SIFT

SIFT

What does it mean?

Sorting = classify

Intolerant = prejudiced

From

Tolerant = forbearin

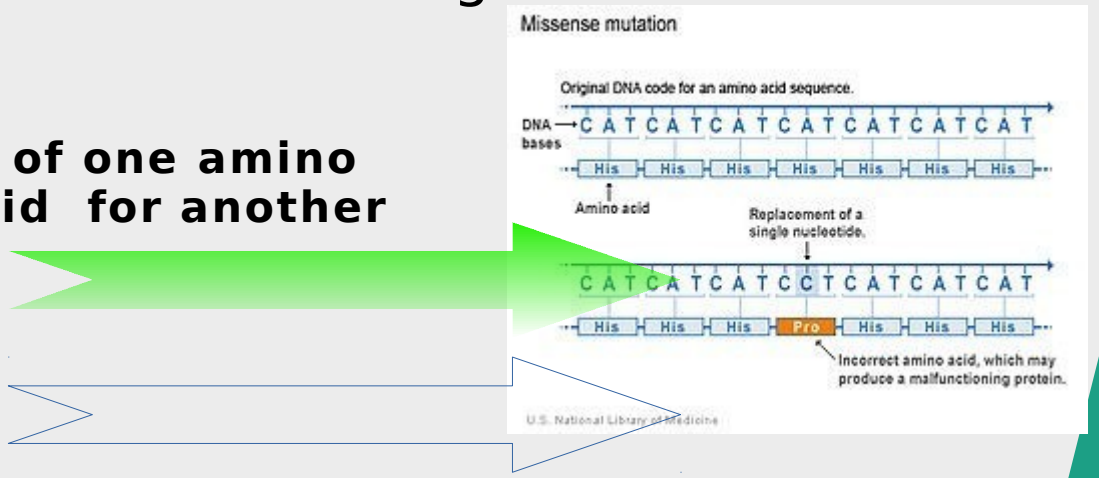
In spanish...

Clasificación de intolerante a tolerante

What does SIFT do?

1. SIFT classifies an amino acid change as tolerated or deleterious to protein function.
2. SIFT is an algorithm that predicts whether an amino acid substitution is deleterious to protein function, and it is often used to prioritize nonsynonymous or missense variants.
3. SIFT takes into account protein conservation with homologous sequences and the severity of the amino acid change.
4. The SIFT 4G algorithm is a faster version of SIFTpredictions quickly and to construct prediction databases for a large number of organisms.

