Tool for Rapid Annotation of Microbial SNPs (TRAMS) Manual

Contact: Vartul Sangal (vartul.sangal@strath.ac.uk); Paul A. Hoskisson (paul.hoskisson@strath.ac.uk)
Bug reports: Richard A. Reumerman (richard.reumerman@strath.ac.uk)

Citation: Reumerman RA, Tucker NP, Herron PR, Hoskisson PA, Sangal V. Tool for Rapid Annotation of Microbial SNPs (TRAMS): a simple program for rapid annotation of genomic variation in prokaryotes, Antonie van Leeuwenhoek Journal of Microbiology (*submitted*)

1. Introduction

TRAMS is a program, written using Python 2.7.3, for functional annotation of genomic SNPs as synonymous, nonsynonymous or nonsense. It separates nonsynonymous SNPs in start and stop codons as non-start and non-stop SNPs, respectively and SNPs in multiple overlapping features are annotated separately for each feature. Multiple nucleotide polymorphisms (MNPs) within a codon are combined before annotation.

2. Installation

TRAMS was developed using Python version 2.7.3, but works under any version of the 2.7 series. The users need to install python, biopython and Numpy to use this program. These are available to download from:

Python 2.7 http://www.python.org/download/,
BioPython http://biopython.org/wiki/Download
NumPy http://sourceforge.net/projects/numpy/files/NumPy/
TRAMS http://sourceforge.net/projects/strathtrams/files/Latest/

The distributions of Python 2.7.3, NumPy and BioPython 1.60 are covered by different third party licences.
We provide a quick guide for installation of Python 2.7.3, NumPy and BioPython 1.60 but the users are highly encouraged to carefully read the instructions and detailed procedure from the download pages (mentioned above) for WINDOWS, Linux and Mac OS platforms.

Windows
TRAMS is available to download as a single file executable for WINDOWS users who do not need to install Python, BioPython or NumPy.

However, if they wish to use TRAMS as Python script, they can download WINDOWS installers for Python 2.7.3, NumPy and BioPython 1.60 from the above websites in the mentioned order.

Linux
Some Linux distributions such as Ubuntu may have Python 2.7.x pre-installed and the users can find out the version by using the command ***python*** in the console. The detailed download and installation instructions for different Linux distributions are available at the Python, BioPython and NumPy download pages (mentioned above). Here, we provide quick commands for Ubuntu using apt-get package manager:

sudo apt-get install python-numpy python-biopython

Mac OS
Mac OS often have Python pre-installed but if you need to install python, NumPy or BioPython, ***fink*** may be useful. The detailed instructions to use ***fink*** are available at http://finkers.wordpress.com/installing-fink/#install-fast.install.

To install Python, NumPy and BioPython on a MacOS, following command can be used:

fink install biopython-py27

That should be sufficient. See the BioPython installation instructions for more information: http://biopython.org/DIST/docs/install/Installation.html#htoc28.

3. Using the program

i. Input file format: TRAMS uses a tab delimited text file containing SNP locations, reference nucleotide and SNPs in variant strains along with a reference genome sequence in GenBank or EMBL format.

