

# 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

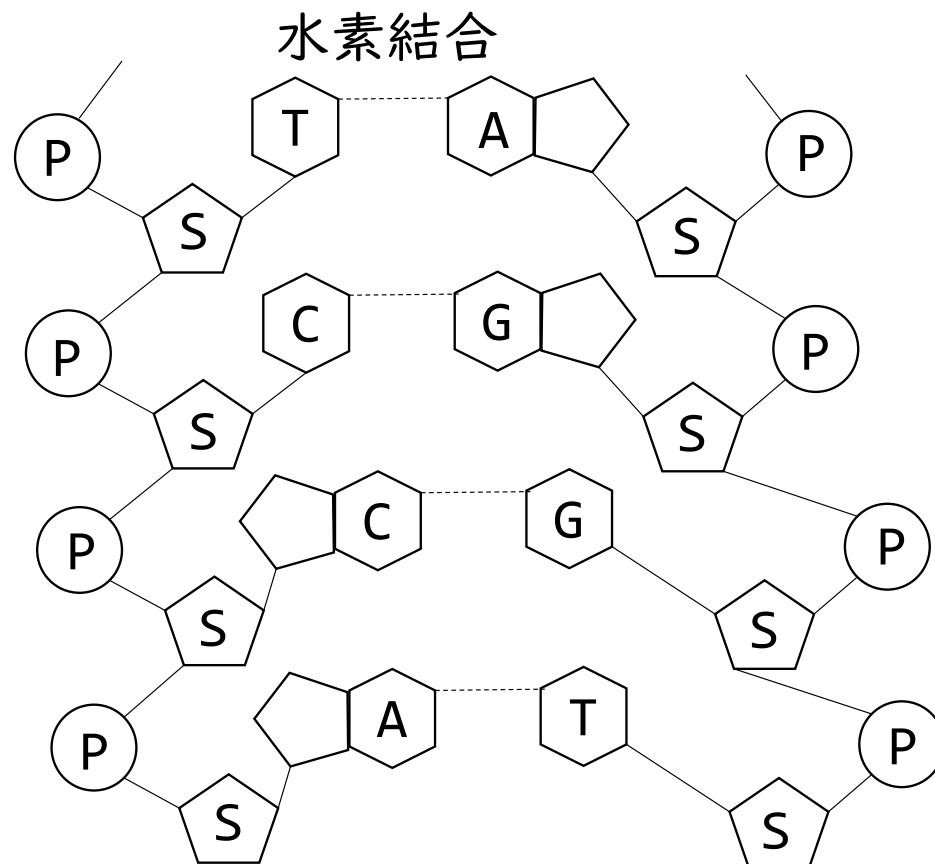


# ビッグデータとデジタルトランスフォーメーション

- ・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產生量の増加

- ・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クエンスが公開される。
- ・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クエンサーを開発することを目的として、2004年に開始

