



INCLUDE
Data Coordinating Center

Analysis of INCLUDE data using CAVATICA

Matthew Galbraith
INCLUDE Workshop October 2022



Linda Crnic Institute for Down Syndrome
UNIVERSITY OF COLORADO ANSCHUTZ MEDICAL CAMPUS

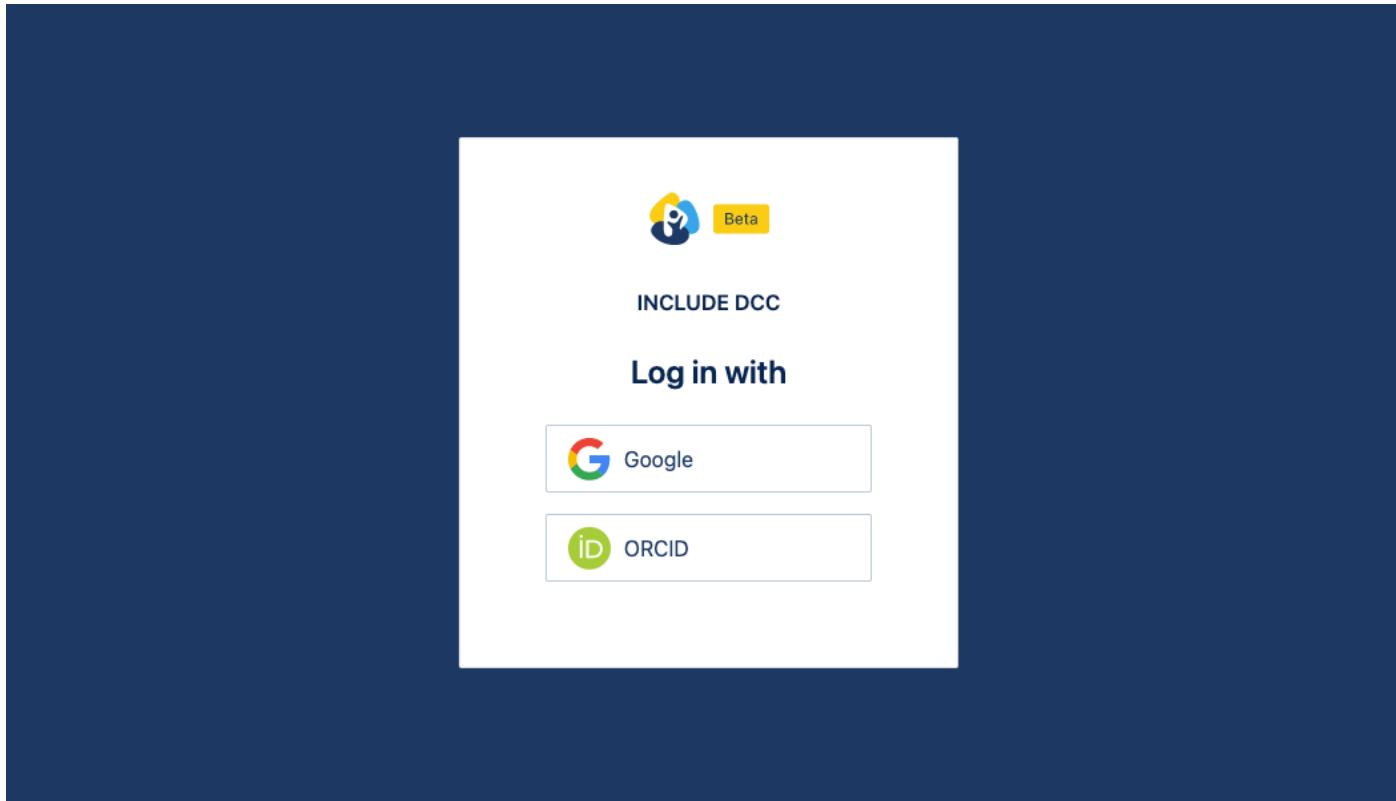




Tutorial slides and R code available at:
https://github.com/mattgalbraith/INCLUDE_examples/

INCLUDE Data Hub

<https://portal.includedcc.org/>



INCLUDE Data Hub: Dashboard



Beta

Dashboard

Studies

Data Exploration

Community

Website

Help

MATTHEW

Hello, MATTHEW

Data Exploration

 5
Studies 6,006
Participants 27.8K
Biospecimens 55TB
Data Files

Authorized Studies

You have access to the following INCLUDE controlled data through your NIH credentials.

[Disconnect](#)

No available studies

Saved Sets

[Participants \(0\)](#)[Biospecimen \(0\)](#)[Files \(1\)](#)

Cavatica Projects

You are connected to the Cavatica cloud environment. [Disconnect](#)

HTP_metab_DS3

1 member

KF_HTP_CRAM_HLA_extractions

2 members

include-test

8 members

Kids First & INCLUDE: Down Syndrome, Heart Defects, and Acute

[+ New project](#)

Saved Filters

HTP plasma metab

Last saved: about 1 hour ago

Feedback

INCLUDE Data Hub: Studies



Beta

Dashboard

Studies

Data Exploration

Community

Website

Help

MATTHEW

Studies

5 Results

| Study Code | Name | Program | dbGaP | Participants | Families | Biospecimens | Genomic | Transcriptomic | Proteomic | Immune Map | Metabolomic |
|----------------------------|---|------------|---------------------------|----------------------|----------|-----------------------|---------|----------------|-----------|------------|-------------|
| DS-PCGC | INCLUDE: (PCGC) Genomic Analysis of Congenital Heart Defects and Acute Lymphoblastic Leukemia in Children with Down Syndrome | INCLUDE/KF | phs002330 | 369 | 270 | 369 | ✓ | | | | |
| DS-COG-ALL | INCLUDE: (Lupo) Genomic Analysis of Congenital Heart Defects and Acute Lymphoblastic Leukemia in Children with Down Syndrome | INCLUDE/KF | phs002330 | 413 | 0 | 625 | ✓ | | | | |
| DS360-CHD | INCLUDE: (Sherman) Genomic Analysis of Congenital Heart Defects and Acute Lymphoblastic Leukemia in Children with Down Syndrome | INCLUDE/KF | phs002330 | 1172 | 809 | 1190 | ✓ | | | | |
| DSC | DS-Connect | INCLUDE | | 3366 | 25 | 0 | | | | | |
| HTP | Crnic Institute Human Trisome Project | INCLUDE/KF | phs002330 | 686 | 518 | 25584 | | ✓ | | | |

INCLUDE Data Hub: Studies

Crnic Institute Human Trisome Project®



LINDA CRNIC INSTITUTE
HUMAN TRISOME PROJECT™
GLOBAL DOWN SYNDROME FOUNDATION

[Visit Website](#)

Introduction

The Crnic Institute Human Trisome Project® (HTP) is an in-depth study of people with Down syndrome using the latest technologies in precision medicine. The goal of the HTP is to enable advanced therapeutic approaches to enhance the quality of life and extend the lifespan of those with trisomy 21 through the study of the co-occurring conditions of Down syndrome.

