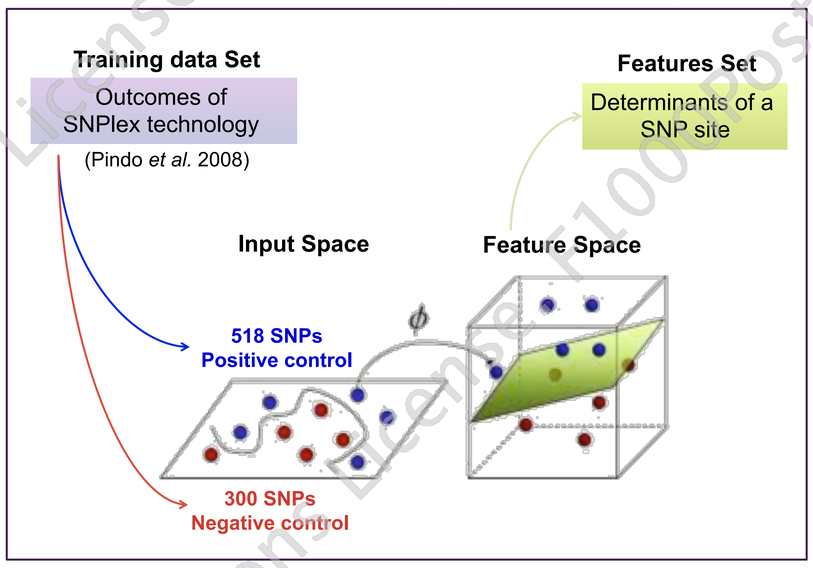
VerySNP: VCF features to train SVM in crop SNP detection

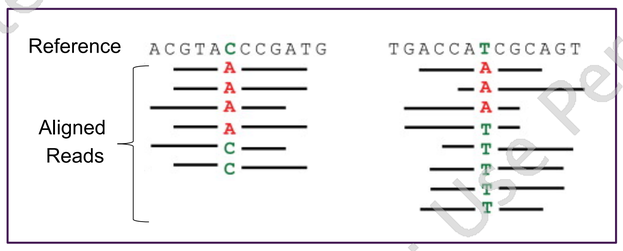
Lorena Leonardelli[1], Alessandro Cestaro[1], Carmen M Livi[2], Patrice This[3], Charles Romieu[3], Enrico Blanzieri[2], Claudio Moser[1]

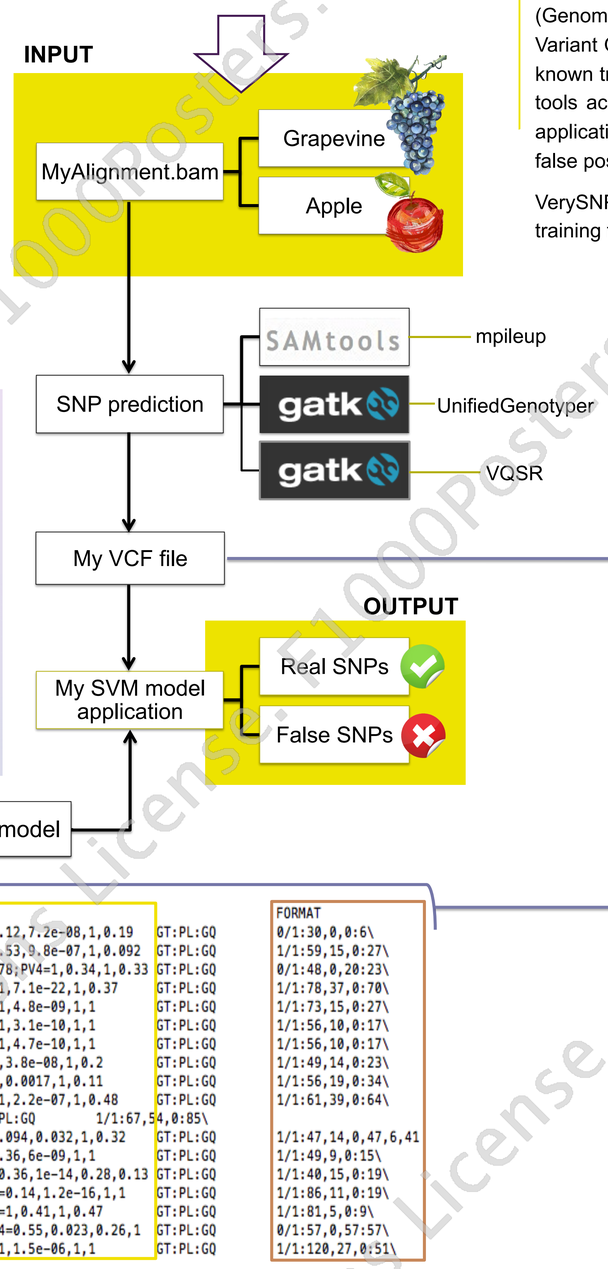
[1] Research and Innovation Centre, Fondazione Edmund Mach, Via Mach 1, 38010 San Michele all'Adige, Italy. *lorena.leonardelli@fmach.it*
[2] DISI, University of Trento, Via Sommarive 18, I-38050 Povo-Trento, Italy.
[3] Montpellier SupAgro - INRA, Unité Mixte de Recherche Amélioration et Génétique de l'Adaptation des Plantes, Montpellier, France

