

Uniform search of public sample and sequence read information  
across NCBI (SRA and dbGaP) and Kids First repositories

**Title:** Uniform search of public sample and sequence read information across NCBI and Kids First repositories. Search UC1 (GitHub NCPI Issue #19)


---

### **Scientific Question/Use Case:**




A researcher/user has a question, a general phenotype, question of interest and wishes to find all nucleotide data and obtain project and sample level details. Researchers in general do not know the specific repositories to search. The goal is consistency in results regardless of the search portal. Search use case is “orofacial cleft”.

# SRA (the world's largest sequence registry by far - coordinated internationally with Japan and EBI)

insdc.org
fer to depo...
Data Repository S...
International INSDC...
Kids First DRC Su...
Kids First DRC Su...
Presentation - I...
Maelstrom Resear...


International Nucleotide Sequence Database Collaboration

ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS

### International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between [DDBJ](#), [EMBL-EBI](#) and [NCBI](#). INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	<a href="#">Sequence Read Archive</a>	European Nucleotide Archive ( <a href="#">ENA</a> )	<a href="#">Sequence Read Archive</a>
Capillary reads	<a href="#">Trace Archive</a>		<a href="#">Trace Archive</a>
Annotated sequences	<a href="#">DDBJ</a>		<a href="#">GenBank</a>
Samples	<a href="#">BioSample</a>		<a href="#">BioSample</a>
Studies	<a href="#">BioProject</a>		<a href="#">BioProject</a>

- The INSDC advisory board, the [International Advisory Committee](#), is made up of members of each of the databases' advisory bodies. The International Advisory Committee published a [paper](#) reiterating the importance of depositing data to INSDC.
- Individuals submitting data to the international sequence databases should be aware of [INSDC policy](#).

### How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- [DDBJ](#), [ENA](#), [GenBank](#)
- The INSDC Feature Table Definition Document is available [here](#).

# SRA (the world's largest sequence registry by far)

NCBI
Resources
How To
adeslat
My NCBI

SRA
SRA
all[sb]
Search
Create alert
Advanced

COVID-19 Information
Public health information (CDC)
Research information (NIH)
SARS-CoV-2 data (NCBI)
Prevention and treatment information (HHS)
Español

Access
Controlled (826,725)
Public (13,166,503)

Source
DNA (9,361,763)
RNA (4,356,464)

Type
exome (384,570)
genome (3,334,850)

Library Layout
paired (10,041,073)
single (4,042,366)

Platform
ABI SOLID (31,971)
BGISEQ (51,967)
Capillary (340,810)
Complete Genomics (5,340)
Helicos (4,484)
Illumina (12,670,412)
Ion Torrent (197,282)
LS454 (344,104)
Oxford Nanopore (169,412)
PacBio SMRT (267,657)

Strategy
EpiGenomics (231,065)
Exome (1,286,737)
Genome (3,812,537)
RNASeq (17,551)
other (12,015,079)

Data in Cloud
GS (13,923,192)
S3 (13,863,624)

Summary
20 per page

Send to:

Filter your results:
All (17362969)
strategy\_mseq(17551)
strategy\_exome(1286737)
strategy\_genome(3812537)
strategy\_epigenomic(231065)
strategy\_other\_other(12015079)
type\_mseq(2795805)
abi\_solid(31971)
access: Controlled(826725)
access: Public(13166503)
aligned\_data(1768969)
cloud ftp ena(0)
cloud ftp ncbi(0)
cloud gs(13923192)
cloud s3(13863624)
filetype bai(51)
filetype bam(1544761)
filetype crai(165046)
filetype cram(1256610)
filetype fastq(9359829)
filetype fastq1(0)
filetype fastq2(0)
filetype illumina native(3464)
filetype illumina native qseq(0)
filetype oxfordnanopore nativ(0)

Search results
Items: 1 to 20 of 17362969

1.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773657

2.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773656

3.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773655

4.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773654

5.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773653

6.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773652

7.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773651

8.
This record has not yet been released. If you found reference to this accession in a publication, please let us know
Accession: SRX12773650

