CSER Consortium Shared Dataset Documentation

CSER Data Coordinating Center (DCC)

Last Updated: July 11th, 2023

# Table of Contents

[Table of Contents 1](#_yp35a0jkyvvp)

[CSER site descriptions 2](#_fzy14m64ok2n)

[Shared dataset descriptions 3](#_4r6f79muv86o)

[Data Access 4](#_i77b0wsxh1av)

[How to obtain data access in AnVIL for CSER and beyond 4](#_ii8857j2za7k)

[How to search dbGAP for CSER data sets 4](#_7vnifpu9h79o)

[Accessing CSER data in AnVIL 6](#_izyqvb2vn8aw)

[Race/ethnicity in the Harmonized Measures, QPR, and Sequence Metadata 8](#_jilyp0fjn2pn)

[Harmonized Measures 14](#_hb8xv4xlve7t)

[Site-Level consent procedures 14](#_g6ww8lsg1bhj)

[Site-level measure adaptations 16](#_ejwne8cxfoi)

[Site-level survey administration protocols 16](#_k60abz88321d)

[Site-level survey QC procedures 17](#_pfuutbuqyl79)

[Site-level survey administration timelines 19](#_105kfnoqya67)

[Harmonized Measure FAQs 19](#_28t6wg8bj4oq)

[What is the difference between variables that end in \_pbl and those that end in \_abl? 19](#_or94hoerbxtb)

[Why is there free-text in the incom2\_\* variable responses? 19](#_dzikluftd8gi)

[What do the three different race measures in the Parent Proxy Baseline survey tell us? 19](#_hcl5a3pixn5x)

[How does each site define “primary parent”? 19](#_q8h1acbfbov2)

[Why are some pediatric ages (in age2\_pbl) 18+? 20](#_2mbvz5tix9nq)

[What reference date(s) do sites use to calculate age in the Harmonized Measures? 21](#_ex01gar44gu7)

[Sequencing Metrics 22](#_s1ghdsiz2kmt)

[Site-level application of case-level diagnostic categories 22](#_qlhzkncepid)

[Sequence Data/Metadata 24](#_7gtcqgbs2ob6)

[Site-level sequence data QC 24](#_1tfb3rns7n7t)

[Phenotype Data 27](#_ubl23e41w5y6)

[HPO Terms 27](#_cuf5mcc5xqyc)

[Calculated Fields 29](#_g6fshbww886y)

[Underserved Framework 29](#_y4rx2729fm71)

[References 29](#_sq9bmyqu09kg)

# CSER site descriptions

Brief descriptions of all CSER2 sites can also be found in the [CSER marker paper](https://pubmed.ncbi.nlm.nih.gov/30193136/) [[1]](https://paperpile.com/c/70t1i7/jyjD).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Institution** | **Abbrev.** | **Patient Population** | **Sequence Modality** |
| [CHARM](https://clinicaltrials.gov/ct2/show/NCT03426878) | Kaiser Permanente Northwest | kfri | Adults at risk for hereditary cancer | Germline ES versus usual care |
| [KidsCanSeq](https://www.bcm.edu/academic-centers/medical-ethics-and-health-policy/research/ethical-legal-and-social-implications-elsi-geneticsgenomics/texas-kidscanseq) | Baylor College of Medicine | bcm | Children with cancer | Germline ES, tumor ES and transcriptome versus targeted panel testing |
| [NCGENES 2](https://clinicaltrials.gov/ct2/show/NCT03548779) | UNC - Chapel Hill | ncch | Children with suspected genetic conditions (developmental disabilities, dysmorphology, neuromuscular disorders) | Germline ES versus usual care |
| [NYCKidSeq](https://clinicaltrials.gov/ct2/show/NCT03738098)/[TeleKidSeq](https://nyckidseq.org) | Mt. Sinai / Montefiore / NYGC | mtsi | Children with suspected neurologic, immunologic and cardiac genetic conditions | Germline GS versus targeted panel |
| [P3EGS](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002324.v1.p1) | University of California San Francisco | ucsf | Infants and children with severe developmental disorders, with or without congenital anomalies; parents whose fetus has a structural anomaly | Germline ES, duos or trios |
| [SouthSeq](https://clinicaltrials.gov/ct2/show/NCT03842995) | HudsonAlpha | haib | Newborns with suspected genetic conditions | Germline GS |
| [ClinSeq](https://clinicaltrials.gov/ct2/show/NCT00410241) | NHGRI | cseq | Healthy adults | Germline ES |

# Shared dataset descriptions

The three types of non-sequence data available for download on the AnVIL are the Harmonized Survey Measures and Outcomes (or **“Harmonized Measures”**), the Participant-level/Case-level sequencing metrics from the Quarterly Progress Report (or **“Sequencing Metrics”**), and sequence data/metadata. All CSER Harmonized measures and Sequencing Metrics are documented in the [Resources](https://anvilproject.org/consortia/cser/resources) section of the CSER home page on the AnVIL website.

Record counts for each data type and subtype in the final version of all datasets (“Freeze 3”) for each CSER project are listed in the following table:

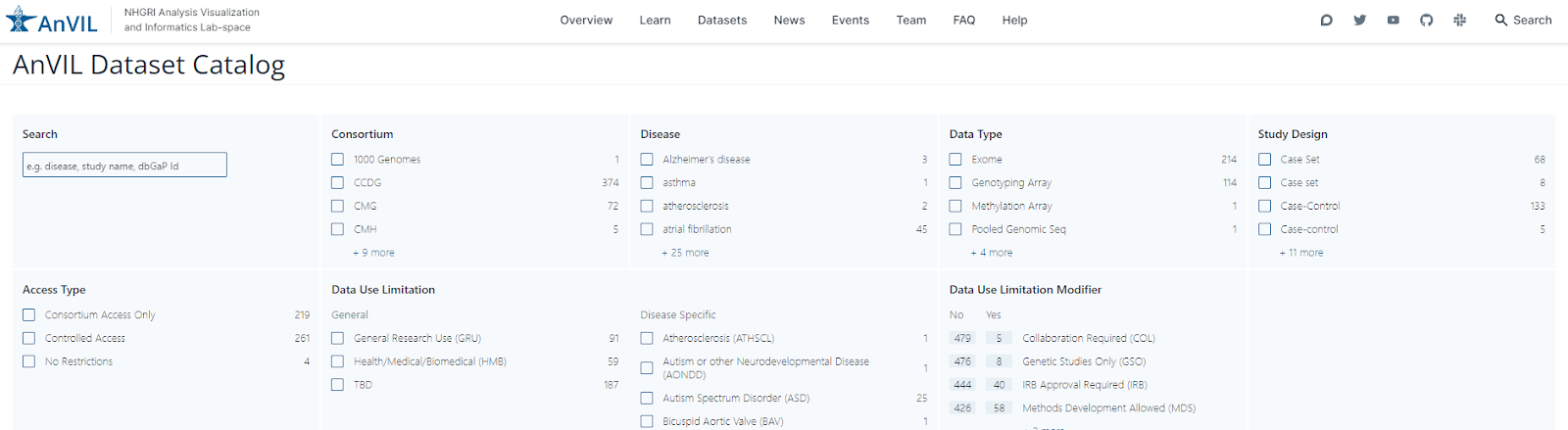
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Type** | **Subtype** | **kfri** | **bcm** | **ncch** | **mtsi** | **ucsf** | **haib** | **cseq** | **Total** |
| Harmonized Measures | Baseline | 941 | 589 | 241 | 718 | 530 | 537 | 408 | 3964 |
| Participant ROR/FU1 | 639 | 224 | 89 | 691 | 301 | 191 | 282 | 2417 |
| Participant ROR/FU2 | 595 | 162 | 107 | 636 | 313 | 154 | 4 | 1971 |
| Provider ROR/FU1 | 164 | 199 | 159 | 8 | 291 | 0 | 0 | 821 |
| Decliner | 86 | 27 | 1 | 0 | 0 | 0 | 0 | 114 |
| Sequencing Metrics | NA | 842 | 623 | 98 | 719 | 534 | 537 | 0 | 3353 |
| Sequence Data | NA | 827 | 623 | 96 | 718 | 534 | 752 | 0 | 3550 |

# Data Access

## How to obtain data access in AnVIL for CSER and beyond

Within Anvil there are three types of data access: open access, controlled access, and consortium-level access. Open access will provide all accessible data from Terra or Gen3. Controlled is just like it sounds, controlled. CSER is available to outside users via controlled access. Consortium-level access is available to users within a given consortium. For example CSER members have consortium access as per their data sharing agreement. More information on dbGap data sharing codes can be found [here](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4721915/).

