
Supplementary information

**Establishing community reference
samples, data and call sets for
benchmarking cancer mutation detection
using whole-genome sequencing**

In the format provided by the
authors and unedited

Establishing reference community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing

1 Somatic Mutations

1.1 Accuracy and Sensitivity of the Machine Learning Frameworks

The high classification accuracies of both SomaticSeq¹ and NeuSomatic are demonstrated in **Supplementary Fig. 1**. SomaticSeq is an ensemble caller that relies on the six callers (MuTect2², SomaticSniper³, VarDict⁴, MuSE⁵, Strelka2⁶, and TNscope⁷) to produce a combined call set, from which true positives and false positives are distinguished. NeuSomatic is a deep convolutional neural network (CNN) based classification algorithm that has two different ways of generating mutation candidates: 1) the Ensemble mode that relies on the same six callers as SomaticSeq (i.e. NeuSomatic-E), and 2) the Standalone mode that relies on its own internal alignment scanner (i.e. NeuSomatic-S). The sensitivity of SomaticSeq and NeuSomatic-E are limited by the combined call set of the six callers. **Supplementary Fig. 2** shows the VAF of variants that were simulated, detected, and classified correctly. Most undetected mutations have low-VAF. The maximum base call quality value is a Phred score of 40, which translates to an error rate of 1/10,000. While that error rate is impressively low, it is still two orders of magnitude greater than the typical somatic mutation rate in cancer⁸. Thus, even under the most ideal situation and with the best sequencing data, variant calling requires a minimum of two variant reads. Given the ~50X average sequencing depth in this study, that translates to a detection limit of roughly 5% VAF. Thus, only a fraction of simulated variants around or below 5% were detected in each data set, and the observed sensitivity in each data set increased with VAF as expected. The sensitivity of variants in the $2.0\% < \text{VAF} \leq 4.0\%$ range was merely 8.6%. The sensitivity increased to 40.3% for variants of $4.0\% < \text{VAF} \leq 6.0\%$, and only reached 90.9% for variants of $10.0\% < \text{VAF} \leq 12.0\%$. Thus, at this coverage level, low-VAF variants would only be called occasionally, resulting in *initially* low confidence classifications.

1.2 SomaticSeq and NeuSomatic Classifications

SomaticSeq and NeuSomatic machine-learning classifiers were used to score each variant candidate in each data set (**Supplementary Fig. 3a**). For NeuSomatic, both NeuSomatic-E and NeuSomatic-S classifiers were trained. NeuSomatic-E employs the same ensemble of callers to generate variant candidates like SomaticSeq does, whereas NeuSomatic-S scans through the alignments with an internal algorithm to generate variant candidates. NeuSomatic-E has higher accuracy with greater need for computational resources. **Supplementary Fig. 3b and c** show that the breakdown of SomaticSeq and NeuSomatic scores for each variant call in real data is qualitatively similar to those found during cross-validations of simulated mutation data, i.e., vast majority of calls are clear PASS or REJECT, with only a small fraction of calls with intermediate scores (**Supplementary Fig. 1**).

More than 90% of HighConf calls were classified as PASS by SomaticSeq in at least 54 of the 63 data sets (Average 60.5 PASSes and 0.15 REJECT. **Supplementary Fig. 3d**). MedConf calls were classified as PASS by SomaticSeq in fewer data sets (average 27.4 PASSes) but also had limited REJECT classifications by SomaticSeq (average 0.8 REJECT). LowConf calls typically were

classified as PASS inconsistently across the 63 data sets (average 8.2 PASSes and 1.7 REJECTs). Unclassified calls were not reproducible and likely false positives, with more REJECT classifications than PASS (average 1.8 PASSes and 3.8 REJECTs). The majority (82%) of the calls in the all-inclusive set were either HighConf or Unclassified. In other words, all-inclusive set calls were either highly reproducible or extremely irreproducible (**Supplementary Fig. 3d**). The challenge was to confirm MedConf calls and rescue LowConf calls to maximize our reference call set while preventing false positives from mixing in.

The 3-D scatter plots for SomaticSeq and NeuSomatic-E classifications are shown in **Extended Data Fig. 1**. These figures demonstrate the consistency between SomaticSeq and NeuSomatic classifications across all data sets, despite the fact that SomaticSeq and NeuSomatic use totally different machine learning algorithms. For indels, upon deeper investigation, many indel calls that had classification conflicts were long deletions (e.g., ≥ 12 -bp deletions). These long deletions were difficult to detect and classify on a per sample basis due to the lack of reads covering the entirety of the deleted bases. In addition, the presence of soft-clipped reads spanning the deleted regions made them seem low quality. However, when multiple reads spanning long deletions were present, they almost always appeared to be true mutations. Thus, we promoted ≥ 12 -bp deletion calls that were initially labelled as MedConf into HighConf. We have manually inspected over 50 such promotions, and the promotions seemed well justified in each and every case (**Supplementary Data File 3**).

1.3 High-confidence regions

The consensus callable regions consist of a total of 2.73 billion base pairs (bps) (**Extended Data Fig. 2b**). In comparison with GIAB NA12878 genome's high-confidence (HC) regions⁹, 88% of our consensus callable regions are in common with GIAB's HC regions. On the other hand, 98% of GIAB's HC regions are a part of our consensus callable regions. GIAB's HC excluded regions with structural variations as well as regions where variant calls were inconsistent with pedigree or regions with unexplained pipeline inconsistencies. Similarly, we further removed three lost chromosome arms (chr6p, chr16q, and chrX) from the consensus callable regions because somatic and germline variants cannot be distinguished, as well as 20-bp regions spanning LowConf and Unclassified somatic mutation calls where we could not confidently call somatic variants. What has remained is now the high-confidence regions which make up >82% of the entire genome. We included the SEQC-II high confidence regions accompanying our reference call set. To use the reference call sets, we recommend limiting the evaluation of calling accuracy to the SEQC-II high-confidence regions to ensure maximum reliability. An application of the reference call set is shown in Supplementary Material Sec. 1.9.

1.4 Variant allele frequencies

In **Extended Data Fig. 8a**, it is apparent that there are VAF peaks for SNVs (VAF estimates for indels, especially long indels, are often underestimated because reads with indels near the ends are often soft-clipped. They do not have indel aligned in them, and thus not counted). Most of these VAF peaks correspond to different copy number states, but peaks around and below 15% can be explained by tumor heterogeneity. While HCC1395BL is a relatively unremarkable genome other than the loss of three chromosomal arms, HCC1395 is a highly rearranged near-

triploid genome with 66 chromosomes. The VAF of germline variants across the genome mostly resemble a typical genome in HCC1395BL, with the exception of the aforementioned three arm losses where only VAF $\approx 100\%$ are observed (**Extended Data Fig. 7a**). **Extended Data Fig. 7b**, on the other hand, shows the VAF of the same germline positions in the HCC1395 tumor cell line. The change in VAF between the same genomic coordinates revealed copy number changes between the normal and tumor genomes. **Extended Data Fig. 7c** shows the VAF of somatic SNVs, which again revealed almost discrete VAFs in different genomic regions with different copy number states; this is consistent with the different VAF peaks shown in **Extended Data Fig. 8b**. In addition, the VAF breaks are also apparent in **Extended Data Fig. 7b and c**. Each implies a structural break in the HCC1395 genome.

We examined the distribution of VAFs and chromosome locations of mutations in all four categories: HighConf, MedConf, LowConf, and Unclassified. While the majority of VAFs in HighConf were between 0.05 and 1, VAFs in MedConf lied between 0.02 and 0.1. Most VAFs in LowConf were less than 0.05. Although most of the VAFs in Unclassified were lower than 0.05, VAFs in Unclassified spanned a full spectrum between 0.01 to 1 (**Extended Data Fig. 1a and b**). Unclassified mutations were most likely artifacts called by a single analysis pipeline or in a single data set and thus were excluded from the reference call set. In addition, over 33% of the mutations in the LowConf and Unclassified groups were outside the consensus callable regions (**Supplementary Table 4**). Therefore, mutations would be classified in the LowConf or Unclassified groups due to: not being called in most of 21 libraries, not being called in every aligner (difficult-to-map regions where the three aligners disagreed), and/or pipeline specific artifacts. The probability of a mutation with $\sim 5\%$ VAF being missed by our all-inclusive set is minuscule (estimated as 2.5×10^{-6} , see Supplementary Materials Sec. 1.7). Hence, calling specificity could be evaluated in the SEQC-II high-confidence regions. Specifically, a novel call in the high-confidence region with VAF higher than 5% should be regarded as a false positive.

The majority of somatic indels were insertions or deletions of a single base but ranged from 58-bp for deletions to 14-bp for insertions in the truth set (**Extended Data Fig. 3d**).

1.5 Orthogonal Validations

The overall validation rates described in the main text were calculated based on the number of SNV and indel calls that were validated by at least one of the three validation experiments. While AmpliSeq enrichment was based on randomly selected positions in our super set and only had 7 overlapping SNVs with WES, Ion Torrent and HiSeq WES had many positions in common because they were both whole exome hybrid captures. We have investigated in detail discrepancies between our confidence labels in the super set and any *one* of the three validation experiments. The details for each manually inspected call (i.e., variant coordinates, confidence labels, automated validation determination, and our comments after visually inspecting the alignments with IGV, etc.) are documented in **Supplementary Data File 2**.

