MGeNDデータ登録説明

※ 2018年10月31日・11月28日・12月20日にAMEDにて開催した説明会資料です

目次

データ登録ポリシー（ELSI対応済）

登録可能なデータ

以下のいずれかを満たすもの

- ◆ 本事業のために、患者の同意を得た上で、取得されたバリアントデータ
- ◆ 本事業以外で取得されたバリアントデータで、以下の内容のいずれかが説明同意文書に入っているもの

（１）解析結果をデータベースに登録すること

（２）解析結果を学術論文で公表すること

- ◆ 学術論文や研究班の報告書、公的データベースなど、学術的に信頼された媒体ですでに発表されているバリアントデータ

バリアントデータの定義

- ◆ 疾患名（標準的な記載方法のもの）
- ◆ 遺伝子名
- ◆ 遺伝型情報（Genotype)
- □ １～数箇所程度のSNV・SNPまたはp値<10^{-4} のSNP すべて
- □ 年齢（層）
- □ 性別（「不明」「混合」等を含む）

データ登録のための倫理対応およびガイダンス

別紙「MGeNDデータ利活用ガイダンス整理表」　20180315 (ver.7)

ガイダンス番号/データベース種別/利用するデータ項目名	ガイダンス項目	提供元側			DB側		
		提供時のデータの状況	個人情報該当性	手続き等の対応	受領時のデータの状況	個人情報該当性	手続き等の対応
① 統合DB MGeND (非制限公開で疾患バリアントを公開)* 「疾患名、遺伝子名、1～数箇所程度のSNP等の遺伝型、(年齢(層))、(性別(不明・混合を含む))」**	A	元データ(診療録や研究データ等)が存在するバリアントデータ。 (いわゆる連結可能匿名化されているデータ、あるいは、連結不可能匿名化されているが照合可能な診療録等が施設内に存在しているデータ。)	提供側施設で、個人情報に該当する。	・本人同意を得ている場合(1): 本DBへの提供を明記した研究計画書や説明同意文書が倫理審査委員会に承認され、本人に説明を行って、同意を得たとき。 →提供可能。 ・本人同意を得ている場合(2): 既存の研究計画書において、データを「データベースに登録する」、成果を「論文等で発表する」等の同意をすでに得ているとき。 →提供可能(新たな倫理審査は不要)。 ただし、超希少疾患等でデータ自体から個人識別性が十分に除けないときは、下記の「同意を得ていない場合」と同様の扱いとする(再同意または、倫理審査委員会の承認を得てオプトアウトのいずれかを行えば提供可能。)。 ・本人同意を得ていない場合: データベース提供や論文発表等によるデータ利用の同意を得ていないとき、診療目的のみで取得し研究同意を得ていないとき。 →再同意、または、再同意困難であれば、非個人情報化(1例情報を誰のものか全く分からなくすること)、あるいは個人識別性が低減され特段の理由がある場合通知・公開(倫理審査要)、もしくはオプトアウト(倫理審査要)によって提供可能(ゲノム指針15(2)ア～ウ)	対応表は受け取らず、以降も対応表を要求しない。 (提供元で付番した個人識別可能な符号・番号(ID等)は、バリアントデータに含まれていても構わない。) ※⑤のDBのデータも同施設に存在する場合がある(→バリアントデータのID等で連結できる場合がある)	多くの場合、データベース側施設で、個人情報に該当する。 (提供元側施設で作成し保有する対応表との間で、照合性が断たれているとはいえないため。 また、⑤のDB用データが存在する場合であってデータに個人識別符号が含まれる場合、それとの照合性によっても、個人情報に該当する。)	提供元施設から対応表を受け取らず、⑤のDB用の個人識別符号を含むデータとの照合性等を適切に管理することとして、特段の理由＋指針に定められた事項を公開することによって、国内提供(ゲノム指針15(2)イ)および国外提供(ゲノム指針11(4)ア(ウ))を行う。(※倫理審査要:ゲノム指針15(1)(2)) 「疾患名、遺伝子名、1～数箇所程度のSNP等の遺伝型、(年齢層)、(性別)」のみを公開する。受領時のデータに含まれるIDやその他の情報は公開しない。
	B	もともと個人の情報を含まないバリアントデータ。 (すでに複数人のデータがまとめられた統計情報になっている、そもそも連結不可能匿名化され誰のものかわからない情報であって個人識別符号を含まないものである、知識データである等。)	もともと個人情報でない。	同意なく提供可能。倫理審査不要。	提供元側に同じ。	データベース側施設でも、個人情報ではない。	・Bだけを収集する場合、特に必要な対応はない。 ・Aの収集もあわせて行う場合、Aの対応でよい。
	C	学術的な価値が定まり、研究実績として十分に認められ、研究用に広く一般に利用され、かつ、一般に入手可能なバリアントデータであって、元データと紐づかないもの。	個人情報に該当しない。	同意なく提供可能。倫理審査不要。 (ゲノム指針22(1)により指針適用範囲外のため。)	提供側元に同じ。	データベース側施設でも、個人情報に該当しない。	・Cだけを収集する場合、特に必要な対応はない。 ・Aの収集もあわせて行う場合、Aの対応でよい。

＊①以降のデータベース(DB)には、②AGD非制限公開DB、③AGD制限公開DB、④AGD制限共有DB、⑤共同研究DB、(全て仮称)がある。本表では①のみを説明している。
＊＊これを本表では「バリアントデータ」とよぶ。