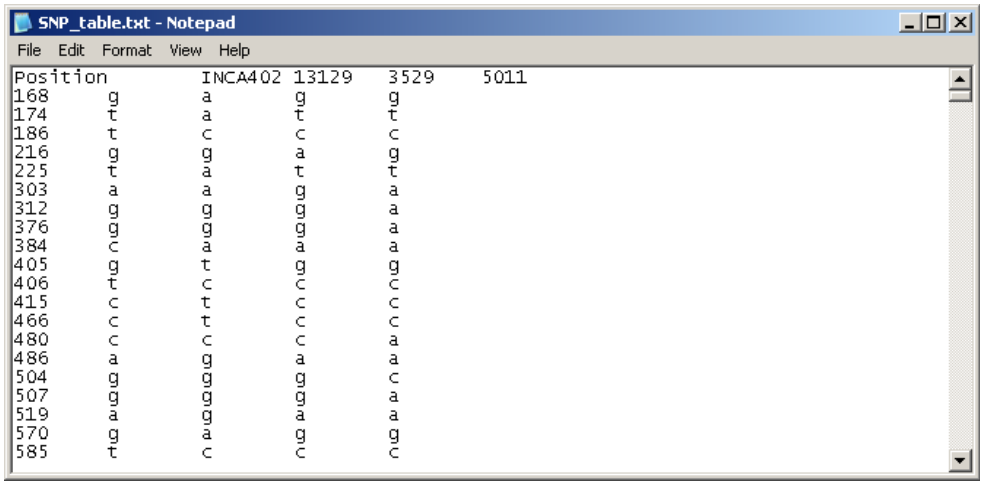


Figure 1: Input file format for TRAMS

As shown in the Fig, 1, the first row in the SNP input file is the header, followed by the data from the second row containing SNP positions in the reference genome (first column), nucleotide in the

reference strain (second column) and SNP alleles in the variant strains (3 column onwards). SNPs in multiple strains can be added to the same input file as separate columns. If a base is not a SNP in a certain strain, it can either be left blank or kept the same as the reference base. In both cases the base will be ignored for that strain in the annotation file. There is no limit to the number of strains that can be added that may be dependent on the memory of your computer.
The GenBank and EMBL formats are recognized directly by the file extensions (.gbk or .embl, respectively).

We have separately provided two scripts to convert the SNPs tables extracted from Mauve (***mauve2trams.py***) and vcf files (***vcf2trams.py***) into TRAMS input. Please see the user manual for generating TRAMS input files from vcf and Mauve output files for details.

ii. Running the program:

Windows: TRAMS Windows executable can be run without any installation requirements. Double-click on the icon to start the program. A console window will open and prompt for the SNP input table and the reference genome file. Make sure the program is placed in the same directory as the input files. The output files will also be placed in this same directory.

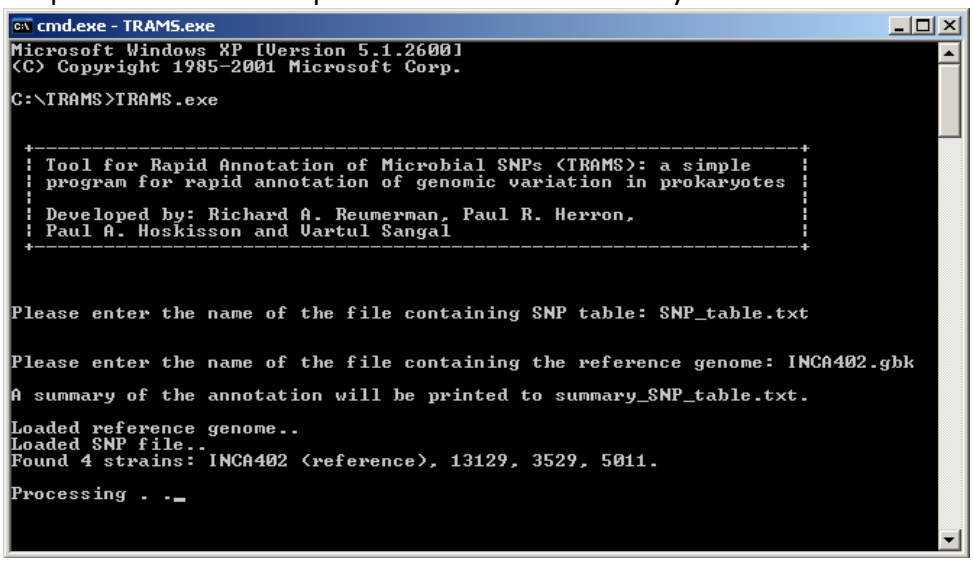


Figure 2: A snapshot of the TRAMS console from the executable file

You can also run the program, both the executable and the python script, using the following command line:

TRAMS.exe -s SNP_Table.txt -r Reference.gbk or

TRAMS.py -s SNP_Table.txt -r Reference.gbk

-s name of the input file
-r Reference genome in GenBank or EMBL format
The users can use “-h” to display the help message.

