バイオインフォマティクス概論

東京科学大学　ILA国府台
中林　潤

JH人材育成課　バイオインフォマティシャン育成講座　①

講義の内容

- バイオインフォマティクス発展の経緯
- バイオインフォマティクスの特徴

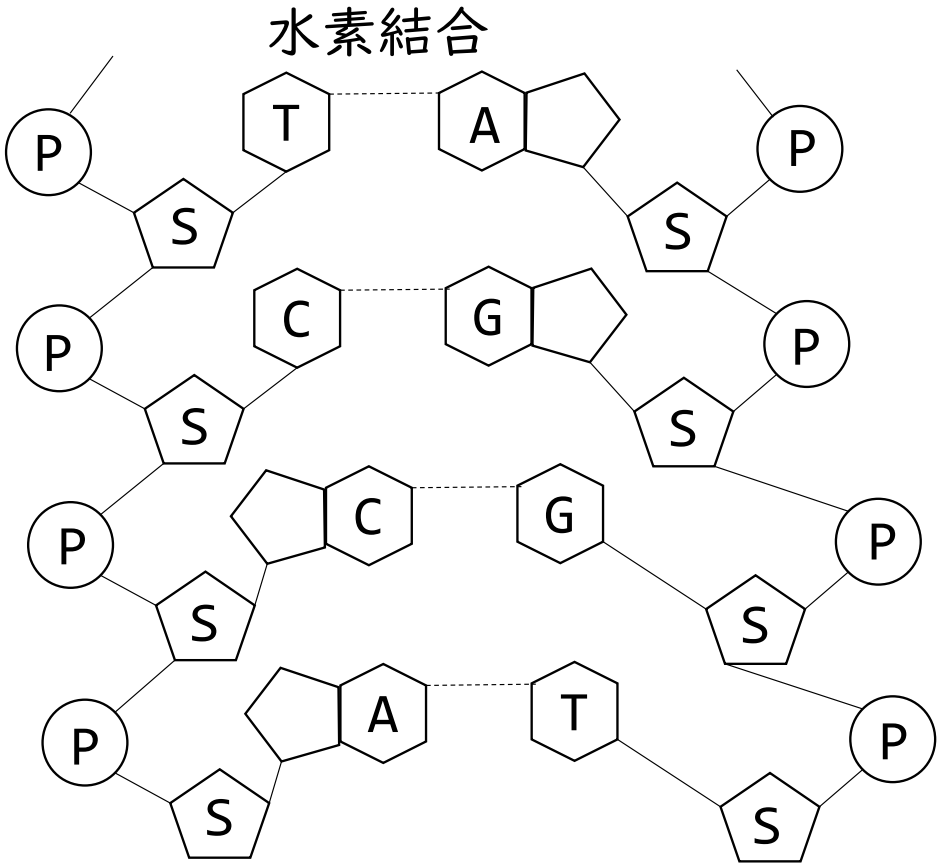
バイオインフォマティクスとは

- 生命を表す“バイオ”と情報科学を表す“インフォマティクス”を合せた造語
- 生命科学と情報科学の融合領域
- 特にゲノム情報のような大規模データを解析するための新しい手法を体系的にまとめたもの
- ここ二十数年で急速に発展してきた新しい分野

ビッグデータとデジタルトランスフォーメーション

- 社会のあらゆる階層で、コンピュータ、通信技術、計測機器の発展により、大規模データ（ビッグデータ）を収集することが容易になってきた。
- 大容量のデジタルデータを効率よく処理、解析することが必要となっている。
- 生命科学の領域でも扱うデータの大規模化が進んでいる。
- 特にヒトゲノム解読以降、顕著になっている。