If you want to quickly access data without an authentication process and preview what data sets are already integrated into AnVIL, you can access the dataset catalog [here](http://anvilproject.org/data). This catalog offers a breakdown of data by useful search terms. A screenshot of these datasets can be found below:



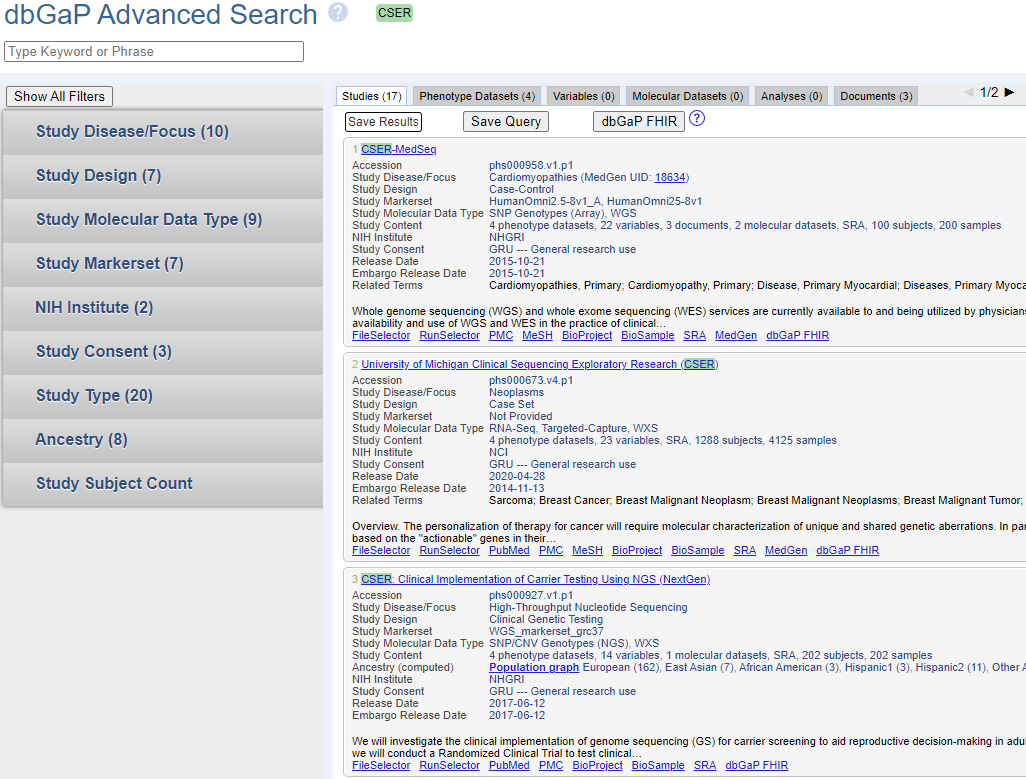
## How to search dbGAP for CSER data sets

If you are not sure which data set you want and want to search for CSER in dbGAP, follow the instructions below:

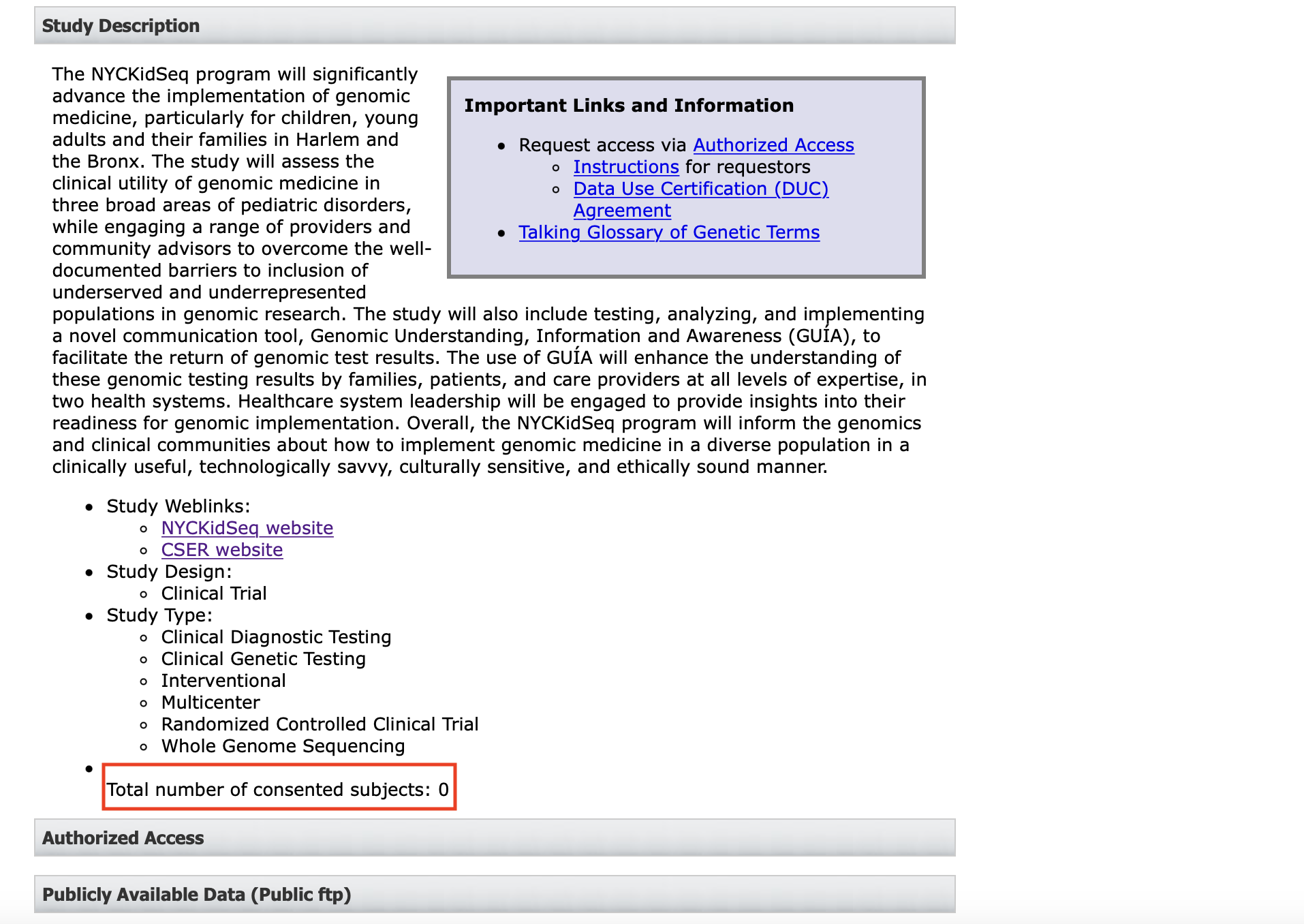
1. Navigate to the [dbGap](https://www.ncbi.nlm.nih.gov/gap/) home page on the NCBI website.
2. Type your keyword (i.e. CSER) into the search box.

