridges' s Cancer Genomics Cloud (CGC) platform.

Author contributions

Study conceived and designed by: W.X., C.W., L.S., H.H., E.D., Z.T., B.N., W.T., R.J.

Biosample preparation: L.K., K.L., M. M., T.H., W.C.

NGS library preparation and sequencing: W.C., Z.C., S.S., K.I., H.J., E.J., G.S., S.S., J.S., P.K., J.B., B.T., V.P., M.S., T.H., E.P., R.K., J.D., P.V., R.M., D.G., S.K., E.R., A.S., J.N., U.L., J.W., J.L., P.PH.

Data analysis: L.T.F., W.X., B.Z., Y.Z., Z.Y., C.L., O.A., L.S., J.L., T.S., K.T., D.M., C.N., M.C., S.M.S., M.M., Y.G., L.Y., H.L., M.P., Z.L.

Data management: W.X., C.X., S. S.

Manuscript writing: L.T.F., W.X., R.K., M.M., C.X., S.S.

Project management: W.X.

References

1. Hofmann, A. L. *et al.* Detailed simulation of cancer exome sequencing data reveals differences

and common limitations of variant callers. *BMC Bioinformatics* **18**, 8 (2017).

2. Krøigård, A. B., Thomassen, M., Lænkholm, A.-V., Kruse, T. A. & Larsen, M. J. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE* **11**, e0151664 (2016).
3. Shi, W. *et al.* Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. *Cell Rep.* **25**, 1446–1457 (2018).
4. Tezak, Z. & Berger, A. Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing (NGS) – Based In Vitro Diagnostics (IVDs) Intended to Aid in the Diagnosis of Suspected Germline Diseases. (2018).
5. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).
6. The International HapMap Project. *Nature* **426**, 789 (2003).
7. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
8. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* 1 (2019). doi:10.1038/s41587-019-0074-6
9. National Institute for Biological Standards and Control. *WHO Reference Panel 1st International Reference Panel for genomic KRAS codons 12 and 13 mutations NIBSC code: 16/250*. (World Health Organization).
10. Oncospan Reference Standard HD827. Available at: https://www.horizondiscovery.com/reference-standards/type/oncospan. (Accessed: 17th April 2019)

11. Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. *Sci. Rep.* **6**, 24607 (2016).

12. Zehnbauer, B. *et al.* MDIC SRS Report: Somatic Variant Reference Samples for NGS. (2019).

13. Popova, T. *et al.* Ploidy and Large-Scale Genomic Instability Consistently Identify Basal-like Breast Carcinomas with BRCA1/2 Inactivation. *Cancer Res.* **72**, 5454–5462 (2012).

14. Gazdar, A. F. *et al.* Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int. J. Cancer* **78**, 766–774 (1998).

15. Kalyana-Sundaram, S. *et al.* Gene Fusions Associated with Recurrent Amplicons Represent a Class of Passenger Aberrations in Breast Cancer. *Neoplasia* **14(8)**, 702–708 (2012).

16. Zhang, J. *et al.* INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Res.* **26**, 108–118 (2016).

17. Robinson, D. R. *et al.* Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat. Med.* **17**, 1646–1651 (2011).

18. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

19. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

20. NovoCraft Technologies Sdn Bhd. NovoAlign. Available at: http://www.novocraft.com/products/novoalign/.

21. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

22. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome

sequencing data. *Bioinformatics* **28**, 311–317 (2012).

23. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
24. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
25. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591 (2018).
26. Freed, D., Pan, R. & Aldana, R. TNscope: Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering. *bioRxiv* 250647 (2018). doi:10.1101/250647
27. Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.* **16**, 197 (2015).
28. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* (2012).
29. Real Time Genomics (RTG) Variant Caller. Available at: https://www.realtimegenomics.com/.
30. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
31. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2017). doi:10.1101/201178
32. Soussi, T., Leroy, B. & Taschner, P. E. M. Recommendations for Analyzing and Reporting TP53 Gene Variants in the High-Throughput Sequencing Era. *Hum. Mutat.* **35**, 766–778 (2014).