大阪大学加藤和人先生・山本先生作

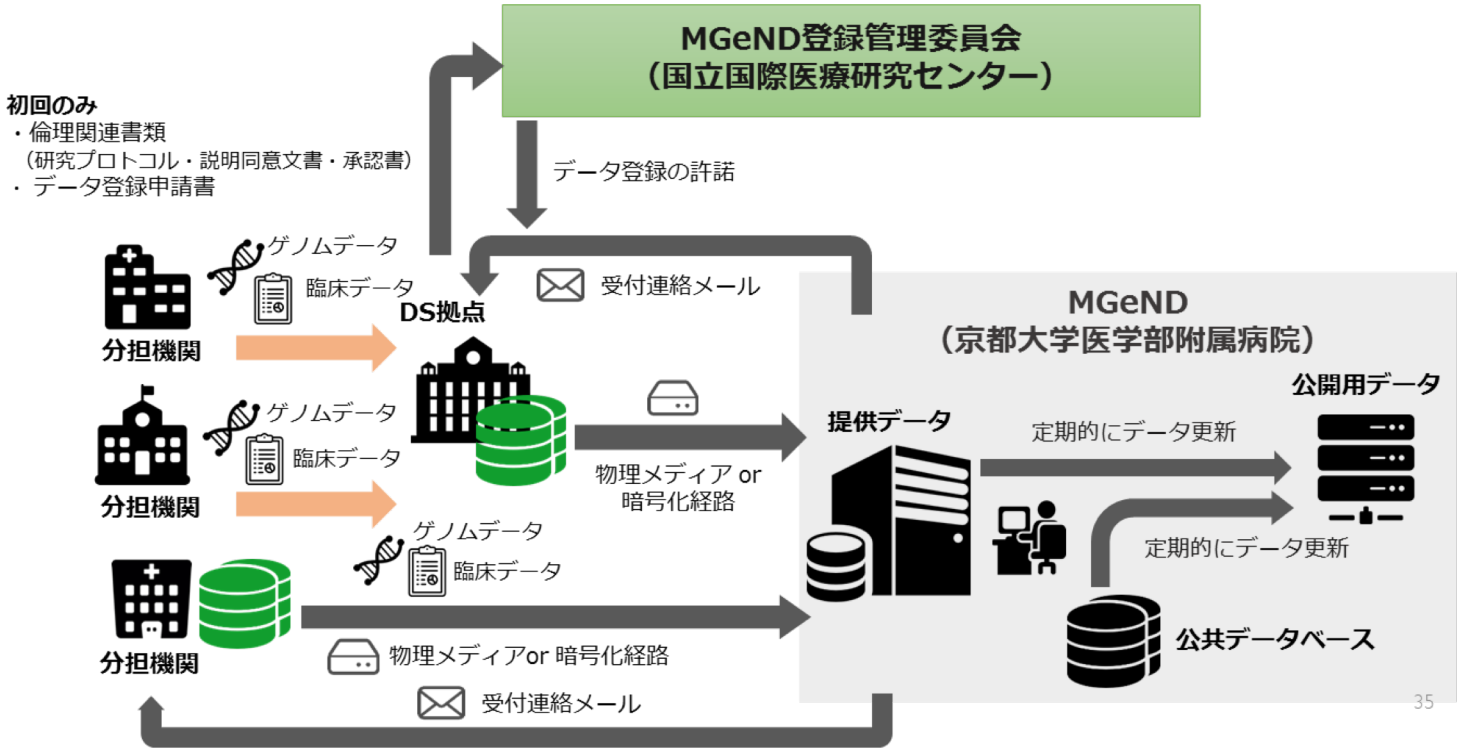
DSからのデータ登録の流れ

データ受付の手順

1. データ登録機関からMGeND登録管理委員会へデータ登録申請書送付
2. MGeND登録管理委員会による倫理関連書類の確認
3. MGeND登録管理委員会によるデータ登録の承認

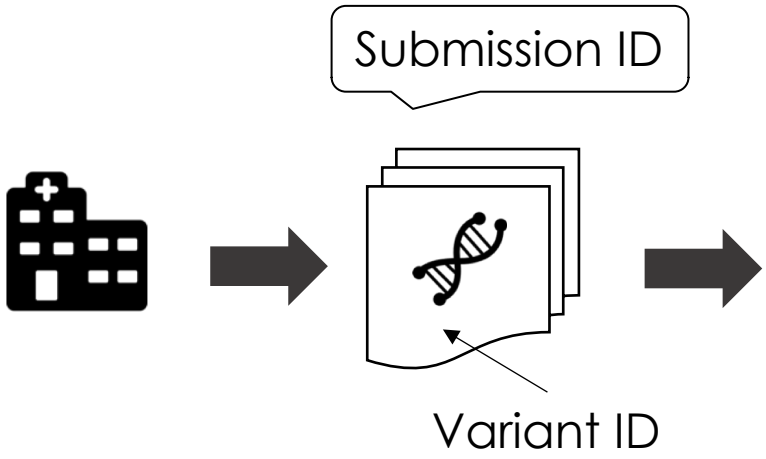
（1から３は初回のみ；ただし異なる研究計画によるデータは毎回必要）

4. DS拠点・データ登録機関からMGeND(京都大)へデータ送付
5. MGeNDにて公開データベースにデータ登録

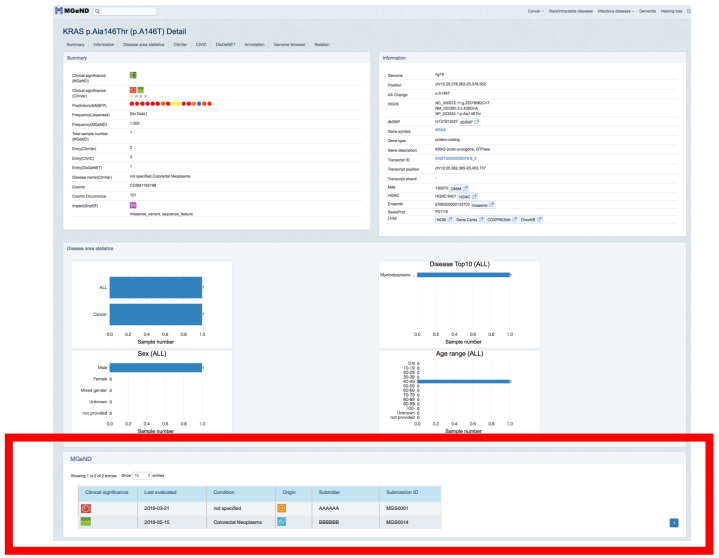


データ登録元情報について

- 現在のMGeNDでは登録者情報を公開していませんが、論文投稿に際し、提供データを参照するためのIDが必要ではないか、との声をただきました。
- そこで Submission ID の発行およびMGeND上での登録者情報の公開を予定しております。
- ただし、提供元が明らかになることによる個人情報保護上の問題も存在することから、提供元機関の判断で非公開とすることが可能です。



