

Sage Bionetworks Data Curation Guidelines

31-March-2011
Version 2.1

Authors: Matt Furia
Solly Sieberts

Data Architects: Xavier Schildwachter
David Henderson

Contents

Introduction	2
Data Packet File Types	2
Data Files	2
Dictionary Files	3
Uniqueness of Identifiers	4
Supplementary Files	4
Data Layers	5
Gene Expression	6
Genotype	6
Phenotype	6
Copy Number Variation	6
miRNA Expression	6
Directory Structure	6
Building a Directory Structure	6
Location of Data and Dictionary Files	6
Location of Supplementary Files	7
A Generalized Data Packet	7
Discussion of Best Practices	9
Deidentified Human Data	9
Data File Column Headers	10
Data File and Dictionary Row Headers	10
Missing Values	10
Dictionary File Contents	10
Feature Dictionary	10
Individuals Dictionary	11
Marker Dictionary	11
Phenotype Dictionary	11
Reporter Dictionary.....	11
Probe Dictionary	11
Signal Data	11
Copy Number Variation	11
Gene Expression.....	11
Genotype	11
miRNA Expression.....	12
Phenotypes.....	12
Use of Supplementary Files	12
Data Integrity Checks	12
Phenotypes	12
Gender Assignment.....	12
Genotypes	12
Gender Inference	12
Replicate Individuals.....	13
Expression	13
CNV, miRNA and other data types	14
Appendix A: Additional Figures and Tables	15

Figures and Tables

Figure 1: Overview of curation process.....	3
Figure 2: Generalized directory structure of a data packet.....	8
Figure 3: Directory listing for a representative data packet.....	15
Figure 4: Screen shot of a phenotype dictionary.....	16
Figure 5: Screen shot of a CNV data file	17
Figure 6: Screen shot of an individuals dictionary	18
Figure 7: Screen shot of a phenotype data file	19
Table 1: Elements used in creating a data packet directory structure	7
Table 2: Examples of the controlled vocabulary to describe the expression data layer.....	20
Table 3: Examples of the controlled vocabulary used to describe genotype, cnv, mirna and phenotype data layers.	21

Introduction

A data packet is a set of files containing data gathered during a scientific experiment or clinical trial. In addition to data files, a packet contains metadata files including licenses, data dictionaries, readme files, citations, and scientific publications. The files in a packet are organized in a controlled directory structure designed to convey information about each file's contents and its relationship to constituent files.

Data contained in a packet have been assembled from public data repositories, or contributed by investigators at universities, research institutions or companies and have been aggregated into a simple, uniform format, making them more widely accessible than in their source files. These data have undergone a series of integrity checks to identify and correct common problems, but are otherwise preserved in the 'rawest' form available.

Herein we describe the directory structure, files and integrity checks that define a "curated" data packet. A cartoon of this process and the resulting data is shown in Figure 1. This work can be done manually or using the newly created sbnRepo R package. This package was created to facilitate data curation at Sage Bionetworks by providing utilities that automate common curation tasks. More detailed, up-to-date information about the package is available by contacting the authors of this document.

Data Packet File Types

Data Files

Data files are tab-delimited text files, containing a rectangular data matrix, that use row and column headers to uniquely identify each cell in the matrix. The first row of a data file contains the column header, with its first element being a text string specifying the type of identifier contained in the row header (see the "Dictionary ID name" column in Table 2 and Table 3). Each subsequent element of the column header contains an identifier that uniquely identifies each column. Similarly, the first column of a data file contains identifiers that uniquely identify each row.

Data files are organized such that each column contains data for a single individual and each row contains a single type of data, whether it is a phenotype, genotype or expression value. See Figure 1(B) and Figures Figure 4Figure 7 for examples.

