centre for population genomics

murdoch children's research institute

Garvan Institute of Medical Research

# Organising project specific rare disease data and metadata for seqr

## *Table of Contents*

⇒ *Download this PDF to use the Table of Contents links*

## i. Purpose

The purpose of this document is to provide instructions on how to prepare project specific data and metadata for use in seqr related to an individual's family history (pedigree).

## ii. Background

The CPG utilises four distinct metadata files to provide information about samples to the variant curation team as they perform variant analysis in seqr.

The four files are described in **Table 1**.

A template for each of these files is provided in section iii. Quick Links, and instructions for filling out each template are included in this document.

***Table 1: Definitions of the metadata files required.***

| Metadata template | Required | Description |
|---|---|---|
| *Pedigree_metadata_template* | **Yes** | Template file used to describe the individuals in each dataset and how they relate to other individuals, mainly their parents in the same dataset.<br>The information in this file is used to generate the participant pedigrees*. |
| *Families_metadata_template* | No | Template file used to describe the families in each dataset. |
| *Individuals_metadata_template* | **Yes** | Template file used to describe the clinical information related to individuals in each dataset. |
| *Sample_mapping_template* | **Yes** | Template file used to map individual IDs AND sample IDs back to the files that have been transferred. |

*A pedigree is a structured description of the phenotypical and familial relationships between samples.

The CPG uses the tool ‘GATK HaplotypeCaller’, which can incorporate pedigree information in the genomic analysis of samples.

## iii. Quick Links

All template files can be found HERE

# 1. Genomic Data

1.1. CPG's bioinformatic pipelines use the following genomic data types:

- CRAM files
- BAM files
- FASTQ files

**Note:** CPG's preference is to use CRAM files. In the absence of CRAM files, BAM or FASTQ files can be transferred.

1.2. For each CRAM/BAM/FASTQ file that is to be transferred, a corresponding MD5 file also needs to be transferred, for data integrity QC to occur after the transfer.

1.3. Ensure that the genomic data files are transferred to a specific directory in the CPG's cloud storage.
Appropriate directories include the date of the transfer in the directory path.

1.4. Further instructions can be found in this document:
*Uploading your data to CPG cloud*

# 2. Pedigree_metadata_template

2.1. Download the *pedigree_metadata_template* file from the CPG Google drive here.

2.2. Information relating to **all** individuals should be documented in a single *pedigree_template* file. If an individual appears in the Paternal ID or Maternal ID column, then that individual needs their own dedicated row.

**Note:** You should only have **one** *pedigree_template* file. This single file can contain as many individuals as described in your cohort/dataset. Do not create separate *Pedigree_metadata_template* files for each individual in your cohort/dataset.

2.3. Populate the *pedigree_metadata_template* according to **Table 2**.
An example is given below in **Table 3**.

2.4. Ensure that the *pedigree_metadata_template* file is shared alongside your transfer.

***Table 2: Data dictionary for pedigree_metadata_template file describing inputs for template fields.***

| Field label | Allowed Values | Notes |
|---|---|---|
| Family ID | Alphanumeric family ID | The combination of family and individual ID should uniquely identify a person. |
| Individual ID | Alphanumeric individual ID | |
| Paternal ID | Alphanumeric paternal ID | Individuals without parental data can use a 0 in these columns or leave them blank. |
| Maternal ID | Alphanumeric maternal ID | |
| Sex | 0 = unknown<br>1 = male<br>2 = female | If an individual's sex is unknown, then a 0 should be used in this column. |
| Affected Status | -9 = missing<br>0 = missing<br>1 = unaffected<br>2 = affected | -9 or 0 can both equally be used to denote a missing affected status for an individual. |

***Table 3: Example of a populated pedigree_metadata_template file.***

| Family ID | Individual ID | Paternal ID | Maternal ID | Sex | Affected Status |
|---|---|---|---|---|---|
| FAM_001 | IND_001 | IND_003 | IND_002 | 1 | 2 |
| FAM_001 | IND_002 | | | 2 | 1 |
| FAM_001 | IND_003 | | | 1 | 2 |
| FAM_002 | IND_004 | | IND_005 | 2 | 2 |
| FAM_002 | IND_005 | | | 2 | 2 |

# 3. Families_metadata_template (Optional)

3.1. Download the *families_metadata_template* file from the CPG Google drive here.

3.2. All information relating to families should be documented in a single *families_metadata_template* file.

**Note:** You should only have **one** *families_metadata_template* file. This single file can contain as many families as described in your cohort/dataset. Do not create separate *families_metadata_template* files for each family in your cohort/dataset.