To summarize, we have concluded that all the HighConf SNVs “not validated” are in fact real somatic SNVs that were *false negatives in the validation experiments*. Two HighConf indels “not validated” had VAF<0.5% germline signal in 21 WGS normal replicates vs. ~40% somatic signal in the tumor, but we manually relabeled them LowConf to err on the side of precision. All MedConf SNVs and indels were “validated.”

On the other hand, the Unclassified SNVs and indels that were “validated” were mostly due to the detection of variant-supporting reads in the middle of large deletions or SVs (likely false positives), having very messy alignments and many mismatches (likely false positives), having germline signal in the WGS data that was not captured in low-depth Ion data (likely false positive), being a part of very complex somatic change (possibly true variants we could not confidently classify), or very low VAF (<1.5%) variants (well below our stated LOD of 5%). All of those issues exist for LowConf SNVs and indels that were “validated.”

1.5.1 Orthogonal Confirmation of Call Set with AmpliSeq

We randomly selected 355 SNV and 21 indel calls of different confidence levels from the super set in the consensus callable regions and performed PCR-based AmpliSeq with approximately 2000X depth for tumor and B lymphocyte cells on an Illumina MiSeq sequencer. Our pre-determined rules dictated that positively confirmed calls must have at least 100 variant-supporting reads in the tumor, and no variant read in the normal with sequencing depths of $\geq 600X$ in the normal. Not confirmed calls had fewer variant-supporting reads than the expected base call errors, and/or had VAF $\geq 10\%$ in the matched normal (germline variant false positives). Uninterpretable calls did not satisfy the criteria for either positive or no validation because they did not have enough read depth (<50) in the tumor and/or normal (See Methods for details).

Both HighConf and MedConf SNV calls had very high average coverage in validation and thus had impressive validation rates (99.6% and 100%, respectively) (**Table 2**). There was one single HighConf SNV (chr3: 71328321 C>T) that was not confirmed by AmpliSeq. In AmpliSeq data, there were zero variant-supporting reads and zero forwardly-oriented reads, indicating issues with PCR enrichment, sequencing, or mapping. However, this variant was called PASS by SomaticSeq, NeuSomatic-E, and NeuSomatic-S in 62 out of 63 WGS data sets. In addition, confirming reads were also found in PacBio data as well. Therefore, based on overwhelming support in the WGS data sets in addition to PacBio data sets, we concluded that it was “not validated” due to *false negative of AmpliSeq*, and *not* due to error of our HighConf label. The six uninterpretable HighConf calls are also documented in **Supplementary Data File 2**.

There were also four Unclassified calls that were “validated,” as shown in **Supplementary Table 6**. The calls at chr3:117937789 (**Supplementary Fig. 4a**) and chr11:24722179 appeared 1-bp away to 13-bp deletions. Adjacent indel is a known false positive marker for SNV calls because it may be a result of misalignment. In addition, indels, as large as 13 bp, create a large edit distance in the alignment, which is one of the top false positive markers in SomaticSeq. The vast majority of calls with large edit distances are false positives, hence SNVs next to large indels present a particular challenge as the SNV is masked among false positive signals. We are not particularly certain if the two cases presented here are actual mutations or false positives, though we’ve kept their “confirmed” status in the validation rate calculation. The other two “confirmed” Unclassified

calls are in the middle of a 38-bp deletion (**Supplementary Fig. 4b**), and very much look like false positives of AmpliSeq. Thus, while AmpliSeq has “confirmed” four out of 22 Unclassified calls, two of them are possibly false positives and the other two are almost certainly false positives. Therefore, the “validation rate” of 18.2% may be an overestimate of false negative rate in Unclassified calls.

The VAF calculated from the WGS data sets correlated highly with the VAF calculated from the AmpliSeq data set for HighConf calls ($R=0.982$, **Fig. 2c**). On the other hand, almost all the data points at the bottom of the graph (i.e., VAF = 0 from AmpliSeq data) are Unclassified calls (black). This means some Unclassified calls have high VAF calculated from 21 WGS data sets, but they were categorized correctly as Unclassified (meaning likely false positives). In addition, a large number of uninterpretable Unclassified calls (black 3-pointed stars) lying at the bottom suggest those were correctly labeled as Unclassified in addition to the not confirmed ones (open black triangles). Moreover, the four “validated” Unclassified calls had dubious supporting evidence as well. Taken together, these results suggest that the actual true positive rate for the Unclassified calls may be even lower than the validation rate (18.2%) we reported here.

The indel equivalent is portrayed in (**Extended Data Fig. 3a**). There were 18 HighConf, 1 MedConf, and 2 Unclassified indel calls re-sequenced by AmpliSeq. The only HighConf indel call (chr1:111980493 GA>G) not confirmed was caused by discovery of germline signal in the matched normal in AmpliSeq data, whereas there was only one single variant-supporting reads in all 21 normal WGS data combined (1 out of 788 total reads). We could not make a definitive conclusion based on PacBio data, either. Thus, we manually re-labeled this indel call from HighConf to LowConf in this release. The lone Unclassified indel call was not confirmed (we expect Unclassified calls not to be confirmed).

1.5.2 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 higher-VAF SNV calls due to its relatively low coverage (**Table 1 and 2**). HighConf and MedConf SNV calls had high positive validation rates (99.1% and 100%). However, because Ion Torrent sequencing was performed at much lower depth, only 54% of the calls were deemed interpretable (compared with 91.5% for AmpliSeq). The trend of higher uninterpretable fraction with lower confidence level calls was also more pronounced in this data set because the coverage was too low to interpret many low-VAF calls.

Upon further investigation of the six HighConf SNVs and four HighConf indels “not validated” by Ion Torrent did indeed have variant-supporting reads in tumor and reference-reads only in normal, thus consistent with somatic mutation status. However, reads in those positions did not meet our pre-determined minimum mapping quality rules, so the variant-supporting reads did not count, resulting in their “not validated” determination. Based on this evidence, we had little reason to “overturn” the HighConf labels for these variant calls.

On the other hand, many of the “validated” Unclassified calls are in low-complexity regions and often have unexplained differences between tumor and normal in the vicinity of the variant calls. While some of them can be real mutations, we cannot currently resolve their status. For the three Unclassified “indels” that were validated, 1) chr10:82232948 TG>T is a few bases away from a 10-bp deletion, so if true it is a complex variant; 2) chr15:30951922 T>TG occurs right next to a SNV, so if true it is also a complex variant; and 3) chr20:63693059 has a 22-bp deletion that has a VAF of 1.4% (called by SomaticSeq 5 times, NueSomatic-E once, and NeuSomatic-S 0 time.), but got captured in the Ion Torrent data. This is probably a true mutation but its VAF is below our LOD. Like all LowConf and Unclassified calls, those positions were removed from our high-confidence regions.

The VAF correlation for HighConf calls between WGS data sets and Ion Torrent WES ($R=0.930$) was lower than that between WGS and AmpliSeq, likely due to low coverage ($34\times$), although the vast majority of the HighConf SNV calls in Ion Torrent data still stayed well within the 95% confidence interval (**Fig. 2d**). There were more uninterpretable Unclassified calls due to possible platform-specific bias (black 3-point stars) at the bottom as well, which is suggestive that the true positive rate for Unclassified calls may be lower than the reported validation rate (15.5%) for Ion Torrent data as well. The indel equivalent is included in **Extended Data Fig. 3b**.

1.5.3 Independent Confirmation of Call Set with WES on HiSeq

We used 16 HiSeq WES replicates from six sequencing centers to evaluate the concordance between these data sets and the WGS data sets employed to construct the super set. While the WES data sets were not sequenced from orthogonal platforms, they provide insights in terms of the reproducibility of our call sets in different library preparations and sequencing strategies. The scatter plots between the super set derived VAF and medium HiSeq WES-derived VAF is presented in **Fig. 2e**. All variants in the reference call set (HighConf and MedConf calls) variants had consistent VAF calculated from both sources.

Again, simple rules were implemented for validation with the WES data as well (**Table 2**). The validation rate for HighConf, MedConf, LowConf, and Unclassified SNV calls by WES were 100%, 100%, 83.5%, and 13.4%, consistent with the previous two methods. The validation rate for HighConf, MedConf, LowConf, and Unclassified indel calls by WES were 98.2%, 100%, 76.9%, and 38.5%. The high-confidence variants were completely reproducible in WES. The one HighConf indel that was “not validated” by WES had a little bit of germline signal: a total of 5-variant supporting reads, 833 in total from 21 WGS replicates, but higher in some WES sets so it was “not validated.” It occurred in a homopolymer region (11A’s), so that it is possible to have indel sequencing error in the normal at 0.6%, against VAF in the tumor of 41% (251/614). This is probably a real indel. Nevertheless, we also removed it from the reference call set as a LowConf because we err on the side minimizing false positives.

