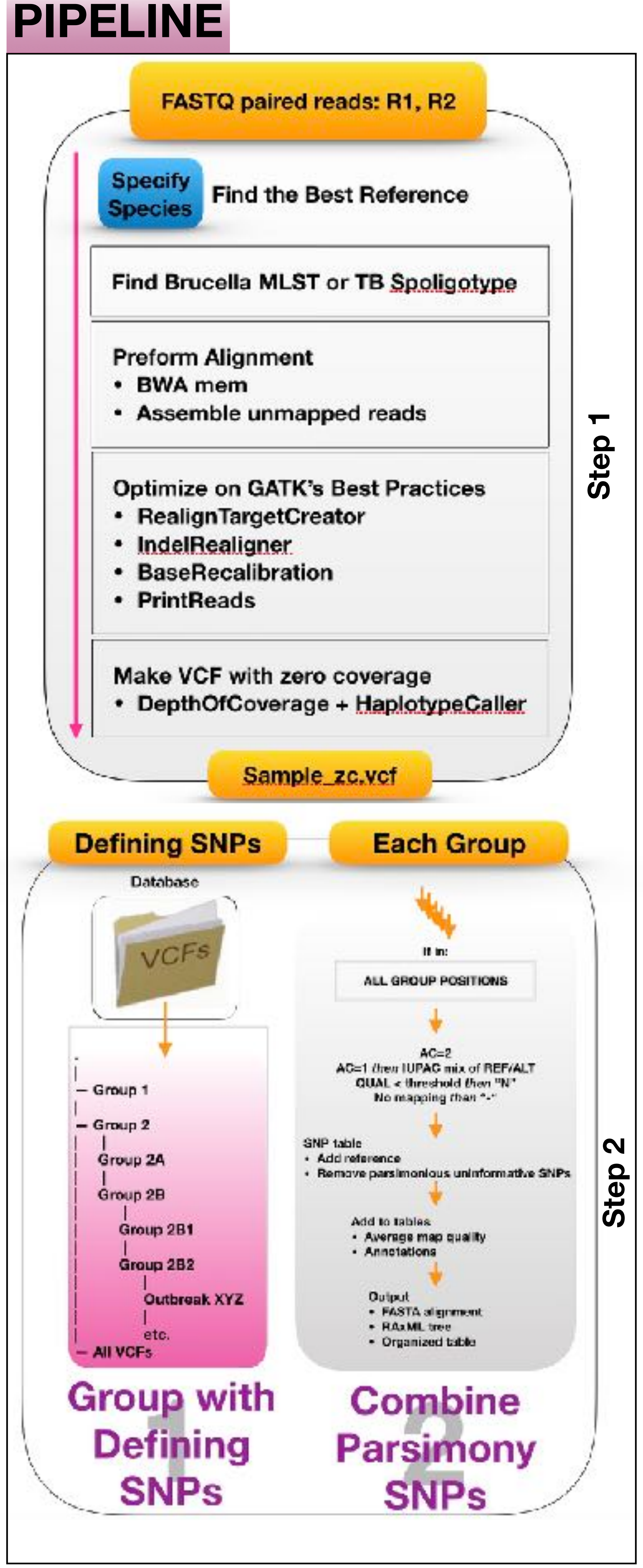
vSNP: Bacterial validation SNP genotyping tool utilizing Mycobacterium and Brucella species

Barcode

Tod Stuber[a], Suelee Robbe-Austerman[a]

[a]United States Department of Agriculture (USDA), Animal & Plant Health Services (APHIS), National Veterinary Services Laboratories (NVSL), Diagnostic Bacteriology Laboratory, Ames Iowa, USA. [b]USDA, APHIS, NVSL, Diagnostic Virology Laboratory, Ames Iowa, USA.
Corresponding authors (515) 337-7388. (tod.p.stuber@usda.gov, suelee.robbe-austerman@aphis.usda.gov)

PIPELINE



Overview

Single nucleotide polymorphism (SNP) analysis of high-throughput sequencing (HTS) data is increasingly the preferred method to genotype bacterial outbreaks. A specific tool was needed to rapidly call, validate and compare SNPs from FASTQ files in a timely manner while utilizing large datasets. vSNP was developed to address these challenges. It is particular well suited for outbreaks or clonal organisms such as Mycobacterium tuberculosis and Brucella species.

vSNP is publicly available at: https://usda-vs.github.io/snp_analysis/

Method

vSNP runs on macOS and Linux systems. It is written in Python 3, utilizing Python packages and system programs available from Anaconda package manager. The program is run from the command-line, requiring minimal experience to setup and execute.

vSNP provides a high-resolution SNP analysis from Illumina pair FASTQ files. vSNP is implemented in a two-step process. Step 1 chooses the best reference, and outputs alignments and SNP calls into VCF files. Step 2 takes in a collection of VCF files and outputs SNP alignments and phylogenetic trees. Predetermined SNP positions are used to create groups. Grouping samples on defining SNPs provides manageable datasets and allows results to be quickly and thoroughly validated. Poor SNP positions are easily targeted for further visual validation using Excel formatted SNP alignment tables. Phylogenetic trees are created from validated tables. SNP tables provide the average map quality and annotation at each position.