3.3. Populate the *families_metadata_template* according to **Table 4**.
An example is given below in **Table 5**.

3.4. Ensure that the *families_metadata_template* file is shared alongside your transfer.

***Table 4: Data dictionary for families_metadata_template file describing inputs for the template fields.***

| Field label | Allowed Values | Notes |
|---|---|---|
| Family ID | Alphanumeric family ID | The IDs are alphanumeric: the family ID should uniquely identify a family. |
| Display Name | Alphanumeric characters | An optional secondary identifier. |
| Description | Alphanumeric characters | Clinical description of the family. |
| Coded Phenotype | Comma-separated list of HPO codes for present phenotypes in this individual | Coded clinical phenotypes related to the clinical description of the family, preferably in HPO terms. |

***Table 5: Example of a populated families_metadata_template file.***

| Family ID | Display name | Description | Coded Phenotype |
|---|---|---|---|
| FAM_001 | | Neurodegeneration, progressive motor degeneration, ataxia, spasticity, dementia, regression, brain atrophy | HP:0002180 |
| FAM_002 | | Dilated cardiomyopathy, leukodystrophy | HP:0002415, HP:0001644 |

# 4. Individuals_metadata_template

4.1. Download the *individuals_metadata_template* file from the CPG Google drive here

4.2. All information relating to individuals should be documented in a single *individuals_metadata_template* file.

**Note:** You should only have **one** *individuals_metadata_template* file. This single file can contain as many individuals as described in your cohort/dataset. Do not create separate *individuals_metadata_template* files for each family in your cohort/dataset.

4.3. Populate the *individuals_metadata_template* according to **Table 6**. An example is given below in **Table 7**.

**Note**: Only populate the fields that you have information for. Not every field needs to be populated in this template file. The more information you provide in the file, the better your experience will be in seqr.

4.4. Ensure that the *individuals_metadata_template* file is shared alongside your transfer.

***Table 6: Data dictionary for individuals_metadata_template file describing inputs for the template fields.***

| Field label | Allowed Values | Notes |
|---|---|---|
| Family ID | Alphanumeric family ID | The combination of family ID and individual ID should uniquely identify an individual. |
| Individual ID | Alphanumeric individual ID | |
| HPO Terms (present) | Comma-separated list of HPO codes for present phenotypes in this individual | This field should have the HPO codes, not the descriptions. |
| HPO Terms (absent) | Comma-separated list of HPO codes for phenotypes *not* present in this individual | This field should have the HPO codes, not the descriptions. |
| Birth Year | Numeric year of birth. E.g. 2010 | If you have collected a DOB, e.g. 01-01-2001, please only include the ***year*** component. |
| Death Year | Numeric year of death, if applicable. Leave blank otherwise. | If you have collected a DOD, e.g. 01-01-2001, please only include the ***year*** component. |
| Age of Onset | **One of the following:**<br>Embryonal onset,<br>Congenital onset,<br>Fetal onset,<br>Neonatal onset,<br>Infantile onset,<br>Childhood onset,<br>Juvenile onset,<br>Adult onset,<br>Young adult onset,<br>Middle age onset, | *This is a rough suggestion, with no clinical source.*<br>**Embryonal onset:** conception to 8 wks gestation<br>**Fetal onset:** 9 wks gestation to birth<br>**Congenital onset:** conception to birth<br>**Neonatal onset:** birth to 1 month,<br>**Infantile onset:** birth to 1 year |

