「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

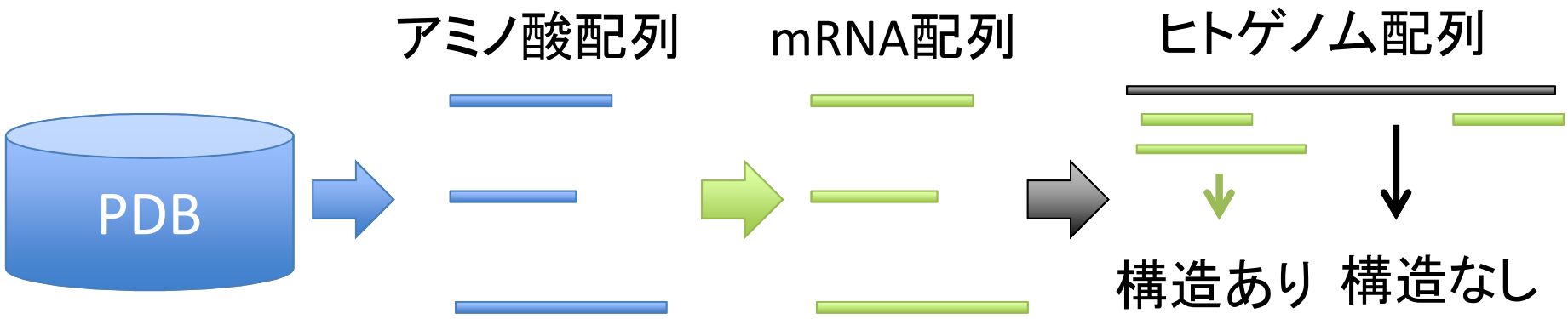
平成26年12月26日（金）
平成26年度「統合データ解析トライアル」中間報告会

本日の発表内容

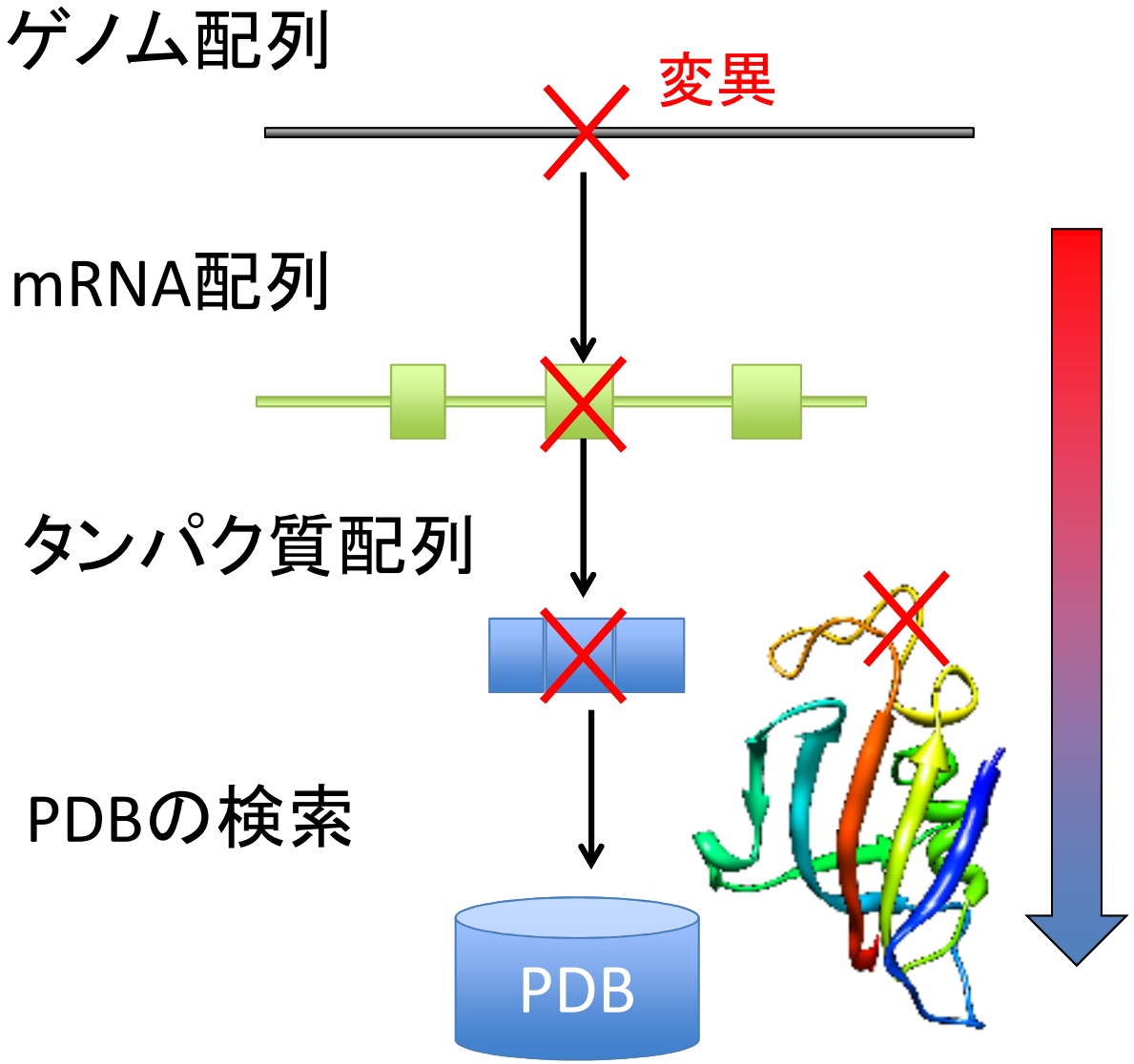
- pdbBAMとは
- ツールの作成
- IGVを用いた個人ゲノム情報とPDB情報の表示
- 今後の方向性

