Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub54 Törn

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
August 7, 2018

Contents

1 Standard Disclaimer

The intent of this DSIC is to provide confidence that the data distributed by the NIDDK repository is a true copy of the study data. Our intent is not to assess the integrity of the statistical analyses reported by study investigators. As with all statistical analyses of complex datasets, complete replication of a set of statistical results should not be expected in secondary analysis. This occurs for a number of reasons including differences in the handling of missing data, restrictions on cases included in samples for a particular analysis, software coding used to define complex variables, etc. Experience suggests that most discrepancies can ordinarily be resolved by consultation with the study data coordinating center (DCC), however this process is labor-intensive for both DCC and Repository staff. It is thus not our policy to resolve every discrepancy that is observed in an integrity check. Specifically, we do not attempt to resolve minor or inconsequential discrepancies with published results or discrepancies that involve complex analyses, unless NIDDK Repository staff suspect that the observed discrepancy suggests that the dataset may have been corrupted in storage, transmission, or processing by repository staff. We do, however, document in footnotes to the integrity check those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

2 Study Background

The TEDDY study was designed to follow children with and without a family history of T1D to understand the environmental factors that contribute to the disease. Newborn children younger than 4 months were screened for high-risk HLA alleles, and those with qualifying haplotypes were eligible for follow-up. Information is collected on medical information (infections, medication, immunizations), exposure to dietary and other environmental factors, negative life events, family history, tap water, and measurements of psychological stress. Biospecimens, including blood, stool, urine, and nail clippings, are taken at baseline and follow-up study visits. The primary outcome measures include two endpoints—the first appearance of one or more islet cell autoantibodies (GADA, IAA, or IA-2A), confirmed at two consecutive visits, and development of T1D. The cohort will be followed for 15 years, or until the occurrence of one of the primary endpoints.

3 Archived Datasets

All the SAS data files, as provided by the Data Coordinating Center (DCC), are located in the TEDDY folder in the data package. For this replication, variables were taken from the "m_54_ctorn_niddk_31jul2013.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Törn et al [1] Diabetes in 2015. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 2 and Table 5 in the publication [1], only the MAF value was checked in this review. In both tables, the MAF value in the data was close to the value in the publication.

6 Conclusions

The NIDDK repository is confident that the TEDDY M54 data files to be distributed are a true copy of the study data.

7 References

[1] Carina Törn, David Hadley, Hye-Seung Lee, William Hagopian, Åke Lernmark, Olli Simell, Marian Rewers, Anette Ziegler, Desmond Schatz, Beena Akolkar, Suna Onengut-Gumuscu, Wei-Min Chen, Jorma Toppari, Juha Mykkänen, Jorma Ilonen, Stephen S. Rich, Jin-Xiong She, Andrea K. Steck, Jeffrey Krischer, and the TEDDY Study Group. “Role of Type 1 Diabetes–Associated SNPs on Risk of Autoantibody Positivity in the TEDDY Study”. Diabetes 2015;64:1818–1829.

Table A: Variables used to replicate Table 2 and Table 5

Table Variable	**dataset.variable**
SNP	m_54_ctorn_niddk_31jul2013.RS##

Table B: Comparison of values computed in Table 2 and Table 5

		MAF	MAF	
Chromosome	SNP	Manuscript	DSIC	Diff
11p15.5	rs1004446	0.3759	0.3759	0.0000
10q23.31	rs10509540	0.2666	0.2666	0.0000
4p15.2	rs10517086	0.2889	0.2889	0.0000
21q22.3	rs11203203	0.3462	0.3462	0.0000
10p15.1	rs11258747	0.2409	0.2409	0.0000
3p21.31	rs11711054	0.3011	0.3014	-0.0003
10p15.1	rs12251307	0.1262	0.1262	0.0000
16p13.13	rs12708716	0.3430	0.3439	-0.0009
14q24.1	rs1465788	0.2874	0.2874	0.0000
2p25.1	rs1534422	0.4574	0.4574	0.0000
17p13.1	rs16956936	0.1187	0.1187	0.0000
6q25.3	rs1738074	0.4131	0.4131	0.0000
18p11.21	rs1893217	0.1641	0.1641	0.0000
2q24.2	rs1990760	0.3976	0.3976	0.0000
1p31.3	rs2269241	0.2294	0.2294	0.0000
20p13	rs2281808	0.3420	0.3420	0.0000
17q12	rs2290400	0.4718	0.4807	-0.0089
12q13.2	rs2292239	0.3271	0.3271	0.0000
22q13.1	rs229541	0.4096	0.4098	-0.0002
6q23.3	rs2327832	0.2120	0.2131	-0.0011
1p13.2	rs2476601	0.1113	0.1114	-0.0001
1q31.2	rs2816316	0.1768	0.1768	0.0000
1q32.1	rs3024505	0.1568	0.1583	-0.0015
2q33.2	rs3087243	0.3978	0.3984	-0.0006
12q24.12	rs3184504	0.4590	0.4590	0.0000
15q25.1	rs3825932	0.3470	0.3471	-0.0001
19q13.32	rs425105	0.1585	0.1585	0.0000
4q27	rs4505848	0.3813	0.3813	0.0000
12p13.31	rs4763879	0.3789	0.3789	0.0000
16p11.2	rs4788084	0.4432	0.4433	-0.0001
14q32.2	rs4900384	0.3247	0.3247	0.0000

