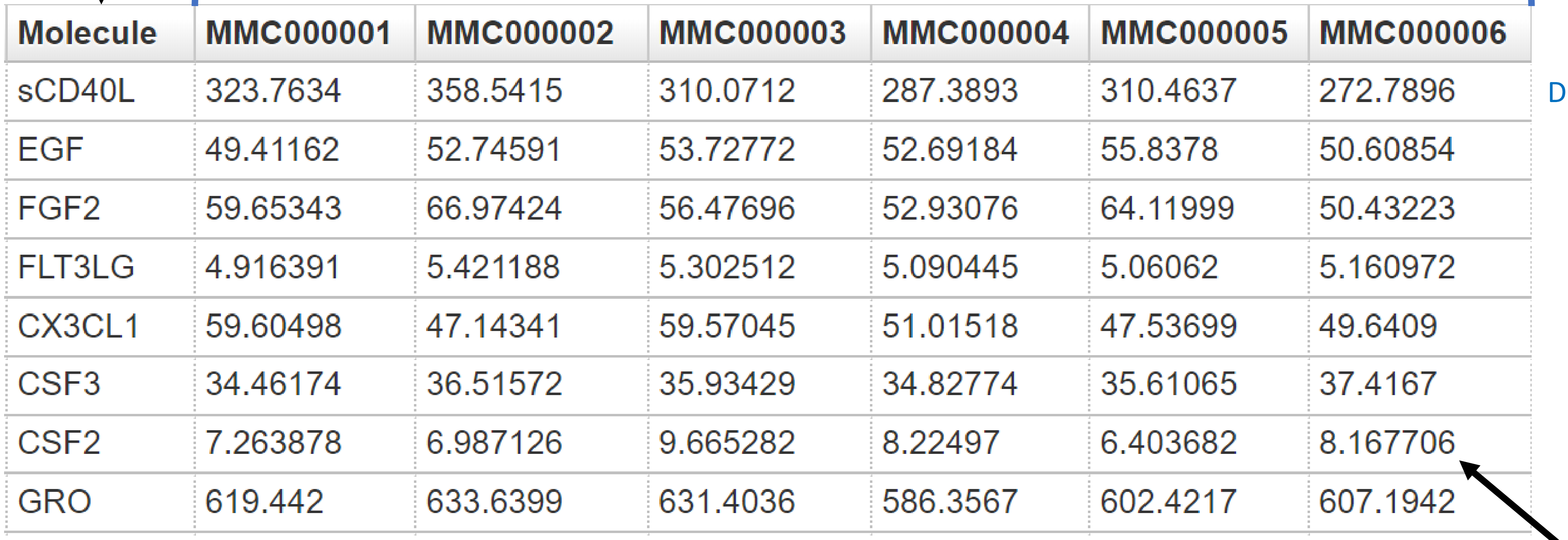
mapMECFS Data Formats

Questions: mapmecfs@rti.org

Cytokine Data File Format

Column 1: Cytokine name.
Column name must be ‘Molecule’.
These must be UNIQUE and
cannot be blank (‘’).

2nd column and onwards
are the participant IDs.
These must be UNIQUE and
match the phenotype file.



Molecule	MMC000001	MMC000002	MMC000003	MMC000004	MMC000005	MMC000006
sCD40L	323.7634	358.5415	310.0712	287.3893	310.4637	272.7896
EGF	49.41162	52.74591	53.72772	52.69184	55.8378	50.60854
FGF2	59.65343	66.97424	56.47696	52.93076	64.11999	50.43223
FLT3LG	4.916391	5.421188	5.302512	5.090445	5.06062	5.160972
CX3CL1	59.60498	47.14341	59.57045	51.01518	47.53699	49.6409
CSF3	34.46174	36.51572	35.93429	34.82774	35.61065	37.4167
CSF2	7.263878	6.987126	9.665282	8.22497	6.403682	8.167706
GRO	619.442	633.6399	631.4036	586.3567	602.4217	607.1942

Data starts on line 2

Each column contains measurements from 1 sample. Missing values are allowed (‘NA’ or leave empty).

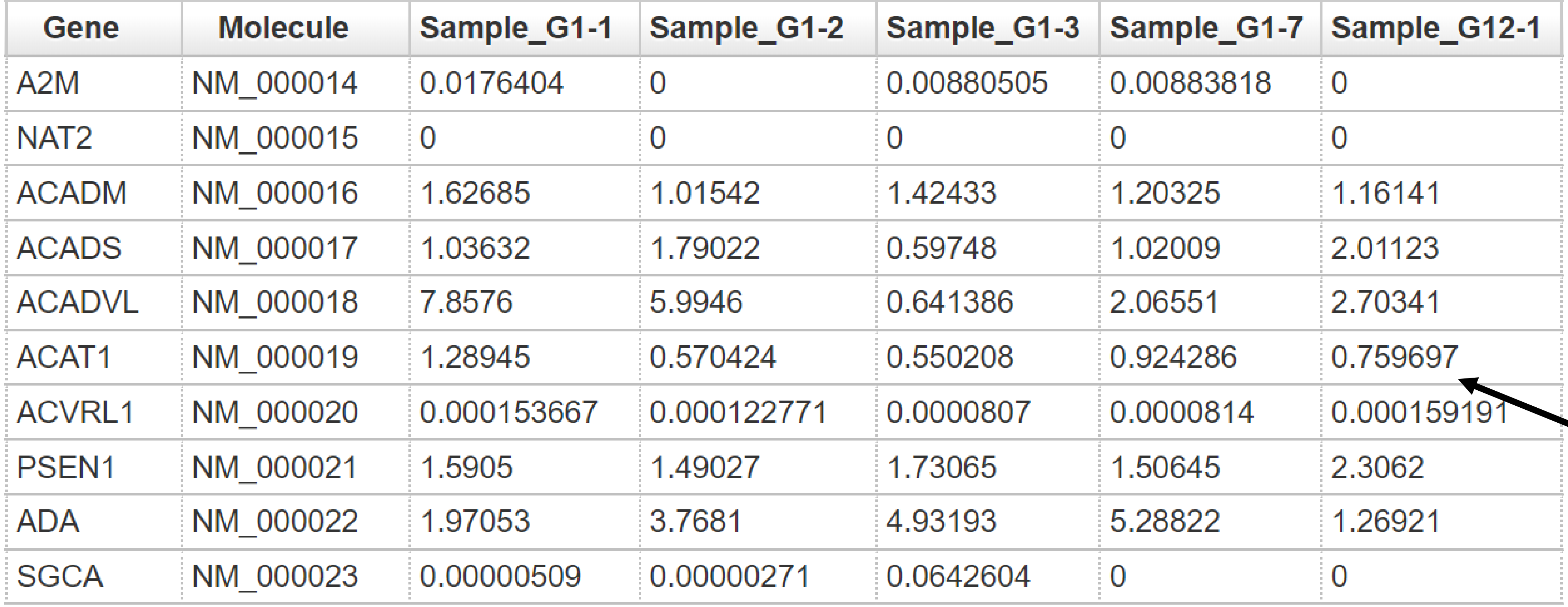
Note: The cytokine data file is formatted as a tab-separated file