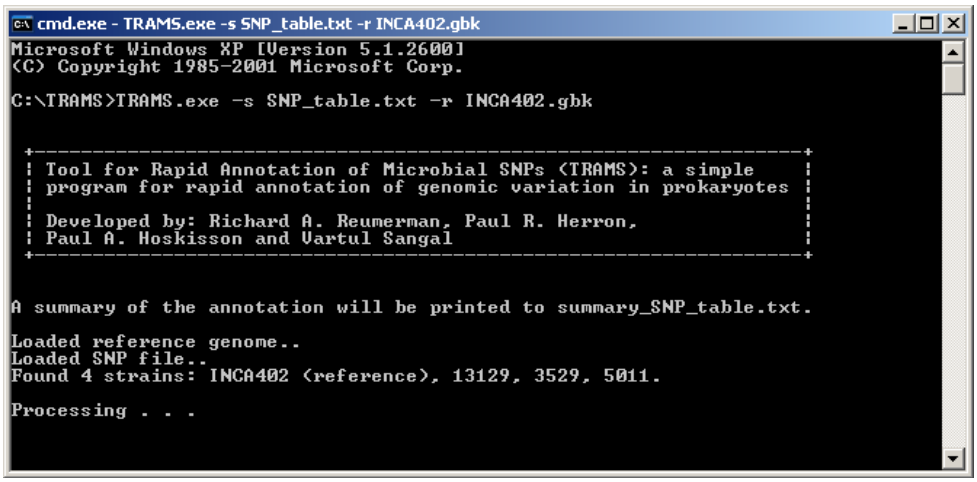


Figure 3: Running TRAMS using the command line

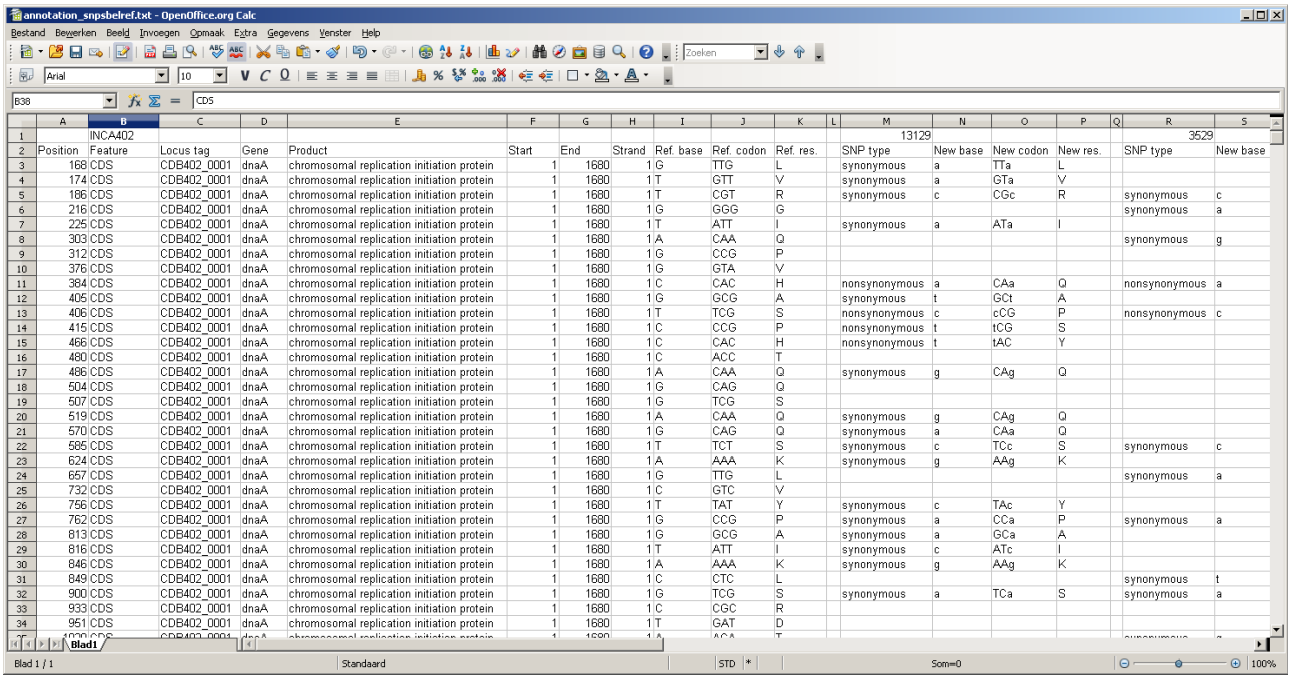
TRAMS script will need the correct versions and the necessary modules (Python 2.7.3, BioPython 1.60 and NumPy) for proper functioning. Using the script with different versions (e.g. Python 3.0) cannot be guaranteed to work. The script will not run under Python 2.6, since TRAMS uses the 'argparse' module, which is absent from 2.6. Once the necessary software is installed the script can be run in the same way as the windows executable, using the same command line code. See the installation instructions for more details on setting up the system for TRAMS.

