

# Minimum Information about a **Genotyping** Experiment (**MIGen**)

Version 1 – May 2011

A standard for outlining the minimum information required to report the experimental details of genotyping experiments.

Jie Huang<sup>1</sup>, Daniel Mirel<sup>6</sup>, Elizabeth Pugh<sup>7</sup>, Chao Xing<sup>5</sup>, Peter N. Robinson<sup>8</sup>, Alexander Pertsemlidis<sup>2</sup>, LiangHao Ding<sup>3</sup>, Julia Kozlitina<sup>2</sup>, Joseph Maher<sup>2</sup>, Jonathan Rios<sup>2</sup>, Michael Story<sup>3</sup>, Nishanth Marthandan<sup>1</sup>, Richard H. Scheuermann<sup>1,5</sup>

<sup>1</sup>Pathology, <sup>2</sup>McDermott Center for Human Growth and Development, <sup>3</sup>Radiation Oncology, <sup>4</sup>Internal Medicine, <sup>5</sup>Clinical Science, U.T. Southwestern Medical Center, Dallas, TX, USA

<sup>6</sup>The Broad Institute Center for Genotyping and Analysis, Broad Institute of MIT & Harvard, Cambridge, MA, USA

<sup>7</sup>Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MA, USA

<sup>8</sup>Institut für Medizinische Genetik, Charité - Universitätsmedizin Berlin, Berlin, Germany

## Abstract

Genotyping experiments are widely used in clinical and basic research laboratories to identify associations between genotypes and normal/abnormal phenotypes. Genotyping assay techniques vary from PCR reactions in a test tube with gel electrophoresis detection to high throughput chip or bead array assays employing various technologies for automated genotype calls. [1-4] Depending on the experiment purpose and design, the resulting genotype data could involve millions of markers for thousands of individuals, requiring substantial data processing and analysis using various statistical, modeling and other methodologies. [5-8] To date, there is no standard for reporting genotyping experiments. The major challenge to developing such standard is how to come up with a specification that is generalized and flexible enough for reporting data from current and future genotyping experiment technologies. Here we present the Minimum Information about a Genotyping Experiment (MIGen) standard. MIGen recommends the standard information required to report a genotyping experiment, covering experiment design, subject recruitment, genotyping method, quality control procedures, and data analysis, that would serve as the experiment metadata intimately linked with the experiment results. A MIGen-compliant genotyping experiment report shall include all relevant information specified in the standard. The goal of MIGen is to set a minimum information reporting standard for adoption by the research community to facilitate consistent data interpretation and independent validation/reproduction [9], and to serve as guidance for database design for storing genotyping experiment data. MIGen is being developed as a collaborative project involving international domain experts and is a registered project under MIBBI: Minimum Information for Biological and Biomedical Investigations.

**Keywords:** genotyping, genetic association, minimum information, MIGen, MIBBI, check list, standard

# Table of Contents

## Introduction

Purpose of MIGen.....	3
Scope of MIGen.....	3
Terminology within this Document.....	3
Manufacturers, Models, and Products Mentioned within this Document.....	4
Vocabulary for Experimental Description.....	4
Units for Experimental Description.....	4
Reference in Experimental Description.....	4
MIGen Glossary of Terms.....	4

## MIGen Components

### 1. Experiment

<b>Overview</b> .....	6
-----------------------	---

- 1.1 Purpose
- 1.2 Features of the Genetic Variants Under Study
- 1.3 Key Words
- 1.4 Organization
- 1.5 Study Personnel
- 1.6 Date
- 1.7 Conclusions
- 1.8 Other Relevant Experiment Information

<b>2. Experiment Subjects Description</b> .....	7
---	---

- 2.1 Study Subject Common Selection Criteria
- 2.2 Study Subject Primary Characteristics
- 2.3 Study Subject Other Characteristics Captured

<b>3. Genotyping Procedure</b> .....	8
--------------------------------------	---

- 3.1 Genomic Variants (Genotyping Analyte) Description
- 3.2 Genotyping Processes Description

<b>4. Data Transformation (Data Analysis)</b> .....	13
---	----

- 4.1 Data Transformation Input
- 4.2 Data Transformation Process
- 4.3 Data Transformation Output
- 4.4 Data Transformation Other Participants

<b>References</b> .....	17
-------------------------	----

# Introduction

## ***Purpose of MIGen:***

The purpose of MIGen document is to specify the minimum information required to report a genotyping experiment, in order to facilitate consistent data interpretation and independent data validation/reproduction.

