

# 演習A

# 本日のスケジュール

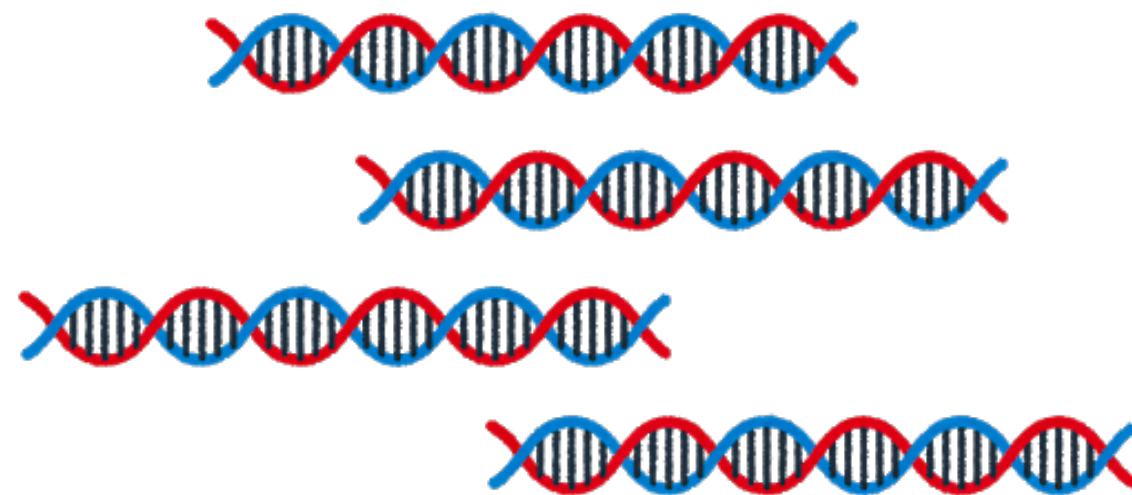
- 9:30-10:00 説明、演習A（リアルタイム）
- 10:00-11:30 演習B（資料＋オンデマンド）
- 11:30-12:00 ラップアップ（リアルタイム）
- 13:30-16:30 演習C（資料＋リアルタイムオンライン）

# データ解析が主戦場の生物学

# 次世代シーケンサー（Next Generation Sequencer; NGS）

- サンガー法（第一世代）以降、2000年代中頃から登場した、新しいDNAシーケンス技術の総称
- ハイスループットな塩基配列読み取り（シーケンシング）ができる
- 「次世代」というには登場から時間が経ちすぎていることもあり、最近ではHigh Throughput Sequencer (HTS) と呼ばれることも

## DNA（物質）



## NGS



## DNA配列（データ）

```
ACGTAAGCCGAAG  
GATTTTAACCCCC  
CAATTCCCGGAGC  
ACGCAAAATTAAG
```

# NGS=汎用超並列塩基配列決定装置

- ある性質を持ったDNAやRNAをライブラリDNAに変換できれば、NGSでさまざまな生命現象を網羅的に測定できる

物質

データ

RNAを逆転写して  
DNAにしたもの

転写されていたRNAの  
塩基配列

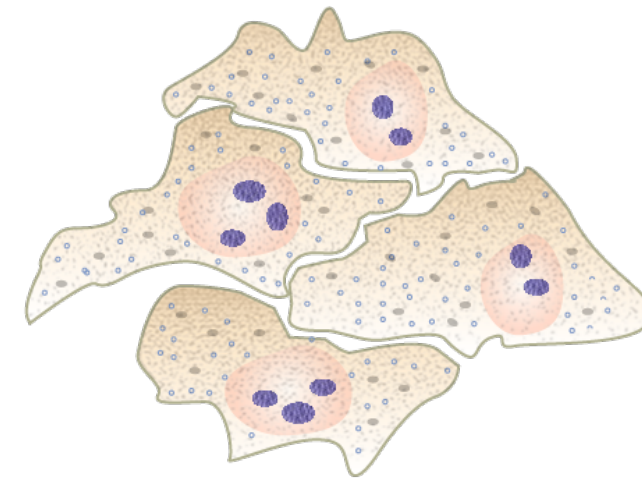
タンパク質が結合している  
短いDNA断片

タンパク質が結合していた  
塩基配列

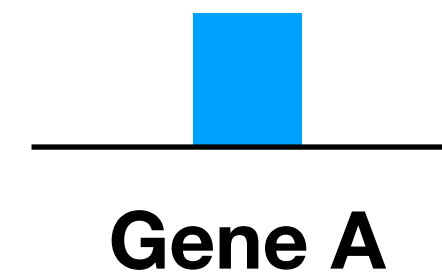
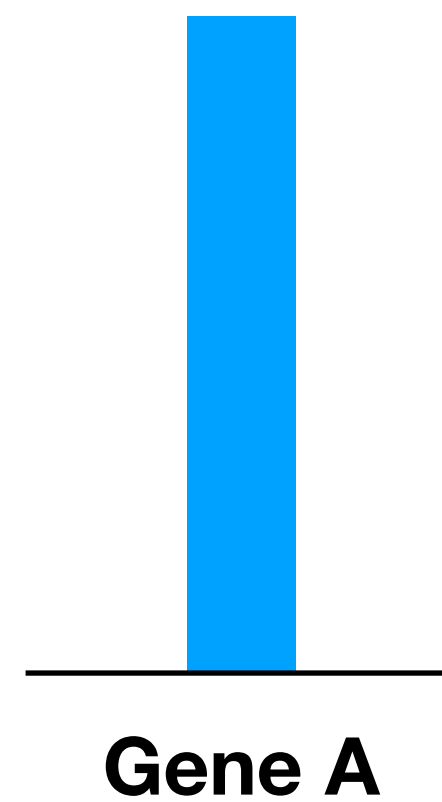
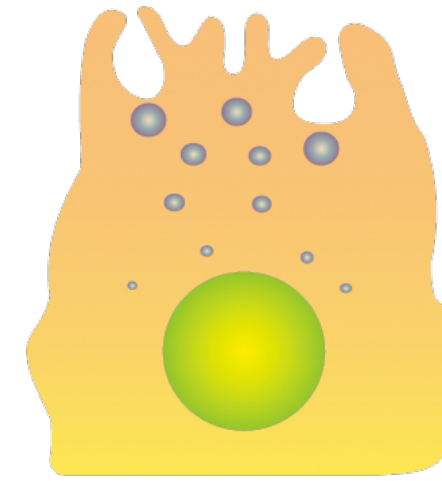