iii. Output files

TRAMS will produce three output files. These are placed in the working directory of the program is running (where the input files are saved). The file names used are the same as the SNP tables, but with prefixes. In the following, the name “SNP_table.txt” is used for demonstration purposes.

Annotation_SNP_table.txt: This file contains the annotation of every SNP in the SNP table in tab delimited format. The first column contains the SNP positions. The following columns describe the feature the SNP is located in (if any), stating feature type, locus tag, gene name, gene product, feature start, feature end, feature strand, base, codon and amino acid. After this the SNP type, mutant base, new codon and new amino acid are given for each strain. If a base in a strain is the same as in the reference strain, the row will be empty for that strain.

If multiple SNPs occur in a single codon, all but the last one in that codon will be labelled “MNP” (Multiple Nucleotide Polymorphism). The annotation of the last SNP in the codon will then take the cumulative effect of all SNPs into account. Mutant types used to annotate SNPs and MNPs in TRAMS are *synonymous* (no change in amino acid), *nonsynonymous* (different amino acid), *nonsense* (normal codon replaced by stop codon), *nonstop* (stop codon replaced by normal codon) and *nonstart* (start codon at start of feature replaced by a normal codon). For this annotation, codon table 11 is used (Bacterial and Plant Plastid) as provided in the BioPython module.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		INCA402											13129					3529	
2	Position	Feature	Locus tag	Gene	Product	Start	End	Strand	Ref. base	Ref. codon	Ref. res.		SNP type	New base	New codon	New res.		SNP type	New base
3	168	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	TTG	L		synonymous	a	TTa	L			
4	174	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	GTT	V		synonymous	a	GTa	V			
5	186	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	CGT	R		synonymous	c	CGc	R		synonymous	c
6	216	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	GGG	G							synonymous	a
7	225	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	ATT	I		synonymous	a	ATa	I			
8	303	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	A	CAA	Q							synonymous	g
9	312	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	CCG	P								
10	376	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	GTA	V								
11	384	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	CAC	H		nonsynonymous	a	CAa	Q		nonsynonymous	a
12	405	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	GCG	A		synonymous	t	GCt	A			
13	406	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	TCG	S		nonsynonymous	c	cCG	P		nonsynonymous	c
14	415	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	CCG	P		nonsynonymous	t	tCG	S			
15	466	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	CAC	H		nonsynonymous	t	tAC	Y			
16	480	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	ACC	T								
17	486	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	A	CAA	Q		synonymous	g	CAg	Q			
18	504	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	CAG	Q								
19	507	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	TCG	S								
20	519	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	A	CAA	Q		synonymous	g	CAg	Q			
21	570	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	CAG	Q		synonymous	a	CAa	Q			
22	585	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	TCT	S		synonymous	c	TCc	S		synonymous	c
23	624	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	A	AAA	K		synonymous	g	AAg	K			
24	657	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	TTG	L							synonymous	a
25	732	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	GTC	V								
26	756	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	TAT	Y		synonymous	c	TAc	Y			
27	762	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	CCG	P		synonymous	a	CCa	P		synonymous	a
28	813	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	GCG	A		synonymous	a	GCa	A			
29	816	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	ATT	I		synonymous	c	ATc	I			
30	846	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	A	AAA	K		synonymous	g	AAg	K			
31	849	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	CTC	L							synonymous	t
32	900	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	G	TCG	S		synonymous	a	TCa	S		synonymous	a
33	933	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	C	CGC	R								
34	951	CDS	CDB402_0001	dnaA	chromosomal replication initiation protein	1	1680	1	T	GAT	D								