Gene Expression at the Transcript Level
Data File Format

Column 1: Gene Symbol. Column name must be ‘Gene’. Other gene identifiers are accepted, although only gene symbol can identify synonyms of the identifier. This is REQUIRED for all data. These are DO NOT have to be UNIQUE.

Column 2: Transcript ID (RefSeq ID, Ensembl ID, Affymetrix ID, etc. are accepted). Column name must be ‘Molecule’. This is REQUIRED only for transcript level data. These must be UNIQUE and cannot be blank (‘’).

3rd column and onwards are the participant IDs. These must be UNIQUE and match the phenotype file.



Gene	Molecule	Sample_G1-1	Sample_G1-2	Sample_G1-3	Sample_G1-7	Sample_G12-1
A2M	NM_000014	0.0176404	0	0.00880505	0.00883818	0
NAT2	NM_000015	0	0	0	0	0
ACADM	NM_000016	1.62685	1.01542	1.42433	1.20325	1.16141
ACADS	NM_000017	1.03632	1.79022	0.59748	1.02009	2.01123
ACADVL	NM_000018	7.8576	5.9946	0.641386	2.06551	2.70341
ACAT1	NM_000019	1.28945	0.570424	0.550208	0.924286	0.759697
ACVRL1	NM_000020	0.000153667	0.000122771	0.0000807	0.0000814	0.000159191
PSEN1	NM_000021	1.5905	1.49027	1.73065	1.50645	2.3062
ADA	NM_000022	1.97053	3.7681	4.93193	5.28822	1.26921
SGCA	NM_000023	0.00000509	0.00000271	0.0642604	0	0

Data starts on line 2

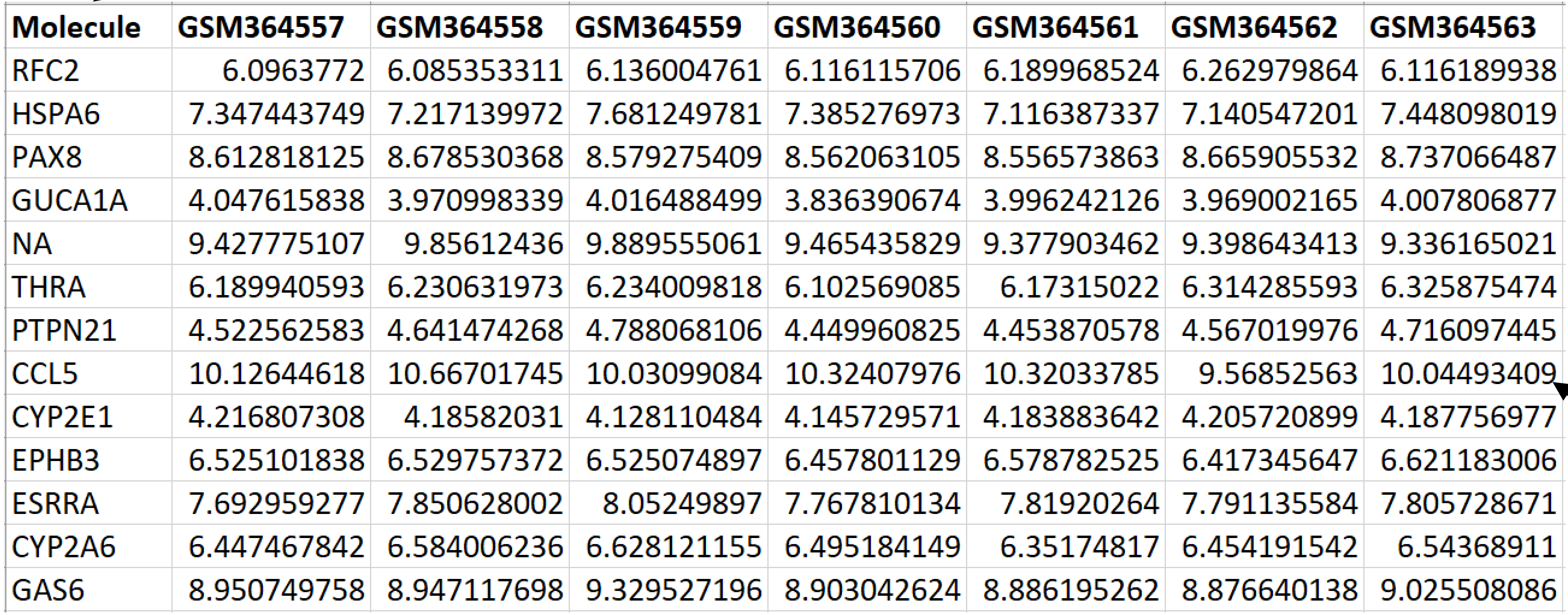
Each column contains measurements from 1 sample. Missing values are allowed (‘NA’ or leave empty).

