Supplementary material to "Feature based classifiers for somatic mutation detection in tumour-normal paired sequencing data"

Jiarui Ding [1,2], Ali Bashashati [1], Andrew Roth [1], Arusha Oloumi [1],
Kane Tse [4], Thomas Zeng [4], Gholamreza Haffari [1], Martin Hirst [4], Marco A. Marra [4],
Anne Condon [2], Samuel Aparicio [1,3]*and Sohrab P. Shah [1,2,3]†
[1]Department of Molecular Oncology, BC Cancer Agency, Vancouver, BC, Canada.
[2]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada.
[3]Department of Pathology, University of British Columbia, Vancouver, BC, Canada.
[4]Canada's Michael Smith Genome Science Centre, Vancouver, BC, Canada.

Details of methods and hyper-parameter selection

Random Forests: RF provides a special out-of-bag (OOB) error. If a feature vector $\mathbf{x}$ is not in a bootstrap sample, we call it an OOB sample. The OOB samples are tested on the trees built without using the OOB samples. The OOB error is the proportion of times that the OOB samples are misclassified. The OOB error is almost identical to the cross-validation error. More importantly, the variable importance can be estimated by the use of OOB samples. The process is as follows: first, when a tree is grown from a bootstrap sample, the OOB samples are tested and the number of correct votes v_1 is recorded. Second, the jth features of the OOB samples are permuted and retested, the new number of correct votes v_{2j} is also recorded. The average of $(v_1 - v_{2j})$ over all trees is the importance measure of feature j. To select the small number of highly relevant features, a backward feature elimination method is used. We first train RFs with all the features, compute the feature importance and the out-of-bag error rate. Then we iteratively remove the least important features, retrain RFs, and compute the OOB error rate. If the OOB error rate increases abruptly (less than the best OOB error rate minus its one standard error), we stop eliminating features and get the final feature set.

To train a RF model, we should provide the number of trees B and the the number of features p selected at each node when growing a tree. The number of pieces M the feature space will be split into are determined by the greedy training algorithms. For RF, because increasing B doesn't cause RF to overfit, we only need to provide a sufficient large number. In our experiments, we fixed $B = 1000$ and only tuned p using cross-validation. p was chosen among $2^0, 2^1, 2^2, 2^3, 2^4, 2^5,$ and 2^6.

*correspondence regarding ethical consent on datasets should be directed to: saparicio@bccrc.ca
†correspondence on the methods and results should be directed to: sshah@bccrc.ca

Bayesian additive regression tree: BART is a fully Bayesian model, and all the parameters are given priors and Markov-chain Mote Carlo sampling is used for inference. Specifically, for tree j, we assume it will split the features into M_j pieces. let $\boldsymbol{\mu}_m = (\mu_1, \ldots, \mu_m, \ldots, \mu_{M_j})$, μ_m with prior:

$$\mu_m \sim N(0, \sigma_\mu^2) \text{ and } \sigma_\mu = 0.5/k\sqrt{B}$$

where k is the parameter which shrinks the response of each individual tree to 0 thus decrease each individual tree's influence to the final prediction (note the dependent variable is rescaled to the interval [-0.5 0.5]). The σ prior (see the main text) is an inverse chi-square distribution. As BART's performance is very robust to the chosen of the parameters of the chi-square distribution. We used the default setting and only used cross-validation to choose B and k. B was chosen among $100, 200$ and k was chosen among $2^{-2}, 2^{-1}, 2^0, 2^1$ and 2^2.

BART estimates the importance of each variable based on the appearance frequency of the variable in growing trees, more precisely, the average appearance frequency of each variable in many MCMC draws. In our case, the features with appearance frequency larger than the average appearance $1/106$ were kept.

Support vector machine: Although SVM can use different kernels, we found that in cross-validations, the non-linear kernels such as the Gaussian kernel were very flat and gave linear separating hyper-planes. The results suggested that linear kernels was sufficient. Also, for the ease of feature selection, it's better to adopt the linear kernel. Linear kernel doesn't have any hyper-parameters so we only need to choose the trade-off hyper-parameter c. c was chosen among $2^{-8}, \ldots, 2^4$.

L1 regularized logistic regression: The Logit model only has one hyper-parameter: the scale parameter ρ. ρ was chosen among $2^{-4}, \ldots, 2^8$.

Select the number of cluster for wildtypes

We use the Bayesian information criteria (BIC) score to select the number of clusters for the Gaussian mixture model because the BIC score is the "standard" tool for this purpose [Fraley and Raftery, 2007]. The BIC score is an approximation to the likelihood given a model with parameter vector θ (in our case, a mixture of K Gaussian distributions with means and covariance matrixes as parameters: $\theta = (\mu_k, \Sigma_k)_{k=1}^K$), and is defined as

$$\text{BIC(K)} \simeq P(Data \mid \theta) \simeq 2 \ln p(Data \mid \theta^*) - M * \ln(N) \tag{1}$$

where $Data$ is the observed data, θ^* is an estimation of θ, M is the number of parameters and N is the number of data points. We use the "standard" package for Gaussian mixture model based clustering MCLUST [Fraley and Raftery, 2006] with default Bayesian regulations parameters (Maximum a posteriori estimation of parameters). We use Bayesian regulation parameters because the traditional BIC score is computed based on maximum likelihood estimation of the parameters of the Gaussian mixture model. The maximum likelihood estimation of the Gaussian mixture model parameters can be poorly behaved especially when K is large and the number of data points is small because of overfitting (recall the Gaussian probabilistic density function can be infinite).

