MOCA - Multi-Omics Concordance Analysis

Jared Huzar

MOCA tools were developed to explore the concordance between genetic and expression evolution. MOCA provides functions which identify genetic ancestries from phylogenetic trees, perform gene expression trajectory analysis followed by a summary of concordance between expression states and genetic ancestry, and assess the robustness of genetic ancestry annotations.

Installation and Dependencies

MOCA tools can be downloaded as a zip file with each function contained within an individual .R file. MOCA requires several packages to be installed as dependencies: ape, phytools, ggtree, stringr, ggplot2, apTreeshape, dplyr, monocle.

This can be accomplished using the following commands.

```
MOCA_packages <- c("ape", "phytools", "stringr", "ggplot2", "apTreeshape",
    "dplyr")

install.packages(MOCA_packages)

if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")

BiocManager::install(c("ggtree", "monocle"), force = TRUE)

MOCA_packages <- append(MOCA_packages, c("monocle", "ggtree"))

lapply(MOCA_packages, library, character.only = TRUE)

setwd("Path To MOCA Main Directory (MOCA-master)")

MOCA_Files <- list.files("MOCA/", pattern = ".R")

for (a in 1:length(MOCA_Files)) {
    source(paste0("MOCA/", MOCA_Files[a]))
}


# R v3.6.2 was tested
```

Example Workflow

Genetic Ancestry Annotation Tools

MOCA consists of three primary sections of tools, the first of which is the genetic ancestry annotation tools.

As input this module first requires a newick tree file. This section consists of three tools: *TreeBalance*, *BalancedAnnotation*, and *UnbalancedAnnotation*.

TreeBalance determines if the input newick tree is of the PDA (unbalanced) model or the Yule (uniform/balanced) model. It will return either "Balanced" or "Unbalanced".

```
tree <- read.tree("MOCA/Example/tree1.nwk")
balanced_tree_res <- TreeBalance(tree)
```

Based on the suggestion of *TreeBalance*, users can select to use the *BalancedAnnotation* function for genetic ancestry annotation, or *UnbalancedAnnotation* function. The *BalancedAnnotation* tool is recommended for trees which *TreeBalance* classifies as balanced, and the *UnbalancedAnnotation* function is recommended for trees which *TreeBalance* function classifies as unbalanced.

As input, *BalancedAnnotation* requires a phylogenetic tree and an integer value for the number of groups to identify from the phylogeny. In addition, for *BalancedAnnotation* it requires a decimal value for percent of total cells to be in first two genetic ancestries, and a decimal value for percent of total cells required in each of first two genetic ancestries. As default, we suggest setting num_groups to 3, the minimum percent of cells combined between the two to be 0.75, and the minimum percent of cells for each to be 0.1.

For *UnbalancedAnnotation*, users are only required to input the phylogenetic tree and an integer value representing the number of genetic ancestries to identify from the tree, as default we suggest 3.

These default parameters are shown in the code chunk below.

```
if(balanced_tree_res == "Balanced"){
  annotate_file1 <- BalancedAnnotation(tree, 0.75, 0.1, 3)
}else{
  annotate_file1  <- UnbalancedAnnotation(tree, 3)
}
```

The functions will then ask if you would like to manually enter Ancestry IDs. You will be prompted to enter 1 to change IDs, or 0 to leave the IDs the same. This implements the *userDefineClades* function, which is explained in greater depth later in the manual. This functionality simply allows users to rename ancestries.

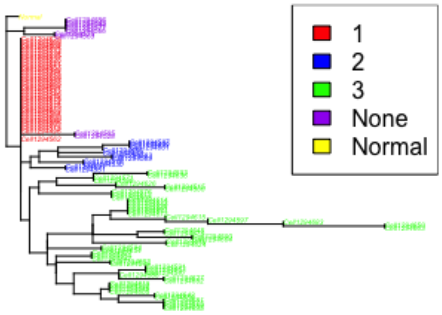
The *BalancedAnnotation* and *UnbalancedAnnotation* functions return an annotation file in your working directory, which takes the form displayed above. The cell names are in the first column, and the ancestry identifications are in the second column The row names are also the same cell names as those in the first column. This format is required for the gene expression trajectory analysis tools. The above code will also plot the tree with each clade colored uniquely.