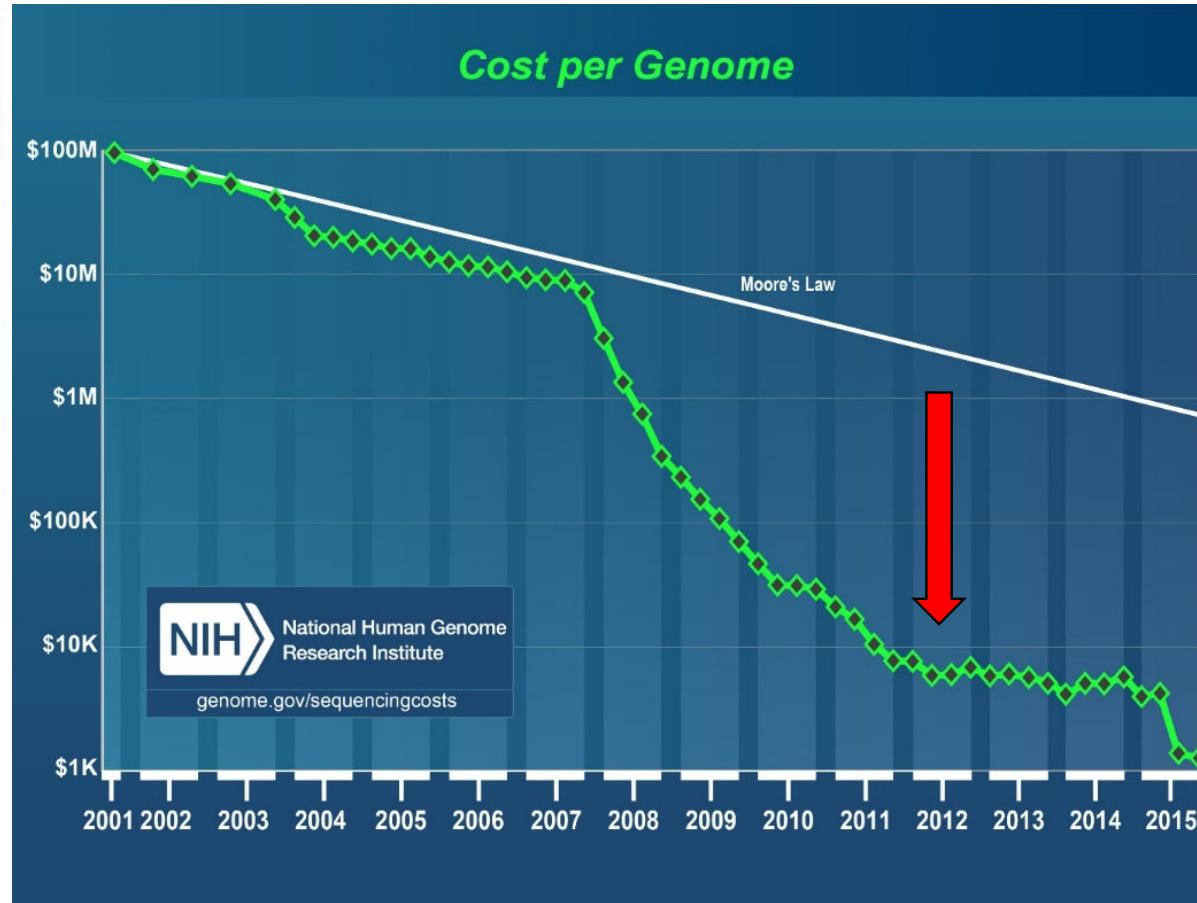
## 従来型シークエンサーと次世代シークエンサーの性能差

	従来型シークエンサー	次世代シークエンサー
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產生量の増加

ヒトゲノムを読むのにかかるコストの変遷



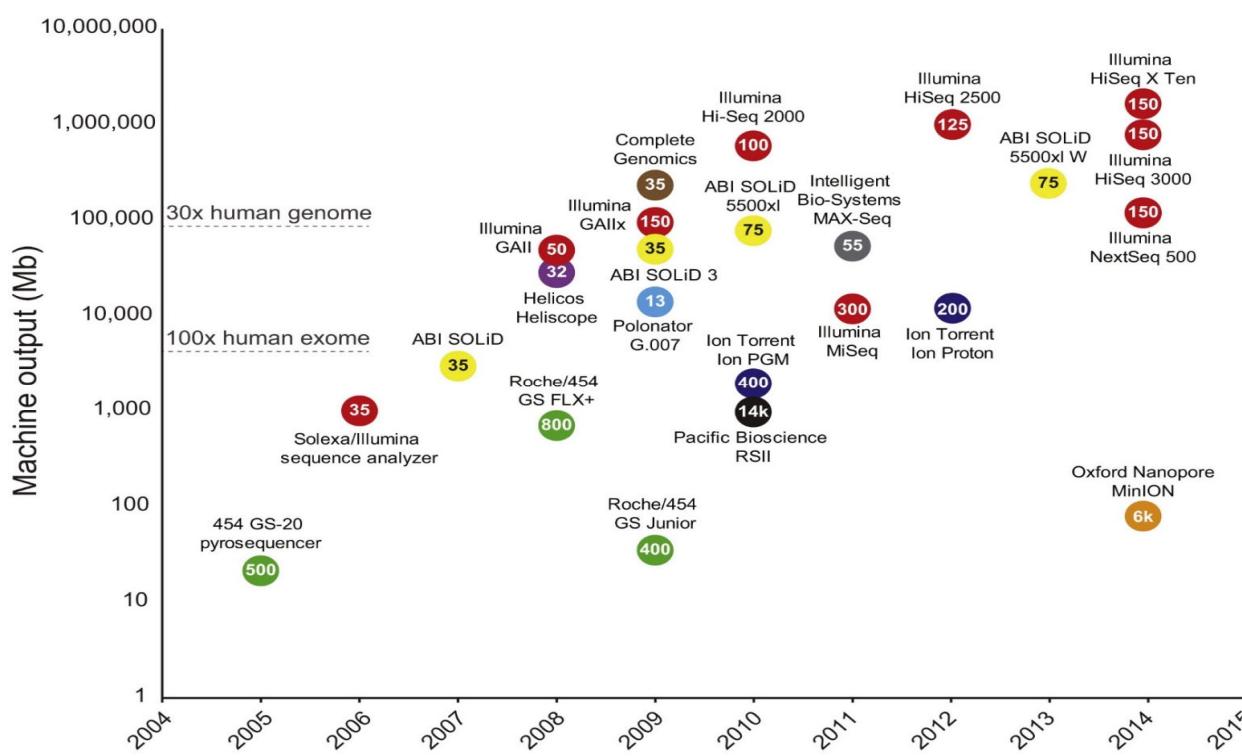
<https://www.genome.gov/sequencingcosts/>