33. OMIM Clinical Synopsis - #604370 - BREAST-OVARIAN CANCER, FAMILIAL, SUSCEPTIBILITY TO, 1; BROVCA1. Available at: https://www.omim.org/clinicalSynopsis/604370. (Accessed: 22nd March 2019)

34. Meléndez, B. *et al.* Methods of measurement for tumor mutational burden in tumor tissue. *Transl. Lung Cancer Res.* **7**, 661–667 (2018).

35. Goodman, A. M. *et al.* Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).

36. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

NGS technologies		platforms	# of reads (coverage)	
			HCC1395	HCC1395BL
Initial	WGS	HiSeq	57 billion (2,800X)	57 billion (2,800X)
		NovaSeq	13 billion (650X)	13 billion (650X)
		10X Genomics	20 billion (1,000X)	20 billion (1,000X)
		PacBio	20 million (50X)	20 million (50X)
Validation	WGS-tumor content	HiSeq	7.6 billion (380X)	7.6 billion (380X)
	WES	HiSeq	5 billion (12,500X)	5 billion (12,500X)
		Ion Torrent	67 million (34X)	82 million (47X)
	AmpliSeq	MiSeq	3.3 million (2000X)	3.3 million (2000X)

Table 1. Massive data from multiple NGS platforms was obtained to derive and confirm germline and somatic variants in HCC1395 and HCC1395BL

Validation Platform	Variant Type	Category	Total Number	Fraction Interpretable	Validation Rate (Interpretable)	Validation Rate (Total)
AmpliSeq Deep Sequencing	SNV	HighConf	247	(237/247) 96.0%	**(234/237) 98.7%**	94.7%
		MedConf	40	(37/40) 92.5%	**(34/37) 91.9%**	85.0%
		LowConf	58	(41/58) 70.7%	**(22/41) 53.7%**	37.9%
		Unclassified	105	(62/105) 59.0%	**(7/62) 11.3%**	6.7%
	INDEL	HighConf	17	(17/17) 100.0%	**(16/17) 94.1%**	94.1%
		MedConf	2	(2/2) 100.0%	**(2/2) 100.0%**	100.0%
		LowConf	1	(0/1) 0.0%	**(0/0) NA**	nan%

		Unclassified	1	(1/1) 100.0%	**(0/1) 0.0%**	0.0%
Ion Torrent WES	**SNV**	**HighConf**	703	(629/703) 89.5%	**(623/629) 99.0%**	88.6%
		MedConf	43	(27/43) 62.8%	**(24/27) 88.9%**	55.8%
		LowConf	134	(39/134) 29.1%	**(28/39) 71.8%**	20.9%
		Unclassified	802	(155/802) 19.3%	**(39/155) 25.2%**	4.9%
	INDEL	**HighConf**	31	(25/31) 80.6%	**(22/25) 88.0%**	71.0%
		MedConf	15	(7/15) 46.7%	**(6/7) 85.7%**	40.0%
		LowConf	8	(0/8) 0.0%	**(0/0) NA**	nan%
		Unclassified	36	(8/36) 22.2%	**(6/8) 75.0%**	16.7%
WES	**SNV**	**HighConf**	1074	(1068/1074) 99.4%	**(1068/1068) 100%**	99.4%
		MedConf	64	(63/64) 98.4%	**(62/63) 98.4%**	96.9%
		LowConf	197	(144/197) 73.1%	**(134/144) 93.1%**	68.0%
		Unclassified	1218	(436/1218) 35.8%	**(184/436) 42.4%**	15.1%
	INDEL	**HighConf**	45	(43/45) 95.6%	**(43/43) 100%**	95.6%
		MedConf	17	(17/17) 100.0%	**(17/17) 100%**	100.0%
		LowConf	13	(10/13) 76.9%	**(9/10) 90%**	69.2%
		Unclassified	54	(19/54) 35.2%	**(14/19) 73.7%**	25.9%