Note: The transcript expression data file is formatted as a tab-separated file.

Gene Expression at the Gene Level
Data File Format

Column 1: Gene Name. Column name must be 'Molecule'. Other gene identifiers are accepted, although only gene symbol can identify synonyms of the identifier. This is REQUIRED. These must be UNIQUE and cannot be blank ('').

2nd column and onwards are the participant IDs. These must be UNIQUE and match the phenotype file.



Molecule	GSM364557	GSM364558	GSM364559	GSM364560	GSM364561	GSM364562	GSM364563
RFC2	6.0963772	6.085353311	6.136004761	6.116115706	6.189968524	6.262979864	6.116189938
HSPA6	7.347443749	7.217139972	7.681249781	7.385276973	7.116387337	7.140547201	7.448098019
PAX8	8.612818125	8.678530368	8.579275409	8.562063105	8.556573863	8.665905532	8.737066487
GUCA1A	4.047615838	3.970998339	4.016488499	3.836390674	3.996242126	3.969002165	4.007806877
NA	9.427775107	9.85612436	9.889555061	9.465435829	9.377903462	9.398643413	9.336165021
THRA	6.189940593	6.230631973	6.234009818	6.102569085	6.17315022	6.314285593	6.325875474
PTPN21	4.522562583	4.641474268	4.788068106	4.449960825	4.453870578	4.567019976	4.716097445
CCL5	10.12644618	10.66701745	10.03099084	10.32407976	10.32033785	9.56852563	10.04493409
CYP2E1	4.216807308	4.18582031	4.128110484	4.145729571	4.183883642	4.205720899	4.187756977
EPHB3	6.525101838	6.529757372	6.525074897	6.457801129	6.578782525	6.417345647	6.621183006
ESRRA	7.692959277	7.850628002	8.05249897	7.767810134	7.81920264	7.791135584	7.805728671
CYP2A6	6.447467842	6.584006236	6.628121155	6.495184149	6.35174817	6.454191542	6.54368911
GAS6	8.950749758	8.947117698	9.329527196	8.903042624	8.886195262	8.876640138	9.025508086

Data starts on line 2

Each column contains measurements from 1 sample. Missing values are allowed ('NA' or leave empty).

Note: The gene expression data file is formatted as a tab-separated file.

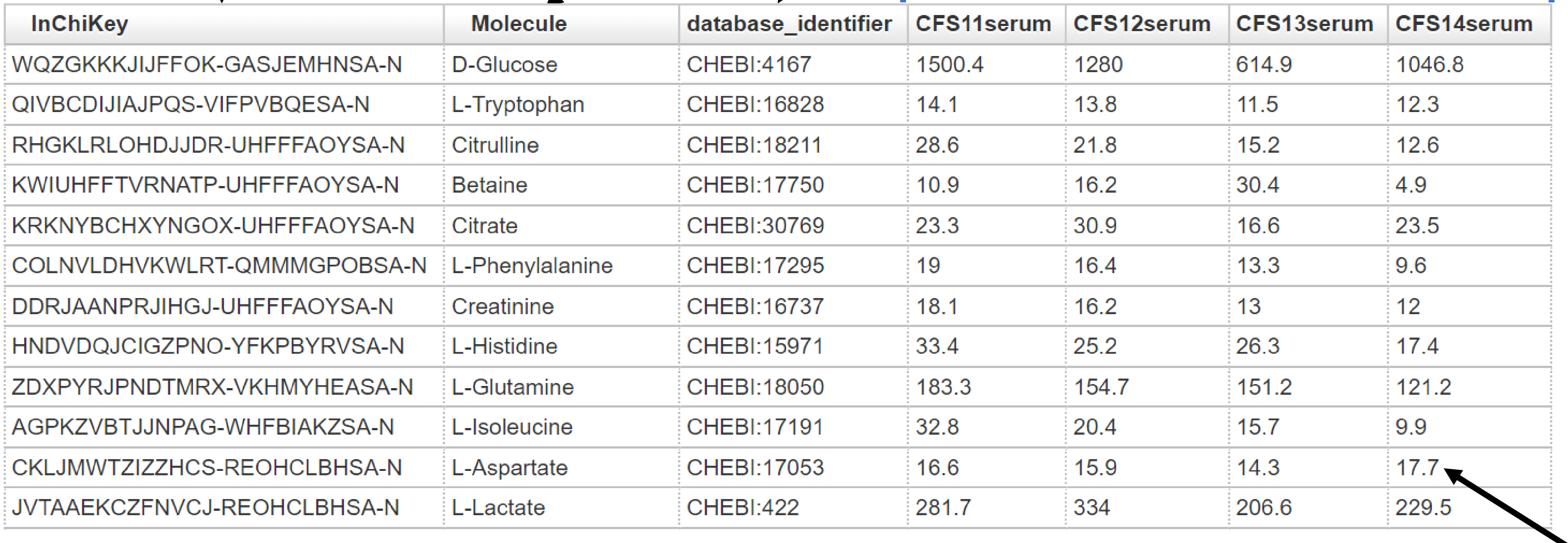
Metabolomics Data File Format

Column 1: InChiKey for the metabolite. Column name must be 'InChiKey'. These must be UNIQUE.

Column 2: Metabolite name. Column name must be 'Molecule'. These must be UNIQUE.

Column 3: CHEBI ID for the metabolite. Column name must be 'database_identifier'. These must be UNIQUE

4th column and onwards are the participant IDs. These must be UNIQUE and match the phenotype file.



