appreci8: A Pipeline for Precise Variant Calling Integrating 8 Tools

1 Data characteristics

Table S1: Data characteristics of samples in the two training sets.

	Training set 1	Training set 2
Number of samples	54	111
Sequencing platform	Illumina HiSeq	Illumina NextSeq
Total reads	4,084,151	7,401,621
Mapped reads	4,045,417	7,360,841
Mapped reads (rel)	0.99	0.99
Qual mean	59.58	59.58
Uniquely mapped reads	4,033,456	7,328,568
Uniquely mapped reads (rel)	1.00	1.00
Target size	42,322	42,322
Reads on target	711,090	2,699,476
Target coverage $\overline{x}$	1,366.58	7,294.20
Target coverage σ	1,019.50	10,329.65
Target bases larger 1x	41,742	24,448
Target bases larger 1x (rel)	0.99	0.58
Target bases larger 50x	39,777	23,592
Target bases larger 50x (rel)	0.94	0.56
Target size (coding)	23,162	23,162
Reads on target	477,613	2,704,763
Target coverage $\overline{x}$	1,484.46	13,134.20
Target coverage σ	1,024.03	10,762.31
Target bases larger 1x	22,889	2,2947
Target bases larger 1x (rel)	0.99	0.99
Target bases larger 50x	22,103	22,542
Target bases larger 50x (rel)	0.95	0.97
Background noise	$5.39 \cdot 10^{-3}$	$6.26 \cdot 10^{-3}$

Table S2: Data characteristics of samples in the five test sets.

	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5
Number of samples	237	46	89	22	119
Sequencing platform	Illumina HiSeq	Illumina HiSeq	Roche 454	Illumina NextSeq	Illumina HiScanSQ
Total reads	3,990,586	5,284,383	56,053	10,483,971	7,811,781
Mapped reads	3,943,585	5,205,918	54,726	10,436,588	7,725,166
Mapped reads (rel)	0.98	0.98	0.98	1.00	0.99
Qual mean	59.60	59.26	59.69	59.78	57.55
Uniquely mapped reads	3,930,573	5,158,725	54,677	10,382,190	7,287,032
Uniquely mapped reads (rel)	1.00	0.99	1.00	0.99	0.94
Target size	42,322	42,322	42,322	139,767	958,547
Reads on target	767,645	939,871	45,105	10,288,590	5,674,948
Target coverage $\overline{x}$	1,473.80	1,789.57	219.73	13,062.47	575.27
Target coverage σ	1,339.67	1,446.52	334.67	10,132.05	701.29
Target bases larger 1x	41,675	41,739	23,330	139,365	915,639
Target bases larger 1x (rel)	0.98	0.99	0.55	1.00	0.96
Target bases larger 50x	37,727	38,639	20,975	138,237	780,972
Target bases larger 50x (rel)	0.89	0.91	0.50	0.99	0.81
Target size (coding)	23,162	23,162	23,162	78,866	218,179
Reads on target	533,553	647,573	45,134	9,111,624	2,724,889
Target coverage $\overline{x}$	1,684.40	2,012.08	388.50	15,092.04	870.10
Target coverage σ	1,366.72	1,476.70	369.78	11,245.09	846.80
Target bases larger 1x	22,835	22,896	21,537	78,746	215,566
Target bases larger 1x (rel)	0.99	0.99	0.93	1.00	0.99
Target bases larger 50x	21,197	21,590	19,563	78,391	204,473
Target bases larger 50x (rel)	0.92	0.93	0.84	0.99	0.94
Background noise	$4.15 \cdot 10^{-3}$	$5.02 \cdot 10^{-3}$	$3.63 \cdot 10^{-3}$	$6.63 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$

2 Variant calls

Table S3: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in training set 1.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	268	49	0.92	0.85	90	49	0.80	0.65
Platypus	272	67	0.93	0.80	94	67	0.84	0.58
VarScan	258	7	0.89	0.97	83	7	0.74	0.92
LoFreq	265	502	0.91	0.35	90	502	0.80	0.15
FreeBayes	290	8,040	1.00	0.03	112	8,040	1.00	0.01
SNVer	271	25	0.93	0.92	95	25	0.85	0.79
SAMtools	248	38	0.85	0.87	71	38	0.63	0.65
VarDict	283	11	0.97	0.96	108	11	0.96	0.91
8 tools	291	8,043	1.00	0.03	112	8,043	1.00	0.01
single-appreci8	285	6	0.98	0.98	106	7	0.95	0.94
appreci8	285	2	0.98	0.99	106	2	0.95	0.98
Biological truth	291				112			

Table S4: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in training set 2.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	516	210	0.82	0.71	154	211	0.67	0.42
Platypus	585	794	0.93	0.42	184	796	0.80	0.19
VarScan	294	108	0.47	0.73	77	108	0.33	0.42
LoFreq	490	1648	0.78	0.23	176	1649	0.77	0.10
FreeBayes	627	40,159	0.99	0.02	226	40,160	0.98	0.01
SNVer	344	4,376	0.55	0.07	130	4,378	0.57	0.03
SAMtools	402	119	0.64	0.77	67	120	0.29	0.36
VarDict	595	3,508	0.94	0.15	205	3,510	0.89	0.06
8 tools	631	40,466	1.00	0.02	230	40,467	1.00	0.01
single-appreci8	622	1,177	0.99	0.35	221	1,184	0.96	0.16
appreci8	620	40	0.98	0.94	219	41	0.95	0.84
Biological truth	631				230			

Table S5: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in test set 1.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	1,120	250	0.92	0.82	322	250	0.77	0.56
Platypus	1,138	272	0.93	0.81	337	272	0.81	0.55
VarScan	1,026	191	0.84	0.84	264	191	0.63	0.58
LoFreq	1,197	2,984	0.98	0.29	402	2,985	0.96	0.12
FreeBayes	1,206	49,731	0.99	0.02	405	49,731	0.97	0.01
SNVer	1,115	115	0.91	0.91	327	115	0.78	0.74
SAMtools	1,004	208	0.82	0.83	223	208	0.53	0.52
VarDict	1,176	339	0.96	0.78	392	340	0.94	0.54
8 tools	1,217	50,362	1.00	0.02	415	50,362	1.00	0.01
single-appreci8	1,207	158	0.99	0.88	405	161	0.98	0.72
appreci8	1,199	16	0.98	0.99	397	16	0.95	0.96
Biological truth	1,219				417			

Table S6: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in test set 2.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	223	83	0.90	0.73	62	85	0.71	0.42
Platypus	231	49	0.93	0.83	70	51	0.80	0.58
VarScan	203	75	0.82	0.73	50	77	0.57	0.39
LoFreq	244	101	0.98	0.71	85	103	0.98	0.45
FreeBayes	247	22,344	1.00	0.01	86	22,346	0.99	0.004
SNVer	233	54	0.94	0.81	73	55	0.84	0.57
SAMtools	201	66	0.81	0.75	42	68	0.48	0.38
VarDict	245	573	0.99	0.30	84	575	0.97	0.13
8 tools	248	22,357	1.00	0.01	87	22,359	1.00	0.004
single-appreci8	248	80	1.00	0.76	87	82	1.00	0.51
appreci8	248	3	1.00	0.99	87	5	1.00	0.95
Biological truth	248				87			

Table S7: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in test set 3.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	370	829	0.95	0.31	69	827	0.85	0.08
Platypus	378	2,465	0.97	0.13	72	2,463	0.89	0.03
VarScan	356	338	0.91	0.51	58	336	0.71	0.15
LoFreq	366	824	0.94	0.31	59	825	0.73	0.07
FreeBayes	388	5,340	0.99	0.07	79	5,340	0.98	0.01
SNVer	380	3,473	0.97	0.10	76	3,472	0.94	0.02
SAMtools	361	168	0.93	0.68	57	169	0.70	0.25
VarDict	388	1,163	0.99	0.25	79	1,166	0.98	0.06
8 tools	388	6,901	0.99	0.05	79	6,899	0.98	0.01
single-appreci8	387	676	0.99	0.36	78	686	0.96	0.10
appreci8	387	121	0.99	0.76	78	119	0.96	0.40
Biological truth	390				81			

Table S8: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in test set 4.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	163	60	0.86	0.73	51	60	0.70	0.46
Platypus	173	397	0.91	0.30	57	396	0.78	0.13
VarScan	139	115	0.73	0.55	45	114	0.62	0.28
LoFreq	138	9,075	0.73	0.01	39	9,070	0.53	0.004
FreeBayes	180	33,709	0.95	0.01	66	33,698	0.90	0.002
SNVer	121	2,571	0.64	0.04	49	2,570	0.67	0.02
SAMtools	133	116	0.70	0.53	33	115	0.45	0.22
VarDict	173	1,727	0.91	0.09	63	1,727	0.86	0.04
8 tools	189	34,079	0.99	0.01	73	34,064	1.00	0.002
single-appreci8	176	762	0.93	0.19	65	871	0.89	0.07
appreci8	176	94	0.93	0.65	65	79	0.89	0.45
Biological truth	190				73			