Principal Investigators

[*Espinosa, Joaquin M*](#)

About

The Human Trisome Project (HTP) leverages a multidisciplinary team of biomedical researchers, clinicians and data scientists located across multiple departments, divisions, institutes and centers at the University of Colorado who work together toward a single goal: to decipher why people with trisomy 21 have a different disease spectrum, being predisposed to some medical conditions while being protected from others.

INCLUDE Data Hub: Data Exploration - Summary

INCLUDE Data Hub Beta

Dashboard Studies Data Exploration

Community Website Help MATTHEW

Participant Biospecimen Data Files

Untitled Filter [Edit](#)

Use the search tools & facets on the left to build a query [My Filters](#)

6,006

Summary Participants (6,006) Biospecimens (27,768) Data Files (11,514)

Demographics

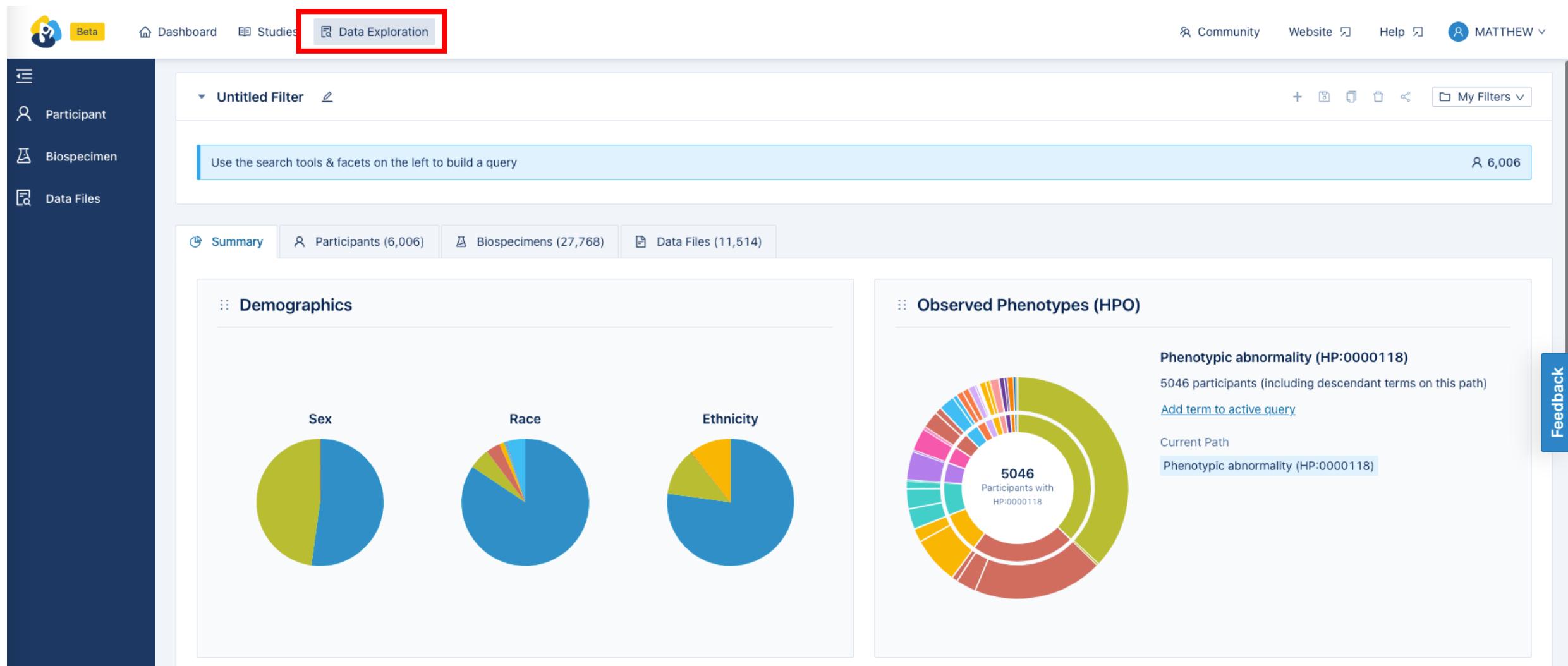
Sex Race Ethnicity

Observed Phenotypes (HPO)

Phenotypic abnormality (HP:0000118)
5046 participants (including descendant terms on this path)
[Add term to active query](#)

Current Path
Phenotypic abnormality (HP:0000118)

Feedback



INCLUDE Data Hub: Participants

The screenshot shows the INCLUDE Data Hub interface with the following details:

- Header:** Includes a yellow "Beta" badge, navigation links for Dashboard, Studies, and Data Exploration, and user profile for MATTHEW.
- Left Sidebar:** Contains icons for Participant, Biospecimen, and Data Files.
- Top Bar:** Features an "Untitled Filter" section, search bar, and "My Filters" dropdown.
- Search Bar:** Encourages users to "Use the search tools & facets on the left to build a query" and displays a result count of 6,006.
- Tab Navigation:** Summary (selected), Participants (6,006) (highlighted with a red box), Biospecimens (27,768), and Data Files (11,514).
- Table:** Displays results for participants 1 - 20 of 6006. The columns are:
 - Checkboxes for selection
 - Participant ID
 - Study Code
 - dbGaP
 - DS Status
 - Sex
 - Diagnosis (Mondo)
 - Phenotype (HPO)
 - Biospecimens
 - Files
- Table Data:** The table lists 8 rows of participant data, each with a "See more" link for detailed information.
- Buttons:** Save participant set, Download clinical data, and a Feedback button.

| | Participant ID | Study Code | dbGaP | DS Status | Sex | Diagnosis (Mondo) | Phenotype (HPO) | Biospecimens | Files |
|--------------------------|----------------|------------|-------|-----------|--------|--|---|--------------|-------|
| <input type="checkbox"/> | 10214 | DSC | - | T21 | Male | Attention deficit-hyperactivity disorder (MONDO: 0007743) See more | Sleep apnea (HP: 0010535) See more | 0 | 0 |
| <input type="checkbox"/> | 10224 | DSC | - | T21 | Female | Down syndrome (MONDO: 0008608) See more | Patent ductus arteriosus (HP: 0001643) See more | 0 | 0 |
| <input type="checkbox"/> | 10243 | DSC | - | T21 | Male | Down syndrome (MONDO: 0008608) | Delayed speech and language development (HP: 0000750) See more | 0 | 0 |
| <input type="checkbox"/> | 10260 | DSC | - | T21 | Male | Down syndrome (MONDO: 0008608) See more | Atrial septal defect (HP: 0001631) See more | 0 | 0 |
| <input type="checkbox"/> | 10261 | DSC | - | T21 | Male | Down syndrome (MONDO: 0008608) See more | Attention deficit hyperactivity disorder (HP: 0007018) See more | 0 | 0 |
| <input type="checkbox"/> | 10263 | DSC | - | T21 | Female | Down syndrome (MONDO: 0008608) | - | 0 | 0 |
| <input type="checkbox"/> | 10267 | DSC | - | T21 | Male | Myopia (disease) (MONDO: 0001384) See more | Myopia (HP: 0000545) See more | 0 | 0 |

INCLUDE Data Hub: Biospecimens

Screenshot of the INCLUDE Data Hub interface showing the Biospecimens section.

The top navigation bar includes: Dashboard, Studies, Data Exploration (selected), Community, Website, Help, and a user profile for MATTHEW.