		MAF	MAF	
Chromosome	SNP	Manuscript	DSIC	Diff
7p12.1	rs4948088	0.0442	0.0443	0.0000
22q12.2	rs5753037	0.3604	0.3604	0.0000
5p13.2	rs6897932	0.2896	0.2896	0.0000
9p24.2	rs7020673	0.4939	0.4939	0.0000
11p15.5	rs7111341	0.2659	0.2659	0.0000
16q23.1	rs7202877	0.1115	0.1115	0.0000
17q21.2	rs7221109	0.3710	0.3711	-0.0001
18q22.2	rs763361	0.4799	0.4799	0.0000
7p15.2	rs7804356	0.2282	0.2282	0.0000
6q22.32	rs9388489	0.4476	0.4477	-0.0001

Attachment A: SAS Code

```
*** TEDDY M54 DSIC;
*** Programmer: Sabrina Chen
*** Date: 6/26/18;

options nocenter validvarname=upcase macrogen;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_54_dsic.sas';
run;


********;
* INPUT ;
********;
libname dat '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_54_CTorn_NIDDK_Submission';

* chromosome to snp ref;
proc import
datafile="/prj/niddk/ims_analysis/TEDDY/private_created_data/TEDDY_M54_chromosome_snp.csv"
  dbms=csv
  out=work.chrom;
run;


*********;
* FORMAT ;
*********;
proc format;
 value noyes
 .  = "no"
 other = "yes"
 ;
run;


********;
* MACRO ;
********;

** this version of the hwe macro expands the final table in a dataset so that it can be exported
directly to excel;
%macro cal_chi_square_v3(datasetname=, varname=  );

data temp;
  set &datasetname;
  if &varname._order ne .;
run;

proc freq data = temp;
*  tables  &varname._order  *&varname  / list  out= freqout noprint;
  tables  &varname._order    / list  out= freqout noprint;
run;

data freqout;
  set freqout;
  subject = 1;
  if &varname._order = 0 then VAR=1;
  else if &varname._order = 1 then VAR=2;
  else if &varname._order = 2 then VAR=3;
run;

*Nucleotide_change;
data oneobs_temp;
  set freqout;
  by subject;
  array xx[*] AA AB BB;
  array YY[*] AA_FREQ AB_FREQ BB_FREQ;
*  array zz[*]$ AA_SNP AB_SNP BB_SNP;
  retain AA AB BB AA_FREQ AB_FREQ BB_FREQ /* AA_SNP AB_SNP BB_SNP */;
```

```
  if first.subject then do i = 1 to 3;
   xx[i] = .;
   YY[i] = .;
  * zz[i] = ' ';
  end;
 xx[VAR] = count;
 YY[VAR] = Percent;
* zz[VAR] = compress(&varname) || '/' || compress(put(&varname._order,1.)) ;
 if last.subject then output;
run;

data oneobs_temp(drop = var    COUNT    PERCENT    subject i);
  length var_label $60;   *** added by SEC, 6Feb2008;
  set oneobs_temp;
  var_label = vlabel(&varname);
  if AA = . then AA = 0;
  if AB = . then AB = 0;
  if BB = . then BB = 0;
  if AA_FREQ = . then AA_FREQ = 0;
  if AB_FREQ = . then AB_FREQ = 0;
  if BB_FREQ = . then BB_FREQ = 0;

  sum=AA+AB+BB;
  p_AA = (2*AA + AB)/(2*(AA+AB+BB));
  p2_AA = p_AA * p_AA;
  q_BB = (2*BB + AB)/(2*(AA+AB+BB));
  q2_BB = q_BB *q_BB;
  pq2 = 2*p_AA*q_BB;
  exp_AA = sum*p2_AA;
  exp_AB = sum*pq2;
  exp_BB = sum*q2_BB;
  sum_exp = exp_AA + exp_AB + exp_BB;
  chi_AA =(AA-exp_AA)*(AA-exp_AA)/exp_AA;
  chi_AB =(AB-exp_AB)*(AB-exp_AB)/exp_AB;
  chi_BB =(BB-exp_BB)*(BB-exp_BB)/exp_BB;
  sum_chi =chi_AA + chi_AB + chi_BB;
  p_value= 1-PROBCHI(sum_chi,1);
  sum_letter = 'Chi, df';
  p_letter = 'p-value';
  exp_AA_freq =  exp_AA/sum_exp * 100;
  exp_AB_freq =  exp_AB/sum_exp * 100;
  exp_BB_freq =  exp_BB/sum_exp * 100;

  * calculate the MAF (minor allele frequency);
  * Minor allele frequency (MAF) = (1*frequency of the heterozygote + 2*frequency of the
homozygote variant) / (2*total N for that SNP)  ;
  maf = (1*AB + 2*BB)/(2*(AA+AB+BB));
run;

*proc print data = oneobs_temp;


** expand the HW table so that we can export to .xls ;
data forxls (keep=sum_chi p_value var_label snpname order /*nucleotides*/ observed_n observed_p
expected_n expected_p maf);
  length snpname $20;
  set oneobs_temp;
 * array genos(3) aa_snp      ab_snp      bb_snp;
  array obs_n(3) aa          ab          bb;
  array obs_p(3) aa_freq     ab_freq     bb_freq;
  array exp_n(3) exp_aa      exp_ab      exp_bb;
  array exp_p(3) exp_aa_freq exp_ab_freq exp_bb_freq;