The Expectation-Maximization (EM) algorithm is used to estimate the parameters of the Gaussian mixture model. The EM algorithm is very sensitive to the initialization parameters. To find the "best" clustering results, normally multiple runs with different initialization are conducted and only the best results are returned. For our experiment, for each number of components K range from three to nine, 5000 runs are conducted for each K and only the results with the best BIC score are returned.

The final results are given in Figure S5. As can be seen from Figure S5(a), when $K = 6$, the corresponding BIC score is the largest. However, one cluster has only 6 wildtypes. By comparing the clustering results when $K = 5$ and $K = 6$, the additional cluster with $K = 6$ includes some "outliers" which have small probability when assigning them to other clusters. The heatmaps with $K = 6$ are given in Figure S5. The R scripts to generate all the results can be abstained from our package.

References

C. Fraley and A.E. Raftery. Mclust version 3 for r: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics, 2006.

C. Fraley and A.E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

Supplementary figures

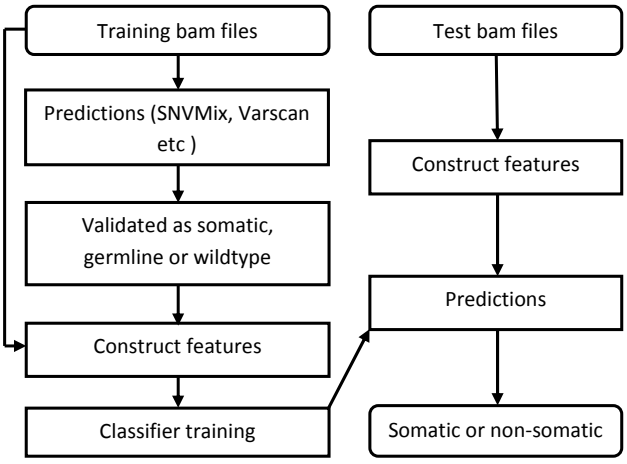


Figure 1: The workflow of the feature-based classifier for somatic mutation prediction from next generation sequencing data. Given test bam files, each candidate site is represented by a feature vector, and the classifier trained on validated ground truth data is applied to make a prediction. The classifier outputs the probability of each site being somatic.

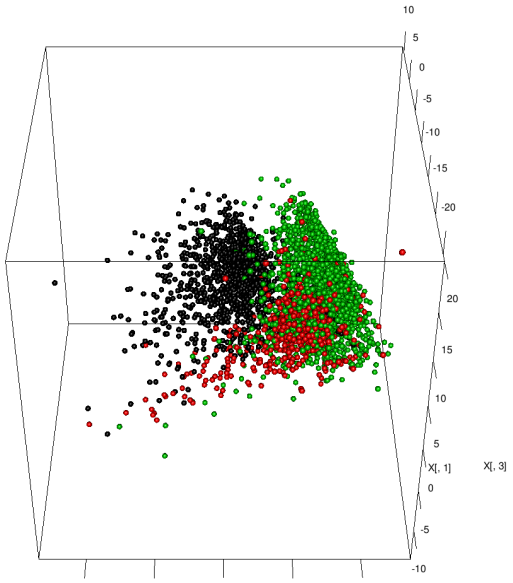


Figure 2: The feature represented training data projected onto the first 3 principal components. The somatic mutations (black) are reasonably well-separated from non-somatic mutations (germline - red, wildtype - green).

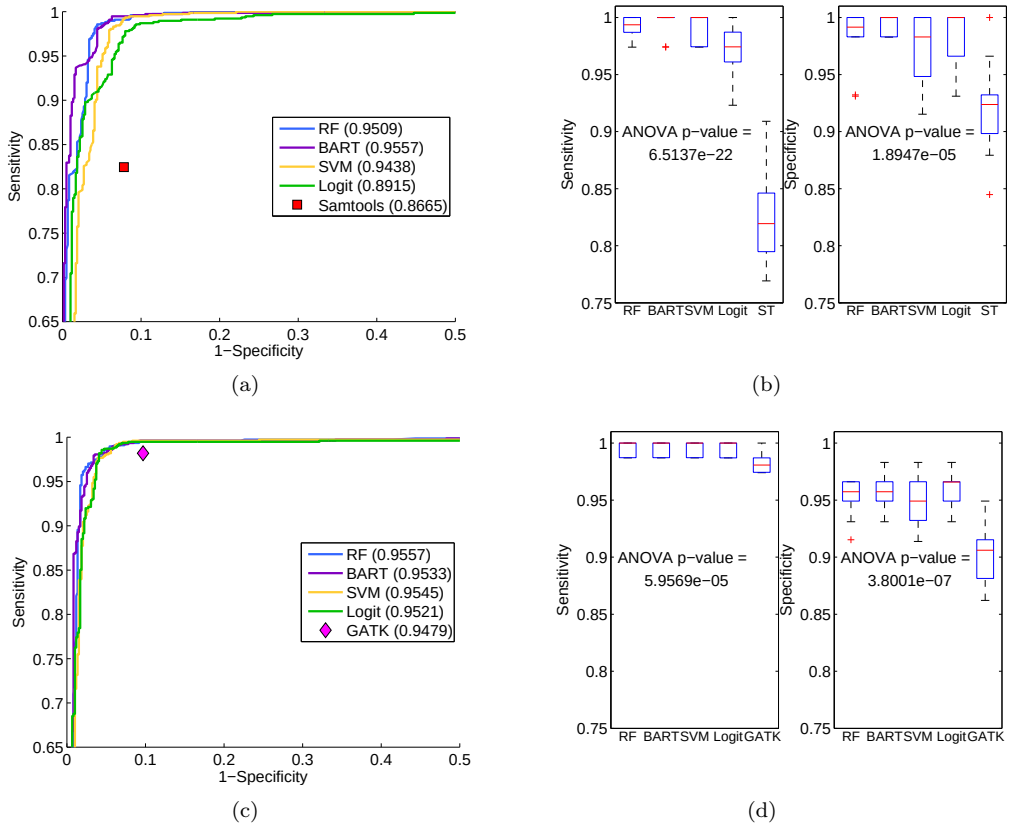


Figure 3: (a) The ROC curves of different classifiers by using only the Samtools' features as well as Samtools' results. (b) Comparison of classifiers with Samtools' results at the specificity and sensitivity level given by Samtools. (c) The ROC curves of different classifiers by using only GATK's features as well as GATK's results. (d) Comparison of classifiers with GATK's results at the specificity and sensitivity level given by GATK.