The left sidebar has filters for Participant, Biospecimen (selected), and Data Files.

The main area shows an "Untitled Filter" with search tools and facets. A message says: "Use the search tools & facets on the left to build a query". The result count is 6,006.

The navigation tabs at the top of the results table are: Summary, Participants (6,006), Biospecimens (27,768) (selected), and Data Files (11,514).

The results table displays 20 rows of biospecimen data. The columns are:

| | Sample ID | Study | Sample Type | Parent Sample ID | Parent Sample Type | Participant ID | Collection ID | Collection Sample Type | Age (days) | Container ID | Volume | Volume Unit | Sample Availability | Laboratory Procedure | Biospecime Storage |
|--------------------------|-----------------|------------|-------------|------------------------------------|------------------------|----------------|--|------------------------|------------|--------------|--------|-------------|---------------------|----------------------|--------------------|
| <input type="checkbox"/> | BS_00BPKP5F_DNA | DS-COG-ALL | DNA | BS_00BPKP5F_Not Reported | Not Reported | PT_8ME5T344 | BS_00BPKP5F_Not Reported | Not Reported | - | - | - | - | No | | |
| <input type="checkbox"/> | BS_011DYZ2J_DNA | DS360-CHD | DNA | BS_011DYZ2J_Peripheral Whole Blood | Peripheral Whole Blood | PT_KJJN4NR4 | BS_011DYZ2J_Peripheral Whole Blood | Peripheral Whole Blood | - | - | - | - | No | | |
| <input type="checkbox"/> | BS_01585F30_DNA | DS-COG-ALL | DNA | BS_01585F30_Not Reported | Not Reported | PT_29604G66 | BS_01585F30_Not Reported | Not Reported | - | - | - | - | No | | |
| <input type="checkbox"/> | BS_01AY2V9A_DNA | DS360-CHD | DNA | BS_01AY2V9A_Peripheral Whole Blood | Peripheral Whole Blood | PT_QYAJZNA4 | BS_01AY2V9A_Peripheral Whole Blood | Peripheral Whole Blood | - | - | - | - | No | | |

A red box highlights the "Biospecimens (27,768)" tab. A blue "Feedback" button is located on the far right edge of the page.

INCLUDE Data Hub: Data Files

The screenshot shows the INCLUDE Data Hub interface with the 'Data Exploration' tab selected. A red box highlights the 'Data Files (11,514)' tab in the navigation bar. The main area displays a table of 11,514 data files, with the first 20 results shown. The columns include File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. Each row has a checkbox, a lock icon, and a 'C' icon.

| File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|-------------|------------|---------------|--------------------------------|-------------------------|--------|-----------|--------------|--------------|
| GF_001CSF26 | DS-PCGC | Genomic | Aligned Reads | Whole Genome Sequencing | cram | 14.48 GB | 1 | 1 |
| GF_0095JPTH | DS-PCGC | Genomic | Aligned Reads | Whole Genome Sequencing | cram | 14.09 GB | 1 | 1 |
| GF_009Z6YHX | DS-COG-ALL | Genomic | Somatic Structural Variations | Whole Genome Sequencing | vcf | 33.72 KB | 1 | 2 |
| GF_00J3C1QG | DS-COG-ALL | Genomic | Simple Nucleotide Variations | Whole Genome Sequencing | maf | 1.22 MB | 1 | 2 |
| GF_00KNEOEN | DS-COG-ALL | Genomic | Somatic Copy Number Variations | Whole Genome Sequencing | png | 97.5 KB | 1 | 2 |
| GF_00M6HPKX | DS360-CHD | Genomic | - | Whole Genome Sequencing | html | 194.26 KB | 1 | 1 |
| GF_00MFRFTC | DS-COG-ALL | Genomic | Simple Nucleotide Variations | Whole Genome Sequencing | maf | 18 MB | 1 | 2 |
| GF_00YQA7WH | DS-COG-ALL | Genomic | Simple Nucleotide Variations | Whole Genome Sequencing | maf | 1.95 MB | 1 | 2 |
| GF_0146HEF6 | DS360-CHD | Genomic | gVCF | Whole Genome Sequencing | gvcf | 5.46 GB | 1 | 1 |
| GF_0147NKA6 | DS360-CHD | Genomic | gVCF | Whole Genome Sequencing | gvcf | 6.63 GB | 1 | 1 |

INCLUDE Data Hub: Filtering

The screenshot illustrates the INCLUDE Data Hub's filtering mechanism. On the left, a sidebar titled "Facets & Filters" contains a red box highlighting the "Participant" facet, which includes a search bar for "Participant ID" and a "Saved Participant Sets" dropdown. Below these are buttons for "Upload a Participant list" and "Expand all". To the right of the sidebar is the main search interface, titled "Untitled Filter". It features a search bar with placeholder text "Use the search tools & facets on the left to build a query" and a result count of "6,006". Below the search bar are tabs for "Summary", "Participants (6,006)", "Biospecimens (27,768)", and "Data Files (11,514)". The "Participants" tab is selected. The main area displays a table of results with columns: Select, Lock, Check, File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. The first seven rows of the table are shown below:

| Select | Lock | Check | File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--------------------------|----------------------|-------------------------------------|-------------|------------|---------------|--------------------------------|-------------------------|--------|-----------|-------------------|-------------------|
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_001CSF26 | DS-PCGC | Genomic | Aligned Reads | Whole Genome Sequencing | cram | 14.48 GB | 1 | 1 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_0095JPTH | DS-PCGC | Genomic | Aligned Reads | Whole Genome Sequencing | cram | 14.09 GB | 1 | 1 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_009Z6YHX | DS-COG-ALL | Genomic | Somatic Structural Variations | Whole Genome Sequencing | vcf | 33.72 KB | 1 | 2 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_00J3C1QG | DS-COG-ALL | Genomic | Simple Nucleotide Variations | Whole Genome Sequencing | maf | 1.22 MB | 1 | 2 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_00KNE0EN | DS-COG-ALL | Genomic | Somatic Copy Number Variations | Whole Genome Sequencing | png | 97.5 KB | 1 | 2 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_00M6HPKX | DS360-CHD | Genomic | - | Whole Genome Sequencing | html | 194.26 KB | 1 | 1 |

INCLUDE Data Hub: Filtering by Study Code

The screenshot shows the INCLUDE Data Hub interface. On the left, a sidebar navigation includes 'Participant' (selected), 'Biospecimen', and 'Data Files'. The main area features a search bar for 'Participant ID' and a 'Saved Participant Sets' dropdown. A prominent 'Untitled Filter' section allows building a query using search tools and facets. Below this, tabs for 'Summary', 'Participants (6,006)', 'Biospecimens (27,768)', and 'Data Files (11,514)' are visible. The 'Data Files' tab is active, displaying results 1-20 of 11514. A table lists data files with columns for File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. The 'HTP' study code filter is highlighted with a red box, and the 'Apply' button is also highlighted. A 'Feedback' button is located on the far right.

Participant

Biospecimen

Data Files

Dashboard Studies Data Exploration

Community Website Help MATTHEW

Untitled Filter

Search by Participant ID
e.g. PT_WFB3TQP4

Saved Participant Sets
Select a saved set

Upload a Participant list

Study Code

All | None

- DSC 3,366
- DS360-CHD 1,172
- HTP 686
- DS-COG-ALL 413
- DS-PCGC 369

Clear Apply ...