Conclusion

To achieve high resolution genotyping, it is necessary to validate SNPs within an outbreak. This must be accomplished quickly to minimize personnel time and to not delay actionable results. We have found it best to validate using SNP tables with fewer than 50 SNP positions. Defining SNPs are used to group similar samples into smaller, more manageable groups. This use of a defining SNP to group closely related isolates is a key element that differentiates vSNP from other pipelines. It allows both a perspective against other isolates and allows the validation necessary to make confident, actionable decisions.

Other clonal organisms, or bacteria or even viruses within an outbreak situation, can also be incorporated into vSNP to provide the same high-level resolution.

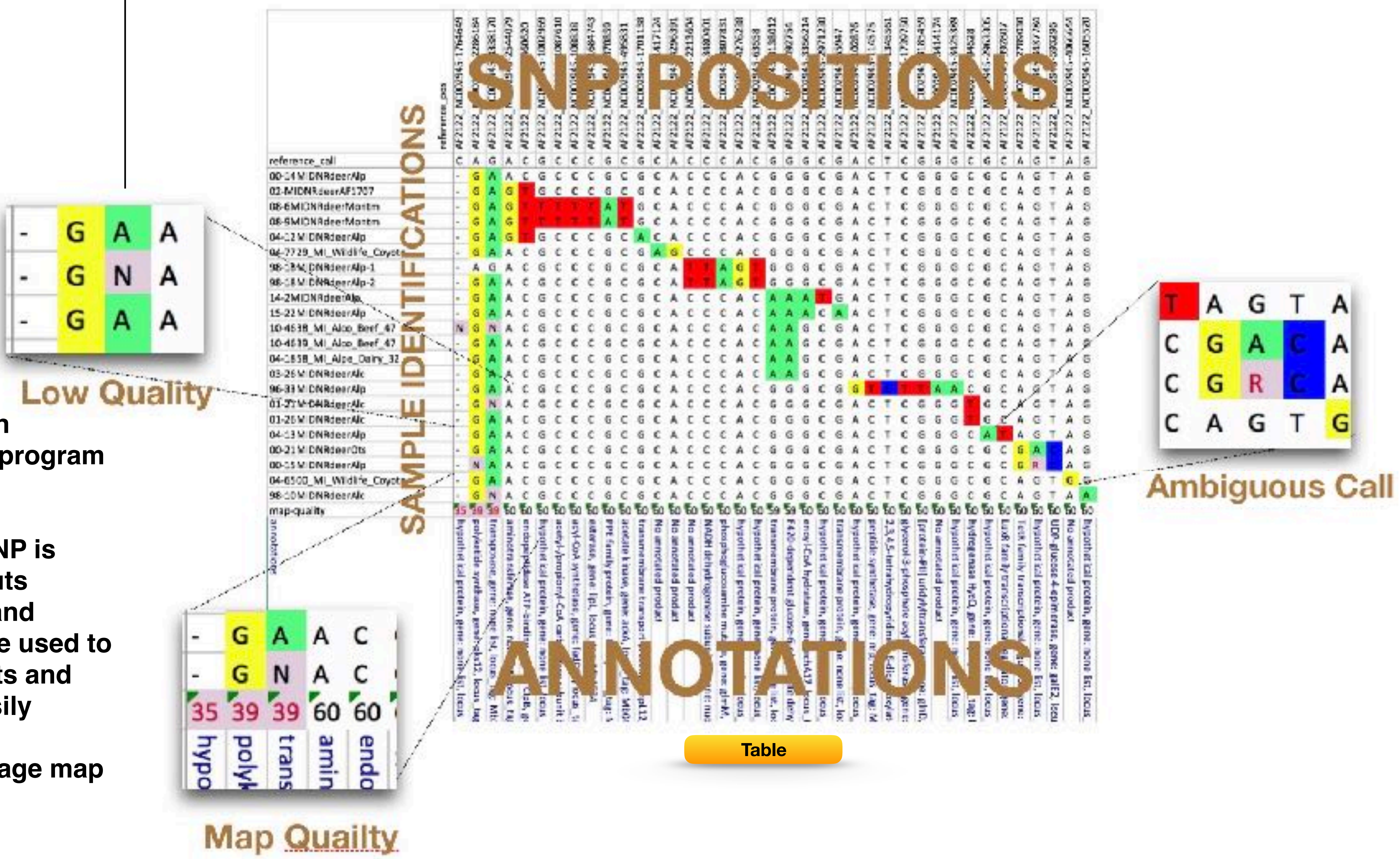
Pros:

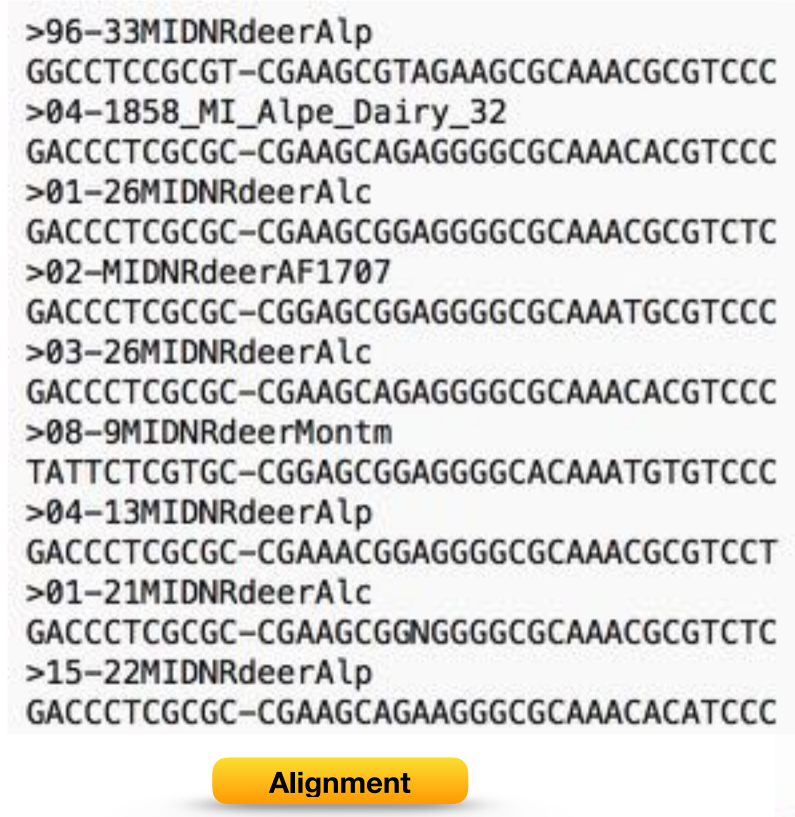
- Proven pipeline workflow
- Maintain sense of data ownership
- High resolution, validated genotyping

Cons:

- Command-line drive
- Unable to run open source tools (GATK) in Windows

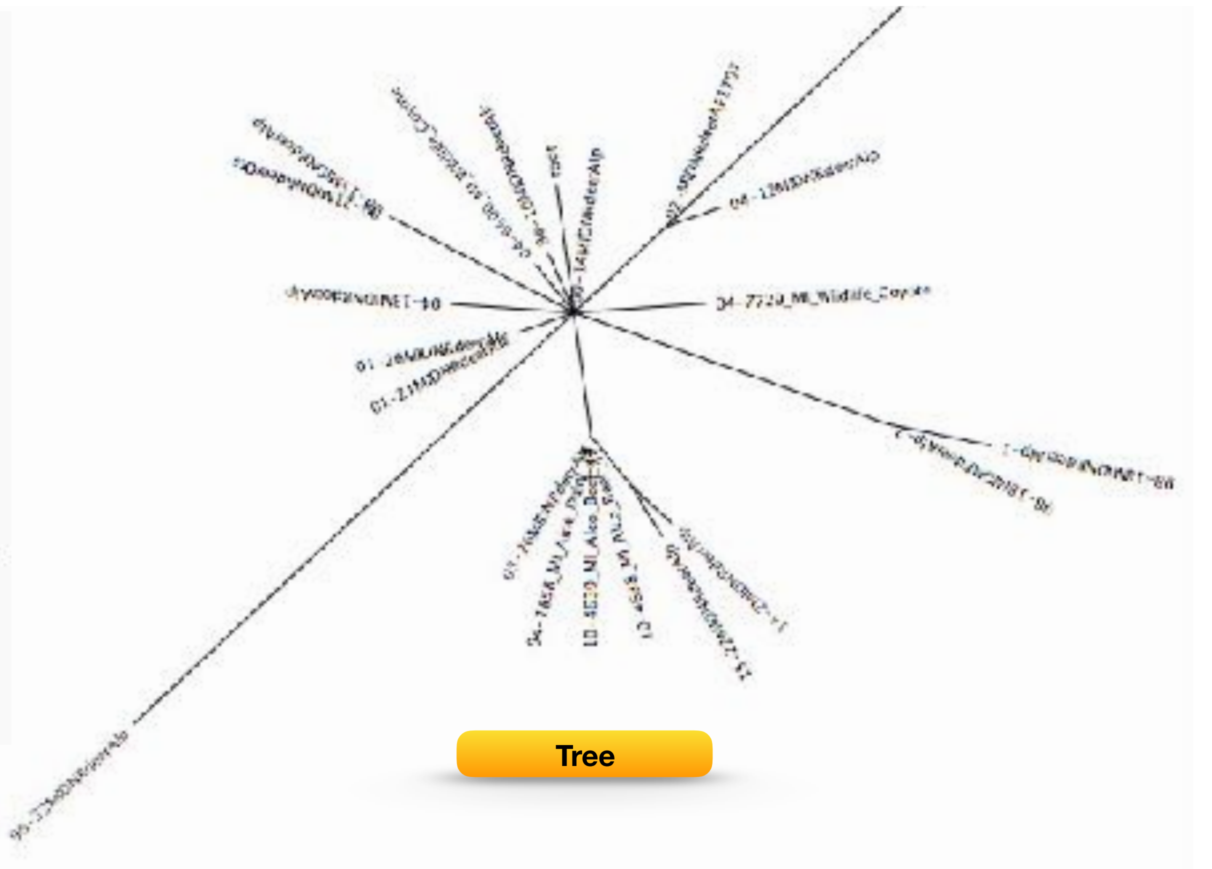
Output



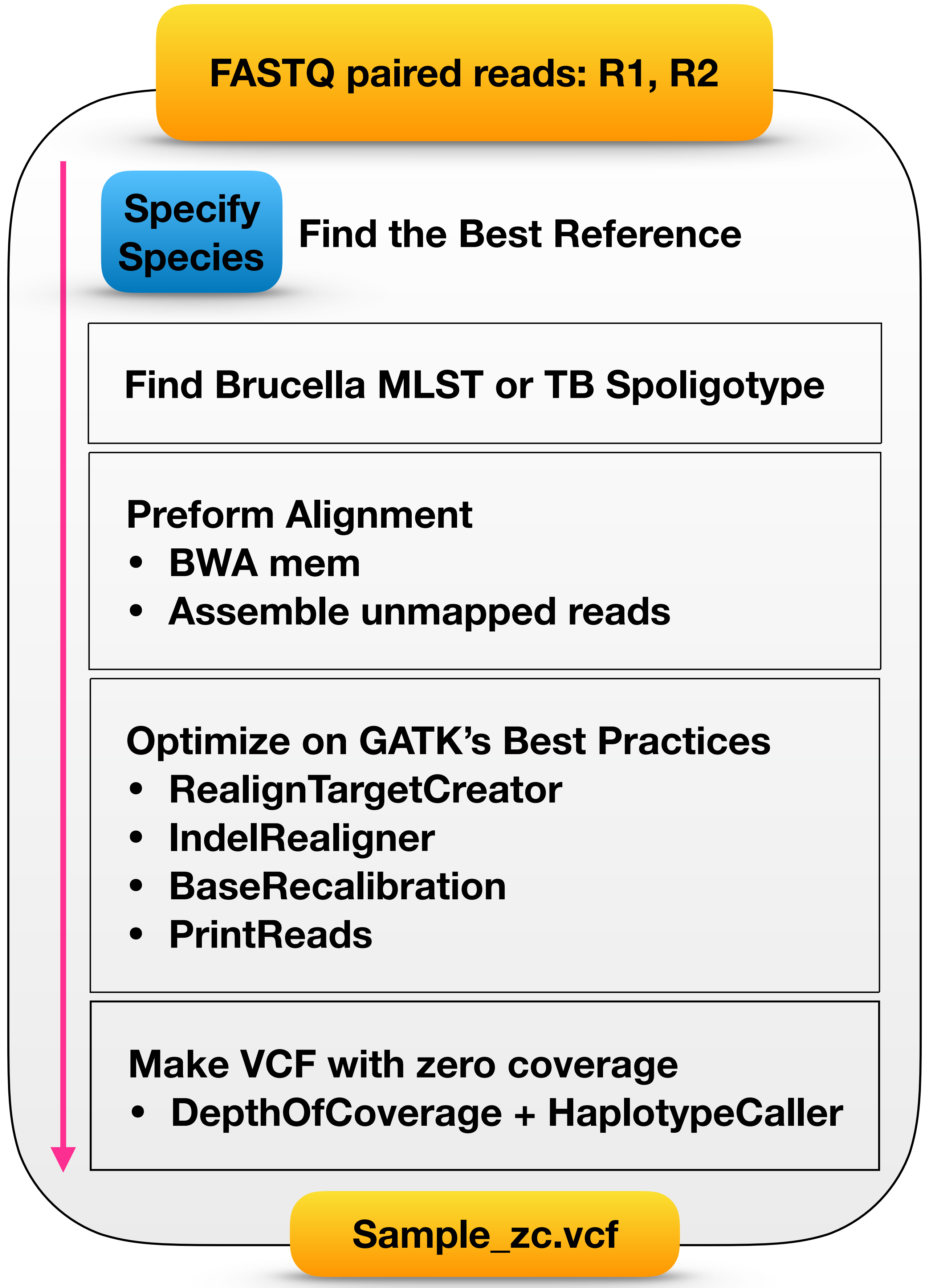


```
>96-33MIDNRdeerAlp
GGCCTCCGCGT-CGAAGCGTAGAAGCGCAAACGCGTCCC
>04-1858_MI_Alpe_Dairy_32
GACCCTCGCGC-CGAAGCAGAGGGGCGCAAACACGTCCC
>01-26MIDNRdeerAlc
GACCCTCGCGC-CGAAGCGGAGGGGCGCAAACGCGTCTC
>02-MIDNRdeerAF1707
GACCCTCGCGC-CGGAGCGGAGGGGCGCAAATGCGTCCC
>03-26MIDNRdeerAlc
GACCCTCGCGC-CGAAGCAGAGGGGCGCAAACACGTCCC
>08-9MIDNRdeerMontm
TATTCTCGTGC-CGGAGCGGAGGGGCACAAATGTGTCCC
>04-13MIDNRdeerAlp
GACCCTCGCGC-CGAAACGGAGGGGCGCAAACGCGTCCT
>01-21MIDNRdeerAlc
GACCCTCGCGC-CGAAGCGGNGGGGCGCAAACGCGTCTC
>15-22MIDNRdeerAlp
GACCCTCGCGC-CGAAGCAGAAGGGCGCAAACACATCCC
```