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

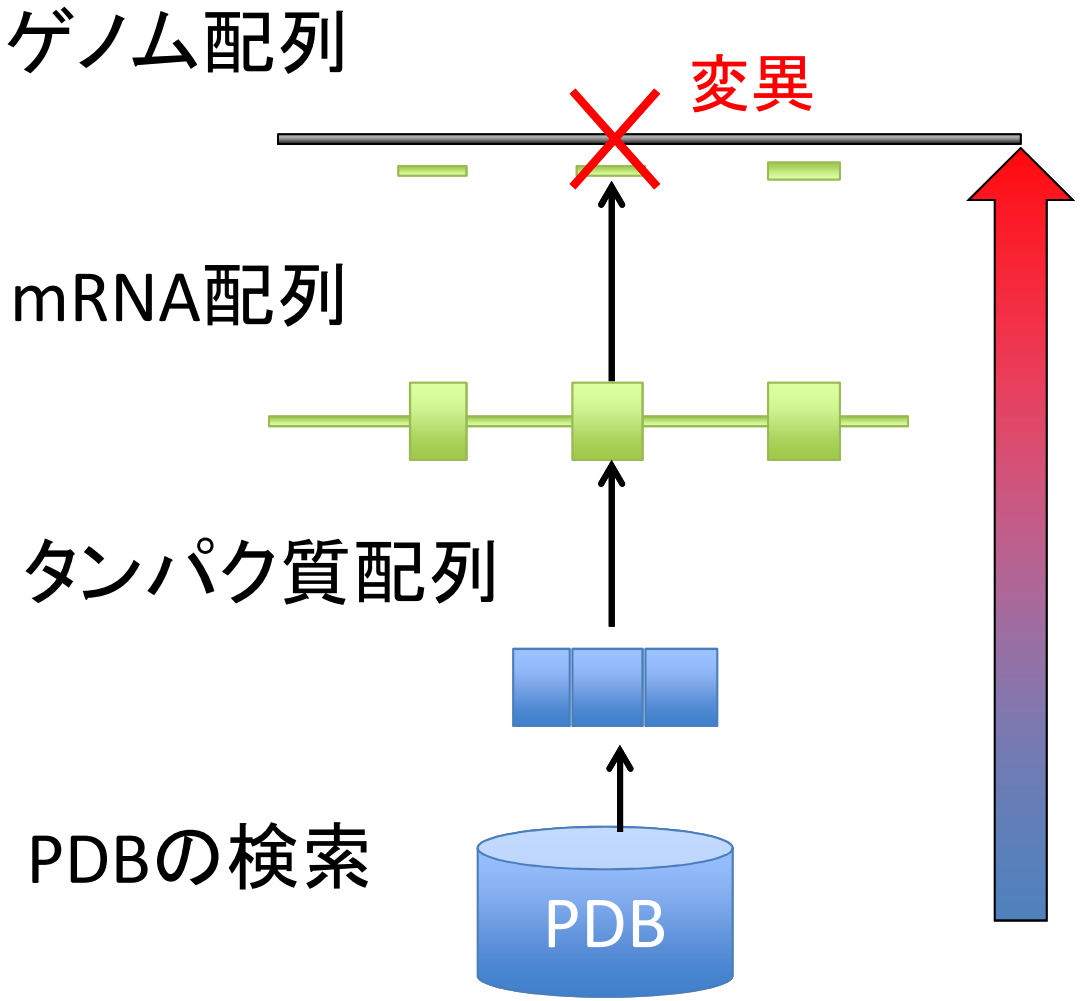


これまでの構造情報の利用



- StSNP
- coliSNP
- MuPIT
  - など