The large number of uninterpretable Unclassified calls at the bottom (black 3-pointed stars) again suggests that the actual true mutation rate among Unclassified calls may indeed be lower than the reported validation rates. The indel equivalent is the subject of **Extended Data Fig. 3c**.

1.6 Overlapping positions between AmpliSeq Target and Ion Torrent WES

There were only seven variant calls that were re-sequenced by both AmpliSeq and Ion Torrent (**Supplementary Table 7**). We highlighted the one call where the validation interpretations were different. The HighConf call at chr15:28815512 was deemed “not confirmed” by Ion Torrent because all the variant-supporting reads had low mapping quality scores only in Ion Torrent data and thus were not counted. It was one of the six “not confirmed” HighConf SNV calls by Ion Torrent that indeed seemed like true positives, and it was “confirmed” by AmpliSeq.

1.7 Rescue low-VAF variant calls

The 21 WGS tumor-normal data sets we relied on to discover somatic mutations had average sequencing depths around 50X. Thus, VAF<5% variants were not expected to be highly reproducible and would be “lucky” to make it as a HighConf call in our *initial* definition. We estimated based on binomial distribution the probability that a VAF=5% variant would be captured as a HighConf mutation using only the 21 50X WGS replicates. With the assumption that three variant reads are necessary and sufficient to make a somatic mutation call (in reality two reads are necessary and can be sufficient, but sufficiency also depends on mapping, alignment, and sequencing qualities that are impossible to estimate accurately), the probability that a VAF=5% variant will be captured in the majority of the replicates and the majority of the sequencing centers (hence a HighConf label) is just 37%. However, the probability that it will escape the all-inclusive call set entirely (one out of 21 replicates will make the super set) is close to zero ($\sim 2.5 \times 10^{-6}$). In other words, while a low-VAF variant will unlikely be labeled as a high-confidence somatic mutation initially, it is unlikely to escape the all-inclusive call set entirely.

To mitigate the shortcoming of 50X sequencing data, we used additional 400X and 350X data sets to promote the confidence level classifications of low-VAF variants in our initial call set, and to rescue previously undetected variants. With the same set of assumptions, even at half of the average depths (i.e., 200X and 175X), a VAF=5% will be captured by both 400X and 350X data sets (and therefore promoted into the high-confidence call set) at 99.4% probability. On the other hand, VAF=3% variants are expected to be “rescued” at only 84.6% probability. Therefore, the limit of detection (LOD) of our reference call set is about 5%. We hereby promoted 1922 SNVs and 98 indels into the high-confidence reference call set at this step. We additionally added 554 LowConf SNVs and 79 LowConf indels into the all-inclusive set, as none of them had consistent SomaticSeq classifications across either the two samples or the three aligners.

Supplementary Fig. 2b shows the VAF for the combined NovaSeq simulation, where 5/9 of the replicates were designated as tumor and 4/9 of the replicates as normal, to create an average sequencing depth of 203X for the designated tumor and 160X for the designated normal. The sensitivities for variants with $2.0\% < \text{VAF} \leq 4.0\%$ and $4.0\% < \text{VAF} \leq 6.0\%$ were 78.4% and 90.3% respectively, a vast improvement from the 50X training set where the sensitivities were 8.6% and 40.3%, respectively (Supplementary Section. 1.1). It stands to reason that the sensitivities for low-VAF variants in 400X and 350X sequencing data are higher.

Finally, to further maximize our sensitivity for low-VAF variants, we created a pair of tumor-normal 1500X WGS data sets by combining all of the short-read WGS data sets into one. We used three aligners and NeuSomatic-S to call variants in them. Because we expected that the 400X and 350X data sets would have captured almost all of the $\text{VAF} \geq 5\%$ variants already, we set up a rather conservative criterion here to promote additional calls into the high-confidence

reference call set. We only promoted calls into the high-confidence call set if in the data sets where the variant was not called, there were ≤ 3 reads. In other words, if a call was LowConf or Unclassified previously due to mapping or alignment, we would not promote them here. We only promoted calls that were LowConf or Unclassified due to low signal. We additionally promoted 915 SNVs and 34 indels into the high-confidence reference call set at this step. We have also added 9 more LowConf SNVs and 6 more LowConf indels that were previously undetected into the super set.

While our binomial estimates did not prove that we have not missed any low-VAF somatic variants, it has shown that the chances of low-VAF false negative due to low sequencing depths are very low due to our use of 21 independent replicates, and then additionally 350X, 400X, and 1500X data sets. The binomial estimates can be found in our code repository.

1.8 Undetected *in silico* variants

We investigated the small number of high-VAF ($\geq 40\%$) *in silico* variants that were not detected by any caller, and thus undetectable by SomaticSeq and NeuSomatic-E as well. We found the reasons preventing detection to be a combination of low coverage and poor mapping (**Supplementary Fig. 5**). The majority of the undetected variants had coverage of fewer than 10 reads (**Supplementary Fig. 5a**), and the vast majority of them had ≤ 1 variant-supporting read (**Supplementary Fig. 5b**). Moreover, these regions tended to have very low mapping quality scores (**Supplementary Fig. 5c**) and with many MQ=0 reads, **Supplementary Fig. 5d**. The mapping qualities were so low for these reads that when BAMSurgeon mutated a base in those reads, they no longer mapped to their original positions. That change in alignment was the reason for zero variant-supporting read in many of these *in silico* mutations. In cases like these, the original mapping was also questionable. Low mappability regions were excluded from our high-confidence regions in this release. These low mappability regions represent limitations of today's short-read sequencing technologies. It also presents opportunities for the development of next-generation technologies, as long-read sequencing technology achieves higher accuracy.

1.9 How to use the truth set to benchmark a call set

Here, we demonstrate an example of how to use our reference call set to evaluate the accuracy of a pipeline. We used the hap.py software recommended by the Global Alliance for Genomics and Health (GA4GH)⁹ to benchmark the Strelka2 call set generated from the tumor BAM file with the highest sequencing depth (87X): NV_T_2 vs. NV_N_2 aligned with NovoAlign.

```
hap.py/bin/som.py -N -o /data/PATH/benchmark_results -r /data/PATH/GRCh38.fa -R /data/PATH/High-Confidence_Regions.bed /data/PATH/sSNV.MSDUKT.superSet.v1.2.vcf.gz /data/PATH/WGS.novo.dedup-NV_T_2_vs_NV_N_2-Strelka.snv.vcf.gz
```

Of the 63 Strelka2 call sets, this call set had the highest sensitivity at nearly 97.2% (1,083 false negatives), which is not surprising since it was generated from the data set with the highest sequencing depth. It had a precision of 86.8% (5,794 false positives).

The false negatives are heavily biased toward MedConf calls. While MedConf calls only consist of 5.5% of the total SNVs in the reference call sets, nearly 60% of the false negatives here are MedConf variants. This is not surprising, since MedConf calls are almost exclusively low-VAF variants that are not reproducible in these 50X WGS data sets.

To show that our reference call set correctly included these MedConf calls into the reference call set, we visually inspected 10 of them on IGV using the 350X tumor-normal data to make sure that they were real mutations that were missed in this Strelka2 call set on NV_T_2 vs. NV_N_2. Indeed, all of them look like real mutations, with VAF ranging between 3% and 7%. In the NV_T_2 replicate on which Strelka2 ran, however, the total number of variant-supporting reads for these calls range from 0 to 3. Thus, it is not surprising that those variants were not detected in NV_T_2. An example of such a false negative is shown in **Supplementary Fig. 6a**.

Conversely, we also randomly inspected 10 false positives to make sure that our reference call set did not miss them, and all of those 10 calls were clearly not real mutations. They either had clear germline signals in the 350X SPP data, or very messy alignments that clearly indicated that they were false positives. Two representative examples are shown in **Supplementary Fig. 6b** for messy alignments and **Supplementary Fig. 6c** for germline false positive. The 10 false negatives and 10 false positives are documented in the **Supplementary Data File 1**.

1.10 Additional functional features and tumor mutation burdens

Tumor mutational burden (TMB) is defined as the number of non-synonymous somatic mutations per Mbps⁴⁴. Recent literature has increasingly reported the correlation between TMB and response to anti-PD(L)-1 immunotherapy⁴⁵. The “gold standard” of measuring TMB is to perform tumor-normal WES and find the total number of non-synonymous mutations (in the coding regions). Due to the higher cost and more time required for WES, researchers are trying to infer TMB using less expensive targeted oncology panels with a small number of driver genes. One way to increase the statistical efficiency of estimating TMB for a targeted panel is to use all somatic mutations, including synonymous mutations which are expected to correlate highly with frequency of non-synonymous mutations if assuming most non-synonymous mutations, especially in high-TMB patients, are not enriched in driver genes compared with synonymous mutations. We inferred TMB with various commercially available target panels. The uncertainties of mutation rate (calculated as the 95% binomial confidence interval) inferred by smaller oncology panels are quite large, so we advise caution when attempting to infer TMB from targeted oncology panels (**Supplementary Table 5**).

2 Germline variants

2.1 Indel validations

The indel equivalent for germline SNV validation with AmpliSeq and Ion WES (shown in the main text **Fig. 4**) is shown in **Extended Data Fig. 4**.