Alignment



Mbovis-01D2



FASTQ paired reads: R1, R2

Specify Species

Find the Best Reference

Find Brucella MLST or TB Spoligotype

Preform Alignment

- BWA mem
- Assemble unmapped reads

Optimize on GATK's Best Practices

- RealignTargetCreator
- IndelRealigner
- BaseRecalibration
- PrintReads

Make VCF with zero coverage

- DepthOfCoverage + HaplotypeCaller

Sample_zc.vcf

FASTQ paired reads: R1, R2

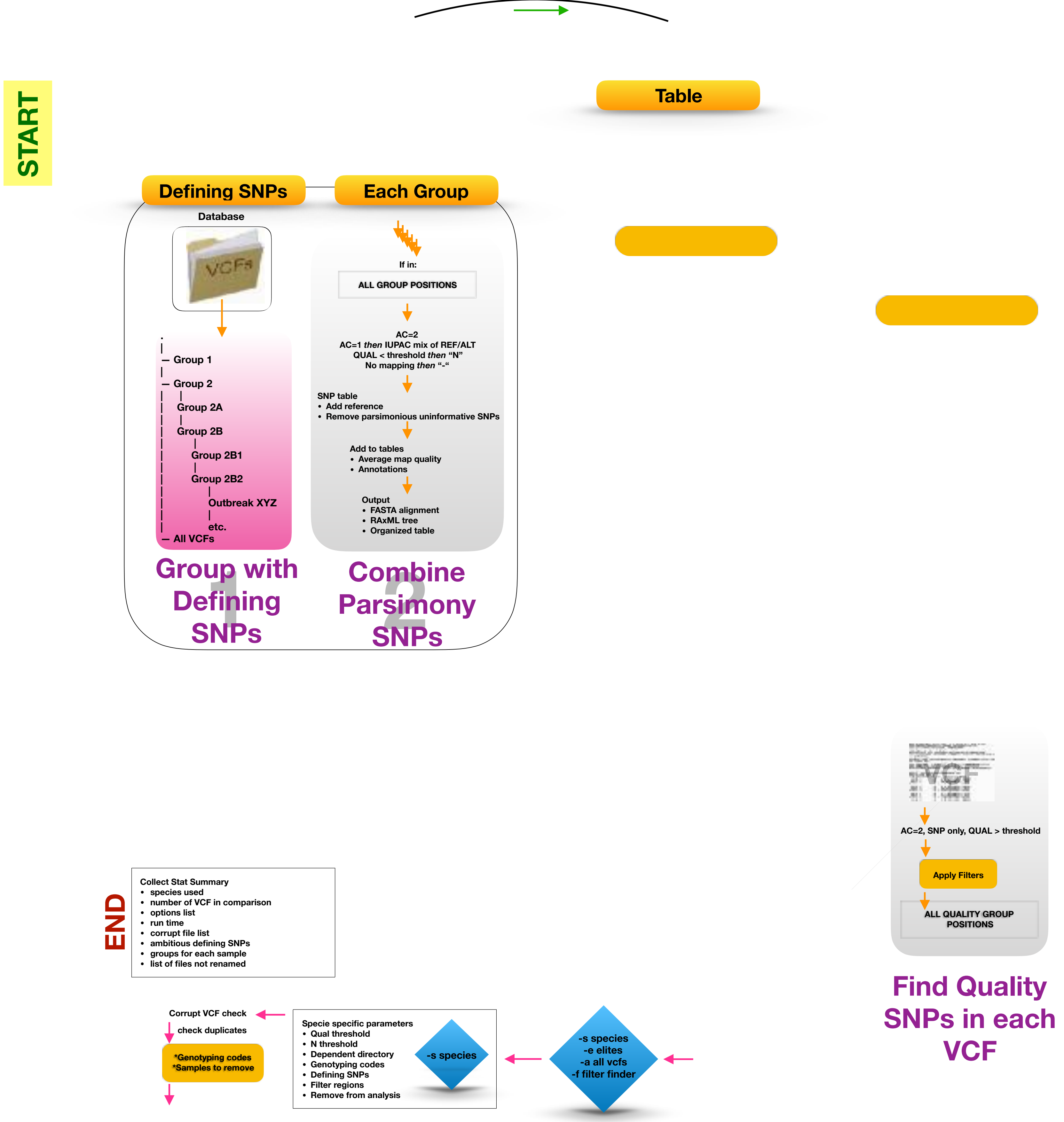
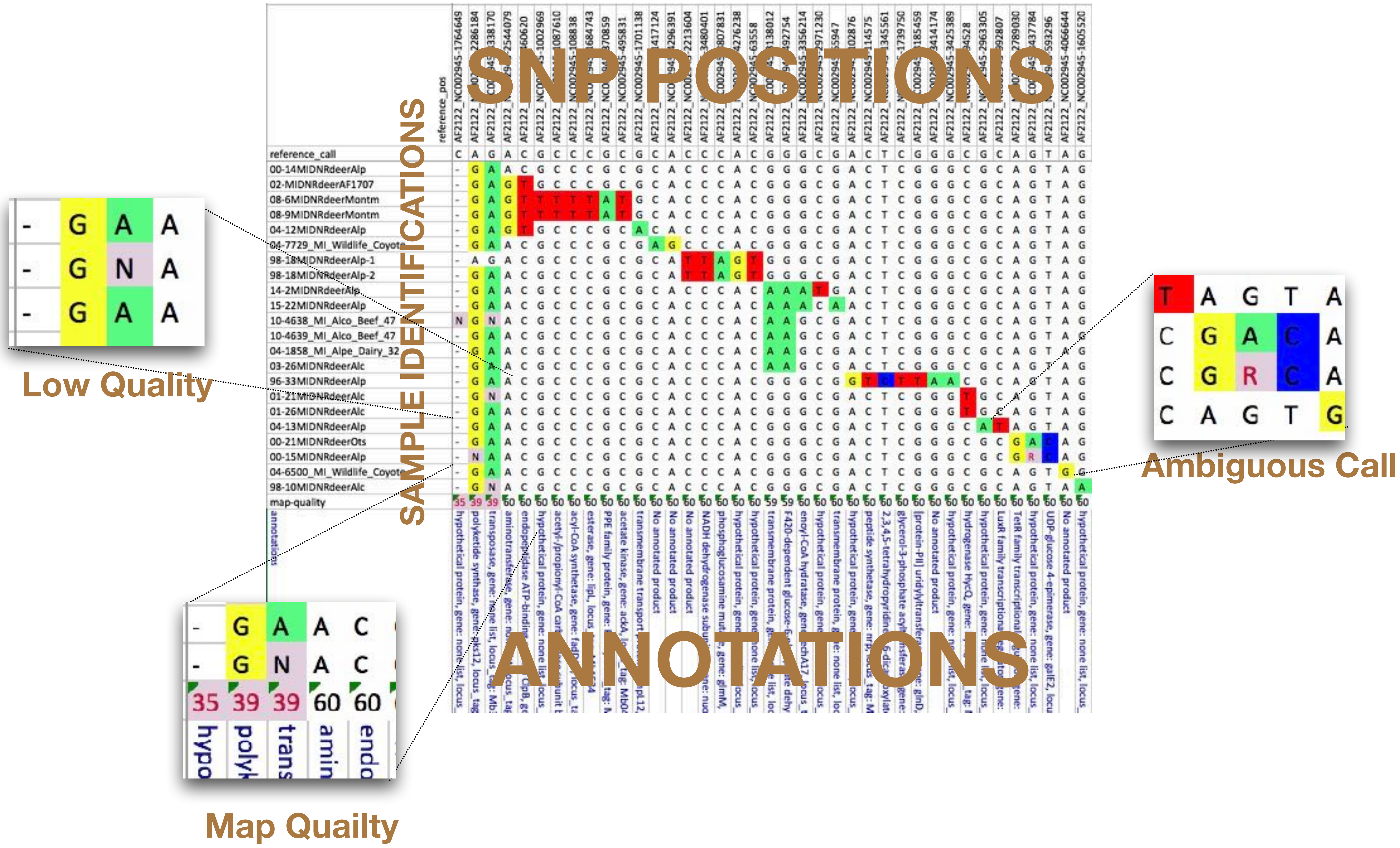


Figure 2: Script 2 workflow — blue diamonds represent script options, yellow boxes represent input from dependency files. Those starred are optional.