InChiKey	Molecule	database_identifier	CFS11serum	CFS12serum	CFS13serum	CFS14serum
WQZGKKKJIJFFOK-GASJEMHNSA-N	D-Glucose	CHEBI:4167	1500.4	1280	614.9	1046.8
QIVBCDIJIAJPQS-VIFPVBQESA-N	L-Tryptophan	CHEBI:16828	14.1	13.8	11.5	12.3
RHGKLRLOHDJJDR-UHFFFAOYSA-N	Citrulline	CHEBI:18211	28.6	21.8	15.2	12.6
KWIUHFFTVRNATP-UHFFFAOYSA-N	Betaine	CHEBI:17750	10.9	16.2	30.4	4.9
KRKNYBCHXYNGOX-UHFFFAOYSA-N	Citrate	CHEBI:30769	23.3	30.9	16.6	23.5
COLNVLDHVKWLRT-QMMMGPOBSA-N	L-Phenylalanine	CHEBI:17295	19	16.4	13.3	9.6
DDRJAANPRJIHGJ-UHFFFAOYSA-N	Creatinine	CHEBI:16737	18.1	16.2	13	12
HNDVDQJCIGZPNO-YFKPBYRVSA-N	L-Histidine	CHEBI:15971	33.4	25.2	26.3	17.4
ZDXPYRJPNDTMRX-VKHMYHEASA-N	L-Glutamine	CHEBI:18050	183.3	154.7	151.2	121.2
AGPKZVBTJJNPAG-WHFBIAKZSA-N	L-Isoleucine	CHEBI:17191	32.8	20.4	15.7	9.9
CKLJMWTZIZZHCS-REOHCLBHSA-N	L-Aspartate	CHEBI:17053	16.6	15.9	14.3	17.7
JVTAAEKCZFNVCJ-REOHCLBHSA-N	L-Lactate	CHEBI:422	281.7	334	206.6	229.5

Data starts on line 2

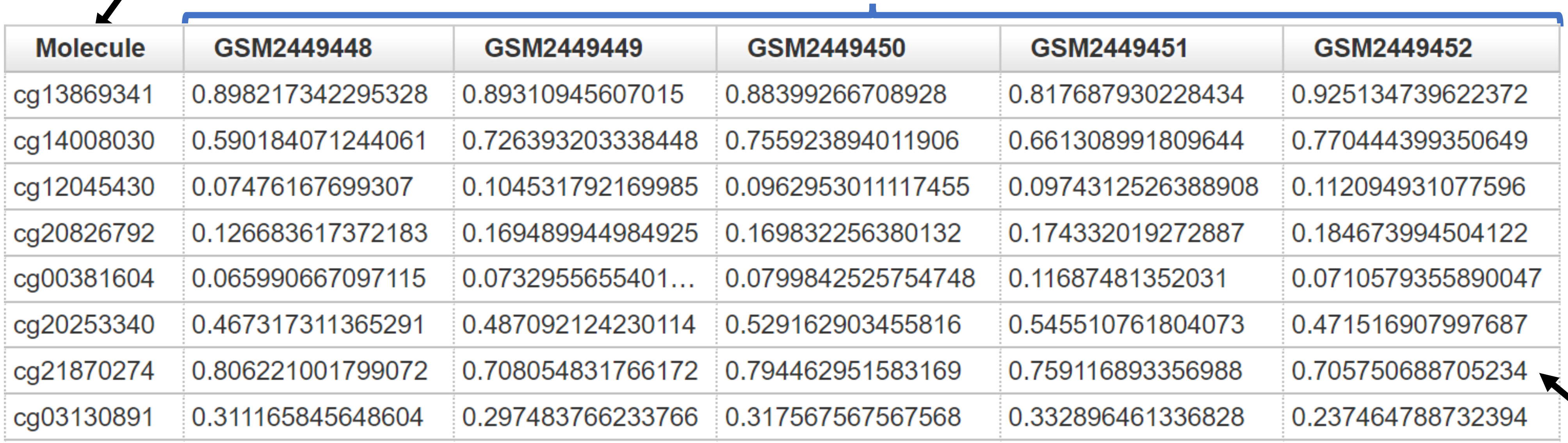
Note: The metabolomics data file is formatted as a tab-separated file.

Each column contains measurements from 1 sample. Missing values are allowed ('NA' or leave empty)

Methylation Data File Format

Column 1: CpG ID. Column name must be ‘Molecule’. This is REQUIRED. These must be UNIQUE

Further columns are the participant IDs. These must be UNIQUE and correspond to match the phenotype file.



Molecule	GSM2449448	GSM2449449	GSM2449450	GSM2449451	GSM2449452
cg13869341	0.898217342295328	0.89310945607015	0.88399266708928	0.817687930228434	0.925134739622372
cg14008030	0.590184071244061	0.726393203338448	0.755923894011906	0.661308991809644	0.770444399350649
cg12045430	0.07476167699307	0.104531792169985	0.0962953011117455	0.0974312526388908	0.112094931077596
cg20826792	0.126683617372183	0.169489944984925	0.169832256380132	0.174332019272887	0.184673994504122
cg00381604	0.065990667097115	0.0732955655401…	0.0799842525754748	0.11687481352031	0.0710579355890047
cg20253340	0.467317311365291	0.487092124230114	0.529162903455816	0.545510761804073	0.471516907997687
cg21870274	0.806221001799072	0.708054831766172	0.794462951583169	0.759116893356988	0.705750688705234
cg03130891	0.311165845648604	0.297483766233766	0.317567567567568	0.332896461336828	0.237464788732394

Data starts on line 2

Each column contains measurements from 1 sample. Missing values are allowed (‘NA’ or leave empty).

