Paper 4680-2020

Identifying "One-Offs": Tuberculosis Genotype Cluster Detection Using the COMPGED Function

Edward Lan, County of Los Angeles Department of Public Health Tuberculosis Control Program

ABSTRACT

In surveillance and epidemiologic functions of public health departments, SAS® software programs are leveraged to identify clusters and outbreaks. Tuberculosis (TB) genotyping is a method used to identify TB patients with closely related strains, indicating potential recent transmission. The genotyping process produces a resulting GENType, comprised of the 15 length spoligotype string and a 24-length mycobacterial interspersed repetitive units (MIRU) string. Each week, the TB Control Program (TBCP) in Los Angeles County (LAC) receives a weekly update of genotyping results of TB isolates. TBCP monitors additions to high priority genotype clusters, i.e. matching GENTypes, and looks for patients with a result string that is one alphanumeric character different than one of the high priority GENTypes. Using a macro %DO loop, the COMPGED function quantifies differences between a patient's TB isolate and a reference library of genotypes for high priority clusters in the County. The quantifiable differences determine exact matches and results that are "one-off" are further investigated by the outbreak team to interrupt TB transmission.

INTRODUCTION

Outbreaks, epidemiology, transmission: these are words that have found its way into our vernacular in recent years. How can we use SAS to make a difference?

When it comes to TB outbreaks, laboratory testing analyzes the genetic content of Mycobacterium tuberculosis. These genetic patterns assist with determining different strains of TB and provide insight on chain of transmission. For a patient suspected of having TB, a specimen is collected and if TB grows in culture media, the isolate is sent to genotyping laboratories. The resulting GENType (complete) is comprised of the 15-lengh spoligotype string and a 24-length MIRU string (partial). Patients with matching GENTypes are considered part of the same genotype cluster.

TBCP has an outbreak and cluster investigation team that conducts patient interviews to determine epidemiologic links in patient history. The Program investigates high priority clusters and monitors any additions to the GENType. Typically, the MIRU results (partial) are returned prior to the full GENType result, which allows for earlier detection if the partial results show either an exact match or a one-off match.

To facilitate the processing early matching, the TBCP team requested that an alert be created for patients that have matched GENTypes, and one-off matches for both complete and partial results. This is where SAS comes in. SAS compares reference strings to the patients' laboratory results. The weekly report provides the team updated information on any new additions to the cluster and any potential new additions based on complete and partial results, respectively.

In this paper, we will discuss how using the COMPGED function in a macro %DO loop quantifies differences between a patient's lab result against the library of high priority genotypes.

CREATING A REFERENCE LIBRARY OF GENOTYPING RESULTS

We created a reference library of genotyping results for the high priority clusters. At TBCP, the library is kept in an Excel spreadsheet containing the GENType, spoligotype, and MIRU values for each high priority cluster (see Figure 1).

GENType	MIRUREF	SPOLMIRUREF
G11610	22342515332224252422 3324	77777737756077122342515332224252422 3324
Grupo	22532513332323223442 3334	77777677000000022532513332323223442 3334
G25340	22531315322123353242 3335	77777772772077122531315322123353242 3335
G11945	12432617322135222412 3227	77777760756077112432617322135222412 3227
G16398	22442515332224242422 4326	77777777776077122442515332224242422 4326

Figure 1. Excel Sheet with Genotyping Results for High Priority Clusters

This library can be easily updated by any investigative staff and be read in using the LIBNAME XLSX DATA step in SAS.

```
libname log xlsx "&Data\Reference.xlsx";
data work.reference;
set log.'Sheet1'n;
run;
libname log clear;
```

Using a %LET statement, we can store the list of GENTypes into a macro variable, for both partial (MIRU) results and complete results (GENType):

```
%let miruref=g11610miruref grupo1miruref g25340miruref g11945miruref
g16398miruref

%let genoref=g11610ref grupo1ref g25340ref g11945ref g16398ref
```

SETTING UP COMPGED FOR COMPARISONS

Now that we have the reference library set up and read into a SAS, we can set up the %DO loop for the COMPGED function comparisons. Two sets of macros are needed: one for the partial results and one for the complete results. The first %DO loop sets up detection of one-offs for only MIRU (partial) results:

```
%macro oneoffmiru(all_ref);
      %let k=1;
      %let dep=%scan(&all_ref,&k);

      %do %while("&dep" ne "");
            *sets up counts of difference between 2 strings;
            diff&dep=compged(&dep,miru24);
            %let k=%eval(&k + 1);
            %let dep=%scan(&all_ref, &k);
      %end;
%mend;
```

