BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(11)

SNPAAMapper*T2K*: A genome-wide SNP downstream analysis and annotation pipeline for species annotated with NCBI.tbl data files

Yongsheng Bai[1, 2]

[1]Department of Biology, [2]The Center for Genomic Advocacy, Indiana State University 600 Chestnut Street, Terre Haute, IN 47809, U.S.A; Yongsheng Bai - Email: Yongsheng.Bai@indstate.edu

Received November 18, 2014; Accepted November 19, 2014; Published November 27, 2014

Abstract:
SNPAAMapper, a genome-wide SNP downstream analysis and annotation pipeline, was designed to classify detected variants according to genomic regions and report the mutation class by processing whole-genome and/or whole-exome sequencing data. A widely used sequence and data annotation table format "knownGene.txt" has not yet been created for many popular model organisms (e.g. Arabidopsis). Instead, NCBI .tbl annotation format files are provided for these species. Therefore, it is of interest to describe SNPAAMapper*T2K*, a genome-wide SNP downstream analysis and annotation pipeline for species annotated with NCBI .tbl data files (e.g. Arabidopsis). The pipeline is tested with a deeply sequenced Arabidopsis thaliana strain (Seattle-0). The SNPAAMapper*T2K* can also annotate and report SNP classes for other species, whose chromosome files are annotated as NCBI .tbl format, but do not have their annotated knownGene.txt files available.

Availability: Perl scripts and required input files are available on the web at http://isu.indstate.edu/ybai2/SNPAAMapperT2K

Background:
Exome sequencing technology is being employed to identify single nucleotide polymorphisms (SNPs) and/or insertions and deletions (INDELs) in genetic disease research. The schema for UCSC Genes (knownGene.txt) **[1]** has been widely employed for use in both standard and customized downstream analysis tools and scripts. However, even for many popular model organisms (e.g. Arabidopsis), sequence and annotation data tables (including knownGene.txt) have not yet been made available to the public. SNPAAMapper **[2]**, a genome-wide SNP analysis and annotation pipeline using whole-genome and/or whole-exome sequencing data, has been developed to perform the downstream annotation for detected variants; this tool can classify variants by regions and report the hit class and requires knownGene.txt as one of its input files. We have developed a tool - Tbl2KnownGene **[3]**, a .tbl file parser that can process the contents of a National Center for Biotechnology Information (NCBI) .tbl file (e.g. the one for Arabidopsis genome (TAIR10)) **[4, 5]** and produce a UCSC Known Genes annotation feature table. Arabidopsis chromosomes are annotated as .tbl files by TAIR, so their knownGene.txt format files are not available.

In this study, we have developed SNPAAMapper*T2K*, a genome-wide SNP analysis pipeline for species that has .tbl but not knownGene.txt files available. We have generated annotation files for Arabidopsis and users can easily download them onto their computers and run their sequence read files against the supporting files.

Our pipeline can be easily extended to analyze SNP annotation for other species which were annotated using .tbl files, but do not have annotated knownGene.txt files available.