# 一つの遺伝子のmRNA量に着目する場合

がん細胞



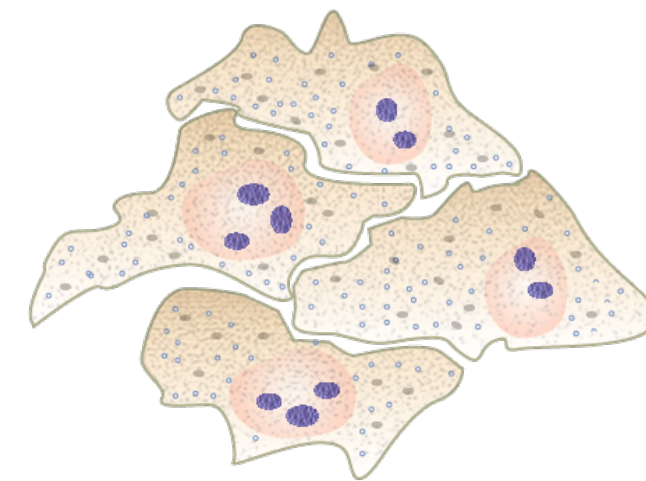
正常細胞



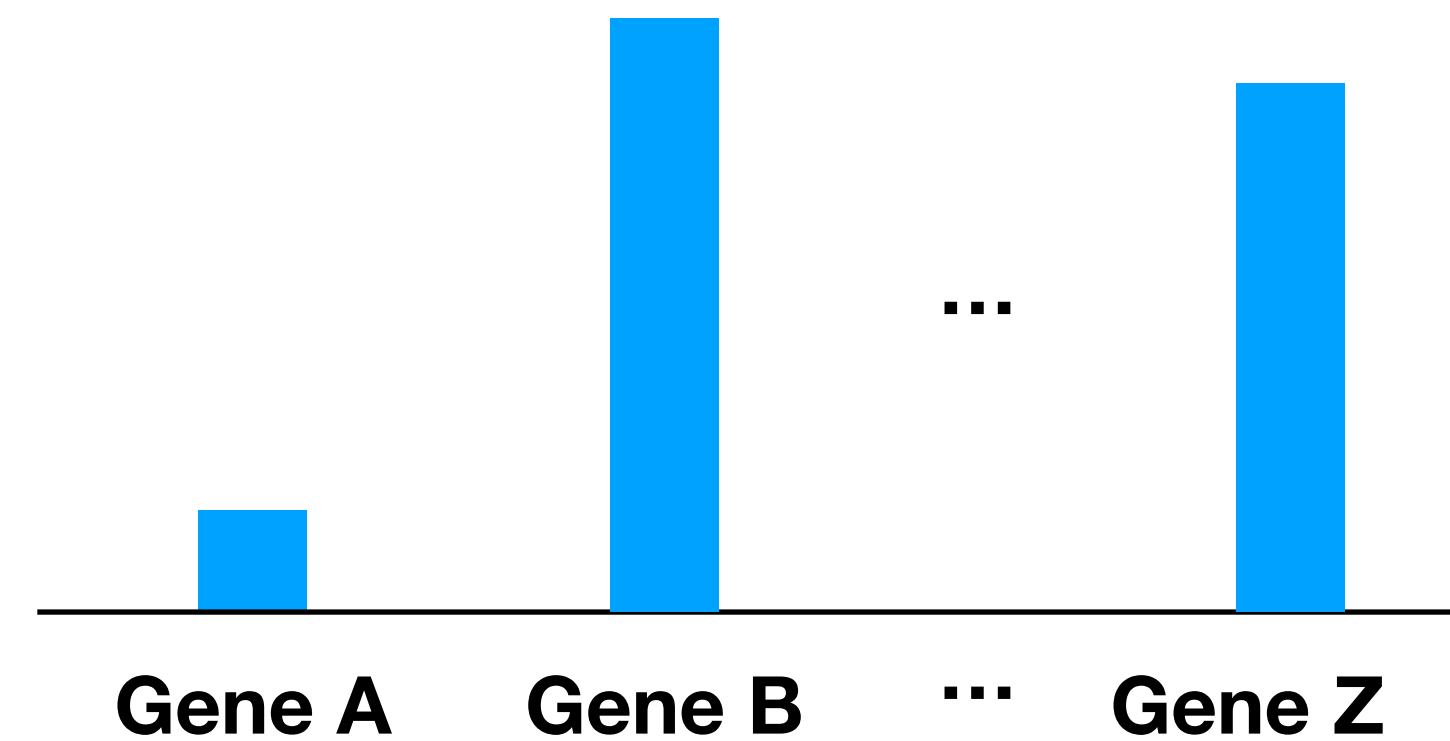
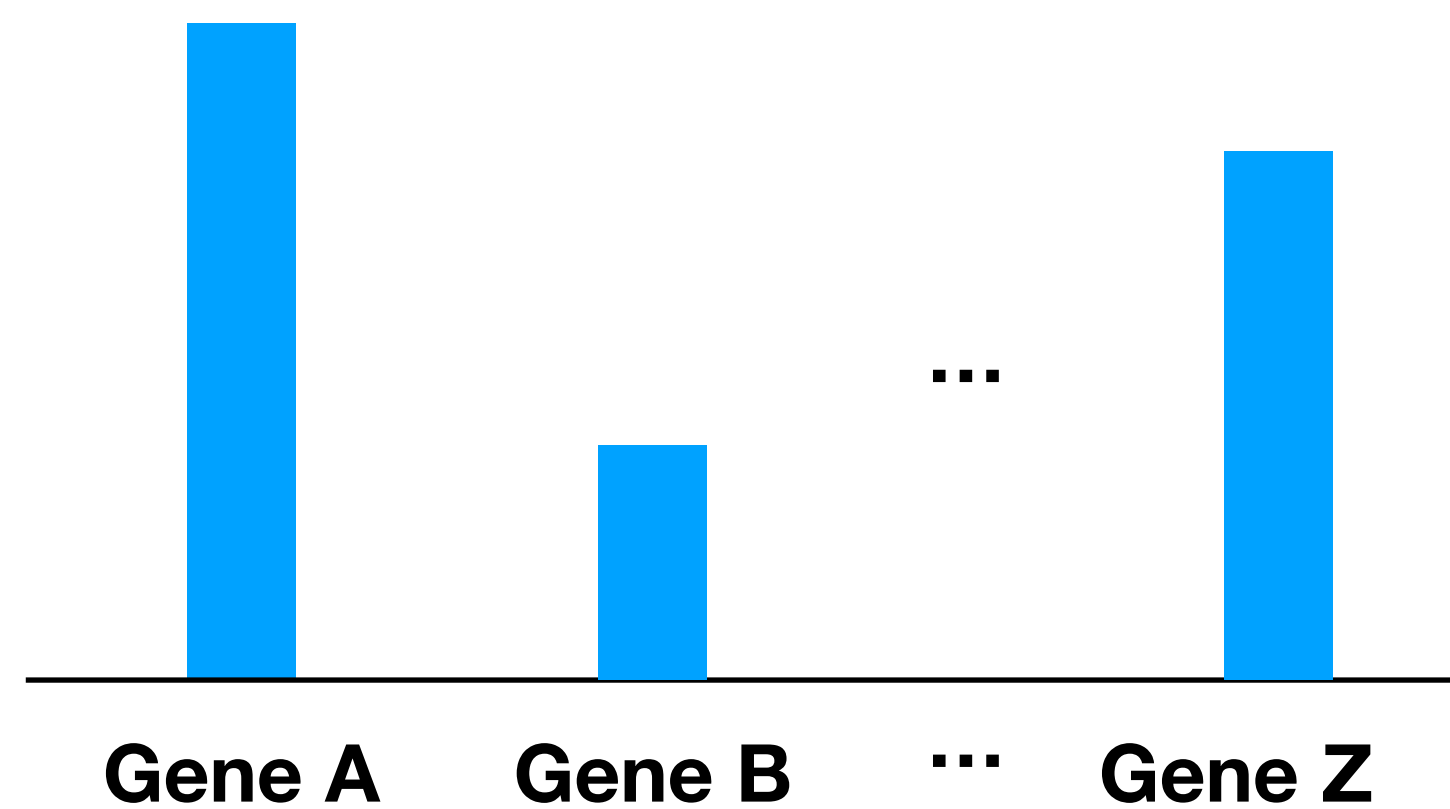
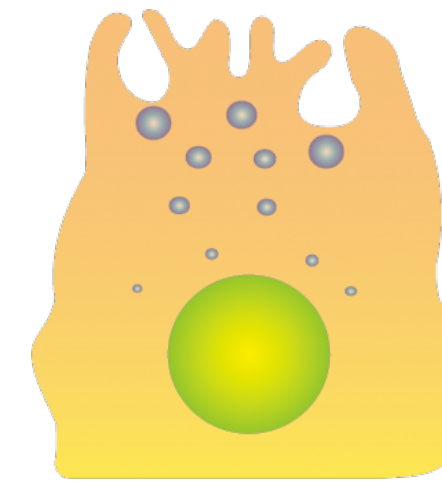
「遺伝子発現量の差がある → 原因かもしれない」という作業仮説

