Minimum **I**nformation about a **Gen**otyping Experiment **(MIGen)**

Experiment Annotation Example

Jie Huang, Richard Scheuermann,
Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX USA

Introduction

This document represents an example application of Minimum Information about a Genotyping Experiment (MIGen) standard on reporting a genotyping experiment. This example is intended for the purpose of demonstrating MIGen only.

A MIGen-compliant genotyping experiment description shall include all relevant information specified in the MIGen standard. MIGen states the content of the provided information; it does not imply the format or structure of the report or whether an item should be directly provided or referenced. Within this example we follow the MIGen structure in order to demonstrate MIGen as clearly as possible.

For illustration purpose this document uses information of the genotyping experiments reported in [1] and [2] as references. Some information was made up when necessary information was not provided in the referral documents or to simplify the example document.

1. Experiment Overview:

1.1 Purpose:

To identify new genetic factors associated with increased the risk of type 1 diabetes (T1D; a.k.a. diabetes mellitus), we performed a genome-wide association study.

1.2 Features of the Genetic Variants Under Study:

1.2.1 Type of Genomic Sequence Feature Variation(s) Assessed: single nucleotide polymorphisms (SNPs).

1.2.2 Number of Genomic Sequence Features Analyzed: We genotyped 500,568 SNPs.

1.2.3 Selection Criteria: SNP genotyping was performed with the commercial release of the GeneChip 500K arrays (#901189 and # 901188).

1.3 Keywords:

type 1 diabetes, genome-wide association, genotyping experiment, genetic association

1.4 Organization(s):

1.4.1 Organization Name: Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory,

Department of Medical Genetics, Cambridge Institute for Medical Research

1.4.2 Organization Address: University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK.

1.4.3 Role the Organization Played: Main study site

1.5 Study Personnel

1.5.1 Name: John A. Todd

1.5.2 Email Address: john.todd@cimr.cam.ac.uk

1.5.3 Role: Principle Investigator

1.6 Date: 2005 - 2007.

1.7 Conclusions: In this study, single-point analyses revealed three novel regions (on chromosomes 12q13, 12q24 and 16p13) showing strong evidence of association ($P<5x10^{-7}$) with increased risk of T1D development; two additional regions attained similar levels of significance through multilocus analyses.

1.8 Other Relevant Experiment Information: The principal funder of this project was the Wellcome Trust. For the 1958 Birth Cohort, venous blood collection was funded by the UK Medical Research Council and cell-line production.

2 Experiment Subjects Description:

2.1 Study Subject Common Selection Criteria:

2.1.1 Subject Sampling Method: case and control design. T1D cases were recruited from pediatric and adult diabetes clinics at 150 National Health Service hospitals across mainland UK. Gender- and geographically-matched controls were recruited accordingly.

2.1.2 Enrollment Inclusion and Exclusion Criteria: Individuals included in the study were living within England, Scotland and Wales and have Caucasian ancestry. Data from individuals with non-Caucasian ancestry were excluded from the final analysis.

2.1.3 Methods for Ascertaining Enrollment Criteria: Study subjects self-identified themselves as white Caucasian.

2.2 Study Subject Primary Characteristics:

2.2 – I Case Subjects

2.2.1 Name of the Characteristics Evaluated: Type I diabetes (T1D).

2.2.2 Methods/Criteria for Evaluating the Characteristics: All T1D cases have an age of diagnosis before 17 years of age and have been insulin dependent since diagnosis (with a minimum period of at least 6 months). Subjects with maturity onset diabetes of the young (MODY) or permanent neonatal diabetes mellitus (PNDM) were excluded.

2.2.3 Number of Subjects: 2000.

2.2 – II Control Subjects

2.2.1 Name of the Characteristics Evaluated: non-T1D.

2.2.2 Methods/Criteria for Evaluating the Characteristics: Control subjects did not have T1D, and were gender and geographical region matched with case subjects.

2.2.3 Number of Subjects: 1500 subjects from 1958 Birth Cohort Controls (58BC) and 1500 from UK Blood Services Controls (UKBS).