※イメージ



MGeND独自の疾患ゲノムデータフォーマット

全DSの疾患に対応できる登録フォーマット・項目を策定

- **変異データ用**（検体単位・検体集団単位）
  - ClinVar登録フォーマットに準拠した項目の設定
  - 個人情報に考慮した統計値（頻度）での登録にも対応
  - TSV/XML/VCF形式に対応
- **GWASデータ用**
  - GWAS解析の共有に必要な項目を策定
  - Xlsx形式に対応
- **HLAタイピングデータ用**（群間比較・集団単位）
  - HLAタイピング解析で標準的に用いられるテーブル形式を踏襲
  - Xlsx形式に対応

前提条件

- データ登録には Study code および DS code が必要となります
  - Study code および DS codeは登録者情報の登録時に発行
  - 登録者は研究代表者
- 各項目名は「アルファベット、数字、アンダーバー、ハイフン」で構成します。
- 原則として日本語（マルチバイト文字）の記載は受け付けておりません。
  ただし、 提供元のDSに関する情報を記載する箇所に関しては一部
  日本語（マルチバイト文字）での記載を認めております。
- 文字コードは基本的にUTF-8ですが、Excel に関しては Shift_JIS とします。
- Excelでは、数式の使用および規定のsheet以外の使用は禁止とします。
- 必須項目が未記入の場合は登録できません。
  - 提供データの登録を中断し、修正依頼等をさせていただきます。
  - 上記の回答をいただくまで登録対象外とさせていただきます。

※ 備考：提供データの修正依頼はMSSからをメールでご連絡させていただくことがあります。
この場合、修正結果を修正依頼の返信に添付されますと、運用規則の範囲外にデータが流出したことになります。そのため、修正依頼の返信に添付することをひかえていただければと存じます。

全フォーマット共通項目

登録機関情報

項目タイプ	項目	説明	形式	必須（※1）	公開（※2）	複数記載（※3）	備考
プロジェクト（DS事業）情報	STUDY_CODE	研究課題コード	String	○	×	×	各DSごとに付与 事前に発行
	STUDY_NAME	研究の名称	String	○	○	×	研究課題名
	DS_CODE	DS登録拠点コード	String	○	×	×	代表機関以外からも登録がある場合、１つの拠点に対しても複数のDS拠点コードを発行（パターン３） 事前に発行
	PROJECT_NAME	プロジェクトの名称	String	×	×	×	登録機関でのプロジェクト名がある場合（パターン３）
	REVIEW_STATUS	研究に関する説明	String	○	×	×	
研究課題代表 / 登録 機関情報	SUBMITTER_TYPE	代表機関か解析（登録）機関か	String	○	×	×	代表機関/分担機関
	SUBMITTER_NAME	登録者氏名	String	○	×	×	
	SUBMITTER_PHONE	電話番号	Numeric	○	×	×	
	SUBMITTER_EMAIL	Emailアドレス	String	○	×	×	
	SUBMITTER_ROLE	組織での職名・役割	String	○	×	×	
	ORGANIZATION	機関名	String	○	△	×	原則として公開推奨（希少疾患などにおいては提供機関の判断に委ねる）
	STREET	住所	String	○	×	×	
	CITY	市町村	String	○	×	×	
	PREFECTURE	都道府県	String	○	×	×	
	COUNTRY	国	String	○	×	×	
	POSTAL_CODE	郵便番号	Numeric	○	×	×	
解析機関情報	ORGANIZATION_SUBMITTER_NAME	担当者氏名	String	×	×	×	必要のある拠点のみ （パターン１のDS分担機関を想定）
	ORGANIZATION_SUBMITTER_PHONE	電話番号	Numeric	×	×	×	
	ORGANIZATION_SUBMITTER_EMAIL	Emailアドレス	String	×	×	×	
	ORGANIZATION_SUBMITTER_ROLE	組織での職名・役割	String	×	×	×	
	ORGANIZATION_ORGANIZATION	機関名	String	×	×	×	
	ORGANIZATION_STREET	住所	String	×	×	×	
	ORGANIZATION_CITY	市町村	String	×	×	×	
	ORGANIZATION_PREFECTURE	都道府県	String	×	×	×	
	ORGANIZATION_COUNTRY	国	String	×	×	×	
	ORGANIZATION_POSTAL_CODE	郵便番号	Numeric	×	×	×	
公開時期	HOLD_RELEASE	すぐに公開するかどうか	String	○	×	×	Hold、 Release immediatelyのどちらかを記載 Holdの場合、下記のRELEASE_DATEを超えるまで登録処理で対象外となります。
	RELEASE_DATE	公開日	Date (yyyy/mm/dd)	△	×	×	Holdの場合は必須となります。 最大で解析実施から2年後までを設定可能です。

全フォーマット共通項目

疾患名情報

項目タイプ	項目	説明
疾患名・診断名	CONDITION_ID_TYPE	コンディションIDタイプ
	CONDITION_ID_VALUE	上記で選択したDB/OntologyのID
	CODE_TYPE	ICD10などのコードIDタイプ
	CODE_VALUE	コードID
	PREFERRED_CONDITION_NAME	特定のDB/Otologyを使用しない場合の疾患名（非推奨）

CONDITION_IDとして下記に対応しています

CONDITION_ID_TYPE	CONDITION_ID_VALUE
OMIM	Phenotypic Series Number
MeSH	Unique ID
MedGen	UID
Orphanet	OrphaNumber
HPO	ID（HP:で始まるもの）