# SRA (the world's largest sequence registry by far)

<p>NCBI Resources</p> <p>SRA</p>	
<p><b>COVID-19</b></p> <p>Public health</p>	
<p><b>Access</b></p> <p>Controlled (826,725)</p> <p>Public (13,166,503)</p>	<p><b>Access</b></p> <p>Controlled (826,725)</p> <p>Public (13,166,503)</p>
<p><b>Source</b></p> <p>DNA (9,361,763)</p> <p>RNA (4,356,464)</p>	<p><b>Source</b></p> <p>DNA (9,361,763)</p> <p>RNA (4,356,464)</p>
<p><b>Type</b></p> <p>exome (384,570)</p> <p>genome (3,334,850)</p>	<p><b>Type</b></p> <p>exome (384,570)</p> <p>genome (3,334,850)</p>
<p><b>Library Layout</b></p> <p>paired (10,041,073)</p> <p>single (4,042,366)</p>	<p><b>Library Layout</b></p> <p>paired (10,041,073)</p> <p>single (4,042,366)</p>
<p><b>Platform</b></p> <p>ABI SOLiD (31,971)</p> <p>BGISeq (51,967)</p> <p>Capillary (340,810)</p> <p>Complete Genomics (5,340)</p> <p>Helicos (4,484)</p> <p>Illumina (12,670,412)</p> <p>Ion Torrent (197,282)</p> <p>LS454 (344,104)</p> <p>Oxford Nanopore (169,412)</p> <p>PacBio SMRT (267,657)</p>	<p><b>Platform</b></p> <p>ABI SOLiD (31,971)</p> <p>BGISeq (51,967)</p> <p>Capillary (340,810)</p> <p>Complete Genomics (5,340)</p> <p>Helicos (4,484)</p> <p>Illumina (12,670,412)</p> <p>Ion Torrent (197,282)</p> <p>LS454 (344,104)</p> <p>Oxford Nanopore (169,412)</p> <p>PacBio SMRT (267,657)</p>
<p><b>Strategy</b></p> <p>EpiGenomics (231,065)</p> <p>Exome (1,286,737)</p> <p>Genome (3,812,537)</p> <p>RNASeq (17,551)</p> <p>other (12,015,079)</p>	<p><b>Strategy</b></p> <p>EpiGenomics (231,065)</p> <p>Exome (1,286,737)</p> <p>Genome (3,812,537)</p> <p>RNASeq (17,551)</p> <p>other (12,015,079)</p>
<p><b>Data in Cloud</b></p> <p>GS (13,923,192)</p> <p>S3 (13,863,624)</p>	<p><b>Data in Cloud</b></p> <p>GS (13,923,192)</p> <p>S3 (13,863,624)</p>

<p>adeslat My NCBI</p> <p>Search</p>	
<p><b>Controlled Access: 826,725 SRR numbers</b></p> <p><b>Public: 13,166,503</b></p>	
<p><b>Platforms: 10 platforms including Oxford Nanopore and PacBio SMRT</b></p>	
<p><b>Data in the Cloud:</b></p> <p><b>GS: 13,923,192</b></p> <p><b>S3: 13,863,624</b></p>	
<p>Send to: Filter your results:</p> <p>All (17362969)</p>	<p>strategy_mseq (17551)</p> <p>strategy_mseq (3812537)</p> <p>strategy_mseq (231065)</p> <p>strategy_other_other (12015079)</p> <p>type_mseq (2795805)</p> <p>abi_solid (31971)</p> <p>access: Controlled (826725)</p> <p>access: Public (13166503)</p> <p>aligned_data (1768969)</p> <p>cloud ftp ena (0)</p> <p>cloud ftp ncbi (0)</p> <p>cloud_gs (13923192)</p> <p>cloud_s3 (13863624)</p> <p>filetype bai (51)</p> <p>filetype bam (1544761)</p> <p>filetype cram (165046)</p> <p>filetype cram (1256610)</p> <p>filetype fastq (9359829)</p> <p>filetype fastq1 (0)</p> <p>filetype fastq2 (0)</p> <p>filetype illumina_native (3464)</p> <p>filetype illumina_native_qseq (0)</p> <p>filetype oxfordnanopore_nativ (0)</p>

# NCPI NLM/NCBI Fact Sheet

## NLM|NCBI

### dbGaP Data

Studies	1,865
Subjects	~2.9 Million
Samples	~3.4 Million
Phenotype: Variables	370,825
Values	~2.5 Billion
Study Documents	7,120
Association Analyses	7,883
Genotype Assays (array)	~2 Million
Genotype Assays (imputed)	543,137
Genotype Assays (seq derived)	399,269
Sequence (WGS SRA)	178,288
Sequence (WXS SRA)	271,447
Sequence (RNAseq SRA)	86,879
Epigenomic (SRA)	~35,000