2.3 Study Subject Other Characteristics Captured:

2.3 - I

2.3.1 Name of the Characteristics Captured: age at study entry.

2.3.2 Value of the Characteristics: see table 1 - summary statistics.

2.3.3 Method Used to Capture the Characteristics: self-report.

2.3 - II

2.3.1 Name of the Characteristics Captured: gender.

2.3.2 Value of the Characteristics: see table 1 - summary statistics.

2.3.3 Method Used to Capture the Characteristics: self-report.

2.3 - III

2.3.1 Name of the Characteristics Captured: Body Mass Index.

2.3.2 Value of the Characteristics: see summary statistics.

2.3.3 Method Used to Capture the Characteristics: BMI(kg/m²) = weight in kilograms / height in meters².

Table 1.

	Case	Control
age (yr)	20.1±7.2	23.2±8.6
male gender (%)	50	49
body mass index (kg/m^2)	25.7±5.3	26.2±4.4

Data are mean ± SD or %

3. Genotyping Procedure

3.1 Genomic Variants (Genotyping Analyte) Description: Annotation of SNPs genotyped in this experiment can be found in GenomeWideSNP_5 Annotations, release 22 (3/9/2007). http://www.affymetrix.com/support/support_result.affx

3.2 Genotyping Processes Description:

3.2.1 Biomaterial Transformation:

3.2.1.1 Biomaterial Transformation Input:

3.2.1.1.1 Type of the Input Material: whole blood

3.2.1.1.2 Amount of the Input Material: 10ml

3.2.1.1.3 Other Attributes: whole blood was EDTA treated.

3.2.1.2 Biomaterial Transformation Process: *Blood collection, DNA isolation and quantification*: A 10 mL EDTA whole blood tube was collected for DNA isolation from each participant by venipuncture. DNA was isolated using the commercially available Puregene DNA Isolation kit (#D-40K, Gentra Systems, Inc., Minneapolis, MN). The isolation process was completed according to the manufacturer's instructions. To determine the DNA concentration of the sample, three 1:20 dilutions were evaluated at A_{260} and A_{280} (triplicate measurement) using the Spectromax spectrophotometer (Molecular Devices, Sunnyvale, CA).

3.2.1.3 Biomaterial Transformation Output:

3.2.1.1.3 Type of Output: DNA.

3.2.1.3.2 Attributes of Output: Genomic DNA; DNA concentration ≥ 50 ng μl^{-1}

3.2.1.4 Biomaterial Transformation Other Participants:

3.2.1.4 - I

3.2.1.4.1 Participant Identifier: EDTA blood collection tube. (PulmoLab #454021)

3.2.1.4.2 Role of Participant: blood collection container

3.2.1.4.3 Attributes of the Participant: N/A

3.2.1.4 - II

3.2.1.4.1 Participant Identifier: Puregene DNA Isolation kit (#D-40K, Gentra Systems, Inc., Minneapolis, MN).

3.2.1.4.2 Role of Participant: DNA isolation

3.2.1.4.3 Attributes of the Participant: N/A

3.2.2 Genotyping Assay

3.2.2.1 Genotyping Assay Input: DNA

3.2.2.1.1 Input Amount: two aliquots of 250 ng DNA were used.

3.2.2.2 Genotyping Assay Process: SNP genotyping was performed with the commercial release of GeneChip 500K arrays. The genotyping assay was completed according to the manufacturer's instructions (user's manual is available at http://www.affymetrix.com/support/downloads/manuals/genome_wide_snp_5_0_manual.pdf).

In brief, two aliquots of 250 ng of DNA each are digested with NspI and StyI, respectively, an adaptor is ligated and molecules are then amplified, fragmented and labelled. Samples were denatured with an Applied Biosystems 2720 Thermal Cycler, then loaded on GeneChip 500K arrays and hybridized using GeneChip® Hybridization Oven 640. Samples were processed in 96-well plate format. Fluidics Station 450 was used to wash and stain arrays, then arrays

were scanned with GeneChip® Scanner 3000 7G. The scanned array images (.dat file) are ready for analysis after this process.

3.2.2.3 Genotyping Assay Output: scanned array image (.dat file).

3.2.2.4 Genotyping Assay Other Participants:

A complete list of reagents, instruments and software applied in the genotyping assay can be found in GeneChip 500k user's manual.