Down Syndrome Status

Diagnosis (MONDO)

Phenotype (HPO)

Family Unit

Results 1 - 20 of 11514

| | | | File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--------------------------|----------------------|-------------------------------------|-------------|------------|---------------|--------------------------------|-------------------------|--------|-----------|--------------|--------------|
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_001CSF26 | DS-PCGC | Genomic | Aligned Reads | Whole Genome Sequencing | cram | 14.48 GB | 1 | 1 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_0095JPTH | DS-PCGC | Genomic | Aligned Reads | Whole Genome Sequencing | cram | 14.09 GB | 1 | 1 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_009Z6YHX | DS-COG-ALL | Genomic | Somatic Structural Variations | Whole Genome Sequencing | vcf | 33.72 KB | 1 | 2 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_00J3C1QG | DS-COG-ALL | Genomic | Simple Nucleotide Variations | Whole Genome Sequencing | maf | 1.22 MB | 1 | 2 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_00KNE0EN | DS-COG-ALL | Genomic | Somatic Copy Number Variations | Whole Genome Sequencing | png | 97.5 KB | 1 | 2 |
| <input type="checkbox"/> | <input type="lock"/> | <input checked="" type="checkbox"/> | GF_00M6HPKX | DS360-CHD | Genomic | - | Whole Genome Sequencing | html | 194.26 KB | 1 | 1 |

INCLUDE Data Hub: Filtering by Study Code

The screenshot shows the INCLUDE Data Hub interface with the following components:

- Header:** Includes a yellow "Beta" badge, navigation links for Dashboard, Studies, and Data Exploration, and a user profile for MATTHEW.
- Left Sidebar:** Shows sections for Participant, Biospecimen, and Data Files. The Participant section is active, displaying a search bar for "Search by Participant ID" (e.g. PT_WFB3TQP4), a "Saved Participant Sets" dropdown, and a button to "Upload a Participant list".
- Filtering Panel:** Titled "Untitled Filter", it contains a query bar with "Study Code = HTP" (highlighted with a red box) and a delete icon. Buttons for "+ New query" and "Labels" are also present.
- Data Summary:** A summary bar showing 686 Participants, 25,584 Biospecimens, and 4,083 Data Files.
- Table View:** A grid showing results 1 - 20 of 4083. The columns include File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Bios. The first row is highlighted with a blue border. The table shows five rows of data, each corresponding to a file named after the HTP study code.
- Feedback:** A vertical "Feedback" button is located on the far right.

INCLUDE Data Hub: Filtering Data Files

The screenshot shows the INCLUDE Data Hub interface. The top navigation bar includes links for Dashboard, Studies, Data Exploration (selected), Community, Website, Help, and a user profile for MATTHEW. The left sidebar has sections for Participant, Biospecimen, and Data Files, with Data Files highlighted by a red box. A search bar for 'File ID' and a 'Saved File Sets' dropdown are also present. The main area displays an 'Untitled Filter' with the query 'Study Code = HTP'. Below this, a table lists 4,083 data files, with the first few rows shown:

| File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Bios |
|--|-------|----------------|---------------------------------|-----------------------|--------|---------|--------------|------|
| HTP.0012022d-c855-4087-9dde-c522f0632024.arriba.fusions.pdf | HTP | Transcriptomic | Gene Fusions | RNA-Seq | pdf | 1.09 MB | 1 | 1 |
| HTP.0012022d-c855-4087-9dde-c522f0632024.arriba.fusions.tsv | HTP | Transcriptomic | Gene Fusions | RNA-Seq | tsv | 6.36 KB | 1 | 1 |
| HTP.0012022d-c855-4087-9dde-c522f0632024.kallisto.abundance.tsv.gz | HTP | Transcriptomic | Gene Expression Quantifications | RNA-Seq | tsv | 2.74 MB | 1 | 1 |
| HTP.0012022d-c855-4087-9dde-c522f0632024.rsem.genes.results.gz | HTP | Transcriptomic | Gene Expression Quantifications | RNA-Seq | tsv | 2.44 MB | 1 | 1 |
| HTP.0012022d-c855-4087-9dde- | HTP | Transcriptomic | Gene | RNA-Seq | tsv | 4.15 | 1 | 1 |

INCLUDE Data Hub: Filtering Data Files

The screenshot shows the INCLUDE Data Hub interface with the following components:

- Header:** Includes a yellow "Beta" badge, navigation links for Dashboard, Studies, and Data Exploration, and a user profile for MATTHEW.
- Left Sidebar:** Shows sections for Participant, Biospecimen, and Data Files. The Data Files section is active, featuring a search bar for "File ID" (e.g. GF_001CSF26), a "Saved File Sets" dropdown, and a "Upload a File list" button. It also includes filters for "Access", "Data Category", "Data Type", "Experimental Strategy", and "File Format". The "Data Category" filter is expanded, showing options like All, Metabolomics (418, checked), Transcriptomic (2,310), Genomics (878), and Proteomics (477). The "Apply" button is highlighted with a red box.
- Filter Bar:** Displays an "Untitled Filter" with the condition "Study Code = HTP and Data Category = Metabolomics". The "Data Category" part is highlighted with a red box.
- Summary and Count:** Shows 418 results for the current filter.
- Table:** A grid displaying 20 results out of 418, with columns for File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. The first row is shown below:

| | File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--------------------------|--|-------|---------------|-----------|-----------------------|--------|---------|--------------|--------------|
| <input type="checkbox"/> | HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 12 |
- Feedback:** A vertical blue sidebar labeled "Feedback" is visible on the right edge.

INCLUDE Data Hub: Saving a filter

The screenshot shows the INCLUDE Data Hub interface. On the left, there's a sidebar with icons for Participant, Biospecimen, and Data Files. The main area has tabs for Dashboard, Studies, and Data Exploration. A yellow 'Beta' badge is visible. The top right includes links for Community, Website, Help, and a user profile for MATTHEW.

In the center, a search bar says "Search by File ID" with placeholder "e.g. GF_001CSF26". Below it, a "Saved File Sets" section says "Select a saved set". A large red box highlights the filter name "HTP plasma metab" in the header, which also features a pencil icon. To the right of the filter name is the text "Saving filters". Another red box highlights the "My Filters" dropdown menu in the top right corner of the header.

The main content area shows a search query: "Study Code = HTP X and Data Category = Metabolomics X". It displays 418 results. Buttons for "+ New query" and "Labels" are present. Below the search bar are tabs for Summary, Participants (418), Biospecimens (3,299), and Data Files (418). A "Feedback" button is on the far right.

