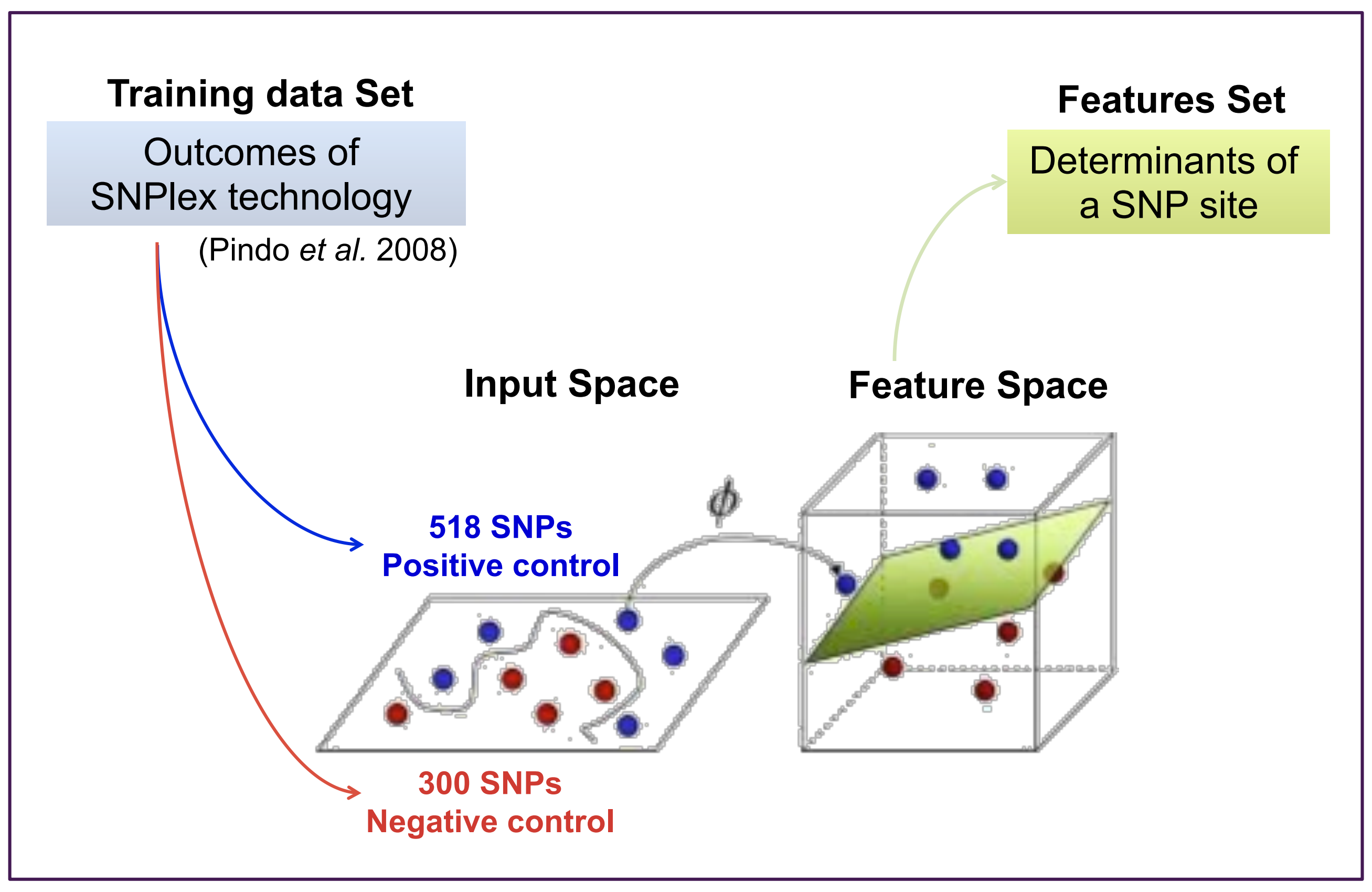
VCF features to train SVM in grapevine SNP detection