Exchange of one amino acid for another

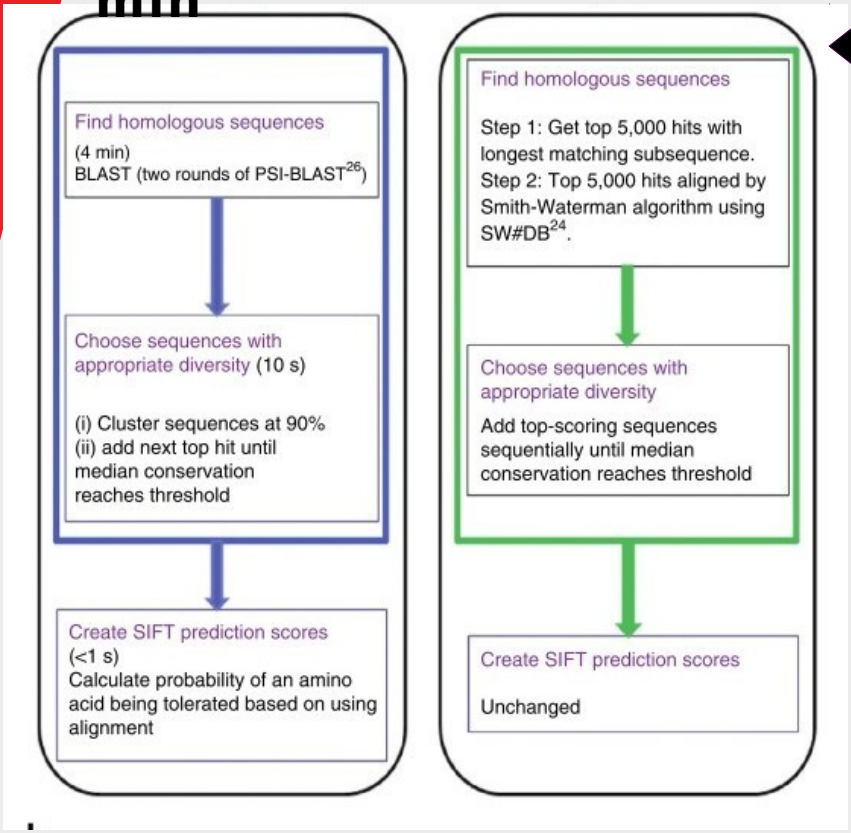


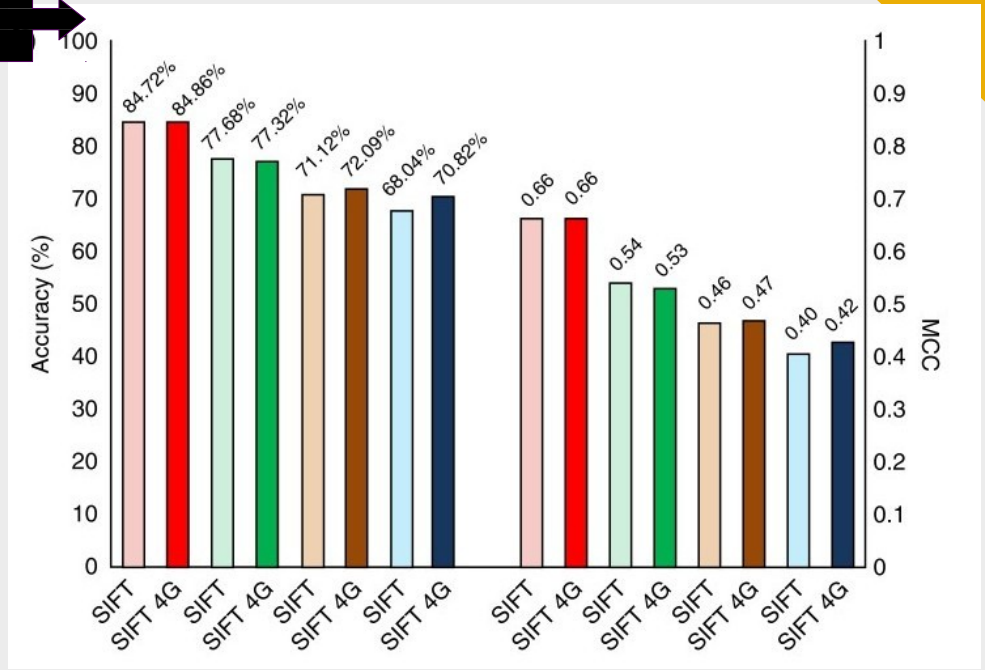
SIFT & SIFT4G

SIFT ~4 min

SIFT ~3 s

The two sofwares used PSI-BLAST .





The steps of the SIFT and SIFT 4G algorithms are shown on the left and right, respectively. The principle of each step has been preserved, but the first two steps have been optimized for speed in the SIFT 4G algorithm.

MCC is a balanced measure of the true and false positives and negatives.

As different heuristic algorithms are used in the first step, the results from SIFT 4G will differ from SIFT.

SIFT & SIFT4G

SIFT 4G's heuristic search algorithm achieves drastic speedup compared with PSI- BLAST at the cost of slightly less sensitivity to distant homologous sequences.

*PSI-BLAST is part of the BLAST family of algorithms and it is a heuristic algorithm, **so optimal answers are not guaranteed**

Materials

MATERIALS

EQUIPMENT

- Computer with Internet connection (see Equipment Setup)
- Data files (see Equipment Setup)