3.2.2.4 - I

3.2.2.4.1 Participant Identifier: Applied Biosystems 2720 Thermal Cycler (#4359659).

3.2.2.4.2 Role of the Participant: thermal cycler for denaturing DNA sample.

3.2.2.4.3 Attributes of Participant:

3.2.2.4.3.1 Parameter Settings of Instruments or Software: thermal program settings: 95°C 10 minutes; 49°C hold.

3.2.2.4 - II

3.2.2.4.1 Participant Identifier: GeneChip® Hybridization Oven 640 (Affymetrix)

3.2.2.4.2 Role of the Participant: microarray hybridization.

3.2.2.4.3 Attributes of Participant:

3.2.2.4.3.1 Parameter Settings of Instruments or Software: temperature 50°C, 60 rpm, rotation for 16-18 hours.

3.2.2.4 - III

3.2.2.4.1 Participant Identifier: Fluidics Station 450 (Affymetrix #00-0213)

3.2.2.4.2 Role of the Participant: wash and stain arrays.

3.2.2.4.3 Attributes of Participant:

3.2.2.4.3.1 Parameter Settings of Instruments or Software: operated by GeneChip® Operating Software (GCOS) 1.4 client.

3.2.2.4 - IV

3.2.2.4.1 Participant Identifier: GeneChip® Scanner 3000 7G (Affymetrix #00-0213)

3.2.2.4.2 Role of the Participant: scan array.

3.2.2.4.3 Attributes of Participant:

3.2.2.4.3.1 Parameter Settings of Instruments or Software: operated by GeneChip® Operating Software (GCOS) 1.4 client.

3.2.2.4 - V

3.2.2.4.1 Participant Identifier: Affymetrix GeneChip® Operating Software (GCOS) 1.4 client.

3.2.2.4.2 Role of the Participant: operate the fluidics station and the scanner.

3.2.2.4.3 Attributes of Participant:

3.2.2.4.3.1 Parameter Settings of Instruments or Software: default settings were used.

4 Data Transformation (Data Analysis)

4 – I

4.1 Data Transformation Input

4.1.1 Input Data Type: .dat file.

4.1.2 Input Data Amount: all SNPs genotyped.

4.2 Data Transformation Process: .dat files were processed with Affymetrix GeneChip® Operating Software (GCOS) to generate .cel files, which provides the intensities of the various probes on each chip. BRLMM (Bayesian Robust Linear Model with Mahalanobis distance classifier) was applied to determine genotype calls.

4.3 Data Transformation Output:

4.3.1 Output Data Type: genotype calls.

4.4 Data Transformation Other Participants: see table 2.

Table 2.

Participant Name	Function of the Participant	Attributes of Participant	
		Parameter Settings of Method or Software	Reference
BRLMM (Bayesian Robust Linear Model with Mahalanobis distance classifier)	Genotype calling algorithm	batch size was set as 96. Used default settings for other parameters.	[3], [4]
GenomeWideSNP_5 Annotations, release 22 (3/9/07)	GeneChip 5.0 annotation file	N/A	http://www.affymetrix.com/analysis/downloads/na22/genotyping/GenomeWideSNP_5.na22.annot.csv.zip
NCBI genome build 35	reference database for genotype calls	N/A	N/A

4 - II

4.1 Data Transformation Input

4.1.1 Input Data Type: SNP genotypes

4.1.2 Input Data Amount: all SNPs genotyped and all subjects enrolled.

4.2 Data Transformation Process: Six quality control filters were applied for sample exclusion: 1. SNP call rate < 97% (missingness). 2. Heterozygosity > 30% or < 23% across all SNPs. 3. External discordance with genotype or phenotype data (such as genotypes from another experiment, blood type or incorrect disease status). 4. Individuals identified as having recent non-European ancestry by the Multidimensional Scaling analysis. 5. Duplicates (the copy with more missing data was removed) 6. Individuals with too much identical-by-state (IBS) sharing (>86%); likely relatives.

4.3 Data Transformation Output:

4.3.1 Output Data Type: genotype of 1934 T1D case subjects and 2963 control subjects that passed the quality control filter. Exclusion summary is provided in table 3 below:

Table 3

Collection	Missingness	Heterozygosity	External discordance	Non-European ancestry	Duplicate	Relative	Total
58C	9	0	4	6	4	1	24
UKBS	8	0	5	14	0	15	42
T1D	7	2	1	18	6	3	37