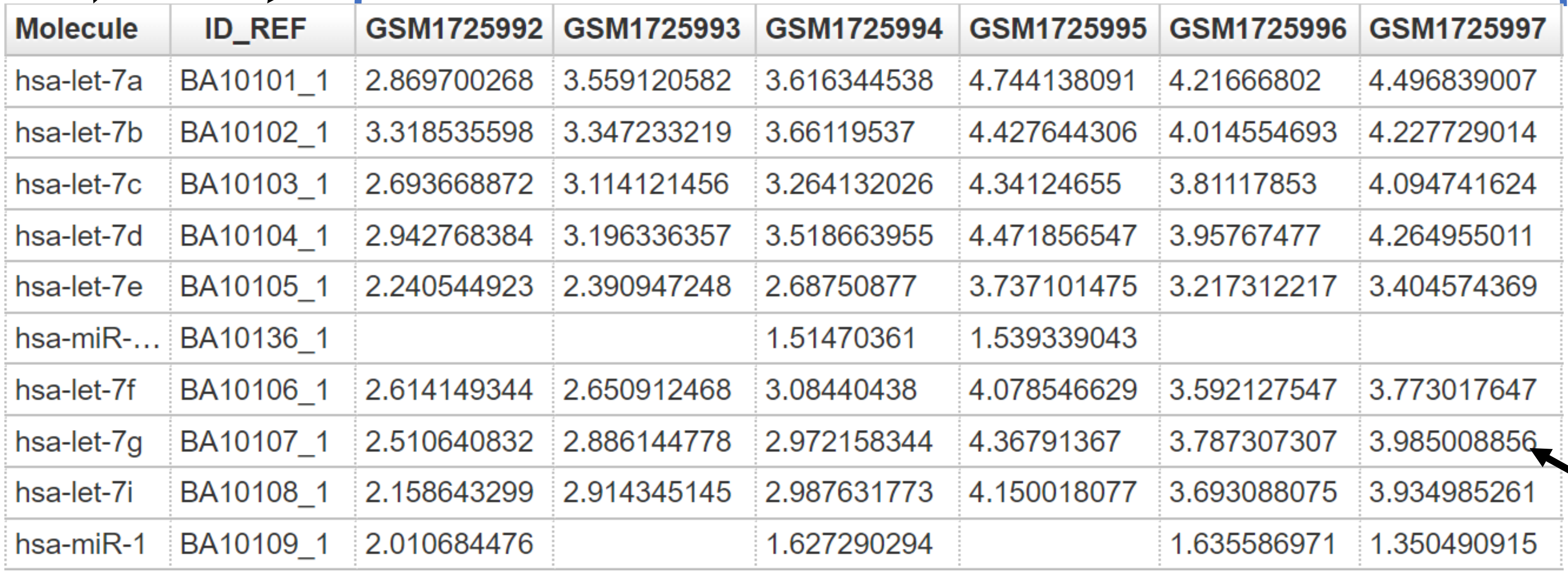
Note: The methylation data file is formatted as a tab-separated file.

miRNA Data File Format

Column 1: miRBase ID for the miRNA. Column name must be 'Molecule'. This is REQUIRED. These must be UNIQUE and cannot be blank ('')

Column 2: Other molecule identifiers. This is OPTIONAL.

2nd column and onwards headings are the participant IDs. These must be UNIQUE and match the phenotype file.



Molecule	ID_REF	GSM1725992	GSM1725993	GSM1725994	GSM1725995	GSM1725996	GSM1725997
hsa-let-7a	BA10101_1	2.869700268	3.559120582	3.616344538	4.744138091	4.21666802	4.496839007
hsa-let-7b	BA10102_1	3.318535598	3.347233219	3.66119537	4.427644306	4.014554693	4.227729014
hsa-let-7c	BA10103_1	2.693668872	3.114121456	3.264132026	4.34124655	3.81117853	4.094741624
hsa-let-7d	BA10104_1	2.942768384	3.196336357	3.518663955	4.471856547	3.95767477	4.264955011
hsa-let-7e	BA10105_1	2.240544923	2.390947248	2.68750877	3.737101475	3.217312217	3.404574369
hsa-miR-…	BA10136_1			1.51470361	1.539339043		
hsa-let-7f	BA10106_1	2.614149344	2.650912468	3.08440438	4.078546629	3.592127547	3.773017647
hsa-let-7g	BA10107_1	2.510640832	2.886144778	2.972158344	4.36791367	3.787307307	3.985008856
hsa-let-7i	BA10108_1	2.158643299	2.914345145	2.987631773	4.150018077	3.693088075	3.934985261
hsa-miR-1	BA10109_1	2.010684476		1.627290294		1.635586971	1.350490915

Data starts on line 2

Each column contains measurements from 1 sample. Missing values are allowed ('NA' or leave empty).

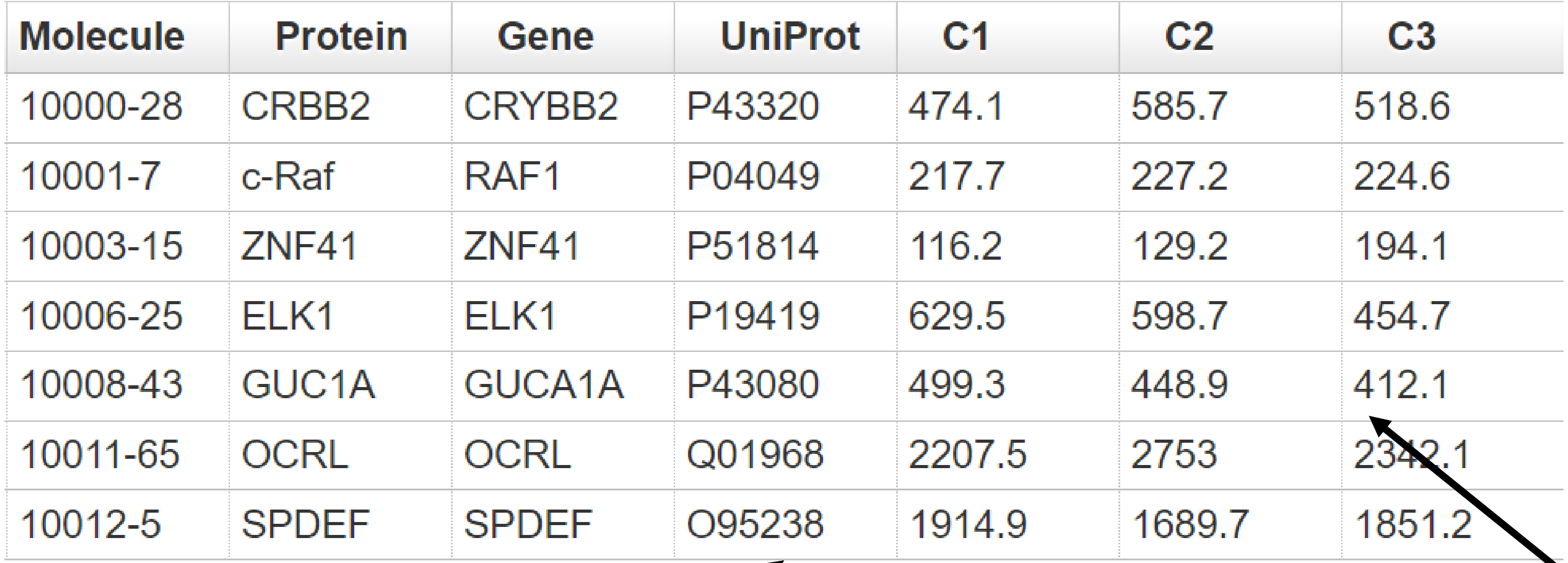
Note: The miRNA data file is formatted as a tab-separated file.