遺伝子DNA



ヒトゲノム＝30億塩基対
ゲノム解析では30億文字分というような膨大な容量のデータを扱う

技術革新によるデータ産生量の増加

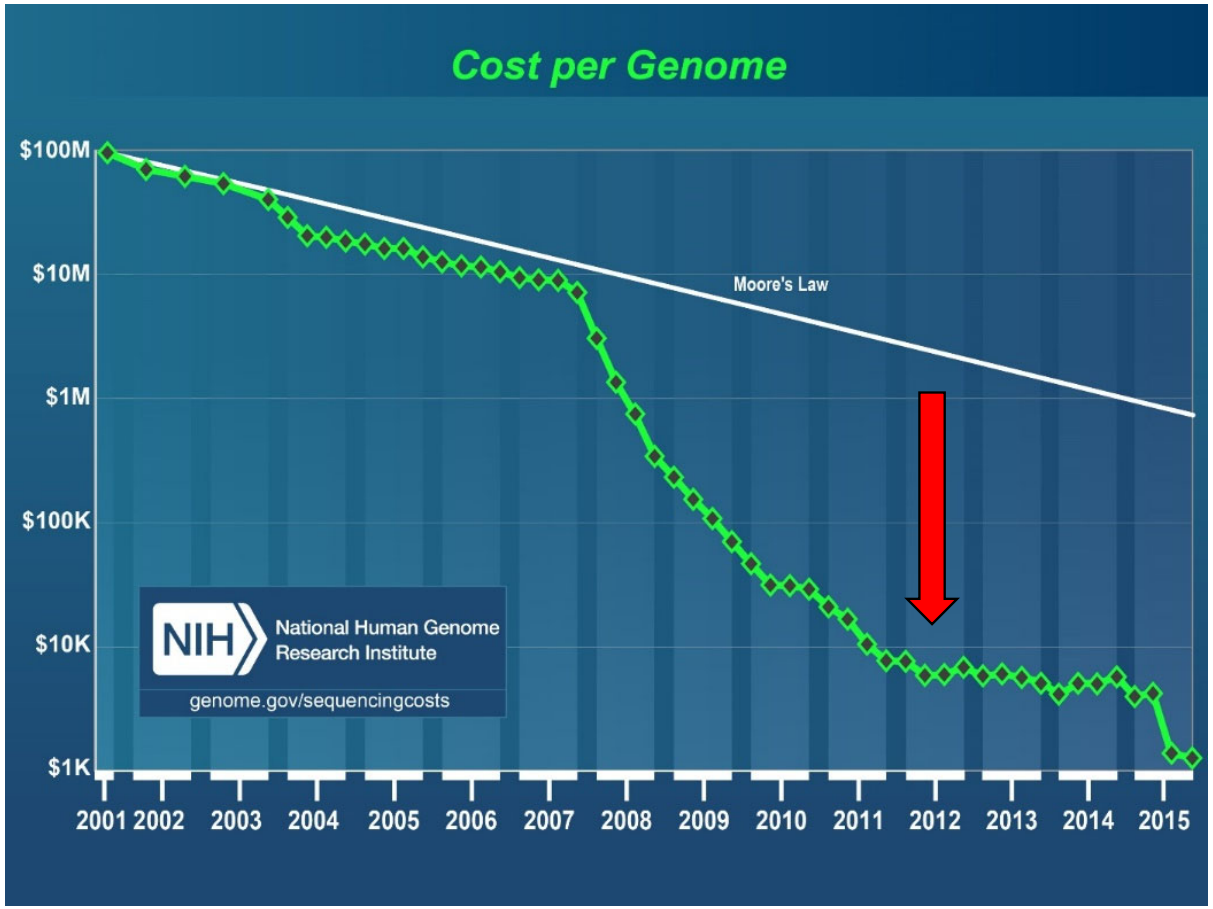
- ヒトゲノムプロジェクト
  1990年に開始、2003年にヒトゲノムシーケンスが公開される。
- 1000＄ゲノムプロジェクト
  ヒトゲノムを1000＄のコストで読めるシーケンサーを開発することを目的として、2004年に開始