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

[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


Training data Set

Outcomes of SNPlex technology

(Pindo *et al.* 2008)

Features Set

Determinants of a SNP site

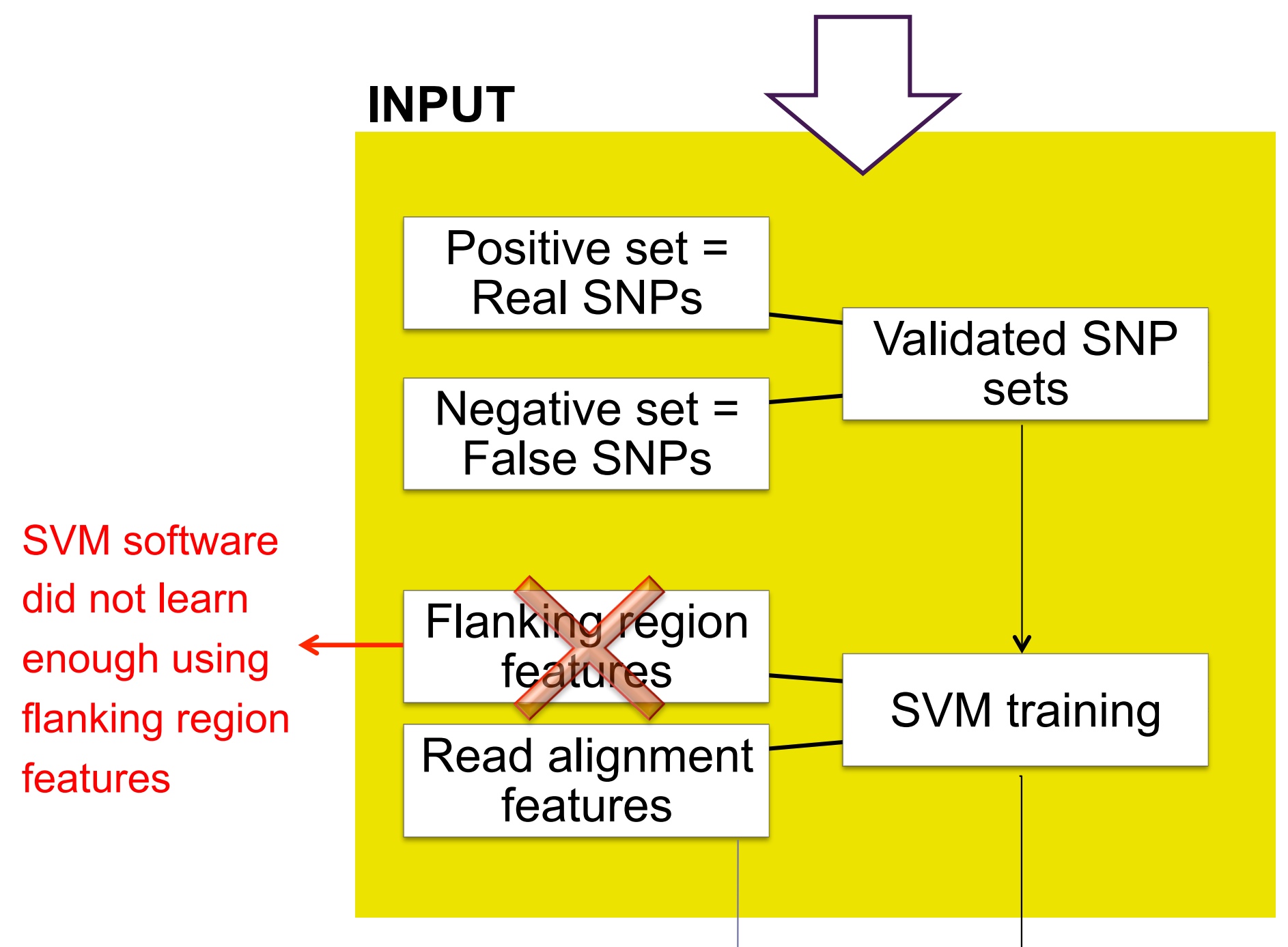
Input Space

Feature Space

ϕ

518 SNPs Positive control