3 Step-by-step procedures to determine the confidence-level of somatic mutations

Here, we document the procedures and commands that defined the four evidence levels (i.e., HighConf, MedConf, LowConf, and Unclassified) in the somatic mutation all-inclusive call set. Instructions to reproduce our results and the location of the files are also documented and will be kept up-to-date at the Somatic Mutation Working Group’s website:
<https://sites.google.com/view/seqc2>

3.1 Creating semi-synthetic BAM files as training data

We have multiple intra-center sequencing replicates in 12 Data Groups: HiSeq @IL, HiSeq @FD, HiSeq @NV, and NovaSeq @IL, each aligned with three aligners (BWA MEM¹⁰, Bowtie2¹¹, and NovoAlign¹²). We employed BAMSurgeon to spike *in silico* SNVs and indels into one of the normal replicates as the *in silico* tumor. These *in silico* tumor-normal pairs served as training data to create SomaticSeq and NeuSomatic classifiers.

An example command to create an *in silico* tumor-normal pair is as follows:

```
somaticseq-2.8.1/utilities/dockered_pipelines/bamSimulator/BamSimulator_multiThreads.sh --output-dir /ABSOLUTE/PATH/bwa.NV1N_vs_NV2N --normal-bam-in /ABSOLUTE/PATH/WGS_NV_N_1.bwa.dedup.bam --tumor-bam-in /ABSOLUTE/PATH/WGS_NV_N_2.bwa.dedup.bam --normal-bam-out normalDesignate1N.bam --tumor-bam-out tumorDesignate2N.bam --genome-reference /ABSOLUTE/PATH/hg38.fa --num-snvs 100000 --num-indels 20000 --num-svs 1000 --min-vaf 0 --max-vaf 1 --left-beta 2 --right-beta 5 --max-depth 1000 --seed 1981 --threads 36 --merge-output-bams
```

We also had two additional deep sequencing WGS data sets: 1) 400X NovaSeq data set created by combining the nine sequencing replicates from NovaSeq @IL, and 2) 350X tumor-normal WGS data sets from GT as a part of the tumor-normal titration series. For each of these two data sets, the same three aligners were used.

The training data sets for the 400X NovaSeq were created by combining 5/9 normal replicates as the pseudo-tumor, and the 4/9 remaining normal replicates as the normal. *In silico* mutations were then spiked into the designated tumor. Two such training data sets were created. The sequencing depth of the training data were about 220X for the *in silico* tumor and 180X for the designated normal. The classifiers are still accurate for 400X depth data (**Supplementary Table 8**).

Supplementary Table 8 shows that using the classifiers created by single NovaSeq replicates, such that their sequencing depths were about 1/4 the target data (training set described in previous paragraph), the accuracies were equivalent to the cross-validation results for the 12 Data Groups (**Supplementary Fig. 1**), even though the cross validation of the two training sets with equivalent depths displayed higher accuracies. This improvement in accuracy could be attributed almost exclusively to low-VAF variants (**Supplementary Fig. 2**). The training sets created for combined NovaSeq were about 1/2 the target data, so the discrepancy in sequencing depth is less than those shown in **Supplementary Table 8**. The training data for the 350X WGS data were created by randomly splitting the normal BAM files into a designated tumor and normal at a 55:45 ratio.

3.2 Building SomaticSeq and NeuSomatic classifiers

For each of the 12 Data Groups with sequencing replicates, two training data sets were created as described in Sec. 3.1. For each Data Group, the two training data sets were combined to create SomaticSeq classifiers specific to that Data Group. Additionally, all the training data sets were combined to create SomaticSeq (for LL, EA, and NC) and NeuSomatic classifiers (for all WGS data sets).

3.3 Classification of each pair of tumor-normal BAM files

For each of the 63 tumor-normal pairs, we have run SomaticSeq, NeuSomatic-Ensemble (E), and NeuSomatic-Standalone (S) pipelines.

3.4 Scoring each Data Group (for each of the three aligners)

The steps in Sec. 3.4 - 3.9 describe in detail the schematics shown in main text's **Fig. 1**, i.e., the algorithm that determined the four confidence levels. Sec. 3.4 – 3.7 rely on SomaticSeq and made up the *initial* definition of our call set. In Sec. 3.8 – 3.9, NeuSomatic classifications of the 63 WGS data sets plus the 1,500X pooled WGS data sets were taken into account to make up our final definition of this v1.2 release. The commands used can be found in our code repository.

There were five Data Groups for each aligner, resulting in 15 Data Groups as shown in main **Fig. 1**. In each Data Group, a somatic mutation candidate got a +1 score if it was classified PASS in a sample, a -1 if it was classified REJECT in a sample, and 0 otherwise (LowQual or undetected). The sample scores were summed up in each Data Group. Then, Data Group scores of +3, +1, 0, or -3 were assigned based on the two tables on the left of **Supplementary Fig. 7**. The upper left table shows all the possible combinations of sample scores for Data Groups with three sequencing replicates, i.e., to get a Data Group score of +3, at minimum two out of three samples had to be PASS, without a single REJECT. While a Data Group score of +1 was not as confident as +3, it still showed positive supporting evidence. A Data Group score of -3 is a “mirror image” of +3, having considerably more contradicting evidence than supporting evidence. The lower left table shows all the possible combinations for the 9-replicate Data Groups, i.e., NovaSeq @IL.

3.5 Aligner-Centric Classifications

For each aligner, the Data Group scores for the five Data Groups were summed together to make an aligner-centric classification. The right table of **Supplementary Fig. 7** shows every possible combination of the five Data Group scores, and the aligner-centric classifications associated with each combination.

3.6 Assigning the initial four Confidence Levels

First, we assigned initial tiers based strictly on the three aligner-centric classifications.

Supplementary Table 9 shows the criteria and the number of total calls that fell into each tier.

The initial tiers provided a starting point to determine how confident we were that a call was a true mutation. In addition, we also took into account certain “red flags” (i.e., high number of MQ0 reads, germline signal, high edit distances, VAF inconsistent with GT's tumor-normal titration data sets, or a complete lack of a supporting call from one of the three aligners) to create the four confidence levels of HighConf, MedConf, LowConf, and Unclassified.

The algorithm to initially assign such confidence levels is at SomaticSeq's public seqc2 branch in Github. A brief description is as follows. First, for each aligner, we established a low-end MQ and a high-end edit distance (BAM tag: NM) threshold with the bottom/top 2% of values among calls that were classified as PASS every single time. Then, we used the following rules to assign the four confidence levels:

1. Tier 1 calls defaulted to HighConf, unless the call was not in the consensus callable region, or if there were more MQ0 reads than there were samples, in which case they got demoted to MedConf.
2. Tier 2 calls also defaulted to HighConf, but there were more opportunities to be demoted to lower confidence-levels, such as zero PASS calls from the third aligner, presence of

germline signal, location outside the consensus callable region, VAF not moving consistently with the tumor-normal titration, or if the MQ or NM exceeded the established thresholds.

3. Tier 3 calls defaulted to MedConf and could be demoted to lower confidence-levels by the same things that demoted Tier 2 calls.
4. Tier 4 calls defaulted to Unclassified but could be promoted to LowConf if they had more sequencing replicates where the call was classified as PASS vs. REJECT.

3.7 Rescue low-VAF and lower confidence level calls

As described previously, for low-VAF ($\leq 15\%$) calls with LowConf and Unclassified labels, two deep sequencing data sets were used to rescue those calls: 1) nine NovaSeq data sets combined into 400X, and 2) a 350X tumor-normal WGS data set that was a part of a tumor-normal titration series. If the calls were deemed PASS in those data sets, they would be promoted into the high-confidence call set.

3.8 Modify confidence labels with NeuSomatic classifications

For each variant call in the super set, we have annotated in how many data sets they were classified PASS by both NeuSomatic modes, i.e., Ensemble mode (NeuSomatic-E) and Standalone mode (NeuSomatic-S). We have examined and re-labeled the confidence of calls when the number of data sets classified as PASS by SomaticSeq deviated significantly ($p < 0.05$, two-proportion z-test) from that of NeuSomatic-E/NeuSomatic-S. We also added an additional 9 LowConf SNVs and 6 LowConf indels into the super set at this step. These calls were previously undetected by SomaticSeq.

3.9 Use 1500X call sets for final rescue of low-VAF calls

We combined all short-read WGS data sets into a pair of combined tumor-normal data sets of 1500X in depth, for each of the three aligners. We used these data sets to promote low-VAF low-signal calls into the high-confidence call set (i.e., if the variant call was rejected in 50X data sets due to low signal).

3.10 Use PacBio data to remove inconsistent variant calls

Our PacBio data's sequencing depth is about 30-40X, and also has error rates higher than many variant calls in the reference call set (**Suppl. Table 2**). Thus, PacBio was not used for variant discovery, but we used it to remove variants from low-mappability regions (i.e., PacBio MQ < 50 , short-read MQ0 reads > 10 , or Phred-scaled linguistic sequence complexity¹³ < 12) from the reference call set if 1) PacBio has no variant-supporting read, and 2) one-sided z-test yields a p-value < 0.05 .

This is the final step in our establishment of the v1.2 reference call set. Variant calls that are high confidence as true positives in the High-Confidence Regions are labeled PASS in the FILTER column of the VCF files.