## ***Scope of MIGen:***

In MIGen a **genotyping experiment** is defined as a study that is designed to elucidate some aspect of the genomic nucleotide sequence structure of an individual or group of individual organism(s).

This document is not intended to specify the report format of the provided information or to address details in implementation of storage of the specified information in databases.

For reporting genotyping experiment for Quantitative Trait Locus (QTL) analysis please refers to MIQAS for minimum information standards.

## ***Terminology within this Document***

The key words “**shall**”, “**should**”, and “**may**” in this document are to be interpreted as described in RFC 2119 [10] and are also compatible with the IEEE Standards Style Manual [11].

The word **shall** is used to indicate mandatory requirements to be followed in order to conform to the standard and from which no deviation is permitted (**shall** equals *is required to*).

The word **should** is used to indicate that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required; or that (in the negative form) a certain course of action is deprecated but not prohibited (**should** equals *is recommended that*).

The word **may** is used to indicate a course of action permissible within the limits of the standard (**may** equals *is permitted to*).

The use of the word “relevant” in this document describes the condition by which information **should** be provided: “relevant” information is information that is necessary for correct understanding of the context of an experiment component.

## ***Manufacturers, Models, and Products Mentioned within this Document***

Manufacturer names, product names, etc. used within this document are intended for illustrative purposes only.

### ***Vocabulary for Experimental Description***

Where possible, experiment details **should** be described using terms from an ontology, controlled vocabulary or other appropriate standard, such as the Ontology for Biomedical Investigations (OBI) [12]; MeSH thesaurus [13], NCBI taxonomy [14], etc. and the source of the terms **should** be noted.

### ***Units for Experimental Description***

MIGen-compliant experiment descriptions **shall** include SI units. Use of SI metric units facilitates scientific communication, especially in international contexts. Descriptions **may** also provide alternative units of measure (e.g., Imperial units).

### ***Reference in Experimental Description***

Each time a database accession number, ID or name is used to describe a biological entity, one **shall** provide the name and build/version of the database (namespace) that serves as the accession number/ID/name source in the report. A hyperlink to the database website **should** also be provided if it is online accessible.

When a manufacturer's manual is referred in the report, the full name of the manufacturer, the full name and the catalog number of the product that the manual is applied to, as well as any version control of the manual document **shall** be provided. A hyperlink to the manufacturer's website where the manual could be accessed online **should** also be provided.

When referencing to previous publications, the investigators of the current report shall verify the information referred is MIGen compliance. Otherwise, missing information shall be added to the current report.

### ***MIGen Glossary of Terms:***

**Reagent:** A test substance that is added to a system in order to bring about a reaction or to see whether a reaction occurs (ID: "CHEBI:33893") [15].

**Planned Process:** A processual entity that realizes a plan which is the concretization of a plan specification (ID: "obo:OBI\_0000011") [12].

**Biomaterial Transformation:** Biomaterial transformation is defined as an event with one or more biomaterials as inputs and outputs [16].

**Assay:** An assay is a planned process with the objective to produce information about some evaluant ((ID: “obo:OBI\_0000070”) [5]. An evaluant role is a role that inheres in an entity that is realized in an assay in which data is generated about the bearer of the evaluant role.

**Data Transformation:** A data transformation is a protocol application which produces output data from input data (ID: “obo:OBI\_0200000”) [12].

**Independent Variables:** A directive information entity that is part of a study design. Independent variables are entities whose values are selected to determine their relationship to an observed phenomenon (the dependent variable) [12] (ID: “obo:OBI\_0000750”).

**Analyte:** Material (e.g., a substance or chemical constituent) plays the role of analyte in an experiment if it is the subject of interest of an analytical procedure within the experiment (i.e., it is being quantified in the experiment) [17].

## MI Gen Components

The following information **shall** be included in a MI Gen-compliant genotyping experiment description:

### 1. Experiment Overview:

This section provides a broad overview of the experiment.

**1.1 Purpose:** a brief description of the goal of the experiment. This **should** include the rationale and hypothesis.

#### 1.2 Features of the Genetic Variants Under Study:

**1.2.1 Types of Genomic Sequence Feature Variation(s) Assessed:** e.g., SNP, microsatellite length variation, gene deletion or insertion, copy number variations.

