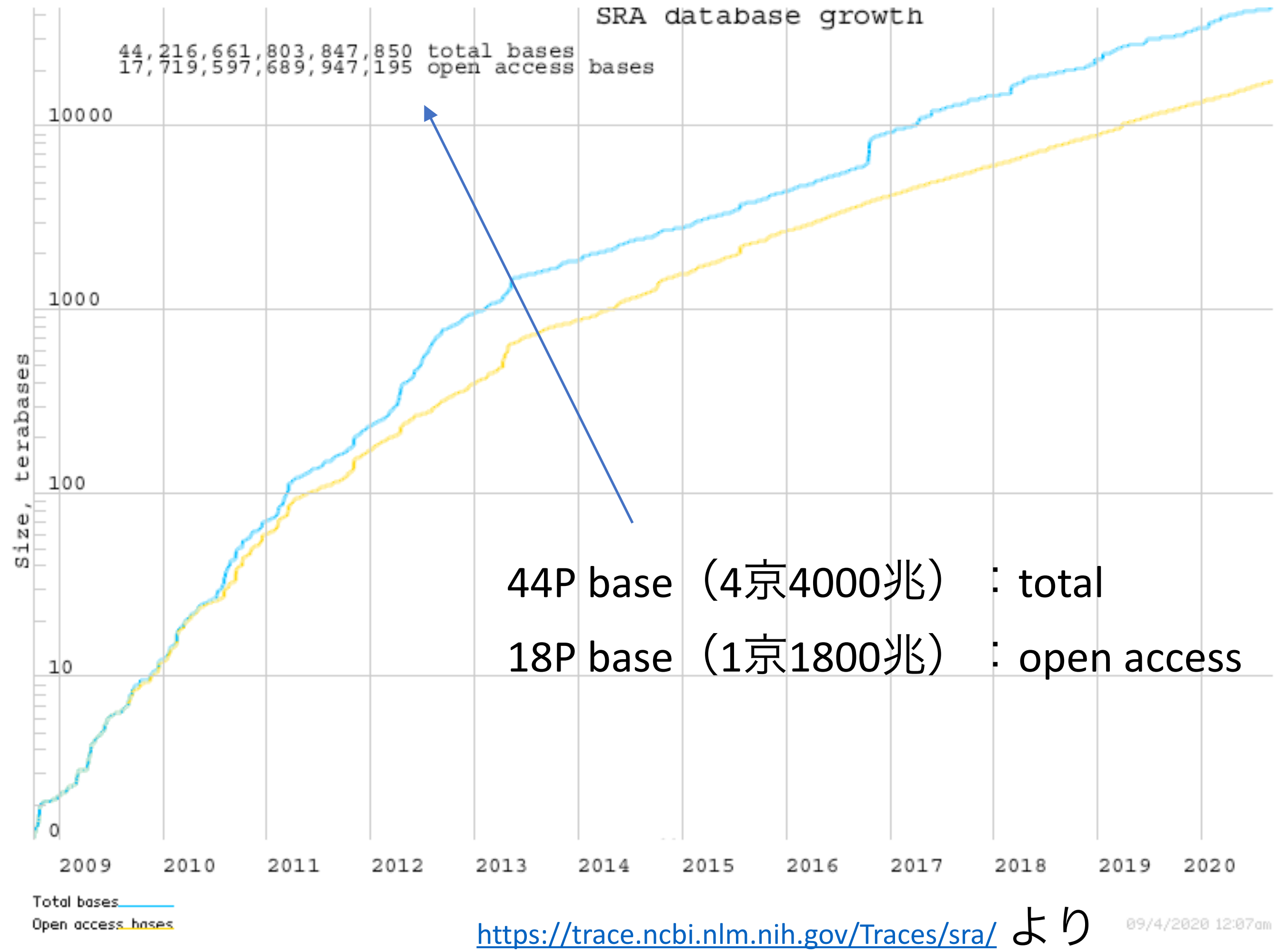
19

公共NGSデータとその関連データの検索 (2020年版)：SRA/BioProject/BioSample

○仲里　猛留、大田　達郎

情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター（DBCLS）

NGSデータの現状：登録塩基数



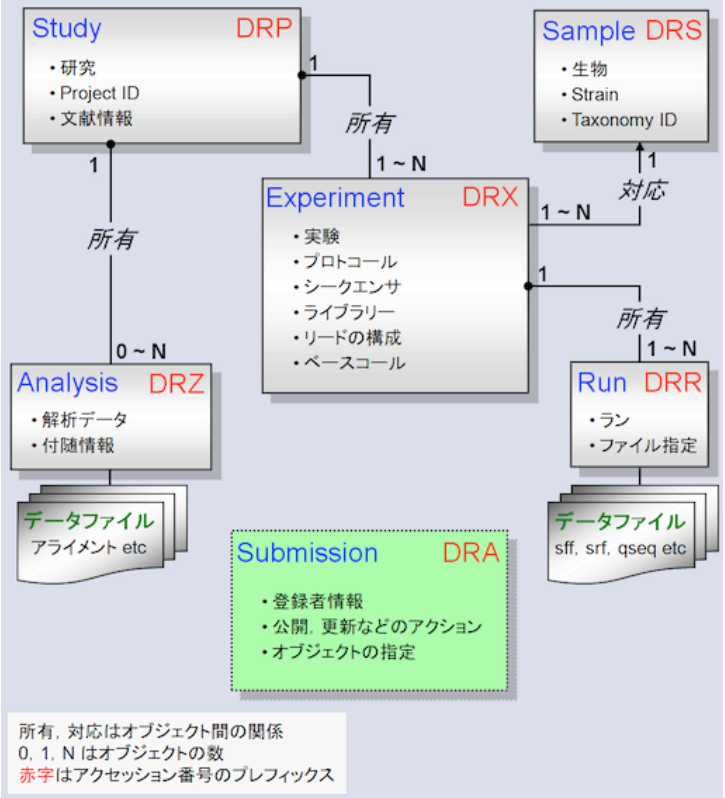
44P base（4京4000兆）：total