4 Software and Version

Software information is presented **Supplementary Table 10**.

5 Statistical methods

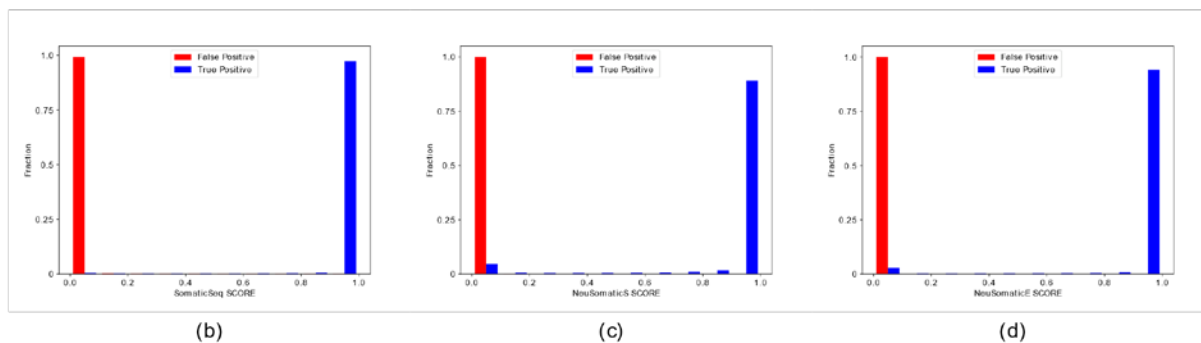
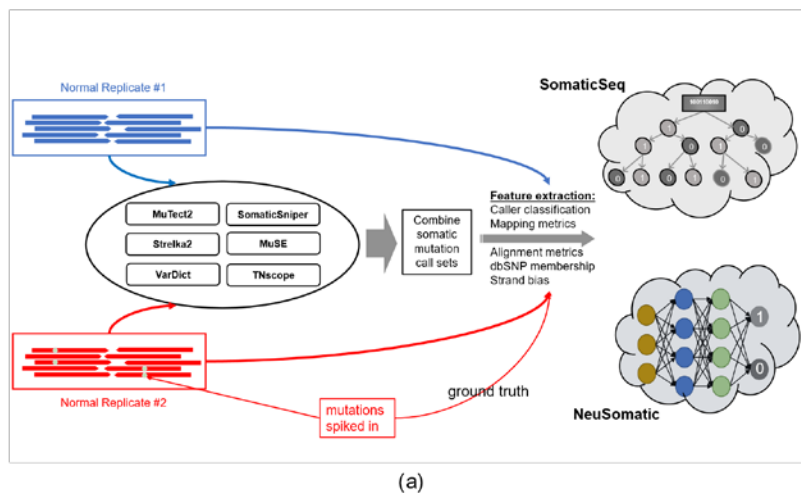
One-sided Z-tests were used to test the difference between two proportions, Pearson's correlation coefficient to calculate the correlation. A linear mixed model was applied to validate somatic mutations with tumor titration series and a generalized linear mixed model was specified to calculate the SNV/indel calling probability (SCP), both of which were detailed at the online method section. One-sided Z-tests and Pearson's correlation were calculated by python with the scipy package. The linear mixed model and generalized linear mixed model were fitted in R.

References

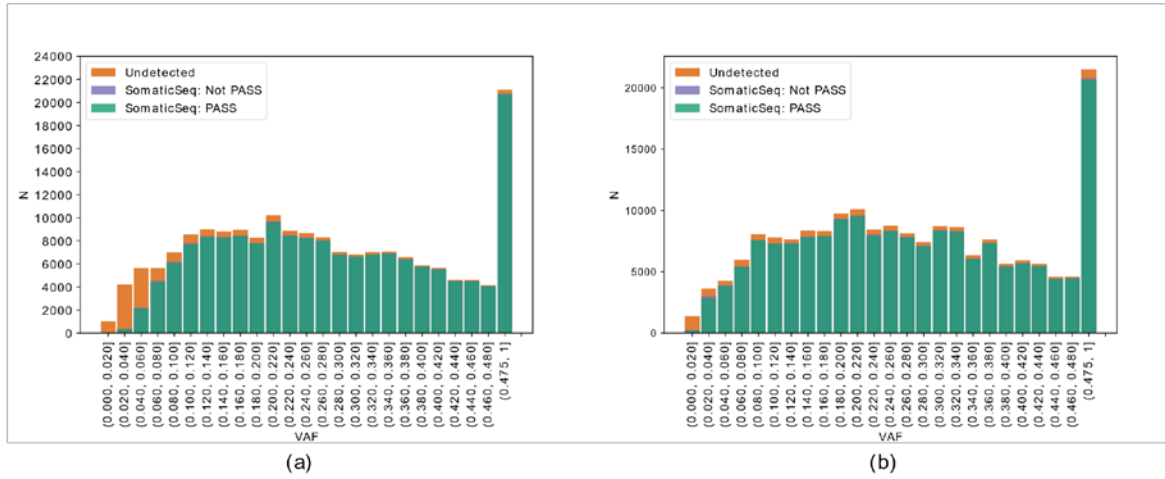
1. Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.* **16**, 197 (2015).
2. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
3. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
4. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
5. 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).
6. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591 (2018).
7. 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.

8. Araten, D. J. *et al.* A Quantitative Measurement of the Human Somatic Mutation Rate. *Cancer Res.* **65**, 8111–8117 (2005).
9. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555 (2019).
10. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
11. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
12. NovoCraft Technologies Sdn Bhd. NovoAlign.
<http://www.novocraft.com/products/novoalign/>.
13. Troyanskaya, O. G., Arbell, O., Koren, Y., Landau, G. M. & Bolshoy, A. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics* **18**, 679–688 (2002).
14. Cheng, D. T. *et al.* Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med. Genomics* **10**, 33 (2017).
15. FoundationOne. FoundationOne CDx Technical Information.
https://www.accessdata.fda.gov/cdrh_docs/pdf17/P170019C.pdf.