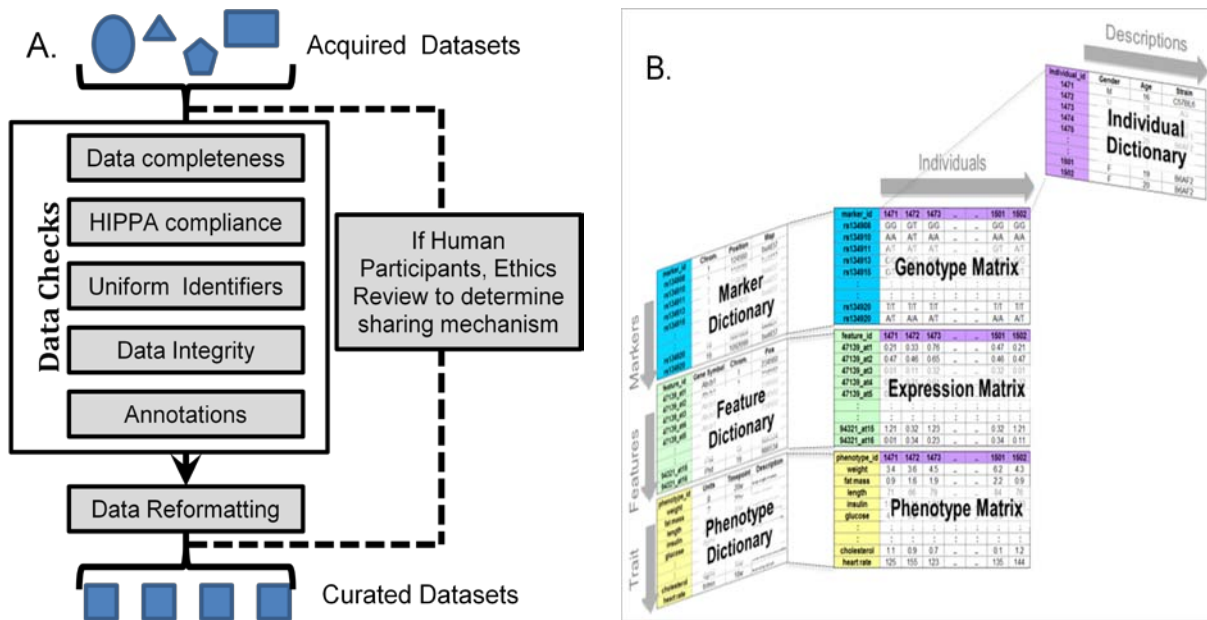


Figure 1: Overview of curation process. (A) Overview of curation process. Datasets, which come in many formats, are processed through a series of data checks and data transformations and presented as curated datasets in a uniform format. Data checks include: data completeness, HIPPA compliance (through the compliance workstream), uniformity of individual identifiers across data types, uniqueness of participant identifiers, uniqueness of genomic and phenotypic trait identifiers, data integrity and proper trait annotation. Data is then reformatted to the standard Sage Bionetworks flat file format shown in (B). Each data type is provided in a single flat file with columns representing individuals and rows representing traits. A separate data file is provided for each data value. A data dictionary file containing annotations by trait id is provided for each data type. An individuals dictionary containing annotations on study participants is also provided.

Dictionary Files

Dictionary files are tab-delimited text files containing metadata that describes the variables contained in the data files. A dictionary file is provided for each data layer within the packet. In some cases, a single dictionary file corresponds to multiple data files, as is the case when multiple statistics are available for a given measurement, but there is a 1:1 relationship between the row identifiers in the data file(s) and the identifiers in the associated dictionary file.

Each packet also contains a dictionary for the column identifiers (i.e. individuals). Because many or all data types may be included for the same individual, a data packet has only a single, study-wide dictionary that contains the superset of individual (column) identifiers used throughout the packet. Since all data files are organized such that the column identifiers uniquely identify an individual, this data dictionary is called the “individuals” dictionary. Each individual maps to exactly one identifier and the same individual identifier is used to associate the individual’s data in all data files throughout the packet. In addition to a list of individual identifiers, this file also enumerates which data (types) are available for each individual (see Figure 6).

Uniqueness of Identifiers

Within each data file, the row identifiers (which represent traits) are required to be unique and are required to have a 1:1 mapping to the identifiers in the associated dictionary file. There is no guarantee, however, of uniqueness between the row identifiers across data files within a packet or between data files in other packets. Future efforts are planned to standardize trait names (row identifiers) across packets by mapping them to a uniform ontology.

In addition, column identifiers (which represent individuals) are required to be unique within a data packet with a guaranteed 1:1 mapping between the identifiers used in a data file and the identifiers described in the individuals dictionary file. However, there is no guarantee that individual identifiers are unique across different data packets.

Supplementary Files

Metadata and additional information are also included in each data packet as supplementary files. These files describe details related to data generation and policies for the proper use of the data packet. All of these files are contained at the top level of the directory structure except item G. Supplementary files include:

- A) Terms of Use: text file that reiterates the terms of use that were agreed to by the end user upon download of the data packet. These require end users to cite and acknowledge the data generator or model builder. In some cases, these also outline restrictions on data use related to human protections or other requirements.
- B) Licenses: Data and models available through the Sage Bionetworks Repository are distributed under the terms of the Creative Commons 3.0 (CCBY 3.0) license. Code is distributed under the terms of the Apache 2.0 license. Both license are included in each data packet.
- C) Abstract: This file contains a brief description of the study from which the data was derived.
- D) Citation: This file contains the appropriate citation for data or models. Under the terms of use, this article should be cited in any publications that result from use of this dataset.
- E) Acknowledgement: This file contains a sentence acknowledging the data source that should be included in the acknowledgement section of any publication that results from use of this dataset. This is of particular importance when data is shared before a citation is available. (See below for details.)
- F) Readme_notes.txt: This file contains details about platforms, genome builds and file content descriptions. In addition, this file contains notes from the data curator that outline any issues or notable items that were discovered during data curation as well as a list and description of all files contained within the data packet.
- G) Curation Notes: When curation notes are specific to a single layer of data (e.g., expression traits), a curation note will be included within that layer of the directory structure.

Other supplementary files may include metadata about the experiments performed, including data of an individual-specific, sample-specific or experiment-specific nature. For individual-specific or sample-specific metadata, it is sometimes subjective what information belongs in the dictionary files and what belongs in supplementary files. As a general rule, information that applies to the entire data packet or data layer (defined below) that is not easily represented in tabular format should be included in a supplementary file. It is important to keep in mind that the tabular data in dictionary files will likely have higher visibility to users than information contained in supplementary files.

Acknowledgements.

Specific language for the acknowledgements is as follows:

The <Dataset Name> was generated and is contributed by <Contributor(s) name(s)> through the Sage Bionetworks Repository. The tissues were provided by <Contributor(s) name(s)> .

For example: The Human Liver cohort was contributed by Merck Pharmaceutical through the Sage Bionetworks Repository. The tissues were provided by Fred Guengerich, Stephen Strom, Erin Schuetz and Merck Research Laboratories.

Or,

The Harvard Brain dataset was contributed by Merck Pharmaceutical through the Sage Bionetworks Repository. The tissues were provided by Harvard Brain Tissue Resource Center which is supported in part by PHS grant R24 MH068855 (<http://www.brainbank.mclean.org/>).

Data Layers

A “data layer” refers to a specific type of data contained in a data packet. Examples of common layers are gene expression, genotype and phenotype data. It is anticipated that there will be increasing diversity in the “intermediate” molecular data layer as proteomic, miRNA and transcriptional sequencing technologies become more accessible. In data packets, controlled vocabulary is used to refer to each data layer type (Table 2 and Table 3).