# 

1. Review the results to determine which data sets you want to request.

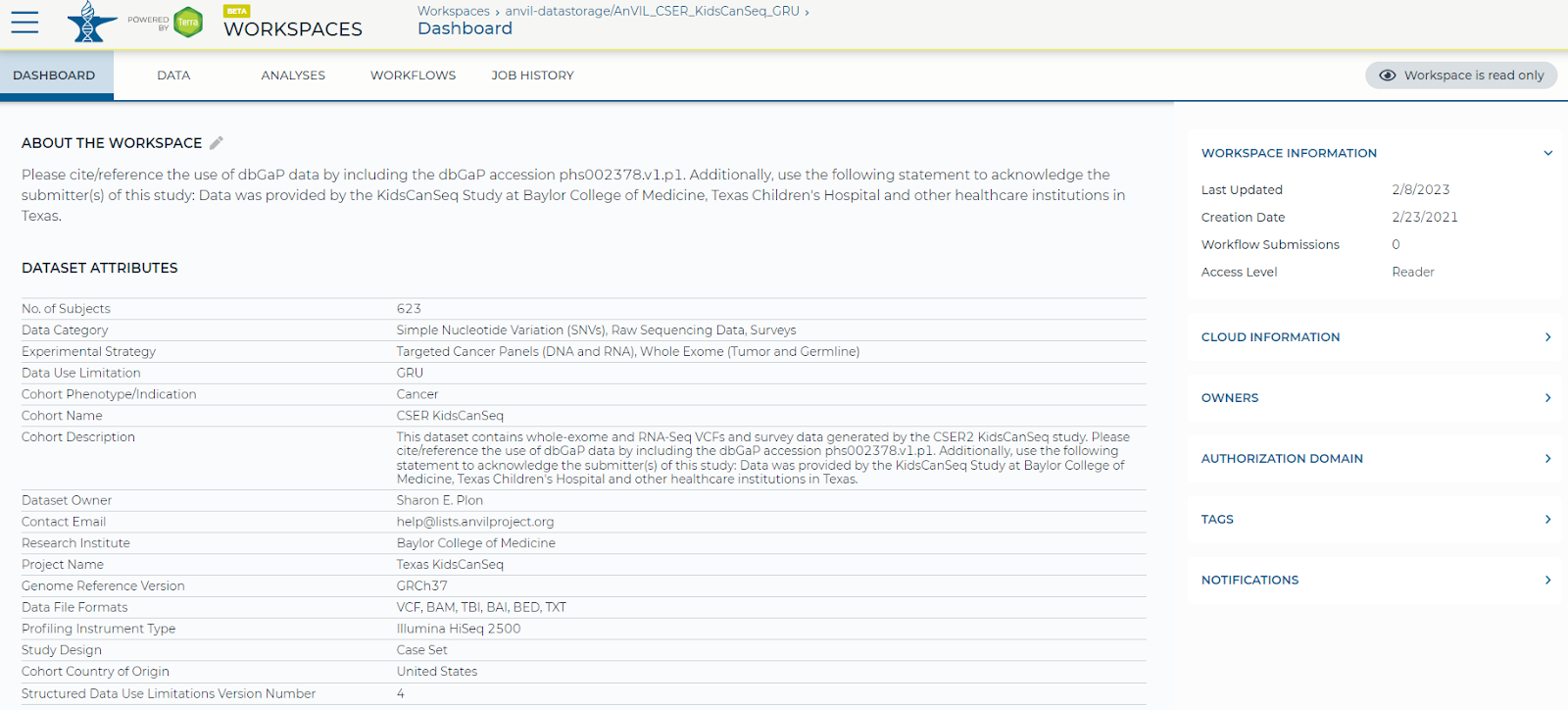


Please note that the total number of consented subjects on the study page will appear as “0” on most CSER study pages in dbGaP (as shown in the screenshot below). This is a known bug that the dbGaP team is currently working to address.

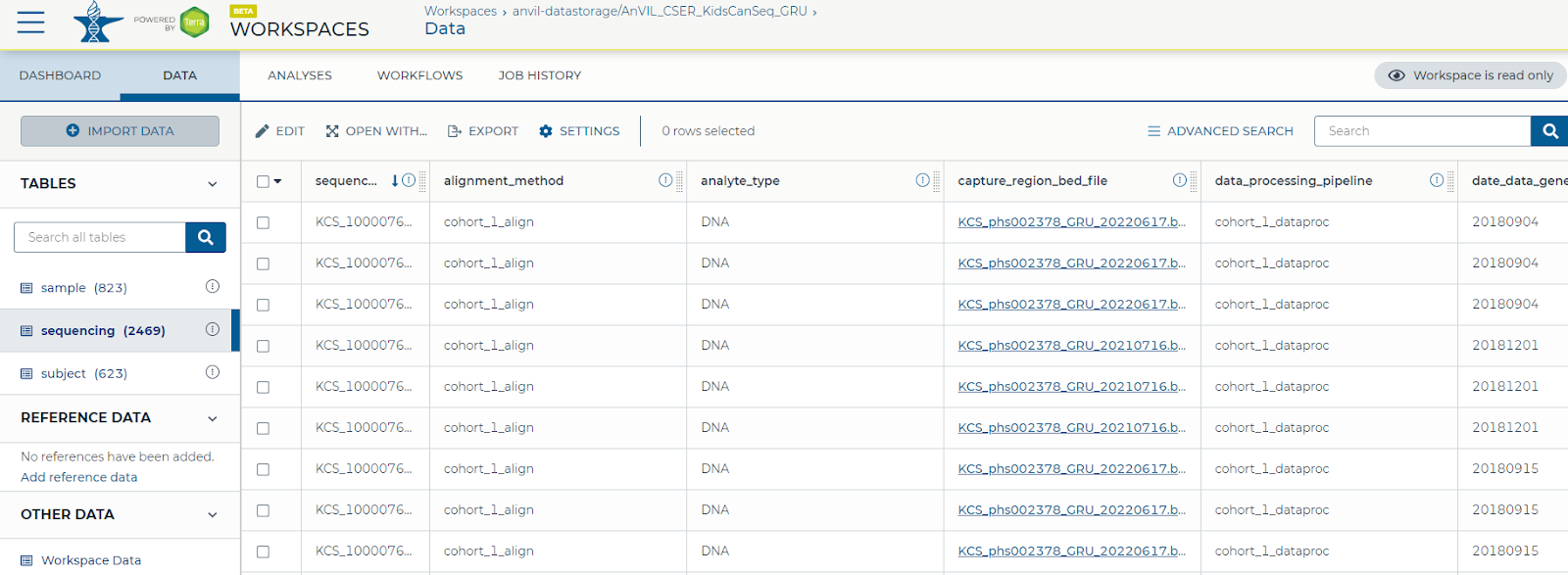


## Accessing CSER data in AnVIL

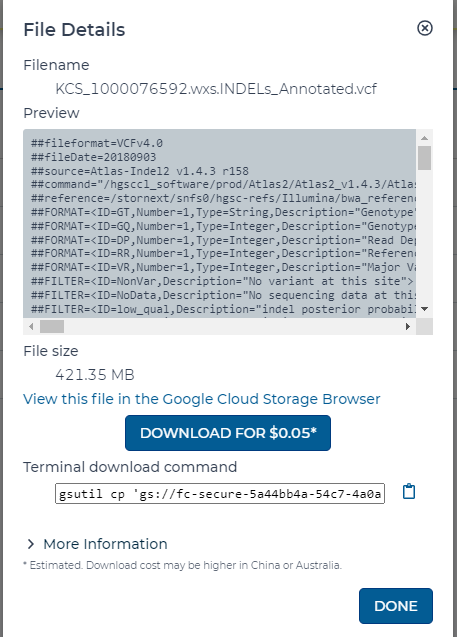
Not specific to CSER, once dbGap has approved your request your data will appear in AnVIL as a workspace on your [Terra workspaces page](https://anvil.terra.bio/#workspaces). An example of how a CSER dataset appears within AnVIL is shown below.