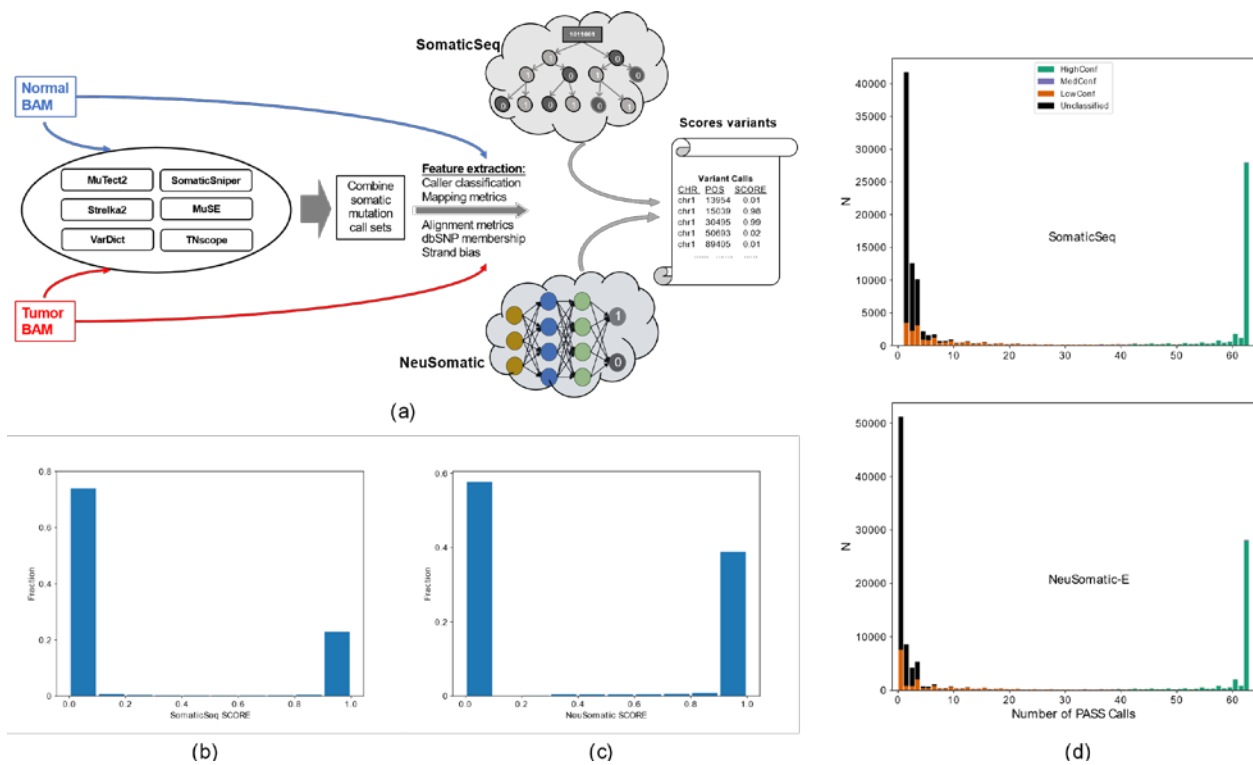
Supplementary Figures



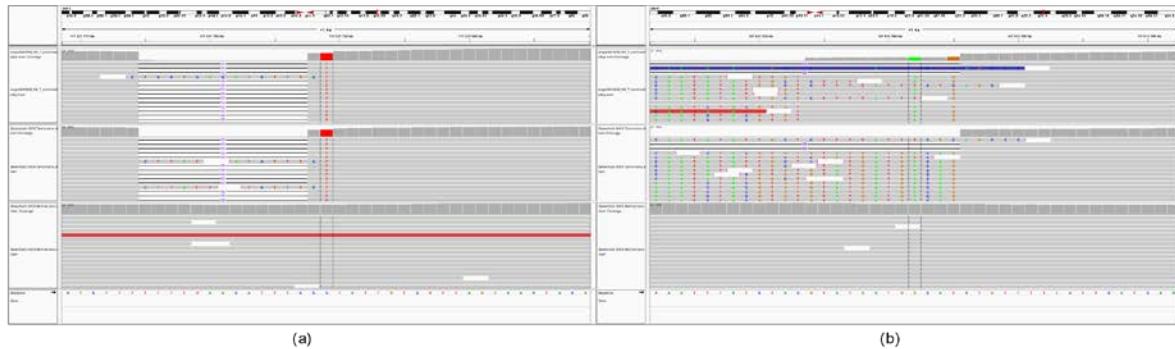
Supplementary Figure 1) Using simulated data to build the classifier for machine learning models: SomaticSeq and NeuSomatic. (a) The schematic of training workflow to create machine learning classifiers. (b) Histogram of SomaticSeq scores assigned to true positives and false positives. Most calls are either > 0.9 (likely true positive) or < 0.1 (likely false positive), with only a tiny fraction of calls in between (1.03%). (c) Histogram of NeuSomatic-S scores. (d) Histogram of NeuSomatic-E scores.



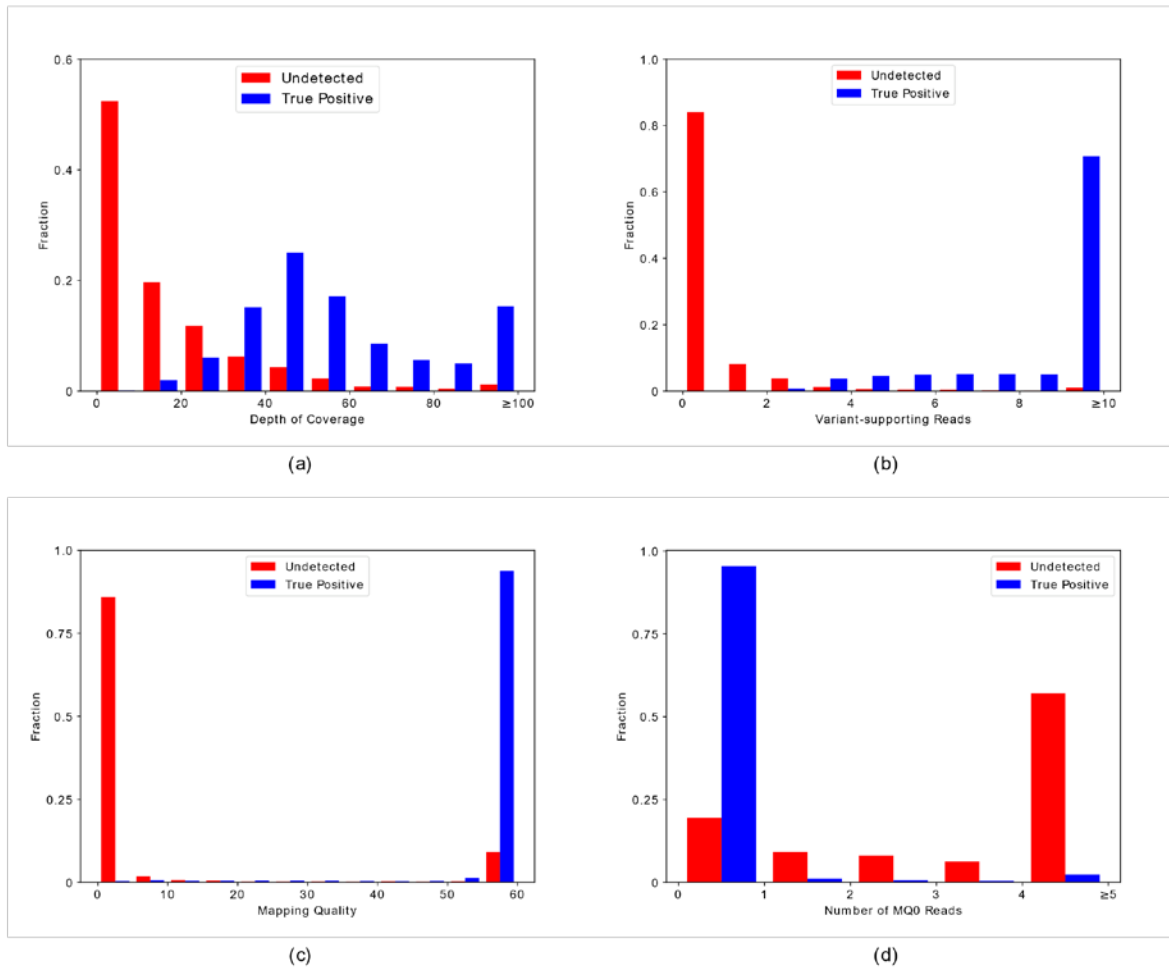
Supplementary Figure 2) Histogram of BAMSurgeon simulated variant. (a) Training data sets with approximately 50X in sequencing depth. (b) Training data set with 203X tumor depth and 160X normal depth. The green bars represent the variants that were classified correctly as true positives by SomaticSeq. The purple bars (on top of green bars), hardly visible in many cases, represent true positives that were detected by at least one caller but were not correctly classified as true positive (PASS) by SomaticSeq. The orange bars represent variants that were not detected by any of the six callers.



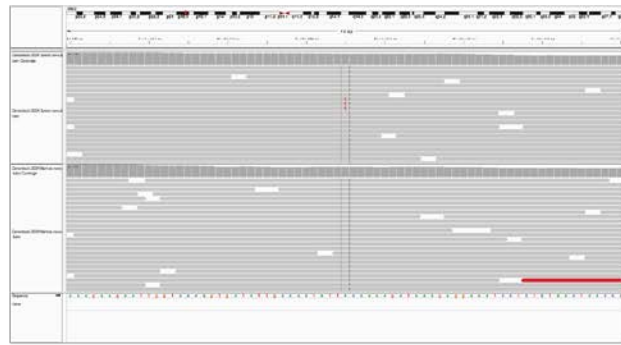
Supplementary Figure 3) Using SomaticSeq and NeuSomatic to classify somatic mutation calls of each pair of BAM files (a) Mutation calling workflow for each pair of tumor-normal BAM files. The SomaticSeq classifiers are specific to each sequencing center and each aligner as depicted in Fig. S1. NeuSomatic classifiers were trained on combined training data sets. (b) Histogram of SomaticSeq scores for our call sets. This is qualitatively similar to the cross-validated histogram shown in Fig S1b. (c) Histogram of NeuSomatic-E scores for our call sets. (d) The number of tumor-normal data sets (out of 63 in total) where SomaticSeq (top) and NeuSomatic-E (bottom) have called a variant PASS. It is apparent that the distribution is bimodal, i.e., HighConf calls have been classified as PASS by both SomaticSeq and NeuSomatic-E in the vast majority of 63 data sets, whereas Unclassified calls have been classified as PASS only a few times.



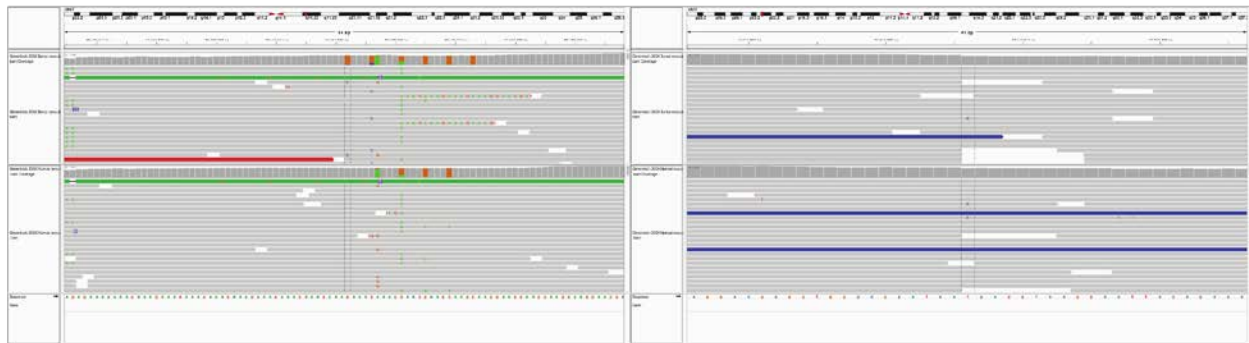
Supplementary Figure 4) Integrative Genome Viewer display of calls of interest. For each of the sub figures, the top alignments are NovaSeq tumor aligned with BWA. The middle alignments are HiSeq tumor aligned with NovoAlign. The bottom alignments are HiSeq normal aligned with NovoAlign. (a) The chr3:117937789 C>T is 1-bp away from the 13-bp deletion, which brings into question the mapping/alignment. It may be a true complex variant. (b) The chr8:107631542 G>A that seemed to be in the middle of a 38-bp deletion. This looks like a false positive even though it was deemed “validated” by our rules; it may be correctly labeled as “Unclassified.”