Please inspect the annotation and adjust it if necessary by simply editing the annotation file above or using the *userDefineClades* function.

Gene Expression Trajectory Analysis

In order to summarize concordance between expression states and genetic ancestries, the file produced from *BalancedAnnotation* or *UnbalancedAnnotation* is used by the *PhyloTrajectory* function.

```
expression_matrix <- read.table("MOCA/Example/expression_matrix1.txt")

annotation_file <- read.table("MOCA/Example/annotate_file1.txt")

concordance_results<- PhyloTrajectory(expression_matrix, annotation_file, "ExampleData")
```



	Cell	GeneticAncestry
Cell1294493	Cell1294493	2
Cell1294494	Cell1294494	1
Cell1294496	Cell1294496	1
Cell1294498	Cell1294498	1
Cell1294499	Cell1294499	2
Cell1294500	Cell1294500	3

Figure 1: Output from Genetic Ancestry Annotation Tools

PhyloTrajectory takes as input an expression matrix, a genetic ancestry annotation file, and a string representing the title of the dataset simply for descriptive figures. The annotation file must be in the format described above, and the expression matrix must be a data frame with row names as gene IDs and column names as cell IDs. The cell IDs in the columns must match with the cell IDs in the genetic ancestry annotation files. This function takes no user defined parameters, but by editing the assignment in the code users can adjust the gene scale vector and the dominant ancestry cutoff value. Both parameters are in the first two lines of the function. Users familiar with Monocle can also edit the Monocle based parameters which are utilized. *PhyloTrajectory* produces the following files in your working directory.

