

# 06 ゲノムデータ解析

尾崎 遼（バイオインフォマティクス研究室） <https://sites.google.com/view/ozakilab-jp>

# この講義の内容

- NGSデータ解析、特に、RNA-seqデータ解析について概略
- 実習の準備をする
  - がんのRNA-seqデータを統合して、間質細胞から腫瘍細胞への細胞間相互作用についての仮説を導出する
- JOIN について学ぶ

# NGS解析

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

- サンガー法以降、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
  - 「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
- 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

# NGSの分類

- 第二世代（塩基配列決定時に電気泳動を必要としない）
  - Illumina の HiSeqシリーズ、MiSeq, NextSeq500, MiniSeq, iSeq100
  - BGI のBGISEQシリーズ
  - Qiagen の GeneReader
- 第三世代（鋳型のPCR増幅が必要ない）
  - PacBio の PacBio RS, Sequel
- 第四世代（蛍光色素を使わない）
  - Oxford Nanopore MinION, GridION X5, PromethION RnD

# NGS＝汎用超並列塩基配列決定装置

- NGSはライブラリDNAを入れるとシーケンシングする
- → 対象とするDNAやRNAをライブラリDNAに変換できれば、多様な生命現象を網羅的に測定可能

# NGS解析の流れ

実験

情報解析

サンプル

ライブラリ調整  
Library preparation

シーケンシング  
Sequencing

データ

データ前処理  
Data preprocessing

データ解析  
Data analysis

知識

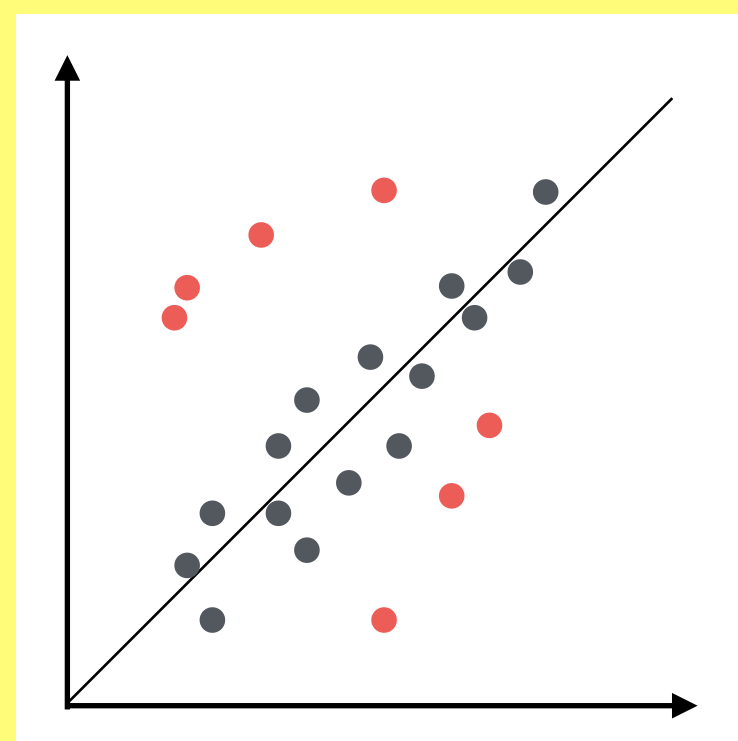
目的別の  
シーケンシング手法

NGSで塩基読み取り

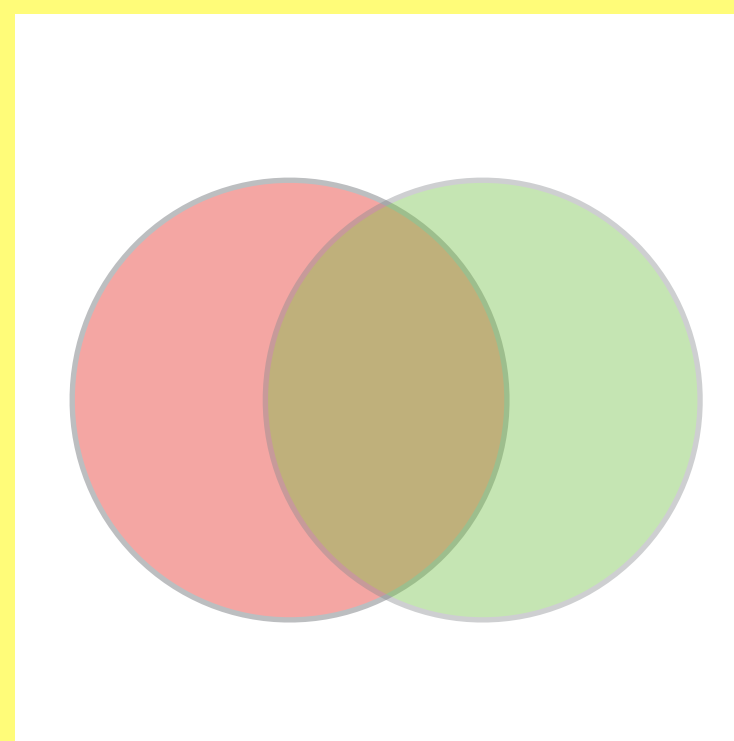
共通の前処理  
目的別の前処理

目的別の解析

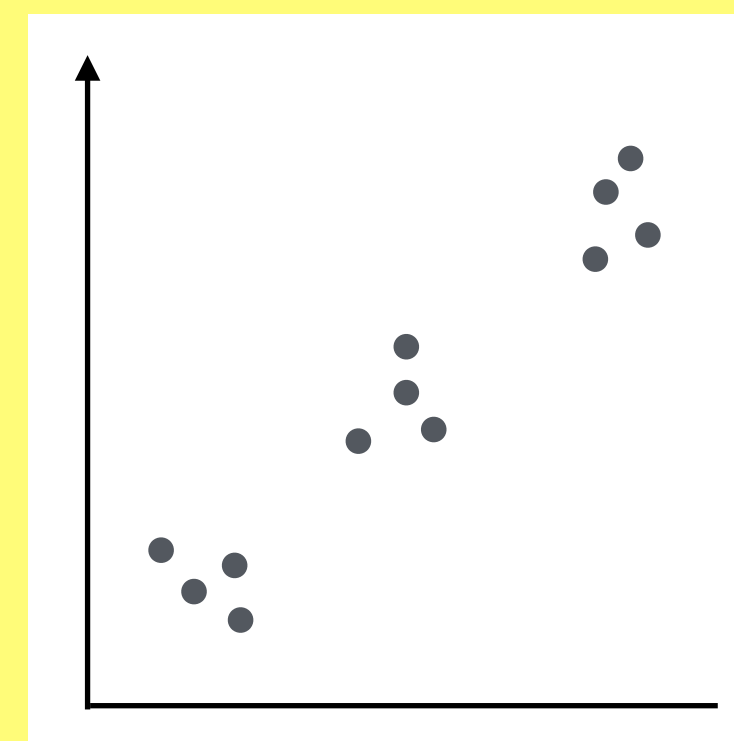
変動パターン解析



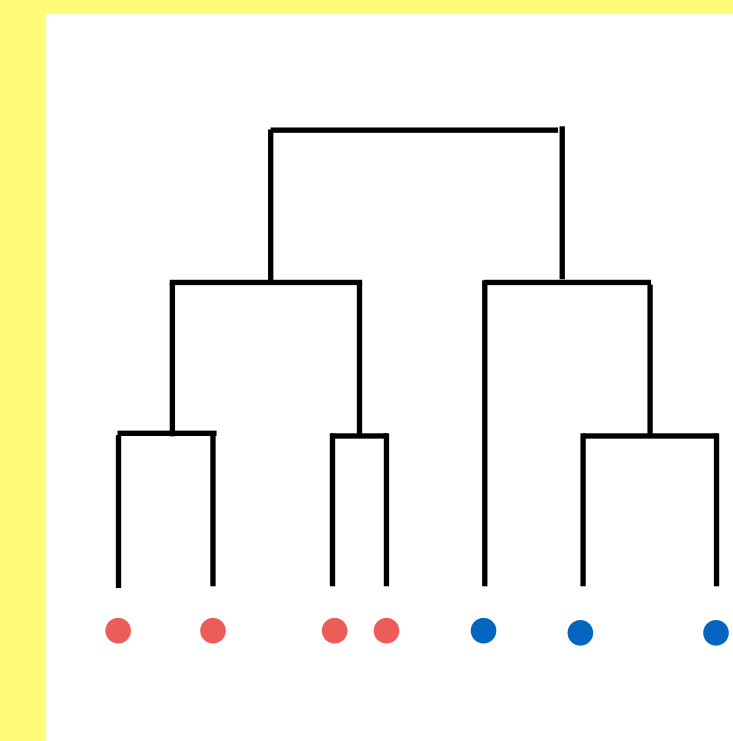
エンリッチメント解析



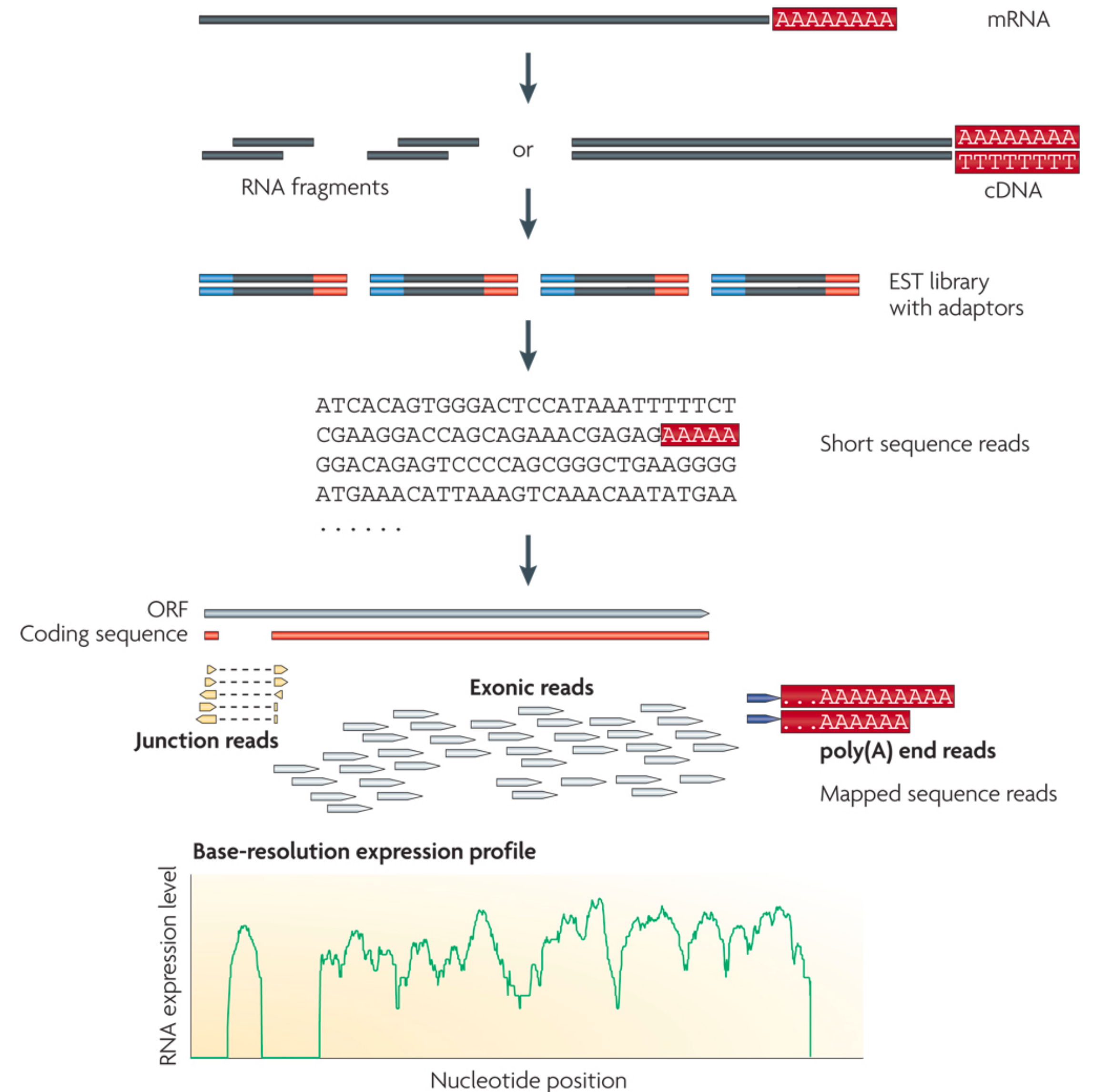
関連性解析



クラスタリング



# RNA-seq: mRNAを試料とし、遺伝子発現量を定量



Zhong Wang et al., RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet. 2009 Jan; 10(1): 57–63. doi: 10.1038/nrg2484