PDB配列からゲノム配列への逆マッピング



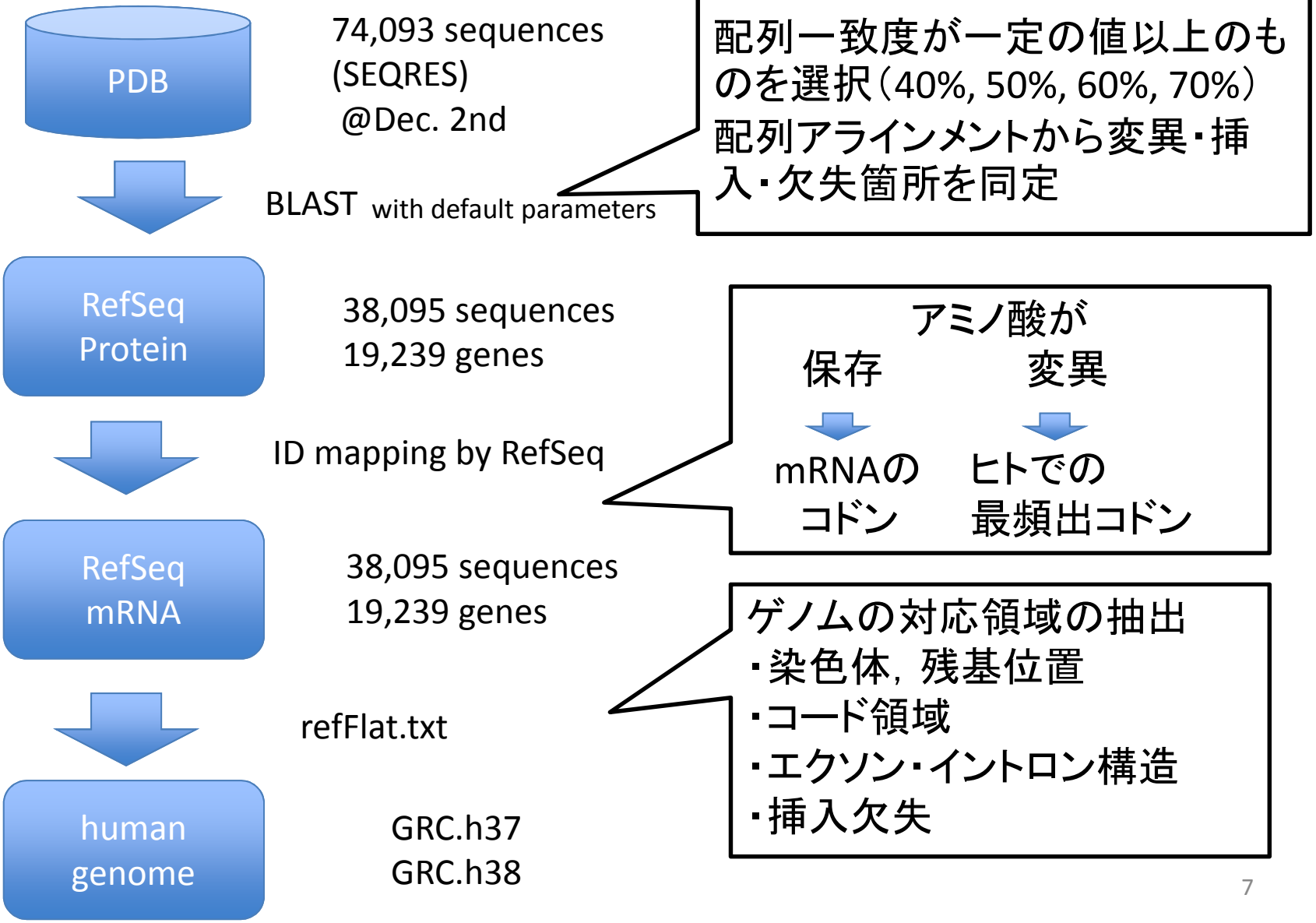
- 利点
  - 変異ごとの検索なしでよい
  - ゲノムを見ればPDB全体が分かる
- 欠点
  - 逆翻訳は曖昧さが残る
  - PDB検索の閾値

