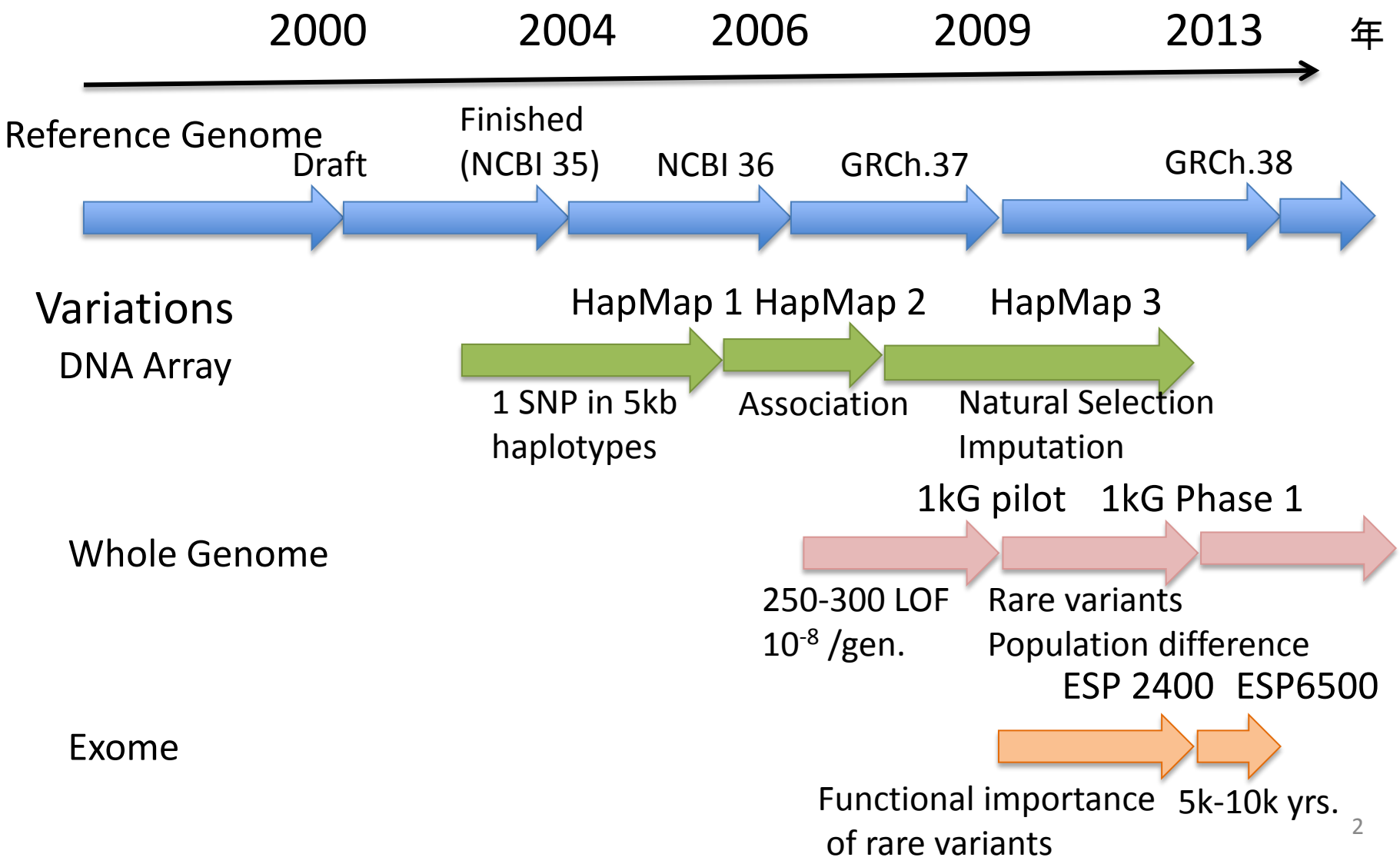
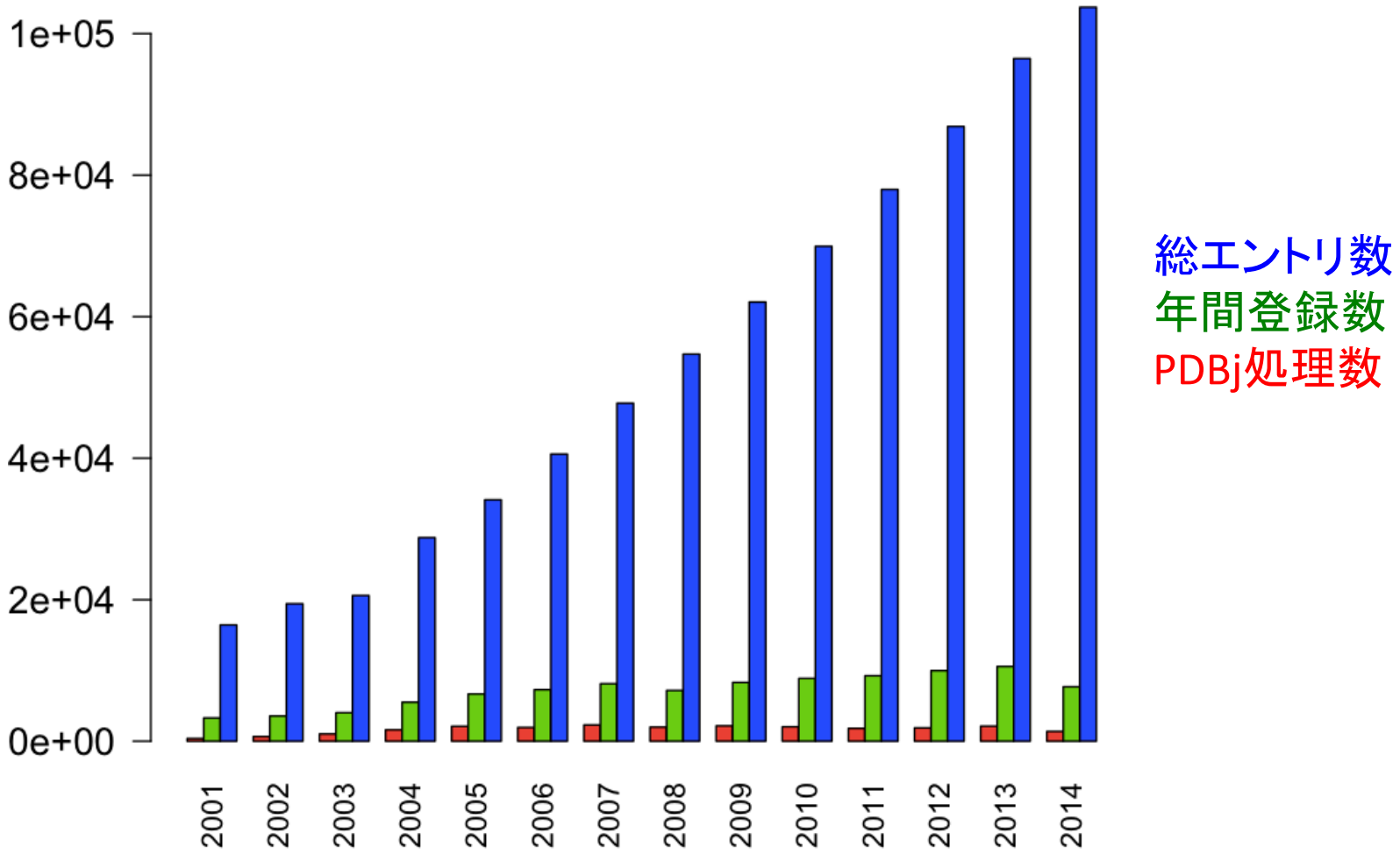
統合データ解析トライアル「PDBjタンパク質をゲノムにマップしたpdbBAMの作成」