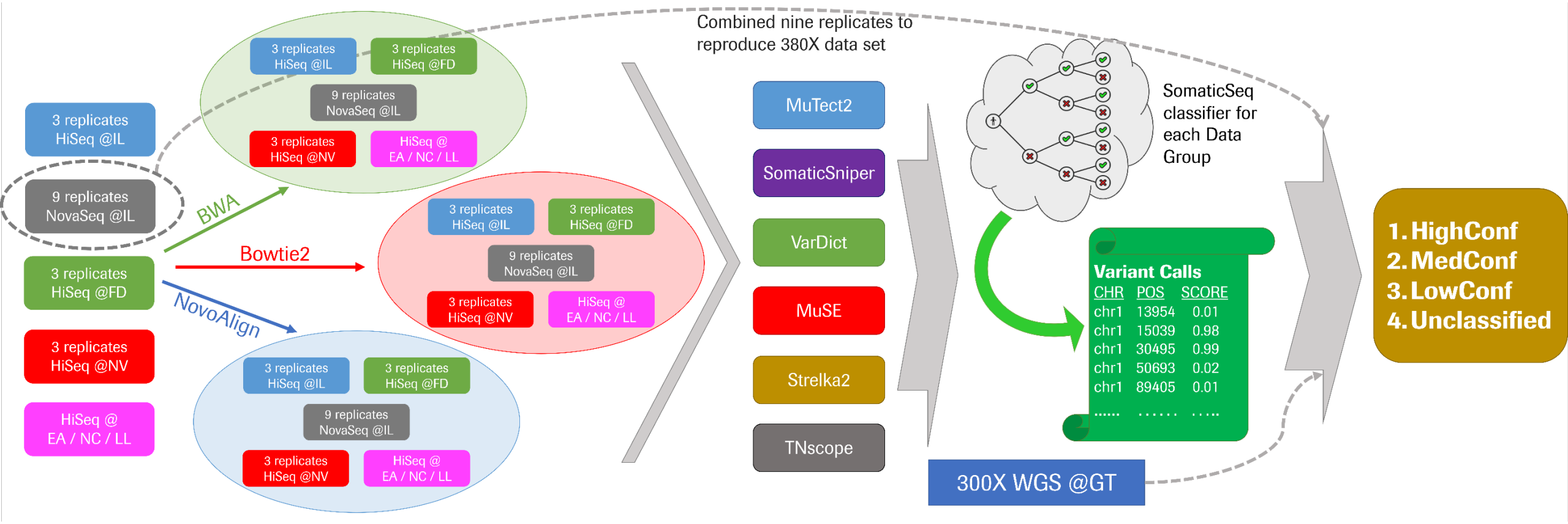
Table 2: Validation of SNVs of different confidence levels by three different methods