  do i=1 to 3;
   * nucleotides = genos(i);
    observed_n = obs_n(i);
    observed_p = obs_p(i);
    expected_n = exp_n(i);
    expected_p = exp_p(i);
    snpname = "&varname";
    order = i;
```

```
    output;
  end;

 * nucleotides = .;
  observed_n = .;
  observed_p = .;
  expected_n = .;
  expected_p = .;
  snpname = "&varname";
  order = 4;
  output;
run;

data xls;
  set xls forxls;
run;

%mend cal_chi_square_v3;



data m_54;
       set dat.m_54_ctorn_niddk_31jul2013;
run;

proc contents data = m_54;
title3 'M 54 file';
run;


proc freq data=m_54;
  tables  PERSIST_CONF_AB
          HLA_CATEGORY
            ALL3AB
/*    ALL3AB_DAYS               */
    CC
    COUNTRY
/*    DIAB_DAYS                 */
    DR33
    DR34
    DR44_48
    FDR
    FEMALE
/*    FID_MASKED                */
    FIN
    GAD_IA2A
/*    GAD_IA2A_DAYS             */
    GAD_MIAA
   /* GAD_MIAA_DAYS             */
    GAD_ONLY
  /*  GAD_ONLY_DAYS              */
    GER
    HLARG
    HLA_CATEGORY
    IA2A_MIAA
   /* IA2A_MIAA_DAYS            */
    IA2A_ONLY
/*    IA2A_ONLY_DAYS            */
    INDETERMINATE
    LAST_GAD_VISIT
    LAST_IA2A_VISIT
    LAST_MIAA_VISIT
    LAST_SERUM_VISIT
/*    MAT_MASKED                */
    MIAA_ONLY
  /*  MIAA_ONLY_DAYS            */
    NAB
 /*   PANT_DAYS                 */
 /*   PAT_MASKED                */
    PCA1
    PCA2
```

```
    PERSIST_CONF_AB
    PERSIST_CONF_GAD
    PERSIST_CONF_GAD_VISIT
    PERSIST_CONF_IA2A
    PERSIST_CONF_IA2A_VISIT
    PERSIST_CONF_MIAA
    PERSIST_CONF_MIAA_VISIT
    PERSIST_VISIT
    PHENOTYPE
    RACE_ETHNICITY
    SEX
    SEX_DEMO
    SWE
    T1D
    T1D_VISIT
   /* TIMETODIAB
    TIME_TO_ALL3AB
    TIME_TO_GAD_IA2A
    TIME_TO_GAD_MIAA
    TIME_TO_GAD_ONLY
    TIME_TO_IA2A_MIAA
    TIME_TO_IA2A_ONLY
    TIME_TO_MIAA_ONLY
    TIME_TO_PANT */
    RS2476601_A /missing;
title3 'Take a look at raw freqs';
run;


proc freq data=m_54;
    tables PERSIST_CONF_AB MIAA_ONLY GAD_ONLY/missing;
    tables PERSIST_CONF_AB*dr33/list missing;
    tables PERSIST_CONF_AB*dr34/list missing;
    tables PERSIST_CONF_AB*DR44_48/list missing;
title3 'Check counts for table columns';
run;

* before calling macros, create ds to append to;
data xls;
  set _null_;
run;

data m_54;
  set m_54 (rename=(RS2476601_A = RS2476601_A_order
                   RS1004446_A = RS1004446_A_order
                   RS10509540_G= RS10509540_G_order
                   RS10517086_A= RS10517086_A_order
                   RS11203203_A= RS11203203_A_order
                   RS11258747_A= RS11258747_A_order
                   RS11711054_G= RS11711054_G_order
                   RS12251307_A= RS12251307_A_order
                   RS12708716_G= RS12708716_G_order
                   RS1465788_A = RS1465788_A_order
                   RS1534422_G = RS1534422_G_order
                   RS16956936_A= RS16956936_A_order
                   RS1738074_A = RS1738074_A_order
                   RS1893217_G = RS1893217_G_order
                   RS1990760_G = RS1990760_G_order
                   RS2269241_G = RS2269241_G_order
                   RS2281808_A = RS2281808_A_order
                   RS2290400_A = RS2290400_A_order
                   RS2292239_A = RS2292239_A_order
                   RS229541_A  = RS229541_A_order
                   RS2327832_G = RS2327832_G_order
                   RS2664170_G = RS2664170_G_order
                   RS2816316_C = RS2816316_C_order
                   RS3024505_A = RS3024505_A_order
                   RS3087243_A = RS3087243_A_order
                   RS3184504_A = RS3184504_A_order
                   RS3825932_A = RS3825932_A_order
                   RS425105_G  = RS425105_G_order
```

```
                    RS4505848_G = RS4505848_G_order
                    RS4763879_A = RS4763879_A_order
