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

NICHD-ODSS

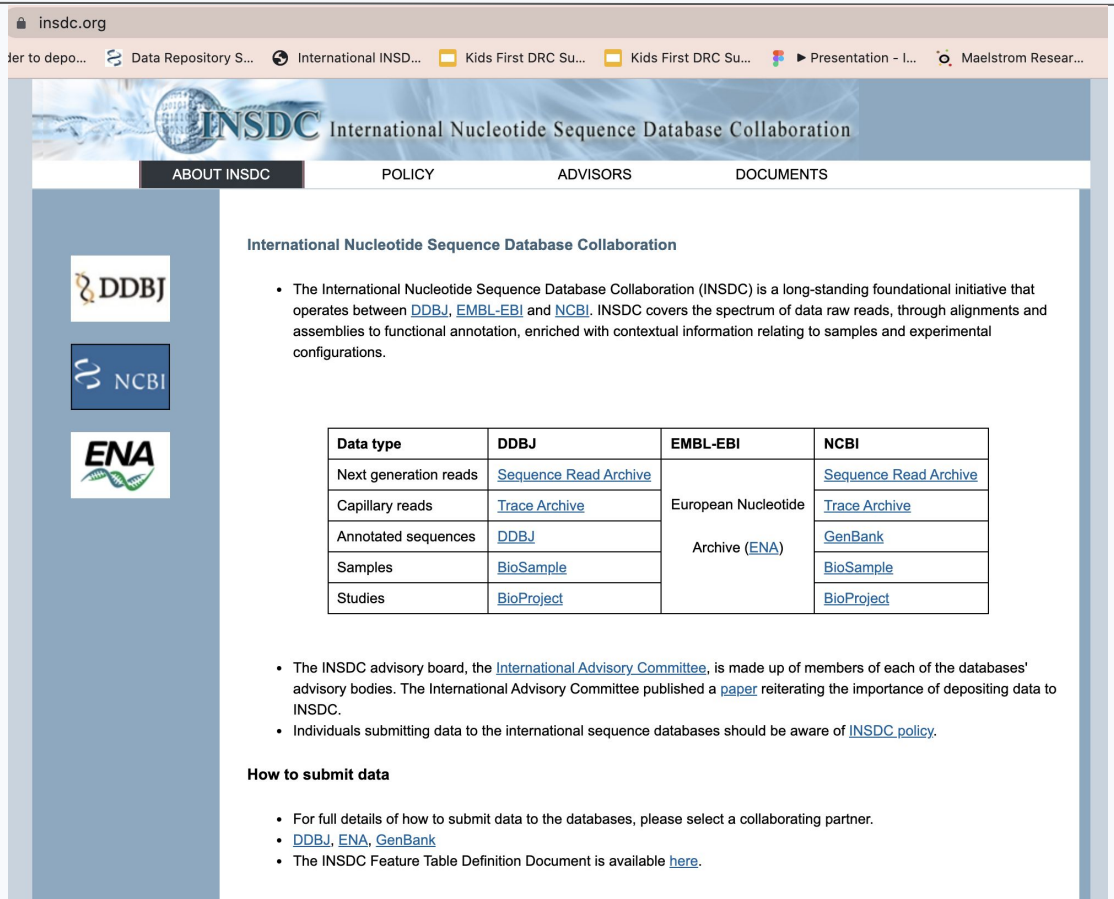
Title: Uniform search of public sample and sequence read information across NCBI and Kids First repositories. Search UC1 (GitHub NCPI Issue #19)

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”.