# RNA-Seqデータの前処理の例

## ソフトウェア・ツール

FastQC

cutadapt

Bowtie2/STAR

RSEM

QC

Quality control

フィルタリング  
Filtering

マッピング  
Mapping

QC

Quality control

発現量推定  
Expression level  
estimation

FASTQ

FASTQ

BAM

テキスト（表形式）

## ファイル形式

# FASTQフォーマット

- NGSから出力される一次データ（primary data, 生データ）はこの断片化された塩基配列（リード）の情報
- 拡張子: .fastq や .fq
- 4行で一つのリードを表す
  - 1. `@` + 配列のID
  - 2. 塩基配列
  - 3. `+`
  - 4. 各塩基の読み取り精度のスコア（Quality score）

```
@USSD-TL1-1227:179:C4E9UACXX:6:1101:12730:2322 1:N:0:GATCAG
CTGGAAGTGTGGAAGGGAACCTTAATCATTGAGTTTCTGTGAAGTATTTGCCATCCTAAAATCCCTGAGAGTGAAACTGTTGAATCATGCTCACTTTCTT
+
BBBFFFFFFFFFIIIIIIIIIIIIIIIIIIIFIIIIIIIIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIBFIIBFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@USSD-TL1-1227:179:C4E9UACXX:6:1101:12519:2371 1:N:0:GATCAG
ATTCTCATCACGTAACACTGATGGATTCCATACCTAATTTATCAATCTAAGACATTACTGGACCACGTAACCTTACATATAACTACCTGACCATATTTTC
+
BBBFFFFFFFFFIIIIIIIIIIIFIIIIIIIIIIIFIIIIIIIIIIIIIIIIIIIIIIIIIFIIFIIIIIIIFFFFFFFIIIIIIIIIIIIIIIIIIIIII
@USSD-TL1-1227:179:C4E9UACXX:6:1101:12546:2486 1:N:0:GATCAG
GTTCCACATTGTTCTGCTGTGCTTTGTCAAATGAACCTTTATGAGCCGGCTGCCATCTAGTTTGACGCGGATTCTCTTGCCCACAATTTCGCTTGGGAA
+
BBBFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIFIIIFIBFIFIIIIIBFFIBFFIIIIIFIIIFFFBFFBFBFBFBFBFBFBFBFBFBFBFBFB
```