The results table has columns for File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. The first few rows show entries like "HTP.HTP001B2_LCMS_Metabolomics.tsv.gz" and "HTP.HTP005A3_LCMS_Metabolomics.tsv.gz", both categorized as Metabolomics with LCMS_Metabolomics experimental strategy and tsv format.

| | | | File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--------------------------|--|--|---------------------------------------|-------|---------------|-----------|-----------------------|--------|---------|--------------|--------------|
| <input type="checkbox"/> | | | HTP.HTP001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 12 |
| <input type="checkbox"/> | | | HTP.HTP005A3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 9 |
| <input type="checkbox"/> | | | HTP.HTP012A2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.41 KB | 1 | 12 |
| <input type="checkbox"/> | | | HTP.HTP015A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.4 KB | 1 | 12 |
| <input type="checkbox"/> | | | HTP.HTP017A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.43 KB | 1 | 12 |
| <input type="checkbox"/> | | | HTP.HTP018B3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.44 | 1 | 12 |

INCLUDE Data Hub: Saving a file set

The screenshot shows the INCLUDE Data Hub interface. On the left, there's a sidebar with icons for Participant, Biospecimen, and Data Files. The main area has tabs for Dashboard, Studies, and Data Exploration, with the latter being active. A search bar at the top allows searching by File ID. Below it, a section for "Saved File Sets" shows a placeholder "Select a saved set". A button to "Upload a File list" is also present. The main content area displays a query titled "HTP plasma metab" with filters: "Study Code = HTP" and "Data Category = Metabolomics". The results show 418 items. A red box highlights a context menu that appears when right-clicking on one of the rows. The menu options are: "Save file set", "Analyze in Cavatica", "418 file selected", "+ Save as new set", "& Add to existing set", and "& Remove from existing set". The menu is overlaid on a table with columns: File ID, Study, Data Category, Data Type, and Experimental Strategy. The table lists several entries, all of which have checkboxes checked in the first column.

Participant

Biospecimen

Data Files

Dashboard Studies Data Exploration

Community Website Help MATTHEW

HTP plasma metab

Study Code = HTP and Data Category = Metabolomics

418

+ New query Labels

Access

Data Category

All | None

Metabolomics 418

Transcriptomic 2,310

Genomics 878

Proteomics 477

Clear Apply ...

418 items selected Clear

Save file set Analyze in Cavatica

418 file selected

+ Save as new set

& Add to existing set

& Remove from existing set

Feedback

| | File ID | Study | Data Category | Data Type | Experimental Strategy |
|-------------------------------------|--|-------|---------------|-----------|-----------------------|
| <input checked="" type="checkbox"/> | HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics |
| <input checked="" type="checkbox"/> | HTP.HTP0005A3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics |
| <input checked="" type="checkbox"/> | HTP.HTP0012A2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics |
| <input checked="" type="checkbox"/> | HTP.HTP0015A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics |
| <input checked="" type="checkbox"/> | HTP.HTP0017A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics |
| <input checked="" type="checkbox"/> | HTP.HTP0018B3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics |

INCLUDE Data Hub: Selecting files

The screenshot shows the INCLUDE Data Hub interface. The top navigation bar includes links for Dashboard, Studies, Data Exploration, Community, Website, Help, and a user profile for MATTHEW. The left sidebar has sections for Participant, Biospecimen, and Data Files, with Data Files currently selected. A search bar at the top left allows searching by File ID (e.g., GF_001CSF26). Below it is a section for Saved File Sets, which is currently empty. A large button labeled "Upload a File list" is present. The main area displays a search result for "HTP plasma metab". The search filters are "Study Code = HTP" and "Data Category = Metabolomics", resulting in 418 items. A red box highlights the "Select all results" link. The results table lists 20 items selected, with columns for File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. The first item in the table is "HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz". A "Feedback" button is located on the far right.

HTP plasma metab

Study Code = HTP and Data Category = Metabolomics

418

20 items selected [Select all results](#) Make sure to select all

| File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--|-------|---------------|-----------|-----------------------|--------|---------|--------------|--------------|
| HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 12 |
| HTP.HTP0005A3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 9 |
| HTP.HTP0012A2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.41 KB | 1 | 12 |
| HTP.HTP0015A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.4 KB | 1 | 12 |
| HTP.HTP0017A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.43 KB | 1 | 12 |
| HTP.HTP0018B3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.44 KB | 1 | 12 |

INCLUDE Data Hub: Selecting files

The screenshot shows the INCLUDE Data Hub interface with the following details:

- Header:** Includes a yellow "Beta" badge, navigation links for Dashboard, Studies, and Data Exploration, and a user profile for MATTHEW.
- Left Sidebar:** Contains icons for Participant, Biospecimen, and Data Files, along with a search bar for "File ID" and a section to "Upload a File list".
- Central Area:** A search bar for "HTP plasma metab" with filters for "Study Code = HTP" and "Data Category = Metabolomics". It shows 418 results.
- Table Header:** Shows categories: Summary, Participants (418), Biospecimens (3,299), and Data Files (418).
- Table Content:** A grid of 418 selected items. The first item is highlighted with a red box around the "418 items selected" message at the top of the table.
- Table Headers:** File ID, Study, Data Category, Data Type, Experimental Strategy, Format, Size, Participants, Biospecimens.
- Feedback:** A blue vertical bar labeled "Feedback" is visible on the right side.

| File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--|-------|---------------|-----------|-----------------------|--------|---------|--------------|--------------|
| HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 12 |
| HTP.HTP0005A3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 9 |
| HTP.HTP0012A2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.41 KB | 1 | 12 |
| HTP.HTP0015A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.4 KB | 1 | 12 |
| HTP.HTP0017A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.43 KB | 1 | 12 |
| HTP.HTP0018B3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.44 KB | 1 | 12 |

INCLUDE Data Hub: Sending files to CAVATICA for analysis

The screenshot shows the INCLUDE Data Hub interface. On the left, there's a sidebar with icons for Participant, Biospecimen, and Data Files. The main area has a search bar for 'File ID' and a 'Saved File Sets' dropdown. A 'Data Exploration' tab is active. At the top right, there are links for Community, Website, Help, and a user profile for MATTHEW.

A search query 'HTP plasma metab' is displayed, with filters 'Study Code = HTP' and 'Data Category = Metabolomics'. The results table shows 418 items selected. A red box highlights the 'Analyze in Cavatica' button in the header of the table.

Sending files to CAVATICA

| File ID | Study | Data Category | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|--|-------|---------------|-----------|-----------------------|--------|---------|--------------|--------------|
| HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 12 |
| HTP.HTP0005A3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.42 KB | 1 | 9 |
| HTP.HTP0012A2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.41 KB | 1 | 12 |
| HTP.HTP0015A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.4 KB | 1 | 12 |
| HTP.HTP0017A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.43 KB | 1 | 12 |
| HTP.HTP0018B3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | - | LCMS_Metabolomics | tsv | 6.44 KB | 1 | 12 |

INCLUDE Data Hub: Sending files to CAVATICA for analysis



Beta

Dashboard

Studies

Data Exploration

Community

Website

Help

MATTHEW

Hello, MATTHEW

Data Exploration

 5
Studies 6,006
Participants 27.8K
Biospecimens 55TB
Data Files

Authorized Studies

You have access to the following INCLUDE controlled data through your NIH credentials.

[Disconnect](#)

No available studies

Cavatica Projects

You are connected to the Cavatica cloud environment. [Disconnect](#)

HTP_metab_DS3

1 member

KF_HTP_CRAM_HLA_extractions

2 members

include-test

8 members

Kids First & INCLUDE: Down Syndrome, Heart Defects, and Acute

[+ New project](#)

Saved Filters

HTP plasma metab

Last saved: about 1 hour ago

Saved Sets

[Participants \(0\)](#)[Biospecimen \(0\)](#)[Files \(1\)](#)

First time: Connect account to CAVATICA

Feedback

CAVATICA: Sending files to CAVATICA for analysis

Ensure the INCLUDE DRS Server is enabled before trying to send files!