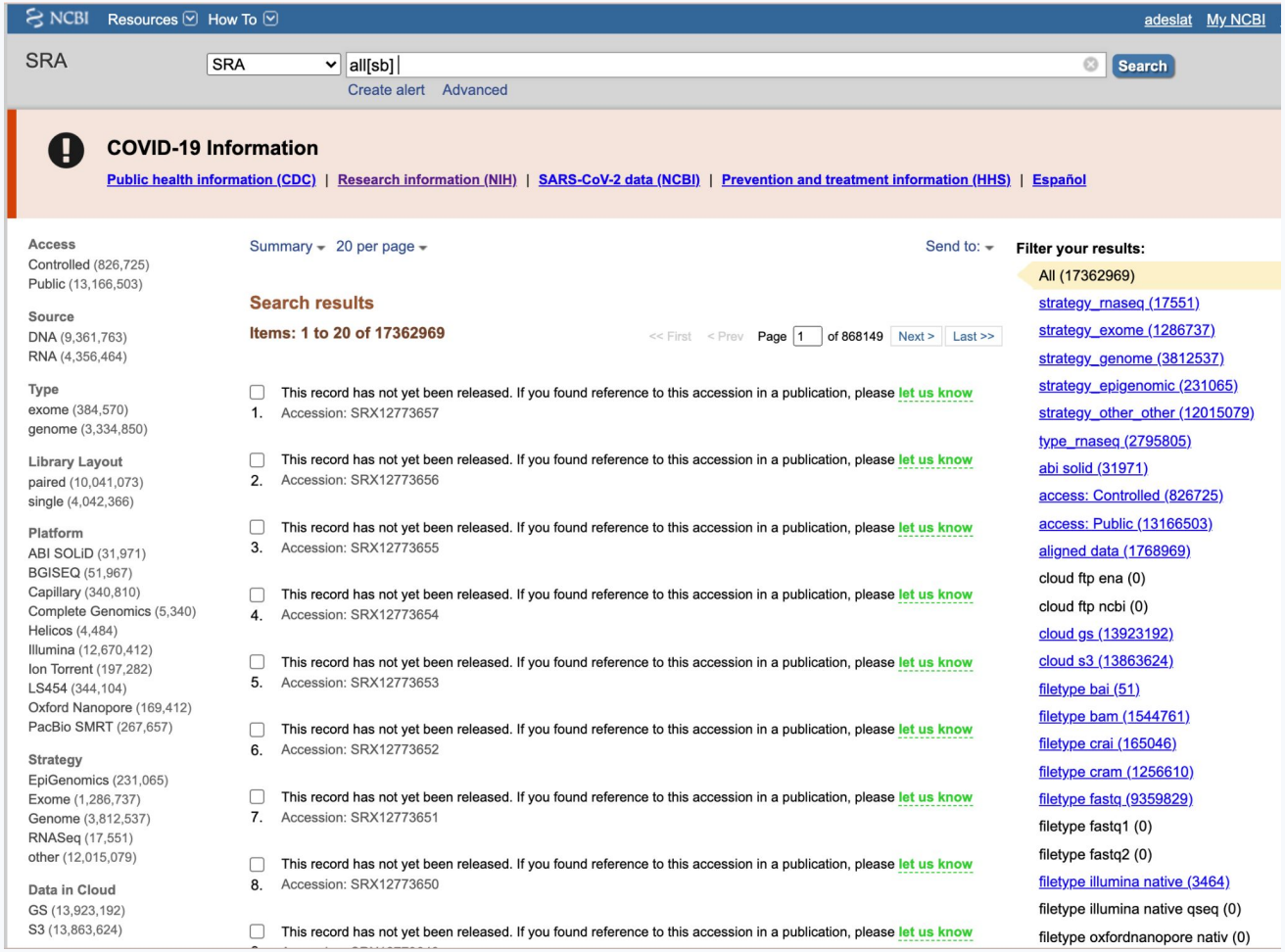
NICHD-ODSS

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



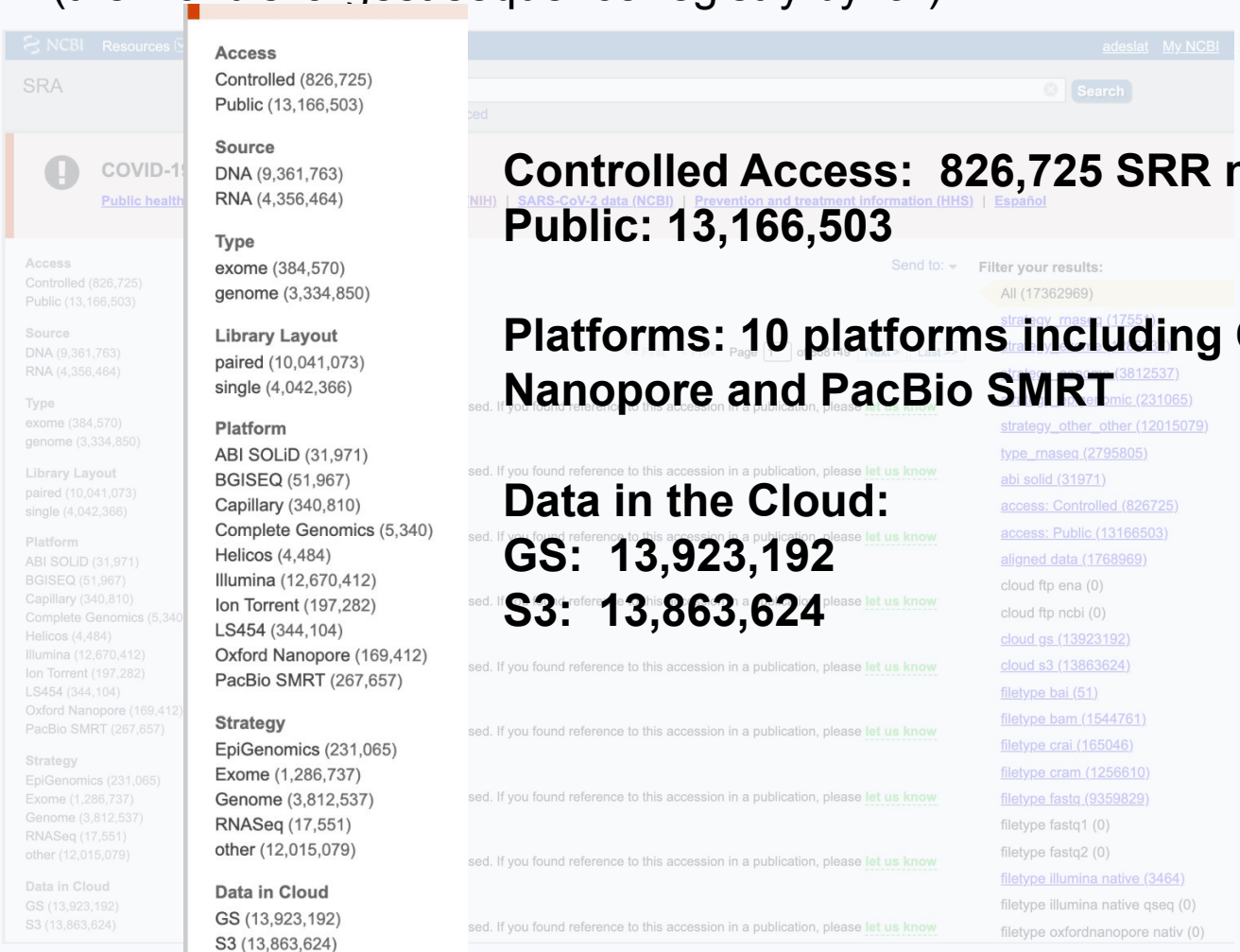
NICHD-ODSS

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