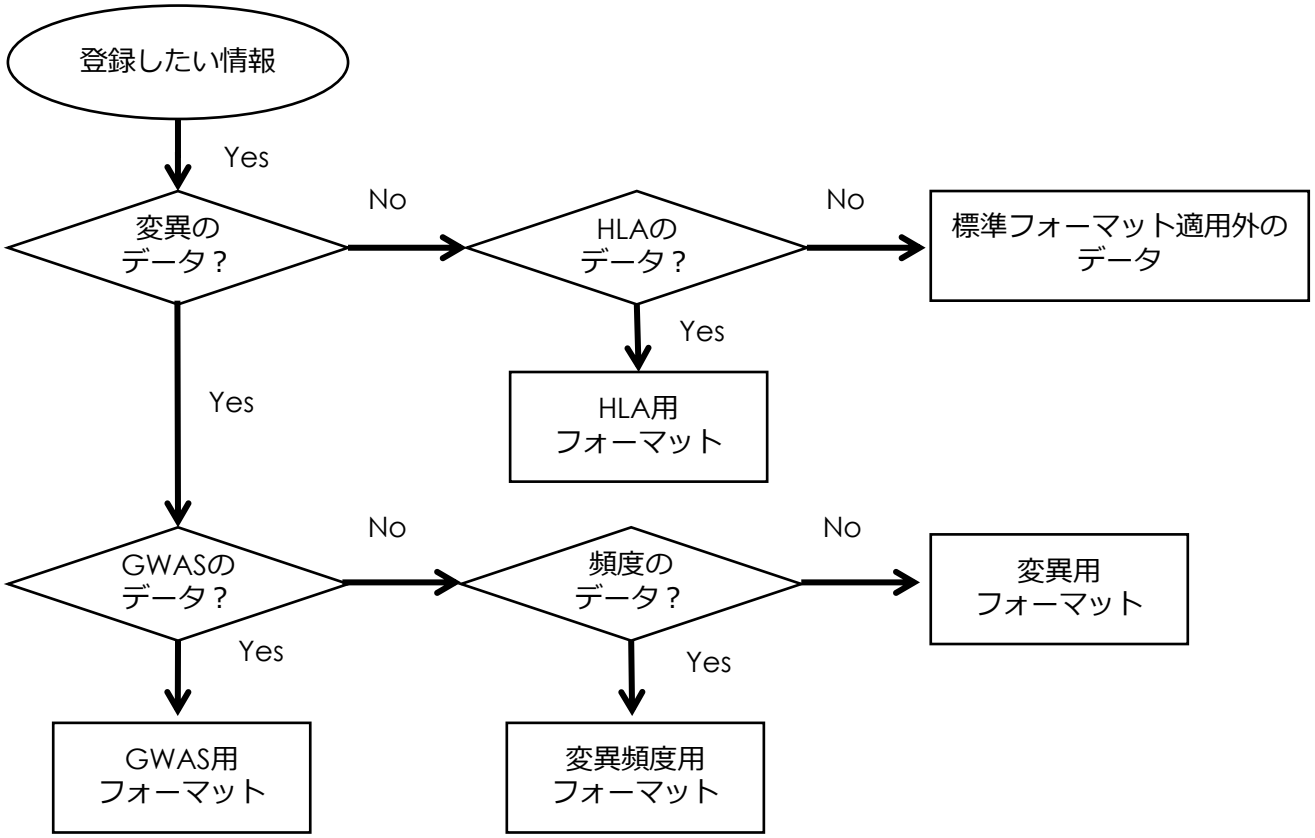
CODE_TYPEは現状ではICD10のみ記載できます

CODE_TYPE	CODE_VALUE
ICD10	WHOの表記のコード e.g. C15.1

標準フォーマット

- 標準フォーマットに記載可能なデータは、

  変異・変異頻度・GWAS・HLA のいずれかになります。



変異・変異頻度用フォーマット（基本事項）

- 登録対象は SNV / short INDEL / 構造変異 情報です。
- **頻度（検体集団単位）での登録**と**検体単位での登録**が可能です。

※ 希少疾患等で個人が特定できる可能性がある場合は、
提供元の判断で該当情報を省略してください。

- 登録用の標準フォーマットとして以下を用意
  - Excel 形式
  - TSV 形式
  - XML 形式
  - VCF 形式 （変異単位での登録のみ）
- Excel 形式・TSV形式・XML形式は全て同じ情報が記載可能
- VCF 形式では構造変異・頻度での登録は対象外
- 同一項目の複数記載は不可です。
- 記載する情報がない場合は対象箇所を**ブランク**にしてください。

変異・変異頻度用フォーマット（基本事項）

項目タイプ	項目	説明
配列名	ASSEMBLY_NAME	参照配列名
遺伝子名	GENE_SYMBOL	HGNC Gene Symbol
	REFERENCE_SEQUENCE	HGVS表記に使用した参照配列
変異情報	HGVS	変異のHGVS表記
	CHROMOSOME	染色体
	START	変異の開始位置
	STOP	変異の終了位置
	REFERENCE_ALLELE	Reference アリル
	ALTERNATE_ALLELE	変異アリル
	REREFENCE_ALLELE_COUNT	Referenceアリル数
	ALTERNATE_ALLELE_COUNT	変異アリル数
	VARIANT_TYPE	構造変異のタイプ or Fusion
構造変異情報	OUTER_START	構造変異の outer start ポジション
	INNER_START	構造変異の inner start ポジション
	INNER_STOP	構造変異の inner stop ポジション
	OUTER_STOP	構造変異の outer stop ポジション
	VARIANT_LENGTH	構造変異の長さ
	COPY_NUMBER	コピー数（for copy-number-loss/gain）
	REFERENCE_COPY_NUMBER	Referenceコピー数（for SV detected by array）
融合遺伝子情報	FUSION_GENES	融合遺伝子名
	GENE_SYMBOL_2	HGNC Gene Symbol
	CHROMOSOME_2	染色体
	START_2	変異の開始位置
	STOP_2	変異の終了位置
遺伝的由来	ALLELE_ORIGIN	Somatic/Germline/unknown/not provided
	ALLELE_ORIGIN_COMMENT	Somatic/Germlineのコメント
変異情報取得方法の詳細	PLATFORM_TYPE	解析プラットフォームの種類
	PLATFORM_NAME	解析プラットフォーム名
	CAPTURE_METHOD	キャプチャメソッド
	PANEL_NAME	解析パネル
	SOFTWARE_NAME_AND_VERSION	解析に使用したソフトウェア名やバージョンなど

