# Data Standards - TKI Indigenous Genomics

Sam Buckberry & Jimmy Breen

2022-03-03

# Contents

# Chapter 1

# Introduction

***Note: These standards are currently in development and are yet to be officially implemented***

In this era of leveraging big data and genomics for precision medicine, small populations such as Indigenous Australians are at high risk for under-representation or complete lack of representation. Our goal is to collect, curate and assemble robust datasets that encompass critical aspects of Indigenous Australian health to ensure equitable outcomes in precision medicine are possible. This includes biological, cultural and socioeconomic data to investigate health outcomes that will deliver meaningful benefit to Indigenous Australians. However, with big data also come significant responsibilities concerning data storage, sovereignty and accessibility. Large collections of data on Indigenous Australians require carefully considered data architecture that maintains sovereignty based on community-defined priorities, protocols, and agreements, with concomitant attendance to matters of accountability and building trust.

Here we outline TKI Indigenous Genomics groups standards for data storage, access and sharing. **The goal of these data standards is to document the policies, rules, and standards governing how data are stored, secured, arranged, integrated, and used in analyses and reporting.** These standards are being developed in consultation with Indigenous Communities. Where possible, these standards will also align with the The Global Alliance for Genomics and Health GA4GH standards. A well-considered approach to data architecture is critical given the national importance of Indigenous genomics data security.

# Chapter 2

# Data descriptors & definitions

This page provides definitions and guidance on the information that should be recorded for primary data files. Data that is clearly, simply and consistently described (named and defined) is much easier for users to identify, understand, use and share the data. This also ensures consistency across projects, interoperability and is complementary to reproducible research practices.

## 2.1 Describing primary data file types

Biological data continue to be produced from an extensive array of assays and technologies. These data are often represented in unique file types and processed by specific software and algorithms. **Below are descriptors that should be used for each file type.**

For example, the description the CGmap file format for DNA methylation data would look:

> **File type:** CGmap
> **Description:** TSV file of stranded pileup base calls for cytosine positions in the reference genome DNA methylation data.
> **Compression:** gzip
> **Permissions:**
> **Access frequency:**
> **Community access:**
> **Identifiable:**
> **Location:**
> **Notes:** This file type is output from BSSeeker2 and used by CGmap

tools
**Sample:**

chr1 G 3000851 CHH CC 0.1 1 10
chr1 C 3001624 CHG CA 0.0 0 9
chr1 C 3001631 CG CG 1.0 5 5

Format descriptions (columns):
(1) chromosome
(2) nucleotide on Watson (+) strand
(3) position
(4) context (CG/CHG/CHH)
(5) dinucleotide-context (CA/CC/CG/CT)
(6) methylation-level = #_of_C / (#_of_C + #_of_T)
(7) #_of_C (methylated C, the count of reads showing C here)
(8) = #_of_C + #_of_T (all Cytosines, the count of reads showing C or T here)

## 2.2 File type descriptors

**File type:** This is the file format to be typically denoted by the file extension.

**Description:** A brief description of the file format.

**Compression:** The algorithm and/or method used to compress the file.

**Permissions:** The groups and users that have access to the data.

**Access frequency:** How often this data type may require access. These descriptions are based on a subset of the Google Cloud storage class descriptors.

- **Standard:** Standard Storage is for data that is frequently accessed and/or stored for only brief periods of time. Standard data files should be able to be regenerated from Nearline or Archive data files without too much time or effort.
- **Nearline:** Nearline Storage is ideal for data you plan to read or modify on average once per month or less. Nearline Storage is appropriate for data backup and short-term archiving.
- **Archive:** highly durable storage for data archiving, backup, and disaster recovery. Only used for infrequently accessed data.

**Community access:** Are their community custodians for these data? If so, who are they?

**Identifiable:** Could this data be used to identify an individual (YES/NO).

**Location:** Location(s) where the data should be stored.

**Notes:** Any additional information that would assist users of the data.

**Sample:** A sample of the file data with descriptions of each field. Useful if this file type is not a widely used and documented standard.

