VCF features to train SVM in grapevine SNP detection

Leonardelli L[1], Cestaro A[1], Livi CM[2], Romieu C[3], This P[3], Moser C[1], Blanzieri E[2].

[1] Research and Innovation Centre, Fondazione Edmund Mach, S. Michele all'Adige, Trento, Italy
[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

Motivation

Although next generation sequencing (NGS) technologies are increasing genomic information at unprecedented pace, they are prone to an error rate even higher than one each 100 bp. Efficient approaches are thus needed to distinguish real polymorphisms from the abundant sequencing artefacts. Many open-source tools have been recently developed to identify Single Nucleotide Polymorphisms (SNPs) in whole-genome data, the most popular being Samtools (Li *et al.*, 2009) and GATK (DePristo *et al.*, 2011). Still they present an unsatisfactory accuracy due to high false positive polymorphism prediction. SNPs are the most abundant type of DNA sequence mutations and they are efficient markers for several biological applications such as cultivar identification, construction of genetic maps, the assessment of genetic diversity, the detection of genotype/phenotype associations, or marker-assisted breeding. The biological importance of finding only true SNPs is evident, considering the expensive cost of SNP validation through re-sequencing or SNP-chip, not only in terms of money but also of time and, above all, sample's availability. Since these small mutations can be responsible of large changes in the physiology or the evolution of an organism, our interest is to define if this category of polymorphisms is the genetic determinant of the low acidity content in the grapevine (*Vitis vinifera* L.) cultivar Gora Chirine. To this aim we are investigating the acidity trait in grapevine by comparative analysis of the genome sequences of Gora and Sultanine, is the latter being a normal acidity grapevine cultivar and genetically a close relative to Gora. Malic acid amount in grape berries is an essential parameter in wine fermentation quality and investigation of new genes involved in grapevine acidic metabolism is a common interest for biologists, enologists and wine makers.

Method

The advent of NGS technologies, such as Illumina/Solexa, AB/SOLiD and Roche/454 (Mardis, 2008) created new raw sequence types and several new tools for read alignment have been developed generating alignments in different formats. A standard alignment format supporting all sequence types and aligners allows an easier connection between read alignment and downstream analyses, including variant detection, genotyping and assembly. New data formats for aligned sequences are SAM/BAM (Sequence Alignment/Map and Binary Alignment/Map), now adopted by the entire genomics community. Calling SNPs from SAM/BAM files with predictors like SAMtools and GATK (Genome Analysis Toolkit) provides as output a Variant Call Format (VCF) file (Danecek *et*

al., 2011). VCF file contains a list of candidate SNPs with relative position on contigs, the nucleotide present on the reference genome and on the alternative alleles, SNP call quality, genotype and many other parameters. It is hard to consider all those values in order to distinguish which SNPs are actually polymorphisms or sequencing errors, but VCF parameters can be a lot more informative if used to train a Support Vector Machine (SVM) approach (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 SNP predictors, SAMtools and GATK, output approximately 400 of the 520 positive SNPs and 40 of the 300 negative SNPs, compelling us to re-balance the SNP sets with the SMOTE algorithm (Chawla *et al.*, 2011) before the SVM training. The SVM training was validated by the 10-fold cross validation method. The resulting model will be applied on the biological study mentioned above as well as on other data sets.

Results

SVM trained with 21 and 12 VCF parameters (Table 1) for GATK and SAMtools, respectively, as features has reached an average accuracy of 94% with SAMtools data and 91% with GATK data. The SVM performance suggests which VCF parameters are determinant to understand if polymorphic sites are real SNP sites or errors due to sequencing as well as to 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%) resulting from the SVM 10-fold cross validation.

Table 1: VCF Features listed within which predictors are shown.

N	Vcf name	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	-

Bibliography

Chawla, N. V, Bowyer, K. W., & Hall, L. O. (2002). SMOTE : Synthetic Minority Over-sampling Technique, *16*, 321–357.

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

DePristo, M., Banks, E., Poplin, R. E., Garimella, K. ., Maguire, J. R., Hartl, C., Philippakis, A. A., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data, *43*(5), 491–498. doi:10.1038/ng.806.A

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9. doi:10.1093/bioinformatics/btp352

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., et al. (1999). A general approach to single-nucleotide polymorphism discovery. *Nature genetics*, *23*(4), 452–6. doi:10.1038/70570

Vapnik, V. N. Statistical Learning Theory. New York: Wiley, 1998, p. 736