NICHD-ODSS

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



Controlled Access: 826,725 SRR numbers
Public: 13,166,503

Platforms: 10 platforms including Oxford Nanopore and PacBio SMRT

Data in the Cloud:
GS: 13,923,192
S3: 13,863,624

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

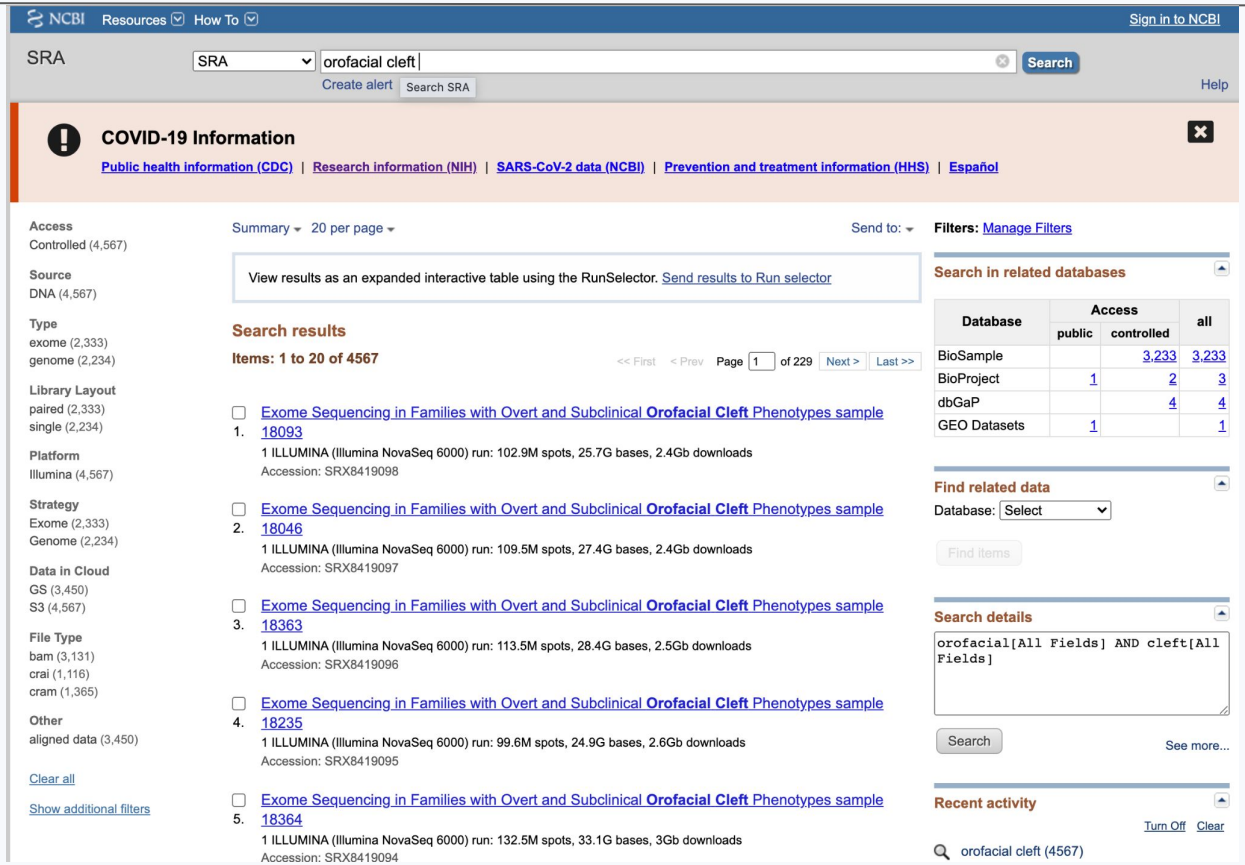
Controlled-Access Sequence Data (Number of Records)