#### 1.2.2 Number Of Genomic Sequence Features Analyzed

**1.2.3 Selection Criteria:** **shall** describe how the genetic markers assayed in the experiment were chosen, e.g., genome wide SNP genotyping, SNPs within a certain range of a list of genes. Selection criteria **may** simply refer to a commercial product, e.g., Affymetrix GeneChip Human Mapping 500K Array Set (catalog # 900768 & 900770) [18].

#### 1.3 Keywords

The keywords should contain a short list of terms to describe the experiment.

**1.4 Organization(s):** the following **shall** be specified for the organization in which the experiment was performed:

##### 1.4.1 Organization Name

##### 1.4.2 Organization Address

**1.4.3 Role the Organization Played:** if it is multi-center collaborative work **shall** indicate the role the personnel of the organization played in the experiment, e.g., subject recruitment; genotyping assay; data analysis.

#### 1.5 Study Personnel

##### 1.5.1 Name

##### 1.5.2 Email Address

**1.5.3 Role:** e.g., principal investigator, experimenter, subject recruiter.

The above information for the primary contact person of the experiment **shall** be provided. **May** include information for additional individuals involved in the experiment.

**1.6 Date:** the date or time period during which the investigation was performed.

**1.7 Conclusions:** a brief summary of the interpretation of the results or outcome of the experiment.

**1.8 Other Relevant Experiment Information:** additional information about the experiment **shall** be provided if relevant. This **may** include genotyping result data availability, funding acknowledgement, related publications (which **should** be referenced by PMID), URLs, databases, etc.

**2 Experiment Subjects Description:** summary level information about the subjects from which the biological specimens were obtained **shall** be described.

**2.1 Study Subject Common Selection Criteria:** the subject common selection criteria are the characteristics that were applied to decide whether an individual could be enrolled in the study. These criteria are applied in general to ALL study subjects during the subject enrollment process, regardless of their study group assignment.

**2.1.1 Subject Sampling Method:** shall indicate the study design for study subject selection, e.g., family trios (consisting of an affected child and both parents); a cohort design; a case-control design. Shall indicate whether it is random sampling from a given population.

**2.1.2 Enrollment Inclusion and Exclusion Criteria:** **shall** list the criteria used for selecting the eligible experiment subjects, e.g., only women between the ages of 17 to 50 years were included in the study; offspring of The Jackson Laboratory, BKS.Cg-Dock7<sup>m</sup> +/- Lepr<sup>db</sup>/J mice.

**2.1.3 Methods for Ascertaining Enrollment Criteria,** e.g., a study of H1N1 influenza infection shall indicate whether any tests were done to differentiate it from other influenza virus infection.

**2.2 Study Subject Primary Characteristics:** shall describe the key characteristics collected for the current study design that were to be used as grouping criteria in data analysis, i.e., the phenotype or trait that is study of interest. These characteristics are usually used to distinguish between cases and controls and/or serve as the main independent variable in the analysis of the study results. For example, in a case-control study on SLE

patients versus control subject, SLE diagnosis will be the primary characteristics evaluated through the study design; in a population prospective study on myopericarditis following smallpox vaccine, myopericarditis will be the primary characteristic evaluated. As opposed to the “common criteria” described in section 2.1, not all subjects enrolled in the study will bear the study subject primary characteristic.

**2.2.1 Name of the Characteristics Evaluated:** e.g. systemic lupus erythematosus, myopericarditis elicited by smallpox vaccination, schizophrenia.

**2.2.2 Methods/Criteria for Evaluating the Characteristics:** e.g., meet the CDC case definition for myopericarditis (shall provide reference to CDC definition here) and occurring within 30 days following smallpox vaccination.

**2.2.3 Number of Subjects:** shall report number of subjects with each of the primary characteristics.

**2.3 Study Subject Other Characteristics Captured:** **should** describe each of the additional study subject characteristics that were captured during the experiment, which could be used to sub-divide the study subjects into groups during secondary data analysis or to help interpret the experiment results.

**2.3.1 Name of the Characteristics Captured:** e.g., age, race, subject’s blood pressure, SLE score, smoking status, cholesterol levels.

**2.3.2 Value of the Characteristics:** the value **may** be reported as summary statistics for each of the study subject sub-groups defined in the study design, e.g., mean and standard deviation of study subject’s age in the case and control groups.

**2.3.3 Method Used to Capture the Characteristics:** for standard methods, **shall** provide the name of the method being used to measure the subject characteristic and reference range (normal range), if applicable. For non-standard methods, a detailed description **should** be provided.

