**Raw Data (Secure Access)**

Alignment files for the various experiments, chip arrays for the SNP genotyping assays and phenotype metadata for the different studies under the consortium; links provided are for the secured data sources on Synapse.

**RAW DATA DESCRIPTION**

Description of all datasets in different formats: [Description in .tsv format (link to www.synapse.org) »](http://www.synapse.org/)

Please clarify. Is this a list of all files with metadata describing samples and file formats. Basically what you have now in the [Capstone file view](https://www.synapse.org/#!Synapse:syn8466658/tables/)). Or, is the intent to describe the studies the data comes from.

**RNA-SEQ FASTQ AND ALIGNMENT FILES**

The RNA-seq fastq and alignment files are available on Synapse under condition of secure and approved access.

1. Fastq files for each sample: [RNA-seq fastqs (link to www.synapse.org) »](http://www.synapse.org/)
2. Bam files for each sample: [RNA-seq bams (link to www.synapse.org) »](http://www.synapse.org/)

According to Dominic, all the Capstone processed RNAseq have both aligned and unaligned reads. Suggest we do not provide fastq in addition

There are 2 bams per sample: Sorted by genome and transcriptome. Do we link to both?

**CHIP-SEQ FASTQ AND ALIGNMENT FILES**

The ChIP-seq fastq and alignment files are available under condition of secure and approved access.

1. Fastq files for each sample: [ChIP-seq fastqs (link to www.synapse.org) »](http://www.synapse.org/)
2. Bam files for each sample: [ChIP-seq bams (link to www.synapse.org) »](http://www.synapse.org/)

According to Eugenio, the Capstone processed ChIPseq bams only have aligned reads. There are 3 options: 1) We provide only the aligned reads fort now. 2) Separate files with the unaligned reads are uploaded to Synapse. 3) We provide the fastq. Option 3 is the least ideal to storage and download cost, and we only have the fastq for 2 of the studies in Synapse (CNON and EpiGABA), the rest were transferred directly to the 2 ChIPseq Capstone pipeline groups.

**SNP GENOTYPES**

The SNP-chip genotyping information for the samples are available under condition of secure and approved access.

1. Chip arrays: [SNP Chip Arrays (link to www.synapse.org) »](http://www.synapse.org/)

All the raw SNP genotypes that have been imputed are in Synapse, but we will need to know if any sample filtering was done in the imputed set so that they are concurrent. Should we provide only the imputed dataset?

**PHENOTYPES**

The sample associated phenotype information are available under condition of secure and approved access.

1. Sample IDs and their corresponding phenotypes: [Phenotypes in .tsv format (link to www.synapse.org) »](http://www.synapse.org/)

This is currently available in 2 tables. The [Capstone data fileview](https://www.synapse.org/#!Synapse:syn8466658/tables/) that has sample names and sample covariates + the individual IDs, and a clinical file with the individual IDs that has age of death and [other information depending on the sample source](https://www.synapse.org/#!Synapse:syn8399269/tables/).

Is what you are looking for here a single table that combines the sample and individual information?

**Pipeline-Processing Results**

RNA-seq quantifications, ChIP-seq signals and peaks, Brain Transcriptionally Active Regions (TARs), Imputed Genotypes (secured),and Phenotypes.

**RNA-SEQ QUANTIFICATIONS**

The annotated region quantifications for the uniformly processed RNA-seq data for each sample: [RNA-seq quantifications in .tsv format »](http://psychencode.gersteinlab.org/)

Should this link be to Synapse? The data is already there available as gene and isoform counts. Should both be included?

**CHIP-SEQ SIGNALS AND PEAKS**

The signal tracks and peak files for the uniformly processed ChIP-seq data for each sample.

1. Signal tracks calculated using PsychENCODE pipeline: [ChIP-seq Signal Tracks in .bw format »](http://psychencode.gersteinlab.org/)

Also in Synapse

1. Peak files calculated using PsychENCODE pipeline: [ChIP-seq Peaks in .bed format »](http://psychencode.gersteinlab.org/)

Also in Synapse, available as bed and ENCODE gappedPeak, narrowPeak and broadPeak. Link all 4?