従来型シーケンサーと次世代シーケンサーの性能差

	従来型シーケンサー	次世代シーケンサー
ヒトゲノムを読むのにかかる日数	13年間	3.5日
ヒトゲノムを読むのにかかる費用	1400万ドル	1000ドル

技術革新によるデータ産生量の増加

ヒトゲノムを読むのにかかるコストの変遷

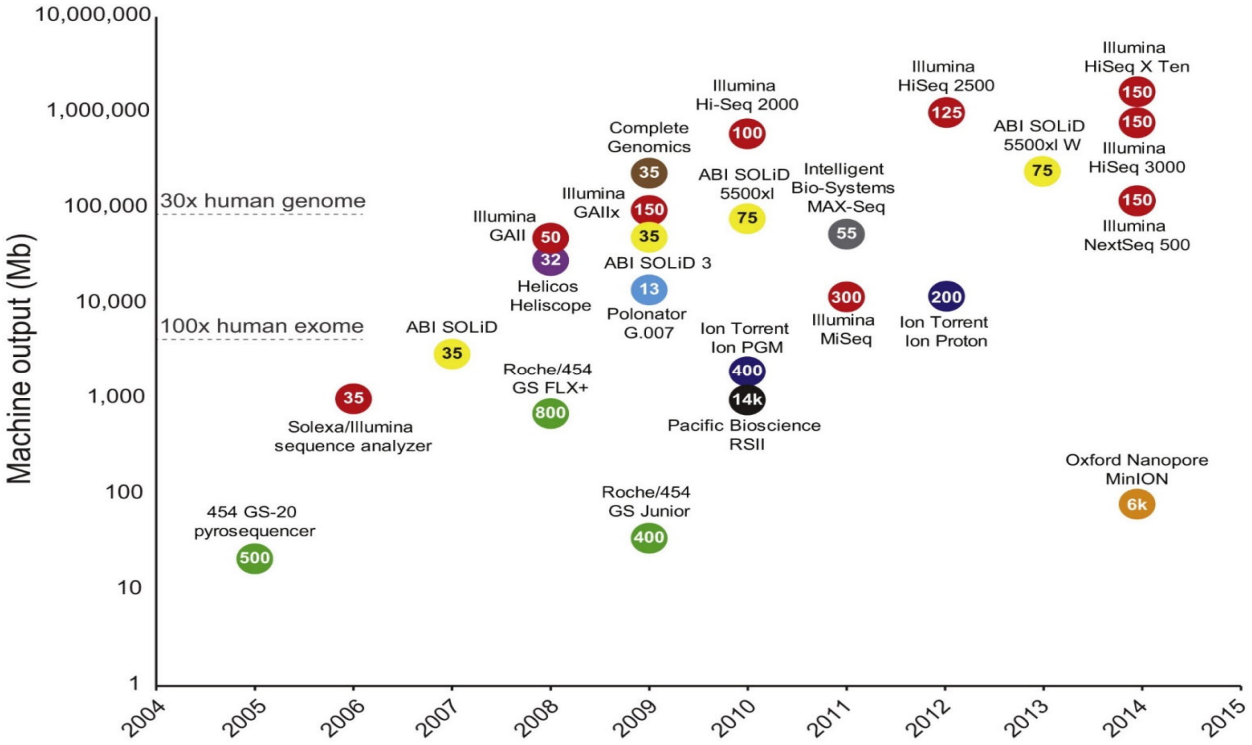


https://www.genome.gov/sequencingcosts/

1000$ゲノムプロジェクトの成果物として次世代シーケンサーが開発され、シーケンスコストが劇的に減少した

技術革新によるデータ産生量の増加

次世代シーケンサー1run当たり出力されるデータ量の変遷



Reuter JA, Spacek DV and Snyder MP ***Mol Cell*** 2015 May 21;58(4):586-97

シーケンスコストの低下に伴い、扱うデータ量が飛躍的に増大した。このような大規模データを解析するために、新しい手法を開発する必要がある。

仮説駆動型研究からデータ駆動型研究へ

- 仮説駆動型研究

  まず研究者が仮説を立てることから研究がスタートする。実験は仮説を検証するために行われる。統計的検定を行い仮説が支持されるか棄却されるか決める。

- データ駆動型研究

  仮説なしにデータを収集し、大量のデータを解析して、そこから新たな法則性やパターンを見出していく。仮説は後付けで得られることがある。機械学習などを使って新しい知見を得たりする。