Figure 4: A snapshot of annotation file

Summary_SNP_table.txt: The summary file provides a brief overview of the results of the annotation process. Most importantly, it states the number of SNPs in each strain and how they are divided over different feature and SNP types.

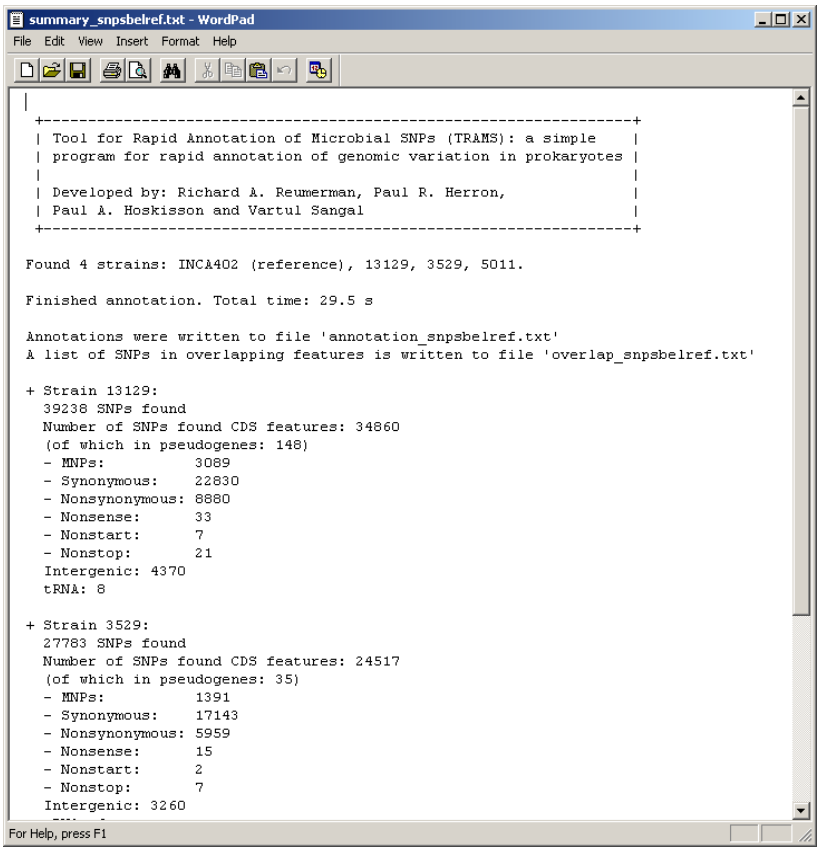
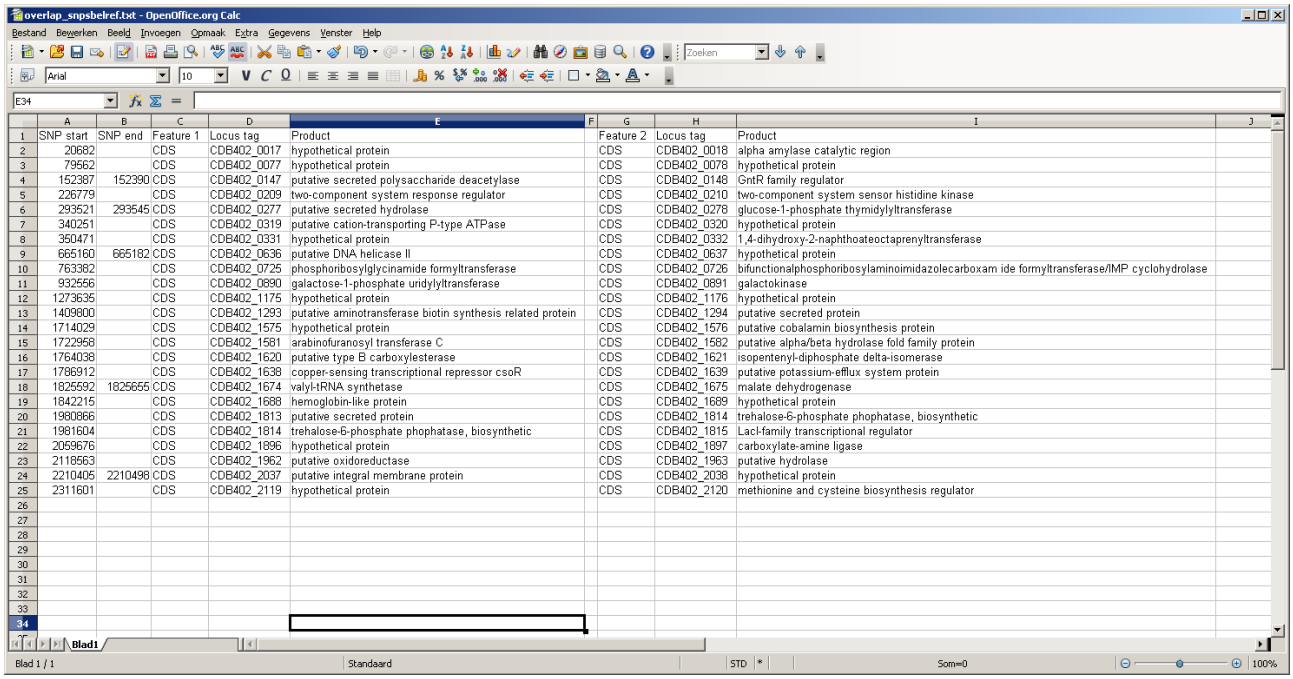