Table S9: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in test set 5.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	4005	2,75	0.92	0.94	433	270	0.77	0.62
Platypus	3,928	233	0.90	0.94	371	252	0.66	0.60
VarScan	3,100	188	0.71	0.94	291	207	0.52	0.58
LoFreq	3,460	482	0.79	0.88	494	500	0.88	0.50
FreeBayes	4,174	12,626	0.96	0.25	535	12,640	0.96	0.04
SNVer	3,169	61	0.73	0.98	457	70	0.82	0.87
SAMtools	3,740	25	0.86	0.99	252	45	0.45	0.85
VarDict	4,223	891	0.97	0.83	534	984	0.96	0.35
8 tools	4,374	13,212	1.00	0.25	553	13,230	0.99	0.04
single-appreci8	4,224	144	0.96	0.97	544	373	0.98	0.59
appreci8	4,371	5	1.00	1.00	550	16	0.99	0.97
Biological truth	4,377				556			

3 Alignment

All training- and test sets were aligned to the reference genome (GRCh37.67) using BWA mem [Li and Durbin, 2009] (version 0.7.8). The default settings were not changed. BAM- and BAI-files were generated from the SAM-files using Picard tools (version 1.118; https://broadinstitute.github.io/picard/).

4 Normalization and combination

Appreci8 performs variant calling by combining and filtering the output of eight variant calling tools. Previous analyses have shown that basically all tools show considerable differences with respect to how variants are reported [Sandmann *et al.*, 2017]. These differences do not only consider the output file format. Main differences can be identified in the way indels and MNVs are reported.

To combine the variant calling results of eight tools, we consider four steps to normalize the raw output of each caller.

1. Converting indels: This step is mainly important to handle the VarScan output. The tool reports deletions with a "-" (e.g. C>-A; normalized: CA>C) and insertions with a "+" (e.g. C>+A; normalized: C>CA).

2. Converting multi nucleotide variants (MNVs): We define an MNV as a variant for which the reported length of the reference is of the same length as the reported variant, e.g. CA>GT, but also ATG>GTC. We decided to break MNVs down to the smallest possible variants, i.e. SNVs. Thus, the MNV CA>GT is converted to a C>G and an A>T variant.

3. Checking alternative bases: If more than one alternate allele is reported, e.g. C>A,T we decided to break this variant down as well (here: C>A and C>T).

4. Finding differing strings: FreeBayes and SAMtools both report indels by considering the surrounding bases as well. If e.g. a CAA>C deletion is located within a homopolymer of 4 A's, it is reported as CAAAAC>CAAC. To normalize these variants, we compare the reference- and the alternative string – first the end, then the beginning – and check the bases for differences. Only the first non-differing base is kept (here: CAA>C). Additionally, this normalization step detects and removes calls where no base is changing. This can happen if an MNV like ATG>GTC is broken down to three distinct variants: A>G, T>T and G>C. As T>T is no variant, it is removed by this normalization step.

Investigating the tools' raw output, we found out that additional left-aligning of the indels is not necessary in case of the eight tools we consider.

Only those variants that share – after normalization – the same position, the same reference string and the same alternate string, are combined.

5 Filtering calls with appreci8

Variant calling with appreci8 involves several steps of filtration. As Figure 1 in the main paper shows, off-target calls are initially removed. Furthermore, we remove all calls located in the 3'-UTR, 5'-UTR, downstream, upstream, intron, intergenic, intragenic, protein-protein-contact and in the splice site region. Silent mutations are removed as well. These filters are exemplary and can be easily adjusted to fit every user's requirements. The remaining list of calls covers all potentially true variants of interest in the analyzed data set. Subsequently, two additional filtration steps are applied, both of them based on the calls' characteristics.

5.1 Filtration based on "Characteristics (1)"

To identify and exclude obvious artifacts due to sequencing errors, a hard filter is applied. The filter is based on parameters characterizing the raw alignment data: depth (DP), the number of alternate reads (#ALT), the variant allele frequency (VAF), the mean base quality (PHRED value) for the reference- (BQ_ref) and alternate allele (BQ_alt).

We exclude all calls with $VAF < 1\%$. Although it is possible to detect variants at lower allelic frequencies with ultra-deep sequencing, we observe background noise between $5.39 \cdot 10^{-3}$ and $6.26 \cdot 10^{-3}$ in our training sets. Furthermore, mean target coverage is only 1,366.58 in training set 1 (1,484.46 when considering only coding bases). Thus, it does not seem feasible to aim for variant calling at lower frequencies.

We also exclude calls with $DP < 50$ and/or $\#ALT < 20$. These thresholds base on the united long-time experience of molecular biology experts. We assume that it is not feasible to evaluate a variant call in targeted NGS data at lower frequencies. Altogether, only eight calls in training set 2 are excluded due to this criteria. These calls include four long deletions of 50 to 100 base pairs on CEBPA, which are covered by 4 to 22 reads. Two short deletions of 1, resp. 3 base pairs on CEBPA, which are covered by 37, resp. 3 reads. Additionally, two SNVs on RUNX1, covered by 8 to 13 reads, are excluded.

Analyzing the distribution of base quality in both training sets reveals huge differences between artifacts and true variants. Figure S1 shows the distribution of the alternate allele's mean base quality.

Although both training sets show considerable differences in the distribution, BQ_alt of true variants is higher compared to BQ_alt of artifacts in case of both training sets. A majority of true variants features BQ_alt between 30 and 40, while many artifacts feature BQ_alt below 20. Considering differences between the two training sets, we opted for excluding all calls with $BQ\_alt < 15$.

However, when considering base quality, it is not only important to consider BQ_alt, but also the difference in BQ_alt and BQ_ref. We assume that in case of a true variant BQ_alt is not expected to be significantly lower compared to BQ_ref. By contrast, BQ_alt may be significantly higher compared to BQ_ref in case of a true variant. Regarding homozygous polymorphisms, 100% of the reads are expected to contain the alternate allele. However, due to sequencing errors, a low number of reads may also contain the reference allele. In this case we expect BQ_ref to be significantly lower compared to BQ_alt.

Figure S2 shows the distribution of $BQ\_diff = BQ\_ref - BQ\_alt$ considering artifacts and true variants.

It can be observed that the distribution of BQ_diff matches our expectations in case of both training sets. Artifacts are on average characterized by a high BQ_diff value. True variants feature a maximum of $BQ\_diff = 0$. A minority of true variants is observed to feature a high negative value of BQ_diff.

Considering both training sets, we opted for excluding all calls with $BQ\_diff > 7$.

5.2 Filtration based on "Characteristics (2)"

Subsequent to the hard filtration based on "Characteristics (1)", a second, more elaborate set of characteristics is determined for the remaining calls. This set includes the presence of a variant in various databases as well as the variant's influence on the corresponding protein based on Provean1.1.5.