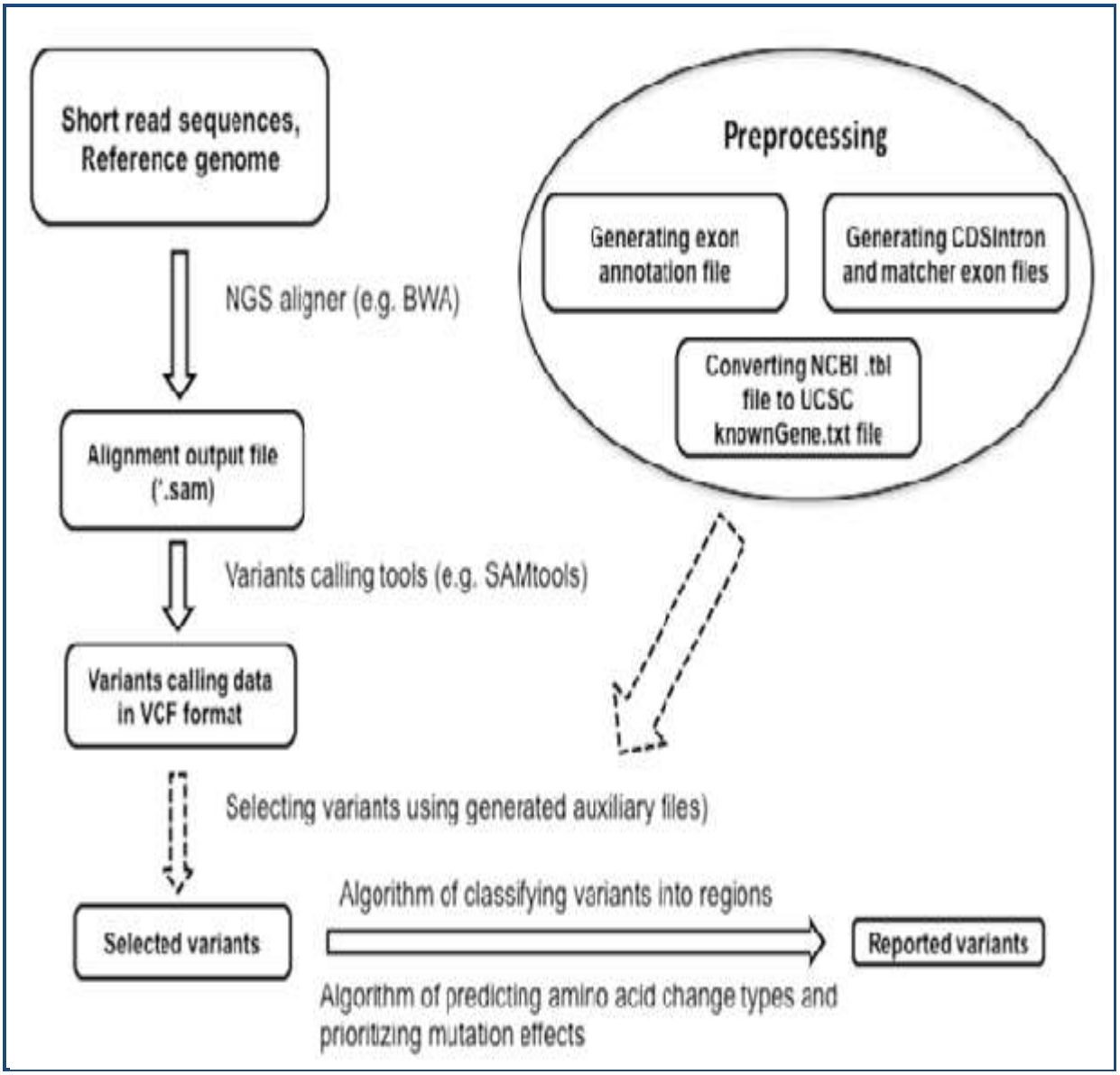

Figure 1: The Workflow of SNPAAMapper*T2K*

Methodology:

The SNPAAMapper*T2K* algorithm consists of two major modules: the first module converts NCBI .tbl file to UCSC knownGene.txt file format, and the second module uses converted KnownGene files and calls BWA **[5, 6]** and SAMTools **[7]** and custom scripts to report the hit class. The workflow of SNPAAMapper*T2K* is shown in **Figure 1.**

SNPAAMapper*T2K* Input and Output:

The inputs are NCBI .tbl files (e.g. the chromosome files of Arabidopsis), TAIR10 sequence annotation files, and short read sequence files. The outputs are annotated variant files. A subset (non-synonymous SNPs) of annotated variants by SNPAAMapper*T2K* is shown in **Table 1 (see supplementary material).**

Conclusions:

Efficient pipelines/tools are needed for downstream genome-wide variant analyses for next-generation sequencing data. We developed a bioinformatics pipeline – SNPAAMapper*T2K* that parses the contents of a NCBI .tbl annotation table, produces a UCSC Known Genes annotation table, and finally calls customized scripts to classify variants and annotate their hit classes. The pipeline was tested with a deeply sequenced Arabidopsis thaliana strain (Seattle-0) from 1001 Genomes Data Center **[8].**

Acknowledgement:

This research was supported by the Indiana State University Start-Up funds to YB.