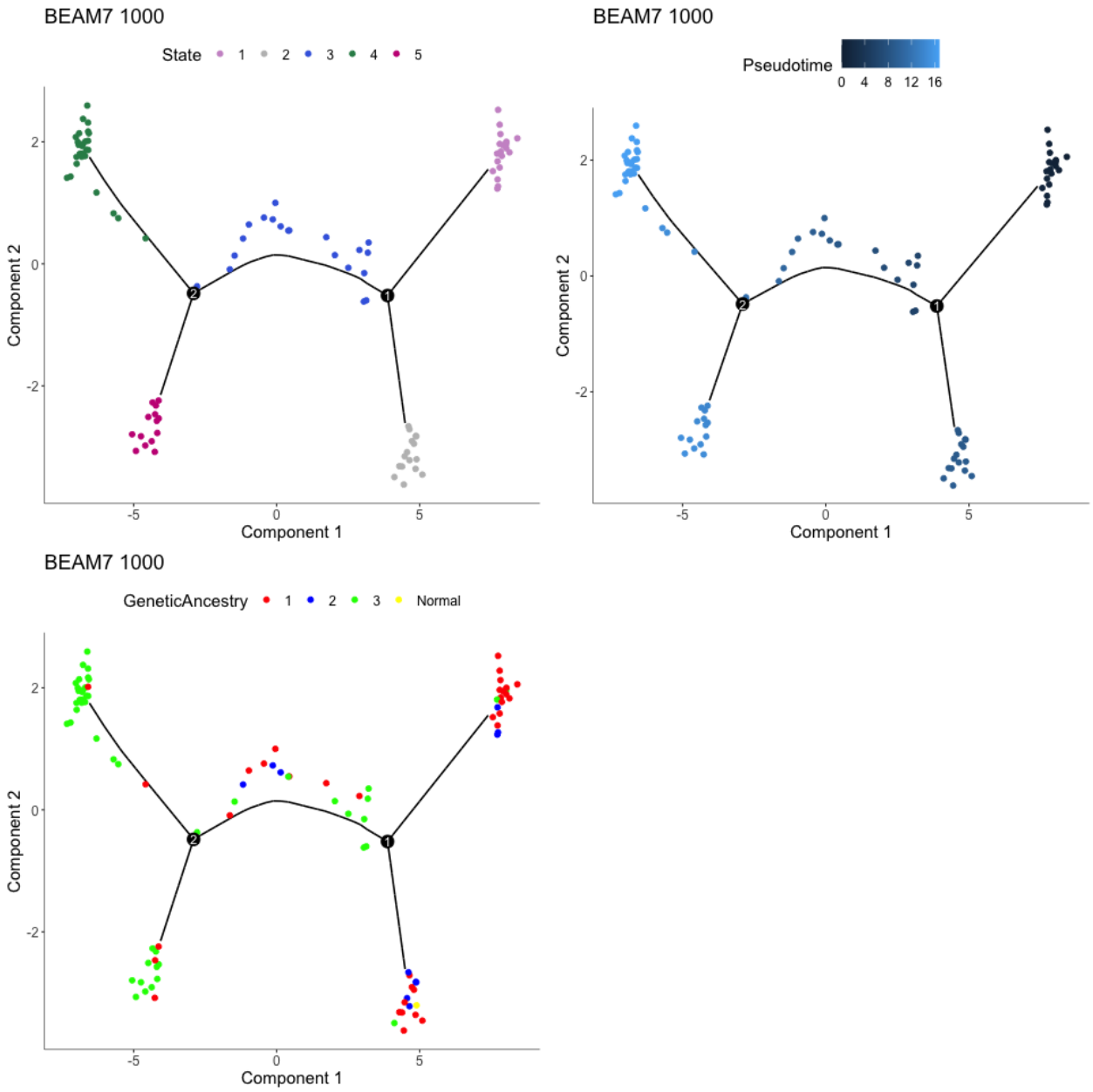
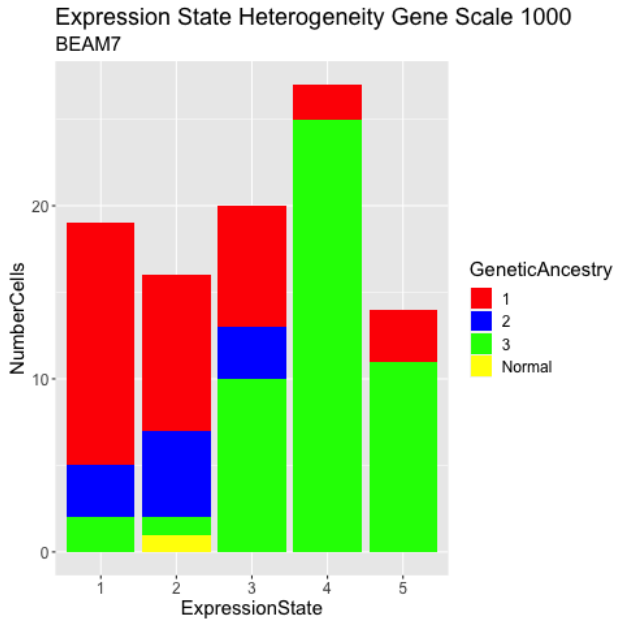


Figure 2: Gene expression trajectory plots for each scale. These are output from Monocle analysis.