Proteomics at the Sequence Level
Data File Format

Column 1: Unique identifier for the sequence (e.g., SeqID). Column name must be ‘Molecule’. This is REQUIRED. These must be UNIQUE.

Columns 2 and 3: Target Gene and Protein Symbol. Column name must be ‘Gene’ and ‘Protein’ as appropriate. This is REQUIRED. These DO NOT have to be UNIQUE.

Further columns are the participant IDs. These must be UNIQUE and correspond to match the phenotype file.



Molecule	Protein	Gene	UniProt	C1	C2	C3
10000-28	CRBB2	CRYBB2	P43320	474.1	585.7	518.6
10001-7	c-Raf	RAF1	P04049	217.7	227.2	224.6
10003-15	ZNF41	ZNF41	P51814	116.2	129.2	194.1
10006-25	ELK1	ELK1	P19419	629.5	598.7	454.7
10008-43	GUC1A	GUCA1A	P43080	499.3	448.9	412.1
10011-65	OCRL	OCRL	Q01968	2207.5	2753	2342.1
10012-5	SPDEF	SPDEF	O95238	1914.9	1689.7	1851.2

Data starts on line 2

Column 4: Database identifier (e.g., UniProt, GI). The column heading must be included in the UniProt ID mapping for synonyms to be identified. This is OPTIONAL. These DO NOT have to be UNIQUE.

Note: The proteomics data file is formatted as a tab-separated file.

Each column contains measurements from 1 sample. Missing values are allowed (‘NA’ or leave empty).

Proteomics at the Protein Level
Data File Format

Column 1: Target Protein Symbol. Column name must be ‘Molecule’. This is REQUIRED. These must be UNIQUE.

Column 3: Database identifier (e.g., UniProt or GI). The column heading must be included in the UniProt ID mapping for synonyms to be identified. This is OPTIONAL. These DO NOT have to be UNIQUE.

Further columns are the participant IDs. These must be UNIQUE and correspond to match the phenotype file.

Molecule	Gene	UniProt	Participant1	Participant2
CRYBB2	CRYBB2	P43320	545.2	…
RAF1	RAF1	P04049	…	
ZNF41	ZNF41	P51814		
ELK1	ELK1	P19419		

Data starts on line 2

Column 2: Target Gene Symbol. Column name must be ‘Gene’. Other gene identifiers are accepted, although only gene symbol can identify synonyms of the identifier. This is REQUIRED. These DO NOT have to be UNIQUE.

Note: The proteomics data file is formatted as a tab-separated file.

Each column contains measurements from 1 sample. Missing values are allowed (‘NA’ or leave empty).

