Galaxy / vSNP

Greg consulted with Dr. Vivek Kapur in the Huck Institutes of the Life Sciences to build the Galaxy / vSNP analysis environment. Whole genome sequencing for disease tracing and outbreak investigations is routinely required for high-consequence disease. The Galaxy / vSNP tools are critical components of several Galaxy workflows that locate and validate SNPs in bacterial samples and produce annotated SNP tables and corresponding phylogenetic trees. The tools are capable of processing very large sets of inputs and efficiently accommodate multiple genome references for mapping samples.

The Galaxy / vSNP toolset consists of the following tools which are components of larger Galaxy workflows.

- **Determine reference from data** - sniffs bacterial genome samples to discover the primary species within the sample
- **Add zero coverage** - adds zero coverage to a VCF file to prepare it for processing by the Get SNPs tool
- **Statistics** - produces an Excel spreadsheet containing statistics for samples and associated metrics files
- **Get SNPs** - collects quality parsimonious SNPs from VCF files and outputs alignment files in FASTA format
- **Build tables** - produces annotated SNP tables in the form of Ecel spreadsheets from outputs produced by the Get SNPs tool

The Galaxy / vSNP environment provides some very interesting tools and workflows. Bacterial samples often consist of a “mixed bag” of species, so it's not always clear to the researcher which is the prominent species in the sample. This makes selecting the correct reference genome for mapping difficult or impossible.

The Galaxy vSNP: determine reference from data tool inspects the bacterial sample data and looks for what we call “dnaprints" that match with the same dnaprints in one of many reference datasets installed into the Galaxy / vSNP environment. The reference dataset containing the same dnaprints will be selected for mapping. The tool outputs a dataset consisting of only the Galaxy dbkey string (metrics and other information is available in a separate output), and this dataset is used as input to one of Galaxy’s workflow parameter tools which can be used to specify the selection of the dynamic select list that the Galaxy mapping tools use for reference genome selection.

Figure 7 shows the workflow that performs this DNA sniffing, reference selection and mapping process.