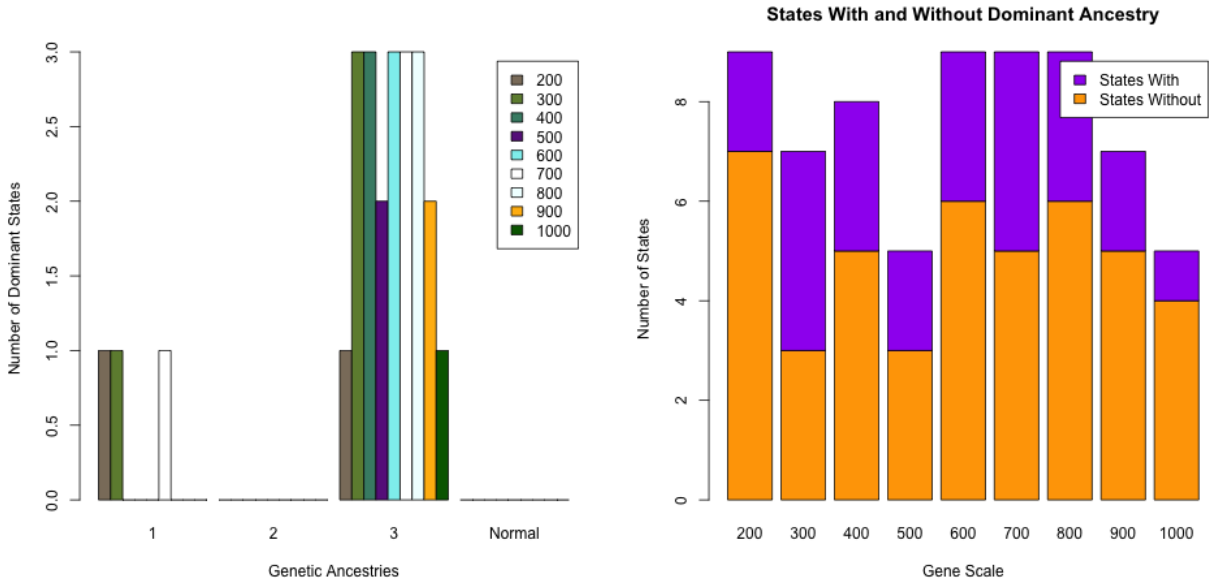
Below are the concordance summary plots and table for each gene scale. The number of cells within each expression state and each genetic ancestry is shown. The table also contains the standard deviation of pseudotime within each ancestry. The major genetic group index represents the percent of cells belonging to the ancestry which is most inhabited by the cells of that particular expression state. The major expression state index is the percent of cells belonging to the state which is most inhabited for that particular genetic ancestry.



	Ancestry.1	Ancestry.2	Ancestry.3	Normal	Major.Genetic.Group
State 1	14.000000	3.0000000	2.0000000	0	0.7368421
State 2	9.000000	5.0000000	1.0000000	1	0.5625000
State 3	7.000000	3.0000000	10.0000000	0	0.5000000
State 4	2.000000	0.0000000	25.0000000	0	0.9259259
State 5	3.000000	0.0000000	11.0000000	0	0.7857143
Major Expression State	0.400000	0.4545455	0.5102041	1	0.0000000
Sequence Type Expression Standard Deviation	5.265511	3.7682706	4.5690821	NA	0.0000000

Figure 3: Concordance summary plots and table.

PhyloTrajectory also produces figures summarizing the concordance across all gene scales.



	1	2	3	Normal
200	1	0	1	0
300	1	0	3	0
400	0	0	3	0
500	0	0	2	0
600	0	0	3	0
700	1	0	3	0
800	0	0	3	0
900	0	0	2	0
1000	0	0	1	0

Figure 4: Overall Concordance Summary Results

The table shows the total number of dominant expression states within a genetic ancestry across all gene scales.

Evaluation of Robustness

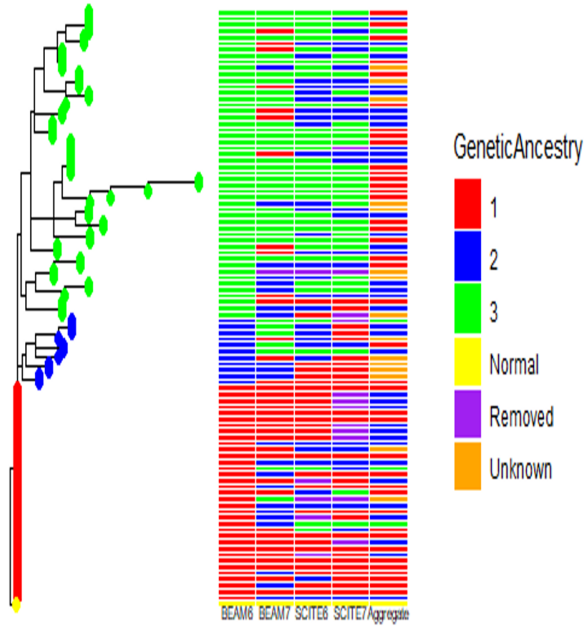
The *AnnotationComparison* function allows users to assess the reliability of ancestry annotation comparison by visualizing the agreement between phylogenies and/or provided annotations.