Figure 5: A snapshot of the summary output

Overlap_SNP_table.txt: Since TRAMS annotates SNPs once for each feature they occur in, they will be annotated more than once when features overlap. This is reported in the overlap file, where a

list of SNPs that occur more than once is given. This file is in tab-delimited format and contains the first and last of the range of overlapping SNPs of each feature overlap. It then gives some information on the two overlapping features.



overlap_snpsbelref.txt - OpenOffice.org Calc

Bestand Bewerken Beeld Invoegen Opmaak Extra Gegevens Venster Help

	A	B	C	D	E	F	G	H	I	J
1	SNP start	SNP end	Feature 1	Locus tag	Product		Feature 2	Locus tag	Product	
2	20682		CDS	CDB402_0017	hypothetical protein		CDS	CDB402_0018	alpha amylase catalytic region	
3	79662		CDS	CDB402_0077	hypothetical protein		CDS	CDB402_0078	hypothetical protein	
4	152387	152390	CDS	CDB402_0147	putative secreted polysaccharide deacetylase		CDS	CDB402_0148	GntR family regulator	
5	226779		CDS	CDB402_0209	two-component system response regulator		CDS	CDB402_0210	two-component system sensor histidine kinase	
6	293521	293545	CDS	CDB402_0277	putative secreted hydrolase		CDS	CDB402_0278	glucose-1-phosphate thymidylyltransferase	
7	340251		CDS	CDB402_0319	putative cation-transporting P-type ATPase		CDS	CDB402_0320	hypothetical protein	
8	350471		CDS	CDB402_0331	hypothetical protein		CDS	CDB402_0332	1,4-dihydroxy-2-naphthoateoctaprenyltransferase	
9	665160	665182	CDS	CDB402_0636	putative DNA helicase II		CDS	CDB402_0637	hypothetical protein	
10	763382		CDS	CDB402_0725	phosphoribosylglycinamide formyltransferase		CDS	CDB402_0726	bifunctionalphosphoribosylaminoimidazolecarboxam ide formyltransferase/IMP cyclohydrolase	
11	932556		CDS	CDB402_0890	galactose-1-phosphate uridylyltransferase		CDS	CDB402_0891	galactokinase	
12	1273635		CDS	CDB402_1175	hypothetical protein		CDS	CDB402_1176	hypothetical protein	
13	1409800		CDS	CDB402_1293	putative aminotransferase biotin synthesis related protein		CDS	CDB402_1294	putative secreted protein	
14	1714029		CDS	CDB402_1575	hypothetical protein		CDS	CDB402_1576	putative cobalamin biosynthesis protein	
15	1722958		CDS	CDB402_1581	arabinofuranosyl transferase C		CDS	CDB402_1582	putative alpha/beta hydrolase fold family protein	
16	1764038		CDS	CDB402_1620	putative type B carboxylesterase		CDS	CDB402_1621	isopentenyl-diphosphate delta-isomerase	
17	1786912		CDS	CDB402_1638	copper-sensing transcriptional repressor csoR		CDS	CDB402_1639	putative potassium-efflux system protein	
18	1825592	1825655	CDS	CDB402_1674	valyl-tRNA synthetase		CDS	CDB402_1675	malate dehydrogenase	
19	1842215		CDS	CDB402_1688	hemoglobin-like protein		CDS	CDB402_1689	hypothetical protein	
20	1980866		CDS	CDB402_1813	putative secreted protein		CDS	CDB402_1814	trehalose-6-phosphate phophatase, biosynthetic	
21	1981604		CDS	CDB402_1814	trehalose-6-phosphate phophatase, biosynthetic		CDS	CDB402_1815	LacI-family transcriptional regulator	
22	2059676		CDS	CDB402_1896	hypothetical protein		CDS	CDB402_1897	carboxylate-amine ligase	
23	2118563		CDS	CDB402_1962	putative oxidoreductase		CDS	CDB402_1963	putative hydrolase	
24	2210405	2210498	CDS	CDB402_2037	putative integral membrane protein		CDS	CDB402_2038	hypothetical protein	
25	2311601		CDS	CDB402_2119	hypothetical protein		CDS	CDB402_2120	methionine and cysteine biosynthesis regulator	
26										
27										
28										
29										
30										
31										
32										
33										
34										