EQUIPMENT SETUP

System requirements

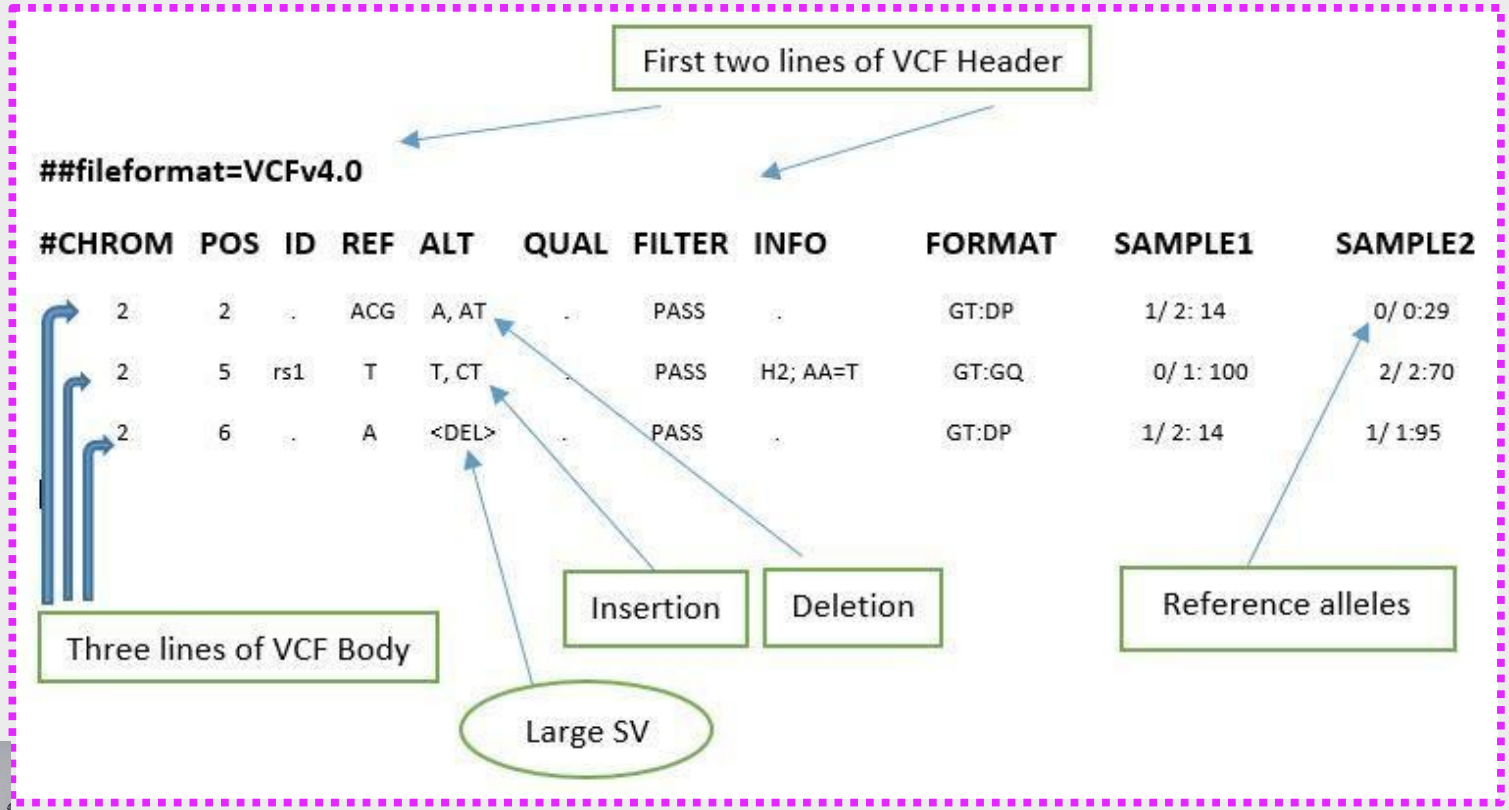
- SIFT 4G annotator: The SIFT 4G annotator requires a computer with Java JRE (Java Runtime Environment) installed (version 1.6 or higher; http://www.java.com/en/) and enough disk space to store the database (which can range from 120 MB for *Escherichia coli* to 3.9 GB for human). The SIFT 4G annotator is platform-independent, and it can run on Windows, Linux and Mac.
- SIFT 4G algorithm: The SIFT 4G algorithm requires any Linux distribution (we have used Ubuntu 12.04) with the following compilers: gcc (version 4 or higher; https://gcc.gnu.org/) and nvcc (version 2 or higher; https://developer.nvidia.com/cuda-downloads). For fast performance, the recommended configuration should include a NVIDIA graphics card (compute capability version 1.3 or higher) and a solid-state drive (SSD).