| | | |
|---|---|---|
| | Late onset | **Childhood onset:** < 5 years<br>**Juvenile onset:** < 17 years<br>**Young adult onset**: < 25 years<br>**Adult onset:** < 36 years<br>**Middle age onset**: < 55 years<br>**Late onset:** > 55 years |
| Individual Notes | Alphanumeric characters | |
| Consanguinity | true, false, or blank if unknown | |
| Other Affected Relatives | true, false, or blank if unknown | |
| Expected Mode of Inheritance | **Comma-separated list of the following:**<br>Sporadic,<br>Autosomal dominant inheritance,<br>Sex-limited autosomal dominant,<br>Male-limited autosomal dominant,<br>Autosomal dominant contiguous gene syndrome,<br>Autosomal recessive inheritance,<br>Gonosomal inheritance,<br>X-linked inheritance,<br>X-linked recessive inheritance,<br>Y-linked inheritance,<br>X-linked dominant inheritance,<br>Multifactorial inheritance,<br>Mitochondrial inheritance | |
| Fertility Medications | true, false, or blank if unknown | |
| Intrauterine Insemination | true, false, or blank if unknown | |
| In Vitro Fertilization | true, false, or blank if unknown | |
| Intra-Cytoplasmic Sperm Injection | true, false, or blank if unknown | |
| Gestational Surrogacy | true, false, or blank if unknown | |
| Donor Egg | true, false, or blank if unknown | |
| Donor Sperm | true, false, or blank if unknown | |
| Maternal Ancestry | comma-separated list of ethnicities | |
| Paternal Ancestry | comma-separated list of ethnicities | |
| Pre-discovery OMIM disorders | comma-separated list of valid OMIM numbers | |
| Previously Tested Genes | comma-separated list of genes | |
| Candidate Genes | comma-separated list of genes | |

***Table 7: Example of a populated individuals_metadata_template file.***

| Family ID | Individual ID | HPO Terms (present) | HPO Terms (absent) | Birth Year | Death Year | Age of Onset | Individual Notes | Consanguinity | Other Affected Relatives | Expected Mode of Inheritance | Fertility medications | Intrauterine insemination | In vitro fertilization | Intra-cytoplasmic sperm injection | Gestational surrogacy | Donor egg | Donor sperm | Maternal Ancestry | Paternal Ancestry | Pre-discovery OMIM disorders | Previously Tested Genes | Candidate Genes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FAM_001 | IND_001 | HP:0002180 |  | 2001 |  | Fetal onset | Neurodegeneration, progressive motor degeneration, ataxia, spasticity, dementia, regression, brain atrophy | True | False | Autosomal dominant inheritance | False |  | True | True | False | True | True | Australian | Australian |  |  |  |

# 5. Sample_mapping_template

5.1. Download the *sample_mapping_template* file from the CPG Google drive here.

5.2. Populate the *sample_mapping_template* file according to **Table 8**.
An example is given below in **Table 9**.

5.3. Ensure that the *sample_mapping_template* file is shared alongside your transfer.

***Table 8: Data dictionary for sample_mapping_template file describing inputs for the template fields.***

| Field label | Allowed Values | Notes |
|---|---|---|
| Individual ID | Alphanumeric individual ID (if different to the Sample ID) | This column can be left blank if the individual ID and the sample ID are identical. |
| Sample ID | Alphanumeric sample ID | A sample ID should be unique within a project. Note that an individual can have multiple samples. |
| Filenames | Comma-separated list of filenames for this sample. | If more than two files are provided, they will be grouped automatically |
| Type | One of the following: WGS, WES | WGS (Whole genome), or (WES) whole-exome sequencing.<br>***If this field is blank the type will default to WGS.***<br>**Note:** If a sample has both WES and WGS sequence data, you should include a row for each type. |

***Table 9: Example of a populated sample_mapping_template file.***

| Individual ID | Sample ID | File names | Type |
|---|---|---|---|
| IND_001 | A0001 | A0001-R1.fastq.gz, A0001-R2.fastq.gz | WGS |
| IND_001 | A0001 | A0001_WES-R1.fastq.gz, A0001_WES-R2.fastq.gz | WES |
| IND_002 | A0002 | A002-R1.fastq.gz, A0002-R2.fastq.gz | WGS |