城田松之
東北大学大学院医学系研究科
創生応用医学研究センター
新医学領域創生分野
平成26年10月8日（水）

個人レベルでの変異情報の蓄積



立体構造の数も年々増加



統計データ：http://pdbj.org/info/statistics

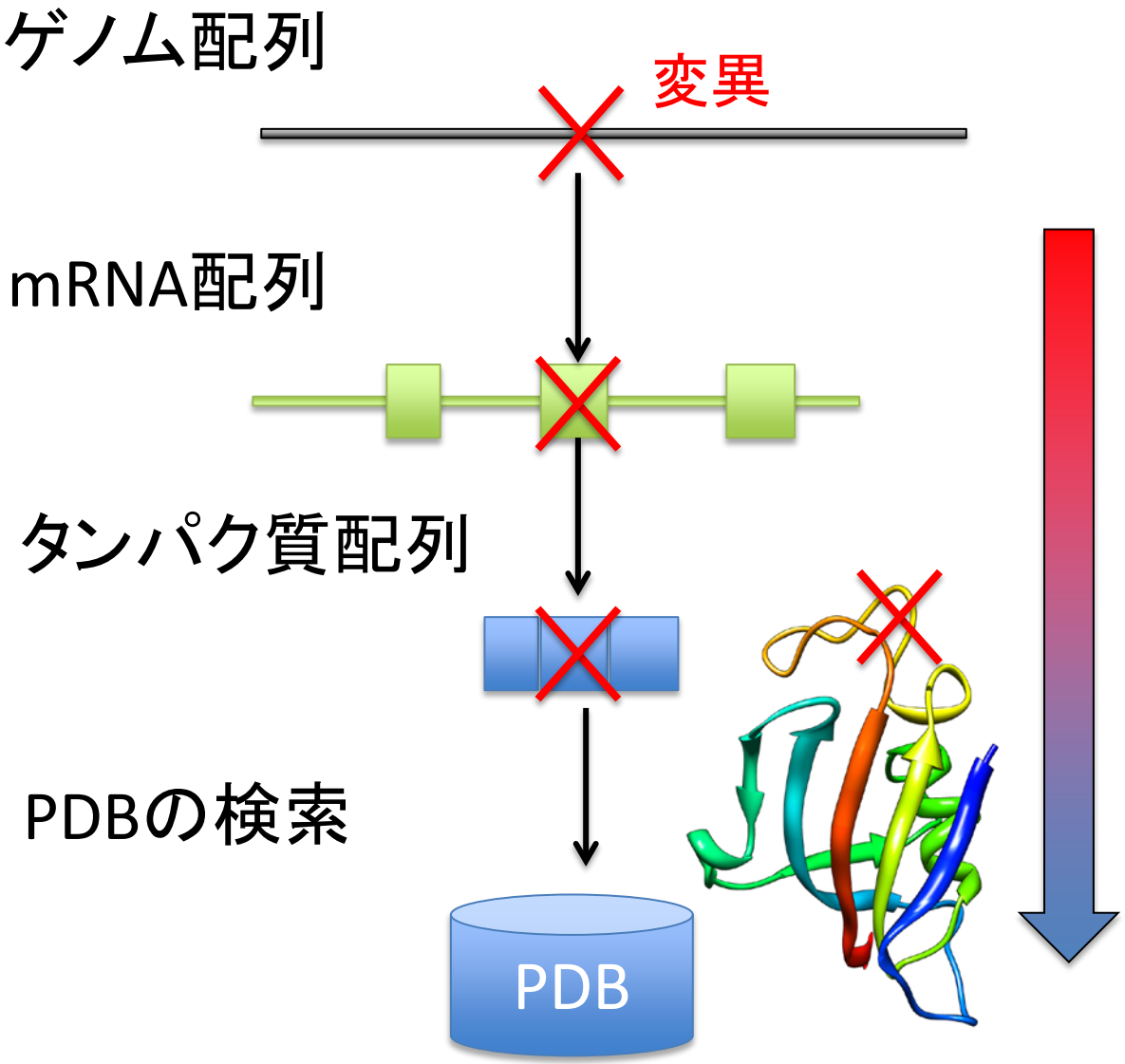
大規模シークエンシングと立体構造

- ゲノム変異の重篤度を考える上で立体構造上の位置は重要な情報となる
  - 埋もれ，表面，活性部位など

- ヒトの遺伝子で近縁種を含めて構造が見つかっているものは少ない
  - 大規模エクソーム解析の変異のうち20%弱

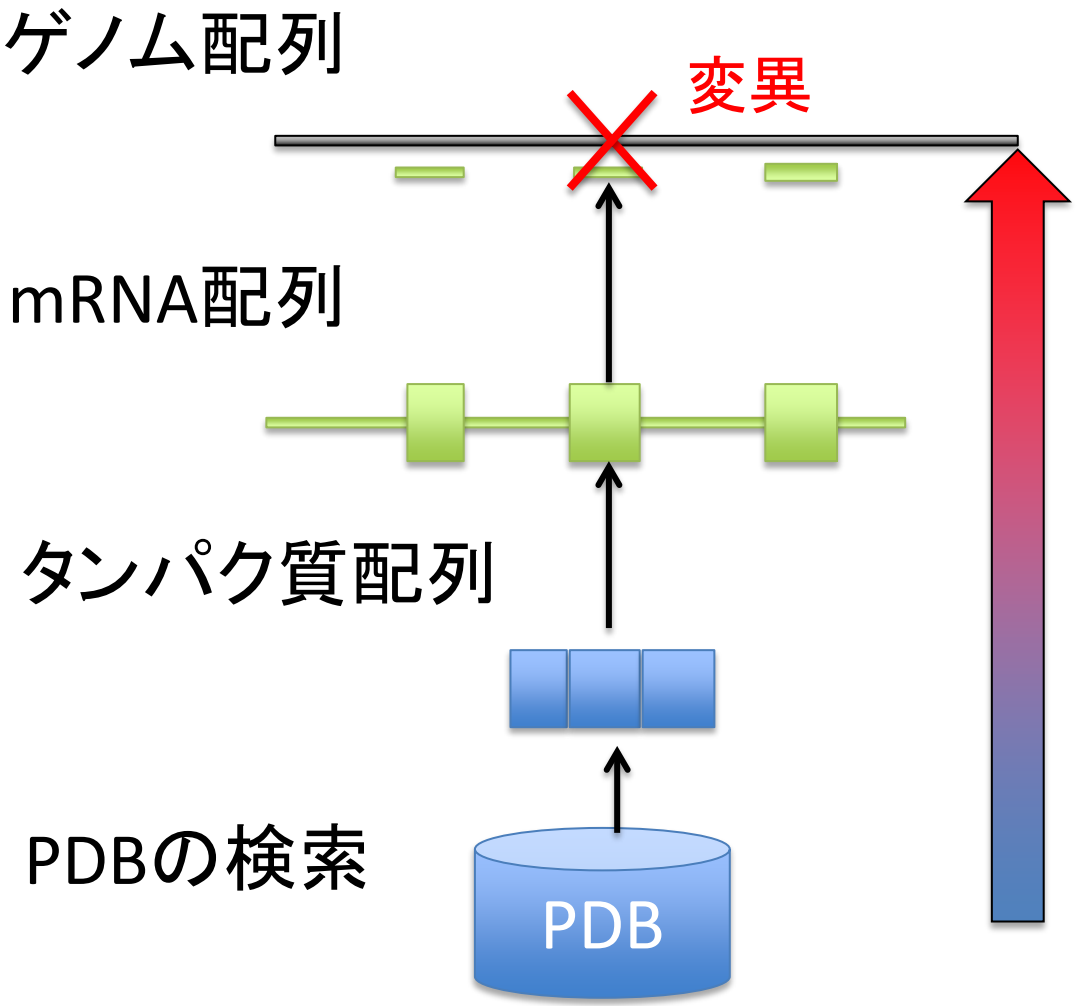
ゲノム上の変異から構造情報を検索する問題

これまでの構造情報の利用



- StSNP
- coliSNP
- MuPIT
  - など

PDB配列からゲノム配列への逆マッピング



- 利点
  - 変異ごとの検索なしでよい
  - ゲノムを見ればPDB全体が分かる
- 欠点
  - 逆翻訳は曖昧さが残る
  - PDB検索の閾値

この向きのマッピングを行うツールはまだ存在しない

SAM/BAMフォーマット

SAM: Sequence Alignment and Map format