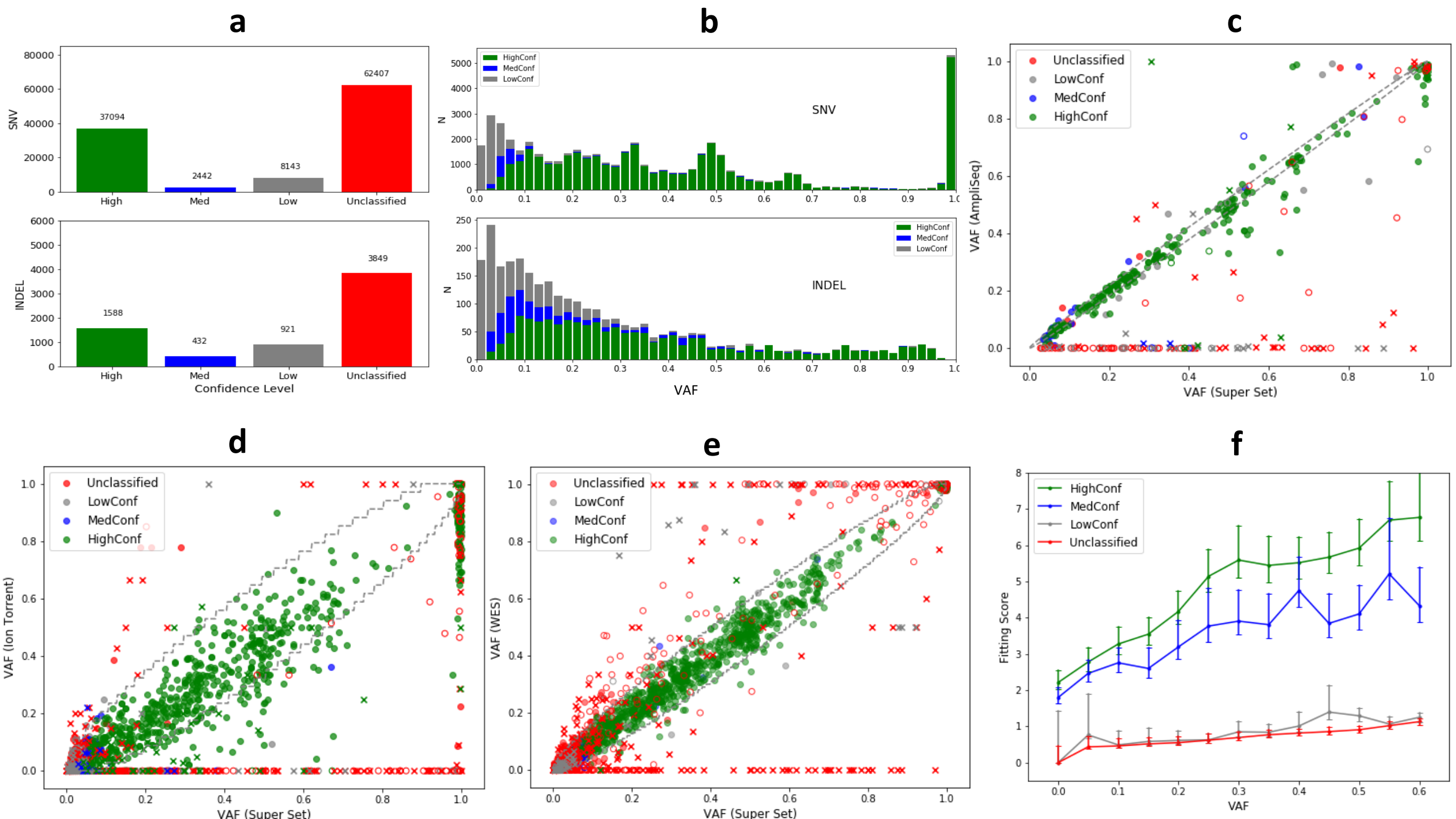
Figure legends

Figure 1: Schematic of the bioinformatics pipeline used to define the confidence levels of the super set and truth set (see Online Methods for detail)