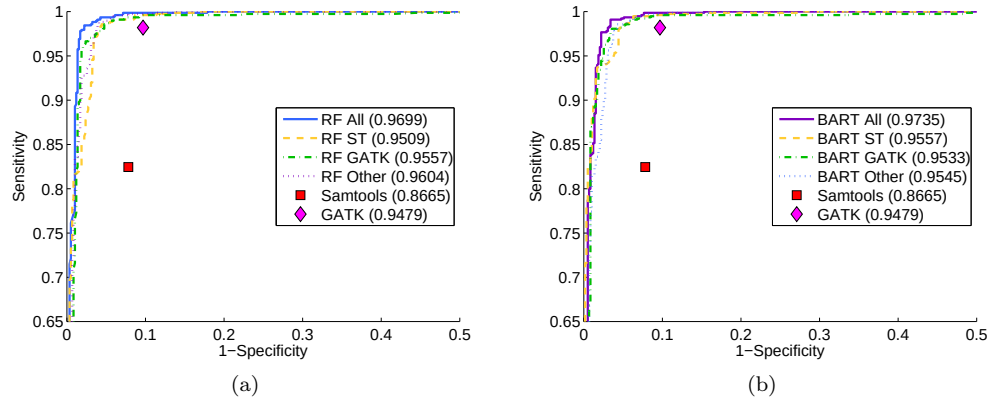
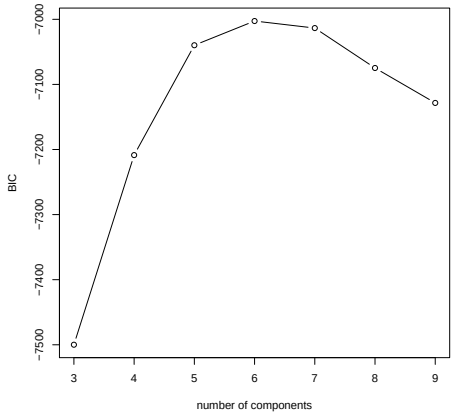
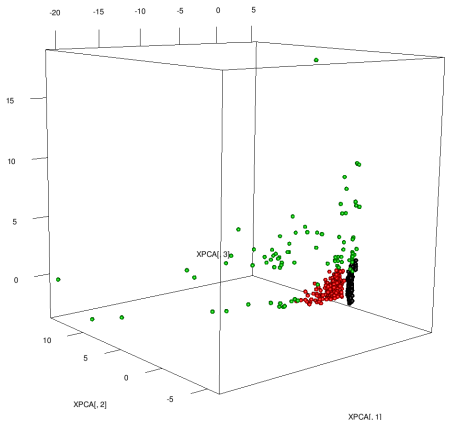


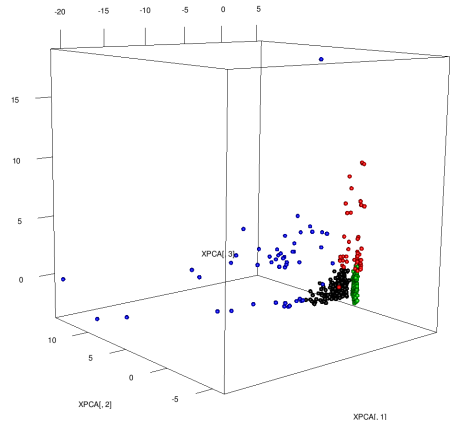
Figure 4: (a) The performance of RF on different feature sets. RF All: RF's accuracy by using all the features, RF ST: RF's accuracy by using only the Samtools features, RF GATK: RF's accuracy by using only the GATK features, RF Other: RF's accuracy by using all the new constructed 26 features. (2) The performance of BART on different feature sets. BART All, BART ST, BART GATK and BART Other are similarly defined.



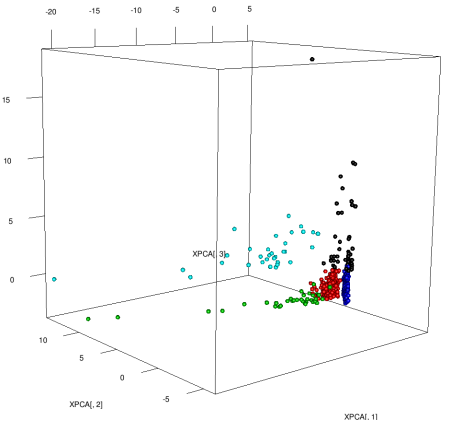
(a) The best BIC scores for different number of Gaussian components K