See: [dbGaP Summary Stats](#); numbers change daily

### SRA Data

#### Public Sequence Data (Number of Records)

Data Format	Google Cloud Platform			Amazon Web Services		
	Public Dataset (Hot)	Commercial (Hot)	Commercial (Cold)	Open Data Program (Hot)	Commercial (Hot)	Commercial (Cold)
Source	0	0	14.2M	0	0	14.2M
SRA Normalized	775,619	7.5M	5.9M	13.4M	825,126	0
SRA Lite	0	8.0M	0	0	0	0

#### Controlled-Access Sequence Data (Number of Records)

Source	0	0	749,915	0	0	749,801
SRA Normalized	0	608,671	141,244	0	705,101	44,700
SRA Lite	0	1,751	0	0	0	0

# SRA “orofacial cleft”

Resources
How To
Sign in to NCBI

SRA

SRA
orofacial cleft
Search

Create alert
Search SRA

**COVID-19 Information**

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Access
Controlled (4,567)
Source
DNA (4,567)
Type
exome (2,333)
genome (2,234)
Library Layout
paired (2,333)
single (2,234)
Platform
Illumina (4,567)
Strategy
Exome (2,333)
Genome (2,234)
Data in Cloud
GS (3,450)
S3 (4,567)
File Type
bam (3,131)
crai (1,116)
cram (1,365)
Other
aligned data (3,450)

Summary
20 per page
Send to

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

**Search results**

Items: 1 to 20 of 4567

<< First
Prev
Page 1 of 229
Next >
Last >>

☐ [Exome Sequencing in Families with Overt and Subclinical Orofacial Cleft Phenotypes sample 18093](#)

1 ILLUMINA (Illumina NovaSeq 6000) run: 102.9M spots, 25.7G bases, 2.4Gb downloads  
Accession: SRX8419098

☐ [Exome Sequencing in Families with Overt and Subclinical Orofacial Cleft Phenotypes sample 18046](#)

2. 1 ILLUMINA (Illumina NovaSeq 6000) run: 109.5M spots, 27.4G bases, 2.4Gb downloads  
Accession: SRX8419097

☐ [Exome Sequencing in Families with Overt and Subclinical Orofacial Cleft Phenotypes sample 18363](#)

3. 1 ILLUMINA (Illumina NovaSeq 6000) run: 113.5M spots, 28.4G bases, 2.5Gb downloads  
Accession: SRX8419096

☐ [Exome Sequencing in Families with Overt and Subclinical Orofacial Cleft Phenotypes sample 18235](#)

4. 1 ILLUMINA (Illumina NovaSeq 6000) run: 99.6M spots, 24.9G bases, 2.6Gb downloads  
Accession: SRX8419095

☐ [Exome Sequencing in Families with Overt and Subclinical Orofacial Cleft Phenotypes sample 18364](#)

5. 1 ILLUMINA (Illumina NovaSeq 6000) run: 132.5M spots, 33.1G bases, 3Gb downloads  
Accession: SRX8419094

Filters: [Manage Filters](#)

Search in related databases

Database	Access		all
	public	controlled	
BioSample			<a href="#">3,233</a>
BioProject	<a href="#">1</a>	<a href="#">2</a>	<a href="#">3</a>
dbGaP		<a href="#">4</a>	<a href="#">4</a>
GEO Datasets	<a href="#">1</a>		<a href="#">1</a>

Find related data

Database: Select

Find items

Search details

orofacial[All Fields] AND cleft[All Fields]

Search

See more...

Recent activity

Turn Off Clear

orofacial cleft (4567)

Clear all

Show additional filters

# SRA “orofacial cleft”

Resources
How To
Sign in to NCBI

SRA
SRA
orofacial cleft
Search
Create alert
Search SRA
Help

COVID-19 Information

Public health information (CDC)
Research information

Access
Controlled (4,567)
Source
DNA (4,567)
Type
exome (2,333)
genome (2,234)
Library Layout
paired (2,333)
single (2,234)
Platform
Illumina (4,567)
Strategy
Exome (2,333)
Genome (2,234)
Data in Cloud
GS (3,450)
S3 (4,567)
File Type
bam (3,131)
crai (1,116)
cram (1,365)
Other
aligned data (3,450)