Data files

- SIFT 4G annotator accepts a list of genomic variants in variant call format (VCF), which is generated by most next-generation sequencing pipelines.

To remember...

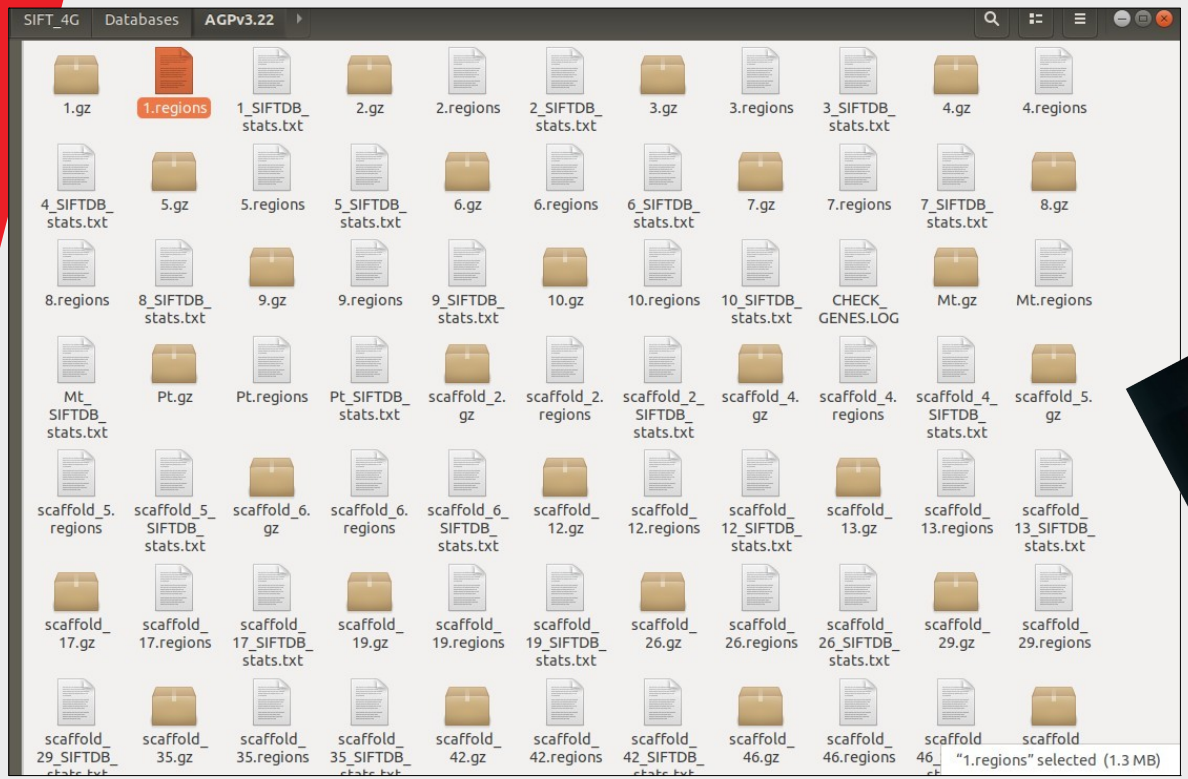
What does a vcf look like?

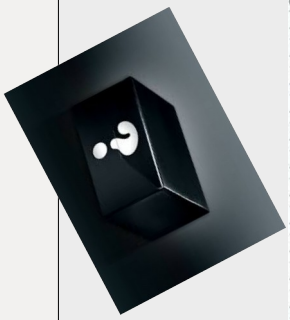


Sure they want to kill me, ja ... long story with this vcf... For me it was important to follow up, I'm sorry ...

Materials

A reference database is needed, I downloaded the corn database suggested by the software. However, it is 2014 data. I suppose I could update the reference base. I'm not sure...