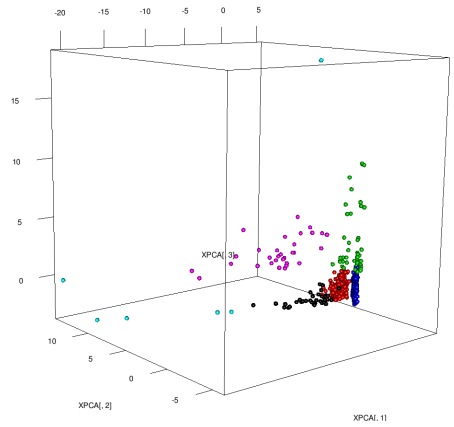
(b) K=3



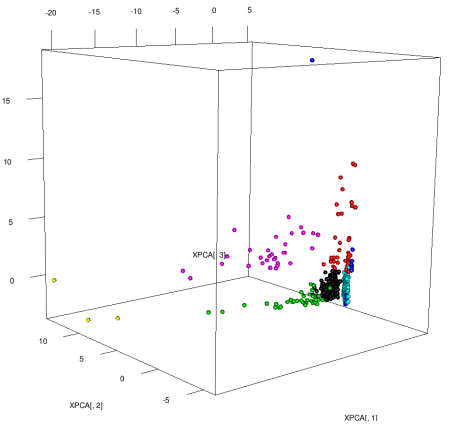
(c) K=4



(d) K=5



(e) K=6



(f) K=7

Figure 5: Using the BIC score to select the number of Gaussian components for model based clustering.

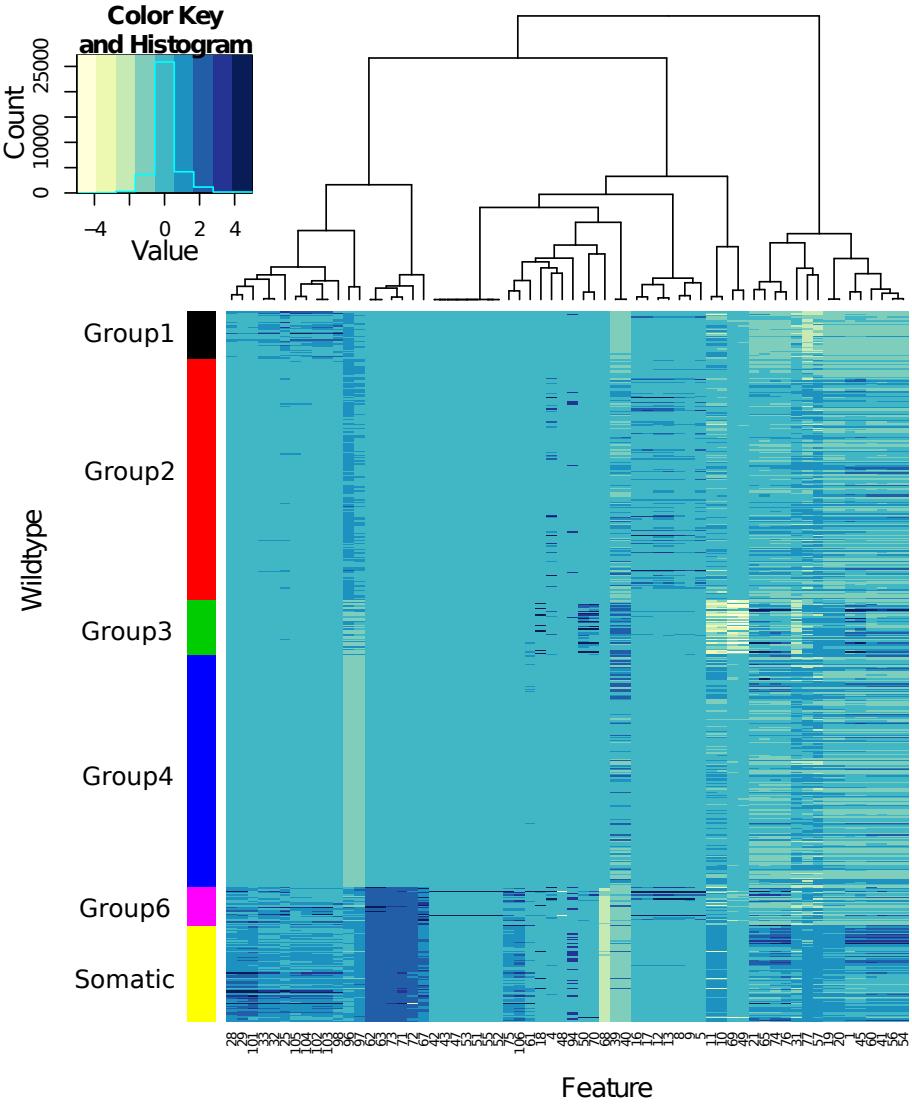
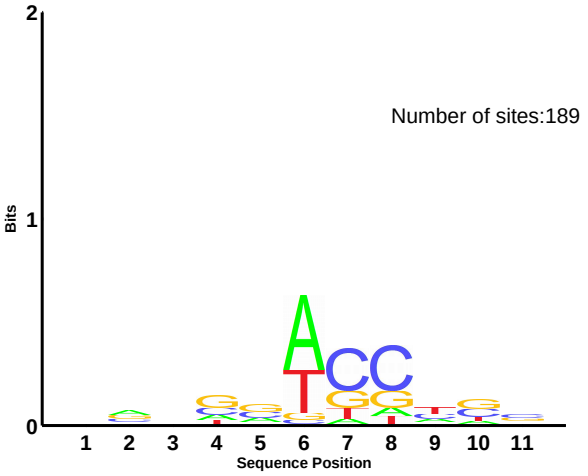
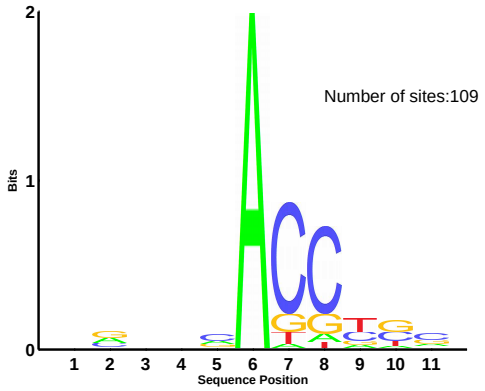