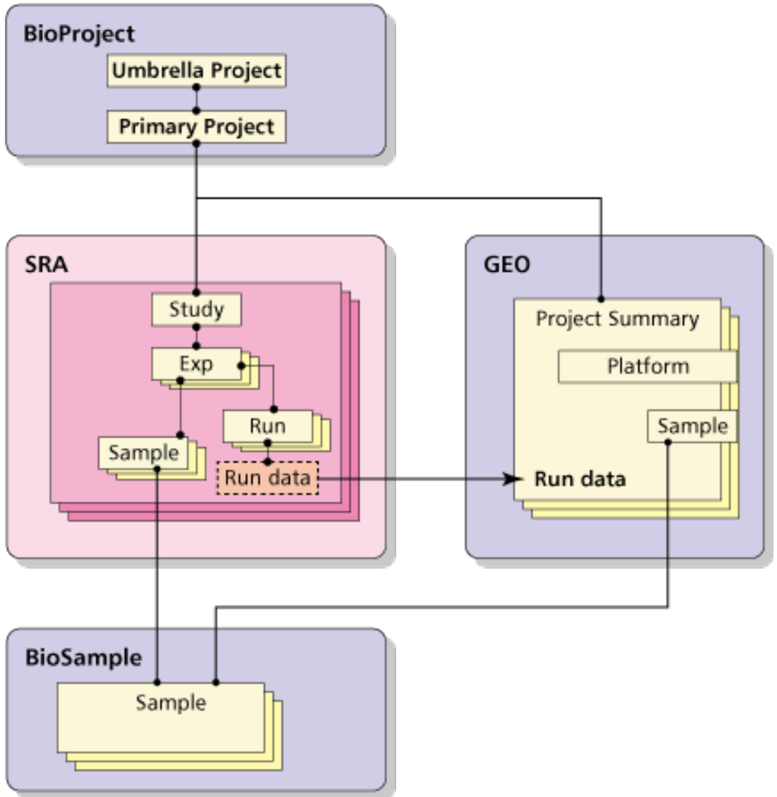
18P base（1京1800兆）：open access

https://trace.ncbi.nlm.nih.gov/Traces/sra/ より

2007年にNCBIがSRAを公開して公共NGSデータを収集し始めて以来、NGSの普及とともにデータ量も爆発的に増加した。

NGSデータにおけるメタデータの関係

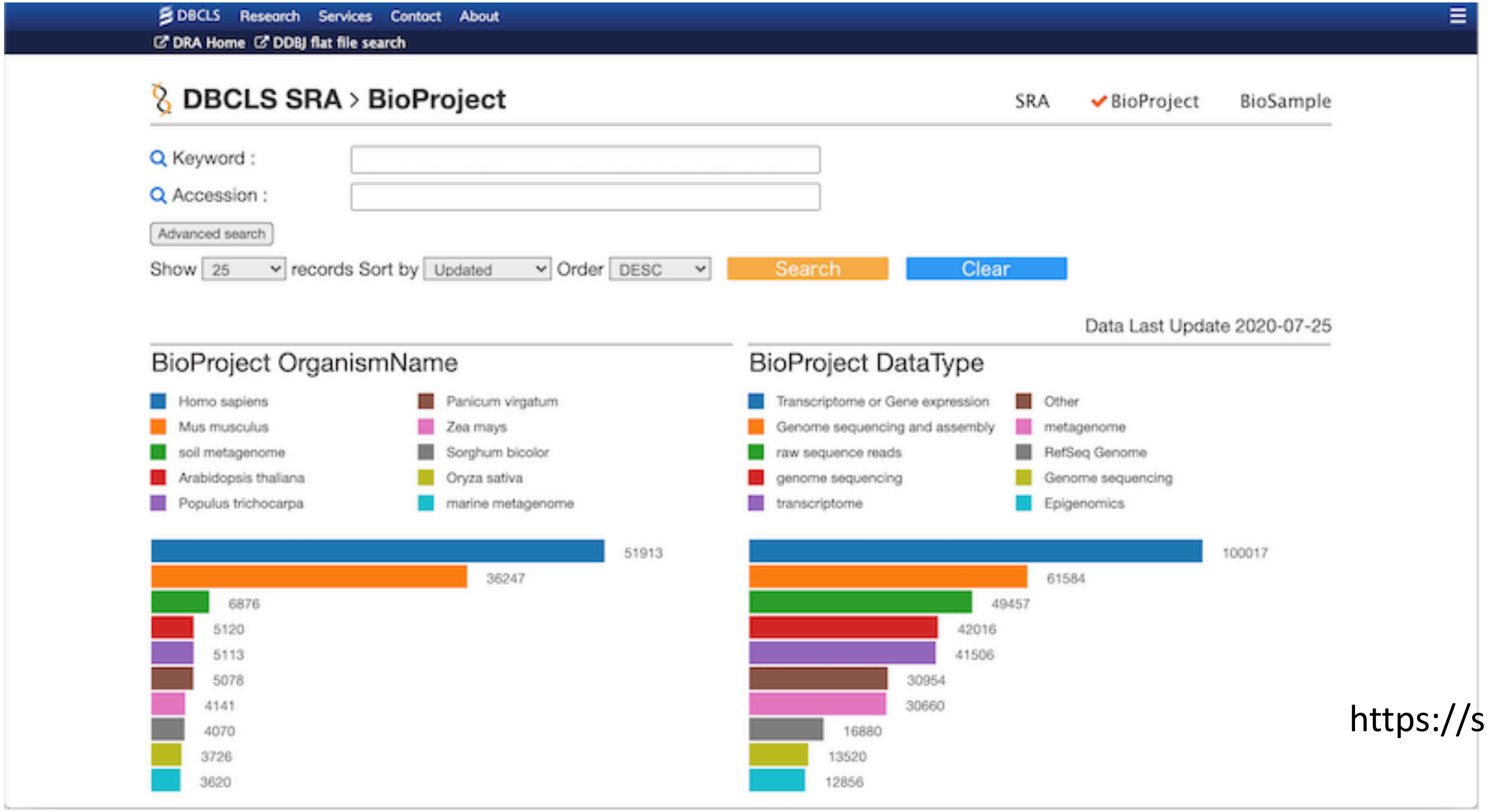