# Chapter 3

# Metadata

# Chapter 4

# Whole genome sequencing

## 4.1 CRAM

**File type:** CRAM
**Description:** CRAM supports a wide range of lossless and lossy sequence data preservation strategies enabling users to choose which data should be preserved. CRAM is the genomics compression standard for GA4GH
**Compression:** NA
**Permissions:**
**Community access:**
**Location: Notes:** When used with a reference genome, this exact reference genome file should be recorded.
CRAM specification

## 4.2 VCF

**File type:** VCF
**Description:**
**Compression:**
**Permissions:**
**Access frequency:**
**Community access:**
**Identifiable:**
**Location:**
**Notes:**

# Chapter 5

# DNA methylation

---

## 5.1 Raw data

### 5.1.1 FASTQ

**File type:** FASTQ
**Description:**
**Compression:**
**Permissions:**
**Access frequency:**
**Community access:**
**Identifiable:**
**Location:**
**Notes:**

---

## 5.2 Mapped data

### 5.2.1 CRAM

**File type:** CRAM
**Description:** Genomic alignments (typically). CRAM supports a wide range of lossless and lossy sequence data preservation strategies enabling users to choose which data should be preserved. CRAM is the genomics compression standard

for GA4GH
**Compression:** NA
**Permissions:**
**Community access:**
**Location: Notes:** When used with a reference genome, this exact reference genome file should be recorded.
CRAM specification

---

## 5.3 Processed data

### 5.3.1 CGmap

**File type:** CGmap
**Description:** TSV file of stranded pileup base calls for cytosine positions in the reference genome DNA methylation data. **Compression:** gzip **Permissions:**
**Access frequency:**
**Community access:**
**Identifiable:**
**Location:**
**Notes:** This file type is output from BSSeeker2 and used by CGmap tools
**Sample:**

```
chr1    G   3000851 CHH CC  0.1 1   10
chr1    C   3001624 CHG CA  0.0 0   9
chr1    C   3001631 CG  CG  1.0 5   5
```

Format descriptions (columns):

```
(1) chromosome \
(2) nucleotide on Watson (+) strand \
(3) position \
(4) context (CG/CHG/CHH) \
(5) dinucleotide-context (CA/CC/CG/CT) \
(6) methylation-level = #_of_C / (#_of_C + #_of_T) \
(7) #_of_C (methylated C, the count of reads showing C here) \
(8) = #_of_C + #_of_T (all Cytosines, the count of reads showing C or T here)
```

### 5.3.2 ATCGmap

**File type:** ATCGmap
**Description:** TSV file of stranded pileup base calls for DNA methylation data.

**Compression:** gzip
**Permissions:**
**Access frequency:**
**Community access:**
**Identifiable:** YES
**Location:**
**Notes:** This file type is output from BSSeeker2 and used by CGmap tools
**Sample:**

```
chr1    T   3009410 --  --  0   10  0   0   0   0   0   0   0   0   na
chr1    C   3009411 CHH CC  0   10  0   0   0   0   0   0   0   0   0.0
chr1    C   3009412 CHG CC  0   10  0   0   0   0   0   0   0   0   0.0
chr1    C   3009413 CG  CG  0   10  50  0   0   0   0   0   0   0   0.83
```

Format descriptions (columns):

```
(1) chromosome \
(2) nucleotide on Watson (+) strand \
(3) position \
(4) context (CG/CHG/CHH) \
(5) dinucleotide-context (CA/CC/CG/CT) \
(6) - (10) plus strand \
(6) # of reads from Watson strand mapped here, support A on Watson strand \
(7) # of reads from Watson strand mapped here, support T on Watson strand \
(8) # of reads from Watson strand mapped here, support C on Watson strand \
(9) # of reads from Watson strand mapped here, support G on Watson strand \
(10) # of reads from Watson strand mapped here, support N \
(11) - (15) minus strand \
(11) # of reads from Crick strand mapped here, support A on Watson strand and T on Crick strand \
(12) # of reads from Crick strand mapped here, support T on Watson strand and A on Crick strand \
(13) # of reads from Crick strand mapped here, support C on Watson strand and G on Crick strand \
(14) # of reads from Crick strand mapped here, support G on Watson strand and C on Crick strand \
(15) # of reads from Crick strand mapped here, support N \
(16) methylation_level = #C/(#C+#T) = C8/(C7+C8) for Watson strand, =C14/(C11+C14) for Crick stra
"nan" means none reads support C/T at this position. \
```

### 5.3.3 Bigwig

**File type:** Bigwig
**Description:**
**Compression:**
**Permissions:**
**Access frequency:**
**Community access:**

**Identifiable:**
**Location:**
**Notes:**

# Chapter 6

# How to contribute

Lorem ipsum