BAM: Binary SAM

- ゲノムなどの長い参照配列（template）に対して次世代シークエンサーなどから得られる多数の短い塩基配列（read）のアラインメントを表現するフォーマット

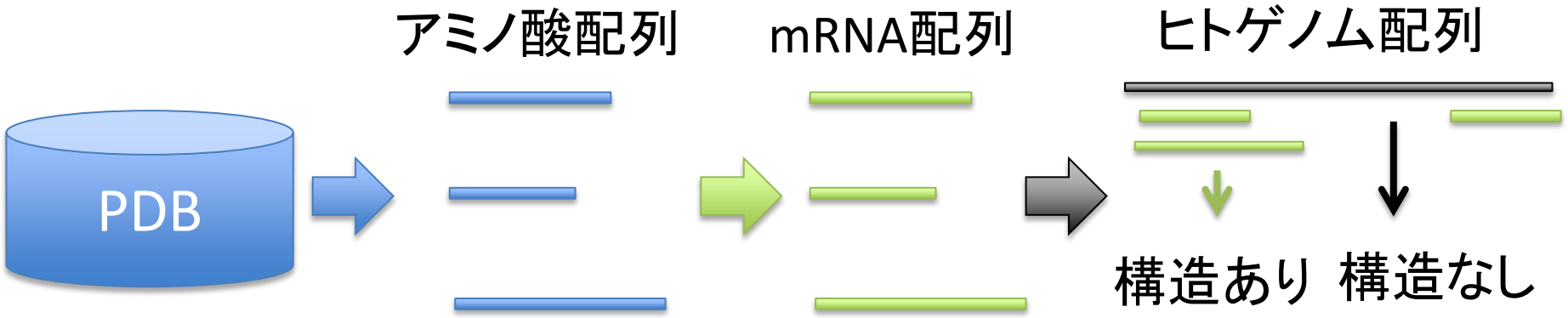
```
Coor   12345678901234 5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002          aaaAGATAA*GGATA
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *R
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                    CAGCGGCAT
```

http://samtools.github.io/hts-specs/SAMv1.pdfより改変

pdbBAMとは何か

- PDBに含まれるタンパク質のアミノ酸配列をmRNA配列を媒介してゲノムにマップしてBAM形式としたもの
- PDB全体をゲノム全体に貼付ける
- ゲノムのどこの遺伝子が構造解析されているかを一目でわかるようにする

