RefSeq の検索と RefEx

(Reference Expression dataset)

2011年 7月 15日

内 藤 雄 樹

(ライフサイエンス統合データベースセンター)

自己紹介

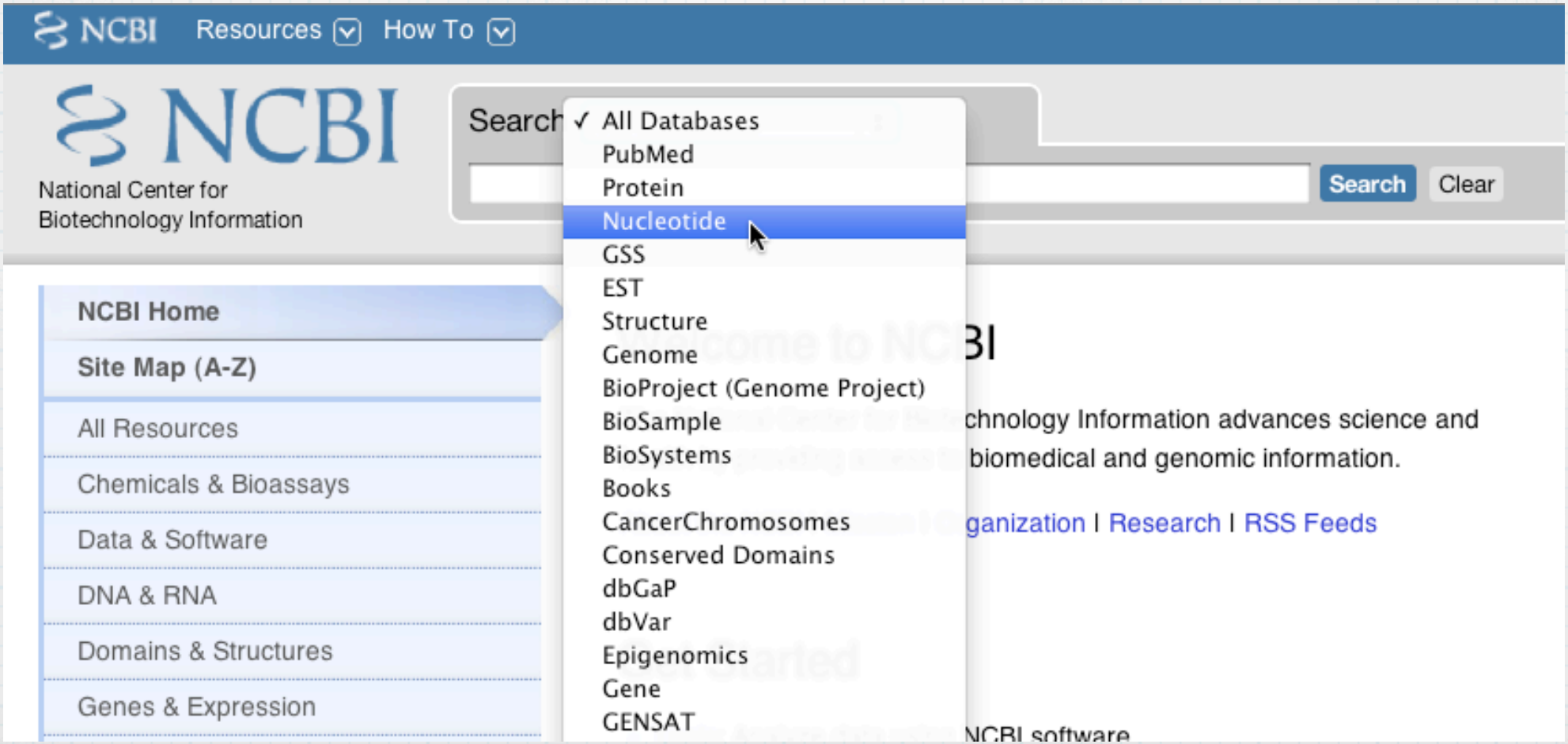
- 内藤 雄樹（ないとう ゆうき）
- ライフサイエンス統合データベースセンター（DBCLS）　特任助教
- 昨年度まで東大理・生物化学専攻でRNAi, miRNA等の研究に従事
  siRNA設計サイト: siDirect
- Twitter: @meso_cacase