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產生量の増加

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



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は、新しい手法を開発する必要がある。

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

# 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。仮説は後付けて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

- ゲノムシークエ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

The screenshot shows the official website for the Sequence Read Archive (SRA) at the National Library of Medicine. The header features the NIH logo and the text "National Library of Medicine" and "National Center for Biotechnology Information". A "Log in" button is in the top right. The main navigation bar includes "SRA" (selected), a dropdown menu, a search bar, and links for "Advanced", "Search", and "Help". Below the header is a large blue banner with the text "SRA - Now available on the cloud" and a description of the archive's purpose. To the left of the banner is a blue circular graphic representing a DNA helix or sequencing data. The main content area is divided into three columns: "Getting Started" (with links to Documentation, How to submit, How to search and download, How to use SRA in the cloud, and Submit to SRA), "Tools and Software" (with links to Download SRA Toolkit, SRA Toolkit Documentation, SRA-BLAST, SRA Run Browser, and SRA Run Selector), and "Related Resources" (with links to Submission Portal, dbGaP Home, BioProject, and BioSample).

**National Library of Medicine**  
National Center for Biotechnology Information

SRA    Advanced Help

**SRA - Now available on the cloud**

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

**Getting Started**

[Documentation](#)  
[How to submit](#)  
[How to search and download](#)  
[How to use SRA in the cloud](#)  
[Submit to SRA](#)

**Tools and Software**

[Download SRA Toolkit](#)  
[SRA Toolkit Documentation](#)  
[SRA-BLAST](#)  
[SRA Run Browser](#)  
[SRA Run Selector](#)

**Related Resources**

[Submission Portal](#)  
[dbGaP Home](#)  
[BioProject](#)  
[BioSample](#)

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



# DRAのホームページ

【DDBJ サービス スパコン 統計 活動 センターについて DDBJ Web Sites 利用規約 問合せ English】

(8/14~15) DDBJ センター夏季休業のお知らせ  
新規作成した DDBJ アカウントで sftp ができない不具合

## Sequence Read Archive

Home Submission ▾ FAQ Search Downloads ▾

ホーム > dra > Sequence Read Archive

DDBJ Sequence Read Archive (DRA) は生シーケンスデータとアライメント情報を保存し、研究の再現性担保、及び、データ解析による新しい発見を支えています。DRA は International Nucleotide Sequence Database Collaboration (INSDC) を構成しており、NCBI Sequence Read Archive ( ) と EBI European Nucleotide Archive ( ) と共に運営されています。

検索 登録方法 登録

### NEWS

新規作成した DDBJ アカウントで sftp ができない不具合  
2025/08/13 お知らせ DDBJ BioProject BioSample DRA GEA JGA AGD MetaboBank DDBJ Center

(8/14~15) DDBJ センター夏季休業のお知らせ  
2025/07/10 お知らせ DDBJ BioProject BioSample DRA GEA JGA AGD MetaboBank DDBJ Center

more ▾

検索 解析 データベース スパコン

DDBJ Search Vector Screening System Annotated/Assembled Sequences (DDBJ)  
getentry WABI (Web API for Biology)  
ARSA DDBJ FTP Site Genomic Expression Archive (GEA)  
TXSearch MetaboBank NIG SuperComputer