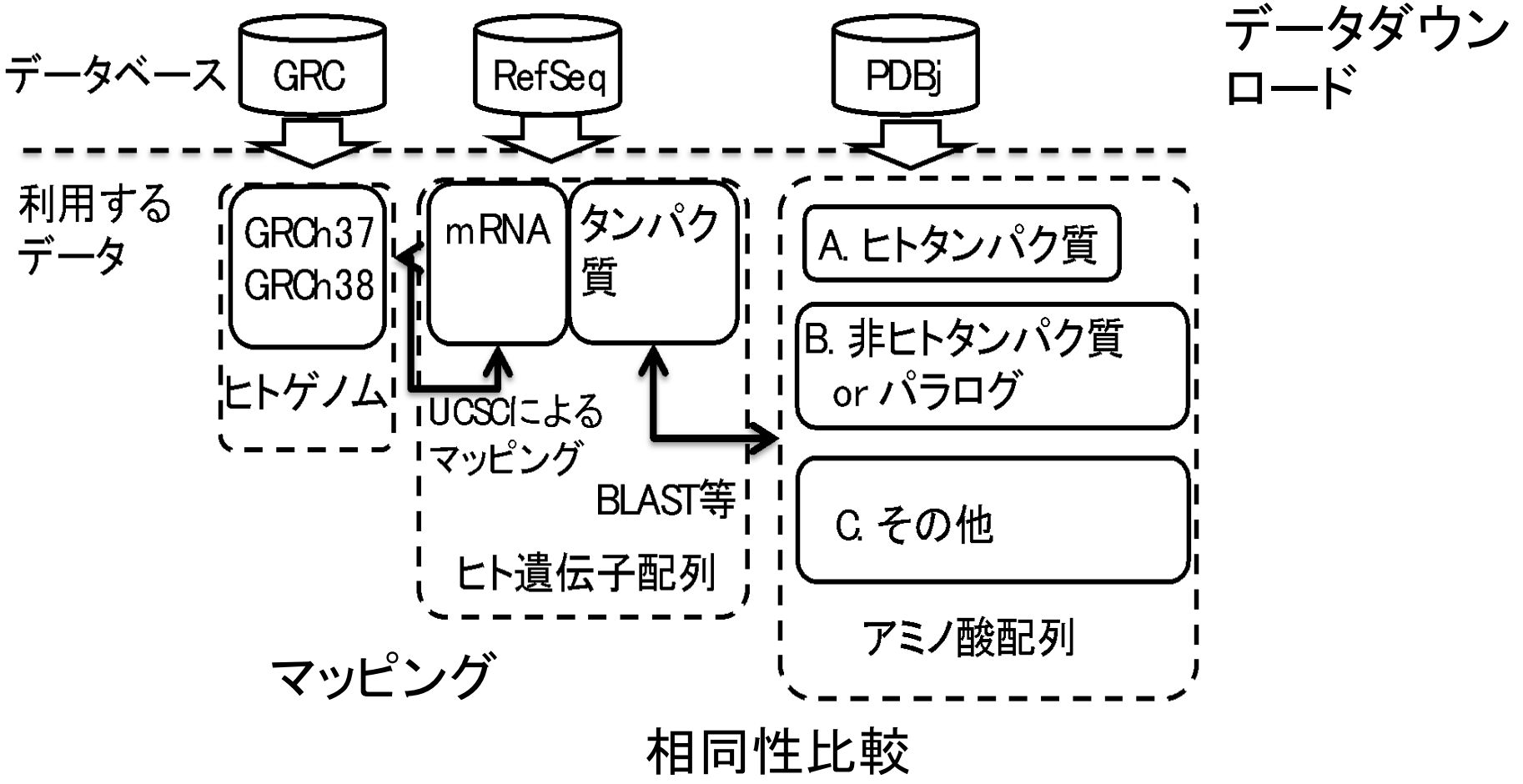


利用するデータベース

- 日本蛋白質構造データバンク(PDBj)
  - タンパク質の立体構造およびアミノ酸配列情報
  - 2014年7月15日現在101741エントリ
- NCBI RefSeq
  - タンパク質アミノ酸配列と対応するmRNA配列
- Genome Reference Consorsium (GRC)
  - ヒトゲノム配列
  - GRC.h38（最新版）とGRC.h37（1つ前の版）