- 生命科学の領域でもデータ駆動型研究が盛んに行われるようになってきた。

網羅的解析（-omics解析）

-ome：全体を表す接尾語

-ics：研究を表す接尾語

- 網羅的ゲノム解析
  genome, genomics
  Whole genome sequencing, Whole exome sequencing

- 網羅的発現解析
  transcriptome, transcriptomics
  RNA-seq, scRNA-seq

- 網羅的エピゲノム解析
  epigenome, epigenomics
  ATAC-seq, ChIP-seq

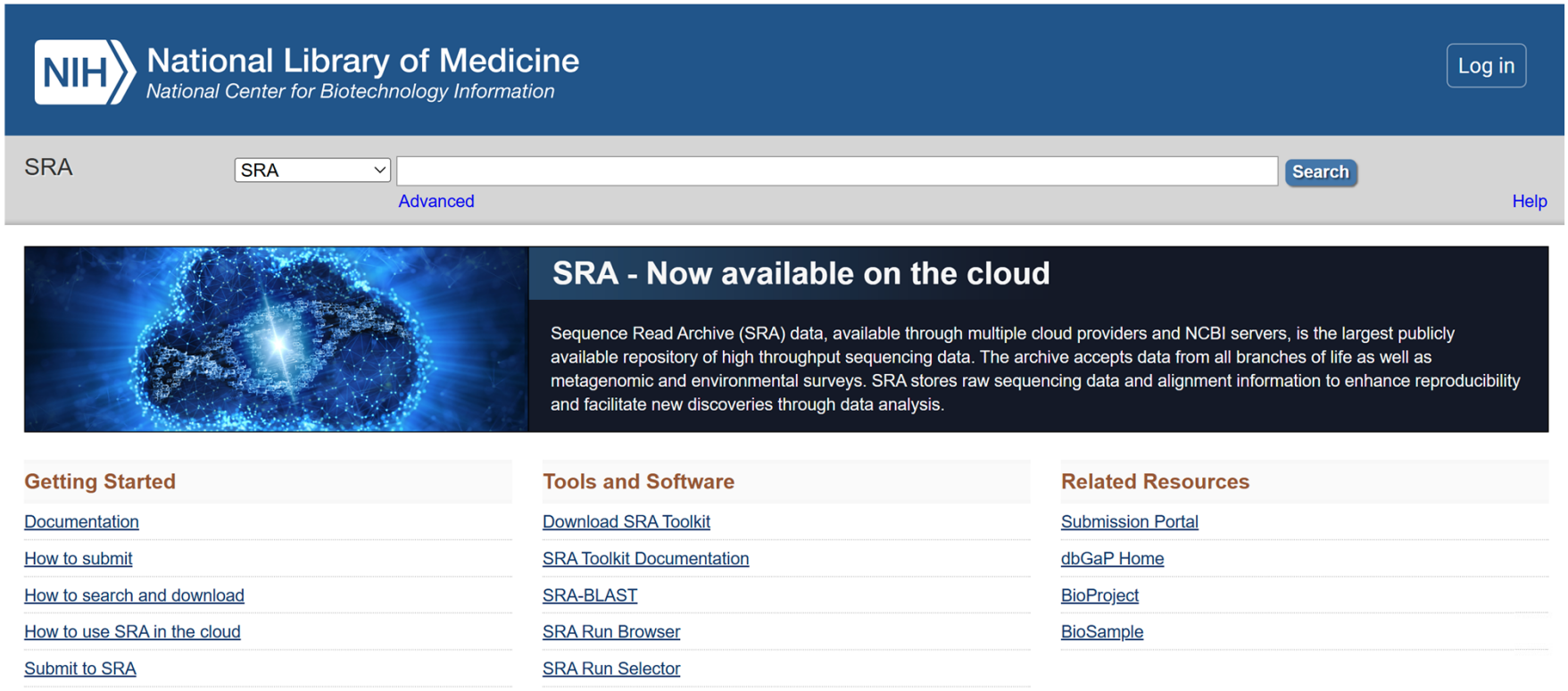
データベースの整備

- ゲノムシーケンスのような大規模データを保存しておくためのデータベースが整備されてきている。
  Sequence Read Archive（SRA）
  DDBJ Sequence Read Archive（DRA）
  European Nucleotide Archive（ENA）
  など

- データを他の研究者が再利用することが出来る。

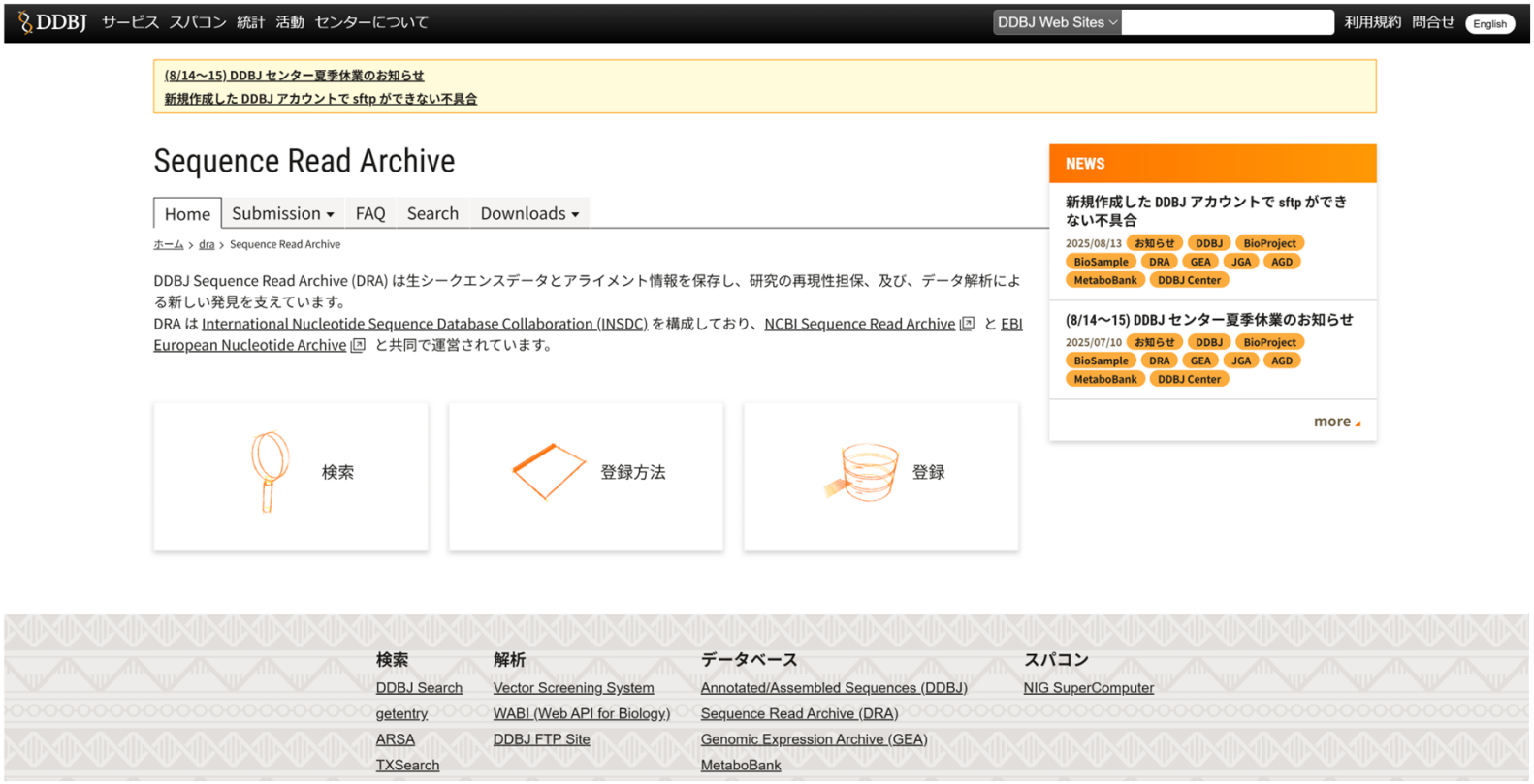
- 検証可能性や研究の透明性を確保するために、論文投稿時にデータの登録を要求されることが増えている。