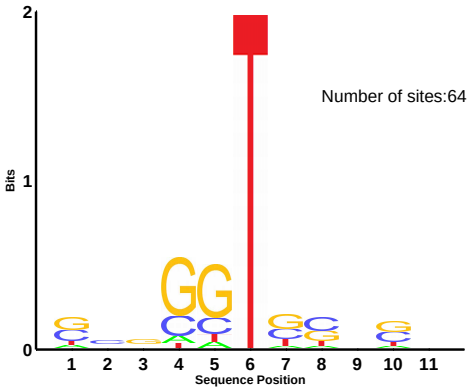
Figure 6: The heatmap obtained with $K = 6$. The 6 events in Group5 were not shown here.



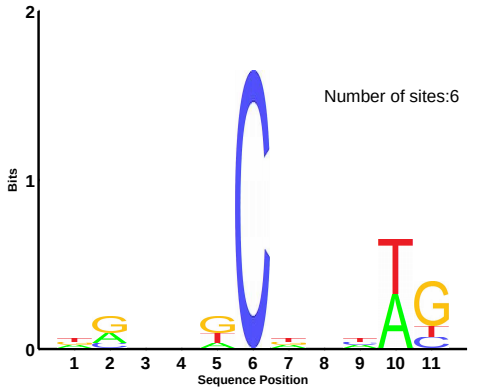
(a)



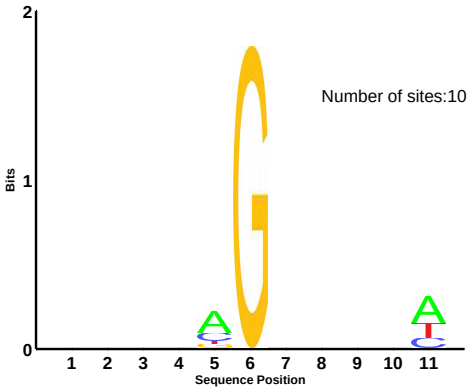
(b)



(c)



(d)



(e)

Figure 7: The sequence motifs at error sites of the group two wildtypes. Here the six base is the position where error occurs. (a) The logo for all the wildtypes in group2, (b) the wildtypes which the errors occur at base 'A', (c) the wildtypes which the errors occur at base base 'T', (d) the wildtypes which the errors occur at base 'C' and (e) the wildtypes which the errors occur at base 'G'.

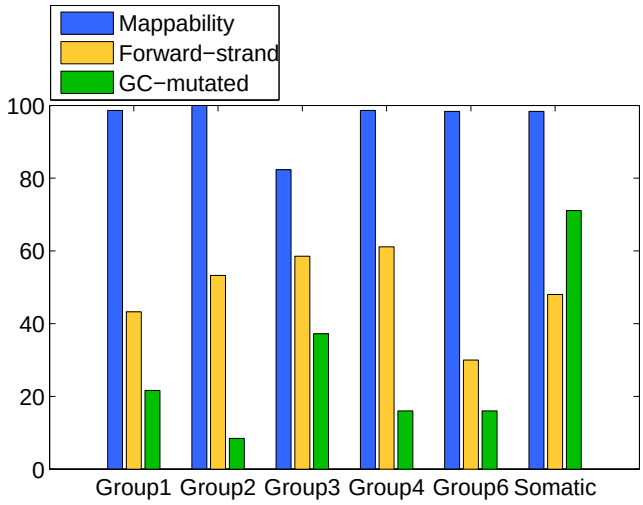
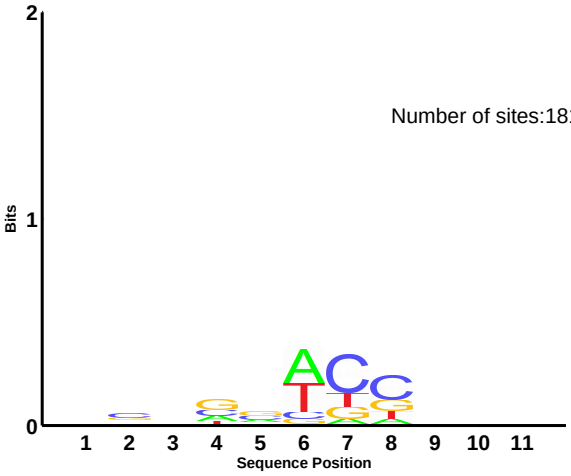
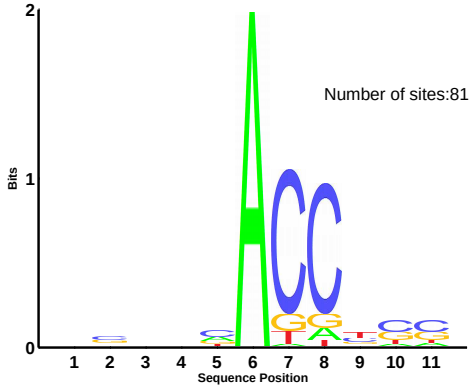


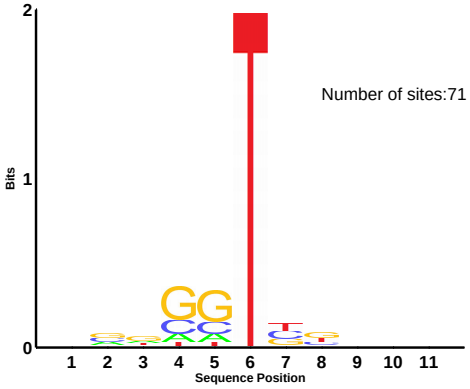
Figure 8: The mean of the mapabilities of each site and its 50 neighbour bases in both sides (denoted by mappabilities), the fraction of wildtypes resides in a gene whose coding strand is the forward strand (denoted by forward strand) and the fraction of G-C bases were the error occur for wildtypes or the G-C bases were mutated for somatic mutations (both denoted by GC mutated).