References:

[1] Kent WJ *et al*. *Genome Res*. 2002 **12:** 996 [PMID: 12045153]
[2] Bai Y & Cavalcoli J, *Bioinformation* 2013 **9**: 870 [PMID: 24250114]
[3] Bai Y, *Bioinformation* 2014 **10**: 544 [PMID: 25258492]
[4] http://www.ncbi.nlm.nih.gov/genbank
[5] http://www.arabidopsis.org/

[6] Li H & Durbin R, *Bioinformatics* 2009 **25:** 1754 [PMID: 19451168]
[7] Li H *et al*. *Bioinformatics* 2009 **25:** 2078 [PMID: 19505943]
[8] http://1001genomes.org/data/SLU/SLUHenning2014/releases/current/strains/Seattle-0/Reads/

Edited by P Kangueane
Citation: **Bai,** Bioinformation 10(11): 711-715 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: A subset (non-synonymous SNPs) of annotated variants by SNPAAMapper*T2K*

Chromosome	Position	Gene	Strand	AA Position	Type	Codon Change	AA Change	Hit Class	Ref	Alt	Read Depth
3	16648869	AT3G45390	-	369	SNP	S(AGT)->I(ATT),S(AGT)->T(ACT)	NSM,NSM	CDSHIT	C	A,G	7
5	16477793	ARABIDOPSIS THALIANA PURINE PERMEASE 12, ATPUP12, MEE6.23, MEE6_23, PUP12, purine permease 12	+	145	SNP	Y(TAT)->S(TCT)	NSM	CDSHIT	A	C	7
2	1688873	F28I8.15, F28I8_15	+	34	SNP	Y(TAC)->S(TCC)	NSM	CDSHIT	A	C	8
5	8197082	K12G2.7, K12G2_7	+	364	SNP	Y(TAC)->C(TGC)	NSM	CDSHIT	A	G	7
1	24322598	FAS1, FASCIATA 1, NFB2, NUCLEOSOME/CHROMATIN ASSEMBLY FACTOR GROUP B	-	281	SNP	V(GTT)->I(ATT)	NSM	CDSHIT	C	T	10
5	2201422	T28J14.20, T28J14_20	+	267	SNP	V(GTG)->E(GAG)	NSM	CDSHIT	T	A	10
1	23642520	F24D7.8, F24D7_8	+	166	SNP	T(ACG)->K(AAG)	NSM	CDSHIT	C	A	5
1	25923838	T6L1.12, T6L1_12	-	466	SNP	S(TCC)->A(GCC)	NSM	CDSHIT	A	C	11
5	5825694	MVA3.30, MVA3_30	+	809	SNP	S(AGT)->N(AAT)	NSM	CDSHIT	G	A	9
4	5348664	C18G5.10, C18G5_10	-	447	SNP	R(CGT)->S(AGT)	NSM	CDSHIT	G	T	12
4	412810	AT4G00955	+	198	SNP	R(CGG)->Q(CAG)	NSM	CDSHIT	G	A	8
5	16834379	MJC20.23, MJC20_23	-	257	SNP	R(AGG)->T(ACG)	NSM	CDSHIT	C	G	6
2	8350309	F27F23.4, F27F23_4	-	507	SNP	R(AGG)->K(AAG)	NSM	CDSHIT	C	T	5
1	11933793	F9L11.10, F9L11_10	-	49	SNP	R(AGA)->S(AGT)	NSM	CDSHIT	T	A	11
5	17561104	MQD19.6, MQD19_6	+	280	SNP	N(AAT)->H(CAT)	NSM	CDSHIT	A	C	17
2	2371960	F5K7.15, F5K7_15	-	19	SNP	N(AAT)->H(CAT)	NSM	CDSHIT	T	G	5
2	13023117	GALACTURONOSYLTRANSFERASE 5, GAUT5, LGT5, los glycosyltransferase 5	-	106	SNP	N(AAC)->T(ACC)	NSM	CDSHIT	T	G	9
5	21320458	T4M5.4, T4M5_4	+	688	SNP	M(ATG)->T(ACG)	NSM	CDSHIT	T	C	13
1	3392354	F14N23.22, F14N23_22	-	43	SNP	M(ATG)->T(ACG)	NSM	CDSHIT	A	G	3
1	7775393	F2E2.13, F2E2_13	-	1545	SNP	M(ATG)->L(CTG)	NSM	CDSHIT	T	G	16
5	22904363	MIK19.1, MIK19_1	+	427	SNP	M(ATG)->L(CTG)	NSM	CDSHIT	A	C	10
1	7628771	AT1G21722	+	74	SNP	L(CTA)->R(CGA)	NSM	CDSHIT	T	G	7
2	18613594	CDS4, T14P1.4,	+	26	SNP	L(CTA)->I(ATA)	NSM	CDSHIT	C	A	3