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

Source: NCBI Mike Feolo

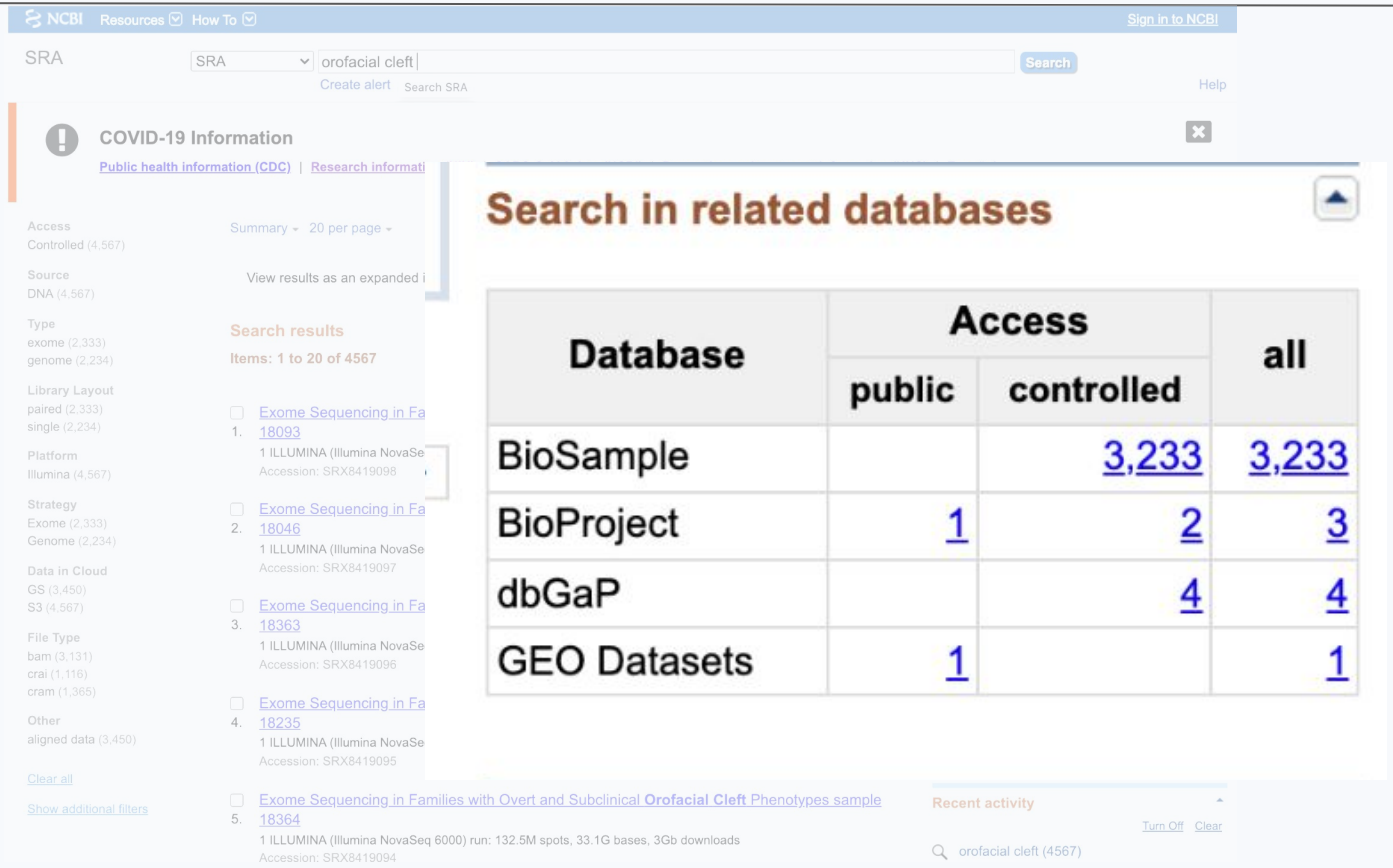
NICHD-ODSS

SRA “orofacial cleft”