# ハイスループット実験としてのオミクス計測

がん細胞



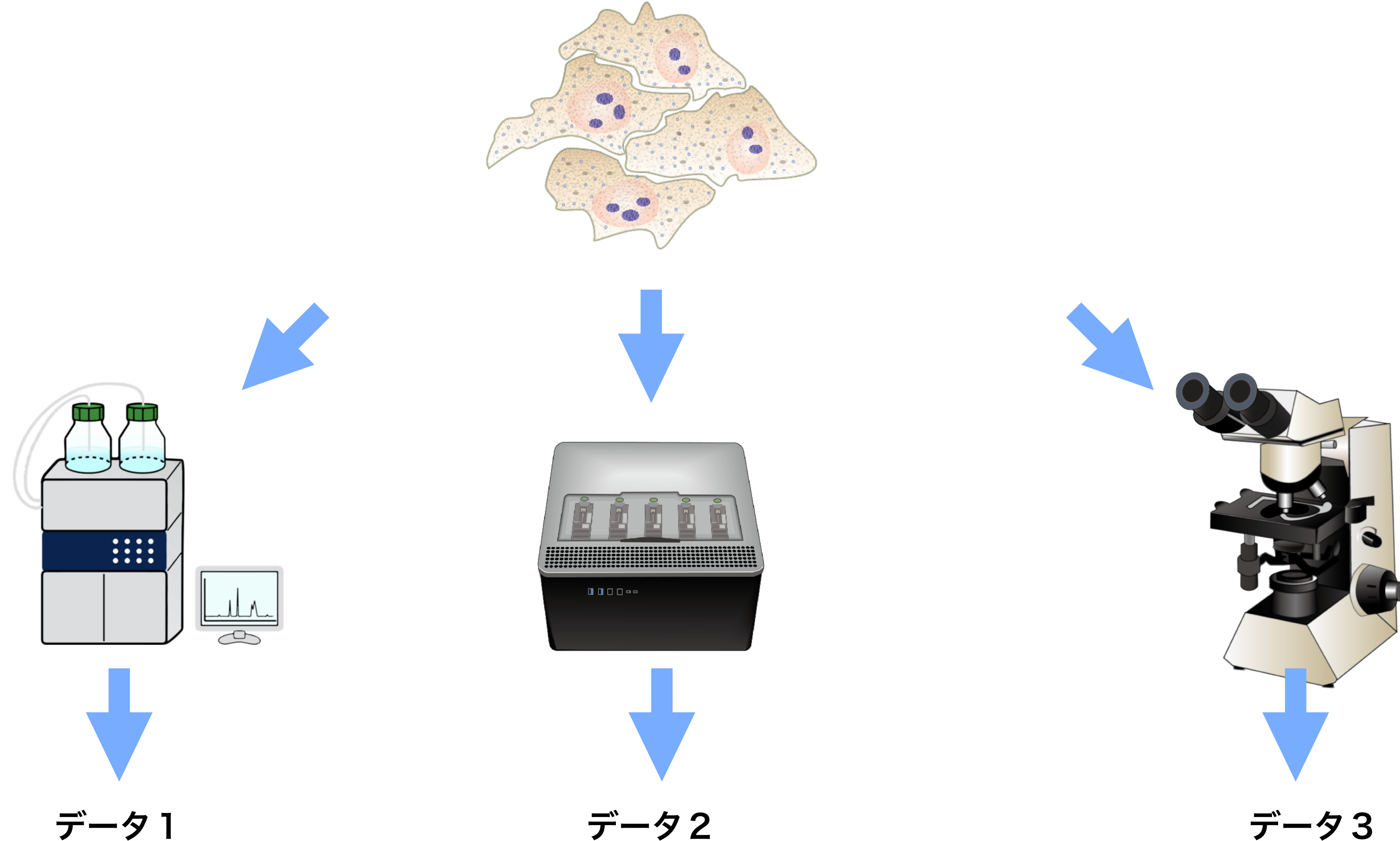
正常細胞



多数の結果を得る→同じ作業仮説で多数の分子をスクリーニングにかけられる



# 様々なモードで計測されるようになった



データ 1

データ 2

データ 3

それぞれの計測系のデータに含まれる情報は異なる

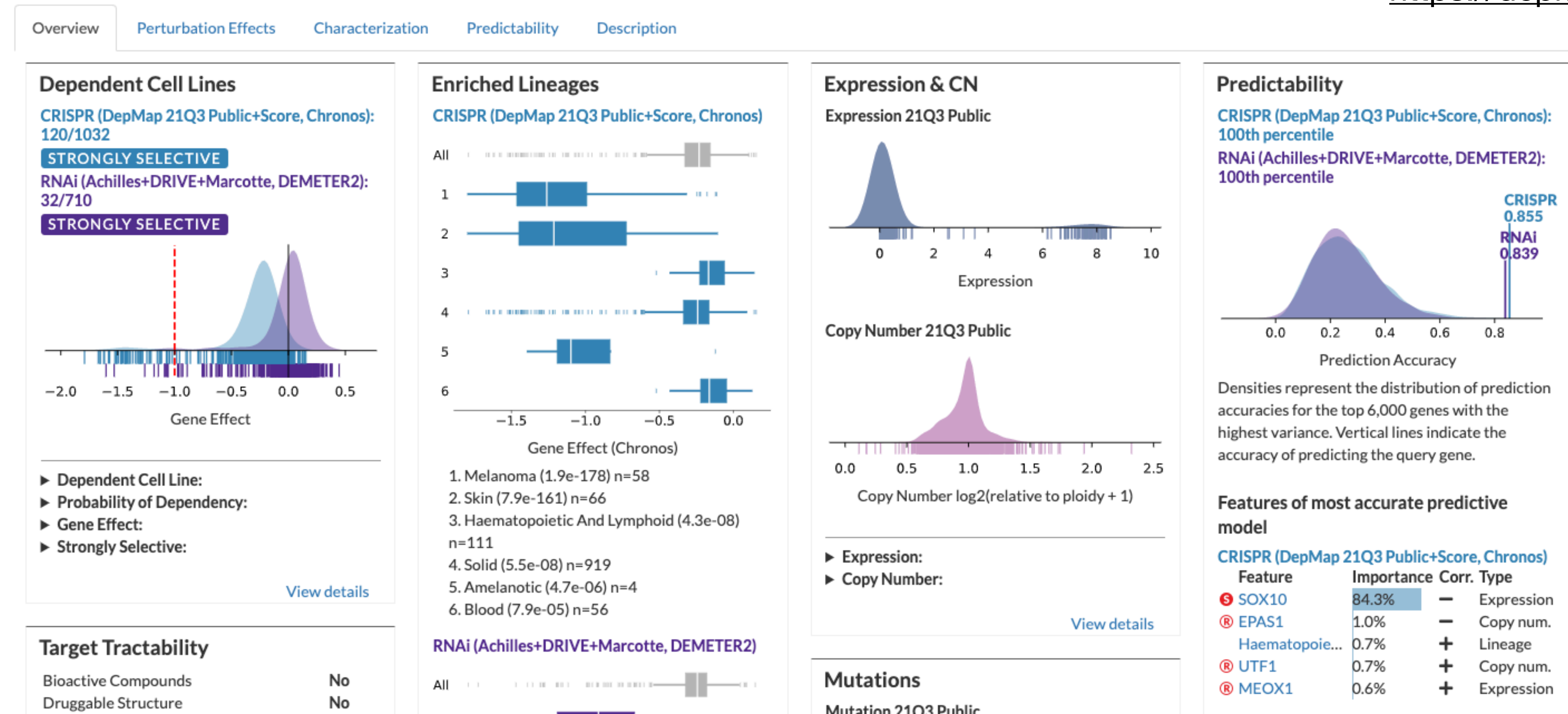


# Dependency map

- 様々ながん種のがん細胞株に対する薬剤や遺伝子ノックアウト実験の結果を収載