The second %DO loop sets up detection of one-offs for complete GENType results (MIRU and spoligotype):

```
%macro oneoffspol(all_ref);
      %let k=1;
      %let dep=%scan(&all_ref,&k);

      %do %while("&dep" ne "");
```

```
                *sets up counts of difference between 2 strings;
                diff&dep=compged(&dep,spolmiru);
                %let k=%eval(&k + 1);
                %let dep=%scan(&all_ref, &k);
        %end;
%mend;
```

In both %DO loops, the COMPGED function calculates the distance, i.e. the difference, between the two text strings. The %DO loop cycles through all the high priority clusters stored in the reference macro variables and evaluates it against the reference library of cluster data:

```
%oneoffmiru(&miruref);
%oneoffspol(&genoref);
```

In a subsequent DATA step, we can set up categories to define the quantity differentials:

```
/*with only miru24 (partial)*/
if diffg11610miruref le 100 then cluster="G11610";
else if diffgrupo1miruref le 100 then cluster="Grupo";
else if diffg25340miruref le 100 then cluster="G25340";
else if diffg11945miruref le 100 then cluster="G11945";
else if diffg16398miruref le 100 then cluster="G16398";

/*w/ spoligotype (complete)*/
if diffg11610ref le 100 then cluster="G11610";
else if diffgrupo1ref le 100 then cluster="Grupo";
else if diffg25340ref le 100 then cluster="G25340";
else if diffg11945ref le 100 then cluster="G11945";
else if diffg16398ref le 100 then cluster="G16398";
```

After reviewing the results, we noticed that a one-off difference in any two strings had a score of COMPGED less than or equal to 100. Consequently, we can establish additional categories in the same DATA step process to define the one-off categorizations:

```
if (cluster="G11610" & ((0 lt diffg11610ref le 100) or (0 lt diffg11610miruref
le 100)))
or (cluster="Grupo" & ((0 lt diffgrupo1ref le 100) or (0 lt diffgrupo1miruref
le 100)))
or (cluster="G25340" & ((0 lt diffg25340ref le 100) or (0 lt diffg25340miruref
le 100)))
or (cluster="G11945" & ((0 lt diffg11945ref le 100) or (0 lt diffg11945miruref
le 100)))
or (cluster="G16398" & ((0 lt diffg16398ref le 100) or (0 lt diffg16398miruref
le 100)))
then oneoff="Yes";
else oneoff="";
```

The program produces an output dataset comprised of a line list of patients, their respective genotyping results, the matched high priority cluster, and one-off categorization. The data compiled into a table would look like the following:

Patient	Full match	One off	Not match	Patient MIRU Result	G11610 MIRU Reference	COMPGED Score - Partial
1	x			22342515332224252422 3324	22342515332224252422 3324	0
2			x	22342515432324252422 3324	22342515332224252422 3324	140
3			x	22342515432024252422 3324	22342515332224252422 3324	200
4	x			22342515332224252422 3324	22342515332224252422 3324	0
5	x			22342515332224252422 3324	22342515332224252422 3324	0
6		x		22341515332224252422 3324	22342515332224252422 3324	100
7		x		22342515432224252422 3324	22342515332224252422 3324	100
8	x			22342515432224252422 3324	22342515332224252422 3324	0

Figure 2. Output data comparison with COMPGED scores

In Figure 2, the patient's TB lab results are listed along with the example high priority cluster reference values with the string differences highlighted in red. Scores of 100 were classified as one-offs, scores of zero were full matches, and anything else was a non-match. Though the display only shows the results for one high priority cluster and only for the partial results, the %DO loop cycles performs the comparison of the patients' results with all the high priority clusters stored in the reference library.