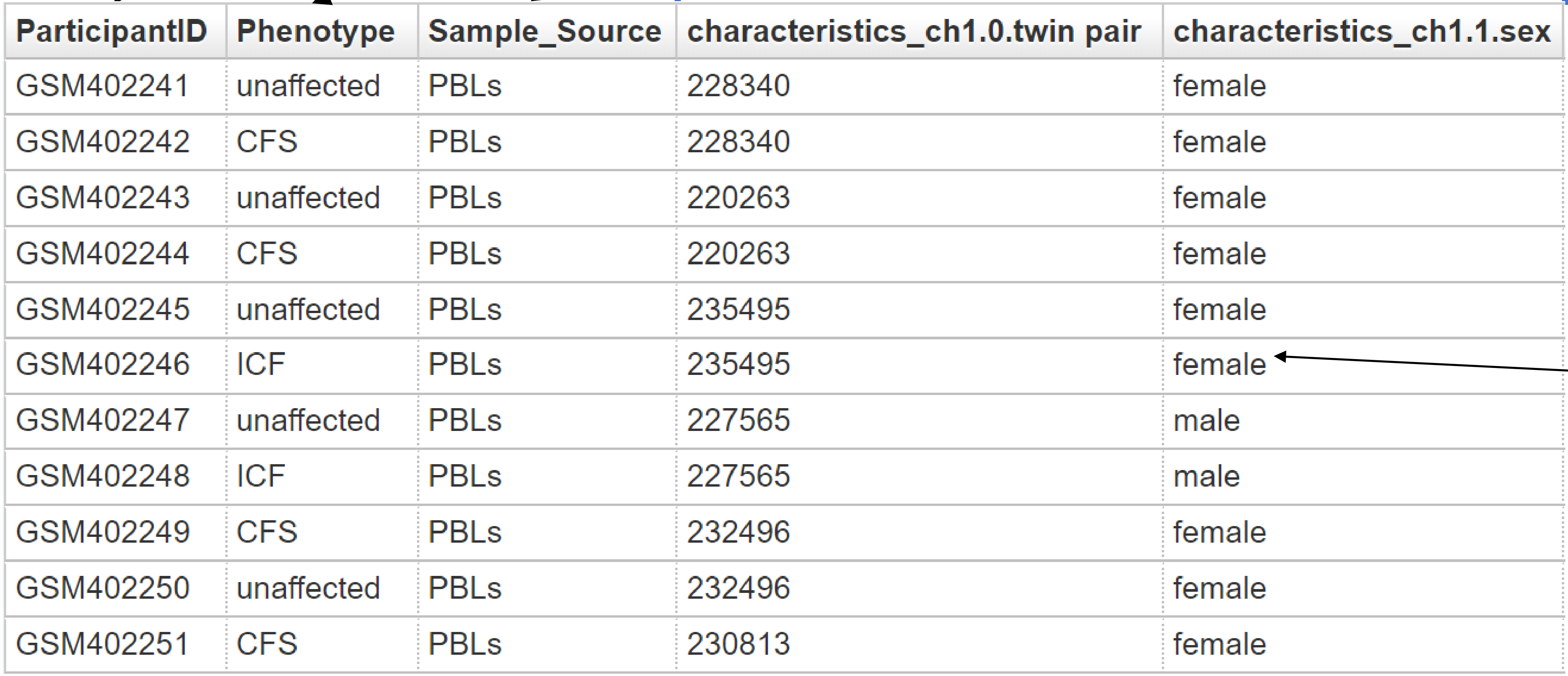
Phenotype File Format

Column 1: ID for each participant. Column header must be ‘ParticipantID’. Column is REQUIRED. These must be UNIQUE and match the data file.

Column 2: Phenotype of interest. Column header must be ‘Phenotype’. Column is REQUIRED.

Column 3: The source or tissue the sample was extracted from. Column header must be ‘Sample_Source’. Column is required. Missing is allowed (‘NA’ or empty string).

Columns are optional. You can substitute with any other information.



ParticipantID	Phenotype	Sample_Source	characteristics_ch1.0.twin pair	characteristics_ch1.1.sex
GSM402241	unaffected	PBLs	228340	female
GSM402242	CFS	PBLs	228340	female
GSM402243	unaffected	PBLs	220263	female
GSM402244	CFS	PBLs	220263	female
GSM402245	unaffected	PBLs	235495	female
GSM402246	ICF	PBLs	235495	female
GSM402247	unaffected	PBLs	227565	male
GSM402248	ICF	PBLs	227565	male
GSM402249	CFS	PBLs	232496	female
GSM402250	unaffected	PBLs	232496	female
GSM402251	CFS	PBLs	230813	female

Data starts on line 2

Each column contains data for 1 participant. Missing values are allowed (‘NA’ or leave empty).

Note: The phenotype file is formatted as a tab-separated file.

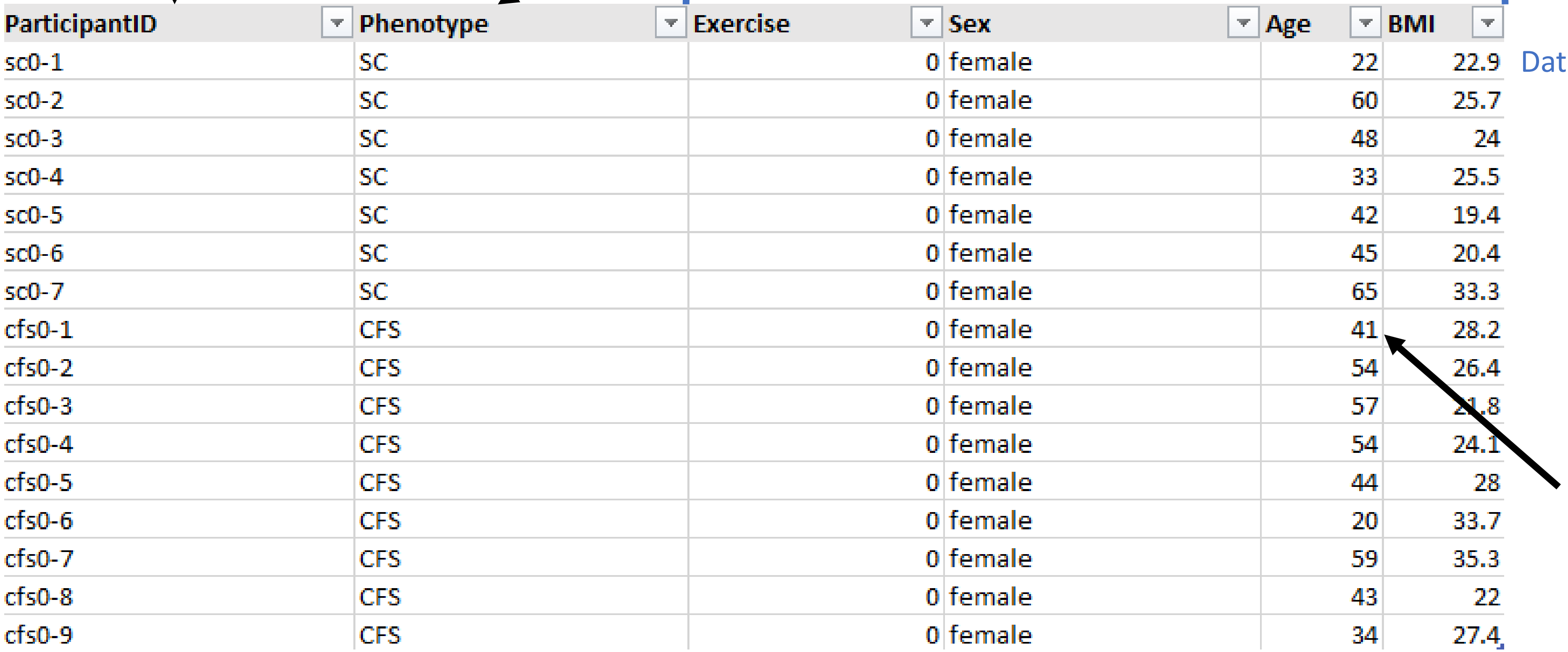
Demographic Health and Survey Phenotype File Format

Note: The DHS phenotype file does not have a required Sample_Source column

Column 1: ID for each participant. Column header must be ‘ParticipantID’. Column is REQUIRED. Values can be repeated for repeated measures.