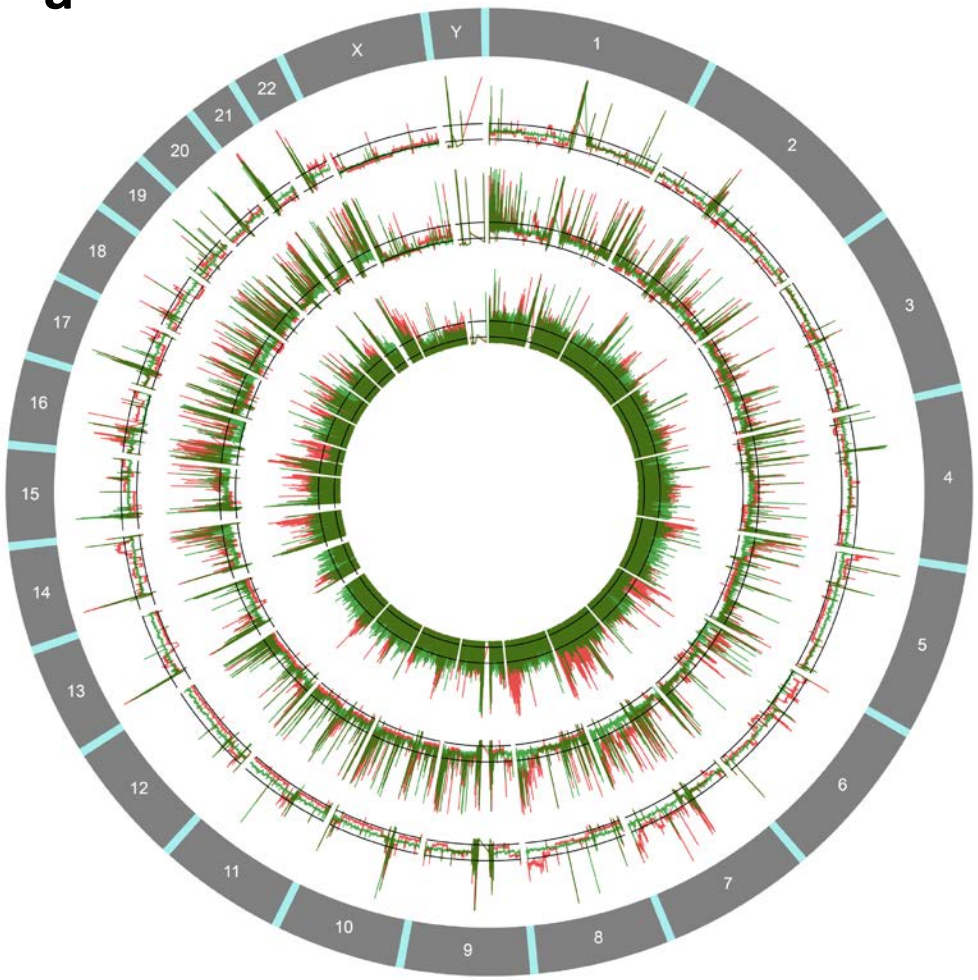
Figure 2: Initial definition of somatic mutation truth set and subsequent validation. (a) A breakdown of the four confidence levels in the super set. (b) Histograms of VAF for SNVs (top) and Indels (bottom) calls. (c) Validation of initial definition of somatic mutation truth set with AmpliSeq. Solid circles are variant calls that were positively confirmed. Open circles are variants that were not confirmed. X's are when validation data were deemed uninterpretable due to low depth or unclear signal. The dashed lines at the diagonal represent the 95% binomial confidence-interval of observed VAF given the actual VAF, calculated based on 2000× depth for AmpliSeq. The figure shows very high correlation between VAF estimated from super set data and validation data for HighConf calls (R=0.958). Many Unclassified data points lie at the bottom, implying that those calls were not real mutations despite the large number of apparent variant-supporting reads in the super set data. X-axis: VAF calculated from the super set. Y-axis: VAF calculated from AmpliSeq data. (d) Validation of the initial definition of the somatic mutation truth set with Ion Torrent WES. The 95% binomial confidence-interval dash lines were calculated based on 34× depth for Ion Torrent. R=0.928 for HighConf calls. (e) Validation of initial definition of somatic mutation truth set with 12 repeats of WES on the HiSeq platform. Y-axis: median VAF calculated based on 12 HiSeq WES replicates. The 95% binomial confidence-interval dashed lines were calculated based on 150× depth for HiSeq WES. R=0.992 for HighConf calls. (f) Average tumor purity fitting scores for the VAF of each SNV across the four different confidence levels vs. the observed VAF in the tumor-normal titration series. The formula for fitting scores is described in

633 Eq. 1 in the Online Methods.
634
635 **Figure 3**: Initial definition of germline variants and validation. (a) Histogram of SNV call probability
636 for germline variants identified by four callers from 63 BAM files. (b) VAF scatter plot of germline
637 SNVs by the truth set and AmpliSeq. R=0.986 for SCP=1 calls. (c) VAF scatter plot of germline SNVs
638 by the truth set and Ion Torrent WES. R=0.758 for SCP=1 calls.
639
640 **Figure 4**: Genome coverage and high-confidence regions on reference genome GRCh38. a)
641 Genome coverage comparison between three technologies. Inner track: PacBio. Middle track:
642 10X Genomics. Outer track: Illumina. Red line: HCC1395. Green line: HCC1395BL. b) Genome
643 regions coverage by Illumina short reads in comparison to NA12878. Inner track: NA12878.
644 Middle track: the callable regions in HCC1395 and HCC1395BL. Outer track: gene density
645