SOX10 SRY-box transcription factor 10

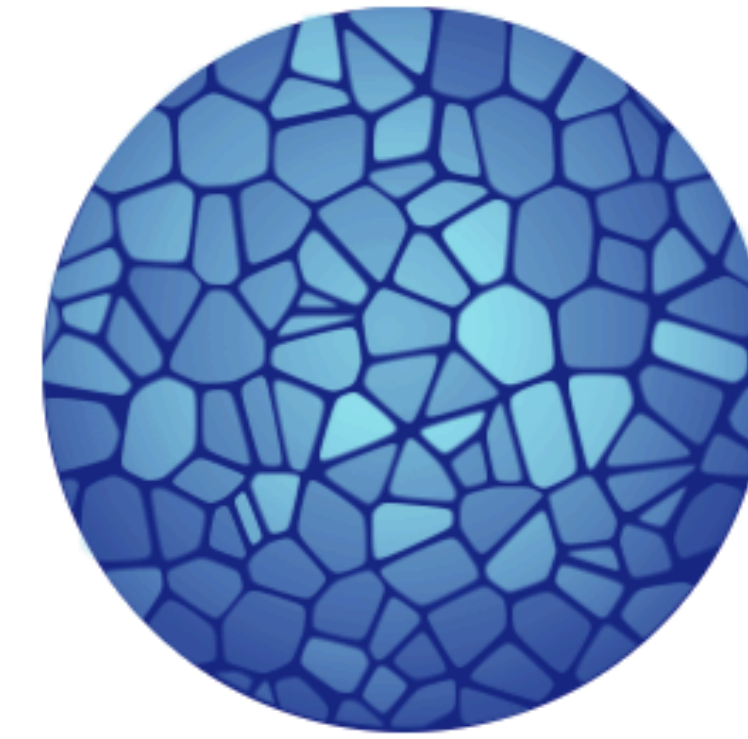
<https://depmap.org/portal/>



摂動実験の結果も手に入る時代

# Human Cell Atlas

- 人間の全身すべての細胞の地図を作る国際プロジェクト
  - Facebookのザッカーバーグ夫妻の財団 Chan Zuckerberg Initiativeから3000億円出資
- シングルセルトランスクリプトーム計測（一細胞レベルでの遺伝子発現計測）が要素技術として用いられている