SRAのホームページ



https://www.ncbi.nlm.nih.gov/sra

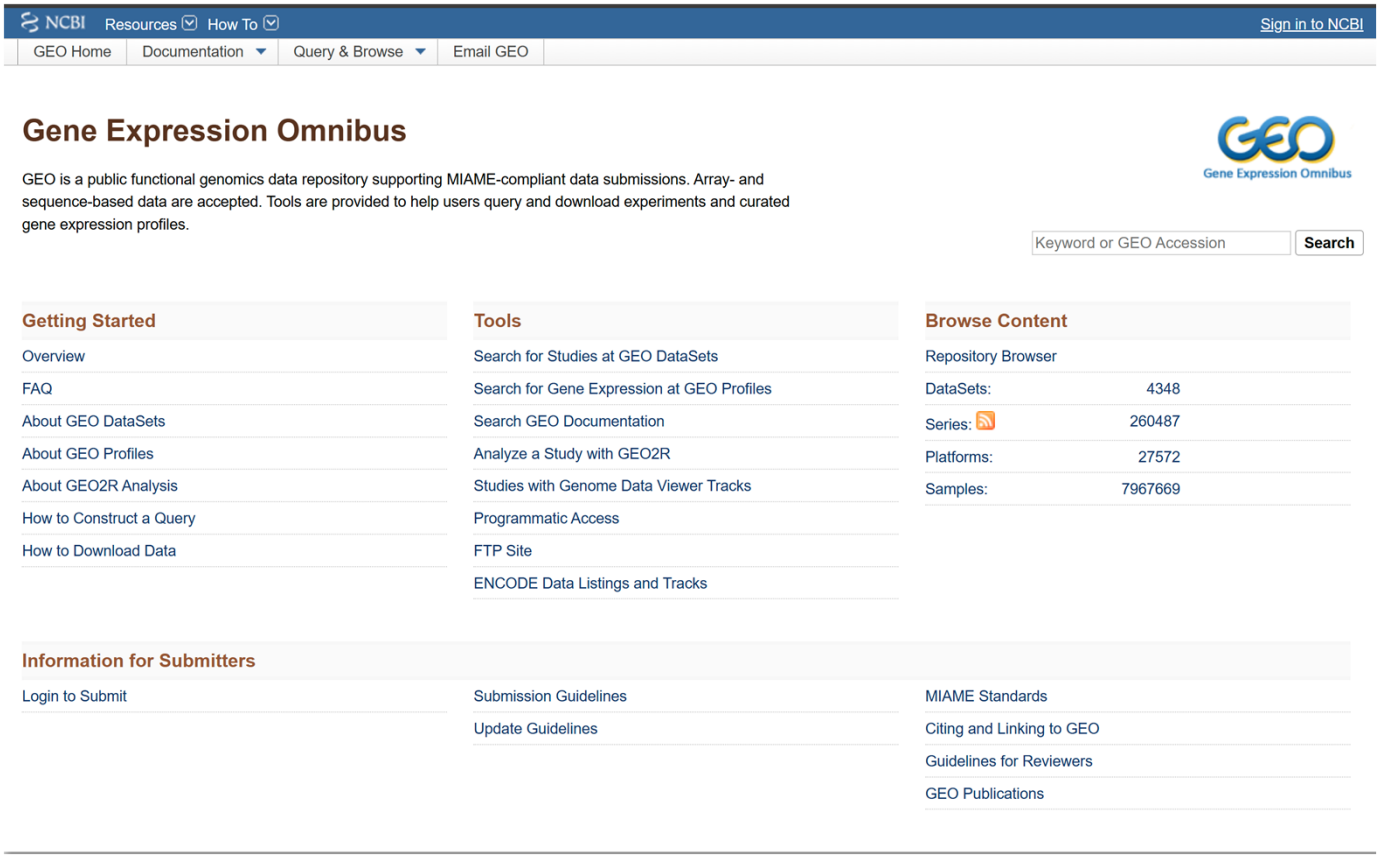
DRAのホームページ



https://www.ddbj.nig.ac.jp/dra/index.html

Gene Expression Omnibus（GEO）のホームページ

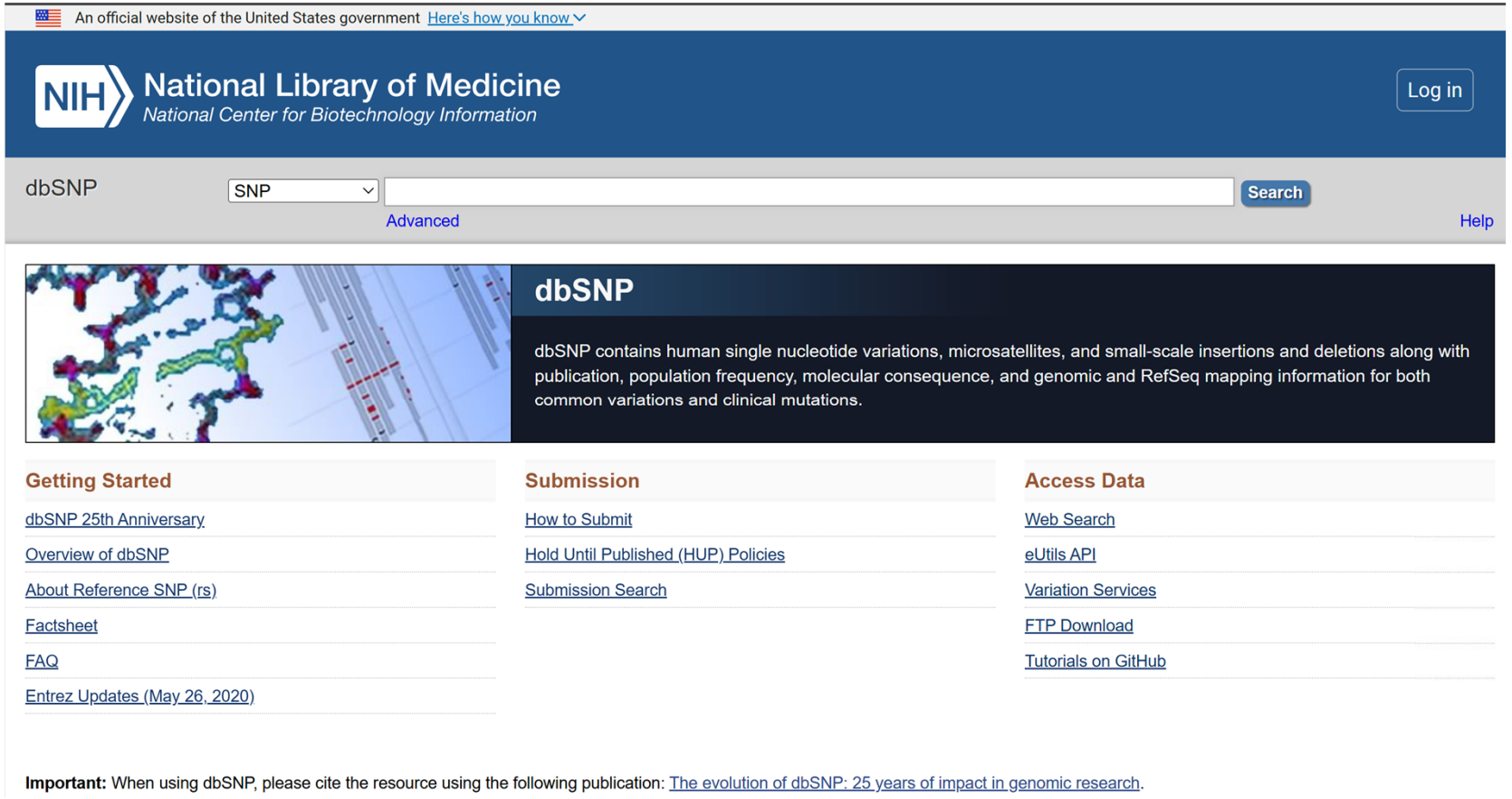
遺伝子発現データのデータベース



https://www.ncbi.nlm.nih.gov/geo/

dbSNPのホームページ

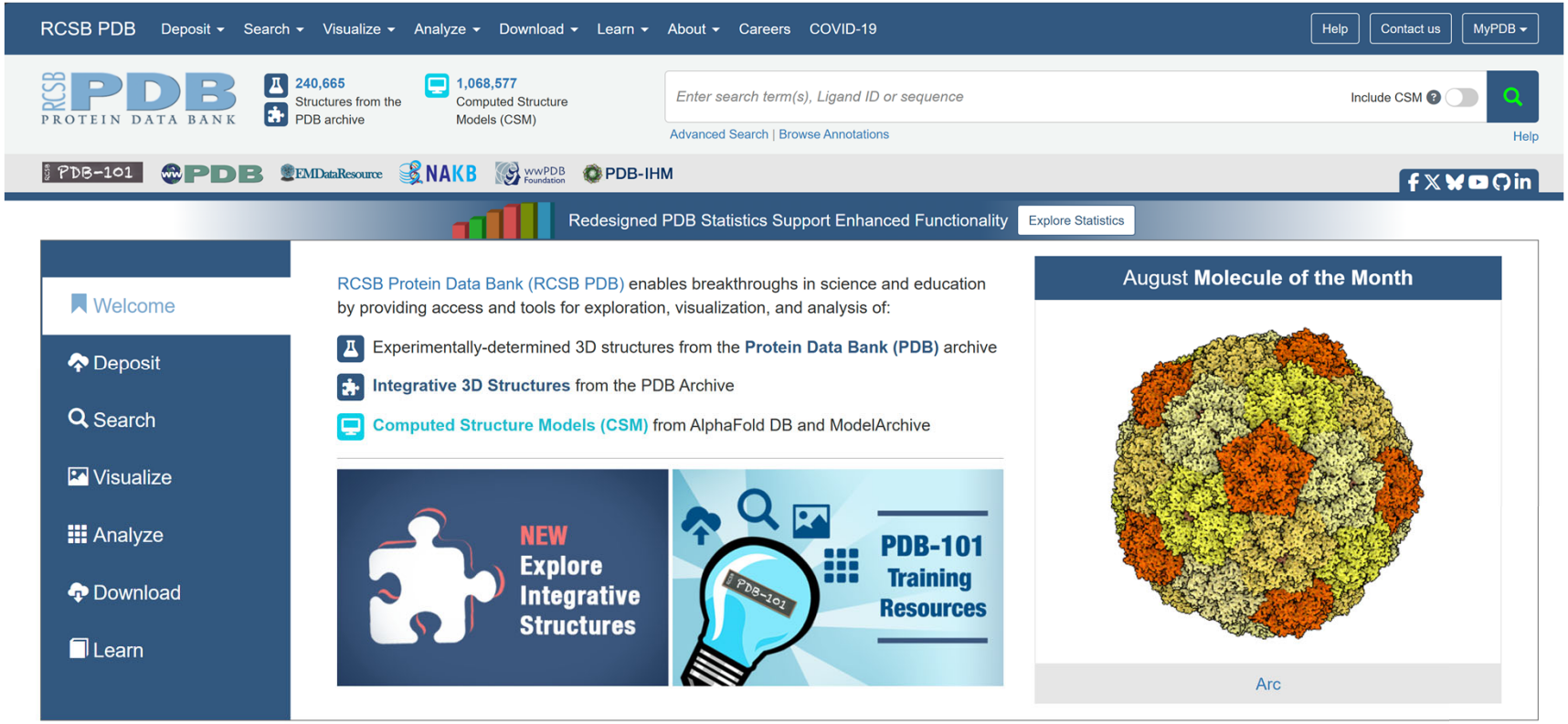
突然変異データのデータベース



https://www.ncbi.nlm.nih.gov/snp/

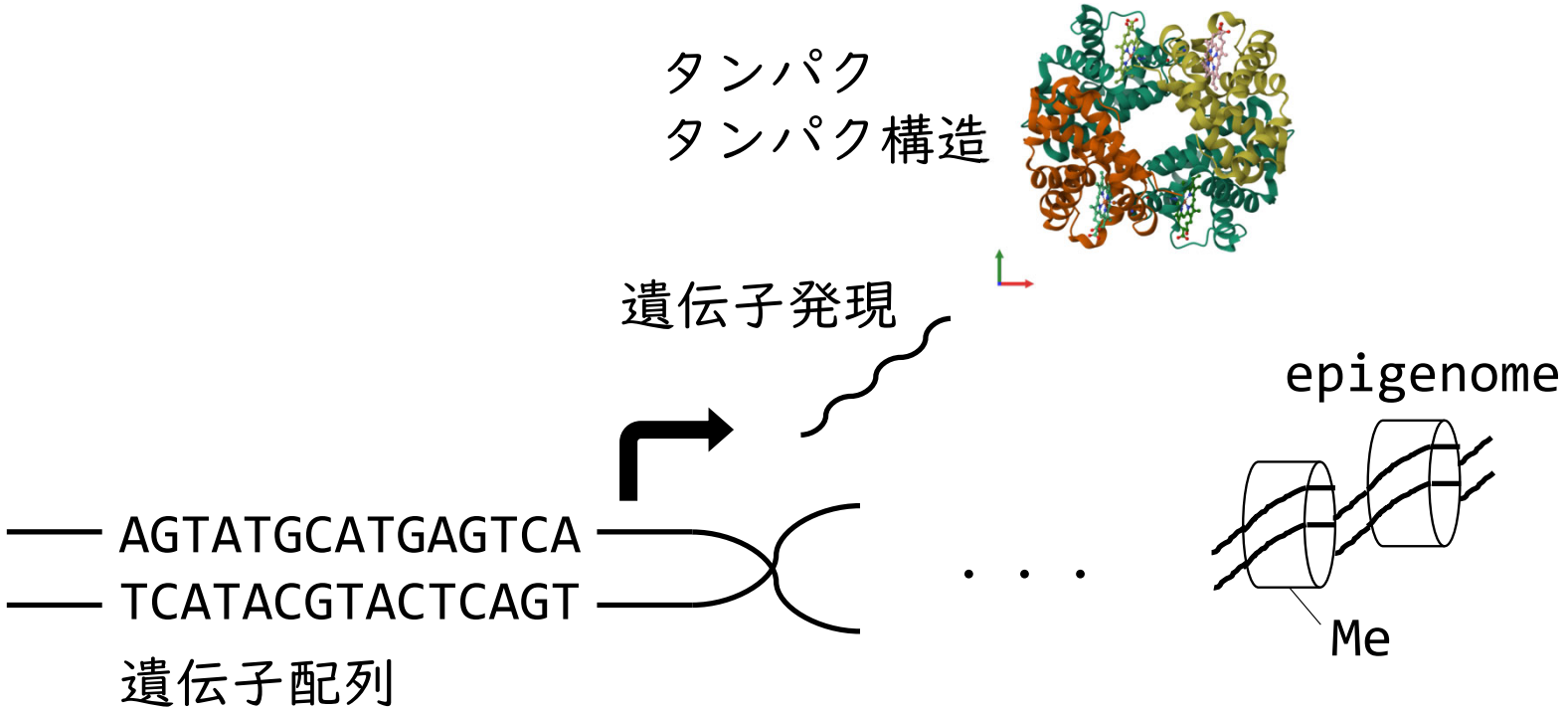
Protein Data Bank（PDB）のホームページ