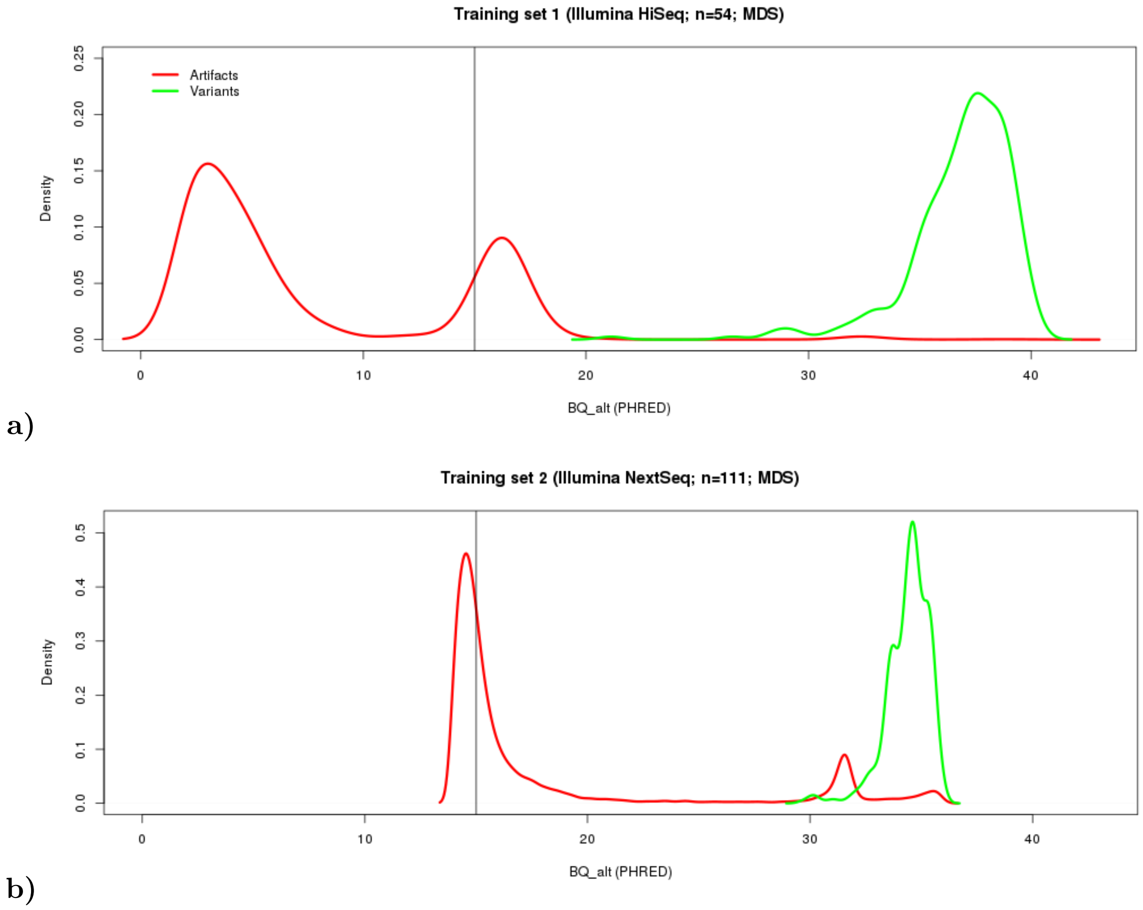

Figure S1: Distribution of mean BQ_alt (PHRED) in a) training set 1, b) training set 2.

For every call present in the two training sets, we consider the determined characteristics and the manually assigned category – mutation, polymorphism or artifact. The categorization has been validated by Sanger sequencing, re-sequencing on the same or another platform as well as manual investigation by independent biological experts. When considering this manual investigation in detail, several call characteristics can be identified that biologists consider as typical for each category. These characteristics can be straight-forward (e.g. presence of a call in the 1000 Genomes database is a typical characteristic for polymorphisms), as well as more complex (e.g. a frame-shift variant that is present in all samples, that features an allelic frequency close to 1.00 and that cannot be found in any database is typically an artifact). This expert knowledge does not only base on the two training sets, but on long-time, everyday routine in this field.

Based on this knowledge, we derived a set of 41 conditions, investigating typical characteristics of mutations, polymorphisms and artifacts. Each condition features an assigned weight indicating its importance. The concrete weight of every condition was determined exploratively, achieving best automatic classification of all calls present in the two training sets.

Twenty-nine conditions serve to separate potential variants from potential artifacts. A positive weight is assigned to those conditions, which we assume to be typical for artifacts. A negative weight is assigned to conditions, which we consider to be typical for actual mutations. Together, they make up the artifact score. For every call, every condition is evaluated. If it is true, the assigned weight is added to the initial artifact score of zero. If it is false, the score does not change. The score is calibrated in a way that a call is classified as being a potential variant, if the artifact score is below zero, i.e. in summary there is more evidence that a call is rather true but than false.

Twelve conditions serve to identify polymorphisms in the initial sets of potential variants and potential artifacts. A positive weight is assigned to those conditions, which we assume to be typical for polymorphisms. A negative weight is assigned to conditions, which we consider to be atypical for

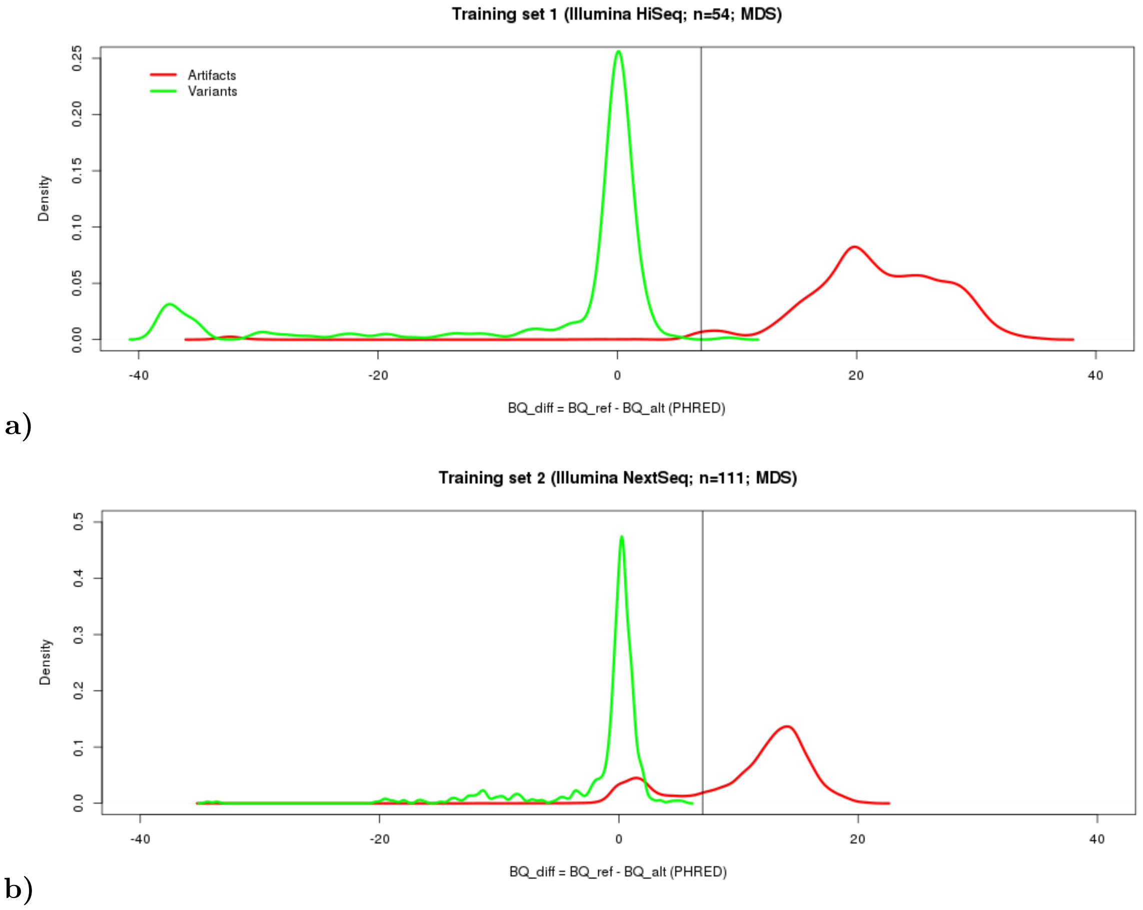

Figure S2: Distribution of $BQ\_diff = BQ\_ref - BQ\_alt$ (PHRED) in a) training set 1, b) training set 2.

polymorphisms. Together, these conditions make up the polymorphism score. Just like in case of the artifact score, every condition is evaluated in case of every call. If it is true, the assigned weight is added to the initial polymorphism score of zero. If it is false, the score does not change. To be classified as a polymorphism, a call needs a polymorphism score of at least 2.

For the appreci8 pipeline, the final analysis step involves combination of all previously determined information and calculation of the artifact- and a polymorphism score for every call by evaluating the set of 41 conditions. Figure S3 shows the conditions evaluated in case of the artifact score, their weight and the percentage of artifacts and true variants that fulfill the conditions for the two training sets.

Although we evaluated a wide set of parameters and conditions, Figure S3 clearly shows that there is no condition that perfectly separates all artifacts from all true variants. Thus, an artifact can only be identified if several conditions are mutually evaluated.

In case of both sets it can be observed that a majority of artifacts does indeed fulfill conditions with a positive weight, while true variants fulfill conditions with a negative weight. For example, $> 60\%$ of the artifacts in training set 1 and $> 40\%$ of the artifacts in training set 2 are not present in any of the considered databases, but are called in more than 50% of the samples ("1: no database AND detected >27x", resp. "1: no database AND detected >55x"). On the contrary, this condition is not fulfilled by true variants.