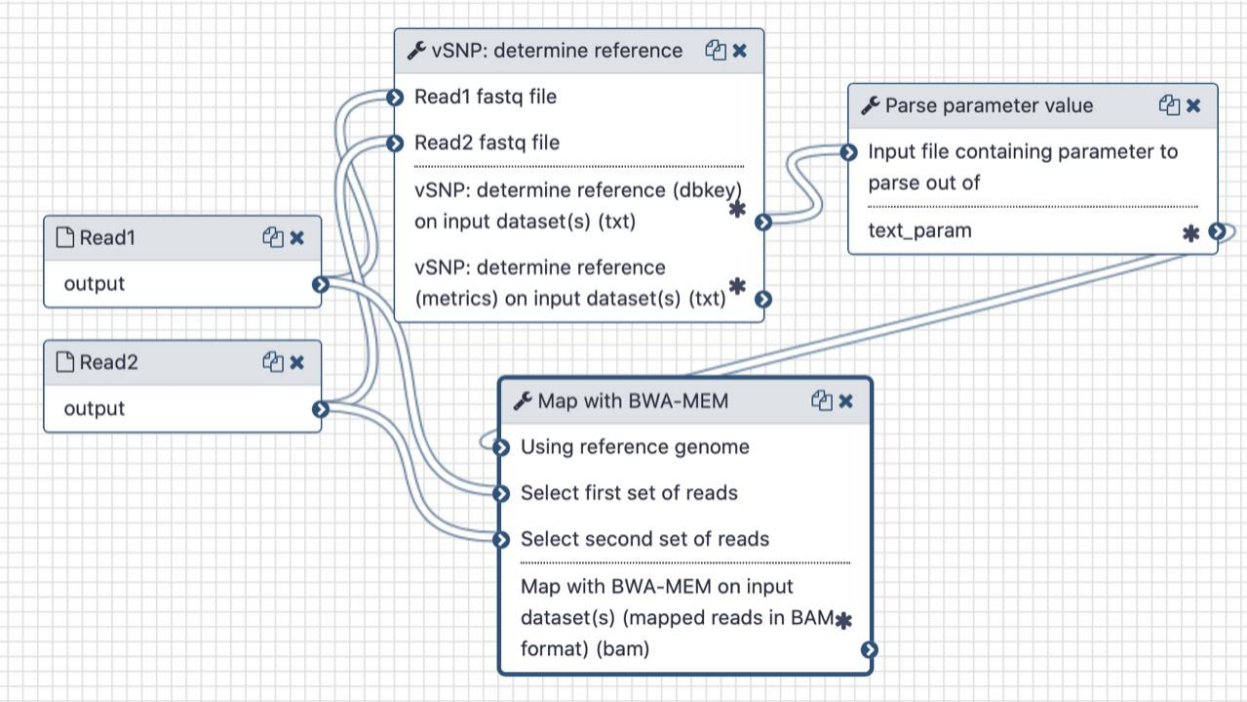

Figure 7: vSNP: determine reference from data, the first DNA sniffer tool for Galaxy

Figure 8 displays an example of an annotated SNP table produced by the vSNP: build tables tool. The tool accepts a combination of single SNPs JSON, average MQ JSON and NEWICK files (or associated collections of each) to produce these annotated SNP tables. The SNPs JSON and average MQ JSON files are typically produced by the Galaxy vSNP: get SNPs tool and the NEWICK files are typically produced by the Galaxy Phyogenetic reconstruction with RaXML tool.

The SNP tables display closely related isolates and enable identification of mixed SNPs when multiple bacterial strains are infecting an organism. The table structure is shown in Figure 8. The columns identify the genome location of the SNP calls and the isolates are contained within the rows. The reference (or ancestral strain if the reference is an outgroup) is listed across the top, identified as the "reference call". SNPs that are not highlighted will match the reference. The map-quality row values are the average of the map quality scores of each isolate in that position. These scores measure the confidence that the read has been mapped to the correct

location on the genome. The maximum score possible is 60, and lower scores lessen the
confidence that the SNP was correctly identified. The annotation of the position is provided at the bottom of the table.