6518	6638	OUT
6639	6797	IN
6798	6917	OUT
6918	7120	IN
7121	7593	OUT
7594	7903	IN
7904	9192	OUT
9193	9652	IN
9653	9881	OUT
9882	10387	IN
10388	109518	OUT
109519	109675	IN
109676	109758	OUT
109759	109935	IN
109936	110768	OUT
110769	111142	IN
111143	111465	OUT
111466	111769	IN
111770	136306	OUT
136307	136634	IN
136635	136718	OUT
136719	137013	IN
137014	137204	OUT
[illegible]	137290	IN
137291	138551	OUT
138552	138929	IN
138930	144360	OUT
144361	144657	IN
144658	144956	OUT
[illegible]	[illegible]	IN

Text Editor

Black boxes behind the software we use. I guess, the important thing would be to understand the software used well. The problem, the time….

To annotate the input variants, the SIFT 4G annotator uses the chromosome (CHROM), position (POS), reference (REF) and alternate (ALT) alleles from the input file and appends the output to the INFO column (the eighth column).

The result in an excel table

	A	B	
1	CHROM	POS	REF_ALLELE
2	1	1034944	CCCA
3	1	1034944	CCCA

	E	F	G	H	I	J	K	L
1	TRANSCRIPT_ID	GENE_ID	GENE_NAME	REGION	VARIANT_TYPE	REF_AMINO	ALT_AMINO	AMINO_POS
2	GRMZM2G05258	GRMZM2G052586	NA	UTR_3	NONFRAMESHIFT DELETION	NA	NA	NA
3	GRMZM2G05258	GRMZM2G052586	NA	UTR_3	NONFRAMESHIFT DELETION	NA	NA	NA
4	GRMZM2G05258	GRMZM2G052586	NA	UTR_3	NONFRAMESHIFT DELETION	NA	NA	NA
5	GRMZM2G05258	GRMZM2G052						
6	GRMZM2G05258	GRMZ						
7	GRMZM2G05258	GRMZM2G052						
8	GRMZM2G05258	GRMZM2G052						
9	GRMZM2G05258	GRMZM2G052						
10	GRMZM2G05258	GRMZM2G052						
11	GRMZM2G08588	GRMZM2G085						
12	GRMZM2G08588	GRMZM2G085						
13	GRMZM2G08588	GRMZM2G085						
14	GRMZM2G08588	GRMZM2G085						
15	GRMZM2G08588	GRMZM2G085						
16	GRMZM2G08588	GRMZM2G085						
17	GRMZM2G04255	GRMZM2G042						
18	GRMZM2G04255	GRMZM2G042						
19	GRMZM2G04255	GRMZM2G042						
20	GRMZM2G04255	GRMZM2G042						

Sheet 1 of 1

Zea_mays.AGPv4.2x_0.8_0.01.NewID

	M	N	O	P	Q
1	SIFT_SCORE	SIFT_MEDIAN	NUM_SEQS	dbSNP	SIFT_PREDICTION
2	NA	NA	NA	NA	NA
3	NA	NA	NA	NA	NA
4	NA	NA	NA	NA	NA
5	NA	NA	NA	NA	NA
6	NA	NA	NA	NA	NA
7	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA
9	NA	NA	NA	NA	NA
10	NA	NA	NA	NA	NA
11	NA	NA	NA	NA	NA
12	NA	NA	NA	NA	NA
13	NA	NA	NA	NA	NA
14	NA	NA	NA	NA	NA
15	NA	NA	NA	NA	NA
16	NA	NA	NA	NA	NA
17	NA	NA	NA	NA	NA
18	NA	NA	NA	NA	NA
19	NA	NA	NA	NA	NA
20	NA	NA	NA	NA	NA
21	NA	NA	NA	NA	NA
22	NA	NA	NA	NA	NA
23	NA	NA	NA	NA	NA
24	NA	NA	NA	NA	NA
25	NA	NA	NA	NA	NA
26	NA	NA	NA	NA	NA
27	NA	NA	NA	NA	NA
28	NA	NA	NA	NA	NA
29	NA	NA	NA	novel	NA
30	NA	NA	NA	novel	NA
31	NA	NA	NA	novel	NA
32	NA	NA	NA	novel	NA
33	NA	NA	NA	novel	NA

.recode_SIFTannotations

At this moment I am not clear...