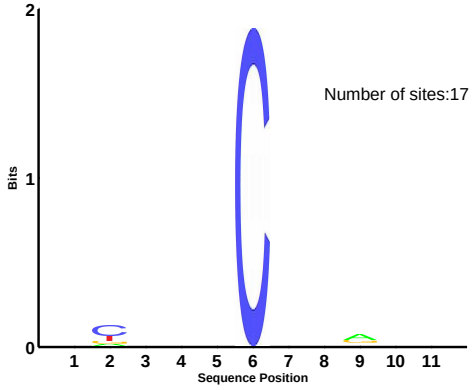
(a)



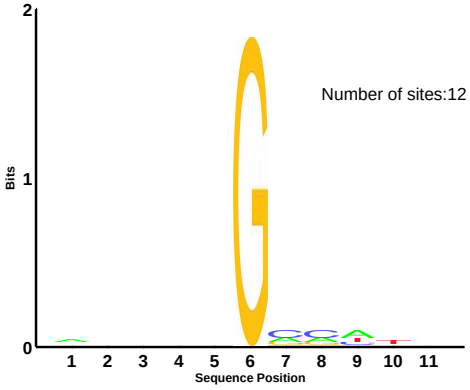
(b)



(c)



(d)



(e)

Figure 9: The sequence motifs at error sites of the group four wildtypes. Here the six base is the position where error occurs. (a) The logo for all the wildtypes, (b) the wildtypes which the errors occur at base 'A', (c) the wildtypes which the errors occur at base base 'T', (d) the wildtypes which the errors occur at base 'C' and (e) the wildtypes which the errors occur at base 'G'.

Supplementary tables

Table 1: The cross-validation results of classifiers and Samtools and GATK's prediction results on exome capture data. Here AUC SE means the standard error of the AUC from the 10 cross-validations. The AUCs are averaged by threshold average.

Model	Sensitivity	Specificity	Accuracy	AUC	AUC SE
RF	0.9901	0.9422	0.9567	0.9968	0.0007
BART	0.9901	0.9584	0.9679	0.9962	0.0013
SVM	0.9901	0.9405	0.9555	0.9954	0.0010
Logit	0.9901	0.8704	0.9065	0.9923	0.0011
Samtools	0.8631	0.7876	0.8103	N/A	N/A
GATK	0.9842	0.6563	0.7551	N/A	N/A

Table 2: The cross-validation results of classifiers and Samtools and GATK's prediction results on SeqVal1 before local realignment around indels and base quality recalibration.

Model	Sensitivity	Specificity	Accuracy	AUC	AUC SE
RF	0.9910	0.9609	0.9699	0.9953	0.0015
BART	0.9910	0.9660	0.9735	0.9939	0.0018
SVM	0.9910	0.9507	0.9628	0.9928	0.0015
Logit	0.9910	0.9524	0.9640	0.9912	0.0022
Samtools	0.8245	0.9218	0.8665	N/A	N/A
GATK	0.9819	0.9031	0.9479	N/A	N/A

Table 3: For SeqVal1, after local-realignment around indels and base quality recalibreation, the cross-validation results of classifiers and Samtools and GATK's prediction results.

Model	Sensitivity	Specificity	Accuracy	AUC	AUC SE
RF	0.9910	0.9660	0.9735	0.9947	0.0016
BART	0.9910	0.9609	0.9699	0.9937	0.0015
SVM	0.9910	0.9507	0.9628	0.9921	0.0015
Logit	0.9910	0.9507	0.9628	0.9914	0.0020
Samtools	0.7123	0.9269	0.8048	N/A	N/A
GATK	0.9587	0.9116	0.9384	N/A	N/A

Table 4: The cross-validation results of classifiers and Samtools and GATK's prediction results on SeqVal2.

Model	Sensitivity	Specificity	Accuracy	AUC	AUC SE
RF	0.9926	0.9004	0.9282	0.9935	0.0017
BART	0.9926	0.9048	0.9312	0.9928	0.0025
SVM	0.9926	0.8830	0.9160	0.9904	0.0026
Logit	0.9926	0.8139	0.8677	0.9838	0.0020
Samtools	0.9851	0.7329	0.7651	N/A	N/A
GATK	0.9926	0.5664	0.6208	N/A	N/A

Table 5: Comparison of the performance of classifiers with that of Samtools and GATK by fixing the sensitivity and specificity as given by Samtools and GATK. The one way ANOVA test was used to test the significance of the difference. Only for SeqVal2, by fixing the specificity given by Samtools and GATK, the sensitivities of classifiers are not statistically better than Samtools and GATK's sensitivities. For all the other cases, the classifiers did statistically better than Samtools and GATK's prediction results.