Within a data layer, multiple statistics may be required in order to fully describe each data point. For example, the expression profile for a given probe may be described by a mean expression, standard deviation, and N beads. In this case, the complete description of the data within this layer would require three data files: one for each statistic. The type of data contained in each file will be described in the file name (e.g., `expression_avg_signal.txt`, `expression_array_stdev.txt`, `expression_nbeads.txt`). These files will have identical dimensions and contain the same row and column identifiers, though there is no guarantee that the rows and columns for each file will be in the same order. In this case, only one data dictionary is required to annotate the row entries of these files, because they are identical across files.

Gene Expression

The gene expression data layer contains data for mRNA expression profiles as measured by array technologies.

Genotype

The genotype data layer contains SNP genotypes and should contain results summarized into genotype calls and/or raw signals from microarrays.

Phenotype

The phenotype data layer contains all clinical traits and covariates. For example, gender and age are phenotypes that are traditionally considered covariates, but they should be included in the phenotypes data layer.

Copy Number Variation

The CNV data layer contains measurements of copy number variation.

miRNA Expression

The miRNA data layer contains data for micro-RNA expression.

Directory Structure

Building a Directory Structure

The directory structure for a data packet contains information describing the identity and characteristics of the data packet (Figure 2 and Figure 3). As such, the folders within the directory are subject to controlled vocabulary defined by Sage Bionetworks (Table 1, Table 2 and Table 3). The directory-naming scheme is as follows:

```
<packet name>/release_<release number>_<release date>/<data layer>/<platform>/<tissue>/<processing method>/
```

Location of Data and Dictionary Files

Dictionary and data files for a data layer reside side-by side under the appropriate data layer directory and are not allowed at higher levels in the directory structure. In the event that data from only one platform is present for a data layer, the platform subdirectory is optional. Similarly, if data are collected in only a single tissue, or if the signal data are not tissue specific, the “tissue” subdirectory is optional.

In the event that processed data are included in the data packet, dictionary and data files for these data should be placed in a subdirectory named using the name of the method used. In this case, the raw data can optionally be placed in a subdirectory named “raw”.

The platform “other” is used when the data platform is not known, or when data do not belong to a specific platform. This means that data meeting this description are placed in a subdirectory named “other”. Alternatively, data and dictionary files that are placed directly in the data layer directory are implicitly assigned to the “other” platform.

Table 1: Elements used in creating a data packet directory structure

Component	Examples	Description
<packet name>	hbtrc	A short string unique to this packet. Determined upon completion of initial curation
<release number>	1, 2, 3	Starts with 1 and increments with each subsequent release
<release date>	2011-01-24	Release date in the form YYYY-MM-DD
<data layer>	expression, genotype	A controlled vocabulary term specifying the type of data contained in the layer
<platform>	agilent, affymetrix, other, etc	Selected from a set of controlled vocabulary terms, specific to each data layer. This can be thought of as specifying the vendor or manufacturer
<tissue>	liver, adipose, prefrontalcortex	The tissue where the signal data was measured
<processing method>	raw, qc, rma, mas5	Selected from a set of controlled vocabulary terms. Specifies the processing method used to generate the enclosed data from the raw data.

Location of Supplementary Files

Every directory can contain supplementary data files that are associated with that “level”, and all sub levels, in the hierarchy. For example, a curation note that applies to all expression data in the data packet, regardless of platform, would be placed in the expression/ directory and a readme file that applies only to expression data generated on a specific platform (e.g. Agilent) should be placed in the expression/agilent/ directory.

While supplementary files are supported at every level in the hierarchy, it is generally a good idea to minimize their use and assemble all supplementary data at the root-level readme file.

A Generalized Data Packet

A generalized representation of a data packet is included in Figure 2. An example of the directory structure from an actual data packet is included in the Appendix (Figure 3).

Figure 2: Generalized directory structure of a data packet

```

+-<packet name>/
  +-<supplementary metadata files general
  |   to the packet layer>
+- release_<release num 1>_<release date 1>/
  +-individuals.txt (dictionary file)
  +-<supplementary metadata files
  |   specific to the release>
  +-<data layer1 name>/
    +-<platform 1>/
      +-<data layer1 name>.txt (data file)
      +-<data layer1 id name>.txt (dictionary file)
      +-<supplementary files specific
      |   to data layer 1, platform 1>
    +-<platform 2>/
      +-<data layer1 name>.txt (data file)
      +-<data layer1 id name>.txt (dictionary file)
      +-<supplementary files specific
      |   to data layer 1, platform 2>
    . . .
  +-<platform "n">/
    +-<layer1 name>.txt (data file)
    +-<data layer1 id name>.txt (dictionary file)
    +-<supplementary files specific
    |   to data layer 1 platform "n">
  +-<data layer2 name>/
    +-<supplementary files specific
    |   to data layer 2>
    +-<platform 1>/
      +-<data layer2 name>.txt (data file)
      +-<data layer2 id name>.txt (dictionary file)
      +-<supplementary files specific
      |   to data layer 2, platform 1>
    . . .
  +-<data layer "n" name>/
    +-<supplementary files specific
    |   to data layer "n">
    +-<platform 1>/
      +-<data layer "n" name>.txt (data file)
      +-<data layer "n" id name>.txt (dictionary file)
      +-<supplementary files specific
      |   to data layer "n", platform 1>
    . . .
  +-release_<release num n>_<release date n>
  . . .

```

Discussion of Best Practices

To improved the readability and consistency of data packets, the Sage Bionetworks Repository Data Curation Team has established a set of guidelines that all curators should follow whenever possible. The contents of this section are not exhaustive, but as a general rule, it is important to keep in mind that the primary goal of data curation is to make genomics data uniform and accessible to the widest possible audience. To this end, it is important to follow these guidelines.