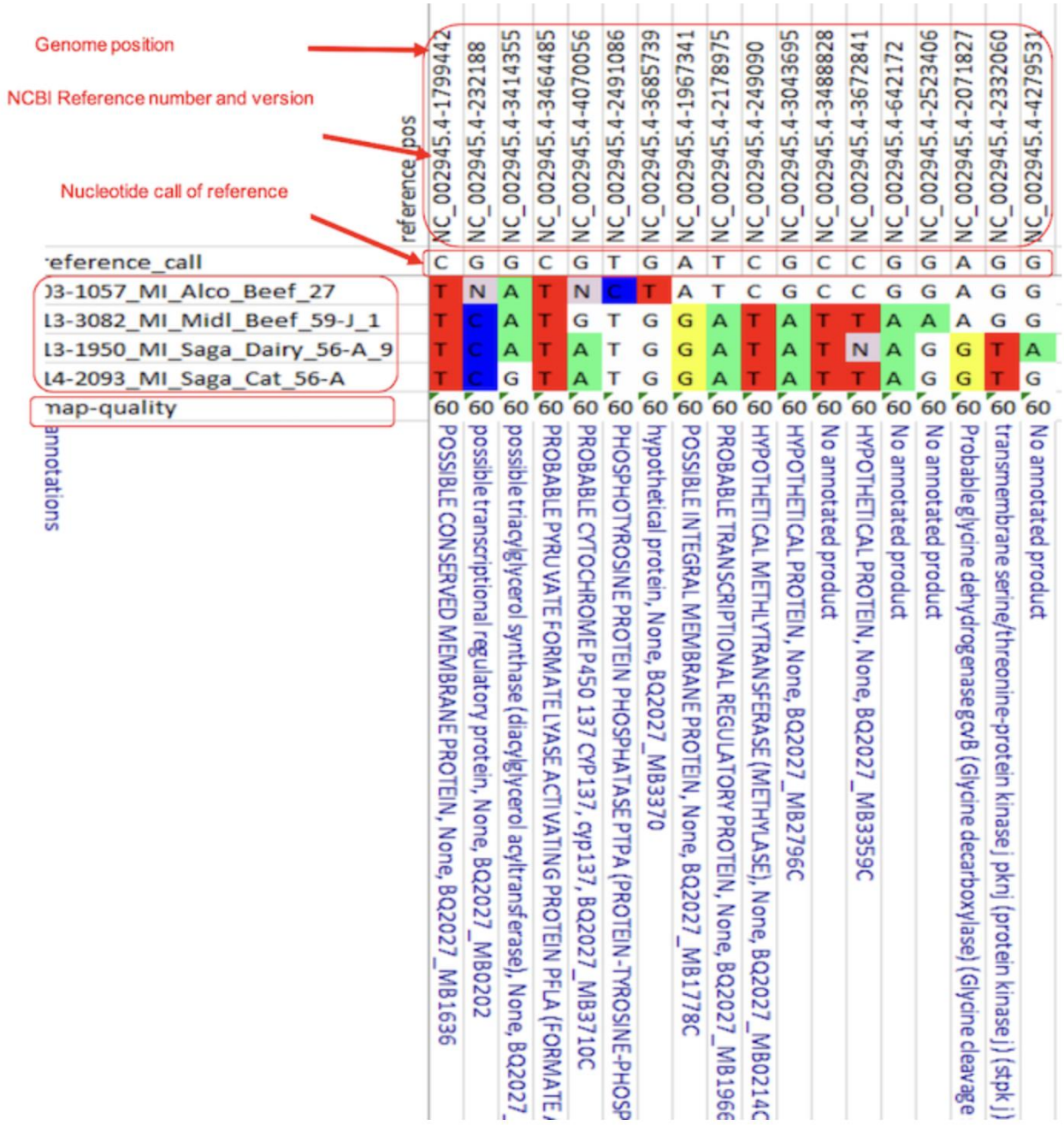

Figure 8: vSNP: build tables produces annotated SNP tables in the form of Excel spreadsheets

Figure 9 displays an example of a phylogenetic tree that would be associated with a SNP table similar3 to the one shown in Figure 8.

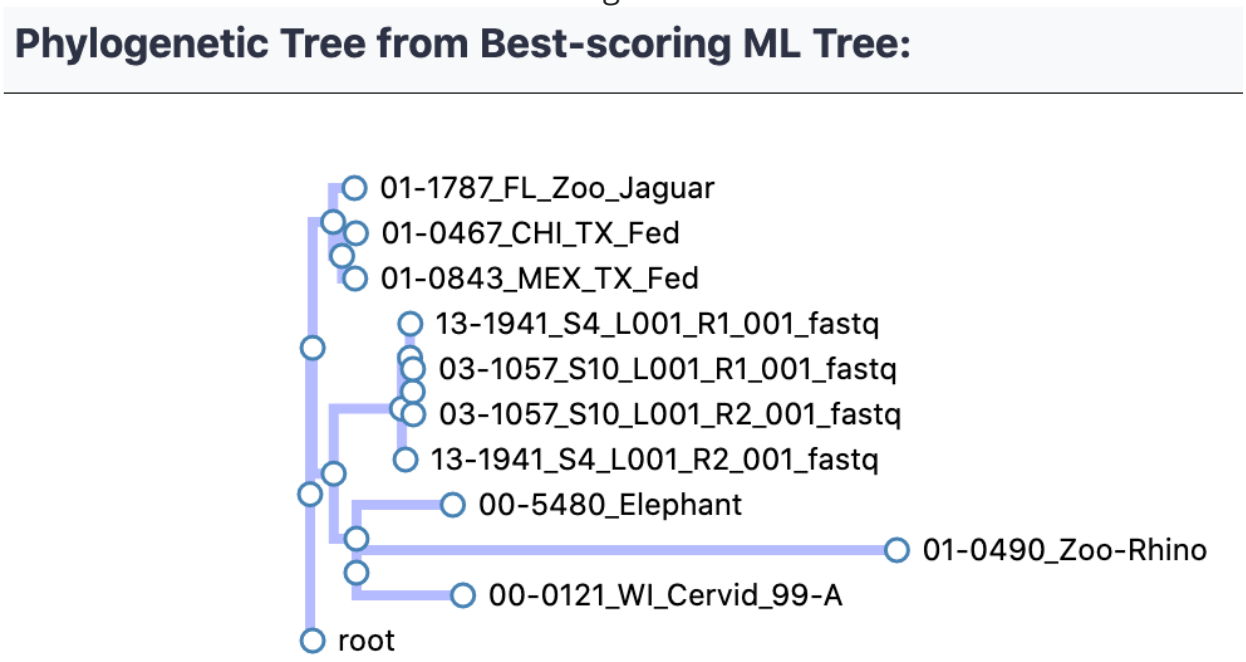

Figure 9: Phylogenetic tree produced by the Galaxy Phyogenetic reconstruction with RaXML tool