今日の内容

- 遺伝子を「さがす」には？
  - 遺伝子名などキーワードで探す
  - 遺伝子のさまざまなID
  - 塩基配列から遺伝子を探す
  - 統合遺伝子検索GGRNAの紹介
- 探した遺伝子を「ながめる」
  - 発現量を調べる: RefExの紹介

遺伝子をさがす　基礎

- NCBI Entrez　http://www.ncbi.nlm.nih.gov/
  （または NCBI でググる）



絞り込み

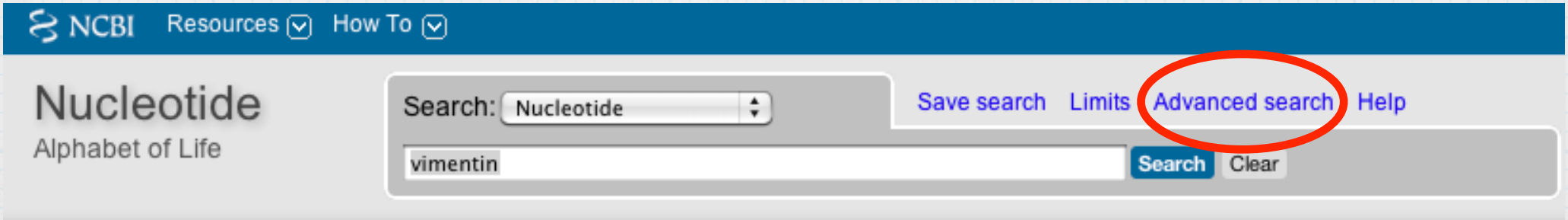
- 検索窓にキーワードを追加していく

  ... AND “Homo sapiens”[Organism]

  ... AND Vimentin[Gene Name]

  ... AND patent[Title]

- または、Advanced searchに行く



遺伝子の ID とは？

- Accession Number
- RefSeq ID
- Gene ID
- Symbol (遺伝子名)

Accession Number

- GenBank/EMBL/DDBJ の三大データバンクに登録された塩基配列のID
- A12345 や AB123456 の形をしている
- A12345.1 のようにバージョンを表示。UTRが延長されたりエラーが修正されてA12345.2 のようにアップデートされる
- GenBankのAccessionと呼ばれることも…

×

RefSeq ID

- 三大データバンクの配列を元にtranscriptごとに1個登録 → RefSeq データベース（遺伝子の百科事典のようなもの）
- 選択的スプライシングで生じるvariantには別々のIDが付与されている
- NM_012345.6 の形式をしている。広義には（実用上は）Accession番号の一種