Deidentified Human Data

All data derived from human participants must be scrubbed of any information that could be used to reidentify participants as per HIPAA privacy rule standards for de-identification found at 45 CFR 164.514. It is Sage Bionetworks policy that this be completed by data contributors prior to transfer of data to Sage Bionetworks. This includes:

- A) Names,
- B) All geographic locations smaller than a state (e.g., city, address, hospital),
- C) All dates smaller than a year (e.g., birth date, admission date, discharge date, date of death). Although ages may be included for most of the population, all individuals greater than 89 years of age must be categorized as 90 years old.
- D) Telephone numbers, fax numbers or electronic mail addresses.
- E) Social Security numbers
- F) Medical record numbers or health beneficiary plan numbers
- G) Account numbers, certification or license numbers
- H) Vehicle or device identifiers
- I) URLs or IP addresses
- J) Biometric identifiers such as fingerprints or voice prints
- K) Full face images

Additional “indirect” identifiers have also been identified that could allow patient re-identification when provided in combination. Indirect identifiers may also pose a risk to patient privacy, as outlined in various guidance documents from funding agencies, ethics resources or institutional statutes. In studies that contain multiple “indirect identifier” traits, guidance from an IRB, either engaged by Sage Bionetworks or the data contributor, may be required and action (e.g. removal, dual-coding, etc) taken prior to data release. These indirect identifiers include:

- a. Sex, ethnicity, age
- b. Rare disease or treatment,
- c. Illicit drug use or “risky behavior”
- d. Socio-economic data: income, education, “rare” occupation, place of work
- e. Multiple pregnancies
- f. Household or family composition
- g. Place of birth

While, Sage Bionetworks requests that data provider strip the data of all potential identifier per HIPAA rules, data curators should verify compliance to this rule and highlight both the non-compliant datasets as well as datasets with >3 indirect identifiers, as noted above, for further assessment **PRIOR** to dataset release to the Sage Bionetworks Commons.

Data File Column Headers

The column headers in a data set are used to uniquely identify a study participant and, as such, the same identifier must be used to describe data from an individual across all data layers within a packet. However, it is common for submitted data to contain multiple types of identifiers per individual. In extreme cases there can be a different identifier type for every data layer or assay. It is up to the curator to decide which of these to use as the primary identifier in the final data packet. As a rule of thumb, when multiple identifiers are present, the primary identifier should be the one that is the most globally unique or the one that is used in the majority of the submitted data files, as long as it does not contain information that could be used to re-identify a human participant.

Data File and Dictionary Row Headers

The row headers in data and dictionary files contain identifiers that uniquely identify a data value. For example, these identifiers could refer to a specific mRNA, SNP marker or clinical trait.

To enhance readability of the data and dictionary files, it is important that row IDs be selected that are as specific as possible while simultaneously being interpretable by the widest possible audience. For SNP data, dbSNP IDs are the preferred choice. For Agilent gene expression data, reporter IDs would be suitable. For Affymetrix expression data, feature IDs fit the criteria.

Special care should be taken when selecting the row headers for clinical phenotypes (and covariates). Phenotype IDs should be descriptive, but concise, text strings that are easily interpretable, and short enough that they are not burdensome to display when loaded into an analysis platform such as R. Additionally, phenotype IDs should not contain any special characters (examples: \$@%#-) or whitespace.

Missing Values

NA should be used to indicate missing values in all data and dictionary files.

Dictionary File Contents

Feature Dictionary

Feature dictionaries are specific to the Affymetrix platform and should include the probe sequence, chromosome, genome position and gene name. Additionally, the dictionary should indicate whether each probe is a perfect-match or mismatch.

Individuals Dictionary

The individuals dictionary should include a column for each identifier type in the data submission. **For human data, identifiers that could be used to reidentify study participant must be excluded.** An example of Individual Dictionary is shown in Appendix A: Additional Figures and Tables, Figure 6.

Marker Dictionary

The marker dictionary contains details about the SNPs in the data packet. As a minimum, it should include the dbSNP ID, chromosome, position and variants.

Phenotype Dictionary

The phenotype dictionary should include a descriptive name, a detailed description including units of measure when available, the timepoint at measurement, transform and adjustment applied. “NA” should be used to indicate when a timepoint or unit of measure is available. When no transform or adjustment has been applied to the trait values, the phenotype dictionary should indicate this by specifying “normal” and “None”, respectively. An example of phenotype dictionary is provide in Figure 4.

Reporter Dictionary

The reporter dictionary contains information about Agilent probes. It should include the probe sequence, chromosome, genome position and gene name.

Probe Dictionary

A probe dictionary is similar to a reporter dictionary, except that probes have a 1:1 mapping to a gene.

Signal Data

Copy Number Variation

CNV data values are signed floating point numbers for ratio-based platforms and non-negative integers for intensity-based platforms. In some cases, CNV data may be summarized as relative, or absolute counts (e.g -1, 2, -3), but raw signal values should be provided when available.

Gene Expression

Expression data values are signed floating point numbers for ratio-based platforms and non-negative integers for intensity-based platforms.

Genotype

Genotype calls should be coded by allele separated by a forward slash (i.e. A/T). Although less ideal, data can also be submitted as a numeric count of the number of reference alleles present (0, 1 or 2) or by specifying the allele genotypes in reference to parent A or B (i.e. AA, AB, or BB).

miRNA Expression

miRNA data values are signed floating point numbers for ratio-based platforms and non-negative integers for intensity-based platforms.

Phenotypes

Phenotypes can be either numeric or character strings.