Summary - 20 per page -
View results as an expanded i

Search results
Items: 1 to 20 of 4567

☐ Exome Sequencing in Fa
1. 18093
1 ILLUMINA (Illumina NovaSe
Accession: SRX8419098

☐ Exome Sequencing in Fa
2. 18046
1 ILLUMINA (Illumina NovaSe
Accession: SRX8419097

☐ Exome Sequencing in Fa
3. 18363
1 ILLUMINA (Illumina NovaSe
Accession: SRX8419095

☐ Exome Sequencing in Fa
4. 18235
1 ILLUMINA (Illumina NovaSe
Accession: SRX8419095

☐ Exome Sequencing in Families with Overt and Subclinical Orofacial Cleft Phenotypes sample
5. 18364
1 ILLUMINA (Illumina NovaSeq 6000) run: 132.5M spots, 33.1G bases, 3Gb downloads
Accession: SRX8419094

Search in related databases

Database	Access		all
	public	controlled	
BioSample		3,233	3,233
BioProject	1	2	3
dbGaP		4	4
GEO Datasets	1		1

Recent activity
Turn Off
Clear

orofacial cleft (4567)



# dbGaP - click through

dbGaP

dbGaP

(orofacial cleft) AND gap\_sra\_all[filter]

Create alert

Limits

Advanced

Search

Help

COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Search results

Items: 4

Search results: 0 Variables, 0 Analyses, 0 Documents, and 0 Datasets in 4 Studies

Studies (4)

Variables (0)

Study Documents (0)

Analyses (0)

Datasets (0)

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
<a href="#">phs000625.v1.p1</a> Targeted Sequencing of GWAS Loci in Cleft Lip and Palate	Version 1: passed embargo	V D A S	6123	Parent-Offspring Trios, Multiplex Families, Twin	<a href="#">Links</a>	HiSeq 2000
<a href="#">phs001168.v2.p2</a> Kids First: Genomic Studies of Orofacial Cleft Birth Defects	Versions 1-2: passed embargo	V D A S	1378	Parent-Offspring Trios	<a href="#">Links</a>	HumanHap550v3.0 Assembler Version 1.2.0; File Format Version: July 2009 HiSeq X v2.5
<a href="#">phs001420.v1.p1</a> Kids First: Genomics of Orofacial Cleft Birth Defects in Latin American Families	Version 1: passed embargo	V D A S	828	Parent-Offspring Trios	<a href="#">Links</a>	HiSeq X Ten
<a href="#">phs001997.v1.p1</a> Kids First: Genomics of African and Asian Orofacial Clefts Triads	Version 1: passed embargo	V D A S	791	Cohort, Parent-Offspring Trios	<a href="#">Links</a>	

Filter your results:

All (4)

disease (0)

dbgap\_type\_studies (4)

Studies having SRA data (4)

SHARE project (0)

Manage Filters

Find related data

Database: Select

Find items

Recent activity

Turn Off Clear

(orofacial cleft) AND gap\_sra\_all[filter] AND 1[s\_discriminator] (4)

dbGaP

orofacial cleft (4567)

SRA

orofacial cleft (8691)

SRA

See more...

Session Started

# dbGaP - click through

dbGaP

dbGaP (orofacial)

Create a new study

COVID-19 Information

Public health information (CDC) | Research

Search results

Items: 4

Search results: 0 Variables, 0 Analyses, 0 Documents

Studies (4) Variables (0) Study Documents (0)

Study	Embargo Release	Details
<a href="#">phs000625.v1.p1 Targeted Sequencing of GWAS Loci in Cleft Lip and Palate</a>	Version 1: passed embargo	V D A S
<a href="#">phs001168.v2.p2 Kids First: Genomic Studies of Orofacial Cleft Birth Defects</a>	Versions 1-2: passed embargo	V D A S
<a href="#">phs001420.v1.p1 Kids First: Genomics of Orofacial Cleft Birth Defects in Latin American Families</a>	Version 1: passed embargo	V D A S
<a href="#">phs001997.v1.p1 Kids First: Genomics of African and Asian Orofacial Clefts Triads</a>	Version 1: passed embargo	V D A S

Search results: 0 Variables, 0 Analyses, 0 Documents, and 0 Datasets in 4 Studies

Studies (4) Variables (0) Study Documents (0) Analyses (0) Datasets (0)