AnVIL provides a graphical user interface (GUI) that allows the user interface with components of the data.



The user may select files to download from the AnVIL workspace. To do this, select the file and follow the instructions below. AnVIL will list the associated cost with downloading the file along with alternate instructions to perform this task using the gsutil command-line tool.



# Race/ethnicity in the Harmonized Measures, QPR, and Sequence Metadata

Some sites report these measures differently. See below for description of how sites reported race/ethnicity measures in the Harmonized Measures vs. the Sequencing Metrics vs. AnVIL Sequence Metadata:

|  |  |  |  |
| --- | --- | --- | --- |
| **Study** | **Harmonized Measures Race/Ethnicity** | **Sequencing Metrics Race/Ethnicity** | **AnVIL Subject Metadata Table** |
| CHARM | The harmonized measure is the direct self report from the baseline survey. Data is missing for people who did not complete the baseline survey or that item on the survey. | For the QPR enrollment summary, we supplemented baseline self-report with race and ethnicity data from the patient's electronic medical record, for patients who were missing baseline data. | Same as QPR |
| ClinSeq | All demographic data, include race/ethnicity, was self-reported | No QPR data provided | NA |
| KidsCanSeq/  TeleKidSeq | Our race and ethnicity measure is asked in the Baseline survey and uses the CSER Harmonized measure that asks the Parent responding which category or categories best describe their child and separately themselves. We do not ask the Parent responding to report on the other Parent’s race and ethnicity. Some of those Parents (secondary parents) did also complete a baseline survey and responded on their own behalf but those data do not go to the DCC.  *What category or categories best describe [your child participating in KidsCanSeq OR you]? Please check all that apply.*  *American Indian, Native American, or Alaskan Native*  *Asian*  *Black or African American*  *Native Hawaiian or Pacific Islander*  *White or European American*  *Middle Eastern or North African/Mediterranean*  *Hispanic or Latino*  *I prefer not to answer*  *Unknown/None of these fully describe me* | Same as Harmonized Measures | Same as Harmonized Measures |
| NCGENES 2 | The harmonized measure is the direct self report from the baseline survey. Data is missing for people who did not complete the baseline survey or that item on the survey. | Same as harmonized measures | Same as harmonized measures |
| NYCKidSeq/  TeleKidSeq | Harmonized measure categories are used and administered within the baseline survey, parent reported. Data is missing for people who did not complete the baseline survey or that item on the survey. NYCKidSeq slightly adapted the questions (below) and added an additional option category “other, text” which is transformed into harmonized options.  *What category or categories best describe your child? Check all that apply. Mapped to* race**a**\_pbl\_\_\_[1-9]  *What category or categories best describe the child's mother? Check all that apply. Mapped to race****b****\_pbl\_\_\_[1-9]*  *What category or categories best describe the child's father? Check all that apply. Mapped to*  *race****c****\_pbl\_\_\_[1-9]* | Same as harmonized measures | Same as the harmonized measure for proband (child) |
| P3EGS | P3EGS Harmonized Measures for self-identified Race/Ethnicity are administered in the Baseline Survey Demographic Questionnaire. Pre-COVID pandemic, parents were asked these questions during in-person clinic visits for enrollment, separately for mothers and fathers. P3EGS does not collect race/ethnicity information for the proband/child. After the COVID pandemic, most of the enrollment was done via telehealth, and the baseline demographic questionnaire was administered by an interviewee during a zoom call, or on the phone. If the baseline survey could not have been completed at the time of enrollment, GCs and CRCs followed up with the patients to complete them with a phone call. P3EGS administered the baseline demographic survey with interpreters in several languages other than English and Spanish.    How would you describe your racial identity? (check all that apply)   * American Indian, Native American, or Alaska Native * Asian * Black or African-American * Native Hawaiian, Samoan, Other Pacific Islander * White or European American * Middle Eastern or North African/Mediterranean * Hispanic/Latino(a) * Prefer not to answer * Unknown/none of these fully describe me | For the QPR Enrollment Summary: Enrollment summary asks for the race/ethnicity of the proband, which we do not collect in P3EGS. In the past we reported the total numbers (separately for mothers and fathers) based on self-identified race/ethnicity which caused some challenges for identifying the total number of probands enrolled. More recently, we have been using the same harmonized self-identified race/ethnicity data, and reporting only mothers’ information when available as the primary parent for both pediatric and prenatal patients. We believe the data from the enrollment summary may require further consideration for any further analysis as a proxy for proband race/ethnicity, because, again, we did not collect race/ethnicity information for the proband/child.    For the QPR activities description, when we report URM status, we use a variable in our screening and tracking form, completed by the enrolling clinician, or by the CRC, either based on patient’s response or based on information in medical records. We collect this per patient, and have used it as a measure of URM, but not to report specific categorization of race/ethnicity. We do not report or use this measure in analysis. | Same as harmonized measures |
| SouthSeq | Our baseline measures are from the "Participant Enrollment" form which is filled out only by families enrolling into the clinical trial arm. Not every family can be enrolled into the clinical trial arm--some opt out, some were from before the trial began, some are ineligible due to language needs, etc. Each enrolled parent gets their own survey. This is worded similarly:  *What category or categories best describe you? Check all that apply.*  *American Indian, Native American, or Alaskan Native*  *Asian*  *Black or African American*  *Native Hawaiian/Pacific Islander*  *White or European American*  *Middle Eastern or North African/Mediterranean*  *Hispanic/Latino(a)*  *Prefer not to answer*  *Unknown/None of these fully describe me*  So, there's the potential for difference if people answer differently with the research nurse, plus the difference in not having a direct option for American Indian / Native American / Alaskan Native in the survey which feeds the QPR. | The QPR info comes from our "Enrollment" survey which is filled out by the research nurses during enrollment into SouthSeq, which is worded:  *Race and Ethnicity of mother? (checkbox, select-multiple)*  *Black or African American*  *White or European American*  *Asian*  *Hispanic or Latino*  *Middle Eastern, North African or Mediterranean*  *Native Hawaiian or Pacific Islander*  *Other: Please Explain (free text box)*  and a separate question for the father. | Same as harmonized measures |

## 

# Harmonized Measures

## Site-Level consent procedures

All CSER sites followed institution-level consent procedures for onboarding patients to their respective studies.