Use of Supplementary Files

While the specification technically allows for supplementary data files at any level in the directory structure, these files should be used sparingly. As a general rule, supplementary files that are deeper in the directory structure have lower visibility than those at higher levels and should be included in the readme located in the top level of the dictionary. Furthermore, information in supplementary data files have lower visibility than data in dictionary files.

Data Integrity Checks

Whenever possible data should be checked for consistency in matching across data layers. Gender can be inferred based on genotype and normal-cell gene expression. (To date, more work needs to be done to assess the reliability of gender inference in tumor genotypes or expression profiles, CNV data and other data types.) When inferred gender does not match across data layers, this indicates an error in sample matching or recordkeeping, depending on the pattern of inconsistency. In addition, genotype data can be checked for unexpected duplicates. Methods and actions are described below.

Phenotypes

Gender Assignment

When the gender inferred from other data layers overwhelmingly disagree with the reported gender, the gender in the phenotype may be changed if there is sufficient reason to support this action. If there is additional evidence supporting the truth of the reported gender, the better action is to remove the sample from all other data layers. Always inquire with the data generator before steps are taken. In the case of mouse data, additional supporting evidence may come from cage data when available (i.e. by checking the consistency of gender within a shared cage).

Genotypes

Gender Inference

Gender inference based on genotypes is done using X-chromosome markers. Since males carry only a single X chromosome, their X-chromosome genotypes should never be heterozygous due to the fact that hemizygous genotypes appear as homozygotes in most genotyping technologies. On the other hand, females usually display both homo and heterozygous genotypes, though the proportion of each may vary heavily depending on the study design. For mouse intercrosses, females with no heterozygotes (or homozygotes) are not unusual. However, for human studies

with sufficient number of X-chromosome markers, females displaying only homozygous genotypes are rare.

For human data, the *-check-sex* option in Plink can be used to identify putative gender errors. By default, this function calls gender based on the (excess) heterozygosity of the X-chromosome markers. Here excess heterozygosity < 0.2 is called female and excess heterozygosity > 0.8 is called male.

For intercross mice, females are expected to be 50% AA and 50% AB, where A is the allele of the maternal strain. Males are expected to be 50% A and 50% B. However, given the relatively few meioses, the variance on these proportions is large and the pseudo-autosomal region can account for some additional heterozygosities.

Reported males with more than 5% heterozygosity should be flagged as problematic. When parental line data is available and cross order is known, females with more than 5% BB genotypes may be flagged as problematic in the absence of AB genotypes.

Replicate Individuals

Genotypes are checked for replicate individuals by looking for samples with identical genotypes within the limits of genotyping error. Pairs of individuals whose non-missing genotypes differ at 10% or less of loci can be considered identical and should be discarded if gender inference has not already excluded one of the pair. An alternate approach for human population data is to use the *--genome* function in Plink, which estimates the proportion of IBD sharing between two individuals. For siblings and parent-offspring pairs, this estimate should be close to 0.5. Any value above 0.75 is highly suspect (unless the pair is known to be related and highly inbred). A useful approach in this case is to use the *-min X* option, which outputs pairs with estimated IBD greater than *X*. Here the appropriate use is:

```
plink --file mydata --genome --min X
```

For studies in which individuals are expected to be unrelated, *X* may be as small as 0.1. However, if there are known to be related individuals, the user may wish to set *X* to be much higher.

Expression

Expression of Y-transcripts can be used to infer gender, though this approach may be unreliable for tumor tissue data. This remains an open area of research so one should use caution when applying these methods in this case. For somatic tissues, the following approach is typically reliable.

1. Identify the Y-transcripts using available annotations, Biomart or similar.
2. Individually examine boxplots by reported gender, to identify transcripts which differentiate males from females remembering that females should have lower expression than males. This step does not require statistical significance, just an “eyeball” difference in distributions.
3. Perform principle component analysis on the “good” Y-transcripts identified in (2). Typically the first (PC1), or first and second (PC2) principle components distinguish gender.

4. Plots of PC1 vs PC2 by reported gender can be used to identify appropriate cutoffs, though some amount of judgment may be necessary.

CNV, miRNA and other data types

CNV of somatic tissue may be used to infer gender, though CNV of tumor tissue may be unreliable for this purpose.

Due to the lack of known miRNA on the Y-chromosome, gender inference based on these data is not currently possible and no standard checks are currently done.

Appendix A: Additional Figures and Tables

Figure 3: Directory listing for a representative data packet

```
+-- pomp_breast_cancer
  +- release_01_2011-01-24
    +-readme.txt
    +-sage_bionetworks_user_agreement.pdf
    +-cc_license.txt
    +-ap_license.txt
    +-version.txt
    +-individuals.txt
    |
    +-expression
      +-illumina
        +-gene_expression_min_signal.txt
        +-gene_expression_avg_signal.txt
        +-gene_expression_max_signal.txt
        +-gene_expression_narrays.txt
        +-gene_expression_array_stdev.txt
        +-gene_expression_bead_stdev.txt
        +-gene_expression_detection.txt
        +-reporter.txt
      |
    +-genotype
      +-marker.txt
      +-genotype_call.txt
    |
    +-phenotype
      +-phenotype.txt
      +-description.txt
    |
    +-cnv
      +-nimblegen
        +-cnv_PM_532nm.txt
        +-cnv_PM_635nm.txt
        +-probe.txt
```