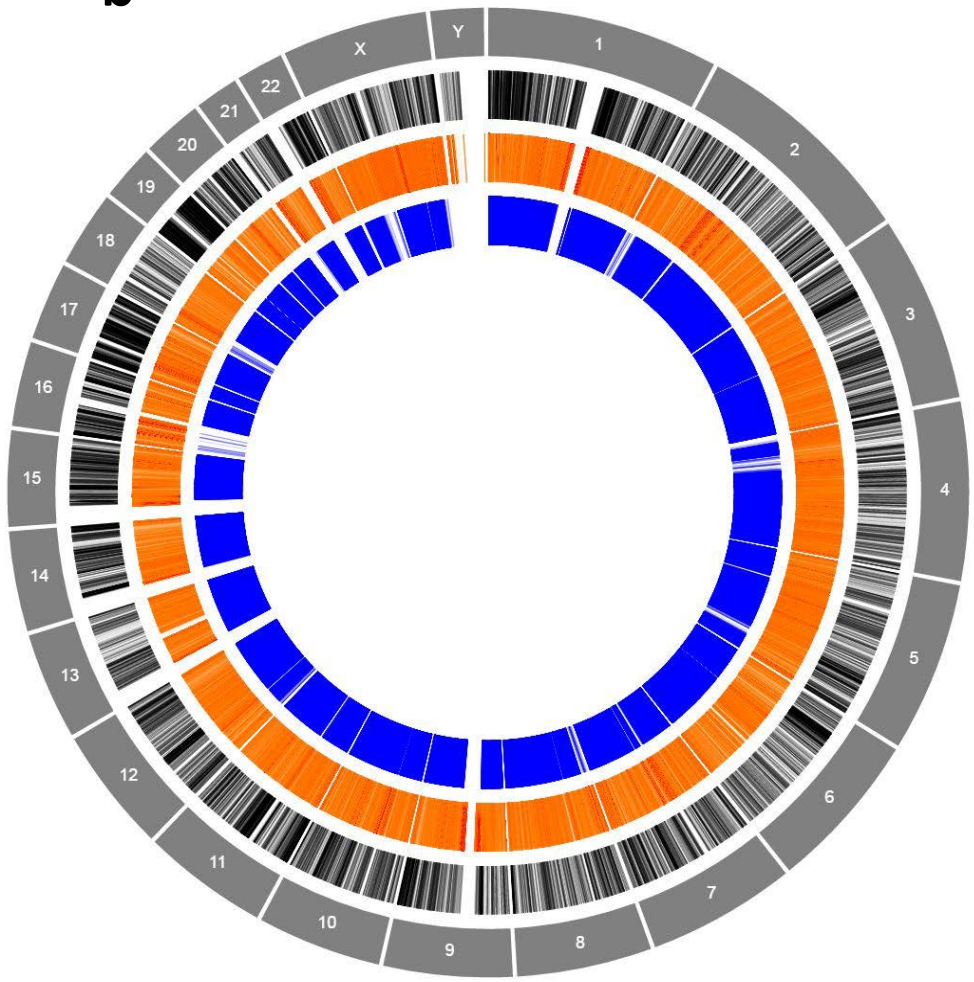

Combined nine replicates to reproduce 380X data set
3 replicates HiSeq @IL
9 replicates NovaSeq @IL
3 replicates HiSeq @FD
3 replicates HiSeq @NV
HiSeq @ EA / NC / LL
BWA
Bowtie2
NovoAlign
MuTect2
SomaticSniper
VarDict
MuSE
Strelka2
TNscope
SomaticSeq classifier for each Data Group
Variant Calls
CHR POS SCORE
chr1 13954 0.01
chr1 15039 0.98
chr1 30495 0.99
chr1 50693 0.02
chr1 89405 0.01
300X WGS @GT
1. HighConf
2. MedConf
3. LowConf
4. Unclassified