However, it can also be observed that some conditions with a positive weight ("2: detected >27x", resp. "2: detected >55x", "2: detected >3x" and "2: detected >3x AND VAF>0.85") are fulfilled by both artifacts and true variants. This observation is due to some especially abundant polymorphisms and hotspots. While misclassification of the hotspots is prevented by conditions like "-3: known hotspot", it can happen that some polymorphisms are initially classified as potential artifacts. However,

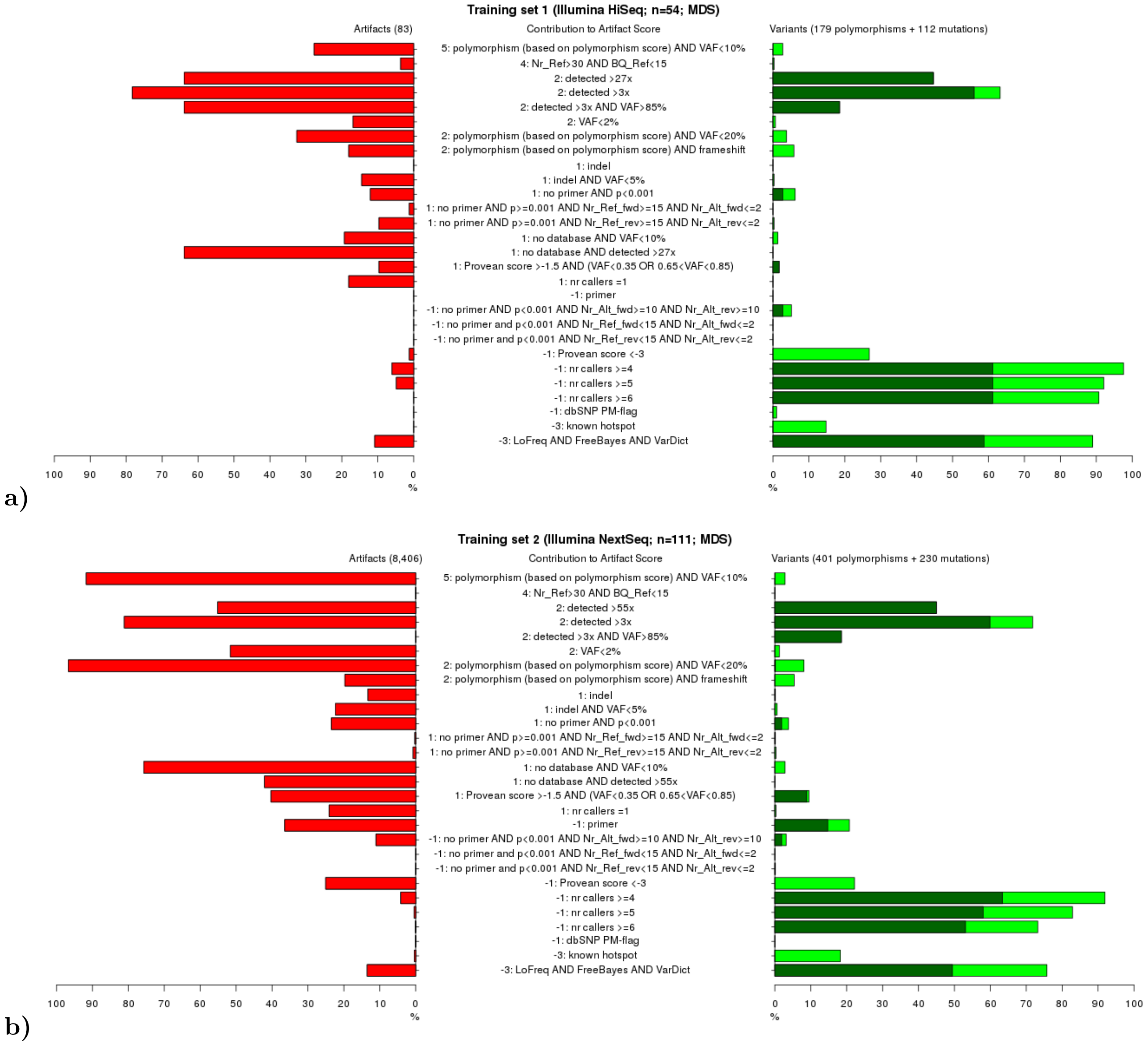

Figure S3: Conditions evaluated and their weight contributing to the artifact score. The percentage of artifacts (red) and true variants, separated into polymorphisms (dark green) and likely pathogenic mutations (light green), meeting the displayed conditions is displayed in a) training set 1, b) training set 2.

as Figure 2 of the main paper shows, evaluation of the polymorphism score in case of all calls leads to the final identification and correct classification of these polymorphisms.

The different conditions that are evaluated in terms of the polymorphism score, their weight and the percentage of artifacts and true variants that fulfill these conditions is displayed in Figure S4.

Altogether, the polymorphism score serves two purposes: separating polymorphisms from (likely pathogenic) mutations in the set of potential variants and identifying polymorphisms in the initial set of potential artifacts.

Considering the first purpose, Figure S4 shows that polymorphisms almost exclusively fulfill conditions with a positive weight. Only a low percentage of polymorphisms is detected in only one sample ("-1: detected 1x") or not present in any polymorphism database ("-1: no polymorphism database"). On the contrary, likely pathogenic mutations do either not fulfill any of the considered conditions or they fulfill conditions with a negative weight. Only a minority of mutations fulfills a condition with a positive weight. However, as the weight is only 1 per condition and a score of at least 2 is necessary

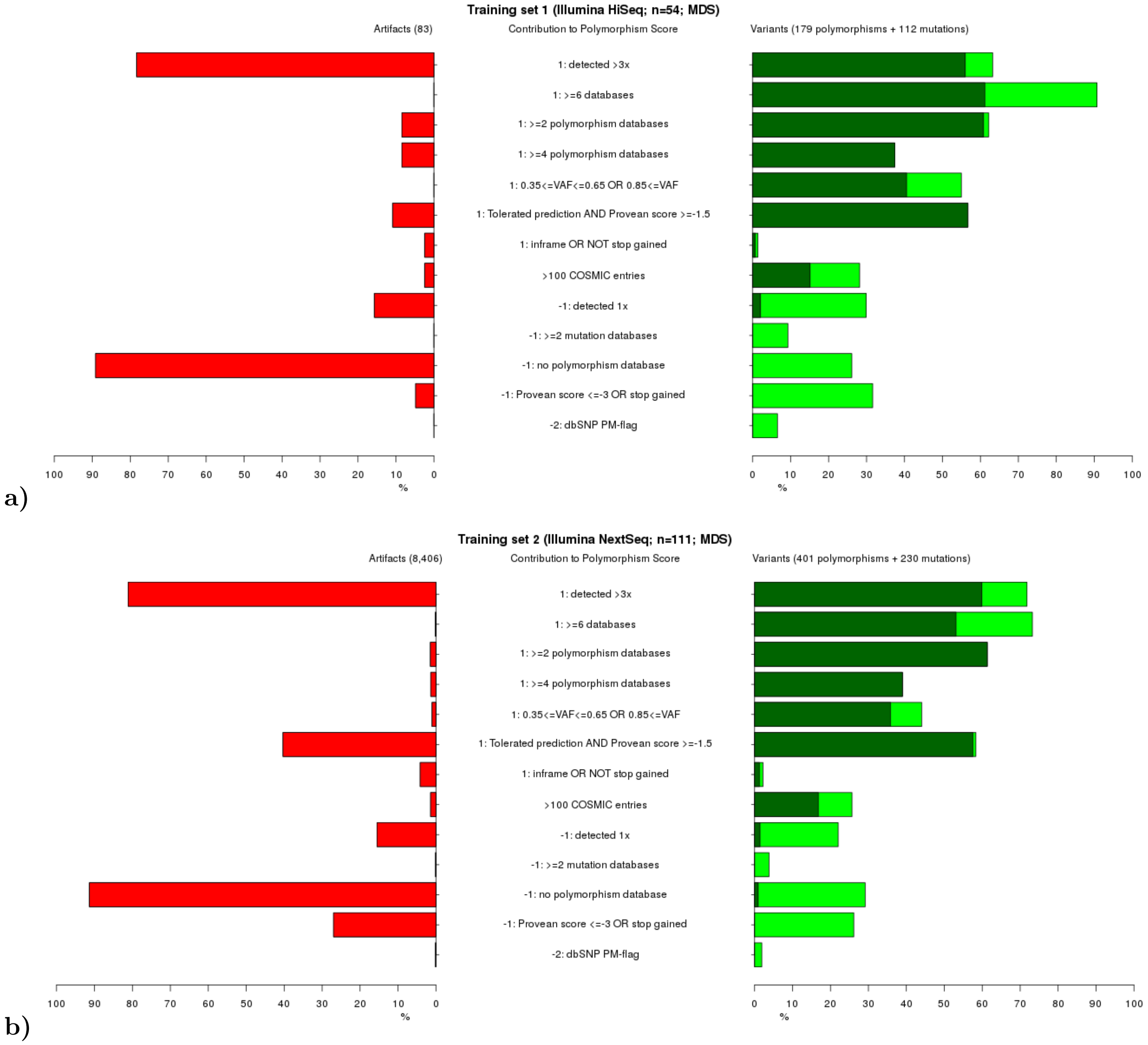

Figure S4: Conditions evaluated and their weight contributing to the polymorphism score. The percentage of artifacts (red) and true variants, separated into polymorphisms (dark green) and likely pathogenic mutations (light green), meeting the displayed conditions is displayed in a) training set 1, b) training set 2.

for classification as a polymorphism, misclassification is not expected to occur.