Mbovis-01D2



SNP POSITIONS

SAMPLE IDENTIFICATIONS

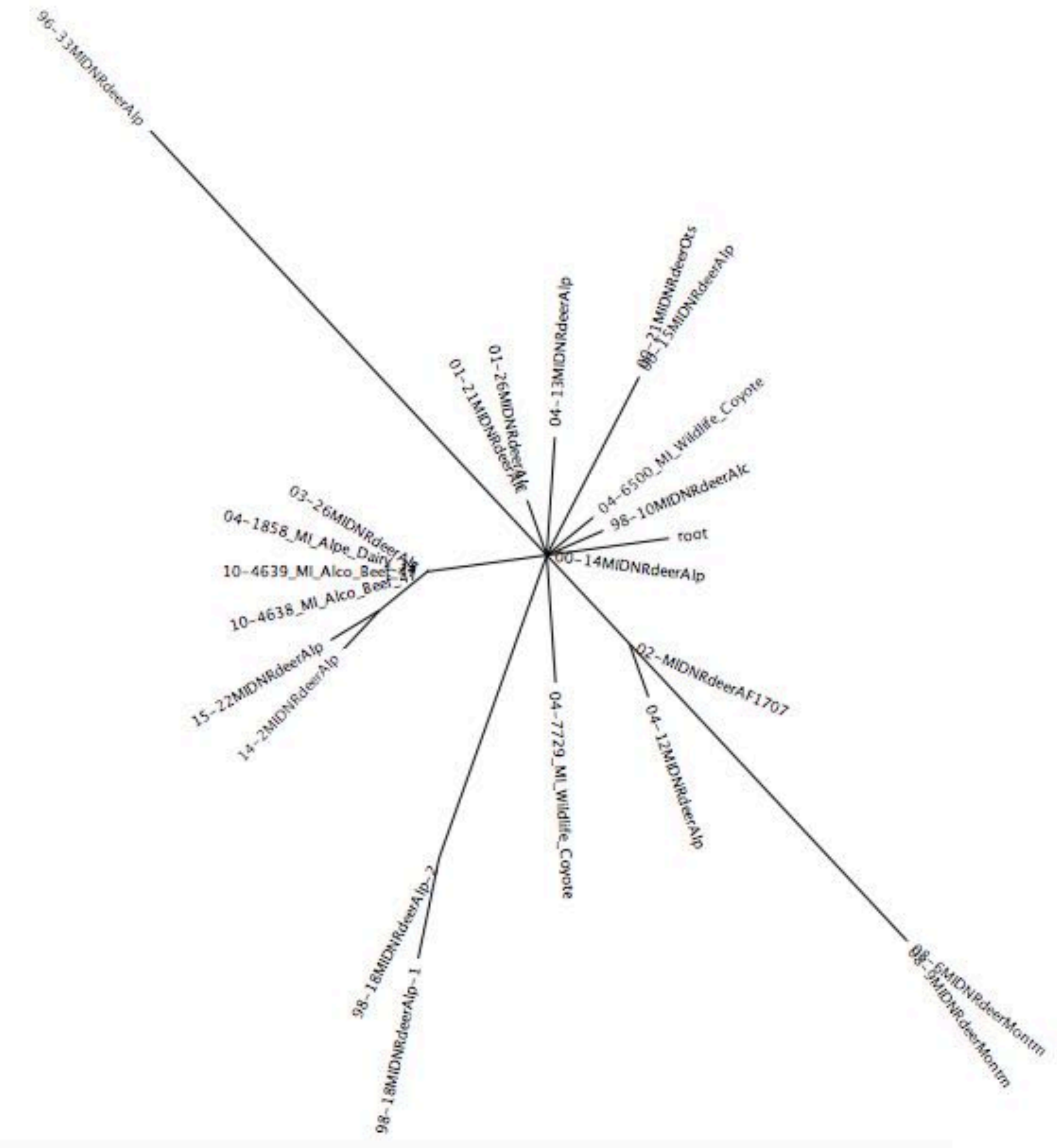
ANNOTATIONS

Low Quality

Ambiguous Call

Map Quailty

reference_pos	AF2122_NC002945-1764649	AF2122_NC002945-2286184	AF2122_NC002945-3338170	AF2122_NC002945-2544079	AF2122_NC002945-460620	AF2122_NC002945-1002969	AF2122_NC002945-1087610	AF2122_NC002945-108838	AF2122_NC002945-1684743	AF2122_NC002945-370859	AF2122_NC002945-495831	AF2122_NC002945-1701138	AF2122_NC002945-4117124	AF2122_NC002945-4296391	AF2122_NC002945-2213604	AF2122_NC002945-3480401	AF2122_NC002945-1807831	AF2122_NC002945-4276238	AF2122_NC002945-63558	AF2122_NC002945-1138012	AF2122_NC002945-492754	AF2122_NC002945-3356214	AF2122_NC002945-2971230	AF2122_NC002945-55947	AF2122_NC002945-102876	AF2122_NC002945-114575	AF2122_NC002945-1345561	AF2122_NC002945-1739750	AF2122_NC002945-1185459	AF2122_NC002945-3414174	AF2122_NC002945-3425389	AF2122_NC002945-94528	AF2122_NC002945-2963305	AF2122_NC002945-992807	AF2122_NC002945-2789030	AF2122_NC002945-1437784	AF2122_NC002945-593296	AF2122_NC002945-4066644	AF2122_NC002945-1605520
reference_call	C	A	G	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
00-14MIDNRdeerAlp	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
02-MIDNRdeerAF1707	-	G	A	G	T	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
08-6MIDNRdeerMontm	-	G	A	G	T	T	T	T	T	A	T	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
08-9MIDNRdeerMontm	-	G	A	G	T	T	T	T	T	A	T	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
04-12MIDNRdeerAlp	-	G	A	G	T	G	C	C	C	G	C	A	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
04-7729_MI_Wildlife_Coyote	-	G	A	A	C	G	C	C	C	G	C	G	A	G	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
98-18MIDNRdeerAlp-1	-	A	G	A	C	G	C	C	C	G	C	G	C	A	T	T	A	G	T	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
98-18MIDNRdeerAlp-2	-	G	A	A	C	G	C	C	C	G	C	G	C	A	T	T	A	G	T	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
14-2MIDNRdeerAlp	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	A	A	A	T	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
15-22MIDNRdeerAlp	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	A	A	A	C	A	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
10-4638_MI_Alco_Beef_47	N	G	N	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	A	A	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
10-4639_MI_Alco_Beef_47	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	A	A	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
04-1858_MI_Alpe_Dairy_32	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	A	A	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
03-26MIDNRdeerAlc	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	A	A	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	G