Support Vector Machine (SVM) training



MyAlignment.bam





#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
VV78X000014.13	1707	.	G	C	3.98	.	DP=18;VDB=0.0399;AF1=1;AC1=2;DP4=4,4,1,9;MQ=10;FQ=-27;PV4=0.12,7.2e-08,1,0.19	GT:PL:GQ 0/1:30,0,0:6\
VV78X000014.13	1734	.	A	G	26.3	.	DP=16;VDB=0.0374;AF1=1;AC1=2;DP4=0,4,3,9;MQ=10;FQ=-42;PV4=0.53,0.8e-07,1,0.092	GT:PL:GQ 1/1:59,15,0:27\
VV78X000014.13	1971	.	A	G	18.1	.	DP=22;VDB=0.0333;AF1=0.5025;AC1=1;DP4=8,2,10,2;MQ=13;FQ=-7.78;PV4=1,0.34,1,0.33	GT:PL:GQ 0/1:48,0,20:23\
VV78X000014.13	2000	.	A	C	45	.	DP=26;VDB=0.0388;AF1=1;AC1=2;DP4=4,0,19,3;MQ=13;FQ=-64;PV4=1,7.1e-22,1,0.37	GT:PL:GQ 1/1:78,37,0:70\
VV78X000014.13	2046	.	C	G	40.3	.	DP=19;VDB=0.0404;AF1=1;AC1=2;DP4=4,0,11,1;MQ=15;FQ=-42;PV4=1,4.8e-09,1,1	GT:PL:GQ 1/1:73,15,0:27\
VV78X000014.13	2050	.	T	A	23.8	.	DP=19;VDB=0.0404;AF1=1;AC1=2;DP4=5,0,10,1;MQ=15;FQ=-37;PV4=1,3.1e-10,1,1	GT:PL:GQ 1/1:56,10,0:17\
VV78X000014.13	2051	.	G	C	23.8	.	DP=19;VDB=0.0404;AF1=1;AC1=2;DP4=5,0,10,1;MQ=15;FQ=-37;PV4=1,4.7e-10,1,1	GT:PL:GQ 1/1:56,10,0:17\
VV78X000014.13	2066	.	T	A	16.4	.	DP=17;VDB=0.0401;AF1=1;AC1=2;DP4=1,0,5,2;MQ=14;FQ=-41;PV4=1,3.8e-08,1,0.2	GT:PL:GQ 1/1:49,14,0:23\
VV78X000014.13	2071	.	T	C	23.1	.	DP=20;VDB=0.0320;AF1=1;AC1=2;DP4=1,0,7,3;MQ=14;FQ=-46;PV4=1,0.0017,1,0.11	GT:PL:GQ 1/1:56,19,0:34\
VV78X000014.13	2138	.	T	C	28	.	DP=19;VDB=0.0147;AF1=1;AC1=2;DP4=1,0,10,7;MQ=10;FQ=-66;PV4=1,2.2e-07,1,0.48	GT:PL:GQ 1/1:61,39,0:64\
VV78X000014.13	2139	.	A	G	34	.	DP=18;VDB=0.0103;AF1=1;AC1=2;DP4=0,0,11,7;MQ=10;FQ=-81	GT:PL:GQ 1/1:67,54,0:85\
VV78X000014.13	2164	.	T	G,A	14.5	.	DP=21;VDB=0.0399;AF1=1;AC1=2;DP4=4,0,7,9;MQ=10;FQ=-41;PV4=0.094,0.032,1,0.32	GT:PL:GQ 1/1:47,14,0,47,6,41
VV78X000014.13	2383	.	C	G	17.1	.	DP=29;VDB=0.0399;AF1=1;AC1=2;DP4=3,4,9,4;MQ=11;FQ=-36;PV4=0.36,6e-09,1,1	GT:PL:GQ 1/1:49,9,0:15\
VV78X000014.13	2393	.	A	G	8.01	.	DP=31;VDB=0.0384;AF1=1;AC1=2;DP4=6,1,12,8;MQ=10;FQ=-42;PV4=0.36,1e-14,0.28,0.13	GT:PL:GQ 1/1:40,15,0:19\
VV78X000014.13	4203	.	C	A	53.6	.	DP=43;VDB=0.0404;AF1=1;AC1=2;DP4=0,13,4,17;MQ=14;FQ=-38;PV4=0.14,1.2e-16,1,1	GT:PL:GQ 1/1:86,11,0:19\
VV78X000014.13	4237	.	A	T	50.1	.	DP=39;VDB=0.0374;AF1=1;AC1=2;DP4=2,21,1,13;MQ=13;FQ=-32;PV4=1,0.41,1,0.47	GT:PL:GQ 1/1:81,5,0:9\
VV78X000014.13	4245	.	T	C	27	.	DP=42;VDB=0.0225;AF1=0.5;AC1=1;DP4=2,14,1,23;MQ=15;FQ=27;PV4=0.55,0.023,0.26,1	GT:PL:GQ 0/1:57,0,57:57\
VV78X000014.13	4682	.	G	A	87	.	DP=20;VDB=0.0172;AF1=1;AC1=2;DP4=0,3,3,13;MQ=23;FQ=-54;PV4=1,1.5e-06,1,1	GT:PL:GQ 1/1:120,27,0:51\