AncestryComparison requires as input the path to a reference tree (or annotation file), the annotation file itself in the form of a data frame, and a list of paths to files (trees or annotation files) which the user would like to compare to the reference. For the reference_tree variable, if the user enters in a path to a phylogenetic tree, the function will plot the reference tree with each cell colored uniquely based on the ancestry assignments in the reference_annotation file. If the user does not enter a tree, but instead the path to an annotation file, *AncestryComparison* will produce only a heatmap for comparing the annotations.

```
reference_tree <- "MOCA/Example/tree1.nwk"

reference_annotations <- annotate_file1

alternates <- list.files('MOCA/Example/alternates')

for(a in 1:length(alternates)){
  alternates[a] <- paste0('MOCA/Example/alternates/', alternates[a])
}

annotation_heatmap <- AncestryComparison(reference_tree, reference_annotations, alternates)
```

If any of the alternate file paths are to a phylogenetic tree, *AncestryComparison* will call *TreeBalance*, *BalancedAnnotation*, *UnbalancedAnnotation*, and *userDefineClades* in order to identify ancestries. In this case, users will be prompted to choose which annotation function they would like to use, to set the parameters, and to manually edit the clades if they choose. In the case the alternate file path is to an annotation file, no additional input is required from the user.

If two or more different alternate annotations are entered, *AncestryComparison* will also produce consensus ancestries. A consensus ancestry is assigned based on the ancestry which is most common for a cell in the user provided annotations. If a cell contains equal number of assignments to two or more different ancestries, the consensus ancestry identification will label the cell as "Unknown." A cell is also labeled as "Unknown" if the cell is most commonly removed from the input ancestry assignments. *AncestryComparison* function returns a categorical heatmap comparing the annotations of each tree/annotation file, and plots a figure which displays a reference tree (if provided) alongside the categorical heatmap.



	BEAM6	BEAM7	SCITE6	SCITE7	Aggregate
CellSRR1295042	1	1	1	1	1
CellSRR1294974	1	1	3	3	Unknown
CellSRR1295067	1	1	3	1	1
CellSRR1295009	1	1	3	3	Unknown
CellSRR1294977	1	1	1	3	1
CellSRR1295053	1	Removed	Removed	Removed	Unknown
CellSRR1294978	1	1	3	3	Unknown

Figure 5: Ouput from AncestryComparison tool.

The Genetic Ancestry Validation module of tools also hosts the *userDefineClades* function.

```
revised_annotation_table <- userDefineClades(annotate_file1)
```

This function takes as input an annotation file data frame and prompts users if they would like to change the assignment name of each clade. *userDefineClades* returns a data frame which is appropriately formatted as an annotation file with the clade assignments updated. This function is implemented in *BalancedAnnotation*, *UnbalancedAnnotation*, and *AncestryComparison* as well, so they are dependent on its presence in the R environment.

Conclusion

MOCA provides three main modules of tools: Genetic ancestry identification, gene expression trajectory analysis, and genetic ancestry validation. MOCA is useful for performing molecular evolution analysis for cells whose genetic and expression profiles can be jointly analyzed. For more information please see the manuscript cited below.

How to cite

If you use MOCA tools in your work, please cite the accompanying publication.

Huzar, J., Kim H., Kumar S., Miura S. Analyzing tumor evolution using genetic variation and gene expression profiles from single cells. XXX (2021)