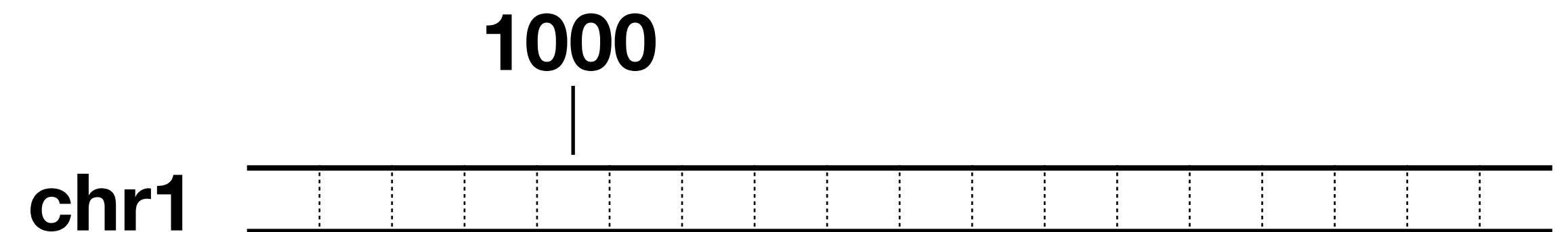
# ゲノム座標 Genome coordinate

## ゲノム配列上の位置（点）を表す

染色体名:整数（端から何塩基目か）

例（1 番染色体の1000塩基目）：

chr1:1000

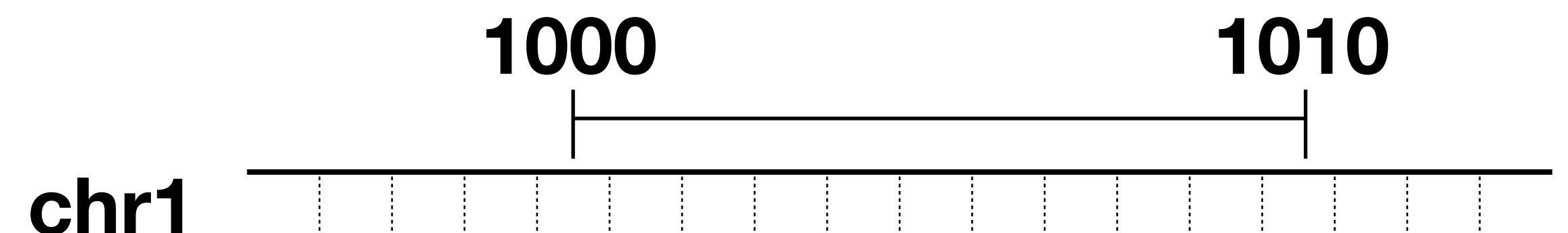


## ゲノム上の区間を表す

染色体名:整数（区間の始まり）-整数（区間の終わり）

例（1 番染色体の1000塩基目から1010塩基まで）：

chr1:1000-1010



# 遺伝子のゲノム上での表現

- ゲノム

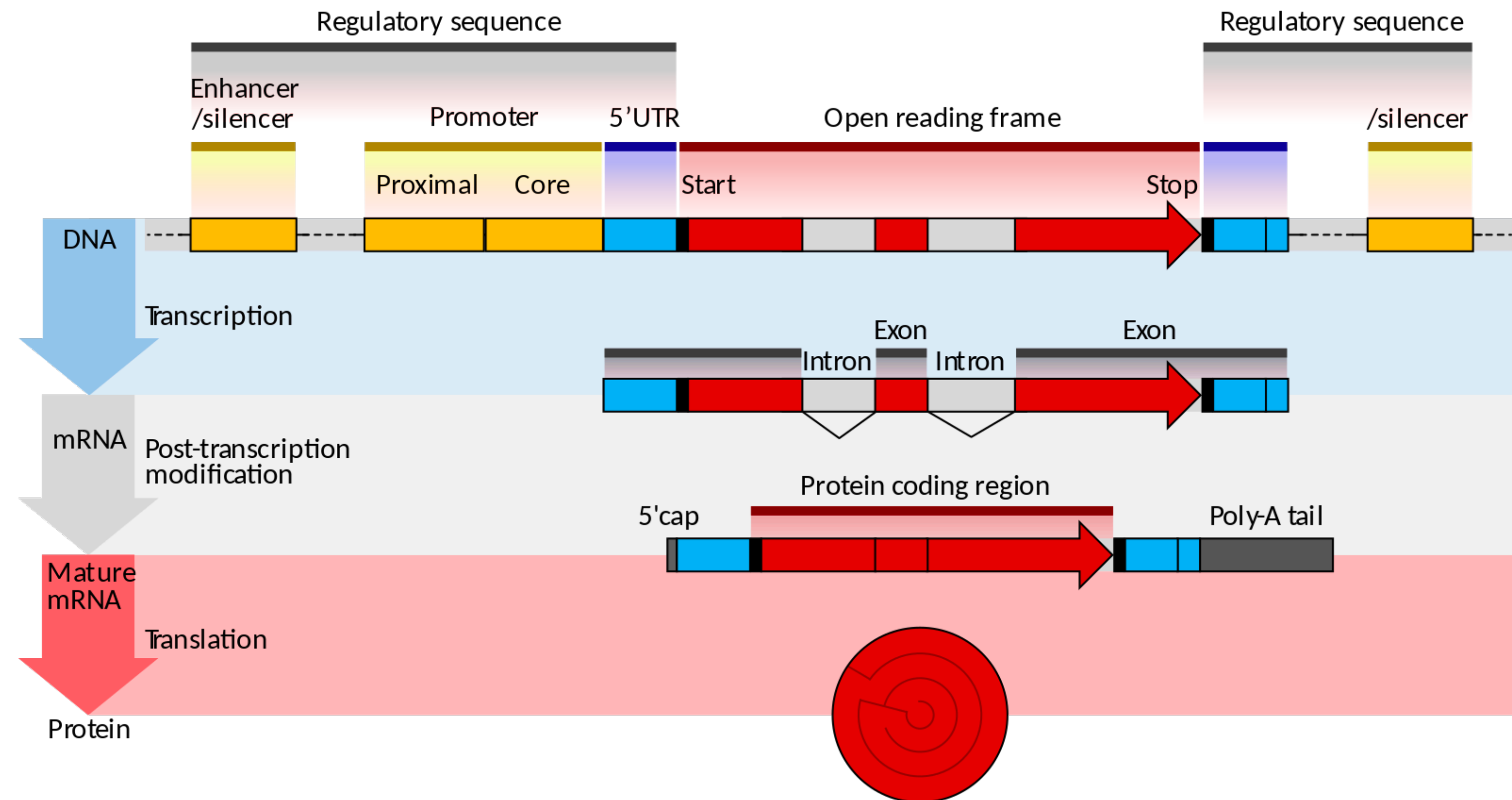
- エキソンとイントロン
- =ゲノム上の区間の集合

- pre-mRNA

- エキソンとイントロン

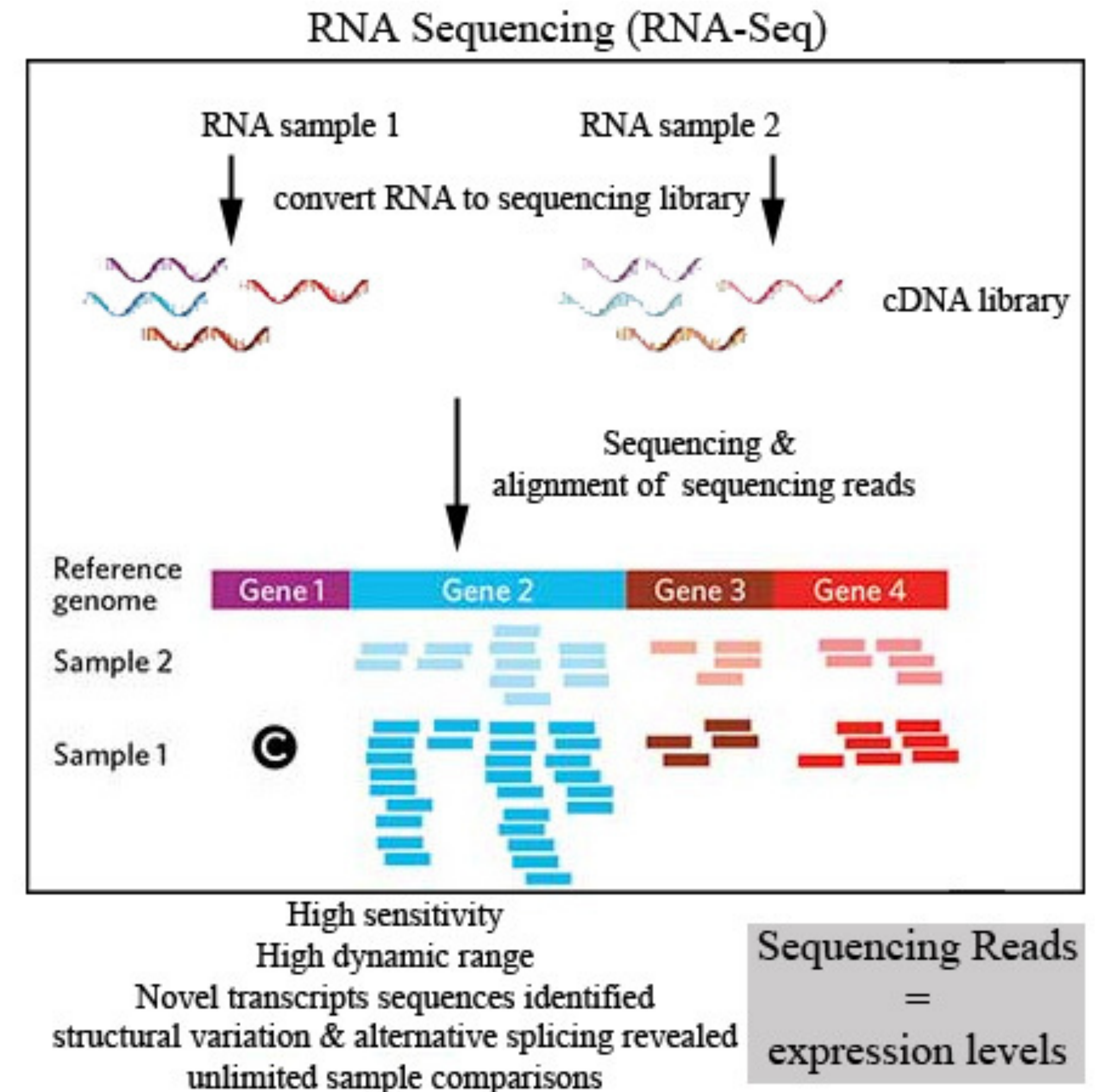
- mRNA

- エキソン



# RNA-seqデータを遺伝子発現データに変換する

- マッピング：RNA-seqのリードを、塩基配列が一致する遺伝子に割り当てる