変異・変異頻度用フォーマット（基本事項）

項目タイプ	項目	説明
疾患名・診断名	CONDITION_ID_TYPE	コンディションIDタイプ
	CONDITION_ID_VALUE	上記で選択したDB/OntologyのID
	CODE_TYPE	ICD10などのコードIDタイプ
	CODE_VALUE	コードID
	PREFERRED_CONDITION_NAME	特定のDB/Otologyを使用しない場合の疾患名（非推奨）
関連疾患名	RELETED_CONDITION	関連する疾患名
臨床的意義	CLINICAL_SIGNIFICANCE	臨床的意義
	DATE_LAST_EVALUATED	最終評価日
	ASSERTION_METHOD	判定方法
	CLINICAL_SIGNIFICANCE_CITATIONS	臨床的優位性のリファレンス
	CITATIONS_OR_URLS_FOR_CLINICAL_SIGNIFICANCE_WITHOUT_DATABASE_IDENTIFIERS	上記に記入できないリファレンス（URLなど）
	COMMENT_ON_CLINICAL_SIGNIFICANCE	
	PUBLICATION	
検体（集団）に関する情報	COLLECTION_METHOD	
	COLLECTION_DATE	
	SAMPLE_ID	
	AFFECTED_STATUS	
	CLINICAL_FEATURES	団）での臨床的特徴 （HPO表記が望ましい）
	COMMENT_ON_CLINICAL_FEATURES	のClinical featuresに関するコメント
	TISSUE	変異検出に用いた細胞（Somatic variantの場合は記載が望ましい）
検体付加情報	SEX	性別
	AGE_RANGE	年齢層（BIN指定）
	AGE_OF_ONSET_RANGE	年齢層（BIN指定）

頻度（検体集団単位）での登録で異なる主な箇所

変異・変異頻度用フォーマット（Excel / TSV）

- Excel および TSV は各行に1変異を記載してください
  - Excel では登録者情報と変異情報を別 Sheet に記載していただきます。
  - TSVの場合、データ提供元情報は全行で同じ内容となります。
  - 同一項目の複数記載は不可

ASSEMBLY_NAME	GENE_SYMBOL	REFERENCE_SEQUENCE	HGVS	CHROMOSOME	START	STOP	REFERENCE_ALLELE	ALTERNATE_ALLELE	VARIANT_TYPE
hg19	MET	NM_0011275(	NC_000007.1	chr7	116339990	116339990	A	G	
hg19	KRAS			chr12	25362365	25403737			AMP

変異・変異頻度用フォーマット（XML）

- XML形式では、データ提供元情報は1ファイルに1つとなります。
- 変異の情報は1ファイルに複数記載できます。
  - 複数サンプルのデータを記載することができます。
- 複数記載が可能な項目は同一項目を複数回記載することできます。
- 記載する情報がない項目を省略することができます。

  ※ 項目を記載して、内容をブランクとしても問題ありません。

```
<?xml version="1.0"?>
<MGENVAR>
 - <STUDY>
      <CODE>0</CODE>
      <NAME>MGeND</NAME>
      <DS_CODE>0</DS_CODE>
      <PROJECT_NAME/>
      <REVIEW_STATUS>Construction of database system</REVIEW_STATUS>
   </STUDY>
 - <SUBMITTER>
      <TYPE>Representative</TYPE>
      <NAME>Yasushi Okuno</NAME>
      <PHONE>0757514881</PHONE>
      <EMAIL>okuno.yasushi.4c@kyoto-u.ac.jp</EMAIL>
      <ROLE>Professor</ROLE>
      <ORGANIZATION>Kyoto University</ORGANIZATION>
      <STREET>54 Syogoin Kawahara-cho, Sakyo-ku</STREET>
      <CITY>Kyoto</CITY>
      <PREFECTURE>Kyoto</PREFECTURE>
      <COUNTRY>Kyoto</COUNTRY>
      <POSTAL_CODE>6068507</POSTAL_CODE>
   </SUBMITTER>
 - <ORGANIZATION>
      <NAME>Yasushi Okuno</NAME>
      <PHONE>0757514881</PHONE>
      <EMAIL>okuno.yasushi.4c@kyoto-u.ac.jp</EMAIL>
      <ROLE>Professor</ROLE>
      <ORGANIZATION>Kyoto University</ORGANIZATION>
      <STREET>54 Syogoin Kawahara-cho, Sakyo-ku</STREET>
      <CITY>Kyoto</CITY>
```

変異・変異頻度用フォーマット（VCF）

- VCF形式は、メタ情報・Genotype fieldsに独自項目を追加することで、Excel形式と同じ内容を記載できるようにしたものです。
- 変異に特異的な情報 ***以外*** は、メタ情報（##MGEND）に記載します。

  ※ ##MGENDは複数行記載することができます。
- 変異に特異的な情報は、**Genotype fields** に記載します。
- 1ファイルに1サンプルの情報を記載してください。
- 構造変異は対象外です。
- 同一項目の複数記載は不可
- 記載する情報がない項目を省略することができます。
  - 項目を記載して、内容をブランクとしても問題ありません。