Supplementary Figure 5) Undetected high-VAF in silico mutations. (a) Depth of coverage for undetected vs. correctly classified simulated variants. For correctly classified variants, the depth of coverage peaked around 50X, which was our target coverage. For the undetected variants; however, the highest peak represented depth of coverage of less than 10. (b) Number of variant-supporting reads. Many of the undetected variants had zero variant-supporting reads, due to low-MQ mutated reads being mapped to a different location after in silico mutation. (c) Most of the undetected variants have BWA mapping quality scores below 5. (d) The majority of undetected variants had at least five MQ=0 reads mapped to that position, indicating low mappability regions.



(a)



(b)

(c)

Supplementary Figure 6) IGV screenshots of a false negative and two false positives that are representative of the 20 visually inspected calls in our benchmarking of Strelka2 called on NV_T_2 vs. NV_N_2 with NovoAlign. The top alignments represent pure tumor of a 350X data. The bottom alignments represent pure normal of a 350X data. (a) A false negative at chr2:51181085 C>T, where there was only one variant read in NT_V_2 so it was not called. In the 350X data, however, there were 9 variant reads with a VAF of 4%. (b) A false positive at chr7:68499352 in a GAAA tandem repeat region with numerous read-to-reference mismatches everywhere in the region. This looks like a clear false positive. (c) A false positive at chr2:32343566. Both tumor **and** normal have 5 variant-supporting reads in the 350X data. There were no variant-supporting reads in NV_N_2, so it was called a somatic mutation in that replicate.

PASS = 1	LowQual = 0	REJECT = -1	SUM	Group Score
3	0	0	3	3
2	1	0	2	
2	0	1	1	1
1	2	0	1	
1	1	1	0	0
0	3	0	0	
0	0	2	-1	
1	0	1	-1	
0	1	2	-2	-3
0	0	3	-3	

9	0	0	9	3
8	1	0	8	
8	0	1	7	
7	2	0	7	
7	1	1	6	
6	3	0	6	
7	0	2	5	
6	2	1	5	
5	4	0	5	
6	1	2	4	
5	3	1	4	
4	5	0	4	
6	0	3	3	1
5	2	2	3	
4	4	1	3	
3	6	0	3	
5	1	3	2	
4	3	2	2	
3	5	1	2	
2	7	0	2	
5	0	4	1	
4	2	3	1	
3	4	2	1	
2	6	1	1	0
1	8	0	1	
4	1	4	0	
3	3	3	0	
2	5	2	0	
1	7	1	0	
0	9	0	0	
4	0	5	-1	
3	2	4	-1	
2	4	3	-1	
1	6	2	-1	
0	8	1	-1	-3
3	1	5	-2	
2	3	4	-2	
1	5	3	-2	
0	7	2	-2	
3	0	6	-3	
2	2	5	-3	
1	4	4	-3	
0	6	3	-3	
2	1	6	-4	
1	3	5	-4	
0	5	4	-4	
2	0	7	-5	-3
1	2	6	-5	
0	4	5	-5	
1	1	7	-6	
0	3	6	-6	
1	0	8	-7	
0	2	7	-7	
0	1	8	-8	
0	0	9	-9	

# of Score=3's	Group Score #1	Group Score #2	Group Score #3	Group Score #4	Group Score #5	SUM	Aligner-Centric Classification
5	3	3	3	3	3	15	Strong Evidence
4	3	3	3	3	1	13	
4	3	3	3	3	0	12	
3	3	3	3	1	1	11	
3	3	3	3	1	0	10	
4	3	3	3	3	-3	9	
3	3	3	3	0	0	9	
2	3	3	1	1	1	9	
2	3	3	1	1	0	8	
3	3	3	3	1	-3	7	
2	3	3	1	0	0	7	
1	3	1	1	1	1	7	
3	3	3	3	0	-3	6	Weak Evidence
2	3	3	0	0	0	6	
1	3	1	1	1	0	6	
2	3	3	1	1	-3	5	
1	3	1	1	0	0	5	
0	1	1	1	1	1	5	
2	3	3	1	0	-3	4	
1	3	1	0	0	0	4	
0	1	1	1	1	0	4	
3	3	3	3	-3	-3	3	
2	3	3	0	0	-3	3	
1	3	1	1	1	-3	3	Neutral Evidence
1	3	0	0	0	0	3	
0	1	1	1	0	0	3	
1	3	1	1	0	-3	2	
0	1	1	0	0	0	2	
2	3	3	1	-3	-3	1	
1	3	1	0	0	-3	1	
0	1	1	1	1	-3	1	
0	1	0	0	0	0	1	
2	3	3	0	-3	-3	0	
1	3	0	0	0	-3	0	
0	1	1	1	0	-3	0	Likely False Positive
0	0	0	0	0	0	0	
1	3	1	1	-3	-3	-1	
0	1	1	0	0	-3	-1	
1	3	1	0	-3	-3	-2	
0	1	0	0	0	-3	-2	
2	3	3	-3	-3	-3	-3	
1	3	0	0	-3	-3	-3	
0	1	1	1	-3	-3	-3	
0	0	0	0	0	-3	-3	
0	1	1	0	-3	-3	-4	
1	3	1	-3	-3	-3	-5	
0	1	0	0	-3	-3	-5	
1	3	0	-3	-3	-3	-6	
0	0	0	0	-3	-3	-6	
0	1	1	-3	-3	-3	-7	
0	1	0	-3	-3	-3	-8	
1	3	-3	-3	-3	-3	-9	
0	0	0	-3	-3	-3	-9	
0	1	-3	-3	-3	-3	-11	
0	0	-3	-3	-3	-3	-12	
0	-3	-3	-3	-3	-3	-15	

Supplementary Figure 7) Intermediate Scoring system. The upper left table depicts all the different combinations of per sample classification when a Data Group has 3 replicates: PASS = +1, REJECT = -1, and 0 otherwise. Their sum was converted to four possible Group Scores of 3, 1, 0, and -3. The lower left table depicts the same thing as the upper left, except it is for Data Groups with 9 replicates. The table on the right depicts the different combinations of the five Group Scores that led to the four aligner-centric classifications for each aligner.

Supplementary Tables

Pre-final label	Number	≥ 1 aligner	≥ 2 aligners
HighConf	37456	99.6%	97.2%
MedConf	1291	93.0%	77.0%
LowConf	15584	24.0%	13.6%
Unclassified	37901	1.7%	0.4%

Supplementary Table 1) This table shows the number of variants in each confidence level and their concordance with the 1,500X NeuSomatic call sets before the confidence labels were further polished. The ≥ 1 aligner column shows the percentage of variant calls that were rediscovered in at least one of the three 1,500X call sets (one call set for each of the three aligners). The ≥ 2 aligners column counts a variant as concordant if it is called in at least two out of three 1,500X call sets.

Platforms	Total Calls	Substitution Errors	Indel Errors	Substitution Error Rate	Indel Error Rate
HiSeq	1,818,768,592,271	2,081,962,721	35,530,669	0.114%	0.002%
NovaSeq	1,009,625,000,339	860,881,365	24,258,829	0.085%	0.002%
PacBio	121,639,927,836	4,464,346,903	8,596,779,700	3.670%	7.067%

Supplementary Table 2) The error rates were estimated as follows: in HCC1395BL sequencing data, every genomic position in the high-confidence region, except for the germline variant positions, is assumed to be a reference base. Error rates were calculated based on the putative reference bases only. In aligned reads (i.e., BAM files), every non-reference base call covering the putative reference base position is assumed to be an error and counted as such. HiSeq and NovaSeq has very similar error profiles, i.e., approximately one substitution error for every 1000 bases and one indel error for every 50,000 bases. The error rate of PacBio long-read sequencing data is higher than the detection limit of our somatic mutation reference call set.

Validation Platform	Variant Type	Category	Total Number	Fraction Interpretable	Validation Rate (Interpretable)	Validation Rate (Total)
AmpliSeq Deep Sequencing	SNV	SCP = 1	422	(419/422) 99.3%	(416/419) 99.3%	98.6%
		$1 > \text{SCP} \geq 0.9$	63	(54/63) 85.7%	(41/54) 75.9%	65.1%
		$0.9 > \text{SCP} \geq 0.5$	68	(56/68) 82.4%	(17/56) 30.4%	25.0%
		$0.5 > \text{SCP} > 0$	11	(11/11) 100.0%	(0/11) 0.0%	0.0%
	INDEL	SCP = 1	48	(47/48) 97.9%	(46/47) 97.9%	95.8%
		$1 > \text{SCP} \geq 0.9$	23	(20/23) 87.0%	(19/20) 95.0%	82.6%
Ion Torrent WES	SNV	SCP = 1	44451	(43148/44451) 97.1%	(42052/43148) 97.5%	94.60%
		$0.5 > \text{SCP} > 0$	4	(4/4) 100%	(0/4) 0%	0%
	INDEL	SCP = 1	2899	(2625/2899) 90.5%	(2536/2625) 96.6%	87.50%

Supplementary Table 3) AmpliSeq and Ion Torrent WES validation rate for germline variants.