4.4 Data Transformation Other Participants: see table 4.

Table 4

Participant Name	Function of the Participant	Attributes of Participant	
		Parameter Settings of Method or Software	Reference
missing genotype rate per sample	low DNA quality control	exclude samples with >0.3% missing rate	N/A
genome-wide heterozygosity	quality control for sample contamination	set empirical thresholds: exclude samples with heterozygosity > 30% or <23% across all SNPs.	N/A
discrepancies between WTCCC information and external identifying information	quality control for samples with external discordance with genotype or phenotype data	N/A	N/A
Multidimensional Scaling analysis	quality control for non-European ancestry	exclude samples that were clearly separate from the main cluster of WTCCC individuals.	N/A
checking duplicates	check for duplicate samples	exclude samples with >99% identity; the copy with more missing data was removed	N/A

genome-wide average identity by state	quality control for related subjects	exclude subject with IBS>86%	N/A

4 - III

4.1 Data Transformation Input

4.1.1 Input Data Type: SNP genotypes

4.1.2 Input Data Amount: all SNPs genotyped and all subjects enrolled.

4.2 Data Transformation Process: We sought to detect individuals with non-Caucasian ancestry using multi-dimensional scaling to provide a two-dimensional projection of the data whose axes represent geographic genetic variation. In the interest of computational efficiency and to avoid confounding of the multi-dimensional scaling by extended linkage disequilibrium we thinned the data to a set of 71,458 SNPs, within which no pair were correlated with r^2>0.2. For this set of nearly independent SNPs we computed genome-wide average identity by state (sum of the number of identical-by-state alleles at each locus divided by twice the number of loci) between each pair of individuals in each sample. We converted these identity-by-state relationships to distances by subtracting them from 1, and the matrix of pairwise identity-by-state values was used as input to multi-dimensional scaling. We excluded samples that were clearly separate from the main cluster of WTCCC individuals.

4.3 Data Transformation Output:

4.3.1 Output Data Type: genotype of all subjects except for the subjects that were clearly separate from the main cluster of WTCCC individuals, i.e., non-European ancestry.

4.4 Data Transformation Other Participants:

4.4.1 Participant Identifier/Name: Calculation of linkage disequilibrium.

4.4.2 Function of the Participant: to increase computational efficiency and to avoid confounding of the multi-dimensional scaling by extended linkage disequilibrium.

4.4.3 Attributes of Participant:

4.4.3.1 Parameter Settings of Method or Software: exclude SNP pairs correlated with r^2>0.2

4 - IV

4.1 Data Transformation Input

4.1.1 Input Data Type: SNP genotypes

4.1.2 Input Data Amount: all SNPs genotyped for all subjects.

4.2 Data Transformation Process: excluded 26,567 SNPs with a study-wide missing data rate >5%, or >1% for SNPs with a study-wide MAF<5%. We additionally excluded 4,351 SNPs with Hardy-Weinberg exact P value $<5.7x10^{-7}$.

4.3 Data Transformation Output:

4.3.1 Output Data Type: In total, 469,557 SNPs passed these quality control filters.

4.4 Data Transformation Other Participants:

4.4 - I

4.4.1 Participant Identifier/Name: study-wide missing data rate

4.4.2 Function of the Participant: Filtering out suboptimal markers.

4.4.3 Attributes of Participant:

4.4.3.1 Parameter Settings of Method or Software: exclude SNPs with study-wide missing data rate >5%, or >1% for SNPs with a study-wide MAF<5%.

4.4 - II

4.4.1 Participant Identifier/Name: Hardy-Weinberg equilibrium

4.4.2 Function of the Participant: Filtering out suboptimal markers.

4.4.3 Attributes of Participant:

4.4.3.1 Parameter Settings of Method or Software: filter out SNPs with Hardy-Weinberg exact P value$<5.7x10^{-7}$.

4 - V

4.1 Data Transformation Input

4.1.1 Input Data Type: SNP genotypes

4.1.2 Input Data Amount: 1934 T1D case subjects and 2963 control subjects passing all project quality control filters and, in addition, had MAF>1%.

4.2 Data Transformation Process: Genotype imputation: we used (1) the genotype data of this study, (2) the HapMap data, and (3) a population genetics model, to simulate genotypes at the HapMap SNPs that are not on the Affymetrix 500K chip. Informally, we determine which haplotypes are present in each individual in a region, and then use HapMap to 'fill in' these haplotypes at untyped SNPs.