### 3. Genotyping Procedure

**3.1 Genomic Variants (Genotyping Analyte) Description:** the following information **shall** be provided to describe each of the genomic variants assayed in the genotyping procedure. Genomic variants description **may** refer to an annotation file, in which case the annotation file and version (if applicable) **shall** be provided as reference.

**3.1.1 Genomic Variants with Record in Public Database:** if the genomic variant genotyped has a record maintained in a public database the following information **shall** be provided.

**3.1.1.1 Identifier:** e.g., SNP ID rs73585716.

**3.1.1.2 Reference Database:** **shall** provide the name and build/version of the reference database in which the genomic variant records were maintained and referred to, e.g. dbSNP 131 [19].

**3.1.2 Genomic Variants Without Public Database Records:** e.g., novel genomic variants whose sequence is to be investigated, or the experiment was designed to detect the genotypes of genetically manipulated genome or genomic region, e.g., transgenic mice, etc. **Shall** report the following:

**3.1.2.1 Genomic Location:** e.g., screen novel SNPs in the IBD6 locus on chromosome 19.

**3.1.2.2 Sequence of Genomic Variant(s):** if applicable by study design, **should** report the expected sequence(s) of the variant(s), strand direction and/or number of base-pairs, e.g., the expected sequence or size of the transgene. The sequences of genomic variants should be named according to current HGVS standards[20].

**3.1.2.3 Reference Database:** **shall** provide the reference database name and build/version to which the sequence of genomic variants were aligned.

## **3.2 Genotyping Processes Description::**

The genotyping procedure starts with the specimen collection process from experiment subject and ends with the process whereby the raw data from genotyping experiment was obtained.

Genotype procedure may be described as a single composite assay process or as sequentially-linked *planned processes*, which might include *biomaterial transformation* and *assay processes*. The investigators **should** use their best judgment to break down the genotyping procedure into smaller steps if they feel that this would be helpful or necessary, so that the report is not labor intensive to generate and all the key information required to reproduce the procedure is reported.

When describing the genotyping processes, one **shall** refer to the appropriate minimum information document if one exists for a specific

experiment technique involved in the genotyping procedure, e.g., **shall** refer to the MIFlowCyt [17] standard if a flow cytometry technique was used; **shall** refer to the MIAME [21] standard if a microarray technique was used.

A manufacturer's manual or other published resources **may** be referred to, as long as the information of the referred resource is MIGen compliant. **Shall** provide the original resource or information to locate the original resource.

For EACH of the genotyping processes described, one **shall** include the following information:

**3.2.1 Biomaterial Transformation:** if the planned process is a biomaterial transformation process one **shall** report the following:

**3.2.1.1 Biomaterial Transformation Input:** the material whose quality/quantity is to be manipulated in the process, and whose processed product is of interest to be carried on into subsequent processes, e.g., blood as input to DNA extraction; DNA as input to PCR amplification; DNA template as input to probe labeling.

**3.2.1.1.1 Type of the Input Material:** e.g., study subject, blood, cell, DNA.

**3.2.1.1.2 Amount of the Input Material:** **shall** report the input amount in value-unit pairs.

**3.2.1.1.3 Other Attributes:** **should** report other key information about the input characteristics if they are critical to interpret or reproduce the result, e.g., sequence of the DNA used as probe template, how the samples were grouped for genotyping (e.g. were cases and controls grouped randomized across lab batches or run separately,) fresh blood sample vs. cryopreserved cell lines.. Additional samples such as hapmap controls, known duplicates of investigator samples)

**3.2.1.2 Biomaterial Transformation Process:** shall describe how the output material was generated, given the input material and other participants (e.g. specific reagents). Shall provide the purpose of the process, e.g., DNA isolation, DNA purification, quality control for biological sample mix-up. May refer to experiment protocols.

**3.2.1.3 Biomaterial Transformation Output:** the material coming out of the process that will be carried on into the next process.

**3.2.1.3.1 Type of Output,** e.g., DNA as output of DNA extraction; PCR amplicons as output of PCR amplification; radioactive labeled probe as output of probe labeling process in Southern genotyping experiment.

**3.2.1.3.2 Attributes of Output:** e.g., quantity of DNA, <sup>32</sup>P labeling

**3.2.1.4 Biomaterial Transformation Other Participants:**

Other participants of biomaterial transformation should be described if considered to be critical in the interpretation of the results and may include critical instruments, software operating the instruments, reagents, and consumables used in the process. For each other participants reported, shall included the following:

**3.2.1.4.1 Participant Identifier:** Shall report the information that will uniquely identify the participant, e.g., reagent name with manufacturer and catalog number; instrument name with manufacturer and catalog number; name of software with manufacturer and its version number if version control exists for the participant.