この向きのマッピングを行うツールはまだ存在しない

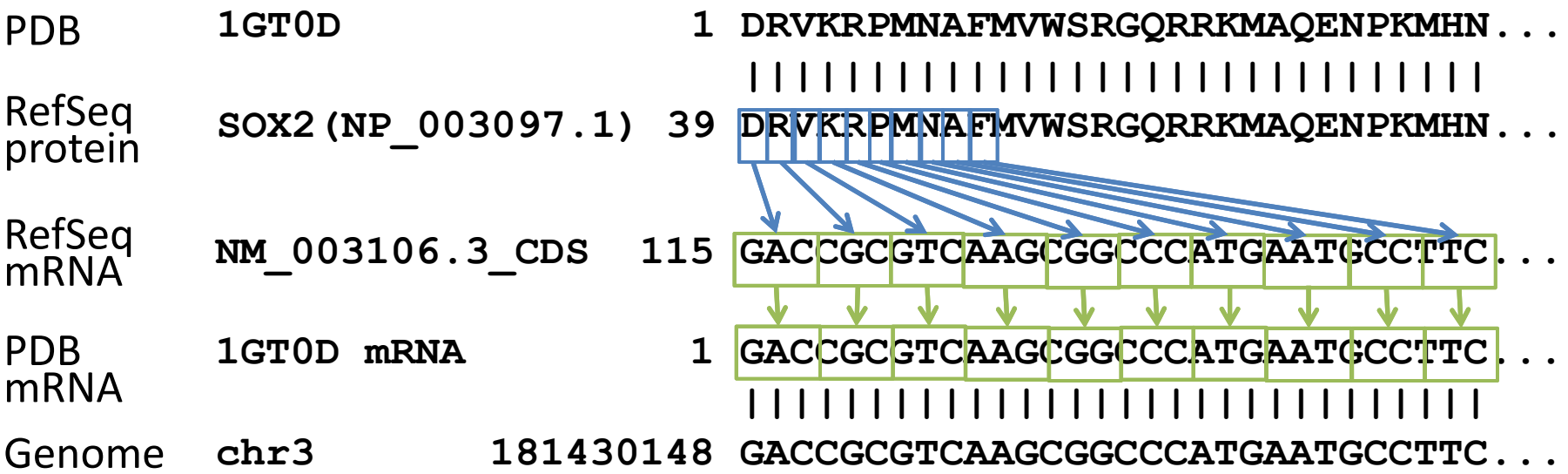
利用したデータベース

- 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
- 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）