300 SNPs Negative control


INPUT

Positive set = Real SNPs

Negative set = False SNPs

Validated SNP sets

Flanking region features

Read alignment features

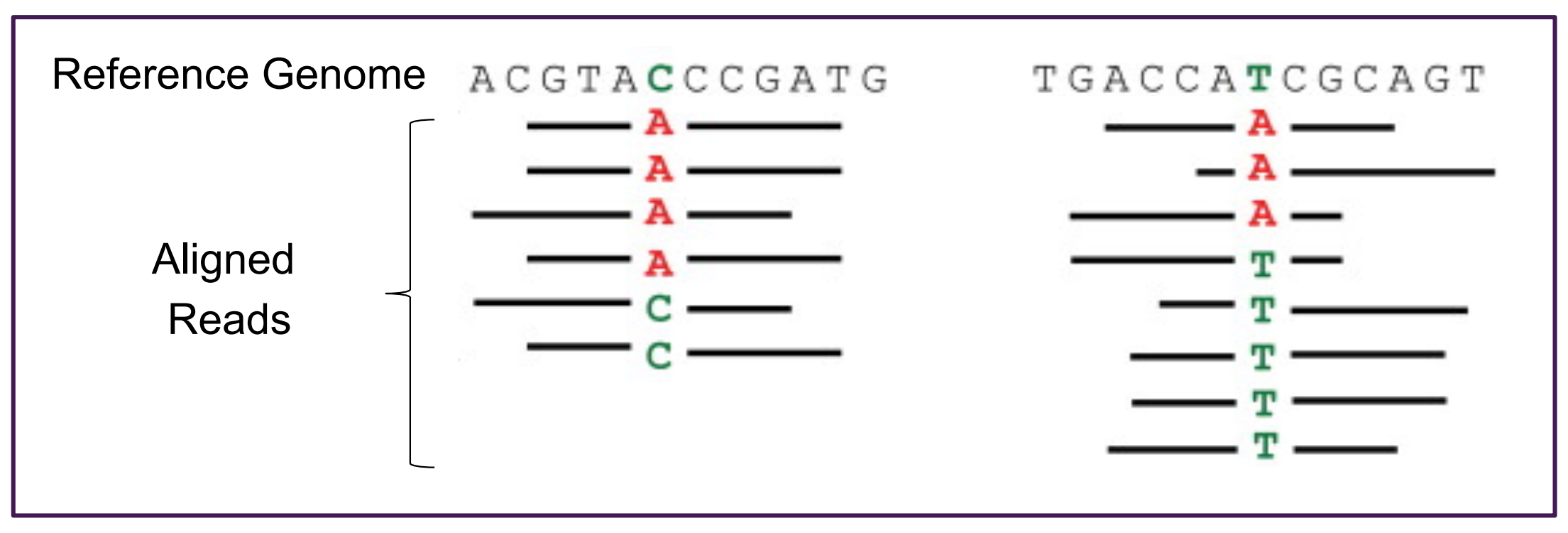
SVM training

SVM software did not learn enough using flanking region features

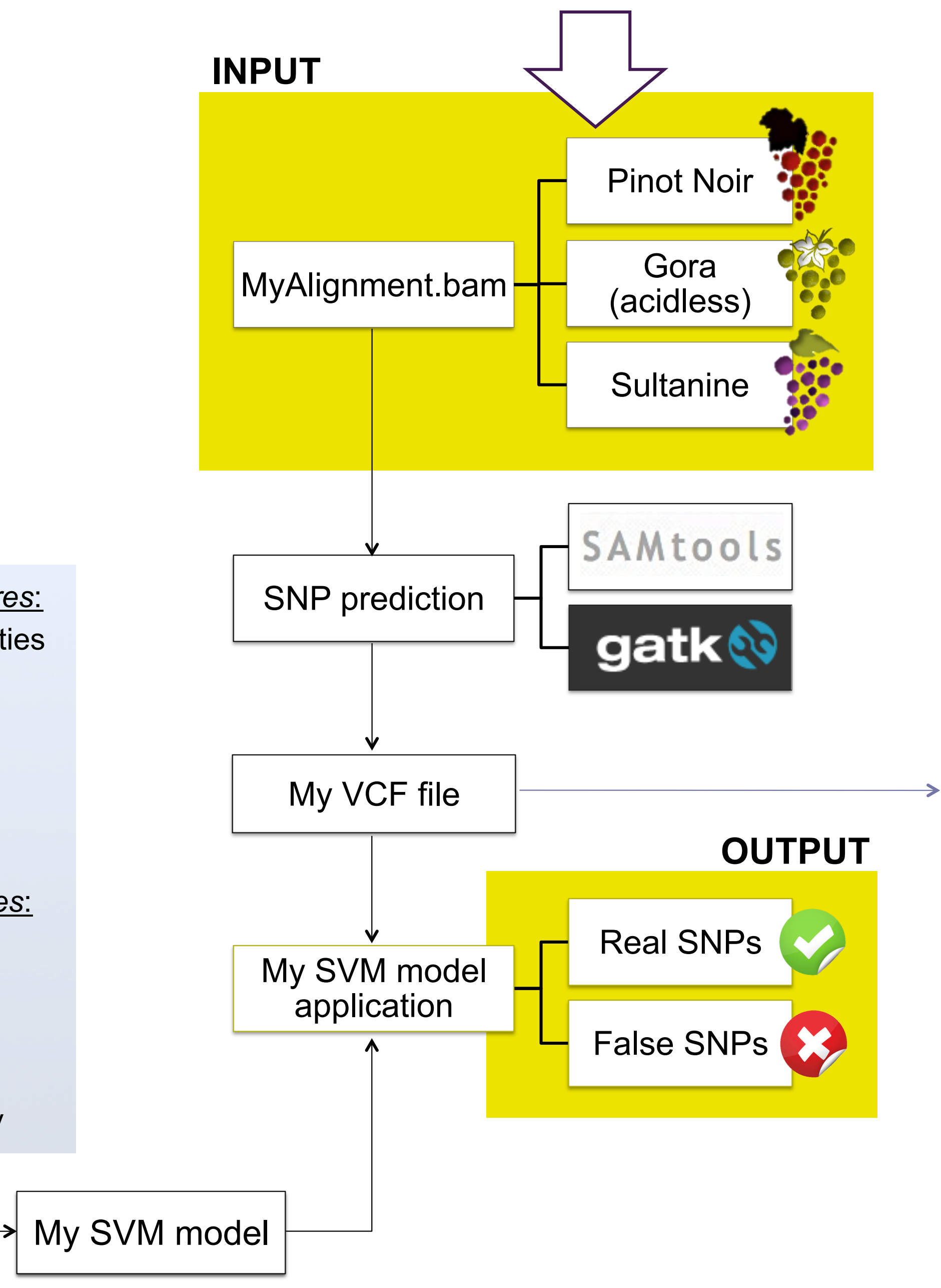
Flanking regions features:
Thermodinamic properties
- T_m=2(A+T)+4(C+G)
- CG content
- Enthalpy (ΔH)
- Entropy (ΔS)
- Free Energy (ΔG)

Read alignment features:
VCF parameters
- Alignment depth
- Allele Frequency
- Mapping Quality
- Nucleotide call Quality

MyAlignment.bam


Reference Genome

Aligned Reads


INPUT

MyAlignment.bam

Pinot Noir

Gora (acidless)