                    RS4788084_A = RS4788084_A_order
                    RS4900384_G = RS4900384_G_order
                    RS4948088_A = RS4948088_A_order
                    RS5753037_A = RS5753037_A_order
                    RS6897932_A = RS6897932_A_order
                    RS7020673_G = RS7020673_G_order
                    RS7111341_A = RS7111341_A_order
                    RS7202877_C = RS7202877_C_order
                    RS7221109_A = RS7221109_A_order
                    RS763361_A  = RS763361_A_order
                    RS7804356_G = RS7804356_G_order
                    RS9388489_G = RS9388489_G_order));
run;

* calculate minor allele freq;
%cal_chi_square_v3(datasetname=m_54, varname=RS2476601_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS1004446_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS10509540_G);
%cal_chi_square_v3(datasetname=m_54, varname=RS10517086_A);
%cal_chi_square_v3(datasetname=m_54, varname=RS11203203_A);
%cal_chi_square_v3(datasetname=m_54, varname=RS11258747_A);
%cal_chi_square_v3(datasetname=m_54, varname=RS11711054_G);
%cal_chi_square_v3(datasetname=m_54, varname=RS12251307_A);
%cal_chi_square_v3(datasetname=m_54, varname=RS12708716_G);
%cal_chi_square_v3(datasetname=m_54, varname=RS1465788_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS1534422_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS16956936_A);
%cal_chi_square_v3(datasetname=m_54, varname=RS1738074_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS1893217_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS1990760_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS2269241_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS2281808_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS2290400_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS2292239_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS229541_A  );
%cal_chi_square_v3(datasetname=m_54, varname=RS2327832_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS2816316_C );
%cal_chi_square_v3(datasetname=m_54, varname=RS3024505_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS3087243_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS3184504_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS3825932_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS425105_G  );
%cal_chi_square_v3(datasetname=m_54, varname=RS4505848_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS4763879_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS4788084_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS4900384_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS4948088_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS5753037_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS6897932_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS7020673_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS7111341_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS7202877_C );
%cal_chi_square_v3(datasetname=m_54, varname=RS7221109_A );
%cal_chi_square_v3(datasetname=m_54, varname=RS763361_A  );
%cal_chi_square_v3(datasetname=m_54, varname=RS7804356_G );
%cal_chi_square_v3(datasetname=m_54, varname=RS9388489_G );


* add chromosome to the list;
data xls;
  set xls;
  * need merge var snp;
  cut = index(snpname,'_');
  snp = substr(snpname,1,cut-1);
run;

*proc print data=xls;
*run;
```

```
data chrom (keep=chromosome snp);
  set chrom;
  snp = upcase(snp);
run;

proc sort data=xls out=xlsper nodupkey;
  by snp maf;
run;

proc sort data=chrom;
  by snp;
run;

data report (keep=chromosome snp maf);
  merge xlsper (in=in1) chrom (in=in2);
  by snp;
  if in1 or in2;
run;

proc sort data=report;
  by chromosome;
run;

proc print data=report;
  var chromosome snp maf;
  title3 'MAF in Tables 2 and 5';
run;
```