例）VCF変異・変異頻度用フォーマット

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##INFO=<ID=GENE,Number=1,Type=String,Description="Gene Symbol">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
##FORMAT=<ID=CS,Number=1,Type=String,Description="Clinical Significance">
##FORMAT=<ID=GS,Number=1,Type=String,Description="Gene Symbol">
##MGEND=<STUDY_CODE="0", STUDY_NAME="MGeND", DS_CODE="0", PROJECT_NAME="MGeND Var", REVIEW_STATUS="Construction of database system">
##MGEND=<SUBMITTER_TYPE="Representative", SUBMITTER_NAME="Yasushi Okuno", SUBMITTER_PHONE="0757514881", SUBMITTER_EMAIL="okuno.yasushi.4c@kyoto-u.ac.jp", SUBMITTER_ROLE="Professor">
##MGEND=<ORGANIZATION="Kyoto University", STREET="54 Syogoin Kawahara-cho, Sakyo-ku", CITY="Kyoto", PREFECTURE="Kyoto", COUNTRY="Kyoto">
##MGEND=<POSTAL_CODE="6068507">
##MGEND=<HOLD_RELEASE="HOLD", RELEASE_DATE="2018/9/10">
##MGEND=<ASSEMBLY_NAME="hg19", ALLELE_ORIGIN="Somatic", ALLELE_ORIGIN_COMMENT="Paired-sample", PLATFORM_TYPE="WES", PLATFORM_NAME="">
##MGEND=<CAPTURE_METHOD="SureSelect", PANEL_NAME="", SOFTWARE_NAME_AND_VERSION="BWA,GATK,VirScan2", COMMENT="", PRIVATE_COMMENT="">
##MGEND=<CITAION="", CONDITION_ID_TYPE="MedGen", CONDITION_ID_VALUE="40104", CODE_TYPE="", CODE_VALUE="">
##MGEND=<PREFERRED_CONDITION_NAME="Non-small cell lung cancer", RELETED_CONDITION="", ASSERTION_METHOD="", COLLECTION_METHOD="clinical testing", COLLECTION_DATE="2017/11/27">
##MGEND=<SAMPLE_ID="", AFFECTED_STATUS="yes", CLINICAL_FEATURES="", COMMENT_ON_CLINICAL_FEATURES="",TISSUE="">
##MGEND=<SEX="Male", AGE_RANGE="60-69", AGE_OF_ONSET_RANGE="60-69">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA00001
1       215847775       .       C       T       50      PASS    NS=3;DP=9;AA=G  GT:GQ:DP:CS:GS  0/1:35:4:Pathogenic:USH2A
1       241667357       .       T       A       50      PASS    NS=3;DP=9;AA=G  GT:GQ:DP:CS:GS  0/1:35:4:not provided:FH
3       178936076       .       C       G       50      PASS    NS=3;DP=13;AA=T GT:GQ:DP:HQ:CS:GS       0|0:54:7:56,60:not provided:PIK3CA
5       112173368       .       A       T       50      PASS    NS=3;DP=9;AA=G  GT:GQ:DP:CS:GS  0/1:35:4:Pathogenic:APC
6       31639845        .       C       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ:CS:GS       0|0:48:1:51,51:Pathogenic:LY6G5B
7       107338528       .       T       G,A     67      PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB       GT:GQ:DP:HQ:CS:GS       1|2:21:6:23,27:other:SLC26A4
7       107338537       .       G       T       67      PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB       GT:GQ:DP:HQ:CS:GS       1|2:21:6:23,27:Pathogenic:SLC26A4
12      25398285        .       C       T       47      PASS    NS=3;DP=13;AA=T GT:GQ:DP:HQ:CS:GS       0|0:54:7:56,60:Likely benign:KRAS
```

GWAS用フォーマット（基本事項）

- 登録対象は多型および変異情報です。
- フォーマットはxlsx形式です。
- 1ファイルに1つのGWAS解析を記載してください。
  - 異なるGWAS解析は別ファイルに記載
- 各行に1多型・変異を記載してください。

  ※ 項目タイプ「変異情報取得方法の詳細」「疾患名」は全ての行で同じ値
- 項目タイプ「臨床情報」はサブ解析ごとに記載してください。
  - サブ解析は異なるフェノタイプで層別化解析を行なった場合などを想定
  - Control となる group は項目タイプ「疾患名」を設定を引き継ぎます。
- 同じ変異（位置・Ref・Altが同じ）でも解析ステージまたは
  サブ解析が異なる場合は別の行に記載します。
- 同一項目の複数記載は不可となります。
- 記載する情報がない場合は対象箇所を**ブランク**にしてください。

GWAS用フォーマット（基本事項）

項目タイプ	項目	説明
配列名	ASSEMBLY_NAME	参照配列名
遺伝子名	GENE_SYMBOL	HGNC Gene Symbol
	REFERENCE_SEQUENCE	VGVSを含むに使用した参照配列
SNP情報 (short indel含む)	RS_NUMBER	dbSNP rs number
	VGVS	VGVS表記
	CHROMOSOME	染色体
	START	ポジション
	END	終端ポジション
	REFERENCE_ALLELE	参照配列(e.g. hg19)の塩基
	ALTERNATE_ALLELE	変異アリル
	EVIDENCE	検証段階
	ANNOTATED_FUNCTIONAL_CLASS	推定される機能クラス
	AFFECTOR_GENE	全量影響を及ぼす遺伝子
GWAS情報 （多型 + 変異に固有な情報） 主に検定の結果得られた情報	STAGE	解析ステージ
	GENOTYPING_METHOD	フェノタイピング手法
	AFFECT_ALLELE	リスク・効果アリル
	NON_EFFECT_ALLELE	リスク・効果アリル
	GENOTYPE_COUNTS_FOR_SUB_GROUP_1	集団1 (sub-group 1)の遺伝型カウント