Category	Callable		Non-Callable	
	Number	Fraction	Number	Fraction
HighConf	37,498	40.65%	0	0.00%
MedConf	2,162	2.34%	336	1.28%
LowConf	14,901	16.15%	3,896	14.80%
Unclassified	37,787	40.96%	22,086	83.92%
Total	92,248	100.00%	26,318	100.00%

Supplementary Table 4) The number of SNV calls of each category in the consensus callable regions and those outside of them. Consensus callable regions made up 92% of the entire genome.

Panels	Region size	Total SNVs	Total non-syn SNVs	95% interval of Inferred SNV/Mbp	95% interval of Inferred non-syn SNV/Mbp
SEQ-II high-confidence regions	2,485,467,324	39,560	---	15.9	---
UCSC Genecode v29 Coding	31,101,689	473	352	14.6 (13.2 – 15.8)	10.9
TruSight One Expanded v2.0	14,607,935	209	141	14.3 (12.4 – 16.4)	9.7 (8.1 – 11.4)
TruSight One v1.1	10,542,342	147	100	13.9 (11.8 – 16.4)	9.5 (7.7 – 11.5)
MSK-IMPACT ¹⁴	934,044	9	7	9.6 (4.4 – 18.3)	7.5 (3.0 – 15.4)
309 cancer-related gene coding regions*	715,933	10	7	14.0 (6.7 -- 25.7)	9.8 (3.9 - 20.1)
TruSight Tumor 170	505,593	8	5	15.8 (6.8 – 31.2)	9.9 (3.2 – 23.0)
NEB Next Direct Cancer HotSpot	35,392	2	2	56.5 (6.8 – 204.1)	56.5 (6.8 – 204.1)
AmpliSeq Cancer Hot Spot v2	21,048	2	2	95.0 (11.5 – 343.2)	95.0 (11.5 – 343.2)
TruSight Tumor 26	13,792	1	1	72.5 (17.6 – 523.7)	72.5 (17.6 – 523.7)

Supplementary Table 5) Tumor mutation burden (TMB) benchmarks on various panels. *These 309 genes are taken from FoundationOne CDx technical information¹⁵, but this is not the same panel as it does not include additional regions. The Region Sizes are calculated as the intersection between the region sizes and the SEQC-II high-confidence regions.

CHROM	POS	REF	ALT	VALIDATION	Comments
chr3	117937789	C	T	YES	One base away from a 13-bp deletion and filled with soft-clipped reads
chr8	107631542	G	A	YES	In the middle of a 38-bp deletion, and the end of AmpliSeq coverage
chr8	107631545	T	G	YES	In the middle of a 38-bp deletion, and the end of AmpliSeq coverage
chr11	24722179	T	G	YES	One base away from a 13-bp deletion and filled with soft-clipped reads

Supplementary Table 6) The four Unclassified SNV calls that were “validated” by AmpliSeq

CHR	POS	VAF	REF	ALT	Category	AmpliSeq					Ion Torrent				
						Validation	Tumor VDP	Tumor DP	Normal VDP	Normal DP	Validation	Tumor VDP	Tumor DP	Normal VDP	Normal DP
chr9	105607844	0.669	G	A	HighConf	Confirmed	1971	4083	2	5611	Confirmed	32	59	0	65
chr9	129102404	0.229	T	C	HighConf	Confirmed	321	1301	0	1096	Confirmed	3	13	0	28
chr12	131844177	0.997	G	C	HighConf	Confirmed	390	396	0	506	Confirmed	28	28	0	58
chr15	28815512	0.471	G	A	HighConf	Confirmed	771	1589	0	1975	Not Confirmed	0	46	0	82
chr17	4545725	0.356	T	C	HighConf	Confirmed	2037	6394	5	5652	Confirmed	3	18	0	38
chr19	35488730	0.506	A	C	LowConf	Confirmed	2642	5087	0	4124	Confirmed	41	83	0	91
chr22	23875150	0.05	G	A	MedConf	Confirmed	200	4832	0	2596	Confirmed	3	82	0	71

Supplementary Table 7) The seven variant calls that were covered by both AmpliSeq Targeted Sequencing and Ion Torrent WES. There was no variant-supporting read for chr15:28815512 in Ion Torrent because none of the 25 variant-supporting reads met the minimum MQ threshold but this was also validated by AmpliSeq.

	SNV		INDEL	
	Truth Set Classification	Cross Validation	Truth Set Classification	Cross Validation
Sensitivity	98.39%	99.21%	93.79%	92.11%
Specificity	99.19%	99.67%	94.64%	97.00%
PPV	99.63%	99.74%	94.64%	96.53%
NPV	99.87%	99.91%	99.10%	98.37%

Supplementary Table 8) Accuracy of the classification for the 203X synthetic tumor vs. 160X normal. Truth set classification denotes the classification accuracies when the variant calls were classified by the classifiers created for single replicates, so their sequencing depth is about 1/4 the test data. Cross validation refers to classifiers for two equivalent training sets, so their sequencing depths are very similar.

Categories	Initial Tiers	Extra Criteria	SNV	INDEL	SNV	INDEL
Tier 1	AllPASS	All 63 data sets are classified as PASS	26,524	421	36,613	1,340
All three Aligner-Centric Classifications are Strong Evidence	Tier 1A	For each Aligner-Centric Classification, each of the 5 Data Groups are deemed “+3”	4,599	467		
	Tier 1B	For each Aligner-Centric Classification, $\geq 3/5$ of the Data Groups are deemed “+3”	4,844	391		
	Tier 1C	No extra criterion	646	61		
Tier 2	Tier 2A	For the 3rd Aligner-Centric Classification that’s not Strong it is classified as Weak	362	86	543	497
Two out of three Aligner-Centric Classifications are Strong Evidence	Tier 2B	The 3rd Aligner-Centric Classification is Neutral.	73	30		
	Tier 2C	No extra criterion: the 3rd Aligner-Centric Classification is deemed Likely False Positive.	108	381		
Tier 3	Tier 3A	The 2 other Aligner-Centric Classification are Weak	125	12	373	133
One out of three Aligner-Centric Classifications is Strong Evidence	Tier 3B	1 of the other 2 Aligner-Centric Classification is Weak	95	60		
	Tier 3C	No extra criterion: none of the other 2 is even Weak	153	60		
Tier 4	Tier 4A	All 3 Aligner-Centric Classifications are Weak	692	40	54,149	3,028
None of the three Aligner-Centric Classifications is Strong Evidence	Tier 4B	2/3 of the Aligner-Centric Classifications are Weak	276	43		
	Tier 4C	1/3 of the Aligner-Centric Classifications are Weak	252	107		
	REJECT	All Aligner-Centric Classification is Neutral/ Likely False Positive	52,929	2,838		

Supplementary Table 9) The criteria for each tier assignment, as well as the number of variant calls that fell into each category. Most calls fell either into Tier 1 or Tier 4, which means these calls either had broad evidence to support their somatic status, or very little evidence at all.

Software	Version	Sources
BAMSurgeon	v1.1	https://github.com/adamewing/bamsurgeon
Bowtie2	v2.2.9	http://bowtie-bio.sourceforge.net/bowtie2
BWA MEM	v0.7.15	https://github.com/lh3/bwa
DeepVariants	v0.5.1	https://github.com/google/deepvariant
Freebayes	v1.2.0	https://github.com/ekg/freebayes
GATK3	v3.8	https://software.broadinstitute.org/gatk/download/archive
hap.py	V0.3.10	https://github.com/Illumina/hap.py
HaplotypeCaller	v3.8.0	https://software.broadinstitute.org/gatk/download/archive
MuSE	v1.0rc_c	https://github.com/danielfan/MuSE
MuTect2	4.beta.6	https://software.broadinstitute.org/gatk/gatk4
NeuSomatic	v0.1.3	https://github.com/bioinform/neusomatic
NovoAlign	v3.07.01	http://www.novocraft.com/products/novoalign/
RTG	V3.8.4	https://github.com/RealTimeGenomics/rtg-tools
SomaticSeq	v2.8.1 / seqc2_v1.2	https://github.com/bioinform/somaticseq
SomaticSniper	v1.0.5.0	https://github.com/genome/somatic-sniper
Strelka2	v2.8.4	https://github.com/Illumina/strelka
TMAP	v3.0.1	https://github.com/iontorrent/TS/tree/master/Analysis/TMAP
TNScope	v201711.02	https://www.sentieon.com/products/#tnscope
Trimmomatic	v0.30	http://www.usadellab.org/cms/index.php?page=trimmomatic
VarDict	v1.5.1	https://github.com/AstraZeneca-NGS/VarDictJava

Supplementary Table 10) Software, version, and download location