EXPORTING RESULTS: ODS EXCEL AND PROC REPORT

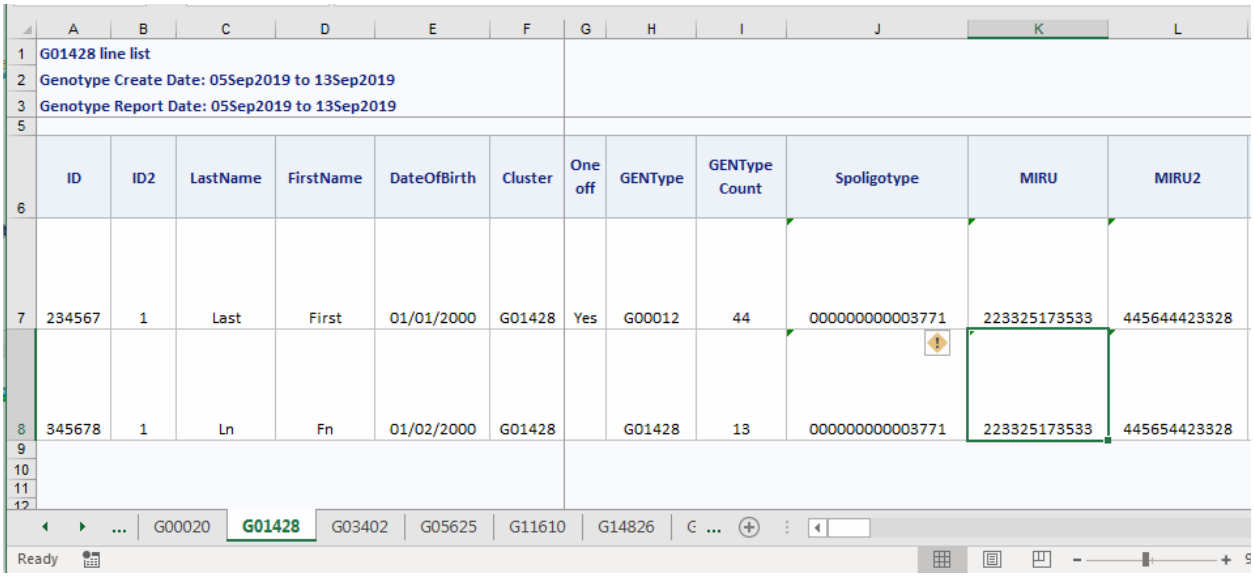
After executing the %DO loops and data cleaning procedures, the data are ready for dissemination to the investigative team. TBCP's investigative team requested Excel spreadsheets as the preferred report format, thus we can utilize ODS EXCEL and the REPORT procedure to accommodate their workflow. If a patient belongs to one of the high priority clusters, then an Excel tab is created in the spreadsheet with the cluster name using a macro with ODS EXCEL:

```
ods excel file="C:\CA Download Cluster Update.xlsx";
%macro create_tab;
	%do i=1 %to &numcmp;

ods excel
	OPTIONS (Orientation = 'landscape'
	sheet name = "&&E&I"
	embedded_titles='yes'
	embedded_footnotes='yes'
	embed_footnotes_once='yes'
	embed_titles_once='on'
	frozen_headers='3'
	row_repeat='2'
	frozen_rowheaders='6'
	);
title1 "&&E&I line list";
footnote1 "Source: CA Download from &date" justify=left;
```

```
proc report data=ClusterUpdateFinal nowd center style=[just=center];
      where cluster="&&E&I";
      column ID ID2 LastName FirstName dateofbirth cluster oneoff GENType
            count Spoligotype miru miru2 pub_hlth_ctr homeless
            Genotype_Report_Date Genotype_Create_Date Genotype_Modified_Date
            county submitter_number;
run;
title1;
footnote1;
%END;
%mend;

%create_tab;
ods excel close;
```

The macro creates Excel sheets for each of the high priority clusters that were matched to a patient's lab result and non-matched clusters are not listed in the report (creation of macro variables is not shown). Each tab provides a line list of patients and any specified variables courtesy of the PROC REPORT.

For instance, if a patient's GENType results matched a high priority cluster, then a spreadsheet tab is created with a line list of patients matched:



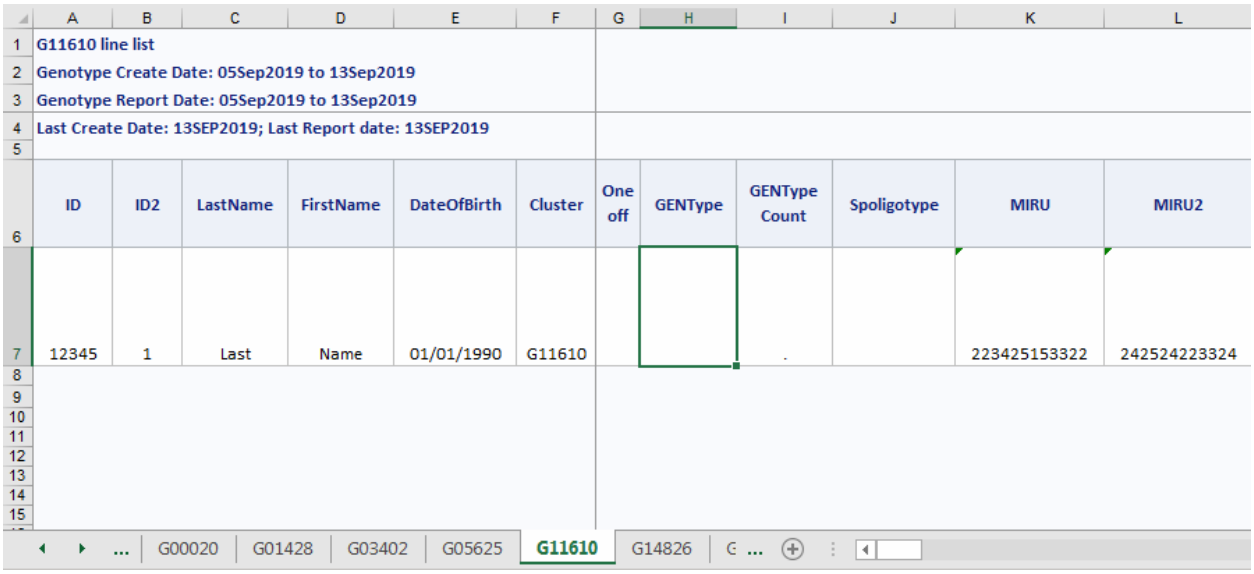
G01428 line list
Genotype Create Date: 05Sep2019 to 13Sep2019
Genotype Report Date: 05Sep2019 to 13Sep2019

ID	ID2	LastName	FirstName	DateOfBirth	Cluster	One off	GENType	GENType Count	Spoligotype	MIRU	MIRU2
234567	1	Last	First	01/01/2000	G01428	Yes	G00012	44	000000000003771	223325173533	445644423328
345678	1	Ln	Fn	01/02/2000	G01428		G01428	13	000000000003771	223325173533	445654423328

Display 1. Sample Output for a High Priority Cluster

In Display 1, the second patient listed is an exact match to high priority cluster "G01428" and the first patient listed is a one-off in the cluster. The PROC REPORT output also provides cumulative counts of each the high priority clusters and can be configured to output any requested variable.

If only the partial MIRU result is available, the patient is included in the cluster tab as either a one-off or a match.



	A	B	C	D	E	F	G	H	I	J	K	L
1	G11610 line list											
2	Genotype Create Date: 05Sep2019 to 13Sep2019											
3	Genotype Report Date: 05Sep2019 to 13Sep2019											
4	Last Create Date: 13SEP2019; Last Report date: 13SEP2019											
5												
6	ID	ID2	LastName	FirstName	DateOfBirth	Cluster	One off	GENType	GENType Count	Spoligotype	MIRU	MIRU2
7	12345	1	Last	Name	01/01/1990	G11610			.		223425153322	242524223324
8												
9												
10												
11												
12												
13												
14												
15												

... | G00020 | G01428 | G03402 | G05625 | G11610 | G14826 | G ...

Display 2. Sample Output for a High Priority Cluster

In Display 2, GENType and spoligotype results are missing, with only the preliminary MIRU results available. The preliminary and partial result is still important. An exact match to the MIRU reference values results in an exact match to the full result later. The report assigns the patient to the cluster even before the full results are resulted, which is helpful for timely cluster investigations. These results are neatly compiled into a single Excel workbook that is distributed to members of the investigative team.

CONCLUSION

For TB genotyping data analysis, the COMPGED function is one approach to identifying exact matches and one-off matches. The %DO loop combined with the COMPGED function is a powerful tool to iterate through the values in a reference library. For TB clusters and outbreaks, timely alerts and early investigations aid in preventing further transmission. The Excel report of COMPGED function results benefit the TBCP investigative team on a weekly basis as the Program aims to eliminate TB in the County. This approach may also be applied to other disease outbreak investigations to prevent further spread of disease within the community.

REFERENCES

Strum, P. 2007. "Fuzzy Matching using the COMPGED Function." *Proceedings of the NorthEast SAS Users Group 2007 Conference*. Baltimore, MD: NESUG. Available at https://www.lexjansen.com/nesug/nesug07/ap/ap23.pdf.

University of California Los Angeles Institute for Digital Research and Education Statistical Consulting. "Introduction to SAS Macro Language." Accessed April 8, 2016. https://stats.idre.ucla.edu/sas/seminars/sas-macros-introduction/.

ACKNOWLEDGMENTS

The author thanks Dr. Ramon Guevara for his encouragement and support of this paper. In addition, this paper would not have been possible without the assistance of Shameer Poonja, who provided the explanation of business needs of the TBCP investigative team.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Edward Lan
County of Los Angeles Department of Public Health Tuberculosis Control Program
(213) 745-0800
elan@ph.lacounty.gov