|  |  |
| --- | --- |
| **Study** | **Consent Process** |
| CHARM | Consent to the study includes sharing data with "federal databases." Participants are required to receive results related to cancer, but carrier findings and secondary findings are both optional. |
| ClinSeq | “The ClinSeq study was reviewed and approved by the NIH National Human Genome Research Institute and Suburban Hospital (Bethesda, Maryland) Institutional Review Boards. In addition, a certificate of confidentiality was obtained for the study to provide additional protection from forced disclosure of research results to third parties. At the initial enrollment of each participant, both the consent form and the associated discussion make it clear that the goal of ClinSeq is to examine the entire genome and to study any and all phenotypes, including a wide spectrum of diseases. The participants are informed that they will be contacted to determine if they are interested in learning about clinically relevant results, if discovered. In addition, participants are consented at the initial visit for permission to contact them to initiate discussions of additional follow-up testing of themselves and/or their family members (the latter being limited to basic phenotype studies and genotyping for cosegregation analyses). To protect the interests of the participants and to provide the investigators with an independent source of advice and review, we plan to establish a sequence variant review panel (i.e., a data-monitoring board). This panel will be comprised of experts in medical and molecular genetics who are otherwise not involved in ClinSeq; their charge will be to periodically review data regarding genes that have not yet been proven to cause human disease. For selected genes, we will pursue hypothesis-testing clinical research to determine if a sufficient data set can be generated to support causation and the return of results to the participants. Such research may involve studies of the population frequency of sequence variants in cases and controls, in vitro or animal models, and participants with and without the variants in response to specific pathophysiologic perturbations. The resulting data will be presented to the sequence variant review panel to assess if the findings warrant the return of individual research information to the participants” [[2]](https://paperpile.com/c/70t1i7/EFjT). |
| KidsCanSeq | There is not an option to share data; it is mandatory. We have an opt in or out to receive secondary findings for patients and parents. We have an opt in or out to share patient tumor tissue from past or future surgeries. We have an opt in or out for patients to participate in our circulating tumor dna substudy. We have an opt in or out for our parents, whose child is eligible, to reach out to the patient/child to participate in our AYA survey. |
| NCGENES 2 | NCGENES 2 collects a verbal consent by phone for participation in intervention 1 (pre-visit prep and surveys). NCGENES 2 has 2 hard copy consent (Consent to Randomization to Research GS and IF RANDOMIZED TO RECEIVE GENOMIC SEQUENCING, then Consent to Research GS), and 2 corresponding assents if relevant (Assent to Randomization to Research GS and Assent to Research GS (used ONLY for children both chronologically and developmentally 7 or older)). NOTE: Assents come in 2 different versions; 1) assents for children over 7 chronologically who are ages 7-14 developmentally, AND 2) assents for children over 7 chronologically who are ages 15-17 developmentally]. NCGENES 2 offers options at the time of each of the 2 hard copy consents. The first is for future use of study data after NCGENES 2 is over - this is on the Consent to Randomization to Research GS. The second and third options are on the Consent to Research GS and include:1) option to receive secondary/incidental sequencing results, and 2) option to allow future research use of specimen (DNA) and related data. These same options are available to assenting children 15-17 but not to assenting children 7-14. The opt in or out items are presented as follows: “Do you agree to the use of you and your/your child’s data collected for the NCGENES study in future studies after NCGENES?” (in consent and assent 15-17-year old to Randomization to Research GS) AND “Do you agree to the analysis of genes that could provide medically actionable information that is unrelated to your child’s current condition be done as part of your child’s genomic sequencing? AND Do you agree to the use of your child’s study specimen (DNA) and related data in future studies after NCGENES?" (in the consent and assent 15-17-year old to Research GS). |
| NYCKidSeq | We have a parental/guardian consent for pediatric participants, a consent for cognitively intact adults (18-21 yrs), and a surrogate consent for cognitively impaired adults (essentially the same as parental/guardian for pediatric participants). They consent to surveys (parents only), blood draws and genomic testing, EHR review, and to storing/use of residual sample and data within NYCKidSeq. Participants can choose to 1) opt-in or opt-out of secondary findings for their child (or themselves if cog intact adult), 2) opt-in to sharing de-identified data/samples with researchers outside of NYCKidSeq (includes CSER) a), 3) opt-in to public sharing of genome date with secure, public databases (e.g., AnVIL, dbGap, ClinVar, other secure external databases). We also ask their preferences for participating in future research: a) to be contacted in the future to learn about new research studies, b) to be contacted in the future if the researchers would like additional samples.  Participants, who are cognitively intact and turn 18 during the study, will be re-consented and data sharing preferences will be noted. Modifications to the initial CSER consent bucket will be made. |
| P3EGS | P3EGS pediatric biological parents are asked to consent on behalf of the proband and themselves. We have IRB approval for the pediatric arm to enroll up to 25 years of age for those patients who have been in the care of UCSF pediatrics department and fulfill enrollment criteria, and were offered testing even after 18 years of age. Consent is requested from both parents in the prenatal arm. Maternal consent form for the prenatal arm is used to consent biological mother and her fetus; paternal consent form is used for fathers only. P3EGS specifically asks participants with whom their genomic and research data can be shared with. They are informed that research and clinical data sharing is mandatory for study participation (sharing sequence data, harmonized survey data, phenotypic data, or result data with P3EGS collaborators, UCSF researchers, CSER/DCC, and ClinVar). We also inform them that participants may choose to opt out of sharing sequence data, which is tracked on their consent forms as consenting to sharing data with “other research databases”. This is explained to patients as “government controlled-access databases” to include dbGap, AnVIL, or similar platforms. This question for broad data sharing is asked separately for the proband, mother and father. If they consent to broad data sharing, we code them as GRU (code label=1), for both sequence and survey data; if not, they are coded as CSER only (code label=6). Parents are also asked for permission to use samples for future research (Yes/No), whether or not to receive secondary findings (Yes/No) and whether or not to be recontacted for future research (Yes/No). |
| SouthSeq | We have consent documents for each site, but in general, a signed consent allows for (1) blood draw for sequencing/Sanger confirmation, (2) completion of study surveys, (3) return of results and (4) sharing findings and de-identified info with ClinVar.  Our participants are currently given the option to opt in or out of the following : (1) storage of specimens at HA for future research, (2) sharing data with NIH-designated repositories like dbGAP and AnVIL, (3) contact for future research opportunities, and (4) consent to receive secondary findings. |

## Site-level measure adaptations

The current version of the cross-site Harmonized Measures **Adaptation Dictionary** can be found in the [Resources](https://anvilproject.org/consortia/cser/resources) section of the CSER home page on the AnVIL website.

## Site-level survey administration protocols

|  |  |
| --- | --- |
| **Study** | **Survey Randomization and/or Administration Protocols** |
| CHARM | Surveys were not randomized. Follow-up surveys were only sent to patients who received their sequencing results. |
| ClinSeq | “Each participant also provided self-reported demographic data including: age, gender, education, income, number of children, marital status, race, and ethnicity. ClinSeq participants who were enrolled for at least one month but had not yet received a sequence result were eligible to complete” an additional “survey on health-related socio-behavioral factors…Participants were recruited by mail, phone, or secure email. The survey was available both online and in a paper version; scales were administered in random order for each participant taking the online survey. Surveys responses were collected from August 2012 to April 2015. The survey stated that participant consent to participate was implicit in their completion of the survey, which was approved by the National Human Genome Institute (NHGRI) Institutional Review Board (IRB)” [[3]](https://paperpile.com/c/70t1i7/HhoA). |
| KidsCanSeq | Survey administration was matched at the provider level - when a provider had a patient with a positive result who got a survey, their next patient with a negative finding also got a survey. |
| NCGENES 2 | Participants were randomized into two interventions simultaneously (pre-visit preparation and genomic sequencing), creating four distinct groups: 1) pre-visit prep intervention and genomic sequencing, 2) pre-visit prep intervention and no genomic sequencing, 3) no pre-visit prep intervention and genomic sequencing, 4) no pre-visit prep intervention and no genomic sequencing. Survey administration was randomized at the site (UNC, Mission, ECU), clinic level (pediatric genetics, pediatric neurology), and by served/underserved and represented/underrepresented status of eligible participants. |
| NYCKidSeq/  TeleKidSeq | Surveys were administered to all participants (or parent proxies) who provided consent. |
| P3EGS | Surveys were administered to all participants who were consented. Baseline demographic survey was administered by an interviewer, either during the clinical visit for enrollment, or later through phone interviews. The two follow up surveys (FU1/Post-ROR and FU2) were administered via e-mail for modified versions for pediatric patients, prenatal patients with ongoing pregnancy, and prenatal patients with terminated pregnancy. |
| SouthSeq | Surveys were administered to all participants who provided consent, regardless of the group to which they were randomized. The participants were randomized into two groups: (1) received genome sequencing results from a genetic counselor or (2) received genome sequencing results from a trained non-genetics provider. The primary outcome of our study was to measure any differences across the two groups by utilizing the GCOS. |