Study	Embargo Release	Details	Participants	Type Of Study	Links	
<a href="#">phs000625.v1.p1 Targeted Sequencing of GWAS Loci in Cleft Lip and Palate</a>	Version 1: passed embargo	V D A S	6123	Parent-Offspring Trios, Multiplex Families, Twin	<a href="#">Links</a>	
<a href="#">phs001168.v2.p2 Kids First: Genomic Studies of Orofacial Cleft Birth Defects</a>	Versions 1-2: passed embargo	V D A S	1378	Parent-Offspring Trios	<a href="#">Links</a>	Ass
<a href="#">phs001420.v1.p1 Kids First: Genomics of Orofacial Cleft Birth Defects in Latin American Families</a>	Version 1: passed embargo	V D A S	828	Parent-Offspring Trios	<a href="#">Links</a>	
<a href="#">phs001997.v1.p1 Kids First: Genomics of African and Asian Orofacial Clefts Triads</a>	Version 1: passed embargo	V D A S	791	Cohort, Parent-Offspring Trios	<a href="#">Links</a>	

# dbGaP search “orofacial cleft” - 11 Studies

## dbGaP Advanced Search ?

"orofacial cleft"

"orofacial cleft"

Show All Filters

Study Disease/Focus (5)

Study Design (4)

Study Molecular Data Type (7)

Study Markers (8)

NIH Institute (3)

Study Consent (51)

Study Type (11)

Ancestry (9)

Study Subject Count

Studies (10)

Phenotype Datasets (18)

Variables (47)

Molecular Datasets (2)

Analyses (0)

Documents (1)

1/1

Save Results

Save Query

### 1 Kids First: Genomic Studies of Orofacial Cleft Birth Defects

Accession phs001168.v2.p2  
Study Disease/Focus Cleft Lip  
Study Design Family/Twin/Trios  
Study Markerset Not Provided  
Study Molecular Data Type WGS  
Study Content 5 phenotype datasets, 23 variables, SRA, 1378 subjects, 1311 samples  
NIH Institute NCI  
Study Consent DS-OBDR-MDS --- Disease-specific (orofacial birth defects and related phenotypes, mds), DS-OBDR-MDS --- Disease-specific (orofacial birth defects, mds), DS-OC-PUB-MDS --- Disease-specific (oral clefts, pub, mds), HMB-MDS --- Health/medical/biomedical (mds), GRU --- General research use  
Release Date 2019-04-05  
Embargo Release Date 2018-03-20  
Related Terms CL - Cleft lip; Cheiloschisis; Cheiloschisis of upper lip; Cleft Lips; Cleft of upper lip; Congenital fissure of lip ...

The Gabriella Miller Kids First Pediatric Research Program (Kids First) is a trans-NIH effort initiated in response to the 2014 Gabriella Miller Kids First Research Act and supported by the ... where other Kids First datasets can also be accessed. The Kids First study of nonsyndromic orofacial cleft birth defects (OFCs) is a whole genome...

[FileSelector](#) [RunSelector](#) [PubMed](#) [PMC](#) [MeSH](#) [BioProject](#) [BioSample](#) [SRA](#)

### 2 Kids First: Genomics of Orofacial Cleft Birth Defects in Latin American Families

Accession phs001420.v1.p1  
Study Disease/Focus Cleft Lip  
Study Design Family/Twin/Trios  
Study Markerset Not Provided  
Study Molecular Data Type Not Provided  
Study Content 2 phenotype datasets, 7 variables, SRA, 828 subjects, 828 samples  
NIH Institute NCI  
Study Consent DS-OBDR-RD --- Disease-specific (orofacial birth defects and related phenotypes, rd)  
Release Date 2019-06-14  
Embargo Release Date 2018-03-14  
Related Terms CL - Cleft lip; Cheiloschisis; Cheiloschisis of upper lip; Cleft Lips; Cleft of upper lip; Congenital fissure of lip ...

The Gabriella Miller Kids First Pediatric Research Program (Kids First) is a trans-NIH effort initiated in response to the 2014 Gabriella Miller Kids First Research Act and supported by the ... KidsFirst.org, where other Kids First datasets can also be accessed. The Kids First study of nonsyndromic orofacial cleft (OFC) birth defects in Latin American...