Sultanine

SNP prediction

SAMtools

gatk

My VCF file

OUTPUT

My SVM model application

Real SNPs

False SNPs

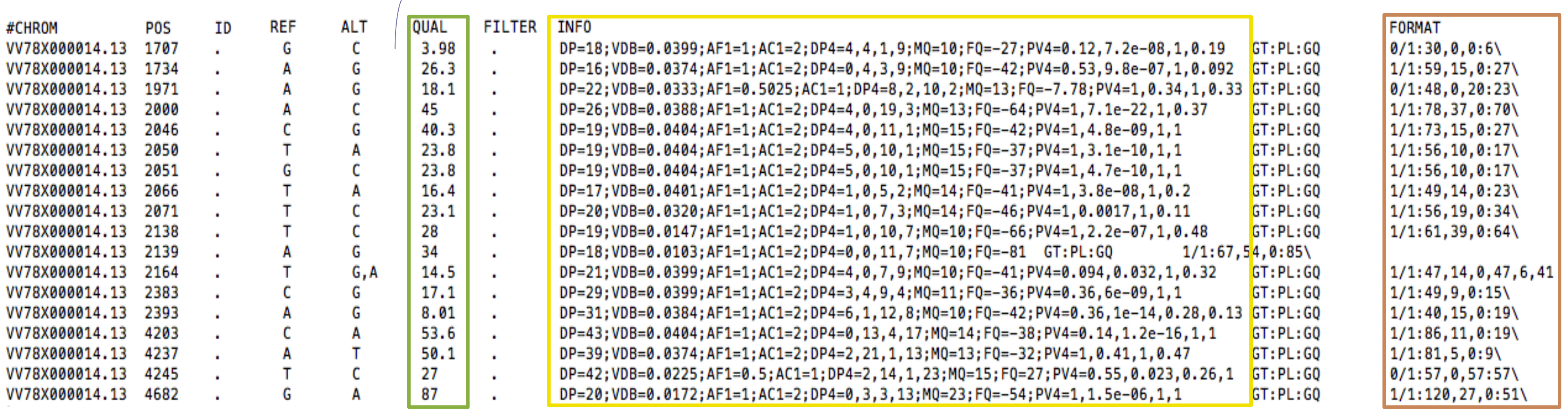
My SVM model

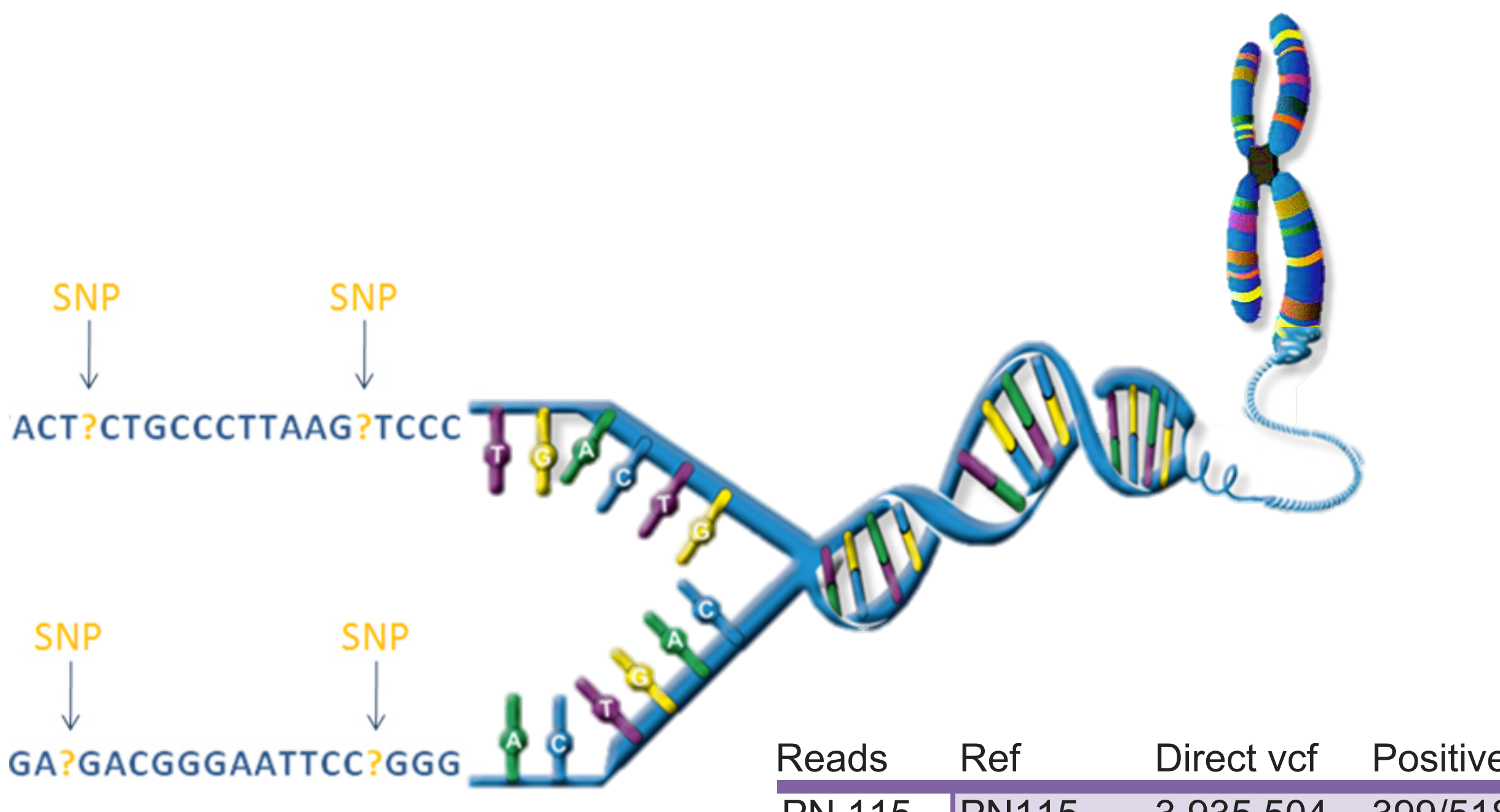
Motivation and Method

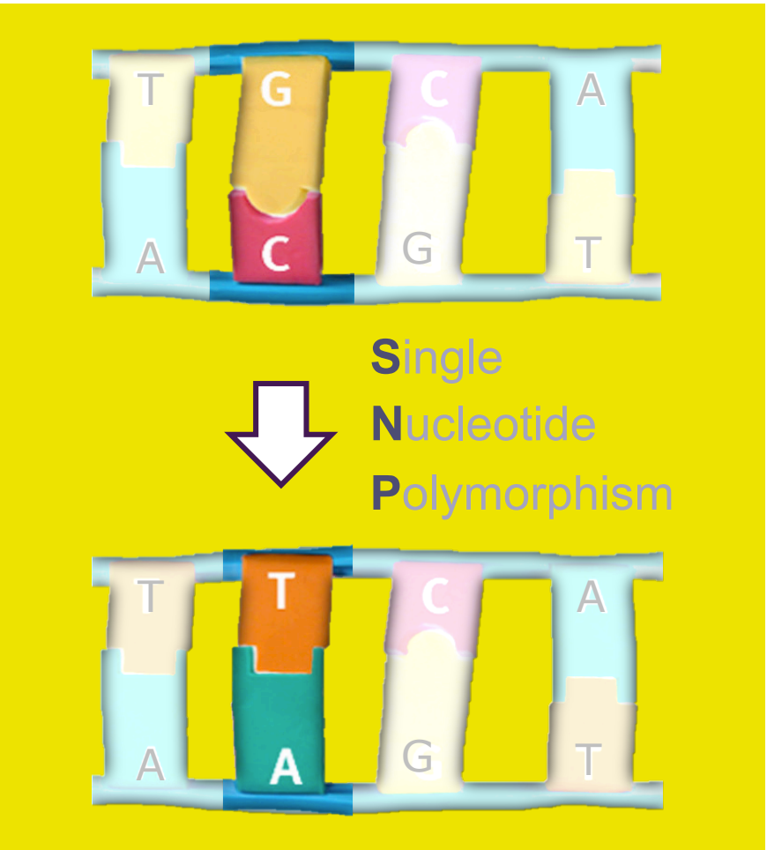
SAM/BAM (Sequence Alignment/Map / Binary Alignment/Map) are in our days the common data formats for aligned sequences since they have been adopted by the entire genomics community. Calling SNPs from SAM/BAM files with predictors like SAMtools and GATK (Genome Analysis Toolkit) provides a Variant Call Format (VCF) file as output (Danecek *et al.*, 2011). The VCF file contains a list of candidate SNPs with several informations such as relative position, the nucleotide present on the reference genome and on alternative alleles, SNP call quality, genotype and many other parameters. It is difficult to distinguish real polymorphisms from sequencing errors by simply looking at these values as such. However VCF parameters can be much more informative if used to train a Support Vector Machine (SVM) (Vapnik *et al.*, 1998) that classifies the list of candidate SNPs in real SNPs and false positive results. SVM is an efficient and reliable machine learning method to distinguish categorical data; it separates the positive and negative training data by constructing a linear classifier or a non-linear classifier with a kernel function. Based on training features, SVM represents the data as points in space, where the data belong to two categories (positive and negative) divided by a gap that is as wide as possible.