Symbol, Gene ID

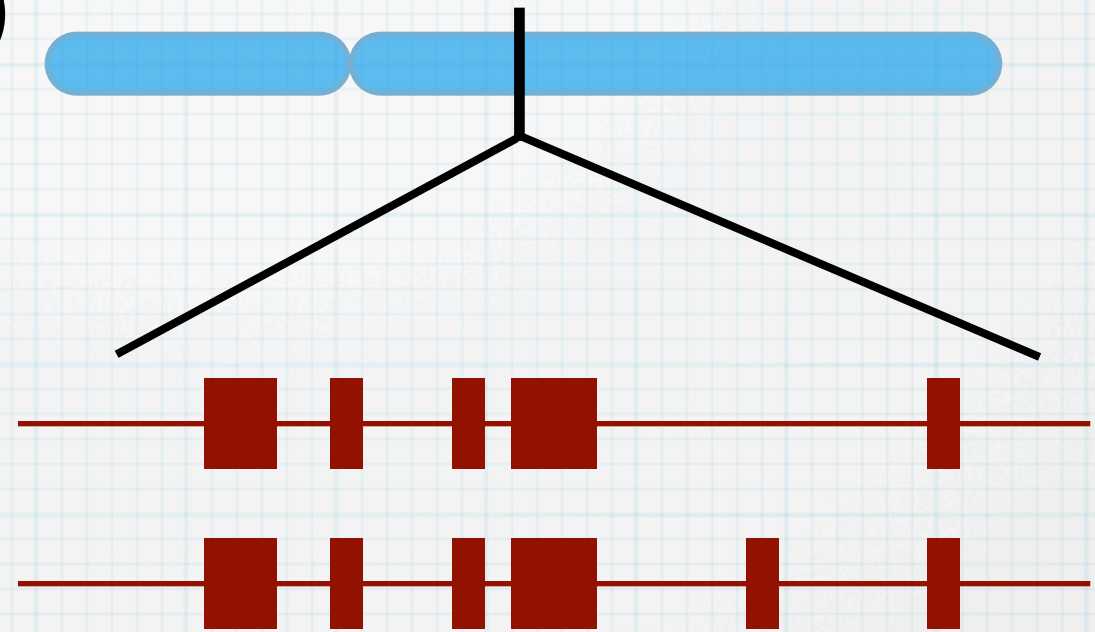
- 遺伝子ごとに付与される遺伝子名と番号

慣用名	Symbol	Gene ID
ヒトcadherin	CDH1	999
マウスcadherin	Cdh1	12550
ラットcadherin	Cdh1	83502

- Symbolは慣用名と一致しないこともあり（ヒトp53 → TP53）種でダブる可能性も
- Gene ID は生物種と遺伝子を特定できる

それぞれの関係

ヒト Chr22 (q11)



RefSeq ID:

NM_001190326

NM_022720

transcriptごと
(塩基配列ごと)

Symbol: DGCR8

Gene ID: 54487

遺伝子 (locus) ごと

配列から遺伝子をさがす

- NCBI BLAST

  http://www.ncbi.nlm.nih.gov/BLAST/
  （または BLAST でググる）

- UCSC BLAT

  http://genome.ucsc.edu/ → BLATへ
  （または BLAT でググる）

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

tgaatgaagacgatcgactcaaattcacagctccacaggatggaattcttcttaacaaagctcgacaattcgga

Query subrange

From

To

Or, upload file ファイルを選択 選択されていません

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Human genomic + transcript | Mouse genomic + transcript | Others (nr etc).

Reference RNA sequences (refseq_rna)

Organism Optional: Enter organism name or id--completions will be suggested | Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: Models (XM/XP) | Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Reference RNA sequence (refseq_rna)

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Reference RNA sequences (refseq_rna) using Megablast (Optimize for highly similar sequences)

Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Home Genomes Tables Gene Sorter PCR Session FAQ Help

C. elegans BLAT Search

BLAT Search Genome

C. elegansを選択

Genome: C. elegans | Assembly: May 2008 (WS190/ce6) | Query type: BLAT's guess | Sort output: query,score | Output type: hyperlink

tgaatgaagacgatcgactcaaattcacagctccacaggatggaattcttcttaacaaagctcgacaattcgga

submit I'm feeling lucky clear

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

File Upload: Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

Upload sequence: ファイルを選択 選択されていません submit file

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 10000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence

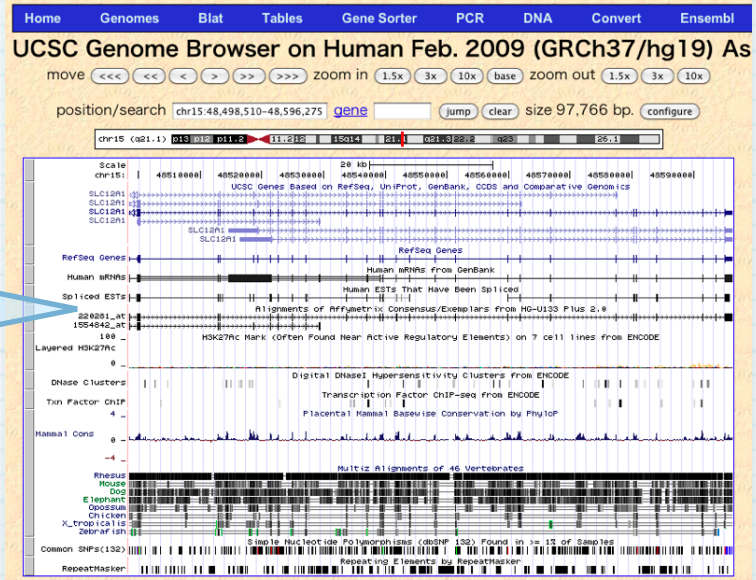
統合遺伝子検索GGRNA

http://bit.ly/GGRNA

- RefSeqを全文検索
- 塩基配列も簡単検索
  ２ミスマッチを許容
- ヒト、マウス、ラット、ニワトリ、ゼブラ、ハエ、線虫、シロイヌナズナ、出芽酵母、分裂酵母