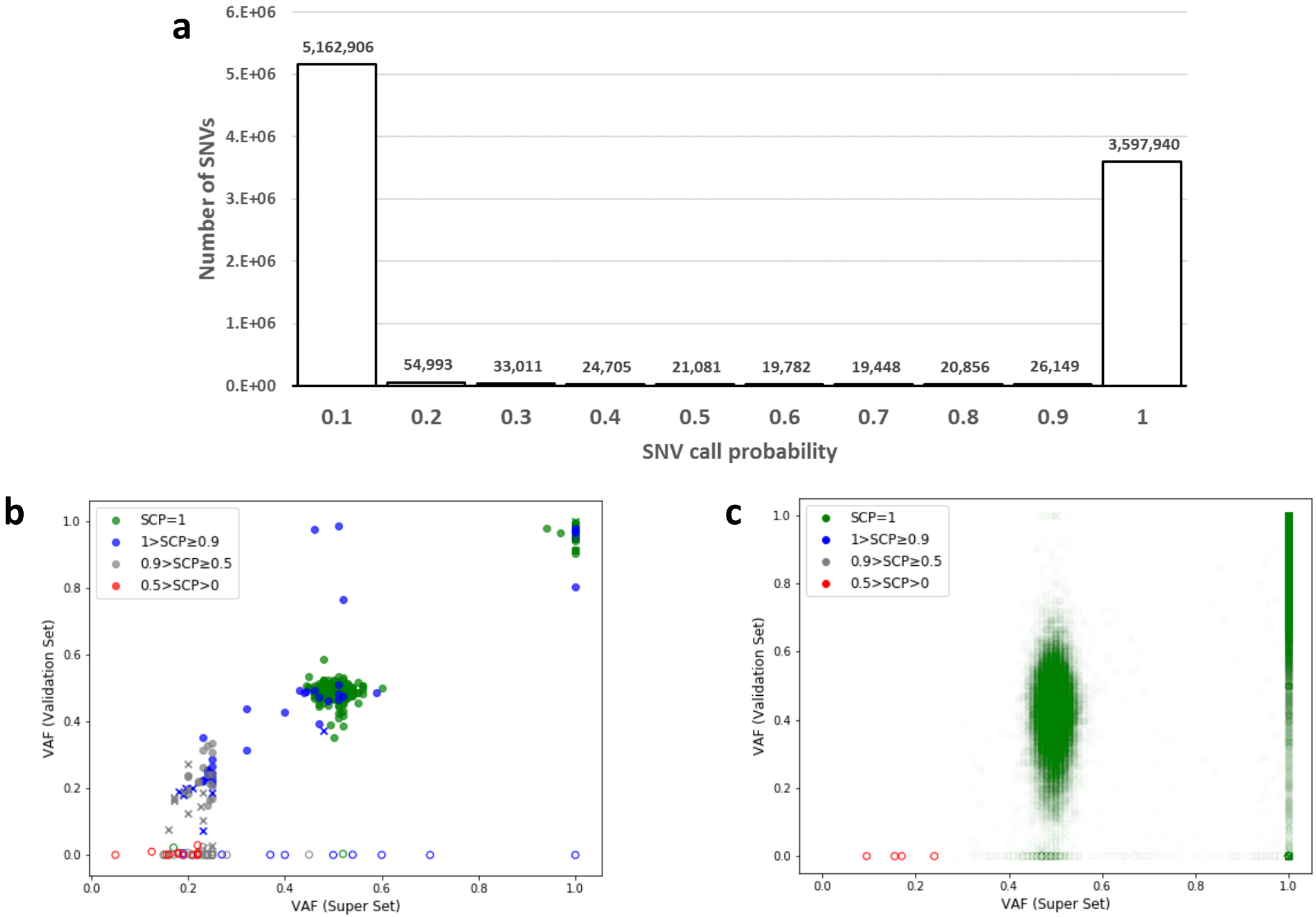

a
SNV
INDEL
37094
2442
8143
62407
1588
432
921
3849
High
Med
Low
Unclassified
Confidence Level
b
HighConf
MedConf
LowConf
SNV
INDEL
N
VAF
c
Unclassified
LowConf
MedConf
HighConf
VAF (AmpliSeq)
VAF (Super Set)
d
VAF (Ion Torrent)
VAF (Super Set)
e
VAF (WES)
VAF (Super Set)
f
Fitting Score
VAF

a


X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

b


X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


a
6.E+06
5.E+06
4.E+06
3.E+06
2.E+06
1.E+06
0.E+00
Number of SNVs
5,162,906
54,993
33,011
24,705
21,081
19,782
19,448
20,856
26,149
3,597,940
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
SNV call probability
b
SCP=1
1>SCP≥0.9
0.9>SCP≥0.5
0.5>SCP>0
VAF (Validation Set)
VAF (Super Set)
c
SCP=1
1>SCP≥0.9
0.9>SCP≥0.5
0.5>SCP>0
VAF (Validation Set)
VAF (Super Set)