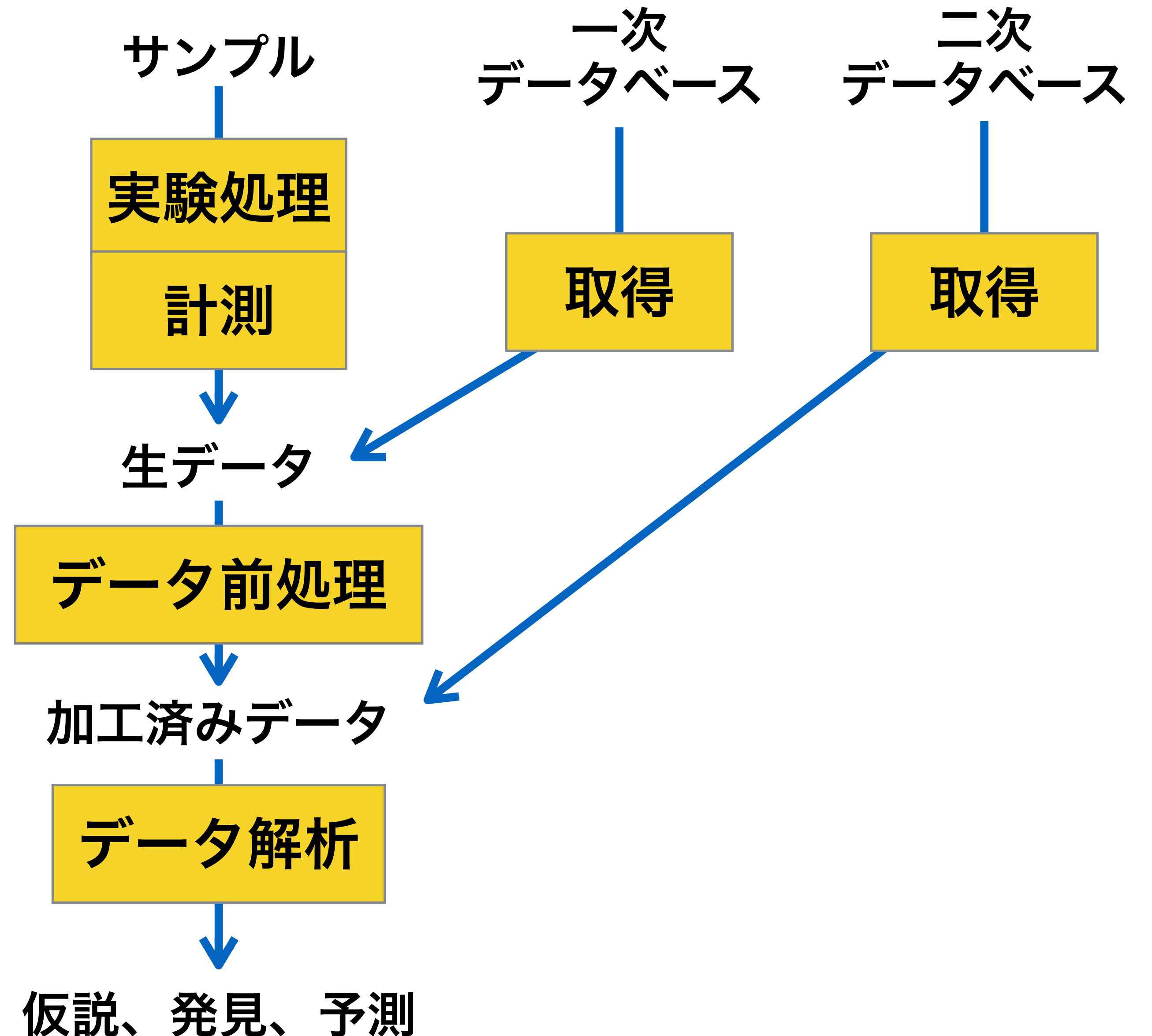
**HUMAN  
CELL  
ATLAS**



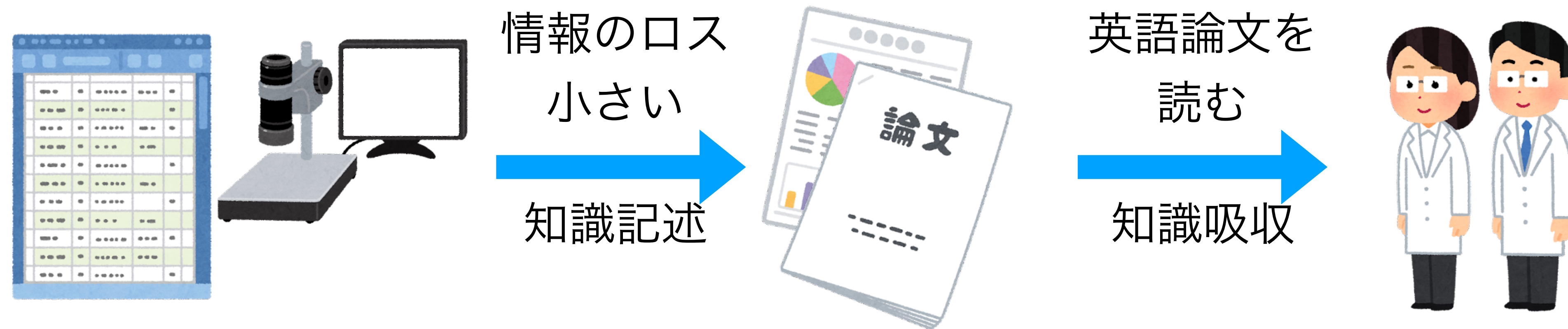


# 公共データベースによるデータの再利用

- 公共データベースから誰でもデータを取得して再利用できる
- 一次データベース：生データを収めたデータベース
- 二次データベース：加工済みデータや解析結果を収めたデータベース