## Site-level survey QC procedures

Before CSER sites upload data to the DCC, they performed a rigorous set of data quality control procedures locally.

|  |  |
| --- | --- |
| **Study** | **Data Quality Control Procedures (Pre-DCC)** |
| CHARM | Participants enter survey data directly into our REDCap database. CHARM data analysts import the REDCap into SAS where we run validity checks for missing variables, outliers and errors, as well as removing ineligible surveys. Our data dictionary has been mapped against the CSER harmonized data dictionary; we recoded the CSER harmonized variables based on the provided CSER data dictionary where relevant and excluded variables where the response options could not be recoded. All programming was reviewed with a second analyst. |
| ClinSeq | None provided |
| KidsCanSeq | The electronic survey data are pulled by the database programmers and recoded to map onto the CSER harmonized variables per the CSER Data Dictionary. The database programmers run a QC check after generating the data as a report from the database. Recoded data is then verified back to our KCS site-specific coded data for QC purposes by a data analyst and approved by the research manager. Any residual issues are brought back to the database programmers for recoding, and the data are checked again before upload to the REDCap database. |
| NCGENES 2 | The NCGENES 2 data collection and tracking system has automated logic and range value checks. Additionally, every attempt is made to ensure appropriate completion of the paper version of the Intake Survey (baseline) returned to the research team at the clinic visit. Additionally, the NCGENES 2 analyst reviews all surveys for missingness including missingness due to 1) skip patterns appropriately applied, and 2) missingness due to non-completion by parents either purposefully skipped or skipped in error (the latter 2 cannot be differentiated). Missing data reports by survey are sent to each site study coordinator for resolution. Resolution includes comparing electronic to paper versions, and either updating the data collection form or indicating that the parent skipped the question. This information is returned to the data analyst who then generates a cleaned analytic data set for upload to CSER at the next quarter. These missing data reports are generated and resolved quarterly. |
| NYCKidSeq/  TeleKidSeq | Survey data is entered into RedCap. QC checks are frequently run to check for missing variables, outliers and errors. Our data dictionary has been mapped against the CSER harmonized data dictionary; we will be extracting the data, and recoding the CSER harmonized variables based on the provided CSER data dictionary. Transformed data is checked against raw data. |
| P3EGS | Survey data are entered in RedCap. Key variables are routinely and systematically reviewed for accuracy/completeness. Project staff and GCs complete monthly data cleaning exercises depending on project needs. Necessary data mapping/transformations are discussed by site PMs/DMs & programmers and effected according to agreed upon rules. Transformed data (for QPR, RPPR, DCC) are compared with P3EGS RedCap data to check for errors, which are discussed and resolved. |
| SouthSeq | Before uploading, we ensure that all mapped variables have at least one response (including from test data) and that all response variables have been properly converted to the DCC response scale. (Test data is NOT submitted to the DCC, so variables with 0 responses are possible.) |

## Site-level survey administration timelines

Details on when follow-up surveys were administered at each CSER site can be found in the [Resources](https://anvilproject.org/consortia/cser/resources) section of the CSER home page on the AnVIL website.

## Harmonized Measure FAQs

### What is the difference between variables that end in \_pbl and those that end in \_abl?

Throughout the dataset, you will see most variables ending in either “\_pbl” or “\_abl”. These suffixes correspond to survey items answered by a parent proxy on behalf of a child or dependent, and to items answered by the patients themselves, respectively.

### Why is there free-text in the incom2\_\* variable responses?

While this question asks participants to provide the number of dependents, participants were allowed to respond using free-text. For example, some responses may include text like “2 adults and 2 children.”

### What do the three different race measures in the Parent Proxy Baseline survey tell us?

* race**a**\_pbl\_\_\_[1-9]: These checkboxes describe the **child’s** race, as reported by the parent completing the Baseline survey.
* race**b**\_pbl\_\_\_[1-9]: These checkboxes describe the race of the **primary parent** (see [How does each site define “primary parent”?](#_q8h1acbfbov2) for details)
* race**c**\_pbl\_\_\_[1-9]: These checkboxes describe the race of the **non-primary parent**.

### How does each site define “primary parent”?

Some response categories in the Parent Proxy measures only allow a response from one parent. For these responses, sites were asked to identify a “primary parent” whose responses were to be used. See below for descriptions of how sites defined the primary parent:

|  |  |
| --- | --- |
| **Study** | **Primary Parent Definition** |
| CHARM | N/A, adults only |
| ClinSeq | N/A, adults only |
| KidsCanSeq | When two parents are enrolled, both are invited to take the surveys separately and one parent is designated as the primary parent who is required to complete the baseline survey. Each parent who completes a baseline survey is also sent a follow-up survey. Although we would expect the primary parent to be the one who will complete surveys longitudinally, either could complete surveys at the follow-up time points. We capture relationship to patient (Mother, Father, Legal Guardian). |
| NCGENES 2 | The primary parent/guardian in NCGENES 2 is the person that will bring their child to the clinic visit, are 18 years of age or older, are able to legally sign documents (e.g., consent) for the child, and are able and willing to complete surveys in either English or Spanish (e.g., consent by phone to the study participation in the first intervention (specifically pre-visit preparation or not – that includes activities such as reading mailed material and completing surveys) |
| NYCKidSeq/  TeleKidSeq | The primary parent (or caregiver) is the one who primarily: administers medications to the child, takes the child to medical appointments, and manages the child’s personal and healthcare needs or is most familiar with these activities. This parent is the caregiver who will attend all research visits / answer all surveys.  The survey question asked to capture relationship:  "What is your relationship to [child's name]?"  1, Mother  2, Father  3, Legal guardian  We asked race and ethnicity for each parent at baseline, see race/ethnicity section. The parent race/ethnicity transformed data was not linked to the primary parent (relationship to child asked at baseline). *race****b****\_pbl\_\_\_[1-9] = mother and race****c****\_pbl\_\_\_[1-9] = father.* |
| P3EGS | P3EGS does not define "Primary Parent", but collects demographic information from both parents separately. (Please see our [note above for the collection of race/ethnicity data](#_jilyp0fjn2pn)). In cases the DCC template includes only one field to upload data for a family, P3EGS’ default is to upload data from mothers. If only the father's responses are available, then we upload father's data. If no answer is available from either parent, we will leave the field blank. |
| SouthSeq | Both parents (or legal guardians) are enrolled, if willing. The primary parent for the clinical trial survey analysis is the mother. |