研究開発内容1

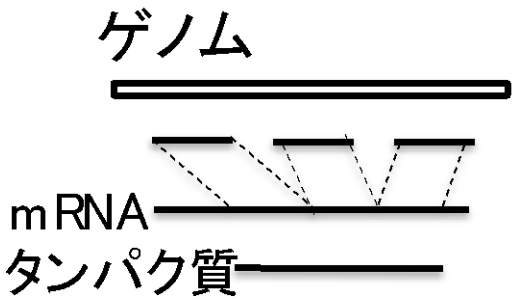
① PDBタンパク質のゲノム上位置の同定



研究開発内容2

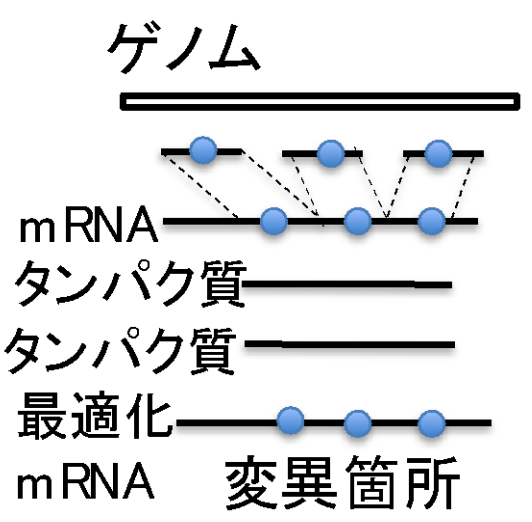
② 変異の有無の同定

A. ヒトタンパク質



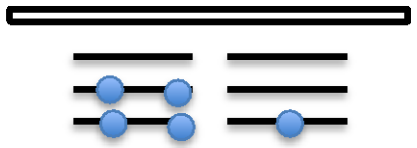
B. ヒト以外タンパク質
or パラログ

C. 未同定



研究開発内容3

③ SAM/BAMの作成



ゲノム上にマップされたmRNA配列と
変異情報からSAM/BAMを作成

1, 2で作成されたPDBjタンパク質とヒトゲノムの
対応関係をもとにSAM/BAMファイルを作成する

研究開発内容4

- 表示方法の検討
  - IGVなどのゲノムブラウザでpdbBAMを表示
  - 1000人ゲノムなどの個人ゲノム解析結果を同時に表示

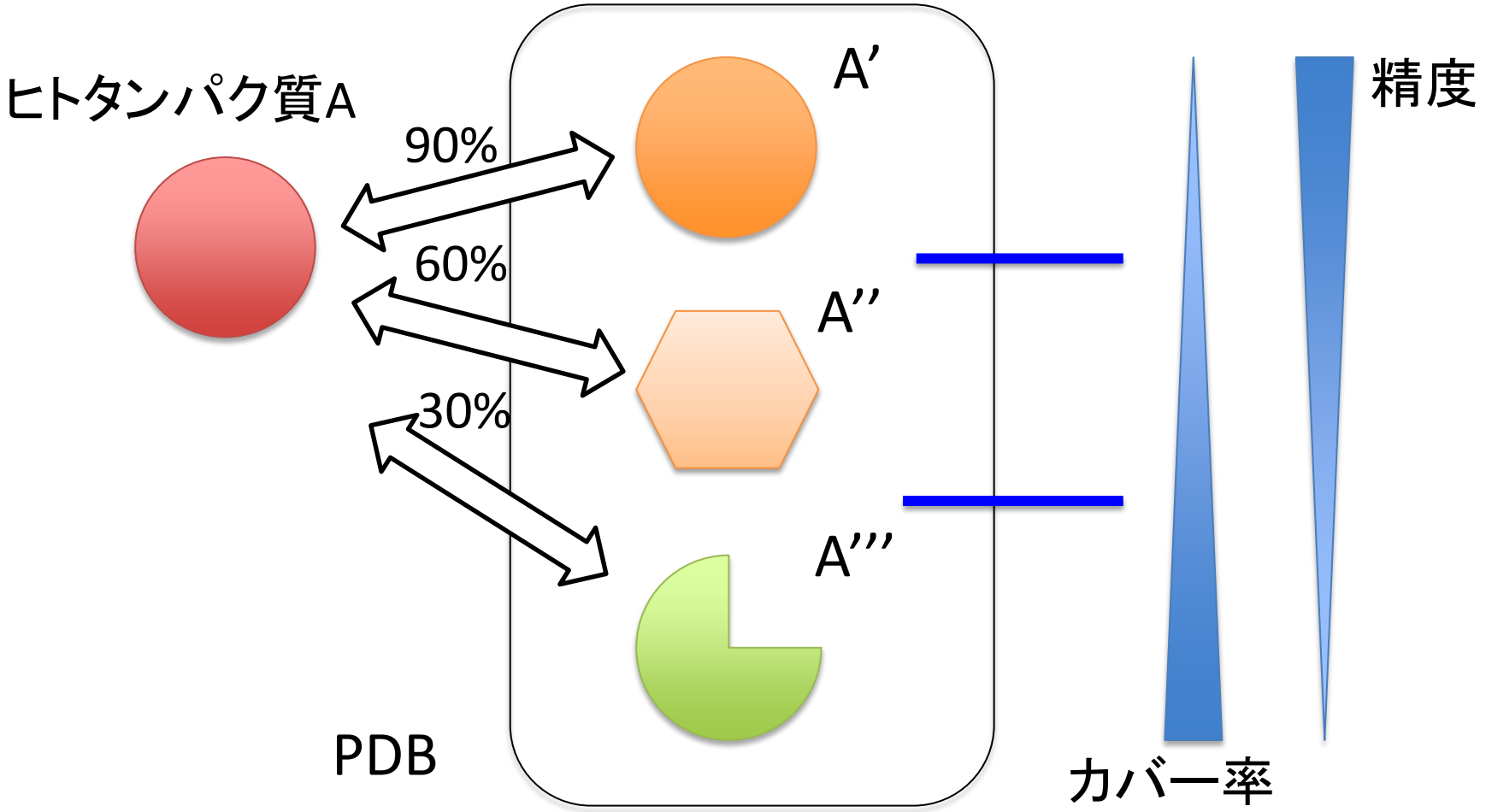
タンパク質配列相同性検索における配列一致度の閾値
アミノ酸配列からmRNA配列への変換方法

を検討

研究開発スケジュール

研究開発項目	平成26年				平成27年	
	9月	10月	11月	12月	1月	2月
1. PDBjタンパク質のヒトゲノムにおける位置の同定	→	→	→			
2. 変異の有無の同定		→	→	→	→	
3. SAM/BAMファイルの作成			→	→	→	→
4. 表示方法の評価					→	→

タンパク質相同性検索における閾値の検討



精度とカバー率のどちらを優先するかで閾値を選べるようにする

タンパク質配列からコドンへの変換

- PDBjタンパク質をmRNA配列に変換した配列と，対応するゲノム領域のDNA配列との一致度が悪い
  - ゲノムブラウザで表示した時に変異が多く入り，視認が難しくなる

- mRNA配列について同義コドンの変換を許すことで，ゲノムDNA配列との一致度をあげる.