Data	Specificity (fix sensitivity given by Samtools)	Sensitivity (fix specificity given by Samtools)	Specificity (fix sensitivity given by GATK)	Sensitivity (fix specifity given by GATK)
SeqVal1+2	2.4803e-42	4.8286e-23	4.3981e-42	8.1410e-05
SeqVal1	5.2109e-09	4.1884e-26	5.3159e-09	9.8084e-09
SeqVal1 realign	1.4307e-10	2.5353e-43	8.3717e-11	1.5456e-11
SeqVal2	1.8212e-20	0.4638	5.0394e-28	0.6394

Table 6: The results of the classifiers trained on the exome capture data and test on the whole genome shotgun data as well as Samtools and GATK's results. The thresholds used here by different classifiers were the same as for those used in doing cross-validations.

Model	Sensitivity	Specificity	Accuracy	AUC
RF	0.8850	0.9518	0.9369	0.9732
BART	0.7876	0.9949	0.9487	0.9627
SVM	0.7876	0.9949	0.9487	0.9510
Logit	0.8496	0.9391	0.9191	0.9501
Samtools	0.7611	0.9467	0.9053	N/A
GATK	0.8230	0.8883	0.8738	N/A

Table 7: Comparison of the performance of classifiers with that of Samtools and GATK on the test whole genome shotgun data by fixing the sensitivity and specificity as given by Samtools and GATK.

Model	Specificity (fix sensitivity at 0.7611)	Sensitivity (fix specificity at 0.9467)	Specificity (fix sensitivity at 0.8230)	Sensitivity (fix specificity at 0.8883)
RF	1.0000	0.8938	0.9772	0.9381
BART	1.0000	0.8761	0.9747	0.9027
SVM	1.0000	0.8761	0.9772	0.9027
Logit	0.9797	0.8319	0.9721	0.8938

Table 8: Results of the RF model on the whole genome shotgun data using the feature sets selected by different classifiers. Here RF_F means the feature selected by RF classifier. BART_F, SVM_F and Logit_F are similarly defined.

Model	Sensitivity	Specificity	Accuracy	AUC
RF_F	0.8407	0.9645	0.9369	0.9638
BART_F	0.8850	0.9670	0.9487	0.9787
SVM_F	0.8938	0.9594	0.9448	0.9664
Logit_F	0.8496	0.9569	0.9329	0.9663

Table 9: Results of the BART model on the whole genome shotgun data using the feature sets selected by different classifiers.

Model	Sensitivity	Specificity	Accuracy	AUC
RF_F	0.8142	0.9721	0.9369	0.9467
BART_F	0.7965	0.9848	0.9428	0.9593
SVM_F	0.7788	0.9822	0.9369	0.9507
Logit_F	0.7876	0.9721	0.9310	0.9621

Table 10: Results of the SVM model on the whole genome shotgun data using the feature sets selected by different classifiers.

Model	Sensitivity	Specificity	Accuracy	AUC
RF_F	0.8496	0.9188	0.9034	0.9328
BART_F	0.7522	0.9949	0.9408	0.9280
SVM_F	0.7699	0.9848	0.9369	0.9453
Logit_F	0.7788	0.9873	0.9408	0.9528

Table 11: Results of the Logit model on the whole genome shotgun data using the feature sets selected by different classifiers.

Model	Sensitivity	Specificity	Accuracy	AUC
RF_F	0.9292	0.8731	0.8856	0.9526
BART_F	0.8761	0.9695	0.9487	0.9622
SVM_F	0.7876	0.9645	0.9250	0.9296
Logit_F	0.8496	0.9543	0.9310	0.9536

Table 12: The meanings of the selected features. As expected, the likelihoods provided by both Samtools and GATK, the base qualities, mapping qualities, strand bias, tail distance features are relevant. The features selected from the normal and tumour are different.

Index	Feature definition	Tumour	Samtools	GATK
6	sum of reference base qualities		√	
10	sum of reference mapping qualities		√	
19	$\max_{G_i \neq aa}(P(D \mid G_i))$		√	
26	sum of reference base qualities	√	√	
28	sum of non-reference base qualities	√	√	
37	sum of squares of tail distance for non-reference bases	√	√	
38	$P(D \mid G_i = aa)$	√	√	
41	QUAL: phred-scaled probability of the call given data			√
53	sumGLbyD			√
57	$P(D \mid G_i = aa)$			√
60	$P(D \mid G_i = bb)$			√
63	AF: allele frequency for each non-ref allele	√		√
69	MQ: root mean square mapping quality	√		√
71	QD: variant confidence/unfiltered depth	√		√
73	sumGLbyD	√		√
77	GQ: genotype quality computed based on the genotype likelihood	√		√
83	the difference between the sum of the base qualities of the current site and the next site			
96	sum of the pooled estimation of strand bias on both strands max(forward, reverse)			
97	sum of the pooled estimation of strand bias on both strands $\sum$(forward, reverse)			
99	Reverse strand non-reference base ratio			
101	sum of squares of non-reference base quality ratio			
102	Sum of non-reference mapping quality ratio			
105	Sum of squares of non-reference tail distance ratio			