**3.2.1.4.2 Role of Participant:** the function of the participant in the process shall be indicated, e.g., centrifugation, incubation, reagent for DNA extraction, etc.

**3.2.1.4.3 Attributes of the Participant:**

**3.2.1.4.3.1 Parameter Settings of instruments or software:** e.g., temperature settings of incubator. May indicate that the default parameter settings were used.

**3.2.1.4.3.2 Amount of Participant Reagent:** for reagent participants shall specify the amount used in the process in value-unit pair. The amount shall be reported as or shall be able to be

inferred as final concentration or weight, e.g., dNTP final concentration 200µM or 2.5µl 2mM dNTP per 25µl reaction.

### 3.2.2 Genotyping Assay

**3.2.2.1 Genotyping Assay Input:** the material used as the source for the genetic material evaluated in the assay, which is often the nucleic acid or transformed nucleic acid that was generated from a biomaterial transformation process. **Shall** report the followings if applicable:

**3.2.2.1.1 Input Amount:** **shall** report the input amount in value-unit pair.

**3.2.2.2 Genotyping Assay Process:** shall describe how the genotyping raw data was generated, given the input material and other participants. May refer to experiment protocols.

**3.2.2.3 Genotyping Assay Output:** the type of genotyping output data **shall** be described, e.g., size of the DNA fragment from gel electrophoresis; the image of emitted florescence; DNA sequence trace chromatogram after capillary electrophoresis; .fcs file generated from flow cytometry analysis, etc.

**3.2.2.4 Genotyping Assay Other Participants:** other participants include critical instruments, reagents, software, and consumables applied in the genotyping assay process. May only report the participants that play key role in generation of the genotyping data, and that will potentially influence the reproduction of genotyping results if not specified..

**3.2.2.4.1 Participant Identifier:** Shall report the information that will uniquely identify the participant, e.g., instrument name with manufacturer and catalog number; name of software and its version number if version control exists for the participant. for example, for genotype assay using Affymetrix GeneChip [18] one **may** report the reference to Affymetrix GeneChip user's manual where key information of the following critical participants can be found:

*Reagents:* R-Phycoerythrin Streptavidin (Vector Laboratories, #BA0500) & Anti-streptavidin antibody (goat), biotinylated (Molecular Probes; #S866);

*Instruments:* GeneChip® Fluidics Station 450 and Affymetrix GeneChip® Scanner 3000 7G (#00-0213);

*Software:* Affymetrix GeneChip® Command Console (AGCC) [18];

*Consumable:* RNase-free, microcentrifuge vials.

**3.2.2.4.2 Role of the Participant:** **shall** indicate the role the other participants played in the assay process, e.g., fluorescence scanner, DNA size standard, etc.

**3.2.2.4.3 Attributes of Participant:**

**3.2.2.4.3.1 Parameter Settings of Instruments or Software:** **should** including critical parameters settings for algorithms or software applied in the assay. **May** indicate that no parameter adjustment was done, i.e., the default parameter settings were used.

**3.2.2.4.3.2 Amount of Participant Reagent:** for reagent participants **shall** specify the amount used in the process in value-unit pairs. The amount **shall** be reported as or **shall** be able to be inferred as final concentration or weight.

## **4 Data Transformation (Data Analysis)**

Data analysis in genotyping experiment is often a sequence of *data transformation* processes starting with raw genotype results and ending with the final conclusions that address the study design hypothesis.

Data transformation processes include the generation of final genotype calls from genotype assay raw data, quality control of genotyping results, data formatting/re-formatting, as well as statistical analysis of genotyping results. Investigators **should** use their best judgment to decide if and how the data

analysis procedure should be broken down into sequential data transformation processes so that all the key information for interpreting and reproducing the experiment result is reported without making the reporting process too labor intensive. For EACH of the processes involved in the data analysis procedure one shall report the following information:

#### **4.1 Data Transformation Input**

**4.1.1 Input Data Type:** e.g., cell signal intensity (.cel file), marker genotypes and subject grouping information. **May** report data value-unit pairs.