As it passes from variants to amino acids. I guess with a blastx, however, I'm not sure.....

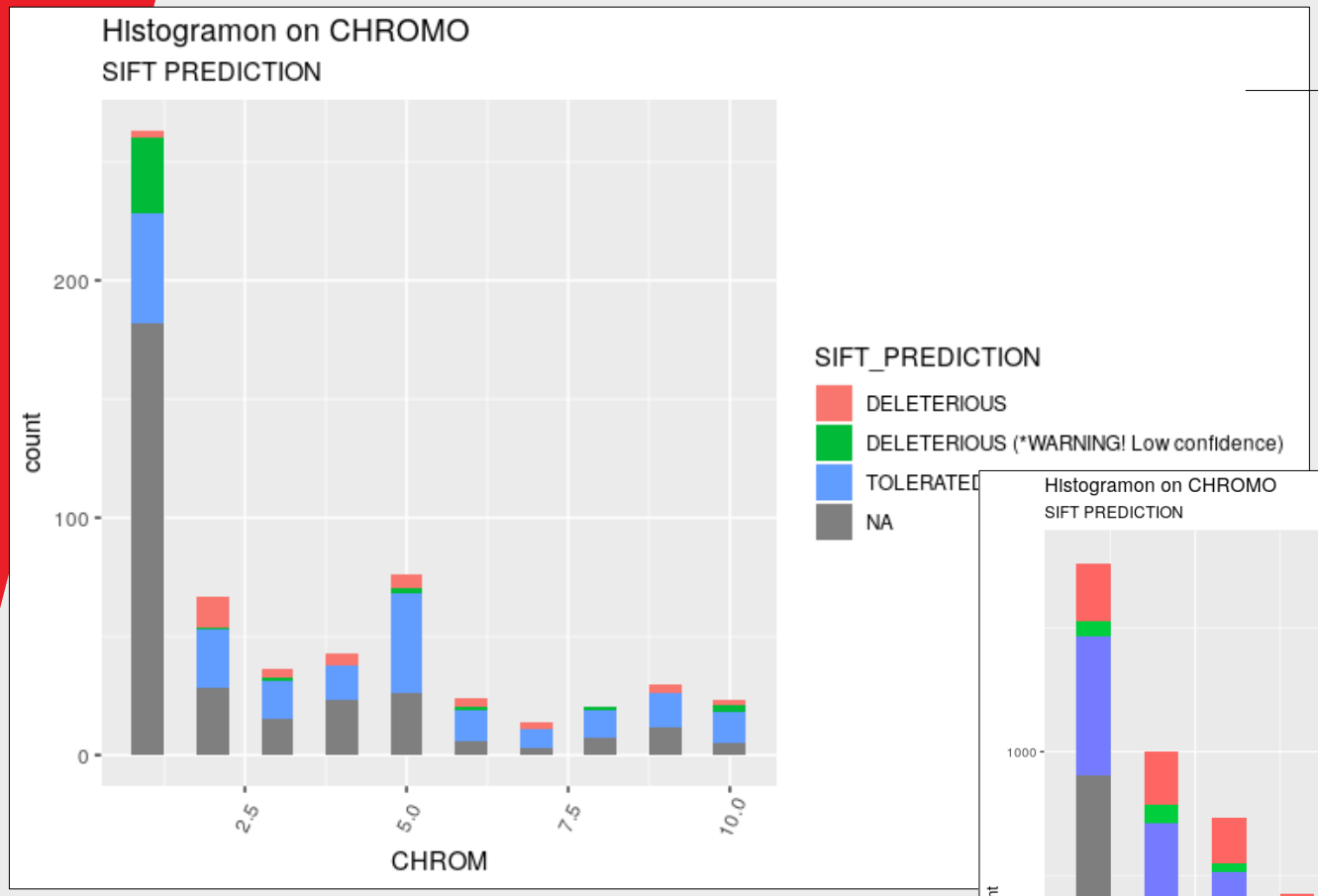
- If the user does not have a VCF file, the user can format their list of variants in a VCF-like file, which should have at least eight columns. A sample VCF is shown in **Supplementary Table 1**. **! CAUTION** If the input file does not have at least eight columns, it will not be annotated as the prediction is appended to the eighth column. If the user's input file contains the chromosome, position, reference and alternate alleles alone, the user can append dummy columns to ensure that the input file will have at least eight columns.