Considering the second purpose of our polymorphism score, Figure S4 shows that artifacts differ considerably form true variants with respect to the conditions evaluated in case of the polymorphism score. A vast majority of artifacts fulfills the conditions “1: detected >3x” and “-1: no polymorphism database”. Thus, the resulting score would be zero. However, as a call needs a polymorphism score of at least 2 to be classified as a polymorphism, the chance of artifacts being misclassified as polymorphisms can assumed to be low. On the contrary, true polymorphisms fulfill considerably more conditions with a positive weight. In training set 1, the top-four conditions with a positive weight are fulfilled by 83-97% of the polymorphisms. In training set 2, even 88-96% of the polymorphisms fulfill these conditions. As hardly any polymorphism fulfills a condition with a negative weight, a score between 3 and 4 can be expected in case of actual polymorphisms, which allows for a clear distinction from artifacts.

5.3 ROC curve

Appreci8's final decision on a call – whether it is classified as a true variant or an artifact – is influenced by several parameters. The artifact score and the threshold applied to this score has got direct influence on sensitivity and PPV. We consider ROC curves for for calling mutations with appreci8, evaluating a successive increase of the artifact score threshold. The results on the two training sets and the five test sets are displayed in Figure S5.

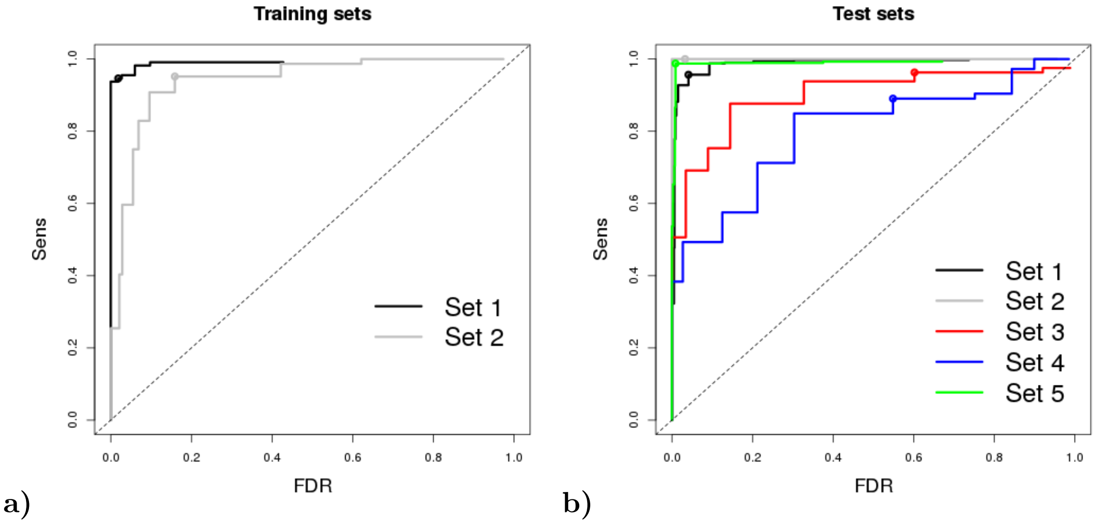

Figure S5: ROC curve for a) training sets 1 and 2, b) test sets 1-5. A successive increase of the artifact score threshold is considered. The chosen threshold – zero – is marked.

Different from classical ROC curves, we decided to plot the false discovery rate FDR, i.e. defined as $FDR = 1 - PPV$. Usually, the false positive rate FPR, that is defined as $FPR = 1 - specificity$, is plotted.

In our evaluation of tools and algorithms, we opted against the analysis of specificity. The reason lies in the considerable amount of true negative calls and the relatively low number of false positive calls. For a vast majority of tools and scenarios, specificity would be close to 1. Thus, this parameter provides little information. In contrast to this, PPV considers the number of true positive and false positive calls. Values between 0 and 1 can usually be observed.

In Figure S5, it can be observed that most ROC curves do not end in (1,1). This observation is due to the filtration based on coverage and base quality that is performed prior to evaluating the artifact score. Some true mutations are filtered by this step. The threshold applied on the artifact score does not have any effect on these calls. In this case, the number of false negative calls is always > 0. Thus, sensitivity is always < 1.00 (see test set 3). Additionally, several artifacts are filtered by this step. In this case, the number of false positive calls is always > 0. Thus, the FDR is always < 1.00 (see e.g. training set 1).

In all cases, the ROC curves are clearly above the diagonal. Consequently, discrimination between true mutations and artifacts by the help of appreci8 is much better than random guess. Furthermore, evaluation of the training sets shows that the selected threshold of zero leads to the best compromise between high sensitivity and low FDR. Similar results can be observed in case of test sets 1, 2 and 5. Both test sets 3 and 4 feature a relatively high FDR. Decreasing the applied threshold to -1 (test set 4), or even -2 (test set 3) would be an option to improve FDR. However, it has to be considered that this increase in FDR would lead to a decrease in sensitivity.

6 8 tools

Appreci8's performance is compared to an alternative approach we refer to as '8 tools'. This approach considers all variants that have been reported by at least one out of eight tools. The approach follows the appreci8-pipeline, but leaves out the last filtration step, involving artifact- and polymorphism score calculation.

The method ‘8 tools’ allows us to consider two important aspects of variant calling: First, the number of true variants. As not a single tool succeeds in detecting all mutations present in the two training sets and in the five test sets, ‘8 tools’ allows us to study if different tools using different algorithms for variant calling tend to miss the same true variants or not. Second, the number of called artifacts. The number of reported artifacts per tools varies considerably. ‘8 tools’ allows us to study if different tools tend to call the same artifacts despite using different algorithms for variant calling.

7 Single-appreci8

In addition to ‘8 tools’, appreci8 is also compared to an alternative approach we refer to as ‘single-appreci8’. It is an experimental variant of our algorithm. The approach follows the original appreci8-pipeline (including the final filtration step). However, every sample is evaluated independently. Information on other samples analyzed in the same run, featuring the same calls is disregarded.

Considering the artifact score, the conditions “2: detected >27x, resp. >55x”, “2: detected >3x”, “2: detected >3x AND VAF>85%” and “1: no database AND detected >27x, resp. >55x” are disabled. As all conditions share a positive weight contributing to the artifact score, it is expected that less artifacts can successfully be identified by ‘single-appreci8’.

Considering the polymorphism score, the condition “1: detected >3x” is disabled, while the condition “-1: detected 1x” is always true for independently analyzed samples.

The method ‘single-appreci8’ allows us to analyze the extend by which different data sets as well as our pipeline are influenced by re-occuring variants that are often systematic sequencing artifacts.

8 Alternative approach

Using eight variant calling tools instead of only one has got negative influence on run-time. Furthermore, appreci8 necessitates complex post-processing of the raw results. A solution involving less than eight tools would be beneficial.

We considered all possible combinations of the eight variant calling tools – from one to eight tools. A call is categorized as true if it is reported by all tools of the considered combination. For both training sets as well as for the five test sets, we determined sensitivity, PPV and the F1 score for all possible combinations. The results can be found in Supplementary Data S9.

Using the training sets as a basis, we determined the best combinations of one to eight tools. The best combination is assumed to feature the highest average F1 score over both training sets. We also took the best combination for training set 1 and the best combination for training set 2 into account. The results can be found in tables S10 and S11.

Considering the best combinations of tools, it becomes obvious that it is not beneficial to combine tools with different detection algorithms, e.g. VarScan (heuristic/statistical method), SNVer (frequentist approach), LoFreq (Poisson-binomial distribution) and GATK, Platypus, FreeBayes or SAMtools (bayesian approach).

Regarding the performance of the different combinations in comparison to appreci8, it can be observed that our solution features a better, more stable performance over all analyzed data sets in comparison to any other solution.

9 Calculating background noise

Background noise is calculated for ever sample in the training set, as well as in the test set.

We consider an exemplary sample s with target region t. Every position t_i with $i = 1, ..., n$ is investigated and compared to the reference. If 100 reads can be mapped to t_1 (reference base G) and 98 out of 100 reads feature a G, while 2 reads feature a C, then the background noise for t_1 would be 2/100=2%. The overall background noise for sample s is the mean over all position-specific background noises.

