Genomic Metadata Data Dictionary for the CP Commons

This metadata data dictionary was developed in 2019.

These metadata are designed for the CP Commons.

There are 17 fields: 6 are Mandatory.

Any problems using this form, please contact Yana Wilson (ywilson@cerebralpalsy.org.au).

Current working version: Version 1.1

Version 1.1 – 17 elements – October 2020

Unique IDs (2)

FIELD NAME	CATEGORY	DESCRIPTION	VALUE	NOTES /LOGIC
submitter_ID	**Mandatory**	Unique ID from submitting team	String, alphanumeric (no special characters)	Unique ID from submitting team, must be two steps removed from any personal information. The SUBMITTER_ID can only be viewed by Data Owners and CPA Data Custodians.
family_ID	**Mandatory**	Family ID	String, alphanumeric, permissible special characters (_)	ID code assigned by the submitting team to a family. FAMILY_ID is not necessary for singleton cases.

Genomic Fields (15)

FIELD NAME	REQUIRED	DESCRIPTION	VALUE	NOTES /LOGIC
filename	**Mandatory**	Name of the genome datafile	String, alphanumeric, permissible special characters (_)	The name of the file
data_access_ restriction	**Mandatory**	Data Access Restrictions	1, Open 2, Consortium 3, Controlled	Data owners can request for embargoes for 6 months on data assets in the Data Submission Request. Open – user is not required to submit a Data Access Request (DAR) to receive these data. Consortium – these data are accessible to consortia members under a consortium data sharing agreement. All other users must submit a DAR Controlled – all users must submit a DAR
consent	**Mandatory**	Do you have consent from the participant to share their de-identified data?	0, No 1, Yes	
consent_no	Core	If no, have you received a consent waiver to share these data without the individuals consent?	0, No 1, Yes	Response required if [consent] has value of 0.

restrictions	Recommended	Does this data have additional data use restrictions?	0, No 1, Yes	If yes, users can add the specific Data Use Restriction [future work]. This will also flag the DAR that they need to refer back to their original consent documentation.
pipeline	Core	Please select the CP Commons harmonisation pipeline	1, SNPs + indels 2, CNV 3, RNAseq short variant discovery 4, gVCF pipeline 5, Pre-processing (BAM only) 6, raw data	Tbc by bioinformatics WG – these pipelines would be packaged up in containers for users to access and run on their datasets.
file_format	Core	Please indicate the file format for this data file:	1, VCF 2, gVCF 3, BAM 4, CRAM 5, SFF 6, FASTQ	Initially only accepting .vcf or .gvcf
molecular_class	Core	Broad categorisation of the molecular data:	1, Genome 2, Methylome 3, Transcriptome	
exp_strategy	Core	Please select the sequencing strategy used to generate the data file:	1, Whole Genome Sequencing (WGS) 2, Whole Exome Sequencing (WES) 3, Targeted Sequencing 4, Array-CGH (CNV) 5, SNP-array 6, Methylation array 7, CpG island array 8, Whole Genome Bisulfite-sequencing (WGBS) 9, Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) 10, Chromatin immunoprecipitation sequencing (ChIP-seq) 11, microRNA sequencing (miRNA-seq)	

			12, Microarray 13, Whole RNA-seq 14, Whole transcriptome shotgun sequencing (WTSS) 15, Targeted RNA-seq amplicon	
instrument	Recommended	Name of platform used for sequencing:	1, Affymetrix 2, Agilent 3, Illumina 4, Ion Torrent 5, Nimblegen 6, PacBio	
instrument_model	Recommended	Name of instrument model used for sequencing:	String, alphanumeric, permissible special characters (_)	
instrument_centre	Recommended	Where was the sequencing performed?	String, alphanumeric, permissible special characters (_)	
raw_data	Core	Is the raw data stored in an open database?	0, No 1, Yes	
raw_data_detail	Core	If yes, please provide the hyperlink:	String, alphanumeric, permissible special characters (_ , . , / , :)	Response required on if [raw_data] has a value of (1).
md5sum	**Mandatory**	MD5 Sum:	Integer	The 128-bit hash value expressed as a 32 digit hexadecimal number (in lower case) used as a file's digital fingerprint.