<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

NCBI Resources How To

GEO Home Documentation Query & Browse Email GEO Sign in to NCBI

### Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

**GEO**  
Gene Expression Omnibus

Keyword or GEO Accession

Getting Started	Tools	Browse Content
Overview	Search for Studies at GEO DataSets	Repository Browser
FAQ	Search for Gene Expression at GEO Profiles	DataSets: 4348
About GEO DataSets	Search GEO Documentation	Series: 260487
About GEO Profiles	Analyze a Study with GEO2R	Platforms: 27572
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples: 7967669
How to Construct a Query	Programmatic Access	
How to Download Data	FTP Site	
	ENCODE Data Listings and Tracks	

Information for Submitters		
Login to Submit	Submission Guidelines	MIAME Standards
	Update Guidelines	Citing and Linking to GEO
		Guidelines for Reviewers
		GEO Publications

Information for Submitters

Login to Submit

Submission Guidelines

Update Guidelines

MIAME Standards

Citing and Linking to GEO

Guidelines for Reviewers

GEO Publications

Speaker icon

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

# dbSNPのホームページ

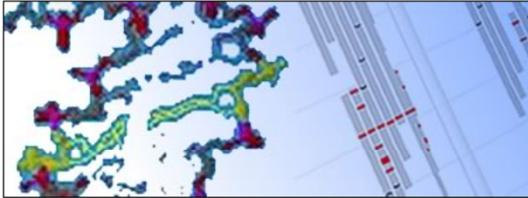
## 突然変異データのデータベース

An official website of the United States government [Here's how you know](#)

 National Library of Medicine  
National Center for Biotechnology Information

Log in

dbSNP    Advanced Help



**dbSNP**

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

**Getting Started**

[dbSNP 25th Anniversary](#)  
[Overview of dbSNP](#)  
[About Reference SNP \(rs\)](#)  
[Factsheet](#)  
[FAQ](#)  
[Entrez Updates \(May 26, 2020\)](#)

**Submission**

[How to Submit](#)  
[Hold Until Published \(HUP\) Policies](#)  
[Submission Search](#)

**Access Data**

[Web Search](#)  
[eUtils API](#)  
[Variation Services](#)  
[FTP Download](#)  
[Tutorials on GitHub](#)



**Important:** When using dbSNP, please cite the resource using the following publication: [The evolution of dbSNP: 25 years of impact in genomic research](#).