Column 2: Phenotype of interest. Column header must be ‘Phenotype’. Column is REQUIRED.

Columns are optional. You can substitute with any other information



ParticipantID	Phenotype	Exercise	Sex	Age	BMI
sc0-1	SC	0	female	22	22.9
sc0-2	SC	0	female	60	25.7
sc0-3	SC	0	female	48	24
sc0-4	SC	0	female	33	25.5
sc0-5	SC	0	female	42	19.4
sc0-6	SC	0	female	45	20.4
sc0-7	SC	0	female	65	33.3
cfs0-1	CFS	0	female	41	28.2
cfs0-2	CFS	0	female	54	26.4
cfs0-3	CFS	0	female	57	21.8
cfs0-4	CFS	0	female	54	24.1
cfs0-5	CFS	0	female	44	28
cfs0-6	CFS	0	female	20	33.7
cfs0-7	CFS	0	female	59	35.3
cfs0-8	CFS	0	female	43	22
cfs0-9	CFS	0	female	34	27.4

Data starts on line 2

Each row contains values for one participant. Missing data is allowed (‘NA’ or leave empty).

Note: The file is formatted as a tab-separated file.

Data Dictionary File Format

Column 1: Human-readable variable name. Column header must be 'Variable'. Column is REQUIRED.

Column 2: Name of variable in dataset. Entries must not contain spaces. Column is REQUIRED.

Column 3: Type of variable in dataset. Allowed values are "character", "integer", and "numeric". Column is REQUIRED.

Column 4: Allowed values of data entries. For **numeric** variables, enter the minimum and maximum values separated by a '-'. For integer and character variables, enter all possible values separated by a ';'. Column is REQUIRED.

Column 5: Free-text description of the variable. Include units of measurement for relevant variables in this column. Avoid using commas if your data dictionary is formatted as a comma-separated file. Column is REQUIRED.

Column 6: Expected for integer values, as-needed for other variable types. Assign values using '=' and separate values using ';'. Include a definition of a missing value, as needed. Column is REQUIRED.

Variable	Variable_Name	Type	Allowed_values	Description	Label
Participant ID number	ParticipantID	character		ID number assigned to participant in sequential order	
ME/CFS status	Case	integer	0;1	Study physician diagnosis of ME/CFS or healthy control	0=Control;1=Case
Birth Sex	Female	integer	0;1	Self-reported sex assigned at birth	0=Male;1=Female
Age at enrolment	Age	numeric	18-75	Age at site visit	
Body Mass Index	BMI	numeric	15-45	Body Mass Index calculated as kg/m2	
Self-reported race	Race	integer	1;2;3	Category of race the participant most closely identifies with	1=white;2=African-American,Black;3=Asian,Pacific Islander;4=American Indian,Alaska Native;5=Unknown,Declined
Self-reported Hispanic ethnicity	Ethnicity	integer	1;2;3;4;5;6	Category of Hispanic ethnicity the participant most closely identifies with	1=Hispanic;2=non Hispanic;3=Unknown,Declined
Recruitement Site	Site	character	california;nevada;utah;florida;new_york	State containing the clinical recruitment site	
Resting heart rate	RHR	numeric	40-120	Heart rate measured after ten minutes of sitting, average of three	
Systolic blood pressure	SBP	numeric	70-140	Systolic blood pressure on dominant arm measured after ten minutes of sitting, average of three	
Diastolic blood pressure	DBP	numeric	40-90	Diastolic blood pressure on dominant arm measured after ten minutes of sitting, average of three	
Duration of ME/CFS symptoms	Duration	numeric	1-75	Self-reported duration of ME/CFS symptoms	
Self-reported IBS dianosis	IBS	integer	0;1;2	Self-reported diagnosis of Irritable Bowel Syndrome	0=no;1=yes,2=Unknown,Decline
Self-reported migraine diagnosis	Migraine	integer	0;1;2	Self-reported diagnosis of migraine	0=no;1=yes,2=Unknown,Decline
Self-reported allergy diagnosis	allergy	integer	0;1;2	Self-reported diagnosis of allergy	0=no;1=yes,2=Unknown,Decline
Thyroid stimulating hormone	TSH	numeric	0-20.0	Thyroid stimulating hormone (mU/L)	

Variables start on line 2

Each row contains information for one variable

Default missing values are 'NA' or empty. Explicitly declare others (eg, '-99') in the Label column

Note: the data dictionary can be formatted as a tab-separated file, comma-separated file, or Excel file.

Results File Format

Column 1: ID for the molecule measured (transcript IDs in this example). Column name must be ‘Molecule’. This is REQUIRED.

Column 2: Other OPTIONAL Identifiers.

Optional Columns: Various analysis results. Columns are flexible to the analysis conducted.

Pvalue: p-value of the test. Column name must be ‘Pvalue’.
PvalueAdj: adjusted p-value for the test. Column name must be ‘PvalueAdj’.

Molecule	Gene	NCases	NControls	baseMean	log2FoldChange	lfcSE	stat	Pvalue	PvalueAdj
NM_000014	A2M	100	100	1889.680119	-0.46319172	0.52404	-0.4848	0.80035	0.984757
NM_000015	NAT2	98	99	3490.874312	3.979896417	0.82957	7.2215	3.43E-05	0.000385
NM_000016	ACADM	99	98	1207.657311	-2.11547061	0.90685	-2.8381	0.07604	0.477411
NM_000017	ACADS	95	94	128.6574772	-5.80762897	0.82892	-7.4583	2.06E-10	2.23E-09
NM_000018	ACADVL	100	100	4203.536292	-0.97538594	0.51438	-0.8381	0.83524	0.874757

Data starts on line 2

Each column contains data for 1 molecule. Missing values are allowed (‘NA’ or leave empty).

Column 3 and 4: Number of cases (‘**NCases**’) and controls (‘**NControls**’) for each molecule. Use these columns for analyses where this varies by molecule. OPTIONAL

Note: The results file is formatted as a tab-separated file.