- 発現量定量：遺伝子ごとに割り当てられたリードの数をカウントする

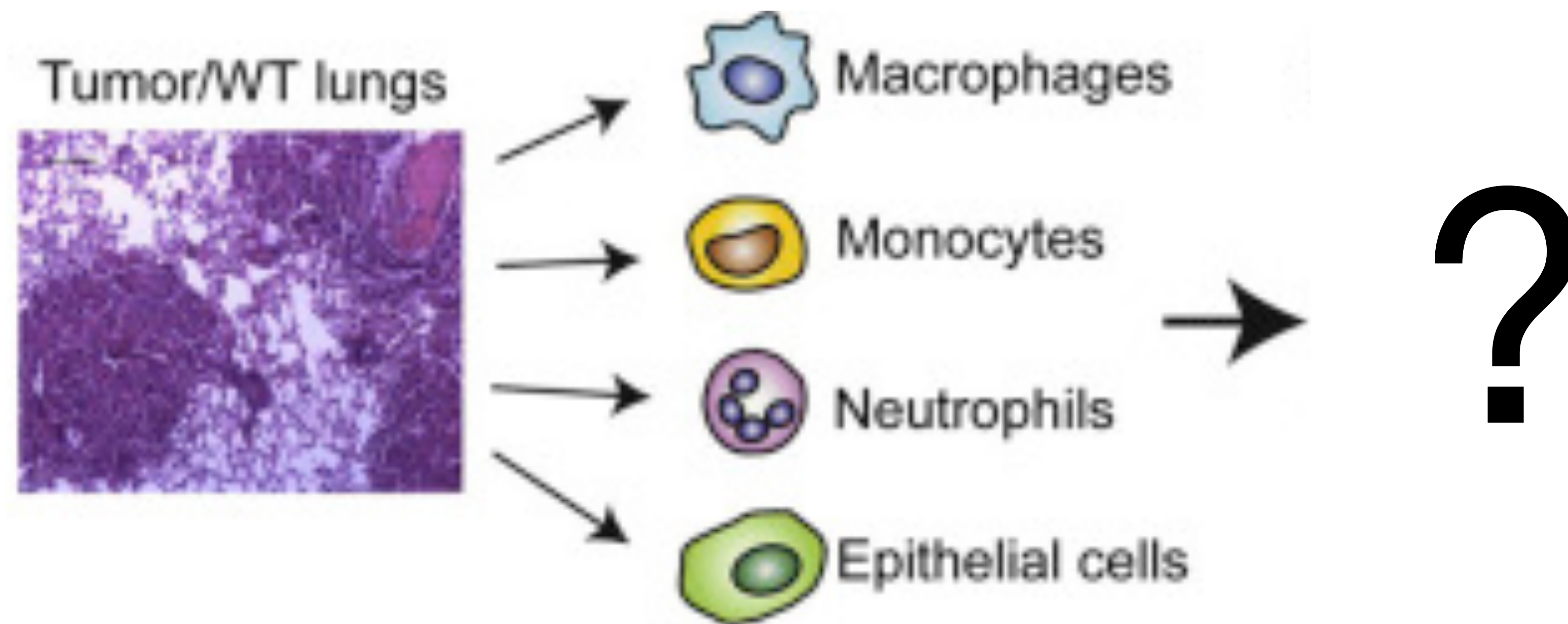


# がんのRNA-seqデータの統合



# 非小細胞肺癌 (non-small-cell lung cancer; NSCLC)

- 問い：NSCLCにおいて、腫瘍組織内に存在する間質細胞から腫瘍細胞にシグナルのクロストークが起こって腫瘍が活性化されるか？



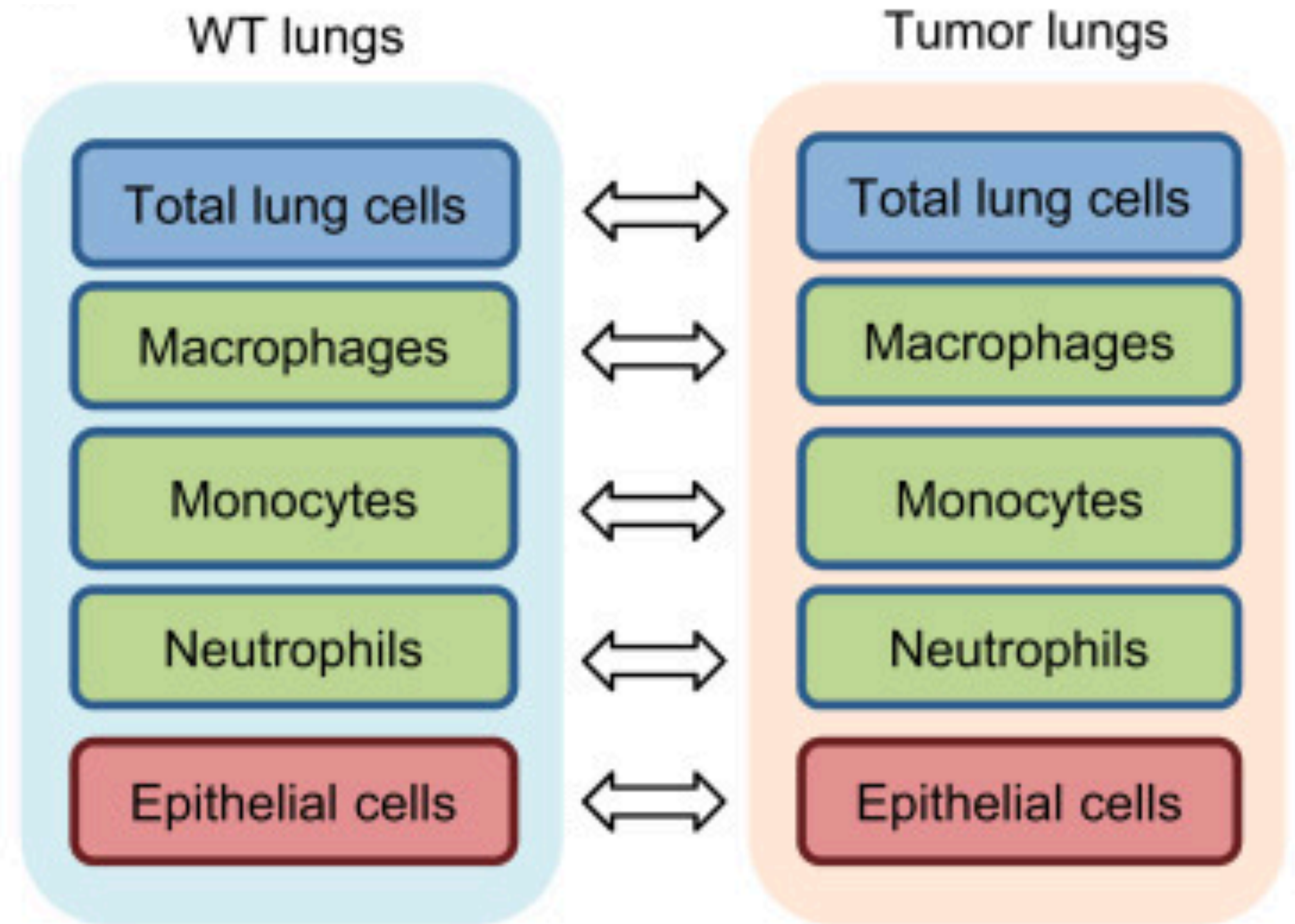
# データの出どころ

- Toi et al., Transcriptome Analysis of Individual Stromal Cell Populations Identifies Stroma-Tumor Crosstalk in Mouse Lung Cancer Model, Cell Reports (2015)
- <https://doi.org/10.1016/j.celrep.2015.01.040>