	GENOTYPE_COUNTS_FOR_SUB_GROUP_2	集団2(sub-group 2)の遺伝型カウント
	ALLELE_COUNTS_FOR_SUB_GROUP_1	集団1 (sub-group 1)のアリルカウント
	ALLELE_COUNTS_FOR_SUB_GROUP_2	集団2(sub-group 2)のアリルカウント
	AFFECT_ALLELE_FREQUENCY_FOR_SUB_GROUP_1	集団1 (sub-group 1)の効果アリルの頻度
	AFFECT_ALLELE_FREQUENCY_FOR_SUB_GROUP_2	集団2(sub-group 2)の効果アリルの頻度
	STATISTICAL_TEST	検定の種類
	TEST_STATISTICS	検定量
	P_VALUE	P値
	ODDS_RATIO	オッズ比
	RS_CONFIDENCE_INTERVAL_OF_OR_UPPER	オッズ比の95%信頼区間
	RS_CONFIDENCE_INTERVAL_OF_OR_LOWER	オッズ比の95%信頼区間
	BETA	効果サイズ
	RS_CONFIDENCE_INTERVAL_OF_BETA_UPPER	効果サイズの95%信頼区間
	RS_CONFIDENCE_INTERVAL_OF_BETA_LOWER	効果サイズの95%信頼区間
	COEFFICIENT_FOR_ALTERNATIVE_ALLELE	Coefficient
	STANDARD_ERROR_OF_COEFFICIENT	Standard error
変異情報取得 方法の詳細	GENOTYPED	genotypeしたものかimputeしたものか
	IMPUTATION_QUALITY	imputedの場合のQuality
	PLATFORM_TYPE	解析プラットフォームの種類
	PLATFORM_NAME	解析プラットフォーム名
	CAPTURE_METHOD	キャプチャメソッド
	SOFTWARE_NAME_AND_VERSION	解析に使用したソフトウェア名やバージョンなど
変異に対するコメント	COMMENT	公開コメント
	PRIVATE_COMMENT	非公開コメント
疾患名 (研究に固有な情報)	CONDITION_ID_TYPE_FOR_SUB_GROUP_1	コンディションIDタイプ
	CONDITION_ID_VALUE_FOR_SUB_GROUP_1	上記で選択したDB/OntologyのID
	CONDITION_ID_TYPE_FOR_SUB_GROUP_2	コンディションIDタイプ
	CONDITION_ID_VALUE_FOR_SUB_GROUP_2	上記で選択したDB/OntologyのID
	CODE_TYPE_FOR_SUB_GROUP_1	ICD10などのコードIDタイプ
	CODE_VALUE_FOR_SUB_GROUP_1	コードID
	PREFERRED_CONDITION_NAME_FOR_SUB_GROUP_1	特定のDB/Ontologyを使用しない場合の疾患名（非推奨）
	CODE_TYPE_FOR_SUB_GROUP_2	ICD10などのコードIDタイプ
	CODE_VALUE_FOR_SUB_GROUP_2	コードID
	PREFERRED_CONDITION_NAME_FOR_SUB_GROUP_2	特定のDB/Ontologyを使用しない場合の疾患名（非推奨）
	SUB_GROUP_FOR_CONTROL	Case/Control の場合は、Control の集団を選択
	CITATION	解析を報告した論文発表情報
臨床情報 （検定に固有な情報） 主に研究デザインに 関わる情報 （サブタイプごとに記載）	SUB_PHENOTYPE	サブ解析
	DESCRIPTION	サブタイプの詳細
	NUMBER_OF_SAMPLE_FOR_SUB_GROUP_1	サンプル数(集団1 (sub-group 1))
	NUMBER_OF_MALES_FOR_SUB_GROUP_1	うち男性
	NUMBER_OF_FEMALES_FOR_SUB_GROUP_1	うち女性
	NUMBER_OF_MIXED_GENDERS_FOR_SUB_GROUP_1	うちmixed gender
	NUMBER_OF_UNKNOWNS_FOR_SUB_GROUP_1	うちunknown
	NUMBER_OF_NOT_PROVIDED_FOR_SUB_GROUP_1	うちnot provided
	AGE_OF_MALES_FOR_SUB_GROUP_1	男性平均分布(mean+SE)
	AGE_OF_FEMALES_FOR_SUB_GROUP_1	女性平均分布(mean+SE)
	AGE_OF_MIXED_GENDERS_FOR_SUB_GROUP_1	mixed gender平均分布(mean+SE)
	AGE_OF_UNKNOWNS_FOR_SUB_GROUP_1	unknown平均分布(mean+SE)
	AGE_OF_NOT_PROVIDED_FOR_SUB_GROUP_1	not provided平均分布(mean+SE)
	NUMBER_OF_SAMPLE_FOR_SUB_GROUP_2	サンプル数(集団2(sub-group 2))
	NUMBER_OF_MALES_FOR_SUB_GROUP_2	うち男性
	NUMBER_OF_FEMALES_FOR_SUB_GROUP_2	うち女性
	NUMBER_OF_MIXED_GENDERS_FOR_SUB_GROUP_2	うちmixed gender
	NUMBER_OF_UNKNOWNS_FOR_SUB_GROUP_2	うちunknown
	NUMBER_OF_NOT_PROVIDED_FOR_SUB_GROUP_2	うちnot provided
	AGE_OF_MALES_FOR_SUB_GROUP_2	男性平均分布(mean+SE)
	AGE_OF_FEMALES_FOR_SUB_GROUP_2	女性平均分布(mean+SE)
	AGE_OF_MIXED_GENDERS_FOR_SUB_GROUP_2	mixed gender平均分布(mean+SE)
	AGE_OF_UNKNOWNS_FOR_SUB_GROUP_2	unknown平均分布(mean+SE)
	AGE_OF_NOT_PROVIDED_FOR_SUB_GROUP_2	not provided平均分布(mean+SE)
	SOURCE_OF_CONTROL	コントロール サンプルの提供元
	ADJUSTED_PARAMETERS	検定における補正パラメーター

配列名

遺伝子名

SNP情報

GWAS情報
（主に検定の結果得られた情報）

変異情報の取得方法の詳細

疾患名（研究に固有な情報）

臨床情報（検定に固有な情報）
主に研究デザインに関わる情報

例）GWASフォーマット

ASSEMBLY_NAME	GENE_SYMBOL	REFERENCE_SEQUENCE	RS_NUMBER	HGVS	CHROMOSOME	START	END	REFERENCE_ALLELE	ALTERNATE_ALLELE	EVIDENCE	ANNOTATED_FUNCTIONAL_CLASS	EFFECTOR_GENE	STAGE	GENOTYPING_METHOD	EFFECT_ALLELE	NON_EFFECT_ALLELE	GENOTYPE_COUNTS_FOR_SUB_GROUP_1	GENOTYPE_COUNTS_FOR_SUB_GROUP_2	ALL
ng19	TBX15		rs2282322		chr1	119247524	119247524	G	A				Discovery	SNP array	G	A	45,90,49	16,93,72	
ng19	SLC8A1		rs13428519		chr2	41279086	41279086	A	C				Discovery	SNP array	A		98,73,13	64,85,32	
ng19	SLC8A1		rs13428519		chr2	41279086	41279086	A	C				Replication	DigiTag	A		54,91,39	31,80,70	
ng19	RER1		rs3736330		chr1	2316935	2316935	C	T				Discovery	SNP array	T		32,88,54	1,17,72	
ng19	TBX15		rs2282322		chr1	119247524	119247524	G	A				Discovery	SNP array	G	A	16,93,72	79,10,1	
ng19	SLC8A1		rs13428519		chr2	41279086	41279086	A	C				Discovery	SNP array	A		64,85,32	15,44,31	