4.3 Data Transformation Output:

4.3.1 Output Data Type: we imputed genotypes at 2,139,483 HapMap SNPs.

4.4 Data Transformation Other Participants:

4.4 - I

4.4.1 Participant Identifier/Name: Phase II HapMap CEU data.

4.4.2 Function of the Participant: reference data for genotype imputation.

4.4.3 Attributes of Participant:

4.4.3.1 Reference: http://hapmap.ncbi.nlm.nih.gov/

4.4 - II

4.4.1 Participant Identifier/Name: population genetics model.

4.4.2 Function of the Participant: model to impute genotype.

4.4.3 Attributes of Participant:

4.4.3.1 Reference: [5]

4 - VI

4.1 Data Transformation Input

4.1.1 Input Data Type: SNP alleles

4.1.2 Input Data Amount: 469,557 genotyped SNPs and 2,139,483 imputed SNPs of the 1934 T1D case subjects and 2963 control subjects that passed the sample and marker quality control filter.

4.2 Data Transformation Process: basic allelic association test was performed using standard Pearson chi-squared test with 1 degree of freedom.

4.3 Data Transformation Output:

4.3.1 Output Data Type: Genetic association analysis results.

4.3.2 Result of Genetic Association Analysis: results for each SNP for all analyses reported will be available from http://www.wtccc.org.uk. The SNPs significantly associated with T1D are provided in table 5 below.

4.4 Data Transformation Other Participants:

4.4.1 Participant Identifier/Name: Pearson chi-squared test

4.4.2 Function of the Participant: Statistical test of allelic genetic association with T1D

4.4.3 Attributes of Participant:

4.4.3.1 Parameter Settings of Method or Software: degree of freedom: 1; p-value cutoff: $<5x10^{-7}$.

4 - VII

4.1 Data Transformation Input

4.1.1 Input Data Type: SNP genotypes

4.1.2 Input Data Amount: 469,557 gentoyped SNPs and 2,139,483 imputed SNPs of the 1934 T1D case subjects and 2963 control subjects that passed the sample and marker quality control filter.

4.2 Data Transformation Process: standard genotypic test was performed using standard Pearson chi-squared test with 2 degree of freedom.

4.3 Data Transformation Output:

4.3.1 Output Data Type: Genetic association analysis results.

4.3.2 Result of Genetic Association Analysis: results for each SNP for all analyses reported will be available from http://www.wtccc.org.uk. The SNPs significantly associated with T1D are provided in table 5 below.

4.4 Data Transformation Other Participants:

4.4.1 Participant Identifier/Name: genotypic Pearson chi-squared test

4.4.2 Function of the Participant: Statistical test of genotypic genetic association with T1D

4.4.3 Attributes of Participant:

4.4.3.1 Parameter Settings of Method or Software: degree of freedom: 2; p-value cutoff: $<5x10^{-7}$.

Table 5. Markers showing significant association:

Marker ID	Significance Level		Allele		Effect Size		Genotype Difference Among Analysis Groups	
SNP	*P value*	*Genotypic P value*	*Risk allele*	*Minor allele*	*Heterozygote odds ratio*	*Homozygote odds ratio*	*Control MAF*	*Case MAF*
rs6679677	1.17x10-26	5.43x10-26	A	A	1.82 (1.59–2.09)	5.19 (3.15–8.55)	0.096	0.169
rs9272346	2.42x10-134	5.47x10-134	A	G	5.49 (4.83–6.24)	18.52 (27.03–12.69)	0.387	0.15
rs11171739	1.14x10-11	9.71x10-11	C	C	1.34 (1.17–1.54)	1.75 (1.48–2.06)	0.423	0.493
rs17696736*	2.17x10-15	1.51x10-14	G	G	1.34 (1.16–1.53)	1.94 (1.65–2.29)	0.424	0.506
rs12708716*	9.24x10-8	4.92x10-7	A	G	1.19 (0.97–1.45)	1.55 (1.27–1.89)	0.35	0.297

Reference:

1. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. 2007. *Nature.* 447(7):661-678.

2. Mueller PW, Rogus JJ, Cleary PA, et al. Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. 2006. J Am Soc Nephrol. 17(7):1782-90.

3. Rabbee, N. & Speed, T. A genotype calling algorithm for affymetrix SNP arrays. 2006. Bioinformatics 22:7-12.

4. Affymetrix. Technical report. BRLMM: an improved genotype calling method for the genechip human mapping 500k array set. 2006. Affymetrix

5. Marchini, J., Howie, B., Myers, S., et al. A new multipoint method for genome-wide association studies by imputation of genotypes. 2007. Nat. Genet. 39: 906-913.