**4.1.2 Input Data Amount:** shall indicate if the whole dataset or a subset of the genotype result was used, in terms of both subjects and markers, e.g., all 550,000 markers genotyped for all case and control subjects were used in the statistical analysis; 32,000 markers didn't pass the data quality controls and were excluded from the statistical data analysis, 2 case subjects and 5 control subjects were excluded because admixture analysis conflicted with reported race values.

**4.2 Data Transformation Process:** shall describe the procedure of how the output data were produced from input data together with other participants. **May** refer to other resources, such as protocol documents, previous publications, etc... Shall state the purpose of the data transformation process, e.g., generation of signal intensity value for each microarray feature; calculation of signal intensity for each transcript; statistical test for genotype disease association, quality control for genotype calls, imputation, etc.

#### **4.3 Data Transformation Output:**

**4.3.1 Output Data Type:** e.g., transcript signal intensity (.chp file), final genotype calls, list of significant markers and their related information as result of genetic association analysis (see next section).

If it is a data quality control process shall report summary statistics of how the input data modified, as well as descriptive statistics on the remaining data from the data quality point of view, such as call rate, duplicate error rate, etc. were modified.

**May** report data as value-unit pairs.

**4.3.2 Result of Genetic Association Analysis:** if genetic association analysis is performed, for each significant result, shall report the following::

##### **4.3.2.1 Marker ID or Genomic Location**

**4.3.2.2 Allele:** the allele that was found to be significantly associated with the trait of interest, e.g., minor allele, over transmitted allele.

**4.3.2.3 Genotype Difference Among Analysis Groups:** e.g., minor allele frequency in the case and control group; ratio of transmitted to untransmitted.

**4.3.2.4 Significance Level:** e.g., p-value.

**Note:** the process of genetic association analysis includes nested data transformation process, e.g., the calculation of minor allele frequency, odds ratio. These data transformation processes are often simple and well-known procedures, so it is the investigator's judgment to whether report them as separate data transformation processes.

**4.4 Data Transformation Other Participants:** other participants include critical software, methods/algorithms, and reference data applied in the data transformation processes.

**4.4.1 Participant Identifier/Name:** Shall report the information that will uniquely identify the participant, e.g., name of software with manufacturer and its version number if version control exists.

**4.4.2 Function of the Participant:** shall indicate the goal or aim accomplished by the participant in the data transformation process, e.g. algorithm for SNP allele calls, probe annotation file, statistical test for genetic association, etc.

**4.4.3 Attributes of Participant:** should apply the following if applicable.

**4.4.3.1 Parameter Settings of Method or Software:** should including critical parameters settings for algorithms or software applied in the data transformation. May indicate that the default parameter settings were used.

**4.4.3.2 Reference:** if a detailed description of the participant is published or publicly available, should indicate its reference.

## References:

1. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447(7145):661-78.
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W and others. Global variation in copy number in the human genome. *Nature* 2006;444(7118):444-54.
3. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;26(10):1135-45.
4. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. *Annu Rev Biomed Eng* 2007;9:289-320.
5. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* 2007;2(10):2492-501.
6. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet* 2009;10:387-406.
7. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7(10):781-91.
8. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ and others. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*;34(6):591-602.
9. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE and others. Replicating genotype-phenotype associations. *Nature* 2007;447(7145):655-60.
10. Bradner S. Key words for use in RFCs to Indicate Requirement Levels. 1997.
11. IEEE. IEEE Standards Style Manual.
12. The Ontology for Biomedical Investigations.
13. United States National Library of Medicine, National Institutes of Health. Medical Subject Headings (MeSH).
14. National Center for Biotechnology Information (NCBI), Entrez Taxonomy.
15. Chemical Entities of Biological Interest (ChEBI)
16. Kong YM, Dahlke C, Xiang Q, Qian Y, Karp D, Scheuermann RH. Toward an ontology-based framework for clinical research databases. *J Biomed Inform*;44(1):48-58.
17. Lee JA, Spidlen J, Boyce K, Cai J, Crosbie N, Dalphin M, Furlong J, Gasparetto M, Goldberg M, Goralczyk EM and others. MIFlowCyt: the minimum information about a Flow Cytometry Experiment. *Cytometry A* 2008;73(10):926-30.

18. Affymetrix. <<http://www.affymetrix.com/estore/>>.
19. National Center for Biotechnology Information (NCBI), SNP database (dbSNP).
20. Human genome Variation Society. <<http://www.hgvs.org/mutnomen/>>.
21. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC and others. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29(4):365-71.