[FileSelector](#) [RunSelector](#) [PubMed](#) [PMC](#) [MeSH](#) [BioProject](#) [BioSample](#) [SRA](#)

### 3 CCGG: Exomes of Orofacial Clefts

Accession phs001675.v1.p1  
Study Disease/Focus Cleft Lip  
Study Design Family/Twin/Trios  
Study Markerset WES\_markerset\_grc37  
Study Molecular Data Type SNP/CNV Genotypes (NGS), WXS  
Study Content 5 phenotype datasets, 30 variables, 1 molecular datasets, SRA, 365 subjects, 319 samples  
NIH Institute NIDCR  
Study Consent DS-OBDR-MDS-RD --- Disease-specific (orofacial birth defects and related phenotypes, mds, rd), DS-OBDR-RD --- Disease-specific (orofacial birth defects and related phenotypes, rd), HMB-MDS --- Health/medical/biomedical (mds), DS-CFR --- Disease-specific (craniofacial research)  
Release Date 2021-04-19  
Embargo Release Date 2021-04-19  
Related Terms CL - Cleft lip; Cheiloschisis; Cheiloschisis of upper lip; Cleft Lips; Cleft of upper lip; Congenital fissure of lip ...

Orofacial Clefts (OFCs) are genetically complex structural birth defects caused by genetic factors, environmental exposures, and their interactions. OFCs are the most common craniofacial anomalies in humans. Affect about 1 ... of Pittsburgh, Pittsburgh, PA, USA Funding Sources: X01-HG010012, Exome sequencing in families with overt and subclinical orofacial cleft phenotypes...

[FileSelector](#) [RunSelector](#) [BioProject](#) [BioSample](#) [SRA](#)

# dbGaP search “orofacial cleft” - 11 Studies

The screenshot displays the dbGaP Advanced Search interface. At the top, the search query "orofacial cleft" is entered in the search bar. Below the search bar, a "Show All Filters" button is visible. The main results area shows a list of filters with their corresponding counts:

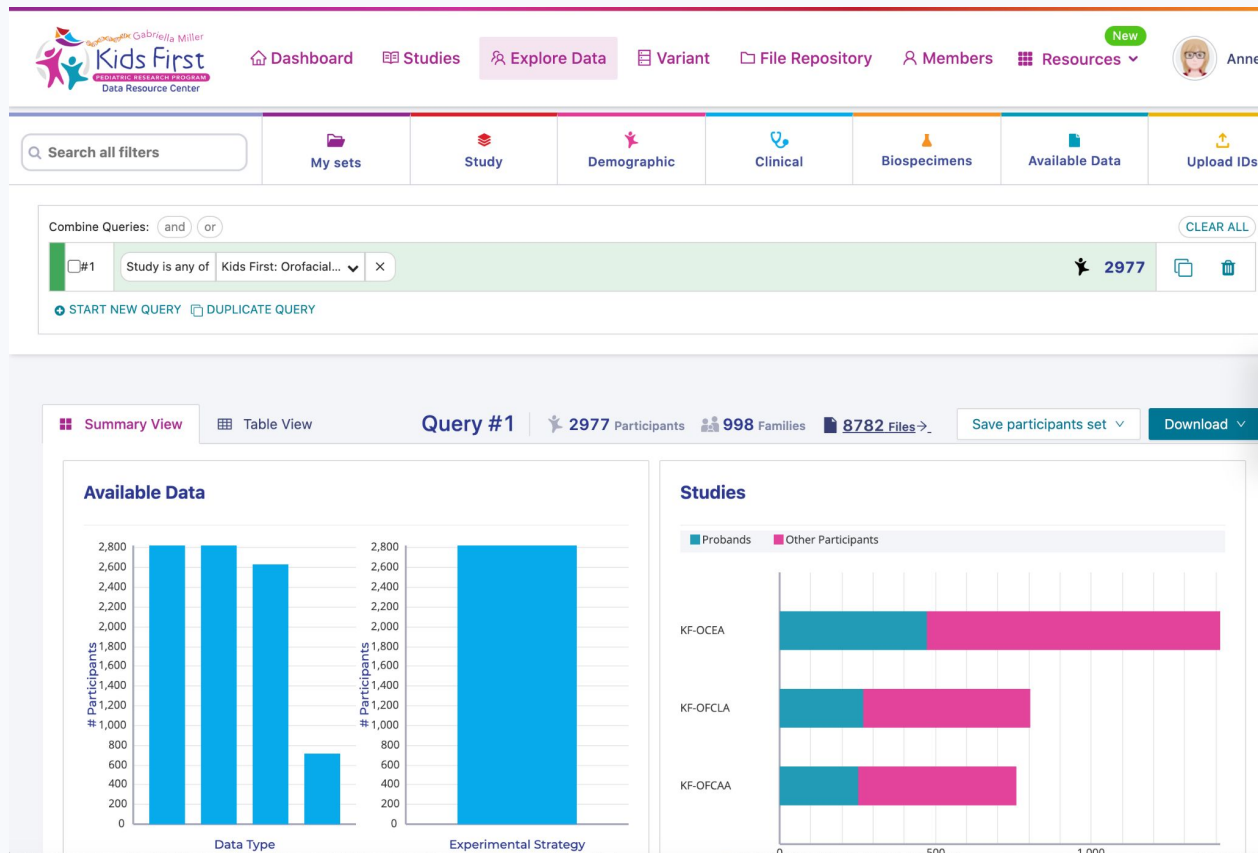
- Study Disease/Focus (5)
- Study Design (4)
- Study Molecular Data Type (7)
- Study Markerset (8)
- NIH Institute (3)
- Study Consent (51)
- Study Type (11)
- Ancestry (9)
- Study Subject Count