Account Settings --> Datasets

INCLUDE DRS Server

Enable your CAVATICA account to import files from the INCLUDE DRS Server.

| DRS Endpoint | Account |
|--|----------------------|
| drs://cavatica-ga4gh-api.sbggenomics.com | No account connected |

Connect



INCLUDE DRS Server

Enable your CAVATICA account to import files from the INCLUDE DRS Server.

| DRS Endpoint | Account | Expires |
|--|-------------------|---------------------|
| drs://cavatica-ga4gh-api.sbggenomics.com | matthew.galbraith | Aug. 14, 2022 09:50 |

Reconnect ...

You may also need to apply for access to certain data files (eg dbGAP)
see <https://help.includedcc.org/docs/applying-for-access>

INCLUDE Data Hub: Sending files to CAVATICA for analysis

The screenshot shows the INCLUDE Data Hub web application. On the left, there's a sidebar with navigation links for Dashboard, Studies, Data Exploration, Participant, Biospecimen, and Data Files. The Data Files section is currently active, showing a search bar for 'File ID' (e.g., GF_001CSF26) and a 'Saved File Sets' dropdown. Below these are buttons for 'Upload a File list' and 'Expand all'. Under 'Data Category', there are filters for 'All' (selected), 'None', 'Metabolomics' (418 items), 'Transcriptomic' (2,310 items), 'Genomics' (878 items), and 'Proteomics' (477 items). There are also 'Clear' and 'Apply ...' buttons. To the right, a main panel displays a table of selected items from the 'Crnic Institute Human Trisome Project'. The table has columns for 'Data Type' (HTP), 'Experimental Strategy' (Metabolomics), 'Format' (tsv), 'Size' (e.g., 6.42 KB), 'Participants' (1), and 'Biospecimens' (12). A modal dialog box titled 'Analyze in Cavatica' is open over the table, prompting the user to 'Copy files to...' a project. It shows a summary of '418 items selected' and a message: 'You are authorized to copy 418 files (out of 418 selected) to your Cavatica workspace.' At the bottom of the dialog are 'Cancel' and 'Copy files' buttons.

| Item ID | File Name | Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|---------|--|-----------|-----------------------|--------|---------|--------------|--------------|
| 1 | HTP.HTP0001B2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | tsv | 6.42 KB | 1 | 12 |
| 2 | HTP.HTP0005A3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | tsv | 6.42 KB | 1 | 9 |
| 3 | HTP.HTP0012A2_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | tsv | 6.41 KB | 1 | 12 |
| 4 | HTP.HTP0015A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | tsv | 6.4 KB | 1 | 12 |
| 5 | HTP.HTP0017A4_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | tsv | 6.43 KB | 1 | 12 |
| 6 | HTP.HTP0018B3_LCMS_Metabolomics.tsv.gz | HTP | Metabolomics | tsv | 6.44 KB | 1 | 12 |

INCLUDE Data Hub: Sending files to CAVATICA for analysis

The screenshot shows the INCLUDE Data Hub dashboard with a modal window titled "Analyze in Cavatica". The modal displays a list of projects where the selected files can be copied:

- HTP_metab_DS3
- KF_HTP_CRAM_HLA_extractions
- include-test
- Kids First & INCLUDE: Down Syndrome, Heart Defects, and Acute Lymphoblastic Leukemia

Below the list is a button labeled "+ New project". The main dashboard background shows a table of 418 selected items, with columns for Data Type, Experimental Strategy, Format, Size, Participants, and Biospecimens. A sidebar on the left provides filtering options for Participant, Biospecimen, and Data Files. The top navigation bar includes links for Dashboard, Studies, Data Exploration, Community, Website, Help, and a user profile for MATTHEW.

| Data Type | Experimental Strategy | Format | Size | Participants | Biospecimens |
|-----------|-----------------------|--------|---------|--------------|--------------|
| HTP | Metabolomics | tsv | 6.42 KB | 1 | 12 |
| HTP | Metabolomics | tsv | 6.42 KB | 1 | 9 |
| HTP | Metabolomics | tsv | 6.41 KB | 1 | 12 |
| HTP | Metabolomics | tsv | 6.4 KB | 1 | 12 |
| HTP | Metabolomics | tsv | 6.43 KB | 1 | 12 |
| HTP | Metabolomics | tsv | 6.44 | 1 | 12 |

CAVATICA: Project dashboard

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects matthew.galbraith

Dashboard **Files** Apps Tasks Data Studio HTP_metab_DS3 ⓘ Interactive Browsers Settings Notes

Description Tags

Welcome to your new project!

Projects are the core building blocks of the Seven Bridges Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

Within your project, you can:

- Start exploring the public pipelines straight away
- Install your tools and create workflows
- Upload your own private data
- Collaborate securely with other researchers

After reviewing the information above, you can continue to use this space for adding notes about your project such as its aims, experimental context, and any other ideas that you'd like to share with your project members as everyone will see the same content. You can also use markdown here to add formatting to your notes.

To start adding your description, click Add Description below.

Remember that details of each pipeline execution you run on the Seven Bridges Platform are logged on the dedicated task page.

Good luck with your research! If you get stuck, take a look at the [Knowledge Center](#).

The Seven Bridges Team

Add description

Members Email notifications

matthew.galbraith OWNER Copy, Write, Execute, Admin

Don't work alone. The best research happens in teams. Invite new members Share your tools, data, and ideas with collaborators

Analysis Search

Tasks Data Studio

Your executions will appear here. Before you start, learn more about them.

CAVATICA: Files

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects matthew.galbraith

Dashboard **Files** Apps Tasks Data Studio HTP_metab_DS3 ⓘ Interactive Browsers Settings Notes

Files

Search Extension: All Sample ID: All Task ID: All Tags: All + Clear filters

New folder Add files ...

| Name | Task ID | Created on | Extension | Size | Sample ID |
|---------------------------------------|---------|---------------------|-----------|----------|-----------|
| test_output.txt | - | July 15, 2022 11:06 | TXT | 12.9 KIB | - |
| DRS HTP0706A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.2 KIB | - |
| DRS HTP0677B_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0676A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0672A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0669B_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0668A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0667A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0666A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0665A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |
| DRS HTP0664A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.2 KIB | - |
| DRS HTP0663A_LCMS_Metabolomics.tsv.gz | - | July 15, 2022 09:52 | TSV.GZ | 6.3 KIB | - |

Refresh Showing 1-100 of 410

CAVATICA: Data Studio (interactive analysis)

CAVATICA Projects Data Public Apps Public Projects Developer Controlled projects matthew.galbraith

Dashboard Files Apps Tasks **Data Studio** HTP_metab_DS3 ⓘ Interactive Browsers Settings Notes

Description Tags

Welcome to your new project!

Projects are the core building blocks of the Seven Bridges Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

Within your project, you can:

- Start exploring the public pipelines straight away
- Install your tools and create workflows
- Upload your own private data
- Collaborate securely with other researchers

After reviewing the information above, you can continue to use this space for adding notes about your project such as its aims, experimental context, and any other ideas that you'd like to share with your project members as everyone will see the same content. You can also use markdown here to add formatting to your notes.