NICHD-ODSS

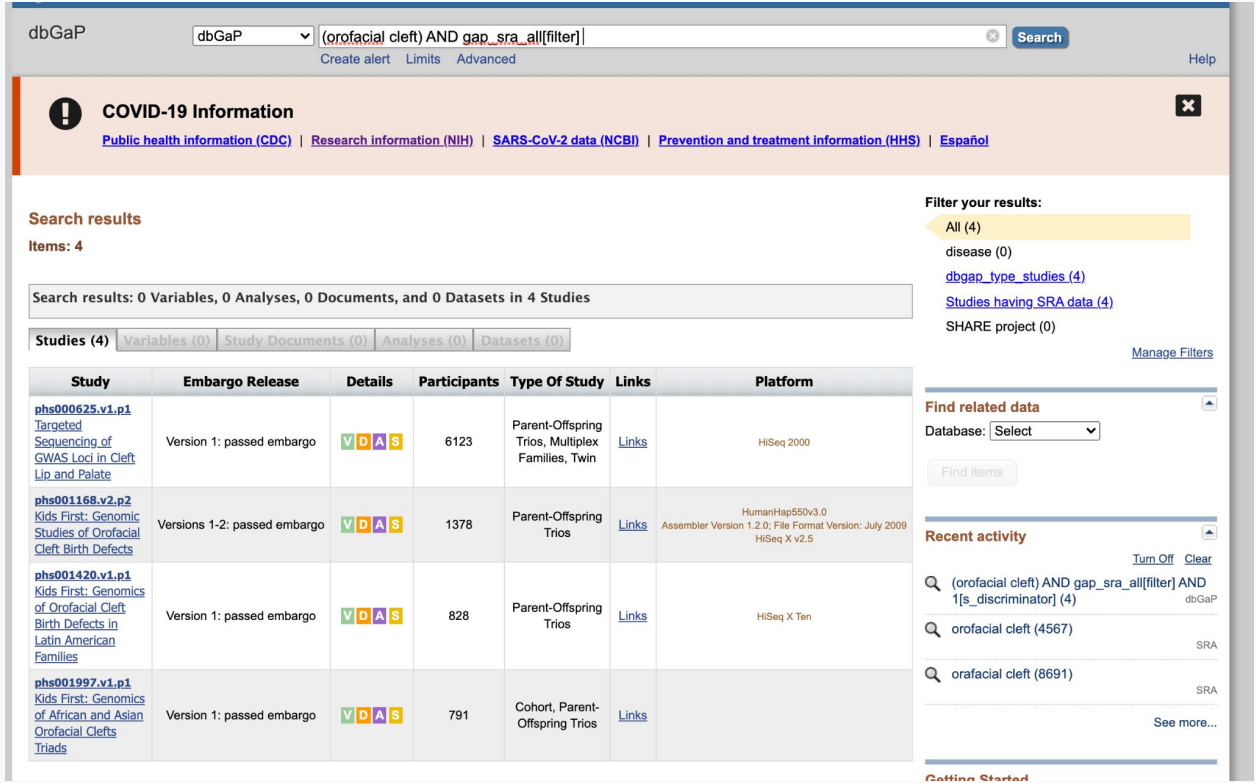
SRA “orofacial cleft”