Table S10: Sensitivity, PPV and F1 score for training sets 1 and 2 considering the best combination of one to eight tools over both training sets, the best combination based on training set 1, the best combination based on training set 2 and appreci8.

Best combination of	Tools	Training set 1			Training set 2		
		Sens	PPV	F1	Sens	PPV	F1
1 tool	GATK	0.92	0.85	0.88	0.82	0.71	0.76
2 tools	Platypus, VarDict	0.91	0.99	0.95	0.88	0.96	0.92
3 tools	Platypus, FreeBayes, VarDict	0.91	0.99	0.95	0.88	0.96	0.92
4 tools	GATK, Platypus, FreeBayes, VarDict	0.90	1.00	0.95	0.78	0.98	0.87
5 tools	GATK, Platypus, LoFreq, FreeBayes, VarDict	0.83	1.00	0.91	0.70	0.98	0.82
6 tools	GATK, Platypus, LoFreq, FreeBayes, SAMtools, VarDict	0.81	1.00	0.90	0.57	0.99	0.72
7 tools	GATK, Platypus, VarScan, LoFreq, FreeBayes, SAMtools, VarDict	0.80	1.00	0.89	0.38	0.99	0.55
8 tools	GATK, Platypus, VarScan, LoFreq, FreeBayes, SNVer SAMtools, VarDict	0.76	1.00	0.87	0.17	0.98	0.28
Training set 1	VarDict	0.97	0.96	0.97	0.94	0.15	0.25
Training set 2	Platypus, FreeBayes, VarDict	0.91	0.99	0.95	0.88	0.96	0.92
appreci8		0.98	0.99	0.98	0.98	0.94	0.96

Table S11: Sensitivity, PPV and F1 score for test sets 1 to 5 considering the best combination of one to eight tools over both training sets, the best combination based on training set 1, the best combination based on training set 2 and appreci8.

Best combination of	Test set 1			Test set 2			Test set 3			Test set 4			Test set 5		
	Sens	PPV	F1	Sens	PPV	F1	Sens	PPV	F1	Sens	PPV	F1	Sens	PPV	F1
1 tool	0.92	0.82	0.87	0.90	0.73	0.81	0.95	0.31	0.47	0.86	0.73	0.79	0.81	0.88	0.85
2 tools	0.91	0.99	0.95	0.92	0.99	0.96	0.97	0.50	0.66	0.85	0.95	0.89	0.83	0.88	0.86
3 tools	0.90	0.99	0.95	0.92	1.00	0.96	0.97	0.65	0.78	0.83	0.95	0.88	0.81	0.88	0.85
4 tools	0.89	0.99	0.94	0.88	1.00	0.94	0.94	0.79	0.86	0.80	0.97	0.88	0.79	0.93	0.85
5 tools	0.88	0.99	0.93	0.87	1.00	0.93	0.89	0.97	0.93	0.62	0.99	0.76	0.61	0.93	0.74
6 tools	0.80	1.00	0.89	0.79	1.00	0.88	0.88	0.99	0.93	0.59	1.00	0.75	0.59	0.98	0.73
7 tools	0.77	1.00	0.87	0.75	1.00	0.86	0.86	1.00	0.92	0.56	1.00	0.72	0.55	0.98	0.71
8 tools	0.73	1.00	0.84	0.72	1.00	0.84	0.85	1.00	0.92	0.32	1.00	0.48	0.31	0.98	0.47
Training set 1	0.97	0.78	0.86	0.99	0.30	0.46	0.99	0.25	0.40	0.91	0.09	0.17	0.90	0.16	0.28
Training set 2	0.90	0.99	0.95	0.92	1.00	0.96	0.97	0.65	0.78	0.83	0.95	0.88	0.84	0.37	0.52
appreci8	0.98	0.99	0.98	1.00	0.99	0.99	0.98	0.76	0.86	0.93	0.65	0.77	1.00	1.00	1.00

To account for true variants, we accept more than one reference base at positions where sample s is known to feature a polymorphism or a pathogenic mutation. For example, if 100 reads can be mapped to t_2 (reference T), sample s is known to feature the heterozygous polymorphism T>G at t_2, and out of 100 reads 50 feature a T, 49 feature a G and 1 features a C, the background noise for t_2 would be 1/100=1% and not 50/100=50%.

The background noise per data set is determined as the mean over the background noise per sample.

10 Calling mutations

The detection of actual mutations is expected to be more challenging compared to polymorphisms. This is mainly due to the on average lower allelic frequencies of pathogenic mutations. Polymorphisms are germline variants and are therefore generally observed in 50 or 100% of the reads, whereas hematological malignancies have a (sub)clonal composition and mutations can occur at considerably lower frequencies.

Appreci8 is able to distinguish likely pathogenic mutations from polymorphisms. To train and test our pipeline, we analyze every individual tool's and our pipeline's ability to call those variants that are categorized as likely pathogenic on the basis of at least two independent experts. To separate likely pathogenic mutations from polymorphisms, information on the VAF, the number of samples featuring the same variant, the predicted effect and the presence of a variant in polymorphism- as well as mutation databases is evaluated. Additionally, we consider information on common hotspot mutations (e.g. [Haferlach *et al.*, 2014], [Papaemmanuil *et al.*, 2013]). This approach allows us to categorize even those variants that are of “unknown significance” according to ClinVar in the absence of matching germline samples.

Variants that have been categorized as likely pathogenic by appreci8, but identified as polymorphisms, are considered as being false positives. All variants that are successfully identified as polymorphisms by appreci8 are excluded from analysis, also when evaluating the individual tools.

10.1 Training appreci8

Sensitivity and PPV of the different variant calling approaches, considering only likely pathogenic mutations in the two training sets, is displayed in Figure S6 (for details see Supplementary Tables S3 and S4).

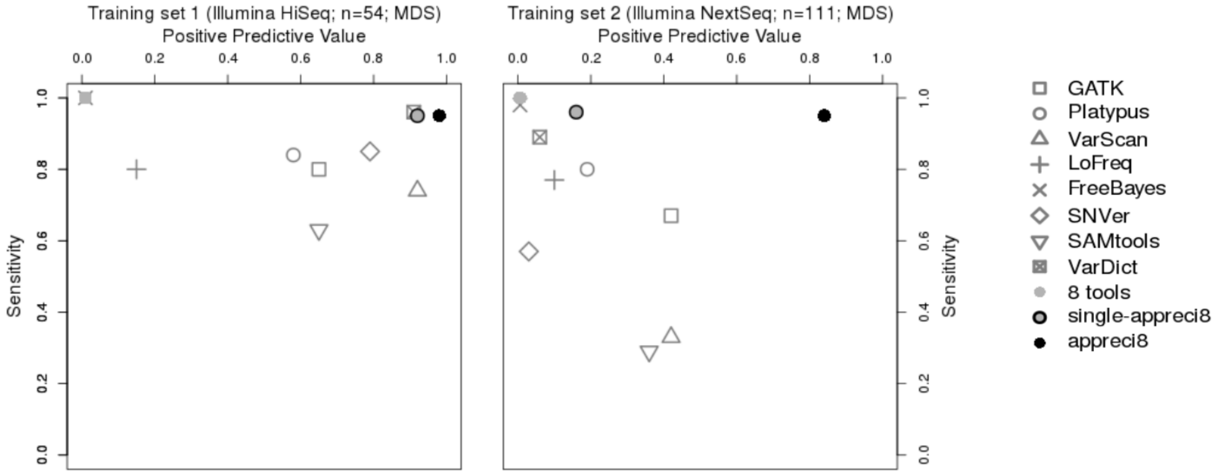

Figure S6: Relation between positive predictive value and sensitivity in case of GATK, Platypus, VarScan, LoFreq, FreeBayes, SNVer, SAMtools, VarDict, the combined output of all tools (8 tools), single-appreci8 and appreci8 in training sets 1 and 2 considering only likely pathogenic mutations.

When comparing Figure S6 to Figure 3 in the main paper, it can be observed that sensitivity and PPV are – on average – lower when concentrating on the detection of likely pathogenic mutations. This observation matches our expectations that the successful calling of mutations is more challenging compared to polymorphisms.

Considering training set 1, every individual tool’s sensitivity ranges between 0.63 and 0.83. The only exceptions from this observation are FreeBayes ($sens = 1.00$) and VarDict ($sens = 0.96$). Regarding PPV, not a single individual tool features a value above 0.95. Instead, PPV ranges between 0.01 (FreeBayes) and 0.92 (VarScan). Combining the output of all tools results in perfect sensitivity and $PPV = 0.01$. Just like in case of calling all variants, application of appreci8 leads to a considerable increase in PPV and only to a minor decrease in sensitivity ($sens = 0.95$, $PPV = 0.98$).