NGSデータはその目的の多様さからSRAだけでなく、GEO（transcriptome）やNucleotide（genomeなど）にも登録されるようになったため、BioProject/BioSampleのデータベースを作成し共通で参照できるようにした。

https://www.ddbj.nig.ac.jp/dra/submission.html を改

DBCLS SRA



DBCLS SRA > BioProject

Keyword :

Accession :

Show 25 records Sort by Updated Order DESC

Data Last Update 2020-07-25

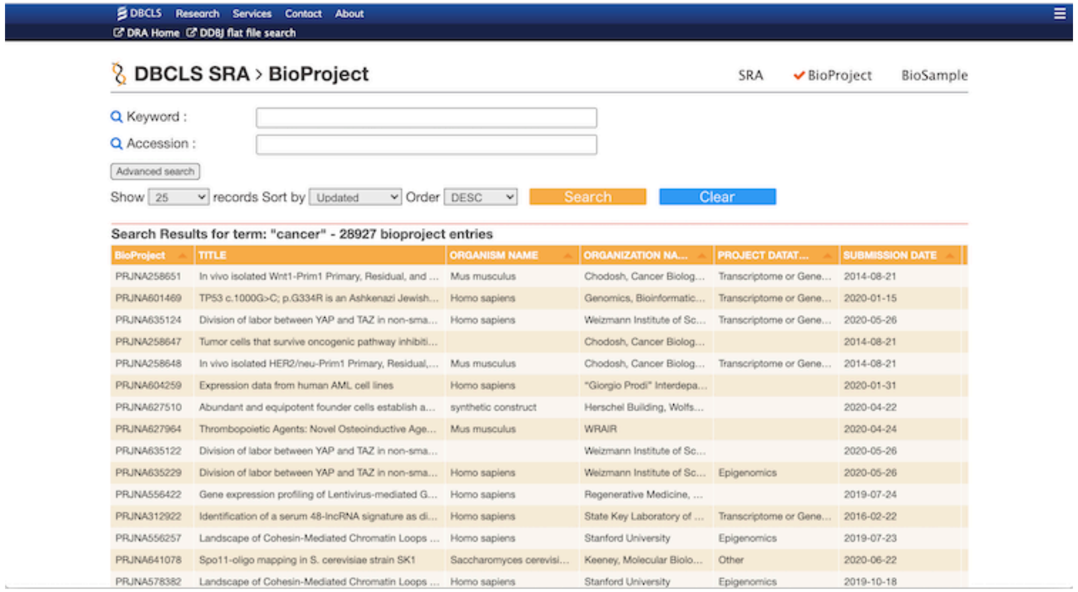
BioProject OrganismName

BioProject DataType

我々はこれまで
DBCLS SRAとしてSRA内の
公共NGSデータに対する
検索エンジンを提供して
きたが、今回 新たに
BioProjectやBioSampleの
データを検索できる
ようにした。

https://sra.dbcls.jp/

DBCLS SRA・リスト画面



DBCLS SRA > BioProject

Keyword :