タンパクのデータベース



https://www.rcsb.org/

バイオインフォマティクスで扱うデータ



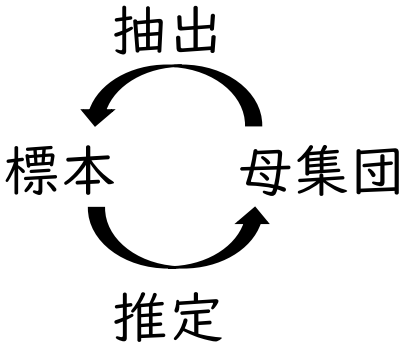
バイオインフォマティクスでは遺伝子の塩基配列、遺伝子発現、タンパクの発現と立体構造、epigenomeなど様々なデータを解析の対象として扱う

バイオインフォマティクス解析の基礎理論

- これまでに統計学、情報科学、データサイエンスの分野では、様々な理論に基づくデータ解析技術が培われてきた。
- これらの理論や解析技術はバイオインフォマティクスの論理的背景となっている。
- 解析方法の背後にある基礎理論を理解したうえで使用することが重要である。

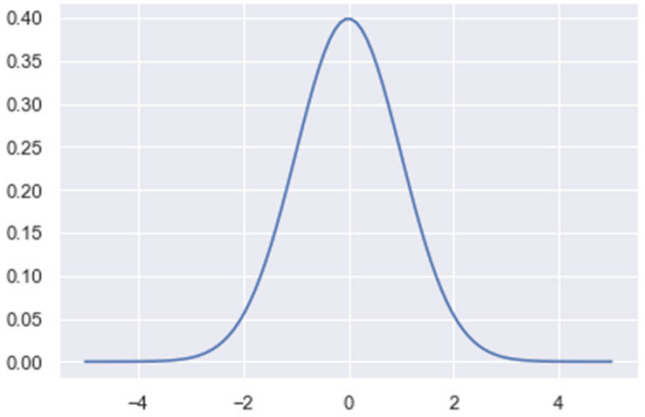
統計モデル

- 統計解析では得られた標本のデータから母集団の分布を推定する。
- 母集団がある確率分布に従うとして統計検定を行う。



正規分布