# 論文は人間が読みやすいフォーマット



元データはあくまで論文の主張の品質保証のため

# 論文にする過程で取りこぼされるものが多くなった

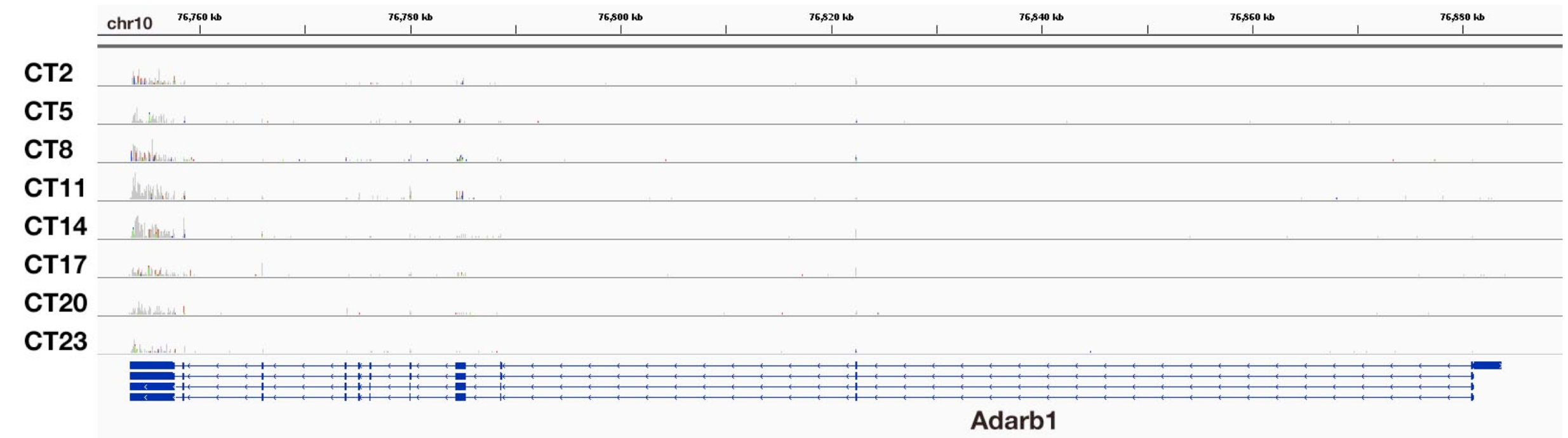


論文だけでなくデータから直接情報を得たい



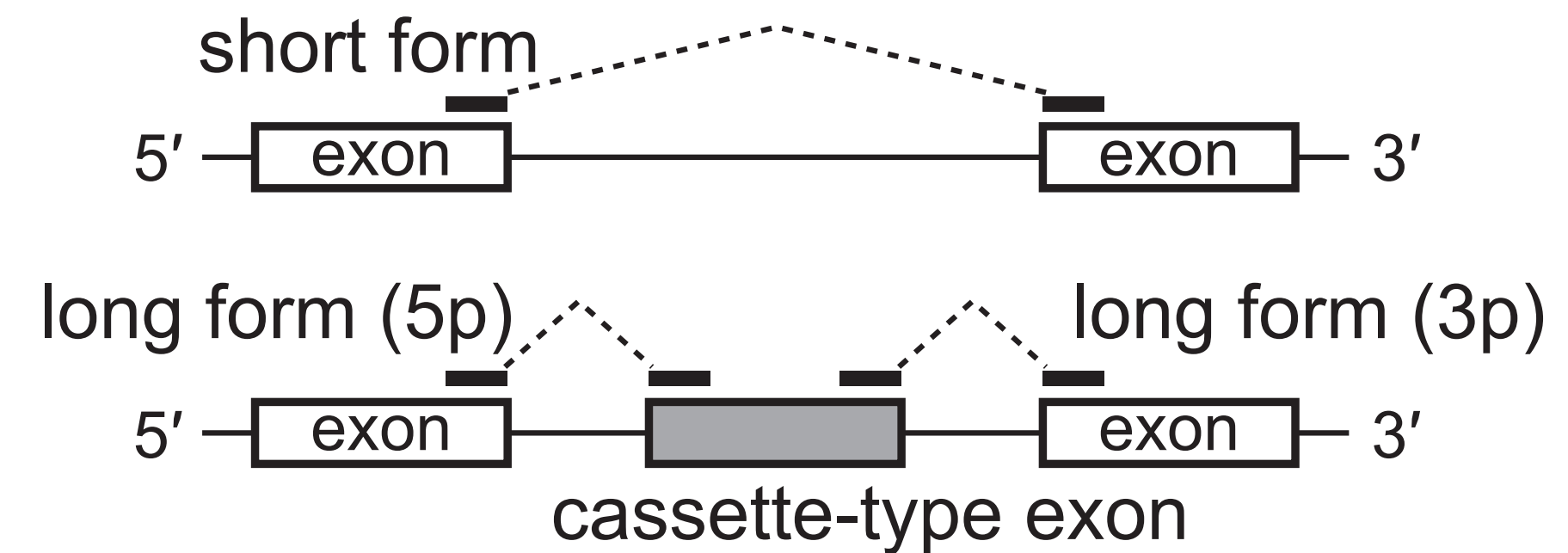
# 1つのデータからいろんな観点で解析ができる

遺伝子発現

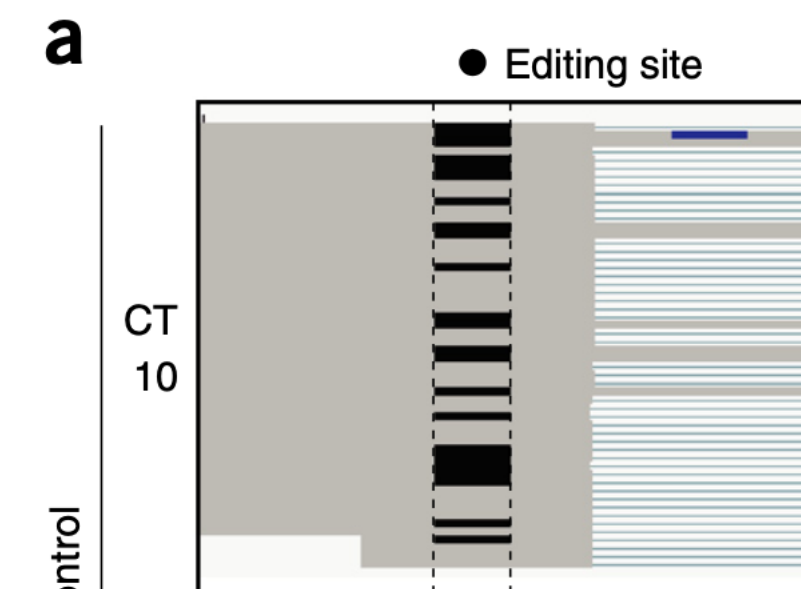


mRNA-seq

スプライシング効率



RNAエディティング





# データ解析が主戦場の生物学

- (1) 計測技術の発展、(2) データベースの整備によって、生命科学の実験データが再利用しやすい形で集積するようになった
- データが集積することで、1つのラボでは不可能な解析が可能となる
- 1つのデータを様々な視点から解析することが重要となってきた