Required inputs

- The SIFT 4G algorithm requires three inputs. The first input is a directory of '.fasta' or '.fa' files where each file contains a protein sequence in FASTA format with the protein name in the description line. The algorithm also requires a companion input file containing a list of amino acid substitutions for each protein sequence, and it will compute predictions for these substitutions. The third required input is the protein sequence database to search homologous sequences—for example, the UniRef90 (ref. 27) or NCBI nonredundant[33] protein databases.

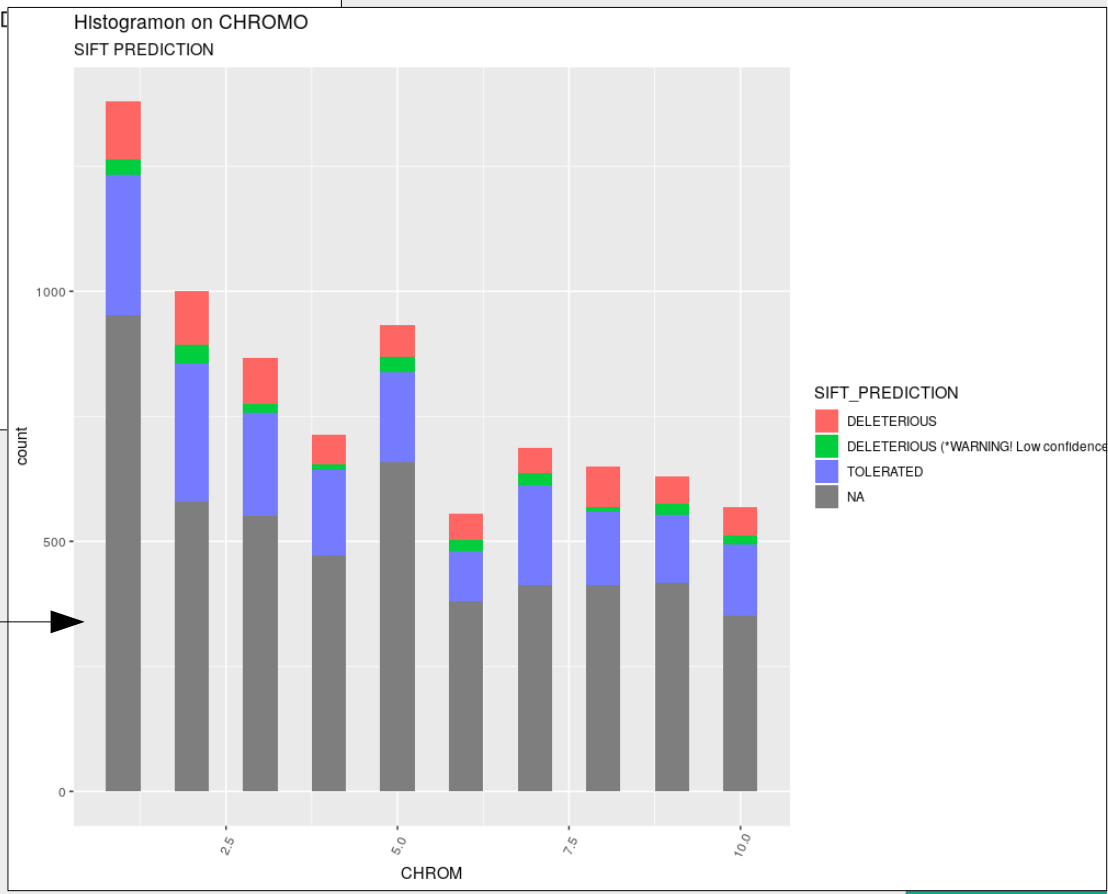
Results SNP array & GBS

They are not the same maices butI think it is very important to discuss...



SNP array (Arteaga et al 2016)

GBS (Rojas et al 2019)




thank
you!