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

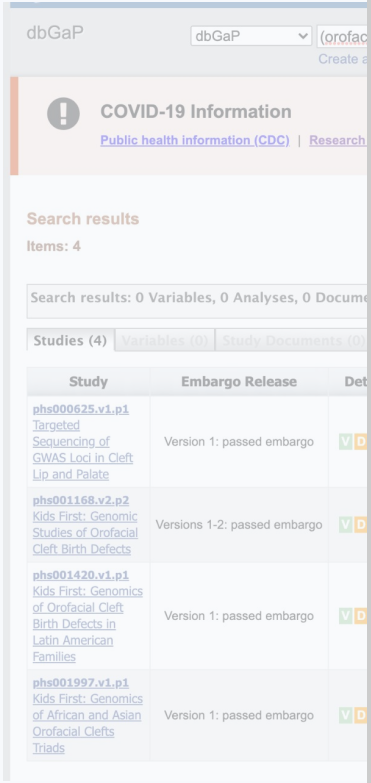
NICHD-ODSS

dbGaP - click through



NICHD-ODSS

dbGaP - click through



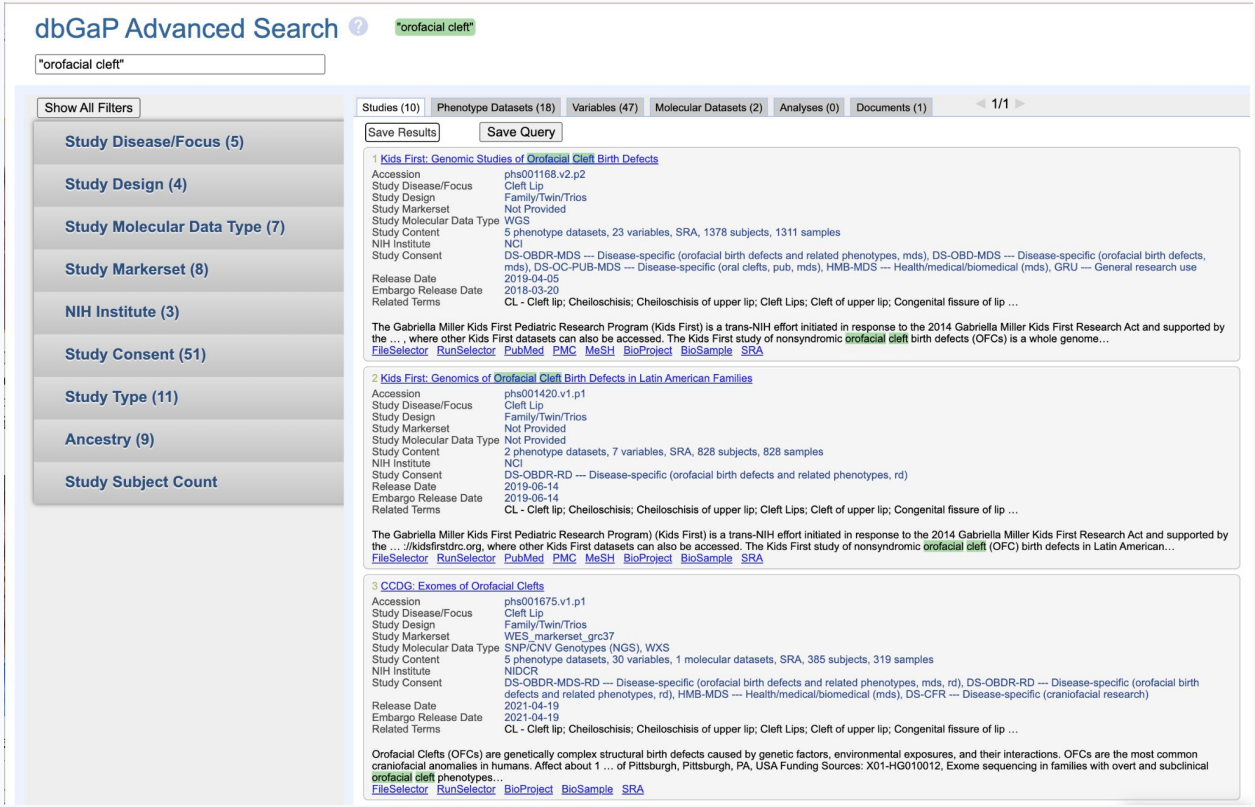
Search results: 0 Variables, 0 Analyses, 0 Documents, and 0 Datasets in 4 Studies

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

Study	Embargo Release	Details	Participants	Type Of Study	Links
phs000625.v1.p1 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	Links
phs001168.v2.p2 Kids First: Genomic Studies of Orofacial Cleft Birth Defects	Versions 1-2: passed embargo	V D A S	1378	Parent-Offspring Trios	Links
phs001420.v1.p1 Kids First: Genomics of Orofacial Cleft Birth Defects in Latin American Families	Version 1: passed embargo	V D A S	828	Parent-Offspring Trios	Links
phs001997.v1.p1 Kids First: Genomics of African and Asian Orofacial Clefts Triads	Version 1: passed embargo	V D A S	791	Cohort, Parent-Offspring Trios	Links