To start adding your description, click Add Description below.

Remember that details of each pipeline execution you run on the Seven Bridges Platform are logged on the dedicated task page.

Good luck with your research! If you get stuck, take a look at the [Knowledge Center](#).

The Seven Bridges Team

Add description

Members Email notifications

matthew.galbraith OWNER Copy, Write, Execute, Admin

Don't work alone. The best research happens in teams. Invite new members Share your tools, data, and ideas with collaborators

Analysis Search

Tasks Data Studio

Your executions will appear here. Before you start, learn more about them.

CAVATICA: Data Studio – new analysis

The screenshot shows the CAVATICA Data Studio interface. At the top, there is a navigation bar with links for Projects, Data, Public Apps, Public Projects, Developer, and Controlled projects. On the right side of the header, there is a notification icon and a user profile for matthew.galbraith.

The main area displays a table of existing analyses. The columns are labeled: Analysis Name, Status, Created by, Environment, Created on, and Action. One row is visible, showing "HTP_metab_karyotype" as the Analysis Name, "SAVED" as the Status, "matthew.galbraith" as the Created by, "RStudio (SB Bioinformatics - R 4.1 - BioC 3..." as the Environment, "Jul. 15, 2022 09:24" as the Created on, and a "Start" button with a three-dot menu next to it.

A search bar is located at the top left, and a "Create new analysis" button is located at the top right, which is highlighted with a red rectangular box.

| Analysis Name | Status | Created by | Environment | Created on | Action |
|---------------------|--------|-------------------|--|---------------------|----------------------------|
| HTP_metab_karyotype | SAVED | matthew.galbraith | RStudio (SB Bioinformatics - R 4.1 - BioC 3... | Jul. 15, 2022 09:24 | <button>Start</button> ... |

CAVATICA: Data Studio – new analysis

The screenshot shows the CAVATICA Data Studio interface. At the top, there is a navigation bar with links for Projects, Data, Public Apps, Public Projects, Developer, and Controlled projects. On the right side of the header, there is a user profile for matthew.galbraith and a notification icon.

The main area shows a table of existing analyses. One analysis, "HTP_metab_karyotype", is listed with a status of "SAVED". To the right of the table, there are buttons for Interactive Browsers, Settings, and Notes.

A search bar is located at the top left, and a "Create new analysis" button is at the top right.

A modal window titled "Create new analysis" is open in the center. It contains fields for "Analysis name" (set to "New analysis"), "Environment" (set to "JupyterLab"), "Environment setup" (set to "SB Bioinformatics - R 4.1 - BioC 3.14 DEFAULT"), "Instance type" (set to "c5.2xlarge (8vCPUs, 16GB RAM)"), "Attached Storage (GB)" (set to "1024"), "Suspend Time" (set to "On" with a value of "30 Minutes"), and a "Start" button.

The "Start" button is highlighted with a red box.

| Analysis Name | Status | Created on | Action |
|---------------------|--------|---------------------------------|-------------|
| HTP_metab_karyotype | SAVED | - BioC 3... Jul. 15, 2022 09:24 | ▶ Start ... |

CAVATICA: Data Studio – existing analysis

The screenshot shows the CAVATICA Data Studio interface. At the top, there is a navigation bar with links for Projects, Data, Public Apps, Public Projects, Developer, and Controlled projects. On the far right, there is a user profile for matthew.galbraith. Below the navigation bar, there is a secondary navigation bar with links for Dashboard, Files, Apps, Tasks, and Data Studio. The Data Studio link is underlined, indicating it is the active page. In the center, there is a title "HTP_metab_DS3" with a help icon. To the left of the title is a search bar with the placeholder "Search". To the right of the title are buttons for "Create new analysis" and a dropdown menu. Below the title, there is a table with the following columns: Analysis Name, Status, Created by, Environment, Created on, and Action. The table contains one row for an analysis named "HTP_metab_karyotype". The status is "SAVED", created by "matthew.galbraith", environment is "RStudio (SB Bioinformatics - R 4.1 - BioC 3...)", created on "Jul. 15, 2022 09:24", and the action button is highlighted with a red box.

| Analysis Name | Status | Created by | Environment | Created on | Action |
|---------------------|--------|-------------------|---|---------------------|------------------------------|
| HTP_metab_karyotype | SAVED | matthew.galbraith | RStudio (SB Bioinformatics - R 4.1 - BioC 3...) | Jul. 15, 2022 09:24 | <button>▶ Start ...</button> |

CAVATICA: Data Studio – Rstudio session

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the `linear_regression_example.R` script. A warning message at the top states: "Packages ggforce, rstatix, and tictoc required but are not installed. [Install](#) [Don't Show Again](#)". The script includes code for installing packages like tidyverse, broom, ggforce, and here, and setting standard colors.
- Terminal:** Shows the R version 4.1.2 startup message, including the license information and a quit command (`q()`).
- File Browser:** Shows the project directory structure with files `linear_regression_example.R` and `test_output.txt`.
- Environment:** Shows the global environment is empty.

Find code at: https://github.com/mattgalbraith/INCLUDE_examples/blob/main/linear_regression_example.R

CAVATICA: Data Studio – input files location

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the project name "HTTP_metab_karyotype", file menu, and "rstudio" logo.
- File Explorer:** Shows a folder structure with "linear_regression_example.R" and "test_output.txt".
- Code Editor:** Displays an R script titled "linear_regression_example.R". A red box highlights the line of code that reads files from the project directory:

```
msd_data_files <- list.files(path = "/sbgenomics/project-files/", pattern = "LCMS_Metabolomics.tsv.gz", full.names = TRUE)
```
- Console:** Displays the R startup message and the command history for the current session.
- Environment:** Shows the global environment is empty.

Files delivered from INCLUDE data hub found here: /sbgenomics/project-files/ (read-only from Data Studio)

CAVATICA: files from INCLUDE data hub need to be combined

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the project name "HTTP_metab_karyotype" and standard menu options: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Code Editor:** Displays the script file "linear_regression_example.R". A red box highlights the code for concatenating individual sample-level files (lines 22-28).

```
15 standard_colors <- c("Control" = "gray60", "T21" = "#009b4e")
16
17
18
19 # get list of files
20 msd_data_files <- list.files(path = "/sbgenomics/project-files/", pattern = "LCMS_Metabolomics.tsv.gz", full.names = TRUE)
21
22 # Concatenate individual sample-level files
23 tic()
24 msd_data <- msd_data_files %>%
25   map_dfr(~read_tsv(., id = "file")) %>%
26   mutate(filename = basename(file)) %>%
27   select(LabID, everything())
28 toc() # ~18.557 sec elapsed
29
30
31 # SHOULD BE IMPORTING CLINICAL DATA (Karyotype, Age, Sex, etc) AND JOINING HERE
32 # but we will cheat and get Karyotype (T21 status) information here from the "LabID"
33
34
35 # Prepare data for linear regression ----
36 # assumes "msd_data" is in long format already joined with clinical data
37 # "LabID" is unique sample identifier; "Analyte" denotes each feature of interest; "Value" is the actual measurement; here "Karyotype"
38 regressions_dat <- msd_data %>%
```
- Console:** Displays the R startup message and help text.

```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```
- File Explorer:** Shows the current directory structure:

| Name | Size | Modified |
|-----------------------------|---------|------------------------|
| linear_regression_example.R | 10.2 KB | Jul 15, 2022, 11:14 AM |
| test_output.txt | 12.9 KB | Jul 15, 2022, 11:14 AM |
- Environment:** Shows the message "Environment is empty".

CAVATICA: example linear regression analysis with HTP metabolomics

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the project name "HTP_metab_karyotype" and standard menu options: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Code Editor:** Displays the script file "linear_regression_example.R". A red box highlights the data preparation and regression analysis code:

```
32 # but we will cheat and get Karyotype (T21 status) information here from the "LabID"
33
34
35 # Prepare data for linear regression -----
36 # assumes "msd_data" is in long format already joined with clinical data
37 # "LabID" is unique sample identifier; "Analyte" denotes each feature of interest; "Value" is the actual measurement; here "Karyotype"
38 regressions_dat <- msd_data %>%
39   # select(LabID, Karyotype, Sex, Age, Analyte, Value) %>%
40   select(LabID, Analyte, Value) %>%
41   mutate(Karyotype = if_else(str_detect(LabID, "A"), "T21", "Control")) %>% # NEED TO REPLACE WITH ACTUAL CLINICAL DATA JOIN
42   mutate(
43     Karyotype = fct_relevel(Karyotype, c("Control", "T21")), # ensure factor levels in correct order
44     #Sex = fct_relevel(Sex, c("Female", "Male")), # ensure factor levels in correct order
45   ) %>%
46   group_by(Analyte, Karyotype) %>% # using both groupings here for categorical testing
47   mutate(extreme = rstatix::is_extreme(log2(Value))) %>%
48   ungroup() %>%
49   filter(extreme != TRUE) %>% # remove extreme outliers
50   # I would usually also check here for:
51   # 1) a minimum number of samples per group (eg 5) and
52   # 2) that there are >1 levels for categorical variables of interest (prevents errors in regression step)
53   nest(-Analyte) # nesting allows for easy testing of all features ~ at once
54
```
- Console:** Displays the R startup message and the command "R 4.1.2 · ~/..." followed by the R version information:

```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```
- File Explorer:** Shows the current directory structure:

| Name | Size | Modified |
|-----------------------------|---------|------------------------|
| linear_regression_example.R | 10.2 KB | Jul 15, 2022, 11:14 AM |
| test_output.txt | 12.9 KB | Jul 15, 2022, 11:14 AM |
- Environment:** Shows the message "Environment is empty".

CAVATICA: example linear regression analysis with HTP metabolomics

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the project name "HTP_metab_karyotype" and standard menu options: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Code Editor:** Displays the R script "linear_regression_example.R". A red box highlights the section from line 56 to line 65, which contains the code for running simple linear regressions. The entire code block is as follows:

```
43 Karyotype = fct_relevel(Karyotype, c("Control", "T21")), # ensure factor levels in correct order
44 #Sex = fct_relevel(Sex, c("Female", "Male")), # ensure factor levels in correct order
45 ) %>%
46 group_by(Analyte, Karyotype) %>% # using both groupings here for categorical testing
47 mutate(extreme = rstatix::is_extreme(log2(Value))) %>%
48 ungroup() %>%
49 filter(extreme != TRUE) %>% # remove extreme outliers
50 # I would usually also check here for:
51 # 1) a minimum number of samples per group (eg 5) and
52 # 2) that there are >1 levels for categorical variables of interest (prevents errors in regression step)
53 nest(-Analyte) # nesting allows for easy testing of all features ~ at once
54
55
56 # Run simple linear regression for each feature with log2(Value) as outcome and "Karyotype" as predictor -----
57 tic("Running linear regressions for simple model...")
58 regressions_simple <- regressions_dat %>%
59   mutate(
60     fit = map(data, ~ lm(log2(Value) ~ Karyotype, data = .x)),
61     tidied = map(fit, broom::tidy), # see ?tidy.lm
62     # glanced = map(fit, broom::glance), # see ?glance.lm # NOT NEEDED FOR DEMO
63     # augmented = map(fit, broom::augment), # see ?augment.lm # NOT NEEDED FOR DEMO
64   )
65 toc()
```

- Console:** Displays the R startup message and the command "R 4.1.2 · ~/" followed by the R version information: R version 4.1.2 (2021-11-01) -- "Bird Hippie".
- File Explorer:** Shows the file structure with "linear_regression_example.R" and "test_output.txt" in the "Home" directory.
- Environment:** Shows the message "Environment is empty".

CAVATICA: example linear regression analysis with HTP metabolomics

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the project name "HTP_metab_karyotype" and standard menu options: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Code Editor:** Displays the R script "linear_regression_example.R". A red box highlights the code from line 79 to 102. The script performs a linear regression analysis on log2(Value) using Karyotype as the predictor and other variables as nuisance variables.
- Console:** Shows the R startup message and the command "Run linear regression for each feature with log2(Value) as outcome, 'Karyotype' as predictor, and Age + Sex and nuisance variables".
- Output Area:** Displays the R version information and the results of the linear regression analysis.
- File Explorer:** Shows the file structure with "linear_regression_example.R" and "test_output.txt".
- Environment:** Shows that the environment is empty.

CAVATICA: example linear regression analysis with HTP metabolomics

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Contains the title "CAVATICA: example linear regression analysis with HTP metabolomics", the project name "HTP_metab_karyotype", and standard menu items: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Code Editor:** Shows the script file "linear_regression_example.R". A red box highlights the code from line 182 to 190, which generates a volcano plot:

```
182 # Volcano plot for simple model ----
183 lm_results_simple %>%
184   volcano_plot_lab_lm(
185     title="Differential abundance in T21 vs. Controls",
186     # subtitle with number significant up/down:
187     subtitle = paste0("[Down: ", .() %>% filter(BHadj_pval < 0.1 & FoldChange <1) %>% nrow(), "; Up: ", .() %>% filter(BHadj_pval <
188     )
189   )
```
- Console:** Displays the R startup message and help text.

```
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```
- Files Tab:** Shows the directory structure with files "linear_regression_example.R" and "test_output.txt".
- Environment Tab:** Shows the environment is empty.

To be visible in “Files” tab, output files must be saved here: /sbgenomics/output-files/

CAVATICA: example linear regression analysis with HTP metabolomics

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Contains the title "CAVATICA: example linear regression analysis with HTP metabolomics", the project name "HTP_metab_karyotype", and standard RStudio menu items like File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins.
- Code Editor:** Displays the R script "linear_regression_example.R". A red box highlights the section from line 194 to 208, which contains the code for creating a combined Sina + Boxplot. The script also includes comments and other R code for data filtering, plotting, and saving the output.
- Console:** Shows the R version information and the start of the R session. It includes the standard R welcome message and help information.
- File Explorer:** Shows the local file system with a folder structure. The "linear_regression_example.R" file is listed with a size of 10.3 KB and a modified date of Jul 15, 2022, 11:26 AM. A "test_output.txt" file is also present.
- Environment:** Shows the global environment, which is currently empty.