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

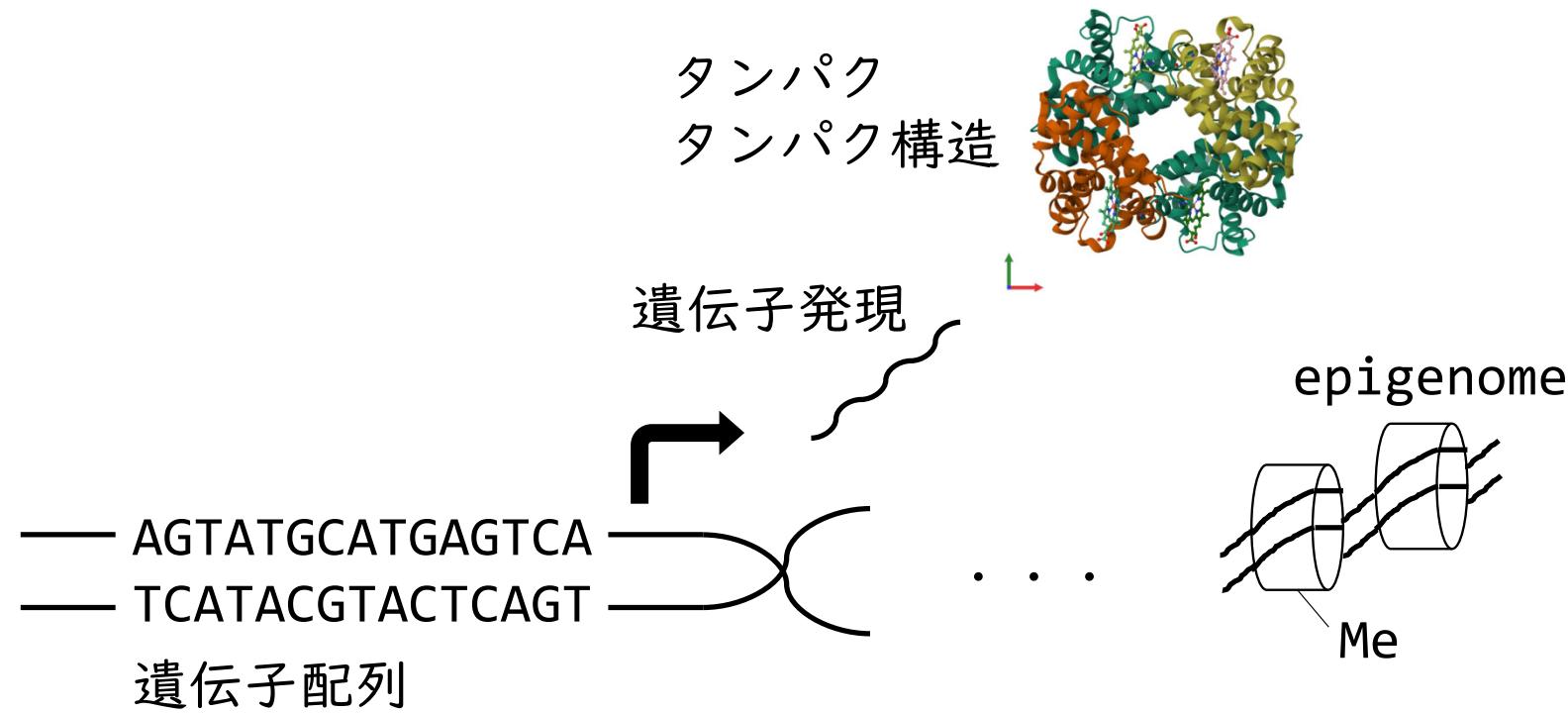
# Protein Data Bank (PDB) のホームページ

## タンパクのデータベース

The screenshot shows the RCSB PDB homepage. The top navigation bar includes links for Deposit, Search, Visualize, Analyze, Download, Learn, About, Careers, COVID-19, Help, Contact us, and MyPDB. The main header features the RCSB PDB logo with statistics: 240,665 Structures from the PDB archive and 1,068,577 Computed Structure Models (CSM). A search bar allows users to enter search terms, ligand ID, or sequence, with options to include CSM and perform an advanced search or browse annotations. Below the header are links to PDB-101, wwPDB, EMDDataResource, NAKB, wwPDB Foundation, and PDB-IHM. A banner at the top right indicates "Redesigned PDB Statistics Support Enhanced Functionality" and provides a link to "Explore Statistics". The left sidebar contains links for Welcome, Deposit, Search, Visualize, Analyze, Download, and Learn. The main content area highlights the RCSB Protein Data Bank's role in science and education, mentioning Experimentally-determined 3D structures, Integrative 3D Structures, and Computed Structure Models (CSM). It also promotes the "NEW Explore Integrative Structures" feature and "PDB-101 Training Resources". A large graphic for "August Molecule of the Month" shows a complex protein structure composed of colored subunits. A speaker icon in the bottom right corner indicates audio content.

<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

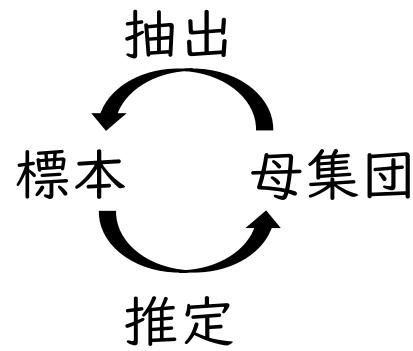
# バイオインフォマティクス解析の基礎理論

- 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

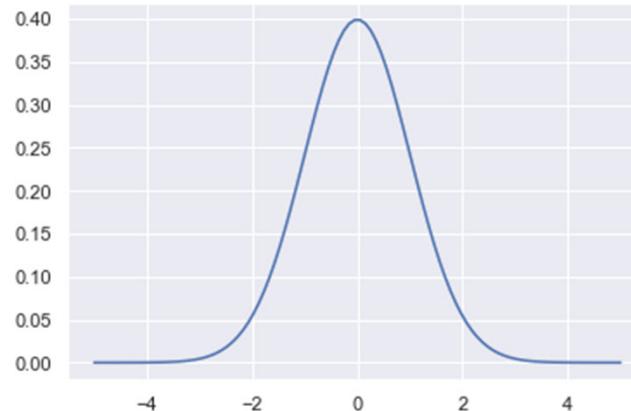
- ・統計解析では得られた標本のデータから母集団の分布を推定する。
- ・母集団がある確率分布に従うとして統計検定を行う。



## 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。

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



# 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。