NICHD-ODSS

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 Markerset (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-OBD-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	2019-06-14
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 … ://kidsfirstdrc.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 CCDG: 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, 385 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

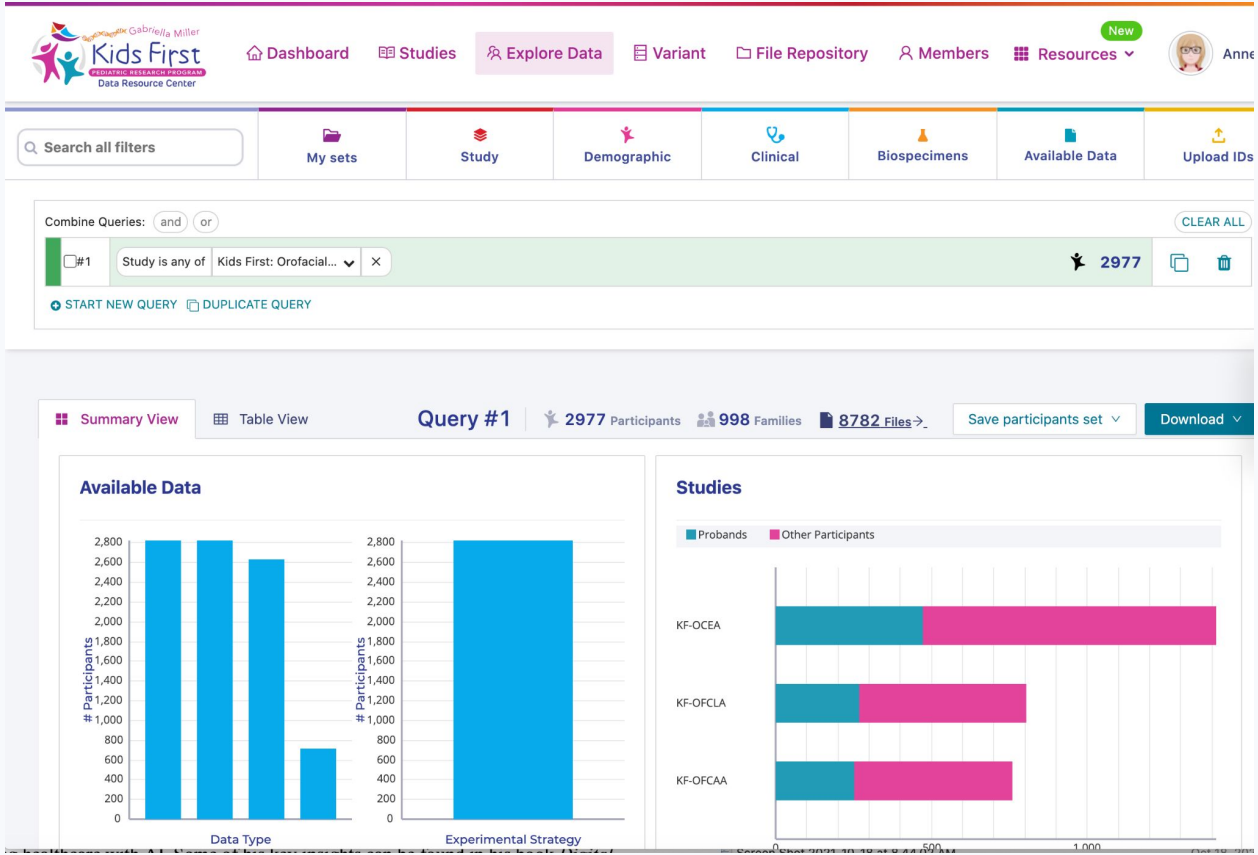
NICHD-ODSS

dbGaP search “orofacial cleft” - 11 Studies

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

NICHD-ODSS

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)

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.