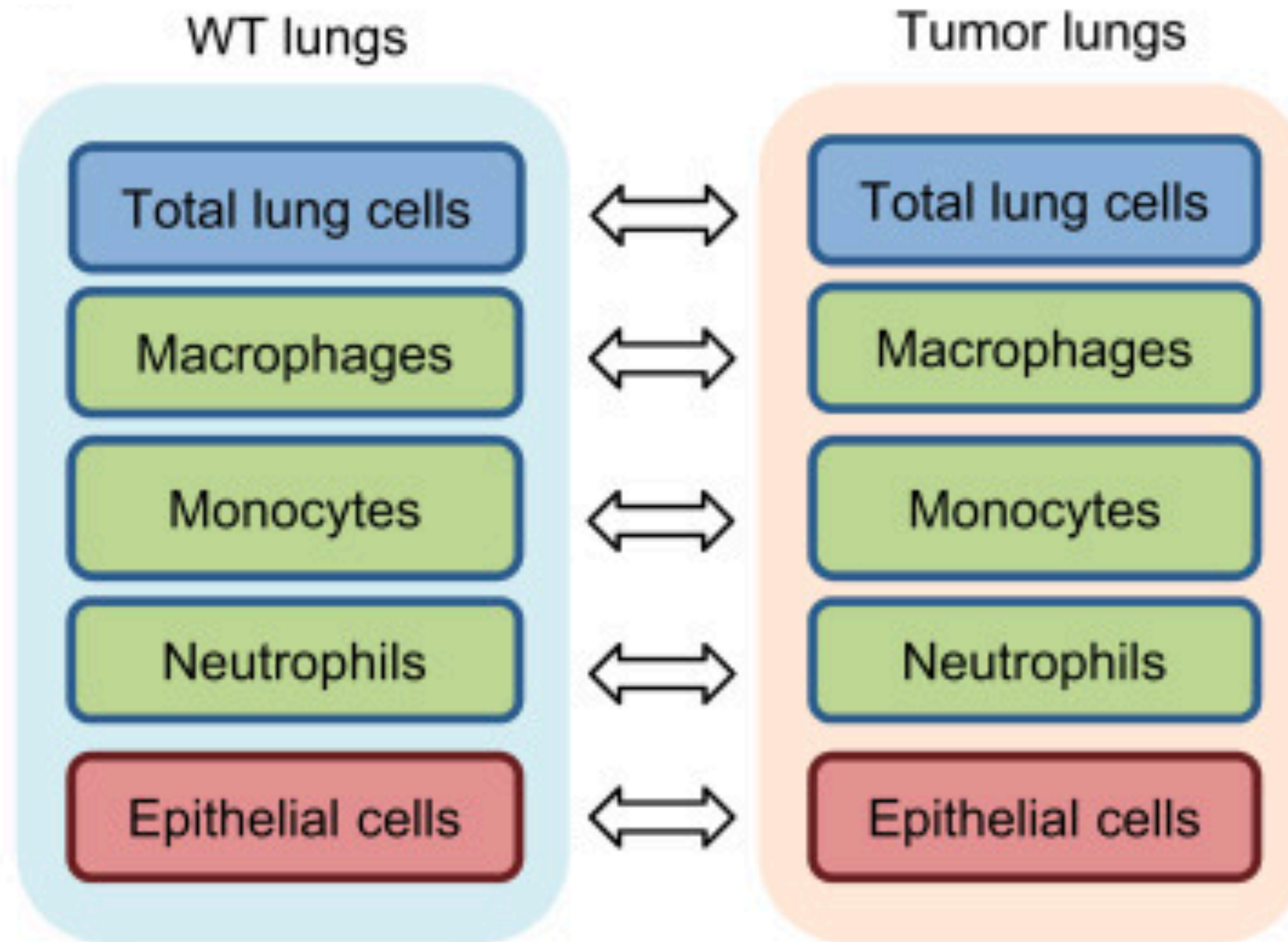


# 非小細胞肺癌 (non-small-cell lung cancer; NSCLC)

- 計測: RNA-seqデータ
- サンプル:
  - マクロファージ、単球細胞、好中球、上皮細胞 (NSCLCモデルマウスと野生型のマウスの肺からセルソーターで分けた)
  - Total lung cells
- 種: マウス