Figure 4: Screen shot of a phenotype dictionary file

trait_id	trait_name	timepoint	transform	adjustment	trait_description	units
rep	rep	NA	NA	NA	NA	NA
diet	diet	NA	NA	NA	diet of the mouse. Either "High" for high fat diet or "Low" for control diet. HFD= 45% of total calories from fat, 20% from protein, and 35% from carbohydrates. Control=10% of total calories from fat, 20% from protein, and 70% from carbohydrates.	NA
met	met	end	normal	None	number of pulmonary metastasis counted at sacrifice	NA
met_resid	met_resid	end	normal	None	NA	NA
ave_met_density	ave_met_density	end	normal	None	average metastatic density	NA
amd_resid	amd_resid	end	normal	None	average metastatic density resid	NA
ttw	ttw	end	normal	None	total weight of all tumors	grams
tumor_ax	tumor_ax	end	normal	None	total weight of the axillary tumors	grams
tumor_ing	tumor_ing	end	normal	None	total weight of the inguinal tumors	grams
weight_at_sacrifice	weight_at_sacrifice	end	normal	None	body weight at sacrifice (~11 weeks for females, ~14 weeks for males)	grams
percent_fat	percent_fat	end	normal	None	percent body fat at sacrifice	NA
tumor_count	tumor_count	end	normal	None	count of tumors at sacrifice	NA
ave_met_area	ave_met_area	end	normal	None	average metastatic area	NA
ave_mets	ave_mets	end	normal	None	average number of metastatic tumors	NA
ave_lung_area	ave_lung_area	end	normal	None	average lung area	NA
pre_tumor_percent_fat	pre_tumor_percent_fat	W7	normal	None	pre tumor percent fat	NA
tumor_onset_in_days	tumor_onset_in_days	NA	normal	None	days until first tumor was observed in mammary glands.	days
lean_mass_pre_tumor	lean_mass_pre_tumor	W7	normal	None	lean mass pre tumor	grams
lean_mass_at_sac	lean_mass_at_sac	end	normal	None	lean mass at sacrifice	grams
change_in_lean	change_in_lean	end	normal	None	change in lean mass	grams
fat_in_grams	fat_in_grams	end	normal	None	fat pad weight	grams
liver_wt	liver_wt	end	normal	None	weight of liver	grams
liver	liver	end	normal	None	weight of liver adjusted for body weight	grams
weight_3w	weight_3w	W3	normal	None	body weight at 3 weeks	grams
weight_6w	weight_6w	W6	normal	None	body weight at 6 weeks	grams
weight_9w	weight_9w	W9	normal	None	body weight at 9 weeks	grams

Figure 5: Screen shot of a CNV data file

	A	B	C	D	E	F	G	H	I	J
1	probe_id	208	306	608-3	708	808	1508	3005-8	3405	3509
2	CHR01FS003001832	2051.67	2222.44	1623.67	2639.11	1455.56	2111.78	2535	1905.56	2528.56
3	CHR01FS003018759	2358.78	2534.33	1983.78	2793.89	1896	2317.33	2797.89	1923.67	2877.78
4	CHR01FS003036253	1974.89	2110.22	1605.78	2834.67	1431.56	1576.44	2058.89	1549.22	2715.11
5	CHR01FS003041992	2514.44	3570.11	3012.89	4498.44	2731	3562	2894.89	3384.22	4504.22
6	CHR01FS003053606	3601.78	3164.78	2303.22	3807.78	2805.56	3420.22	2935.67	1891.11	3906.56
7	CHR01FS003065156	2285.56	2270.67	1598.78	2895.33	1748.56	1985.44	2438.56	1897.78	3128.44
8	CHR01FS003076536	3410.44	3664.56	2881.33	4104	2787.22	4061.56	3066.56	2716.22	3455.78
9	CHR01FS003087994	5185.89	4048	3393.89	3983	3741.56	5023.44	2956.67	4398.33	2554.44
10	CHR01FS003093673	5488.56	5167.11	4089.33	6820.67	4618.56	7096	5071.67	6626.44	4900.56
11	CHR01FS003105423	7739.22	6941.11	5710.78	8645.56	5414.56	7634.56	5721.78	5962.11	6765.11
12	CHR01FS003110919	2610	2574.33	1820.44	3299.22	1855.33	2876.78	2475.33	2035.44	3071.22
13	CHR01FS003116559	2561	3017.56	2278	3135.33	2018.22	2780.11	3403	2608.44	3674
14	CHR01FS003122452	2679.89	2978.89	2031.33	3317.44	2267.22	3220.11	3152.33	2588.11	3466.22
15	CHR01FS003134224	2094.78	2169.56	2135	3617.56	2316.78	1854.56	1689.67	2127.33	3005.11
16	CHR01FS003156928	2963.22	3808.67	2229.89	4726.67	2323.44	3479.22	3089.78	2940.56	3726.56
17	CHR01FS003162762	2981.11	4104.67	2898	4755.44	2277.22	3132.44	2885.78	3085.44	4691.56
18	CHR01FS003174487	1700.44	2028.33	1387.33	2401.33	1479.67	1763.56	1684.33	1339	3236.78
19	CHR01FS003180251	2045.89	2506	1585.67	3124	2140.78	2092.56	1813	1459.89	2543.44
20	CHR01FS003185980	2095.44	2906	2078.56	3340.11	1881.11	2472.11	2762.33	1658.22	3460.22
21	CHR01FS003191525	2578.89	2880.44	1732.22	3024.22	2101.78	2470.56	3251.78	1897.11	2851.33
22	CHR01FS003197138	3503.11	3528.56	2558.56	4453.89	3107.89	3620.33	2740.67	2428.11	4290.78
23	CHR01FS003203195	2465.67	3019.56	1593.89	2946	1749.22	2221.33	3155.56	2016.33	2882.44
24	CHR01FS003208636	4215.89	4776.22	2406.89	5563.22	3278.56	3573.89	3873.56	3017.67	4379.78
25	CHR01FS003214407	5055.56	4491.22	3624.22	7297.56	3647.44	4251	3456.56	2898	6260.89
26	CHR01FS003220004	1550.44	2245.56	1561.56	2947.78	1172.44	1642.78	2031.44	1527	2957.11
27	CHR01FS003237597	5313.33	5769.56	5752.89	8602.44	6573.78	6394.22	4090.78	4721.56	7261.44
28	CHR01FS003243214	2450	2856.89	1984	2604.44	1874	2951.22	2872.22	2417.67	3206.33
29	CHR01FS003254921	3830.67	4506.44	2523.67	4956.67	2794.11	3392.89	3112.56	3066.11	4835.11
30	CHR01FS003260419	4099.56	5091	4497.89	7546.67	3890.22	5622.78	4565.11	4534.44	5387.56
31	CHR01FS003277959	2672.33	2988.44	2570.89	3788.44	1992.33	3252.78	3272.78	2654.11	3175.89
32	CHR01FS003283516	7081.11	8614.33	6570.22	9861.56	6153.67	9314.89	8655.67	9122.11	8425.56
33	CHR01FS003289244	3677	3008.56	3226.56	5246.67	2828	3725.67	2322.33	3180.89	4173.67
34	CHR01FS003300897	2479.44	2745.78	1602.78	3358.67	2301.11	2109.44	3369.11	1812.44	2762
35	CHR01FS003312249	3450.22	3145.22	2748	4889.56	2200.89	3609.33	2912.44	2764	4653.44
36	CHR01FS003323571	3951.56	3831.56	2271.22	3995.89	2402.56	3349.56	3063.33	2970.11	3577.78
37	CHR01FS003329326	3927.44	3929	5821.44	5701.22	4369.89	5909.33	3169.11	4874.89	3696.33