Motivation and Method

Although next generation sequencing (NGS) technologies are increasing genomic information at unprecedented pace, still the application of NGS data to *in silico* SNP identification is problematic. Major problems come from the inaccurate mapping of data on reference genome (either due to the short average read length or poor base quality) and from distortion respect to the sample population, due to biases from the chosen sequence technology or from reverse DNA transcription and PCR processes required to generate cDNA libraries.

Efficient approaches are thus needed to distinguish real polymorphisms from the abundant sequencing artefacts. Several SNP/variant predictors have been developed since the NGS advent and the most popular to process large-scale datasets are the mpileup function in SAMtools package (Li *et al.*, 2009) and UnifiedGenotyper in GATK (Genome Analysis ToolKit; McKenna *et al.*, 2010), which are binomial-based methods. GATK includes also the Variant Quality Score Recalibration (VQSR; DePristo *et al.*, 2011), which is a machine-learning technique based on known true SNP sites and fits new potential SNPs into a multidimensional Gaussian distribution. Even though those tools accurately call true variants, they still show high false positive polymorphism prediction. Quite recently, the application of a Supported Vector Machine (SVM; Vapnik, 1998) approach has been proven to be efficient to reduce false positive SNP predictions (O'Fallon *et al.*, 2013).

VerySNP is a tool that applies an SVM approach to VCF files highlighting the use of VCF information as SVM training features (Table 1) to detect true SNPs in crop genomes.

Table 1. VCF features description

N.	Vcf name	Feature description	GATK	SAMtools
1	QUAL	SNP call quality	Yes	Yes
2	AC	Allele count in genotypes, for each ALT allele, in the same order as listed	Yes	Yes
3	AF	Allele Frequency, for each ALT allele, in the same order as listed	Yes	Yes
4	GQ	Genotype Quality	Yes	Yes
5	PL	Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification	Yes	Yes
6	MQ	RMS Mapping Quality	Yes	Yes
7	GT	Genotype	Yes	Yes
8	DP	ApproYesimate read depth (reads with MQ=255 or with bad mates are filtered)	Yes	Yes
9	FQ	Phred probability of all samples being the same	-	Yes
10	VDB	Variant Distance Bias	-	Yes
11	DP4	High-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases	-	Yes
12	PV4	P-values for strand bias, baseQ bias, mapQ bias and tail distance bias	-	Yes
13	AN	Total number of alleles in called genotypes	Yes	-
14	BaseQRankSum	Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities	Yes	-
15	DP	Approximate read depth; some reads may have been filtered	Yes	-
16	Dels	Fraction of Reads Containing Spanning Deletions	Yes	-
17	FS	Phred-scaled p-value using Fisher's exact test to detect strand bias	Yes	-
18	HaplotypeScore	Consistency of the site with at most two segregating haplotypes	Yes	-
19	MLEAC	Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed	Yes	-
20	MLEAF	Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT allele, in the same order as listed	Yes	-
21	MQ0	Total Mapping Quality Zero Reads	Yes	-
22	MQRankSum	Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities	Yes	-
23	QD	Variant Confidence/Quality by Depth	Yes	-
24	ReadPosRankSum	Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias	Yes	-
25	AD	Allelic depths for the ref and alt alleles in the order listed	Yes	-

Results and Discussion

VerySNP training takes as input the positive and negative SNP sets and builds a model, based on the VCF features, to separate the known data into two classes. The 10-fold-cross validation estimates the model performance in terms of accuracy, specificity, sensitivity and precision. Among the 10 proposed models the best performing one is applied to unknown data, for instance if a list of candidate SNPs in VCF file is the model input, then the output provided is the very same list of polymorphisms classified as true (+1) and false (-1) variants.