**BRAIN TRANSCRIPTIONALLY ACTIVE REGIONS (TARS)**

TARs refer to the non-canonical transcription in the regions excluding protein-coding exons, annotated ncRNAs and pseudogenes. TARs were found by first finding all unannotated regions within each PsychENCODE sample that have RNA-seq RPM values greater than a certain threshold, and then considering those regions that have expression above the given threshold in a significant fraction of the samples.

1. Genome-wide TARs within each sample: [Sample-specific Genome-wide TARs in .bed format »](http://psychencode.gersteinlab.org/)
2. Genome-wide TARs that are found within at least 70% of the individuals: [Consensus Genome-wide TARs in .bed format »](http://psychencode.gersteinlab.org/)

Are these [the following files](https://www.synapse.org/#!Synapse:syn8466658/tables/query/eyJzcWwiOiJTRUxFQ1QgKiBGUk9NIHN5bjg0NjY2NTgiLCAic2VsZWN0ZWRGYWNldHMiOlt7ImNvbmNyZXRlVHlwZSI6Im9yZy5zYWdlYmlvbmV0d29ya3MucmVwby5tb2RlbC50YWJsZS5GYWNldENvbHVtblZhbHVlc1JlcXVlc3QiLCAiY29sdW1uTmFtZSI6InJ1blR5cGUiLCAiZmFjZXRWYW). There are multiple files per sample. Link all?

**IMPUTED GENOTYPES (SECURED)**

The imputed genotypes derived using the raw SNP genotyping data are available from Synapse, but only under secure and approved access.

1. Sets of imputed genotypes for all individuals in the analyses: [Imputed Genotypes (link to www.synapse.org) »](http://psychencode.gersteinlab.org/www.synapse.org)

The imputed genotypes are available as a file with multiple samples per chromosome. I suggest we [add this table that lists](https://www.synapse.org/#!Synapse:syn10909366/tables/) the individual IDs genotyped per study, the genotypingID and the platform

* HiC Maps
* What are the exact samples considered as being part of the adult studies. For example, the LIBD\_sczControl study has both fetal and adult samples and the CMC\_HBCC study individuals down to ~ 2 years. Are all samples from those studies considered 'adult'?
* Ideally we would like to have a list of all files used in each Capstone study. We can then annotate files based on which study they are part of and then pull that subset from the common [Capstone file view](https://urldefense.proofpoint.com/v2/url?u=https-3A__www.synapse.org_-23-21Synapse-3Asyn8466658_tables_&d=DwMFaQ&c=cjytLXgP8ixuoHflwc-poQ&r=S2riYg2F1ex2GXYU57QwKPjxerdh3GLV2ytZDPzhBm0&m=bxb9lu9FZTWh1C83sA0FojbpkOyC_NAvzTjsdbMNv4g&s=-FOgiVzHU2xkOM80wU0pvgoQHL6YAJm6gVYOXOZyvvk&e=)
* Should the reference tissue RNAseq and ChIPseq data be included? All the RNAseq has completed Capstone pipeline processing and are linked in the current Capstone file view. It's my understanding that there is a plan to process the ChIPseq using the Capstone pipeline. Or, should this dataset be provided separately at a future time from the other data?
* We need to give reviewers data access. This can be done in a couple of ways, and can discuss this directly with the journal
* For the review I assume we treat the raw and pipeline processing results the same (i.e., reviewers access the data through anonymous accounts that are either set up by us or the editor), and that we revisit the issue of pipeline processed results being public with Geetha before public release. Please confirm.
* Can you confirm that with adult you mean; BipSeq, BrainGVEX, CMC, CMC\_HBCC, EpiDiff, EpiGABA, EpiMAP, UCLA-ASD and Yale-ASD and LIBD\_szControl[as linked here](https://www.synapse.org/#!Synapse:syn8466658/tables/query/eyJzcWwiOiJTRUxFQ1QgKiBGUk9NIHN5bjg0NjY2NTgiLCAic2VsZWN0ZWRGYWNldHMiOlt7ImNvbmNyZXRlVHlwZSI6Im9yZy5zYWdlYmlvbmV0d29ya3MucmVwby5tb2RlbC50YWJsZS5GYWNldENvbHVtblZhbHVlc1JlcXVlc3QiLCAiY29sdW1uTmFtZSI6InN0dWR5IiwgImZhY2V0VmFsdW)