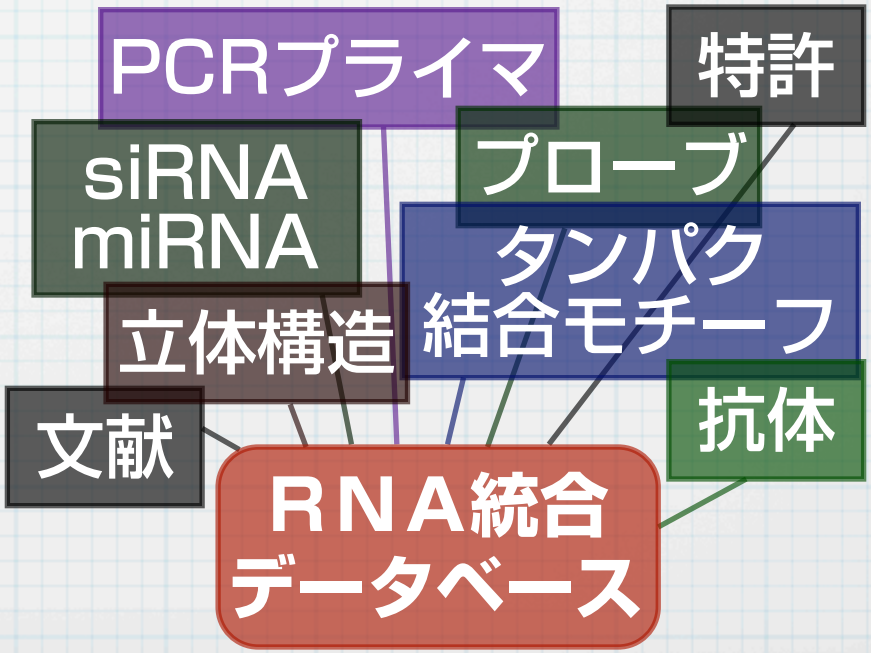
統合遺伝子検索GGRNA

従来はゲノム上に種々の情報をマッピングしていた。しかし、biologistは遺伝子や転写産物を単位にものごとを考えている。



そこで、

遺伝子（より正確にはRNA）の上に知識を統合し、簡単に**さがす** & **ながめる** ことが可能なサービスを目指す。



実習：簡単な検索例

- 遺伝子名、フレーズ、各種IDで検索
  例）claudin, “RNA interference”,
  NM_001518, 10579,
  VIM(ヒット多し)→ symbol:VIM
- 塩基配列から検索。seq1: や seq2: を付けると1, 2ミスマッチを含む配列もヒット
  → 完全一致と比べてどのくらい増える？
- さきほどの線虫の配列をさがしてみる