The training features were calculated on an experimentally validated set of SNPs (550 positive data set) and on monomorphic SNP positions (300 negative control data set). The SVM training was validated by the 10-fold cross validation method. The resulting model was applied on genomic data of three different grapevine cultivars aligned against both the available reference genomes: Pinot Noir ENTAV 115 (Velasco *et al.*, 2007) and Pinot Noir 40024 (Jaillon *et al.*, 2007).

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	-


```
#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=0.53,9.8e-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\
```





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

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

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

Table 1. GATK pipeline

	GATK functions	INPUT	OUTPUT
1	RealignerTargetCreator	myAlignment.bam	myRealignment.intervals
2	IndelRealigner	myRealignment.intervals	myRealignment.bam
3	UnifiedGenotyper	myRealignment.bam	realigned_snp.vcf
4	--- none ---	realigned_snp.vcf	known_snp.vcf
5	BaseRecalibrator	known_snp.vcf	recalibration_data.grp
6	PrintReads	recalibration_data.grp	myRecalibration.bam
7	UnifiedGenotyper	myRecalibration.bam	dbSNP.vcf
8	UnifiedGenotyper	myRecalibration.bam	recalibrated_snp.vcf

Table 2. GATK pipeline results

Reads	Ref	Direct vcf	Positive	Neg	Realigned vcf	Positive	Neg	Recalibrated vcf	Positive	Neg
PN 115	PN115	3.935.504	399/518	39/300	3.935.403	401/518	39/300	3.196.719	407/518	39/300
PN 115	PN40024	4.597.656	413/502	21/276	4.596.428	413/502	20/276	3.792.196	401/502	19/276
Gora	PN115	4.612.556	255/518	25/300	4.619.858	255/518	24/300	4.204.112	265/518	33/300
Sultanine	PN115	4.425.275	253/518	23/300	4.433.344	254/518	23/300	4.009.233	254/518	34/300

Table 3. GATK VS SAMtools

Grapevine Cultivar	Reference Genome	GATK Total	SAMtools Total	GATK Positive	SAMtools Positive	GATK Negative	SAMtools Negative
PN 115	PN 115	3.935.504	2.428.738	399/518	398/518	39/300	32/300
PN 115	PN 40024	4.597.656	2.916.135	413/502	392/502	21/276	17/276
Gora	PN 115	4.612.556	1.526.193	255/518	144/518	25/300	8/300
Sultanine	PN 115	4.425.275	1.824.658	253/518	155/518	23/300	9/300