### Why are some pediatric ages (in age2\_pbl) 18+?

CSER sites used different procedures for collecting and recording age in some cases. See the table below for more details:

|  |  |
| --- | --- |
| **Study** | **Procedure for pediatric age data collection (age2\_pbl)** |
| CHARM | N/A, adults only |
| ClinSeq | N/A, adults only |
| KidsCanSeq | No pediatric ages >18 submitted. |
| NCGENES 2 | N/A, pediatric participants only, <16 at eligibility. |
| NYCKidSeq/  TeleKidSeq | “Pediatric” patients with ages 18+ had conditions with childhood onset and have been in follow-up through our pediatric geneticists. We have IRB approval to enroll up through21 years of age who otherwise fulfill the inclusion criteria. Parents fill out surveys for all, including age 18+. |
| P3EGS | P3EGS has enrolled a few “pediatric” patients over the age of 18. We have IRB approval to enroll up to 25 years of age. The 18+ patients enrolled were receiving care from UCSF Pediatrics department, and eligible for the study, and thus were offered enrollment for exome sequencing. |
| SouthSeq | Parents/legal guardians completed the surveys as the probands in our study were newborns-12 months of age. |

### What reference date(s) do sites use to calculate age in the Harmonized Measures?

While participant age is reported in the Baseline measures, not all sites report age using the survey completion date. Age is not reported again in the ROR/FU1 or ROR/FU2 measures, but is used to determine the version of particular measures (e.g. the PEDSQL) that participants are administered.

|  |  |  |  |
| --- | --- | --- | --- |
| **Site** | **Baseline** | **ROR/FU1** | **ROR/FU2** |
| UCSF P3EGS | Eligibility date (peds only) | Haven't used, but would use survey completion date | Haven't used, but would use survey completion date |
| BCM KidsCanSeq | Consent date | Consent date | Consent date |
| NYCKidSeq/  TeleKidSeq | Survey completion date | N/A | Survey completion date |
| SouthSeq | Consent date | Consent date | Consent date |
| NCGENES 2 | Eligibility date | N/A | Date of FU2 survey mailing |
| CHARM | Survey completion date | Haven't used, but would use survey completion date | Haven't used, but would use survey completion date |

# Sequencing Metrics

## Site-level application of case-level diagnostic categories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CSER project** | **Procedures and personnel assigning phenotypic category** | **Study personnel making case-level classification** | **Time point(s) at which classification is made** | **Protocols for updating classifications** |
| CHARM | All phenotypes were reported as “Other”, since participants were recruited if they were determined to be at increased risk to have a hereditary cancer syndrome rather than having cancer at the time of the study | Genetic counselor | During preparation of submission of sequencing metrics data | Case level diagnostic categories were updated if a variant was reclassified during the course of the study. |
| KidsCanSeq | All subjects required to have cancer as a phenotype. Medical records were reviewed by site genetic counselors at each subsite and entered into a phenotype checklist in the study database. Lauren Desrosiers (GC) and study PI (ABMGG clinical geneticist) reviewed and made final determination.. | Weekly review of exome test results with study PI (ABMGG), Lauren Desrosiers (GC) and GC staff from each subsite. | Determination made at time of exome report review or prior to result disclosure. | Need for update raised at weekly meeting to review exome results. |
| NCGENES2 | Selected from available categories by enrolling clinician at initial visit | Molecular pathologist (in consultation with multidisciplinary sign-out committee) | Prior to initial result reporting. Re-assessed after discussion with study clinician and/or after secondary testing performed (e.g., parental testing to determine phase) | Study clinicians may prompt re-evaluation with additional clinical information. Sign-out committee meets with study clinicians on semi-annual basis to review results and case-level classification |
| NYCKidSeq/  TeleKidSeq | Referring provider selects phenotypes. This is input into a study defined phenotype checklist for our primary phenotypes and also includes open text fields for other congenital systems. | Genetic counselor in consultation with the referring provider and/or discordance committee. | Prior to initial result reporting. | Study clinicians may prompt re-evaluation with additional available parental sample(s) or clinical information. In addition, the lab may prompt re-evaluation based on changes to technology. Depending on the outcome of changes, the result will be returned to the family by the study team and/or inform the referring provider and update EHR and study database with amended report data. |
| P3EGS | Providers (pediatric and prenatal clinicians) select phenotypes. | Pediatric and prenatal PIs, in consultation with multidisciplinary exome sign-out committee | Prior to initial result reporting. | Study clinicians may prompt re-evaluation with additional clinical information. |
| SouthSeq | The SouthSeq nurse coordinator entered phenotypes from the child’s medical record into the enrollment software used for our study. The nurses were allotted a free test as well as a list of common HPO terms to which they could select relevant conditions.The list was created by our team of researchers and clinicians to capture the major conditions/features observed in the target population. | The research team, with input from the clinical team, when necessary, made case-level designations for findings identified via genome sequencing. The clinical team would review results and voice any comments to the research team prior to returning results to the enrolled participant families. | Prior to initial result reporting. | If the clinical team asked for re-evaluation of a case or if the phenotypic information was updated and disclosed to the research team after the initial result disclosure, the case may be reanalyzed and classifications may be updated accordingly. |

# Sequence Data/Metadata

## Site-level sequence data QC

|  |  |
| --- | --- |
| **Study** | **Sequence Data QC Process** |
| CHARM | Next generation DNA sequencing identified variants, including single base substitutions, small insertions/deletions (<50bp) and deletions/duplications (single exon to whole gene) in the coding portions of the targeted genes. >95% of the coding regions and canonical splice sites were sequenced to a read coverage of >20X and compared to the human reference genome. This assay does not detect variants located: 1) outside the captured target, 2) in regions of insufficient coverage, 3) in regions containing paralogous genes or pseudogenes and 4) where the reference genome is inaccurate or contains gaps and insertions. This test has limited ability to detect low-level mosaicism and is not capable of detecting repeat expansions (ie. tri-nucleotide). Deletions and duplications, larger than a single gene, will not be detected. |
| ClinSeq | “We have implemented a number of steps for the quality control of ClinSeq-generated sequence data. For example, to ensure that a particular DNA sample being sequenced is, in fact, the same DNA sample analyzed at another point in time, we use a novel sample-tagging strategy that we developed. Specifically, upon arrival at the NIH Intramural Sequencing Center (NISC), each human DNA sample is “spiked” with an aliquot of a plasmid containing a unique segment of nonprimate DNA whose sequence is known not to be present in the human genome. Upon every use of that human DNA sample, the insert of the spiked plasmid is PCR-amplified and sequenced, so as to confirm that the correct spike DNA is present (and thus that the same ClinSeq participant DNA was used as previously). This approach does not tie the research DNA sample to the CLIA (Clinical Laboratory Improvement Amendments) DNA sample; in fact, ClinSeq has been designed with an information and data firewall that separates testing of the research DNA sample from the clinical data.  Our quality control measures for detecting sequence variants involve the routine analysis of HapMap ([The International HapMap Consortium 2007](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2752125/#B15)) DNA samples, which are associated with extensive genotype data at known variant sites. Specifically, for every 30 ClinSeq participant DNA samples analyzed, we sequence the same genomic regions in one HapMap sample. The sequence generated with the HapMap samples is interrogated for the presence of known variants (from the available genotype data) to assess our overall false-negative rate. Such quality control steps facilitate the generation of reliable data, the estimation of false-negative rates, and accurate variant detection” [[2]](https://paperpile.com/c/70t1i7/EFjT). |
| KidsCanSeq | The following sequencing quality metrics were assessed for performance and used as the criteria for passing.  >50 % of the reads aligned to target/buffer  >95% of targeted bases covered at 20x  >85% of targeted bases covered at 40x  Average coverage >100x  In addition to the above sequencing quality metrics, the HGSC developed quality control software ERIS was used to assess contamination and validate the identity of samples by comparing sequencing reads to the Fluidigm SNPtrace array data.  Concordance between SNPtrace and sequencing data >90%  Contamination rate <10% |
| NCGENES 2 | Sequencing quality metrics were reviewed including mean coverage of the targeted exome region (goal >= 80x, minimum >= 50x), percent of targeted region covered at at least 20x (minimum >= 90%, goal >= 95%), and percent of reads mapped (minimum 97%). Correspondence of research/exome sample data to clinical laboratory samples was verified by comparision with Sanger sequencing of the clinical laboratory samples at 8 polymorphic markers. |
| NYCKidSeq/  TeleKidSeq | Sequencing was performed at two clinical laboratories and their quality metrics are included here.  New York Genome Center (n=972): Sequence quality metrics that were reviewed and used as criteria for passing include mean genome coverage (>= 27x), percentage of bases achieving at least 20x coverage ( >= 85%), AT and GC dropout rates (< 5 for both), duplication rate (< 20%), base quality score (75% bases above Q30). To ensure sample identity, array-sequencing genotype concordance (> 99%), and cross-sample contamination (< 2%) were also measured. Manual review and corrective actions for any failed metric.  Rady Children’s Institute Genomic Medicine (n=80): *to be completed.* |
| P3EGS | Sequencing quality metrics at UCSF include mean target coverage of the exome region (goal >= 100x, minimum >= 55x), percent of targeted region covered >10x (minimum >= 90%, goal >= 98%), and percent of reads aligned (minimum 94%). UCSF verified the expected sex (via chr X heterozygous counts in comparison to test requisition form and electronic health records) of all samples and expected relatedness among all families. Sample contamination was assessed by FreeMix. |
| SouthSeq | Sequencing QC was performed by the HudsonAlpha CSL prior to release of data to SouthSeq investigators. SouthSeq required mean coverage >= 30x and the percentage of bases in the human genome covered at >= 20x to be at least 80%. Current CSL processes also monitor numerous metrics including percentage of reads >= Q30, Qscore, and % aligned reads as part of their SOP for sequence acceptance. SouthSeq investigators verified expected sex (via chrX heterozygous counts in comparison to sample order forms and clinical records) of all samples and expected relatedness among all families (via KING or somalier software). Before upload to AnVIL, all samples were re-aligned to hg38 using the latest CSL pipeline, including generation of a PDF report with detailed sequencing and alignment metrics (including Reads, duplications, and pass rates; bases, adapters rates and Q scores; read groups, alignments, insert size, aligned/duplicate/discordant mappings, and coverage). |