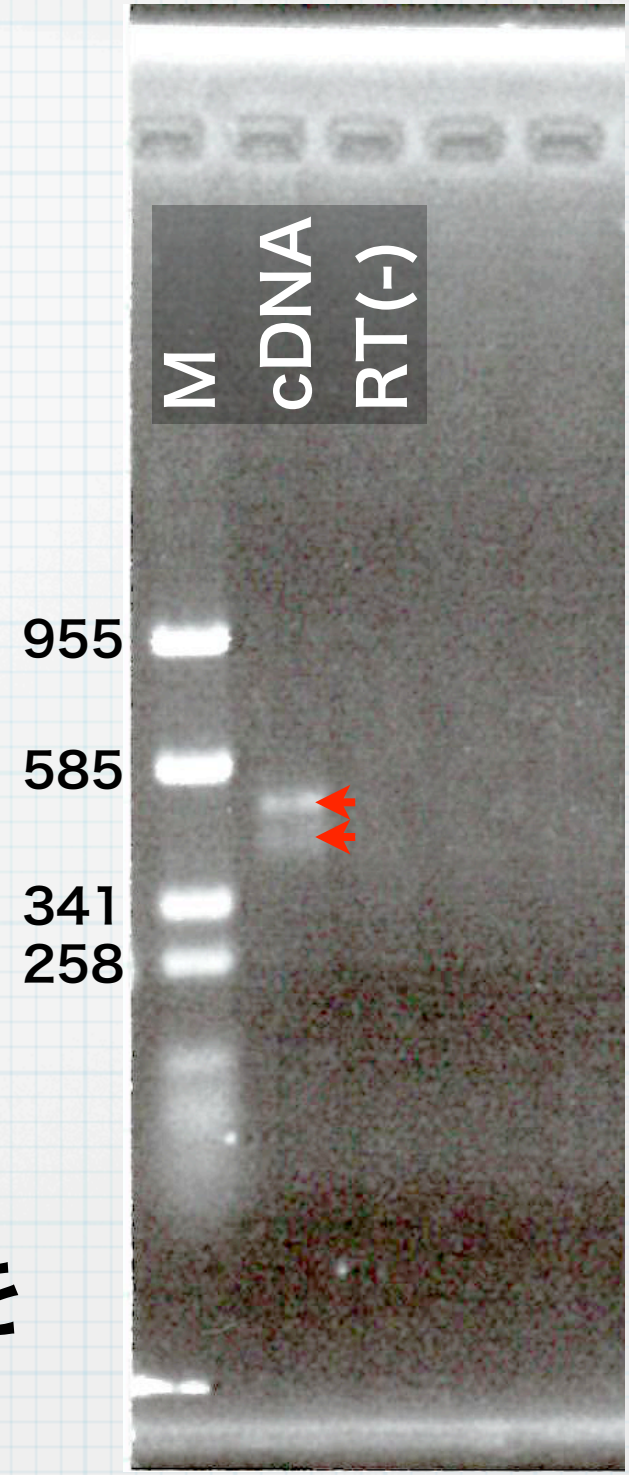
実 習

PCRプライマ配列から増幅領域とサイズを調べる

ヒトのある遺伝子に対して
RT-PCRをかけようとしたら
なぜかバンドが2本・・・

primer-F:
`agctcattactttatcagtgca`

primer-R:
`tgacgtattcactcttctggtt`



増幅遺伝子のSymbol、
Refseq ID、バンドのサイズを
調べる

※架空のデータです

実 習

遺伝子の発現量を RefExで調べてみる (GGRNA→RefEx)

RefEx

- 遺伝子発現データのリファレンスとなるデータセット
- EST/GeneChip/CAGE/RNA-seqの発現量データをまとめて一覧できる
- 組織ごとの発現量の一覧
  ヒト３Ｄ人体マップ上にも表示
- ヒト、マウス、ラット

統合遺伝子検索
GGRNA

symbol:VIM 再検索 Homo sapiens (human)

Summary:

search term:	hits:	results:
symbol:VIM	1	NM_003380
[AND]	1	NM_003380

Results:

検索語に色がつきます。重なると色が濃く表示されます。

Homo sapiens vimentin (VIM), mRNA. (2151 bp)
Summary: This gene encodes a member of the intermediate filament family. Intermediate filamentents, along with microtubules and actin microfilaments, make up the cytoskeleton. The protein encoded by this gene is responsible for maintaining cell shape, integrity of the cytoplasm, and stabilizing cytoskeletal interactions. It is also involved in the .....
NM_003380.3 - Homo sapiens (human) - NCBI - UCSC - RefEx発現量

RefEx
発現量

RefEx: Reference Expression Dataset

ヒト　マウス　ラット

トップ　発現パターンから探す　染色体領域で眺める　遺伝子ファミリーで眺める　機能で眺める　ダウンロード　ヘルプ

絞込み検索

Gene Name or Symbol　Search

" key1='Gene ID'　keyword1='7431' " での検索結果

1 遺伝子 存在します

Download results as a tsv file

#	Refseq ID	遺伝子名	最上位遺伝子ファミリー	染色体上位置	絶対発現量 (low – high) EST	Gene Chip	CAGE	RNA-seq	相対発現量 (脳 血 結 殖 筋 消 肝 肺 腎 泌) EST	Gene Chip	CAGE	RNA-seq	3Dmap
1	NM_003380	vimentin	Intermediate filament protein	10p13(+)	1.4	12.28	113,501	360					