Table 4. SVM 10-fold cross validation

	TP	TN	FP	FN	Sensitivity	Specificity	Precision	Accuracy
GATK	37,3	2,6	1,3	2,5	94%	63%	97%	91%
SAMtools	37,4	1,8	1,3	1,4	96%	65%	97%	94%

Results and Discussion

The GATK pipeline (Table 1) may involve the use of known SNPs as training for the following SNP prediction. We compared the amount of real/false SNPs detected by GATK with and without the use of known polymorphic sites as training in different grapevine alignments. By performing GATK pipeline three VCF files are produced and for each one we calculated the number of called SNPs and the ratio of positive/negative SNP (Table 2). Table 3 shows a comparative analysis between GATK, (with no SNPs training) and SAMtools, (no SNPs training is possible) in SNP detection when three grape cultivars are aligned against the same reference genome (PN ENTAV 115) and when Pinot Noir ENTAV 115 reads are aligned on Pinot Noir 40024. In general GATK predicts many more SNPs than SAMtools, but they give quite similar results when predicting the known SNPs (e.g. PN ENTAV 115 aligned on it-self: 398 SNPs predicted by SAMtools and 399 by GATK) or the known false ones (e.g GATK 39 and SAMtools 32). Although few, GATK and SAMtools predicted some known false polymorphisms.

As summarized in table 4, a linear SVM trained with VCF parameters as features has reached an average accuracy of 94% starting with SAMtools data and 91% with GATK data in the 10-fold cross validation on PN ENTAV 115 data aligned against the PN ENTAV 115 reference. The really high SVM performance suggests that the VCF parameters are sufficiently informative to discriminate whether polymorphic sites are real SNPs or sequencing errors or low quality nucleotide alignment. SVM can efficiently recognize true SNPs from false positive predictions as shown by high sensitivity (GATK 94%, SAMtools 96%), specificity (GATK 63%, SAMtools 65%), and precision (GATK 97%, SAMtools 97%).

Conclusion

The SVM model can be applied to recognize real SNPs in VCF file generated by SNP prediction. Although many SNP predicting tools are available depending on the data set specific properties (genomic or transcriptomic or gene specific), several of them can output a VCF file. When the sample has a high genomic distance from the reference sequence a new training with known positive and negative SNPs is likely required.

Work in progress

For now, the SVM approach have been applied to PN ENTAV 115 reads aligned on the PN ENTAV 115 *de novo* assembly with a 107X depth of coverage, which is a sort of optimal situation. We will perform the same SVM approach to the other grapevine alignments (Gora and Sultanine reads aligned on PN ENTAV 115 as reference genome and PN ENTAV 115 reads aligned on PN 40024 as reference genome). Since we do not have a set of positive/negative SNPs in Gora/Sultanine, we are going to predict SNPs through SVM methodology in Gora/Sultanine using the SVM model we got from the PN 115 training. Another possibility is to train the SVM with the positive/negative SNP subset predicted by GATK and SAMtools in PN 115 and in Gora/Sultanine as well, assuming to confirm the real nature of those SNPs with experimental techniques in the next future. Other validation inter/intra species have been planned for the next months, thanks to other *Vitis vinifera* and *Malus domestica* (apple) SNPs data we have recently retrieved.

Bibliography

- **Danecek**, P., Auton, A., Abecasis, G., Albers, C. a, Banks, E., DePristo, M. a, Handsaker, R. E., et al. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–8. doi:10.1093/bioinformatics/btr330
- **Jaillon**, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–7. doi:10.1038/nature06148
- **Vapnik**, V. N. Statistical Learning Theory. New York: Wiley, 1998, p. 736
- **Velasco**, R., Zharkikh, A., Troggio, M., Cartwright, D. a, Cestaro, A., Pruss, D., Pindo, M., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PloS one*, *2*(12), e1326. doi:10.1371/journal.pone.0001326