With respect to training set 2, similar results are observed. FreeBayes features highest sensitivity ($sens = 0.98$), while sensitivity of most of the other tool ranges between 0.56 and 0.89. Especially low sensitivity can be observed in case of VarScan (0.33) and SAMtools (0.29). PPV ranges between 0.01 (FreeBayes) and 0.42 (GATK). By combining the output of all tools, sensitivity increases to 1.00, while PPV decreases to 0.01. Application of appreci8 leads to a considerable improvement of the results, especially in case of its actual functionality. Sensitivity decreases slightly (0.95), while PPV shows a significant increase to 0.84, which is twice as high as the highest PPV that can be observed in case of the individual tools.

10.2 Testing appreci8

In order to test appreci8's performance regarding the automatic identification of likely pathogenic mutations, we consider the same five data sets that have already been analyzed in section "Testing appreci8" of the main paper. The results regarding sensitivity and PPV are summed up in Figure S7 (for details see Supplementary Tables S5 to S9).

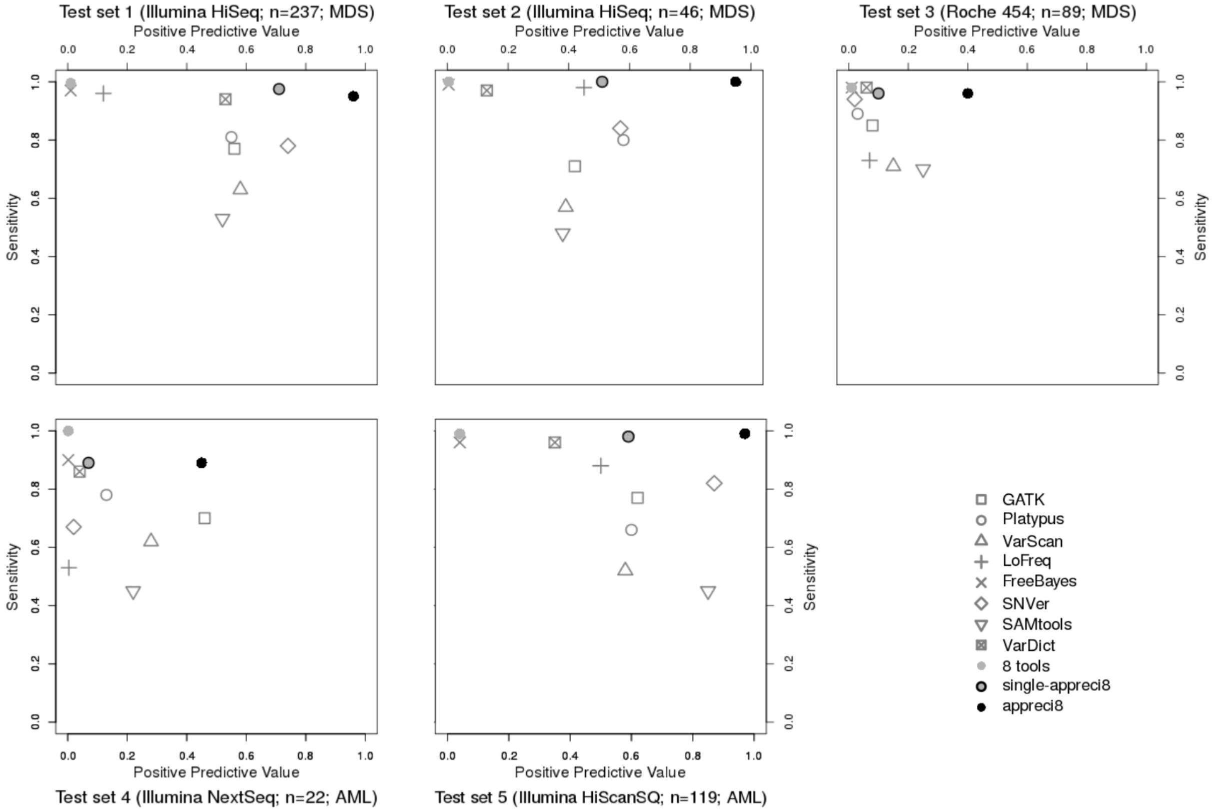

Figure S7: Relation between positive predictive value and sensitivity in case of GATK, Platypus, VarScan, LoFreq, FreeBayes, SNVer, SAMtools, VarDict, the combined output of all tools (8 tools), single-appreci8 and appreci8 in test sets 1-5 considering only likely pathogenic mutations.

When comparing Figure S7 to Figure 4 in the main paper, an overall shift to lower sensitivity and PPV is observed. Considering test sets 1 and 2, FreeBayes, LoFreq and VarDict still feature highest sensitivity by far, ranging between 0.94 and 0.99, while the remaining individual tools partly feature sensitivity as low as 0.48 (test set 2, SAMtools). PPV ranges between 0.004 (test set 2, FreeBayes) and 0.74 (test set 1, SNVer). In contrast, the application of appreci8 leads to results with equally high sensitivity and PPV (test set 1: $sens = 0.95$, $PPV = 0.96$; test set 2: $sens = 1.00$, $PPV = 0.95$).

Considering the detection of actual mutations in Roche 454 data (test set 3) a considerable decrease in PPV can be observed in comparison to Figure 4. Sensitivity ranges between 0.70 (SAMtools) and 0.98 (FreeBayes and VarDict), while PPV ranges between 0.01 (FreeBayes) and 0.25 (SAMtools). Although appreci8's performance on Roche 454 data is worse than in case of Illumina data, it still outperforms every individual tool ($sens = 0.96$, $PPV = 0.40$).

Similar observations can be made when considering test sets 4 and 5. In general, the individual tools show a decrease in sensitivity and PPV. However, appreci8 still succeeds in calling mutations with sensitivity comparable to the best individual tools (test set 4: Freebayes $sens = 0.90$, appreci8 $sens = 0.89$; test set 5: FreeBayes $sens = 0.96$, appreci8 $sens = 0.99$), while PPV is at least comparable (test set 4: GATK $PPV = 0.46$, appreci8 $PPV = 0.45$) or even better than the best individual tool (test set 5: SAMtools $PPV = 0.87$, appreci8 $PPV = 0.97$).

11 Calling variants in an Ion Torrent data set

We developed appreci8 by analyzing two well-characterized sets of Illumina data. In our manuscript, we tested appreci8 on four independent sets of Illumina data and one set of Roche 454 data.

Another widely used sequencing platform is Ion Torrent. The sequencing technique differs from Illumina as well as Roche 454. Data is characterized by a different error profile, conditioned by the sequencing technique itself. Nevertheless, we decided to test our method on an exemplary Ion Torrent data set. We analyzed 24 targeted sequencing samples, sequenced on Ion Torrent PGM. All samples result from patients with AML. The target region covers 30,671 bp. All samples are available as duplicates. Additionally, all samples were re-sequenced on Illumina NextSeq.

The Ion Torrent data set was analyzed according to the standard appreci8 pipeline. Thresholds were not changed ($\#ALT \geq 20$, DP$\geq$ 50, VAF$\geq$ 1%, $BQ\_alt \geq 15$ and $BQ\_diff \leq 7$). The results of the eight individual tools and our appreci8 approach are summed up in table S12. Detailed results can be found in Supplementary Data S10.

Table S12: Number of variants, artifacts, sensitivity and PPV considering variant calling in general and mutation calling in particular (expert-based classification) in the Ion Torrent data set.

Approach	Calling variants				Calling mutations			
	Variants	Artifacts	Sens	PPV	Mutations	Artifacts	Sens	PPV
GATK	165	2,249	0.85	0.07	67	2,249	0.77	0.03
Platypus	157	3,007	0.81	0.05	60	3,007	0.69	0.02
VarScan	161	694	0.83	0.19	63	694	0.72	0.08
LoFreq	154	21	0.80	0.88	56	21	0.64	0.73
FreeBayes	185	20,001	0.96	0.01	79	20,001	0.91	0.00
SNVer	182	6,392	0.94	0.03	80	6,392	0.92	0.01
SAMtools	136	66	0.70	0.67	44	66	0.51	0.40
VarDict	165	1,172	0.85	0.12	69	1,172	0.79	0.06
appreci8	177	3	0.92	0.98	73	3	0.84	0.96
Biological truth	193				87			

Table S12 indicates that appreci8 is not only applicable on Illumina or Roche 454 data, but also on Ion Torrent data. Two tools, FreeBayes and SNVer perform slightly better with respect to sensitivity. However, PPV is lowest for these tools. Furthermore, it has to be noted that that six out of 14 mutations missed by appreci8 are highly difficult to detect with the help of Ion Torrent sequencing data (insertion of a G in a homopolymeric stretch of eight G's).