Table 13: The significant features

Group number	feature	Tumour	Samtools	GATK	p-value
Group1	28 : sum of non-reference base qualities	√	√		0.0000
	29 : sum of squares of non-reference base qualities	√	√		0.0000
	103 : Sum of squares of non-reference mapping quality ratio F_{33}/F_{13}				0.0000
	102 : Sum of non-reference mapping quality ratio F_{32}/F_{12}				0.0000
	101 : Sum of squares of non-reference base quality ratio F_{29}/F_9				0.0000
	104 : Sum of non-reference tail distance ratio F_{36}/F_{16}				0.0000
	105 : Sum of squares of non-reference tail distance ratio F_{37}/F_{17}				0.0000
	25 : number of non-reference Q13 bases on the reverse strand	√	√		0.0000
	98 : Forward strand non-reference base ratio F_{24}/F_4				0.0000
	77 : GQ: genotype quality computed based on the genotype likelihood	√		√	0.0000
	45 : total (unfiltered) depth over all samples			√	0.0000
	20 : $\sum_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled		√		0.0000
	19 : $\max_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled		√		0.0000
	57 : GQ: genotype quality computed based on the genotype likelihood			√	0.0000
	65 : total (unfiltered) depth over all samples	√		√	0.0000
	1 : number of reads covering or bridging the site		√		0.0000
	54 : allelic depths for the ref-allele			√	0.0000
	56 : DP: read depth (only filtered reads used for calling)			√	0.0000
	60 : $P(D \mid G_i = bb)$, phred-scaled			√	0.0000
	21 : number of reads covering or bridging the site	√	√		0.0000
	76 : DP: read depth (only filtered reads used for calling)	√		√	0.0000
	41 : QUAL: phred-scaled probability of the call given data			√	0.0000
	74 : allelic depths for the ref-allele	√		√	0.0000
Group2	33 : sum of squares of non-reference mapping qualities	√	√		0.0000
	32 : sum of non-reference mapping qualities	√	√		0.0000
	40 : $\sum_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled	√	√		0.0000
	39 : $\max_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled	√	√		0.0000
Group3	96 : max of the pooled estimation of strand bias on both strands max(forward, reverse)				0.0000
	97 : sum of the pooled estimation of strand bias on both strands $\sum$(forward, reverse)				0.0000
	69 : MQ: root mean square mapping quality	√		√	0.0000
	49 : MQ: root mean square mapping quality			√	0.0000
	70 : MQ0: total number of reads with mapping quality zero	√		√	0.0000
	50 : MQ0: total number of reads with mapping quality zero			√	0.0000
	11 : sum of squares of reference mapping qualities		√		0.0000
	10 : sum of reference mapping qualities		√		0.0000
	31 : sum of squares of reference mapping qualities	√	√		0.0000
	40 : $\sum_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled	√	√		0.0000
	39 : $\max_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled	√	√		0.0000
	45 : DP: total (unfiltered) depth over all samples			√	0.0000
	18 : $P(D \mid G_i = aa)$, phred-scaled		√		0.0000
	20 : $\sum_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled		√		0.0000
	19 : $\max_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled		√		0.0000
	1 : The number of reads covering or bridging the site		√		0.0000
Group4	96 : max of the pooled estimation of strand bias on both strands max(forward, reverse)				0.0000
	97 : sum of the pooled estimation of strand bias on both strands $\sum$(forward, reverse)				0.0000
	40 : $\sum_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled	√	√		0.0000
	39 : $\max_{G_i \neq aa}(P(D \mid G_i))$, phred-scaled	√	√		0.0000
	71 : QD: variant confidence/unfiltered depth	√		√	0.0000
	72 : SB: strand bias (the variation being seen on only the forward or only the reverse strand)	√		√	0.0000
	73 : sumGLbyD	√		√	0.0000
	68 : HaplotypeScore: estimate the probability that the reads at this locus are coming from no more than 2 local haplotypes	√		√	0.0000
	63 : AF: allele frequency for each non-ref allele	√		√	0.0000
	62 : AC: allele count for non-ref allele in genotypes	√		√	0.0000
	67 : HRun: largest contiguous homopolymer run of variant allele in either direction	√		√	0.0000

28 : sum of non-reference base qualities	√	√		0.0000
29 : sum of squares of non-reference base qualities	√	√		0.0000
103 : Sum of squares of non-reference mapping quality ratio F_{33}/F_{13}				0.0000
102 : Sum of non-reference mapping quality ratio F_{32}/F_{12}				0.0000
106 : Non-reference allele depth ratio F_{75}/F_{55}				0.0000
101 : Sum of squares of non-reference base quality ratio F_{29}/F_{9}				0.0000
104 : Sum of non-reference tail distance ratio F_{36}/F_{16}				0.0000
105 : Sum of squares of non-reference tail distance ratio F_{37}/F_{17}				0.0000
25 : number of non-reference Q13 bases on the reverse strand	√	√		0.0000
75 : AD2: allelic depths for the non-ref allele	√		√	0.0000
98 : Forward strand non-reference base ratio F24/F4				0.0000
13 : sum of squares of non-reference mapping qualities		√		0.0000
12 : sum of non-reference mapping qualities		√		0.0000
8 : sum of non-reference base qualities		√		0.0000
9 : sum of squares of non-reference base qualities		√		0.0000
16 : sum of tail distances for non-reference bases		√		0.0000
17 : sum of squares of tail distance for non-reference bases		√		0.0000
5 : number of non-reference Q13 bases on the reverse strand		√		0.0000
4 : number of non-reference Q13 bases on the forward strand		√		0.0000
42 : allele count for non-ref allele in genotypes			√	0.0000
47 : HRun: largest contiguous homopolymer run of variant allele in either direction			√	0.0000
43 : AF: allele frequency for each non-ref allele			√	0.0000
53 : sumGLbyD			√	0.0000
51 : QD: variant confidence/unfiltered depth			√	0.0000
48 : HaplotypeScore: estimate the probability that the reads at this locus are coming from no more than 2 local haplotypes			√	0.0000
55 : allelic depths for the non-ref allele			√	0.0000
52 : SB: strand bias (the variation being seen on only the forward or only the reverse strand)			√	0.0000
94 : The alternative base is T				0.0000
61 : QUAL: phred-scaled probability of the call given data	√		√	0.0001