Rについて

# なぜRが生物学研究に有用なのか

- パッケージ：まとまった機能を持ったコマンド群
- 統計検定やデータの可視化などに強いプログラミング言語
- 統計解析や可視化のパッケージが多数利用可能
  - 生命科学系データの解析に特化したRパッケージも多い：NGS解析、マススペクトロメトリー解析、サイトメトリーデータ解析、etc.
- 一般に広く使われているためサポート情報・コミュニティの情報・日本語の情報が充実している

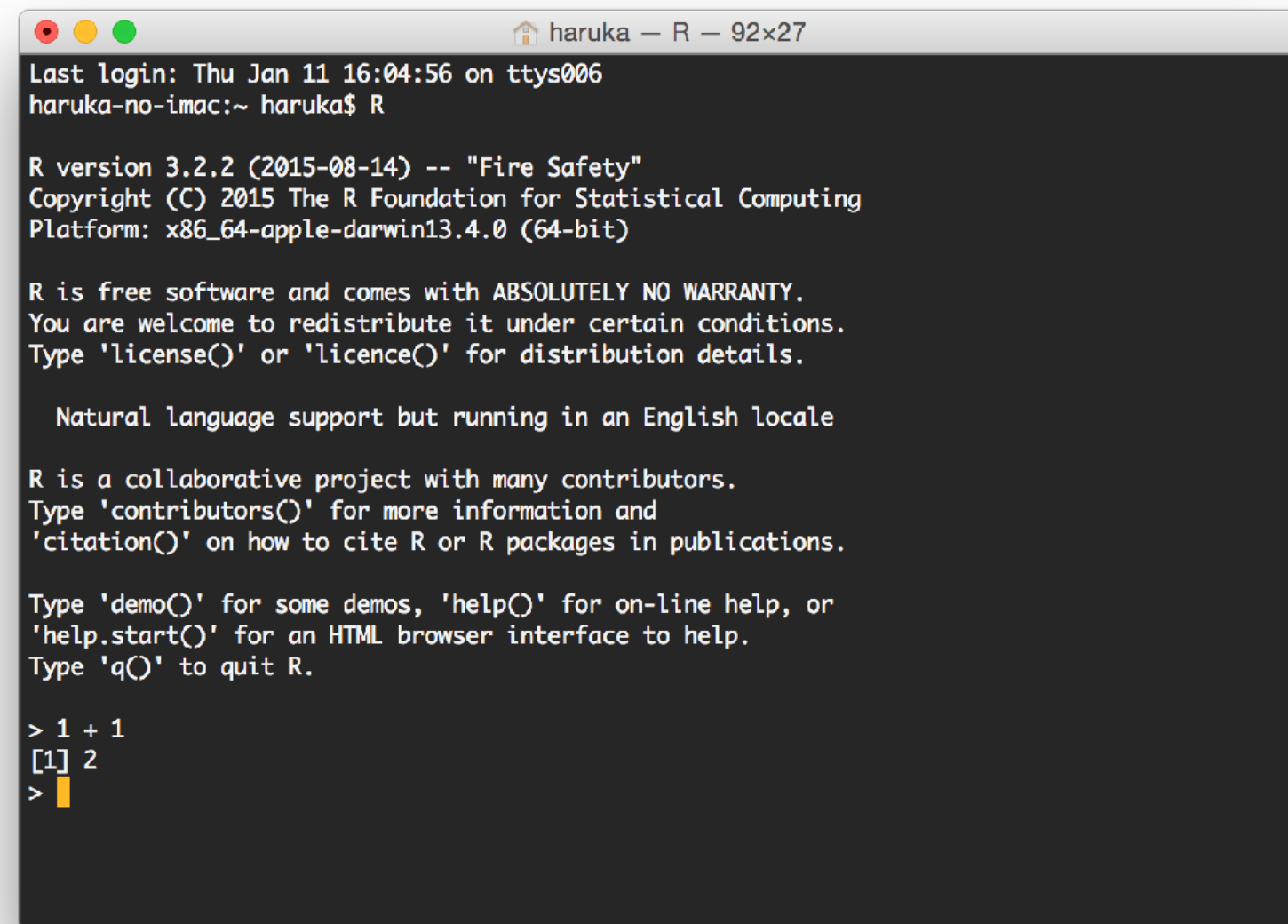
# レポジトリ：パッケージを配布するために集積する場所

- **CRAN**（しーらん）
  - Rパッケージの代表的なレポジトリ
  - `install.packages()` 関数でインストール
- **Bioconductor**（ばいおこんだくたー）
  - 生命科学系のデータ解析用のパッケージに特化したレポジトリ
  - `BiocManager::install()` 関数でインストール
- **GitHub**（ぎっとはぶ）
  - R言語に限らず汎用的なレポジトリ
  - `devtools`パッケージの `install_github()` 関数でインストール

# Rを使うインターフェース

REPL  
(対話型実行環境)