# 非小細胞肺癌 (non-small-cell lung cancer; NSCLC)

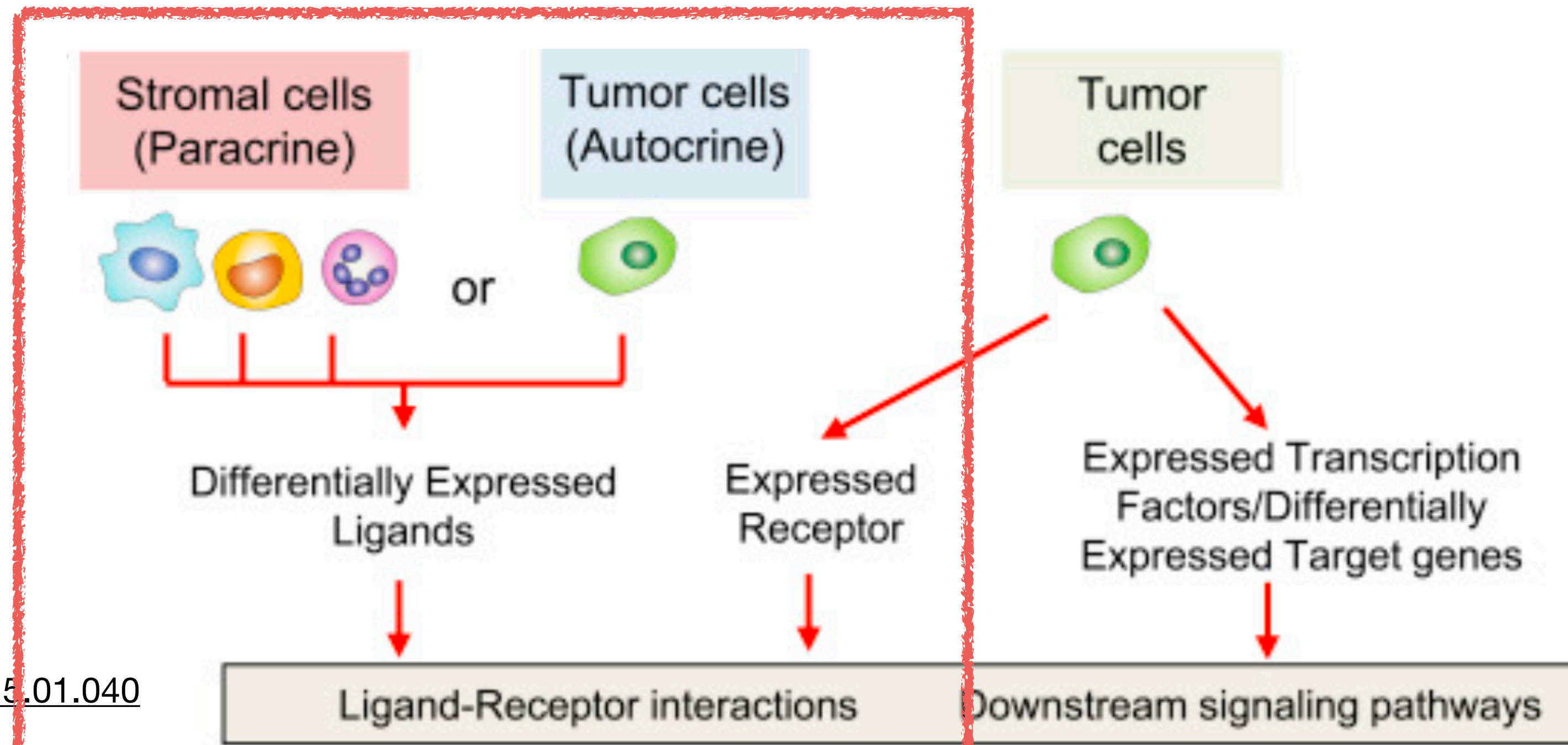


Cells	Surface markers
Total lung cells	-
Macrophages (Tumor)	CD11c+ CD11b+
Macrophages (WT)	CD11c+ CD11b-
Neutrophils	CD11b+ Ly6G+
Monocytes	CD11b+ Ly6C+
Epithelial cells	Epcam+ CD11c-



# 方針

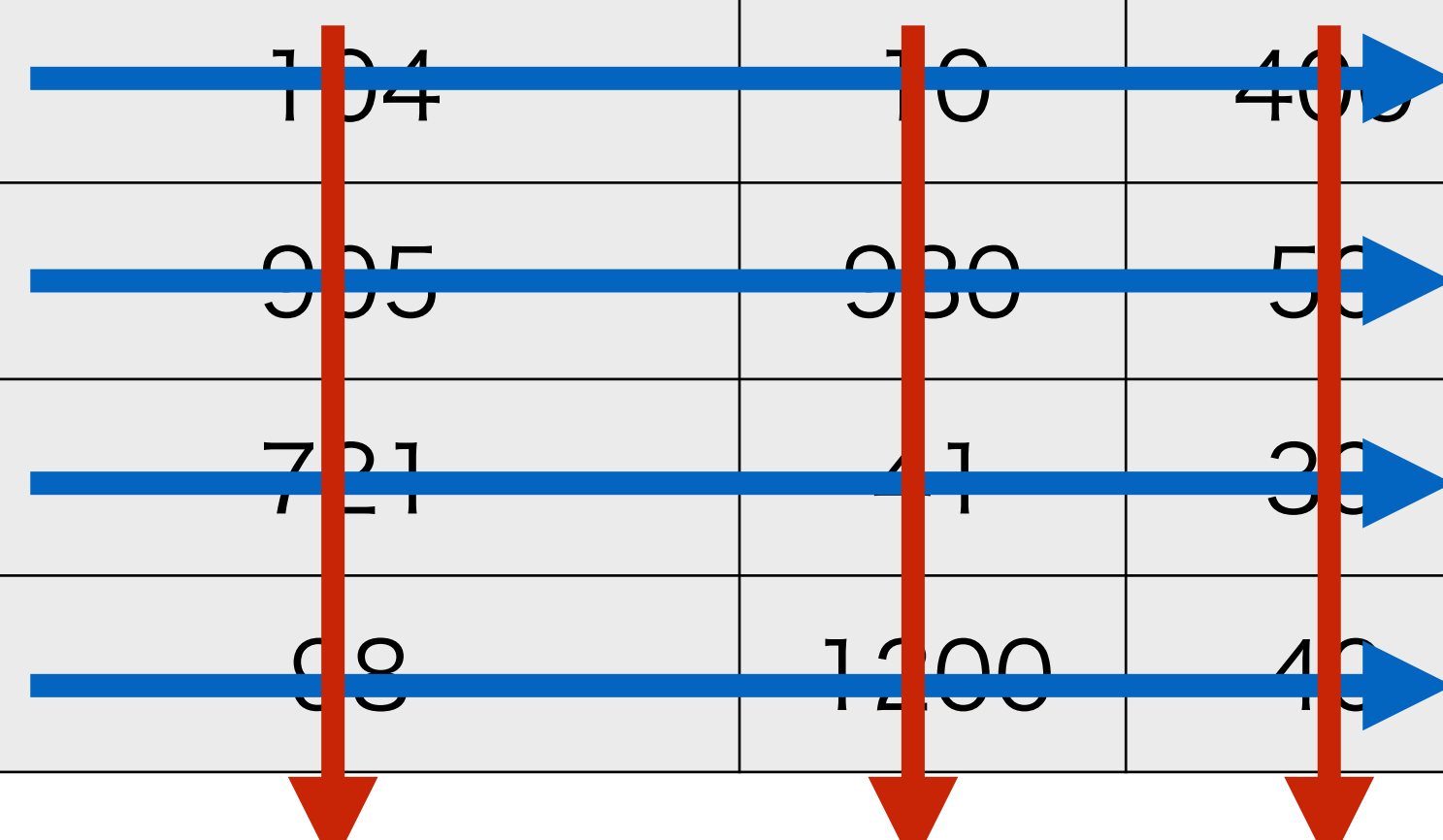
- 1. 間質細胞において「正常組織由来のサンプル」に比べて「腫瘍組織由来のサンプル」で発現量が増加している遺伝子群を抽出する
- 2. 腫瘍組織において正常組織において発現量が増加している遺伝子群を抽出する
- 3. 「1.に含まれるリガンド」と「2.に含まれる受容体（レセプター）」をリガンド・受容体のペアのデータベースと比較する