平均値の頻度が最も高く、外れるに従って
左右対称性に頻度が減少するような分布。
様々な測定値で正規分布が仮定される。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-(x-\mu)^2/2\sigma^2\right)}$$



遺伝子発現量の統計モデル

- 遺伝子発現量は負の二項分布に従うと仮定されている。
- 遺伝子発現量は平均値よりも分散が大きいため（過分散）、正規分布と仮定することが難しい。
- データの性質に応じて適切な分布を仮定することが重要である。

負の二項分布：「成功確率pの独立な試行がk回成功するまでに必要な試行回数Xが従う確率分布」

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

生命科学領域で扱うデータの特徴

- 生命科学領域で扱うデータは他の分野にない特徴を持つものが見られる。
  非線形性、非対称性、離散的、分散が大きい、0が多いなど

- ポアソン分布、指数分布、ベータ分布など様々な確率分布を適宜用いる。

- 特に網羅的解析では分布の選択が結果に大きな影響を及ぼすことがあるので注意を要する。

バイオインフォマティクス解析

- バイオインフォマティクス解析では非常に多種多様な大容量のデータを扱う。
- 効率的に解析するため、高速に演算処理するアルゴリズムが必要とされる。
- 解析用のアルゴリズムをまとめたものがライブラリとして公開されている。
  統計解析ソフトRのedgeRパッケージやSeuratパッケージなど
  pythonのScanpyライブラリなど
- これらのライブラリを使って解析を行う。

まとめ

- バイオインフォマティクスは生物学と情報科学の融合科学である。
- 統計学、情報科学、データサイエンスの理論に基づいて、大規模なデータを効率よく高い精度で解析する。
- 遺伝子配列、遺伝子発現、タンパクの発現と立体構造、epigenomeなど様々なデータを解析の対象として扱う。
- 背景にある理論を理解したうえで解析を行うことが重要である。