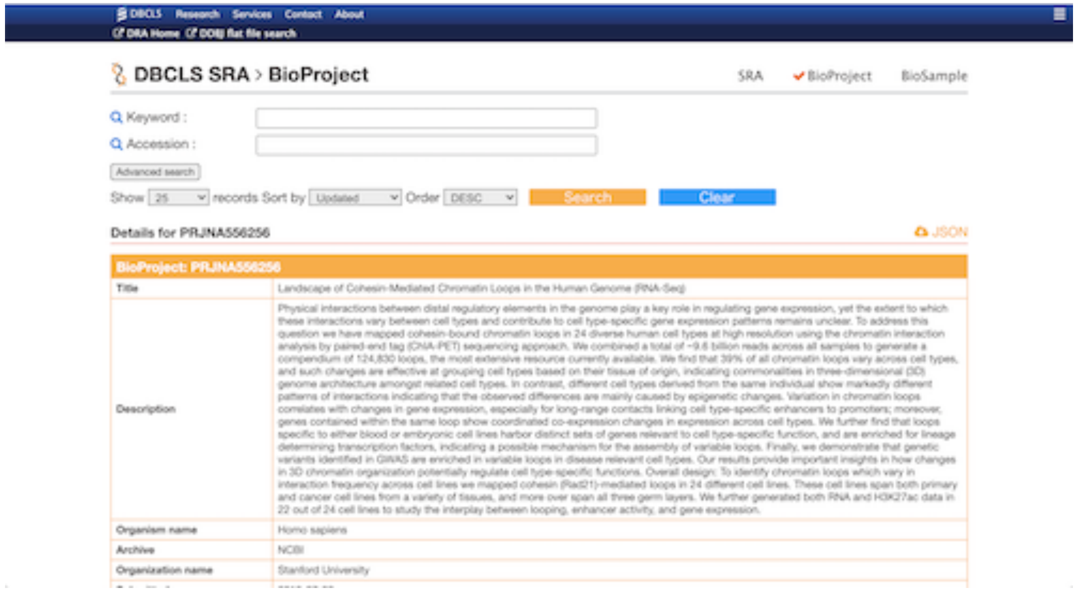
Accession :

Show 25 records Sort by Updated Order DESC

Search Results for term: "cancer" - 28927 bioproject entries

BioProject	TITLE	ORGANISM NAME	ORGANIZATION NA...	PROJECT DATAT...	SUBMISSION DATE
PRJNA258651	In vivo isolated Wnt1-Prim1 Primary, Residual, and ...	Mus musculus	Chodosh, Cancer Biolog...	Transcriptome or Gene...	2014-08-21
PRJNA601469	TP53 c.1000G>C; p.G334R is an Ashkenazi Jewish...	Homo sapiens	Genomics, Bioinformatic...	Transcriptome or Gene...	2020-01-15
PRJNA635124	Division of labor between YAP and TAZ in non-sma...	Homo sapiens	Weizmann Institute of Sc...	Transcriptome or Gene...	2020-05-26
PRJNA258647	Tumor cells that survive oncogenic pathway inhibiti...		Chodosh, Cancer Biolog...		2014-08-21
PRJNA258648	In vivo isolated HER2/neu-Prim1 Primary, Residual,...	Mus musculus	Chodosh, Cancer Biolog...	Transcriptome or Gene...	2014-08-21
PRJNA604259	Expression data from human AML cell lines	Homo sapiens	"Giorgio Prodi" Interdepa...		2020-01-31
PRJNA627510	Abundant and equipotent founder cells establish a...	synthetic construct	Herschel Building, Wolfs...		2020-04-22
PRJNA627964	Thrombopoietic Agents: Novel Osteoinductive Age...	Mus musculus	WRAIR		2020-04-24
PRJNA635122	Division of labor between YAP and TAZ in non-sma...		Weizmann Institute of Sc...		2020-05-26
PRJNA635229	Division of labor between YAP and TAZ in non-sma...	Homo sapiens	Weizmann Institute of Sc...	Epigenomics	2020-05-26
PRJNA556422	Gene expression profiling of Lentivirus-mediated G...	Homo sapiens	Regenerative Medicine, ...		2019-07-24
PRJNA312922	Identification of a serum 48-lncRNA signature as di...	Homo sapiens	State Key Laboratory of ...	Transcriptome or Gene...	2016-02-22
PRJNA556257	Landscape of Cohesin-Mediated Chromatin Loops ...	Homo sapiens	Stanford University	Epigenomics	2019-07-23
PRJNA641078	Spo11-oligo mapping in S. cerevisiae strain SK1	Saccharomyces cerevisi...	Keeney, Molecular Biolo...	Other	2020-06-22
PRJNA578382	Landscape of Cohesin-Mediated Chromatin Loops ...	Homo sapiens	Stanford University	Epigenomics	2019-10-18