JOINについて

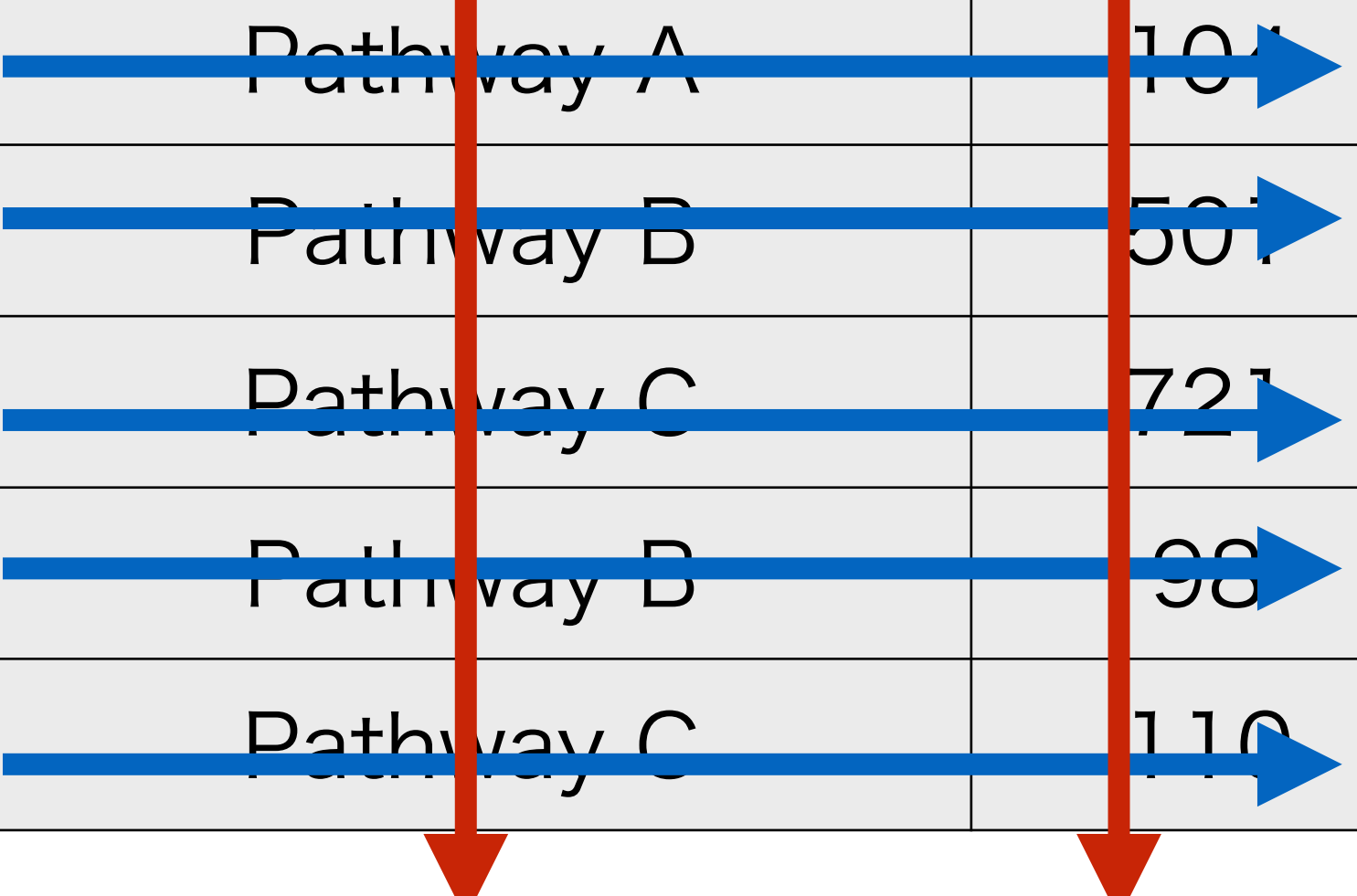
# 表形式データについて

Gene ID	Case	Control
104	10	400
905	900	50
721	41	30
99	1200	40



行：サンプル

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110



列：変数（属性）

# JOIN は2つの表を結合すること

Gene ID	Case	Control
104	10	400
905	980	50
721	41	39
98	1200	40

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110

キー：2つの表に共通に存在し、  
結合するために使う列

# 2つの表をどう結合するか

Gene ID	Case	Control
104	10	400
905	980	50
721	41	39
98	1200	40

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110

**Left join**

**Right join**

**Inner join**

**Outer join**

# Inner join: 両方の表に存在するキーの行だけが残る

Gene ID	Case	Control
104	10	400
905	980	50
721	41	39
98	1200	40

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110

Gene ID	Case	Control	Pathway
104	10	400	Pathway A
721	41	39	Pathway C
98	1200	40	Pathway B

欠損値



# Left join: 左の表に存在する行が残る

Gene ID	Case	Control
104	10	400
905	980	50
721	41	39
98	1200	40

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110

Gene ID	Case	Control	Pathway
104	10	400	Pathway A
905	980	50	-
721	41	39	Pathway C
98	1200	40	Pathway B

欠損値

# Right join: 右の表に存在する行が残る

Gene ID	Case	Control
104	10	400
905	980	50
721	41	39
98	1200	40

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110

Gene ID	Case	Control	Pathway
104	10	400	Pathway A
721	41	39	Pathway C
98	1200	40	Pathway B
501	-	-	Pathway B
110	-	-	Pathway C

欠損値

# Outer join: 少なくとも一方に存在する行が残る

Gene ID	Case	Control
104	10	400
905	980	50
721	41	39
98	1200	40

Pathway	Gene ID
Pathway A	104
Pathway B	501
Pathway C	721
Pathway B	98
Pathway C	110

Gene ID	Case	Control	Pathway
104	10	400	Pathway A
501	-	-	Pathway B
905	980	50	-
721	41	39	Pathway C
98	1200	40	Pathway B
110	-	-	Pathway C

欠損値

# pandas で join ができる

- pandas.merge
  - [https://www.tutorialspoint.com/python\\_pandas/python\\_pandas\\_merging\\_joining.htm](https://www.tutorialspoint.com/python_pandas/python_pandas_merging_joining.htm)

```
pd.merge(left, right, how='inner', on=None)
```

DataFrame オブジェクト

DataFrame オブジェクト

join の種類  
(inner, left, right, outer  
のいずれか)

キーとなる列名