On the left side of the interface, there is a sidebar with additional filters and a "Show All Filters" button. The sidebar includes the following filters:

- Study Disease/Focus (5)
- Study Design (4)
- Study Molecular Data Type
- Study Markerset (8)
- NIH Institute (3)
- Study Consent (51)
- Study Type (11)
- Ancestry (9)
- Study Subject Count

- **5 Disease/Focus**
- **4 Study Designs**
- **7 Molecular Data Types**
- **8 Study Markersets**
- **3 NIH Institutes**
- **9 Ancestry**

# KidsFirst Data Portal - 3 Studies



## NICHD-ODSS and NCBI Uniform Search Use Case Goal

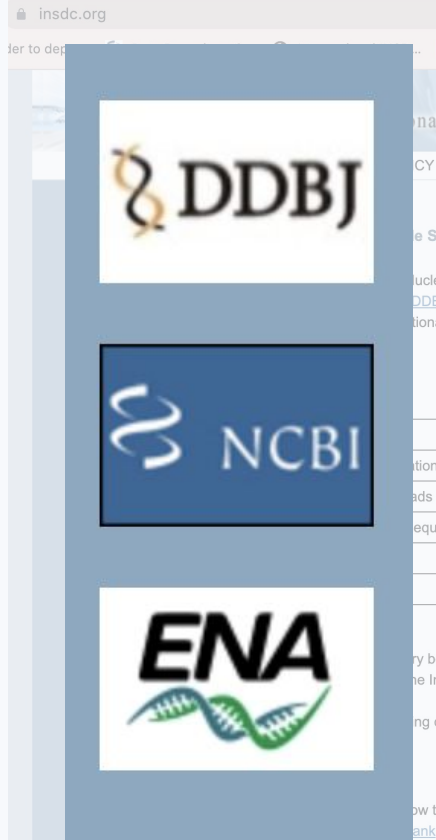
---

**Goal:** The goal is consistency in results regardless of the search portal.

**What should be returned:** Search should return consistent results regardless of search origin (search results should be similar in terms of number of results and dataset details regardless of point of search entry)

## **BACK UP SLIDES**

SRA (the world's largest sequence registry by far - coordinated internationally with Japan and EBI)



• The INSDC Feature Ta

Like NCBI, DDBJ Center (Japan's equivalent to NCBI) and ENA (Europe's equivalent to NCBI) is officially certified to collect nucleotide sequences from researchers and to issue the internationally recognized accession number to data submitters.

**The accession number issued for each sequence data is unique on the database and internationally recognized to guarantee the submitter the property of the submitted and published data.**

Since DDBJ Center exchanges the released data with ENA/EBI and NCBI on a daily basis, the three data centers share virtually the same data at any given time. The virtually unified database is called INSD; **International Nucleotide Sequence Database.**



Accession Definition: <https://www.ncbi.nlm.nih.gov/books/NBK470040/>

### **accession number**

The accession number is a **unique identifier** assigned to a record in sequence databases such as GenBank. Several NCBI databases use the format **[alphabetical prefix][series of digits]**. A change in the record in some databases (e.g. GenBank) is tracked by an integer extension of the accession number, an Accession.version identifier. The initial version of a sequence has the extension “.1”. When a change is made to a sequence in a GenBank record, the version extension of the Accession.version identifier is incremented. For the sequence NM\_000245.3, “.3” indicates that the record has been updated twice. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the original Accession.version identifiers.