Figure 6: Screen shot of an individuals dictionary

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G
	individual_id	cnv_cy3_sample_name	cnv_cy5_sample_name	gene_expression	genotype	phenotype	cnv
2	105	NA	NA	yes	yes	yes	no
3	106	106a	106t	yes	yes	yes	yes
4	1104	1104a	1104t	yes	yes	yes	yes
5	1107	NA	NA	yes	yes	yes	no
6	1204	1204a	1204t	yes	yes	yes	yes
7	1206	1206a	1206t	yes	yes	yes	yes
8	1308	NA	NA	yes	yes	yes	no
9	1409	NA	NA	yes	yes	yes	no
10	1507	NA	NA	yes	yes	yes	no
11	1508	1508-6	1508-26	yes	yes	yes	yes
12	1609	NA	NA	yes	yes	yes	no
13	1804	NA	NA	yes	yes	yes	no
14	1808	NA	NA	yes	yes	yes	no
15	1905	1905a	1905t	yes	yes	yes	yes
16	208	208-1	208-21	yes	yes	yes	yes
17	2206	NA	NA	yes	yes	yes	no
18	2209	NA	NA	yes	yes	yes	no
19	2303	NA	NA	yes	yes	yes	no
20	2304	NA	NA	yes	yes	yes	no
21	2406	NA	NA	yes	yes	yes	no

Figure 7: Screen shot of a phenotype data file

	A	B	C	D	E	F	G	H	I	J	K
1	phenotype_id	105	106	208	303	306	308	603	604	606	608
2	rep	A	A	A	A	A	A	A	A	A	A
3	diet	Low	High	Low	High	Low	Low	High	High	Low	Low
4	met	0	6	6	0	9	2	9	4	2	7
5	met_resid	3.2109888	6.01885507	3.2109888	6.01885507	3.2109888	3.2109888	6.01885507	6.01885507	3.2109888	3.2109888
6	ave_met_density	0	2	14.08	13.06	10.11	1.25	3.75	2.67	2.36	0.83
7	amd_resid	2.1007416	4.2048902	2.1007416	4.2048902	2.1007416	2.1007416	4.2048902	4.2048902	2.1007416	2.1007416
8	ttw	1.511	8.548	11.129	8.829	5.972	2.663	13.361	5.356	7.483	5.649
9	tumor_ax	1.011	4.531	6.987	6.111	4.489	2.142	6.987	3.497	4.46	3.388
10	tumor_ing	0.5	4.017	4.142	2.718	1.483	0.521	6.374	1.859	3.023	2.261
11	weight_at_sacrifice	32.01	38.49	48.86	44.14	40.04	31.76	50.21	40.14	43.64	36.92
12	percent_fat	24.92	13.84	15.07	18.9	19.93	18.94	14.24	17.14	16.67	16.97
13	tumor_count	7	10	10	10	7	6	10	10	10	10
14	ave_met_area	0	26384.71	1381.02	968.56	5831.86	949	4483.38	5863.6	6119.63	10589.67
15	ave_mets	0	0.7	4.8	3.9	2.9	0.4	1.3	1	0.8	0.3
16	ave_lung_area	0	350300	341000	298700	286754.5	321100	346900	374100	339000	363400
17	pre_tumor_percent_fat	14.5	13.7	16.4	18.4	15.1	12.7	15	11.5	17.6	14
18	tumor_onset_in_days	65	41	NA	NA	41	52	40	45	45	40
19	lean_mass_pre_tumor	20.35517	20.66645	21.47749	23.0768	20.16679	20.06337	21.61463	23.34723	24.07383	19.84418
20	lean_mass_at_sac	21.29269	27.91584	34.95719	31.5479	26.69534	20.27311	36.15642	27.27751	27.76556	25.83063
21	change_in_lean	0.937519	7.249385	13.479698	8.471099	6.528548	0.209733	14.54179	3.930279	3.691728	5.986449
22	fat_in_grams	7.976892	5.327016	7.363202	8.34246	7.979972	6.015344	7.149904	6.879996	7.274788	6.265324
23	liver_wt	1.682	1.77	2.602	2.129	2.244	1.462	3.038	2.132	2.447	1.915
24	liver	5.254608	4.598597	5.32542	4.82329	5.604396	4.603275	6.050588	5.31141	5.607241	5.186891
25	weight_3w	12.851	11.491	12.367	14.859	14.677	11.075	11.973	12.182	13.82	10.922
26	weight_6w	27.481	25.275	29.651	31.9	30.381	24.645	29.785	28.24	31.085	26.165
27	weight_9w	29.972	30.836	38.079	35.258	32.981	27.551	35.013	32.344	34.879	29.786
28											
29											
30											