Download results as a tsv file

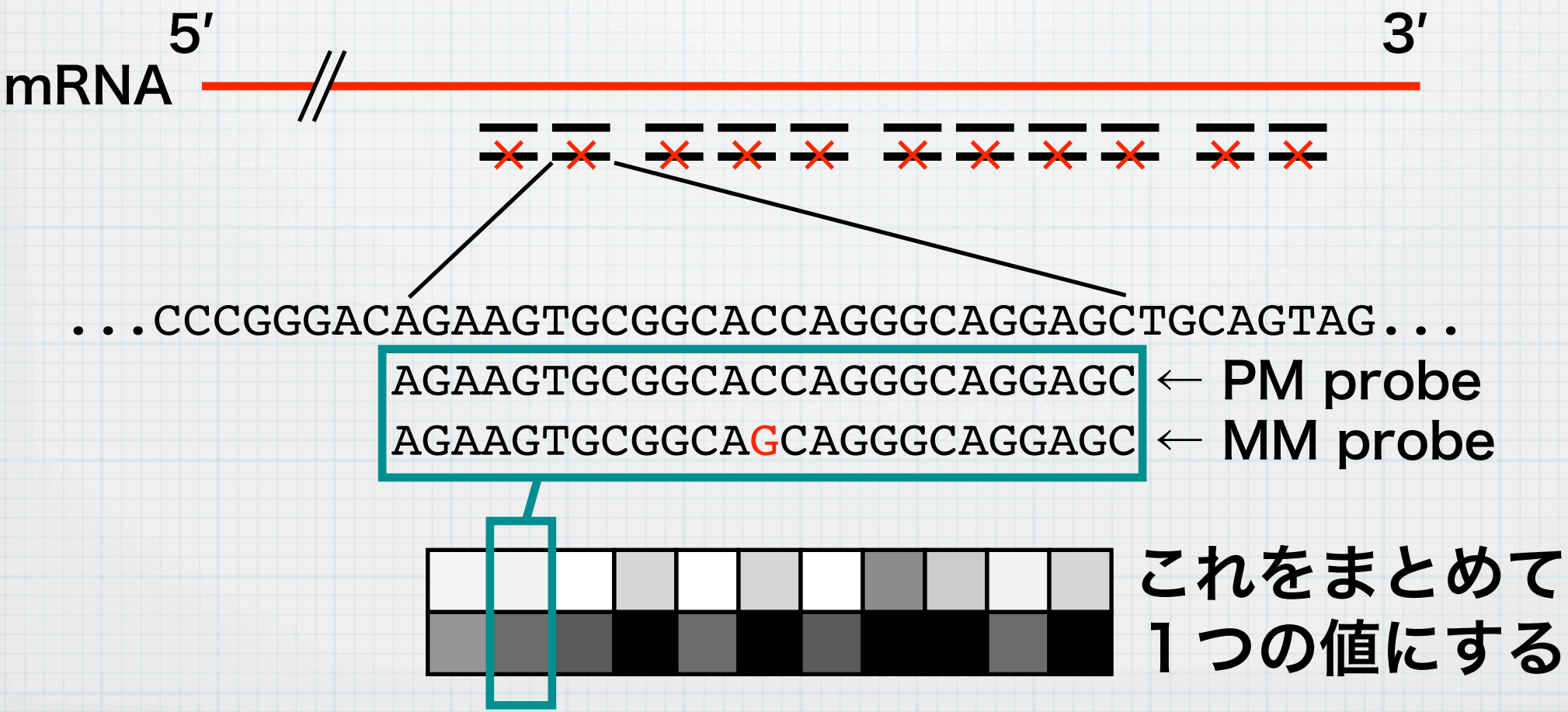
©2011 DBCLS (cc) BY

遺伝子の情報　発現量　組織ごとの発現量とヒト人体マップ

Affymetrix社GeneChipマイクロアレイ

1遺伝子につき25塩基 × 11箇所のプローブ

→ **「プローブセット」** 例) 1552311_a_at



実 習

マイクロアレイの
プローブ配列検索
(RefEx→GGRNA)