VerySNP was tested using positive and negative SNPs coming from Pinot Noir ENTAV 115 genome and validated by SNPlex (Pindo *et al.*, 2008) and a set of paired-end reads from Illumina sequencing technology. Once that Illumina reads were aligned we obtained a VCF files from 3 different strategies (mpileup by SAMtools, UnifiedGenotyper by GATK and VQSR by GATK); and then we applied our software to all the 3 sets of predictions improving the predictions of 96.6 % in precision and 92.3 % of accuracy (Table 2).

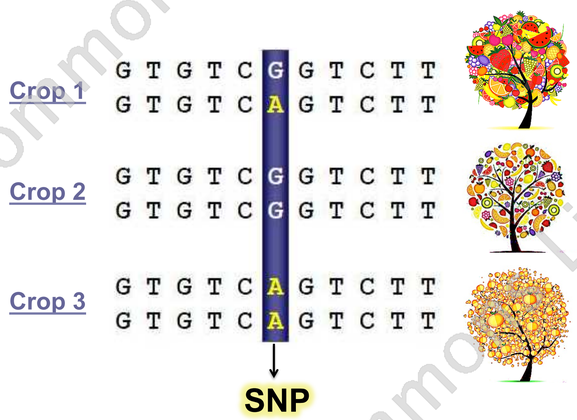


Table 2. SVM 10-fold cross validation

	Sensitivity	Specificity	Precision	Accuracy
GATK	94%	63%	97%	91%
VQSR	95%	67%	95%	91%
SAMtools	96%	65%	98%	95%
Average	95%	65%	96,6%	92,3%

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

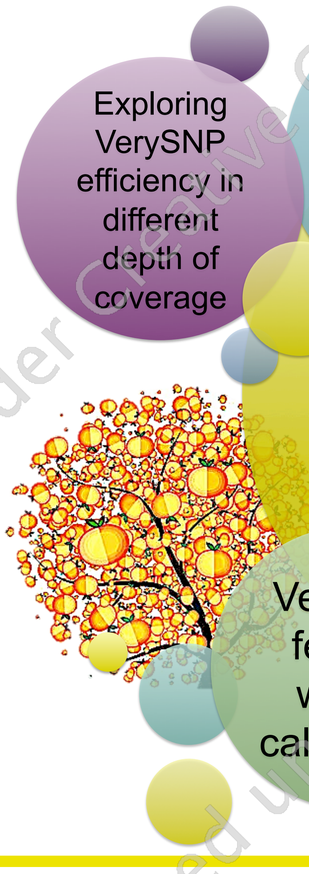
$$\text{Specificity} = \frac{TN}{TN + FP} \qquad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Conclusion

SNP prediction is a challenging operation especially in non-model organisms, which are really problematic due to the lack of large validated set and the high complexity of the genome sequence. For those reasons we provide a software that helps the researcher in handling the valuable information of VCF files by means of an SVM approach. Once the 10 fold cross-validation is completed the scientist will be able to choose the proper model for his dataset, visualizing the efficiency of each model with the training data sets or to use the best performing model for his dataset. Finally, the system is ready to be applied to unknown data the user needs to analyse.

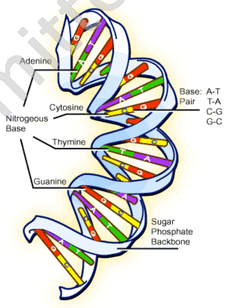
Future Work with FruitBreedomics data

VerySNP has been designed to identify mutation in vegetal genomes given the high complexity of those sequences. While plants have a totally different evolutionary history from the other two most studied kingdoms, animals and microorganisms, the interest in crops is exponentially rising as consequence of biological problems caused by climate changes and the increasing social demand of a more sustainable production. The enhancement of fruit quality and crops resistance to biotic stress are major purposes in the modern agricultural field, where the first solution would come from a more efficient fruit breeding, which is strategic but limited in tree breeding: long term, low efficiency and hence high cost. In this scenario the European project FruiBreedomics takes place suggesting to bridge the existing gap between scientific molecular genetics research and application in breeding and although FruiBreedomics initially focused on apple and peach as major fruits in Europe, results will be extended to other rosaceae fruit tree species. VerySNP has the potential to help the scientific molecular genetics research in finding the most likely true variants in Apple's genome, thanks to the availability of a big amount of data collected in the FruiBreedomics project, retrieved through 18K SNP-chip analysis. FruitBreedomics has already planned to produce other two chip, 18K and 300K, respectively.



Bibliography

- **DePristo**, M., *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498.
- **Li**, H., Handsaker, B. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9.
- **Mckenna**, A., *et al.* (2010). The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data, 1297–1303.
- **O'Fallon**, B.D., *et al.* (2013). A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics (Oxford, England)*, 1–6.
- **Vapnik**, V. N. Statistical Learning Theory. New York: Wiley, 1998, p. 736



Contacts: Lorena Leonardelli - lorena.leonardelli@fmach.it