		cytidinediphosphate diacylglycerol synthase 4									
1	11350555	F27M3.9, F27M3_9	+	234	SNP	K(AAG)->R(AGG)	NSM	CDSHIT	A	G	3
3	17756972	ATEDS1, EDS1, EDS1 PROTEIN, enhanced disease susceptibility 1	-	107	SNP	K(AAG)->N(AAT)	NSM	CDSHIT	C	A	9
2	12440955	T9I4.4, T9I4_4	-	331	SNP	K(AAA)->N(AAC)	NSM	CDSHIT	T	G	6
5	23531310	MCK7.2, MCK7_2	-	422	SNP	I(ATT)->V(GTT)	NSM	CDSHIT	T	C	7
5	52951	F7J8.130, F7J8_130	+	294	SNP	I(ATT)->L(CTT)	NSM	CDSHIT	A	C	6
5	23914653	APUM16, MNC17.19, MNC17_19, PUM16, pumilio 16	-	163	SNP	G(GGT)->R(CGT)	NSM	CDSHIT	C	G	7
3	11540933	AT3G29690	-	127	SNP	G(GGG)->E(GAG)	NSM	CDSHIT	C	T	6
1	23445234	F16M19.22, F16M19_22	+	516	SNP	G(GGA)->V(GTA)	NSM	CDSHIT	G	T	4
2	13462172	ATX1, SDG27, SET DOMAIN PROTEIN 27, T9H9.17, T9H9_17, homologue of trithorax	-	4	SNP	F(TTT)->V(GTT)	NSM	CDSHIT	A	C	12
3	11474203	AT3G29638	+	112	SNP	D(GAT)->G(GGT)	NSM	CDSHIT	A	G	8
1	30037515	F19K16.20, F19K16_20, GL2, GLABRA 2, HOMEOBOX PROTEIN GLABRA 2	+	26	SNP	D(GAC)->G(GGC)	NSM	CDSHIT	A	G	10
1	3764105	T28P6.11, T28P6_11	-	120	SNP	D(GAC)->E(GAA)	NSM	CDSHIT	G	T	6
3	11048091	AT3G29060	-	125	SNP	D(GAC)->E(GAA)	NSM	CDSHIT	G	T	2
1	12539668	ATPS1, PARALLEL SPINDLE 1, PS1	+	930	SNP	D(GAC)->A(GCC)	NSM	CDSHIT	A	C	9
3	16948733	AT3G46140	+	215	SNP	C(TGC)->Y(TAC)	NSM	CDSHIT	G	A	11
2	15752431	F13M22.4, F13M22_4	-	194	SNP	A(GCT)->V(GTT)	NSM	CDSHIT	G	A	2
4	9718808	DL4740W, FCAALL.426	+	379	SNP	A(GCT)->V(GTT)	NSM	CDSHIT	C	T	10
5	23811769	K19M22.22, K19M22_22	-	628	SNP	A(GCT)->T(ACT)	NSM	CDSHIT	C	T	8
5	16354658	MHK7.6, MHK7_6	-	400	SNP	A(GCT)->T(ACT)	NSM	CDSHIT	C	T	8
3	5097337	AT3G15130	-	629	SNP	A(GCA)->V(GTA)	NSM	CDSHIT	G	A	7