Table 2: Examples of the controlled vocabulary to describe the expression data layer

Data layer	Platform	Data file prefix	Dictionary ID name	Dictionary file name	Required data values	Data file name	Data description
expression	agilent	expression	reporter_id	reporter.txt	g_intensity	expression_g_intensity.txt	green intensity
					r_intensity	expression_r_intensity.txt	red intensity
					mlavg	expression_mlavg.txt	mean log average intensity between channels across a flour-reversed pair
					mlratio	expression_mlratio.txt	mean log expression ratio across a flour-reversed pair
					pvals	expression_pvals.txt	pvalue associated with mlratio ratio
	affymetrix	expression	feature_id	feature.txt	pm	expression_pm.txt	perfect match intensity value
					mm	expression_mm.txt	mismatch intensity value
					sd	expression_sd.txt	standard deviation
	illumina	expression	reporter_id	reporter.txt	detection	expression_detection.txt	detection
					avg_signal	expression_avg_signal.txt	average signal
					avg_nbeads	expression_avg_nbeads.txt	average number of beads
					min_signal	expression_min_signal.txt	minimum signal
					max_signal	expression_max_signal.txt	maximum signal
array_stdev					expression_array_stdev.txt	array standard deviation	
bead_stdev					expression_bead_stdev.txt	bead standard deviation	
narrays	expression_narrays.txt	number of arrays					

Table 3: Examples of the controlled vocabulary used to describe genotype, cnv, mirna and phenotype data layers.
Please contact the authors for controlled vocabulary to describe additional types of data.

Data layer	Platform	Data file prefix	Dictionary ID name	Dictionary file name	Required data values	Data file name	Data description
genotype	affymetrix	genotype	marker_id	marker.txt	call	genotype_call.txt	genotype call
	perlegen	genotype	marker_id	marker.txt	call	genotype_call.txt	genotype call
	nimblegen	genotype	marker_id	marker.txt	call	genotype_call.txt	genotype call
cnv	agilent	cnv	probe_id	probe.txt	rawrat	cnv_rawrat.txt	raw ratio
					log2rat	cnv_log2rat.txt	log base 2 transformed ratio
					log2stddev	cnv_log2stddev.txt	standard deviation of the log base 2 ratio
					nreplic	cnv_nreplic.txt	number of replicates
					bad_p	cnv_bad_p.txt	bat probe flag
	nimblegen	cnv	probe_id	probe.txt	PM_532nm	cnv_PM_532nm.txt	perfect match intensity at 532nm
					PM_635nm	cnv_PM_635nm.txt	perfect match intensity at 635nm
mirna	agilent	mirna	reporter_id	reporter.txt	glsGeneDetected	mirna_glsGeneDetected.txt	flag indicating whether the gene was detected
					gTotalGeneSignal	mirna_gTotalGeneSignal.txt	total gene signal
					gMeanSignal	mirna_gMeanSignal.txt	mean signal
					gMedianSignal	mirna_gMedianSignal.txt	median signal
					glsSaturated	mirna_glsSaturated.txt	flag indicating where reporter was saturated
phenotype	other	NA	phenotype_id	description.txt	NA	phenotype.txt	clinical phenotypes and covariates

Appendix B: Checklist for External Curators

The curated dataset release check list

Please use this as a guide to verify datasets are complete and curated.

Dataset Name: _____

Curator (and contact): _____

Date: _____

This dataset contains (list all data types):

Type of data:	Yes/No	Platform	Tissue	N
Genotypes				
Phenotype Traits				
Intermediate Traits 1				

Confirmation of Data Curation:

- Participant identifiers match across genotype, intermediate and phenotype datasets.
- Genotypes were checked for gender consistency
- All intermediate traits were checked for gender consistency (if possible)
- A genotype annotation file was created that maps the genotypes to the genome
- Is the genome build used for genotype file known? If so: _____
- An expression annotation file was created that maps the expression probes to the genome
- Is the genome build used for expression file known? If so: _____
- A phenotype description file was created that provides a description of each trait, listed by trait ID.
- An individual.txt file was created that links all study participants across datasets and provides a description of the type of data available for each participant.

A README_note.txt file was created that describes:

- The platforms used to generate each type of data, (eg, "genotypes were measured using the illumina 610k quad beadarray. Expression traits were measured in adipose and brain tissue using the Illumina Ref8v3 expression beadarray").
- A description of the type of data and source of data for each data file (eg, "gene_expression_average_intensity.txt contains the average intensity across all probe replicates for each expression trait probe as measured by Illumina's GenomeStudio").
- Any notable deviations or issues identified during the curation process (eg, "data generator noted that there was a big effect of batch within the adipose gene expression traits. Within this dataset, we found and discarded 10 samples that demonstrated for which the inferred gender did not match the described gender").

Blanks are coded as NAs within:

- Genotypes
- Intermediate traits
- Phenotypes
- Replicate samples have been removed from genotype files.
- Replicate samples have been removed from intermediate files.