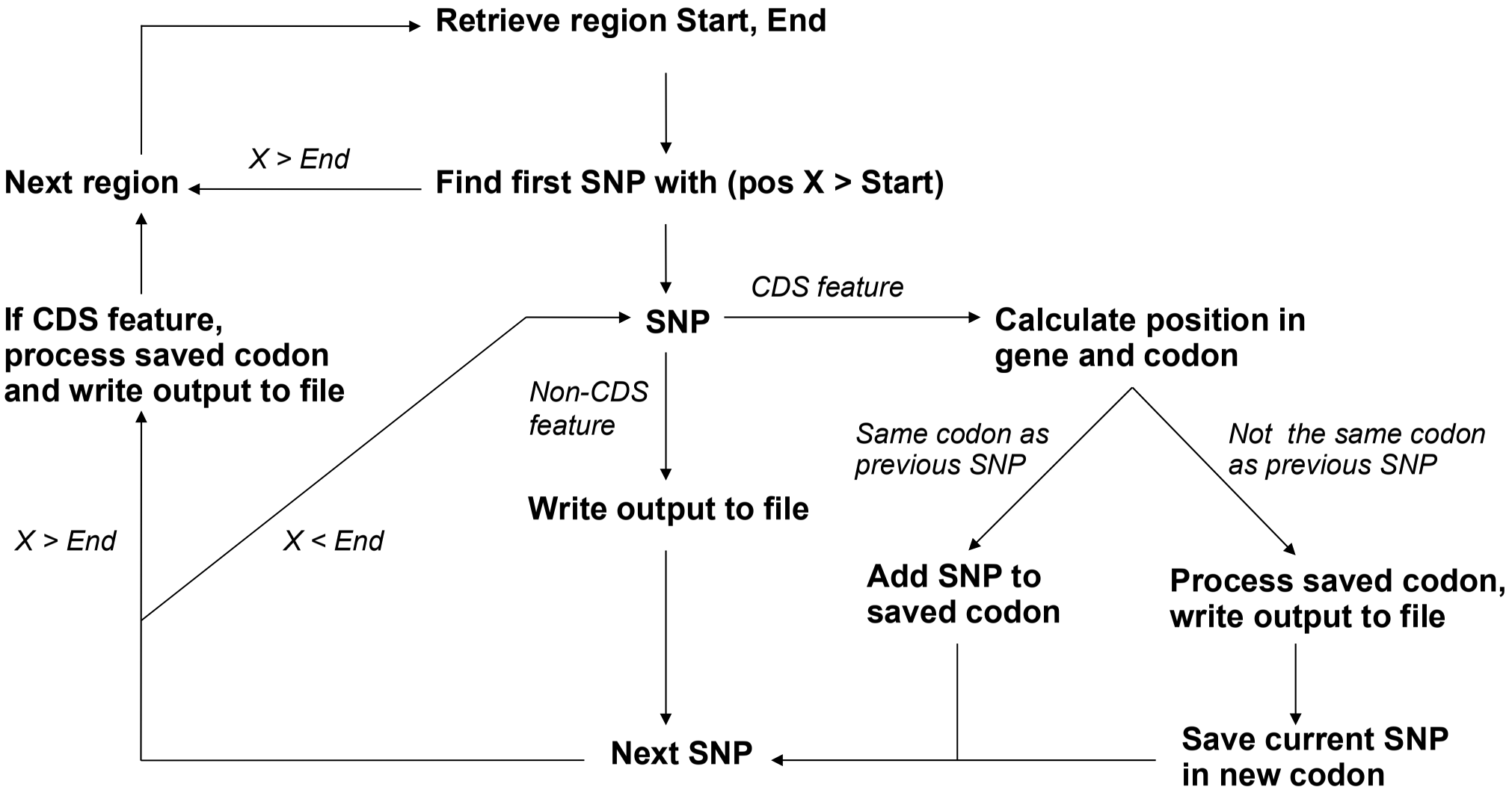
Blad1

Blad 1 / 1 Standaard STD Som=0 100%

Figure 6: A snapshot of list containing overlapping SNPs

Algorithm

The way TRAMS annotates SNPs is outlined in the diagram below. The sequence is divided into regions containing either a feature or a featureless region (intergenic). SNPs are annotated region by region, ensuring that SNPs in overlaping features are annotated once for every feature that spans their position. Multiple SNPs in a single codon are combined, resulting in the annotation of their cumulative effect on the codon rather than their effect in isolation.


Retrieve region Start, End
Find first SNP with (pos X > Start)
X > End
Next region
If CDS feature, process saved codon and write output to file
X > End
X < End
SNP
CDS feature
Non-CDS feature
Write output to file
Calculate position in gene and codon
Same codon as previous SNP
Not the same codon as previous SNP
Add SNP to saved codon
Process saved codon, write output to file
Save current SNP in new codon
Next SNP