# Phenotype Data

## HPO Terms

Some CSER sites provided HPO terms (present and/or absent) in addition to the broad phenotype groups reported in the Sequencing Metrics. These terms can be found in the **hpo\_absent** and **hpo\_present** fields in the Subject metadata table in each AnVIL workspace. Please see below for details on whether each site provided HPO terms, and if so, how they were collected and reported:

|  |  |  |
| --- | --- | --- |
| **Study** | **Provided HPO terms?** | **Collection and Reporting Procedures (if HPO provided)** |
| CHARM | Not provided | NA |
| ClinSeq | Not provided | NA |
| KidsCanSeq | Provided | As part of the KidsCanSeq protocol the genetic counselor or coordinator at each site reviewed the medical record to complete the germline exome requisition form for the pediatric cancer patients enrolled in the study. The form includes questions about the patient’s cancer diagnosis and other phenotypes using HPO terms for major organ systems. Based on information available in the medical record, HPO terms were added to the requisition form, submitted with the sample for testing, recorded in our study database and transmitted to the CSER DCC/ANViL. Three notes, (1) HPO descriptors were not used to describe the patient’s cancer diagnosis (Orphanet IDs were provided for that) or medical problems that appeared to be directly related to cancer treatment, e.g., neutropenia. Thus, the majority of KidsCanSeq patients have no HPO terms noted, (2) No attempt was made to record HPO terms that were absent, (3) the GC could also enter free text describing the patient’s phenotype or family history if HPO terms weren’t available but that information was not transmitted to the DCC. |
| NCGENES 2 | Provided | HPO information was entered by the enrolling clinician (or study clinical geneticist reviewing the clinic note from initial visit) directly into a locally-hosted instance of Phenotips. |
| NYCKidSeq/  TeleKidSeq | Provided | The NYCKidSeq project provided HPO terms on a subset of cases that were sequenced via the RADY lab. As per the RADY lab: Phenotypic information from the affected children receiving WGS was utilized to derive associated Human Phenotype Ontology terms through a comprehensive review of the of the clinical notes provided to the laboratory. The HPO terms were used to prioritize and analyze WGS data by qualified genome analysts and licensed genetic counselors. Cases are reviewed and signed out as a clinical WGS report by certified and licensed laboratory directors. |
| P3EGS | Provided | HPO data was generated using the  ClinPhen NLP program to extract HPO terms from case notes recorded  before the diagnosis was made. Textract was used to extract text from PDF notes, then the text was processed by ClinPhen to produce a list of HPO terms, which were then formatted for the subject table. Our ClinPhen run used the HPO.obo version releases/2018-12-21. |
| SouthSeq | Provided | SouthSeq curated a list of potential HPO terms and included them on a phenotyping survey administered to the site enrollment staff (research nurses or genetic counselors). HPO terms and codes were presented on the phenotype survey as select-multiple checkboxes and organized by body system. Phenotypes were primarily collected at the time of enrollment and often were updated throughout the period between enrollment and ROR based on the emergence of additional relevant phenotypes and site staff capacity. HudsonAlpha analysts reviewed free text fields from the phenotyping survey and generated a list of common findings and their associated HPO terms and additional HPO terms were generated based on simple text-matching against the phenotyping survey free-text fields. The free-text matching was likely subject to omissions due to formatting, spelling, and varying use of multiple free-text fields. |

# Calculated Fields

## Underserved Framework

The CSER ELSI & Diversity WG is currently developing a manuscript that will describe the Underserved Framework 2.0 in detail, and can be used to reference the framework in future CSER publications. In the meantime, the Underserved Framework can be referenced in CSER publications using the following (or similar) phrasing:

*The Underserved Framework 2.0 employs different combinations of demographic factors (including language, income, insurance status, residence, race and ethnicity) to form 9 distinct risk groups, each of which reflects direct barriers to medical care access and/or contextual factors that may exacerbate barriers to access. The framework was not developed to be a reflection of how participants were recruited into each CSER study, but rather to offer guidance on the implementation of variable “underserved” criteria in clinical research.*

# References

1. [Amendola LM, Berg JS, Horowitz CR, Angelo F, Bensen JT, Biesecker BB, et al. The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. Am J Hum Genet. 2018;103: 319–327.](http://paperpile.com/b/70t1i7/jyjD)

2. [Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, et al. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome Res. 2009;19: 1665–1674.](http://paperpile.com/b/70t1i7/EFjT)

3. [Lewis KL, Han PKJ, Hooker GW, Klein WMP, Biesecker LG, Biesecker BB. Characterizing Participants in the ClinSeq Genome Sequencing Cohort as Early Adopters of a New Health Technology. PLoS One. 2015;10: e0132690.](http://paperpile.com/b/70t1i7/HhoA)

4. [Mittendorf KF, Kauffman TL, Amendola LM, Anderson KP, Biesecker BB, Dorschner MO, et al. Cancer Health Assessments Reaching Many (CHARM): A clinical trial assessing a multimodal cancer genetics services delivery program and its impact on diverse populations. Contemp Clin Trials. 2021;106: 106432.](http://paperpile.com/b/70t1i7/E6qS)