リスト画面



DBCLS SRA > BioProject

Keyword :

Accession :

Show 25 records Sort by Updated Order DESC

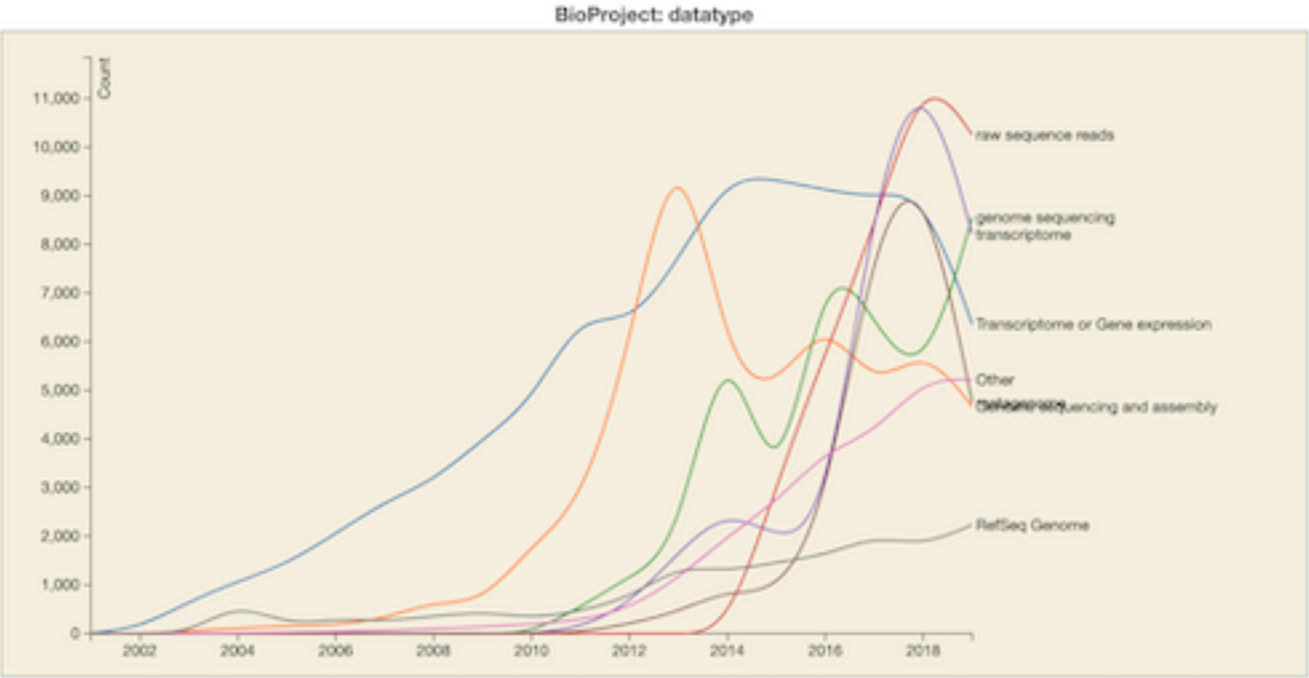
Details for PRJNA556256

BioProject: PRJNA556256	
Title	Landscape of Cohesin-Mediated Chromatin Loops in the Human Genome (RNA-Seq)
Description	Physical interactions between distal regulatory elements in the genome play a key role in regulating gene expression, yet the extent to which these interactions vary between cell types and contribute to cell type-specific gene expression patterns remains unclear. To address this question we have mapped cohesin-bound chromatin loops in 24 diverse human cell types at high resolution using the chromatin interaction analysis by paired-end tag (ChIA-PET) sequencing approach. We combined a total of ~9.8 billion reads across all samples to generate a compendium of 124,830 loops, the most extensive resource currently available. We find that 39% of all chromatin loops vary across cell types, and such changes are effective at grouping cell types based on their tissue of origin, indicating commonalities in three-dimensional (3D) genome architecture amongst related cell types. In contrast, different cell types derived from the same individual show markedly different patterns of interactions indicating that the observed differences are mainly caused by epigenetic changes. Variation in chromatin loops correlates with changes in gene expression, especially for long-range contacts linking cell type-specific enhancers to promoters; moreover, genes contained within the same loop show coordinated co-expression changes in expression across cell types. We further find that loops specific to either blood or embryonic cell lines harbor distinct sets of genes relevant to cell type-specific function, and are enriched for lineage determining transcription factors, indicating a possible mechanism for the assembly of variable loops. Finally, we demonstrate that genetic variants identified in GWAS are enriched in variable loops in disease relevant cell types. Our results provide important insights in how changes in 3D chromatin organization potentially regulate cell type-specific functions. Overall design: To identify chromatin loops which vary in interaction frequency across cell lines we mapped cohesin (Rad21)-mediated loops in 24 different cell lines. These cell lines span both primary and cancer cell lines from a variety of tissues, and more over span all three germ layers. We further generated both RNA and H3K27ac data in 22 out of 24 cell lines to study the interplay between looping, enhancer activity, and gene expression.
Organism name	Homo sapiens
Archive	NCBI
Organization name	Stanford University