Considering both, sensitivity and PPV, appreci8 outperforms the eight individual tools with respect to variant calling in general as well as mutation calling in particular.

12 Calling variants in a public data set

To test our method on a completely independent data set, we analyzed a public targeted sequencing data set available at the Sequence Read Archive (PRJEB14077). Colon Adenoma Cells have been studied using – among others – a targeted colorectal cancer DNA sequencing panel with 71 different oncogenes and tumor suppressor genes often mutated in colorectal cancers [Dame *et al.*, 2018].

As the panel itself was not published, we had to deduce a target region on the basis of the reported mutations. We conclude that if a mutation in a certain gene is reported, the gene had to be part of the target region. Altogether, we could identify 41 genes that had to be part of the target panel. For these genes, we include the corresponding coding regions.

The public data set differs from our training- and test sets in a way that matching controls were available. For alignment and variant calling, a commercial matched-sample approach was used (QIAseq DNA enrichment portal; allele frequency threshold for variant calling 5%). Thus, no information on germline calls or systematic artifacts, present in the tumor and control samples, was available.

When analyzing the data, we followed our usual pipeline – using BWA mem for alignment and appreci8 for variant calling. The detection threshold was increased to 5% - just like in case of the commercial software. We decided to analyze the normal samples and the tumor samples independently.

Subsequently, the results were combined and evaluated manually. This way, we are able to report somatic as well as germline calls. The results of the eight individual tools and our appreci8 approach can be found in table S13. Detailed results can be found in Supplementary Data S8.

Table S13: Number of somatic variants, germline variants, artifacts, sensitivity and PPV considering variant calling in a public targeted sequencing data set.

Approach	Somatic variants	Germline variants	Artifacts	Sensitivity	PPV
GATK	57	950	1	0.98	1.00
Platypus	60	953	0	0.99	1.00
VarScan	34	737	4	0.75	0.99
LoFreq	41	748	152	0.77	0.84
FreeBayes	70	954	746	1.00	0.58
SNVer	69	956	142	1.00	0.88
SAMtools	21	898	3	0.90	1.00
VarDict	62	945	138	0.98	0.88
appreci8	62	950	0	0.99	1.00
Biological truth	70	956			

A call is classified as a somatic variant if it is present on the published list of mutations and cannot be found in the normal control. It was interesting to observe that five reported calls were also detected in the normal samples with high coverage and a comparable frequency. We inspected all calls manually. As data quality was high, we decided to categorize these five calls as germline variants.

In addition to the reported somatic variants, we found two additional variants that are likely to be somatic mutations as well.

Regarding the germline variants, no information was published. Assuming that a true germline call has to be present in both, the control and the tumor sample, the two files per sample can be interpreted and analyzed as re-sequencing experiments. If a variant has been called in the tumor and in the normal sample with a comparable allele frequency, the call can be a germline call. Taking the presence of the variant in databases, the prediction, the allele frequency itself and the base quality into account, we came to a decision. Additional manual inspection was performed with the help of the IGV.

For calling variants with appreci8, we neither used a hotspot list nor primer information. If this information was available, we expect it to have positive influence on variant calling with appreci8. Still, our pipeline performs as good as the best individual tool, i.e. in this case Platypus. In general, it can be observed that all tools perform better with respect to PPV, compared to our training- and test sets. This is likely to be due to the fact that we increased the detection threshold to 5%.

13 Analyzing the influence of reoccurring variants

Three out of five test sets feature the same target region compared to our training sets. Every test set features mutations and polymorphisms that have also been detected in the training sets. We investigate, whether our appreci8 pipeline performs better in case of reoccurring variants, i.e. mutations or polymorphisms that have already been called in the training sets, compared to unknown variants. Furthermore, we consider the frequency of the variant in the training sets. The results are summed up in table S14.

Altogether, we consider five different frequency categories: 0% for variants that have not been detected in the training sets; 0%< x ≤1% for unique variants (1/165=0.61%); 1%< x ≤10% for rare variants (2 to 16 out of 165 samples); 10%< x ≤50% for relatively common variants (17 to 82 out of 165 samples); >50% for common variants (more than 82 out of 165 samples).

The combined results of training sets 1 and 2 indicate that unique variants are more difficult to classify ($sens$ = 0.93) compared to variants present in more than 10% of the samples ($sens$ = 1.00). However, when considering the five test sets, only one out of five sets supports this hypothesis. In case of all the other test sets, perfect sensitivity can be observed, when considering unique variants. It has

Table S14: Number of true positives, false negatives and sensitivity in both training sets and test sets 1 to 5. Results are reported, dependent on the variant frequency in the training sets.

Data set	Variant frequency in training sets	True positives	False negatives	Sens
Training sets 1 and 2	0%	/	/	/
	0%< x ≤1%	190	15	0.93
	1%< x ≤10%	204	2	0.99
	10%< x ≤50%	97	0	1.00
	>50%	414	0	1.00
Test set 1	0%	238	15	0.94
	0%< x ≤1%	58	2	0.97
	1%< x ≤10%	203	3	0.99
	10%< x ≤50%	128	0	1.00
	>50%	572	0	1.00
Test set 2	0%	42	0	1.00
	0%< x ≤1%	10	0	1.00
	1%< x ≤10%	44	0	1.00
	10%< x ≤50%	38	0	1.00
	>50%	114	0	1.00
Test set 3	0%	35	2	0.95
	0%< x ≤1%	32	0	1.00
	1%< x ≤10%	67	1	0.99
	10%< x ≤50%	29	0	1.00
	>50%	224	0	1.00
Test set 4	0%	90	12	0.88
	0%< x ≤1%	6	0	1.00
	1%< x ≤10%	17	2	0.89
	10%< x ≤50%	8	0	1.00
	>50%	55	0	1.00
Test set 5	0%	3,88	6	1.00
	0%< x ≤1%	23	0	1.00
	1%< x ≤10%	136	1	0.99
	10%< x ≤50%	46	0	1.00
	>50%	286	0	1.00

to be noted that the low number of false negative calls makes interpretation of the results difficult. This also affects the analysis of variants that have not been detected in the training sets. In test sets 1, 3 and 4, sensitivity is lowest for this category, indicating that our pipeline works best for variants known from the training sets. However, sensitivity is only slightly lower compared to all the other analyzed scenarios. Furthermore, test sets 2 and 5 show perfect sensitivity for this category.

Altogether, our results indicate that appreci8's performance is only marginally influenced by the variant frequency in the training sets. The correct classification of variants that have not been detected in the training sets does, in general, not result in inferior performance of appreci8.

Supplementary Data S1

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in training set 1.

Supplementary Data S2

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in training set 2.

Supplementary Data S3

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in test set 1.

Supplementary Data S4

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in test set 2.

Supplementary Data S5

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in test set 3.

Supplementary Data S6

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in test set 4.

Supplementary Data S7

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in test set 5.

Supplementary Data S8

Tables containing all true positive somatic variants, false negative somatic variants, false positive somatic variants, true positive germline variants, false negative germline variants, false positive germline variants and true negative calls (at two stages of filtration) detected by appreci8 in a public data set. Additionally, variants excluded from consideration are reported.

Supplementary Data S9

Sensitivity, PPV and F1 score for all possible combinations of 1 to 8 tools considering training sets 1 and 2 and test sets 1 to 5.

Supplementary Data S10

Tables containing all true positive mutations, true positive polymorphisms, false negative mutations, false negative polymorphisms, false positive mutations, false positive polymorphisms and true negative mutations (at two stages of filtration) detected by appreci8 in the Ion Torrent data set.

References

[Dame *et al.*, 2018] Dame,M.K., Attili,D., McClintock,S.D., Dedhia,P.H., Ouillette,P., Hardt,O., Chin,A.M., Xue,X., Laliberte,J., Katz,E.L. *et al.* (2018) Identification, Isolation, and Characterization of Human LGR5-positive Colon Adenoma Cells. *Development*, **145**, dev.153049.

[Haferlach *et al.*, 2014] Haferlach T, Nagata Y, Grossmann V, Okuno Y, Bacher U, Nagae G, Schnittger S, Sanada M, Kon A, Alpermann T, *et al.* (2014) Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, **28**, 241–247.

[Li and Durbin, 2009] Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 14.

[Papaemmanuil *et al.*, 2013] Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, Yoon CJ, Ellis P, Wedge DC, Pellagatti A, *et al.* (2013) Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, **122**, 3616-3627.

[Sandmann *et al.*, 2017] Sandmann,S., de Graaf,A.O., Karimi,M., van der Reijden,B.A., Hellström-Lindberg,E., Jansen,J.H. and Dugas,M. (2017) Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci Rep*, **7**, 43169.