れぷる



```

haruka — R — 92x27
Last login: Thu Jan 11 16:04:56 on ttys006
haruka-no-imac:~ haruka$ R

R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

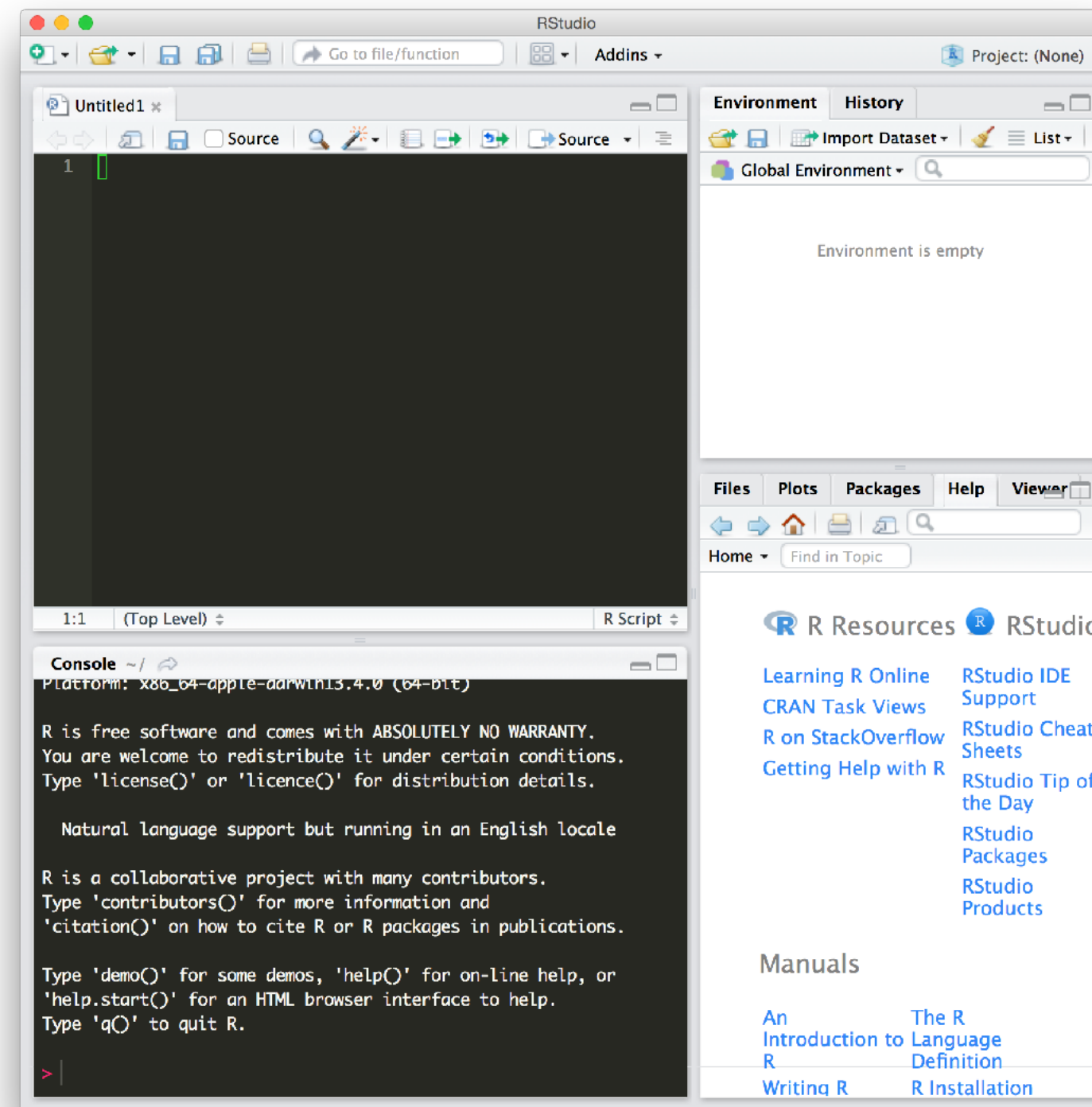
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 1 + 1
[1] 2
>
  
```

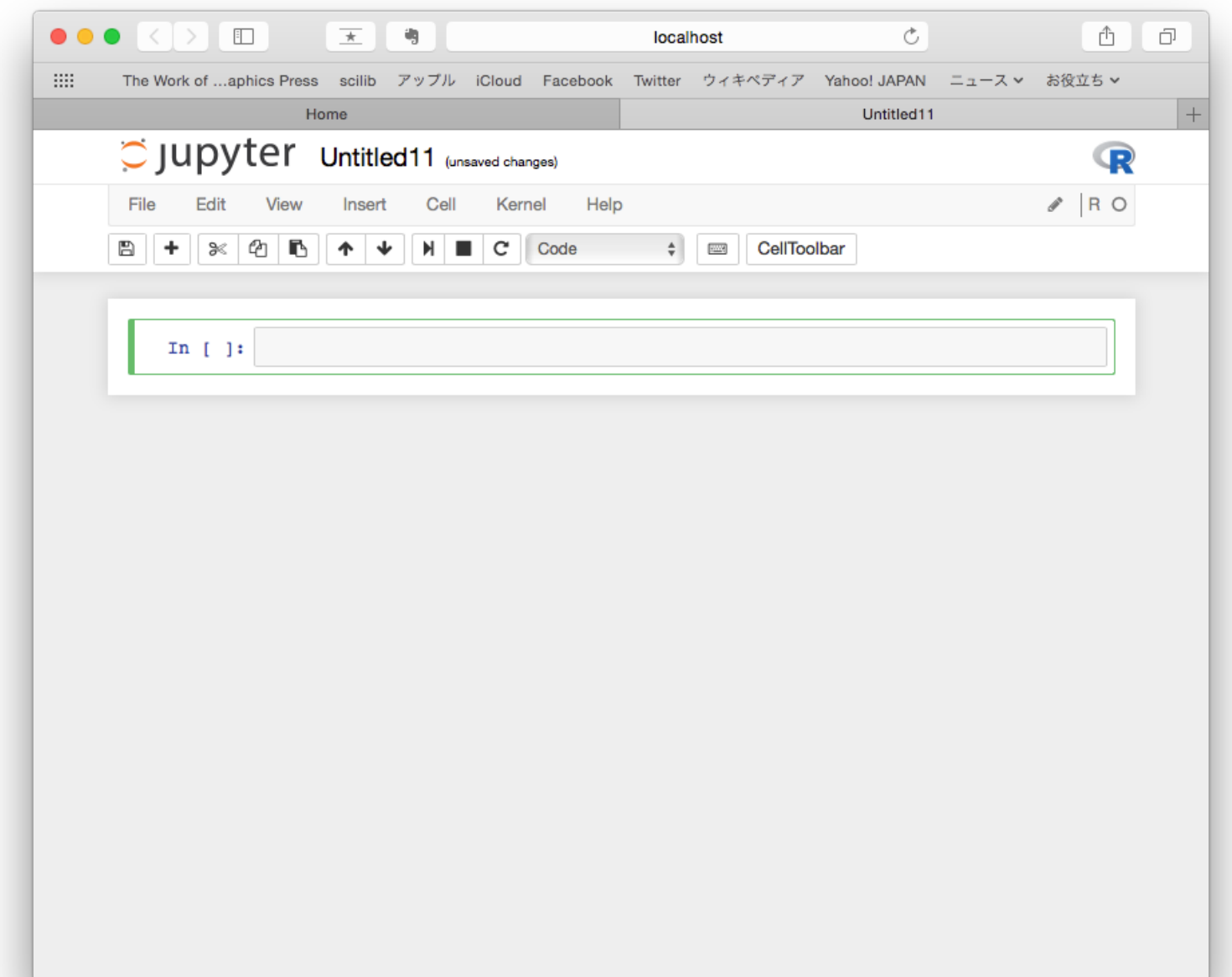
CUI

RStudio  
あーるすたじお



GUI

Jupyter  
じゅぱいたー



GUI

# 演習A

<https://github.com/bioinfo-tsukuba/FY2022-EB62104-Bioinformatics/tree/main/%E6%BC%94%E7%BF%92A>