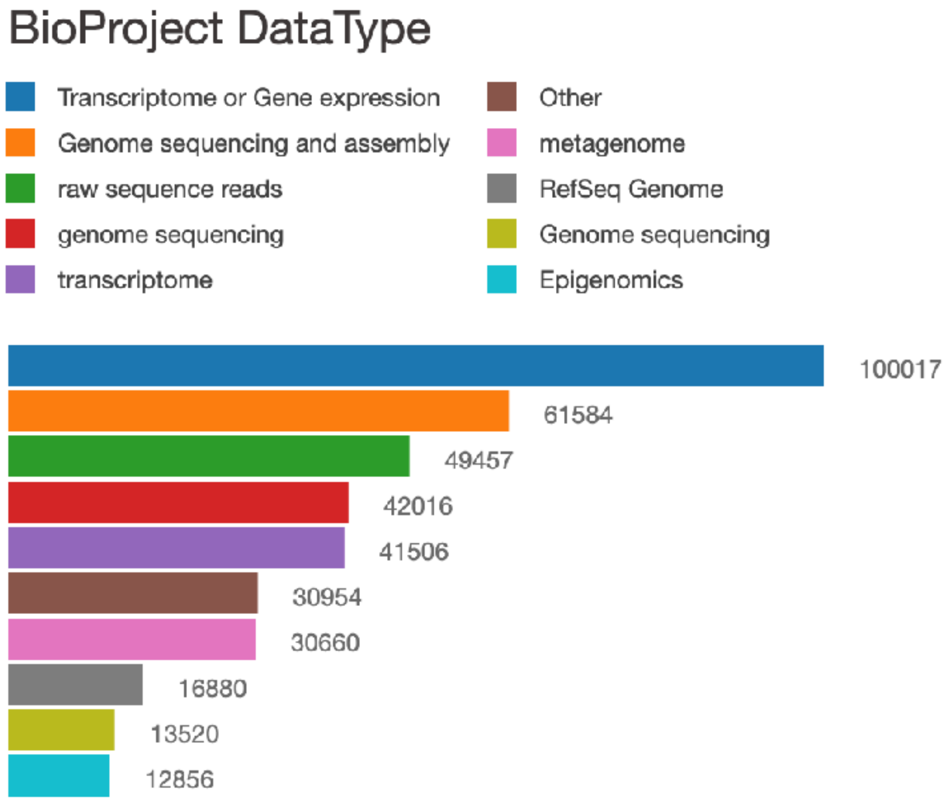
詳細画面

これまでのDBCLS SRAでの公共NGSデータ検索と同様にBioProjectデータを検索し、そのリスト表示や個々の登録の詳細の確認などを行うことができる。

NGSデータの現状：目的別



各年の登録数（累積ではない！）

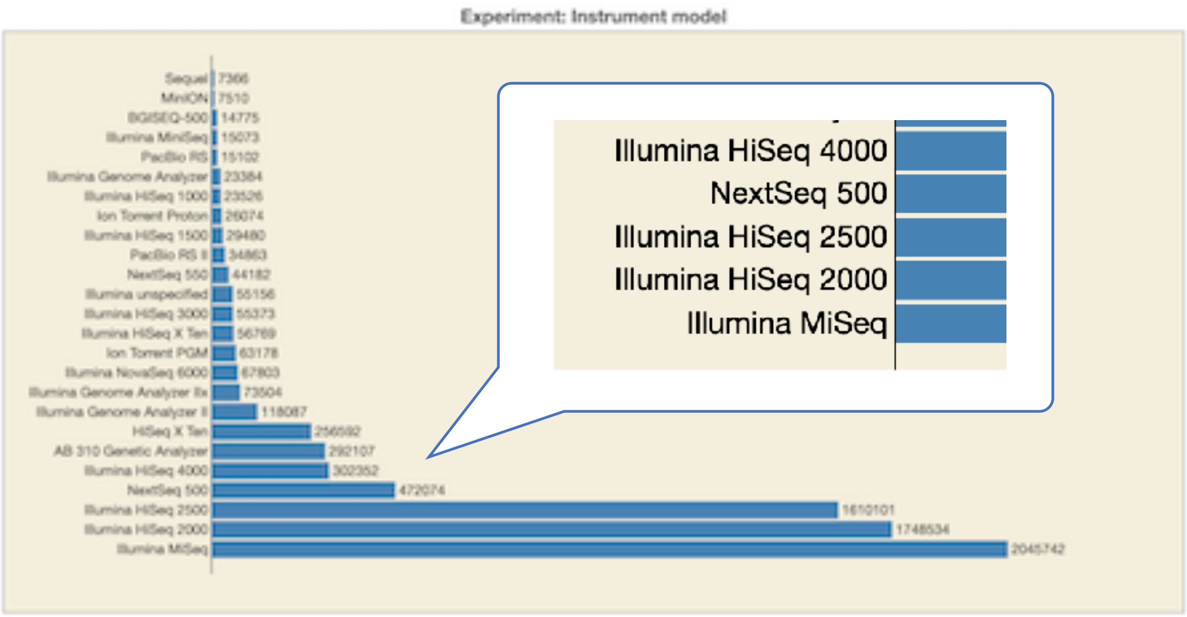


登録数の現状

BioProjectについての統計情報も提供している。

図は、目的別の各年と全体の目的別登録数。全体としてはtranscriptomeの登録が多い。

NGSデータの現状：platform別



登録数の現状

SRAも公開から10年以上経ち、その間に多くのシーケンサーが出現し、そして消えていった。
数年前まではIllumina HiSeq 2000がSRAデータで1位を占めていたが、今はMiSeqが1位となっている。