96-33MIDNRdeerAlp	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	G	T	C	T	T	A	A	C	G	C	A	G	T	A	G
01-21MIDNRdeerAlc	-	G	N	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	T	G	C	A	G	T	A	G
01-26MIDNRdeerAlc	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	T	G	C	A	G	T	A	G
04-13MIDNRdeerAlp	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	A	T	A	G	T	A	G
00-21MIDNRdeerOts	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	G	A	C	A	G
00-15MIDNRdeerAlp	-	N	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	G	R	C	A	G
04-6500_MI_Wildlife_Coyote	-	G	A	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	G	G
98-10MIDNRdeerAlc	-	G	N	A	C	G	C	C	C	G	C	G	C	A	C	C	C	A	C	G	G	G	C	G	A	C	T	C	G	G	G	C	G	C	A	G	T	A	A
map-quality	35	39	39	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	59	59	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60	60
annotations	hypothetical protein, gene: none list, locus_	polyketide synthase, gene: pks12, locus_tag	transposase, gene: none list, locus_tag: Mb	aminotransferase, gene: no locus_tag	endopeptidase ATP-binding ClpB, ge	hypothetical protein, gene: none list, locus_	acetyl-/propionyl-CoA carb subunit b	acyl-CoA synthetase, gene: fadD , locus_ta	esterase, gene: lipL, locus_ 4	PPE family protein, gene: _tag: M	acetate kinase, gene: ackA, lo tag: Mb0	transmembrane transport pl12,	No annotated product	No annotated product	No annotated product	NADH dehydrogenase subunit gene: nuo	phosphoglucosamine mutase, gene: glmM,	hypothetical protein, gene locus_	hypothetical protein, gene: none list, locus_	transmembrane protein, ge e list, loc	F420-dependent glucose-6- ate dehy	enoyl-CoA hydratase, gene: echA17, locus_t	hypothetical protein, gene locus_	transmembrane protein, gene: none list, loc	hypothetical protein, gene locus_	peptide synthetase, gene: nrp, locus_tag: M	2,3,4,5-tetrahydropyridine-2,6-dicarboxylat	glycerol-3-phosphate acyltransferase, gene:	[protein-PII] uridylyltransferase, gene: glnD,	No annotated product	hypothetical protein, gene: none list, locus_	hydrogenase HycQ, gene: tag: M	hypothetical protein, gene: none list, locus_	LuxR family transcriptional regulator, gene:	TetR family transcriptional gene:	hypothetical protein, gene: none list, locus_	UDP-glucose 4-epimerase, gene: galE2, locu	No annotated product	hypothetical protein, gene: none list, locus_


96-33MIDNRdeerAlp
01-26MIDNRdeerAlc
01-21MIDNRdeerAlc
04-13MIDNRdeerAlp
00-21MIDNRdeerOts
98-15MIDNRdeerAlp
04-6500_MI_Wildlife_Coyote
98-10MIDNRdeerAlc
root
00-14MIDNRdeerAlp
03-26MIDNRdeerAlp
04-1858_MI_Alpe_Dairy
10-4639_MI_Alco_Beef
10-4638_MI_Alco_Beef
15-22MIDNRdeerAlp
14-2MIDNRdeerAlp
02-MIDNRdeerAF1707
04-12MIDNRdeerAlp
04-7729_MI_Wildlife_Coyote
98-18MIDNRdeerAlp-2
98-18MIDNRdeerAlp-1
98-6MIDNRdeerMontm
98-9MIDNRdeerMontm