HLA 用フォーマット（基本事項）

- 登録対象は HLA allele の頻度情報です。
- HLA haplotype は対象外となります。
- 対象の Field は 2~4 field とします。
  - Field の混在は可能
- 登録用標準フォーマットとして、以下の2通り用意しています。

  HLAフォーマット1：HLAの情報をLocusごとにsheetを分けて記載するもの

  HLAフォーマット2：HLAの情報を1 sheetにまとめて記載するもの
- 2群比較の場合は、HLAフォーマット1を使用してください。
- 解析ごとに "1ファイル" 作成してください。
  - 同じデータを用いた場合でも異なる解析条件によるものは

    別ファイルとしてください。
  - STUDY_NAME が同じ場合は共通データの別解析と判断します。
- 解析情報 とHLA情報はそれぞれ別のSheetに記載していただきます。

HLA 用フォーマット（解析情報）

項目タイプ	項目	説明	形式	必須（※1）	公開（※2）	複数記載（※3）	備考
解析情報	STUDY_NAME	研究の名称	String	○	○	×	研究課題名。STUDY_NAMEが同じデータは共通のデータに対する異なる解析手法の結果と認識します。
	SUB_STUDY_NAME	研究のサブ名称	String	△	○	×	Group AとGroup Bの組合せの名称。 複数の組合せが存在するStudyの場合は必須となります。 STUDY_NAMEとSUB_STUDY_NAMEの組合せがユニークである必要があります。
	DESCRIPTION	研究の詳細	String	×	○	×	
	PHENOTYPE	フェノタイプ	String	×	○	×	
	REGION	データ取得を実施した地域	String	×	○	×	
	TYPING_METHOD	タイピング手法	String	×	○	×	
	TYPING_ALGORITHM	タイピングアルゴリズム	String	×	○	×	
	TYPING_KIT	キット名	String	×	○	×	
	SEQUENCING_LIBRARY_STRATEGY	Sequencing Library Strategy	String	×	○	×	
	PLATFORM	解析プラットフォーム名	String	×	○	○	選択肢に存在しない記載の場合、Otherとなります。
	IPD_IMGT_HLA_VERSION	IPD-IMGT/HLA version	String	×	○	×	

HLA用フォーマット（ HLAフォーマット１）

- ファイルはxlsx形式（エクセル）です
- 2群比較データ用に作成したフォーマット

※ 1群のデータ登録にも使用することができます。

- HLA情報 sheet には Locus ごと（e.g. HLA-A）の情報を記載
    - Locusごとのため、sheet数は不定
    - LocusごとのHLAの情報を記載
    - ヘッダー及び複数行のHLA情報で構成
- 同一項目の複数記載は不可となります。
- 記載する情報がない場合は対象箇所を**ブランク**にしてください。

項目タイプ	項目	説明	形式	必須（※1）	公開（※２）
頻度情報	ALLELE	アレル	String	○	○
	GROUP_A_COUNTS	Group Aの件数	Integer	○	○
	GROUP_A_FREQUENCY	Group Aでの割合	Numeric	○	○
	GROUP_B_COUNTS	Group Bの件数	Integer	△	○
	GROUP_B_FREQUENCY	Group Bでの割合	Numeric	△	○
	P_VALUE	p値	Numeric	×	○
	ODD_RATIO	オッズ比	Numeric	×	○
	ODDS_RATIO_95_CI_LOWER	オッズ比が95%信頼区間	Numeric	×	○
	ODDS_RATIO_95_CI_UPPER		Numeric	×	○

例）HLAフォーマット1

	A	B	C	D	E	F	G	H	I
1	ALLELE	GROUP_A_COUNTS	GROUP_A_FREQUENCY	GROUP_B_COUNTS	GROUP_B_FREQUENCY	P_VALUE	ODD_RATIO	ODDS_RATIO_95_CI_LOWER	ODDS_RATIO_95_CI_UPPER
2	A*02:10	0	0	18	0.394563788	0.011094359			
3	A*02:18	0	0	2	0.043840421	0.3978451			
4	A*24:20	0	0	22	0.48224463	0.004974693			
5	A*26:01	136	8.343558282	348	7.628233231	0.355759119	1.102309622	0.896372401	1.355559923
6	A*26:02	1	0.061349693	66	1.446733889	3.47533E-06	0.04181781	0.005799179	0.301547709
7	A*26:03	36	2.208588957	105	2.301622096	0.82889169	0.958666428	0.65373293	1.40583605
8	A*26:05	0	0	4	0.087680842	0.231744599			
9	A*30:01	0	0	4	0.087680842	0.231744599			
10	A*31:01	164	10.06134969	379	8.307759754	0.031676451	1.23469171	1.018376104	1.496955412
11	A*33:03	118	7.239263804	380	8.329679965	0.164683442	0.858876358	0.692880981	1.064639697
12									
13									
14									

Submission | HLA-A | HLA-C | HLA-B | HLA-DRB1 | HLA-DQB1 | HLA-DPB1

HLA用フォーマット（ HLAフォーマット２）

- ファイルはxlsx形式（エクセル）です
- 1群データ用に作成したフォーマット
- HLA情報 Sheet には書くアレルおよびその頻度等を記載
    - HLAの情報を記載
    - 1行のヘッダー及び複数行のHLAの情報で構成
- 同一項目の複数記載は不可となります。
- 記載する情報がない場合は対象箇所を**ブランク**にしてください。

項目タイプ	項目	説明	形式	備考
頻度情報	Gene	遺伝子座	String	例えば「HLA-A」のような形式で記載する。
	Allele	アレル	String	例えば「A*02:01」のような形式で記載する。（2fieldの記載に限定しない。）
	Counts	件数	Integer	
	Frequency	割合	Numeric	
	Sample size	サンプルサイズ	Integer	
	field	フィールド	Integer	フィールド数はAlleleの記載から得られるフィールド数よりもこの値を優先する。 例えばAlleleが「A*02:01:01」、fieldが2の場合、「A*02:01」として登録する。 また、Alleleが「A*02:01」、fieldが3の場合、「3fieldのデータで、3つ目のfieldが不明」の情報と判断する。これは登録の対象外となる。

例）HLAフォーマット２

	A	B	C	D	E	F
1	GENE	ALLELE	COUNTS	FREQUENCY	SAMPLE_SIZE	FIELD
2	HLA-A	A*02:01	383	0.110184	3476	2
3	HLA-A	A*02:01:01	383	0.110184	3476	3
4	HLA-A	A*31:01	280	0.0805524	3476	2
5	HLA-A	A*31:01:02	280	0.0805524	3476	3
6	HLA-A	A*26:01	254	0.0730725	3476	2
7	HLA-A	A*26:01:01	254	0.0730725	3476	3
8	HLA-A	A*24:02	1360	0.391254	3476	2
9	HLA-A	A*24:02:01	1356	0.390104	3476	3
10	HLA-A	A*02:07	116	0.0333717	3476	2
11	HLA-A	A*02:07:01	116	0.0333717	3476	3