pdbBAM作成の流れ



PDB配列からゲノムへのアラインメント
～完全一致の場合～



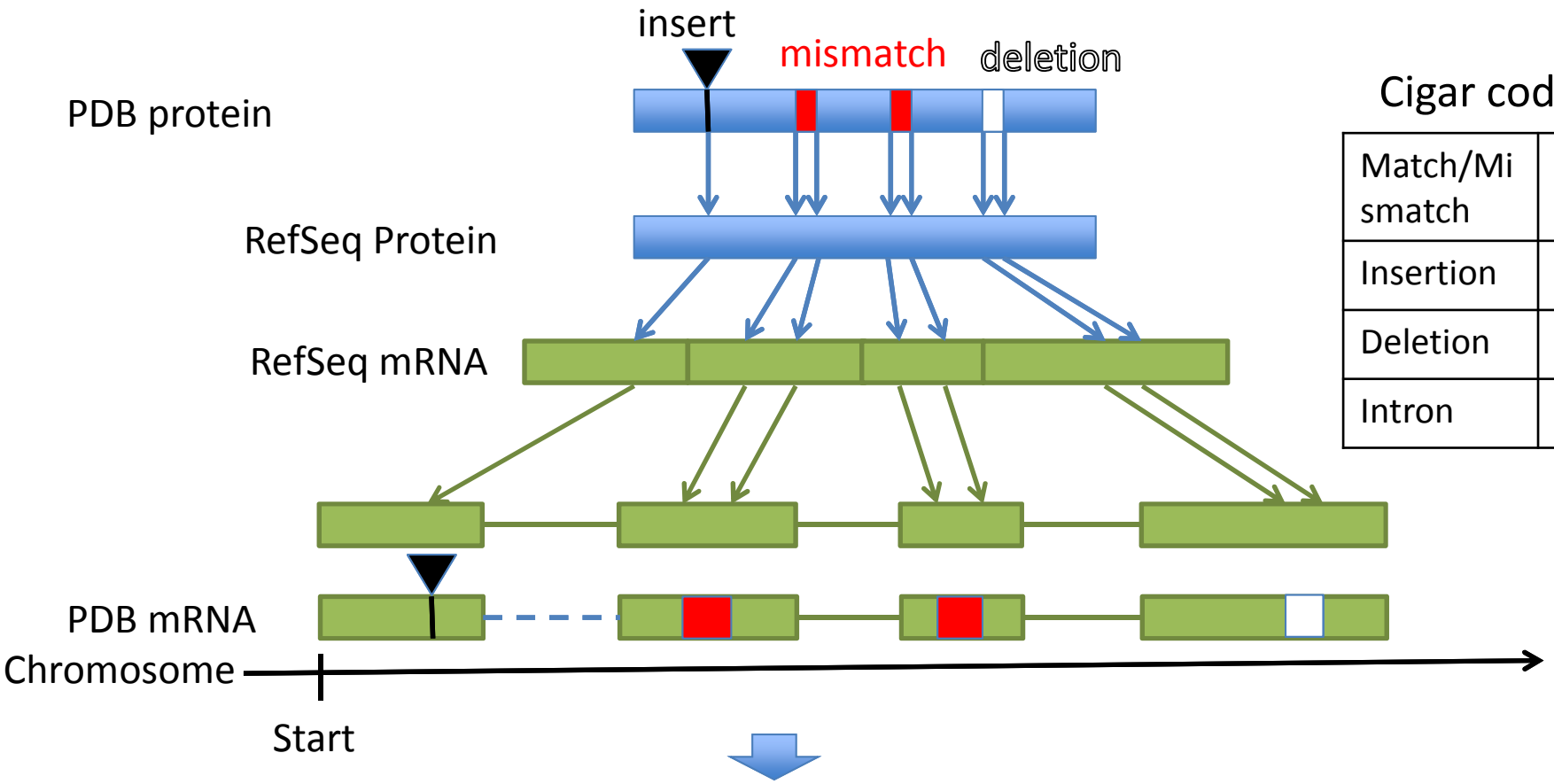
mRNAと一致する場合はmRNAのコドンを利用

PDB配列からゲノムへのアラインメント
～変異を含む場合～

mRNAと一致しない場合はアミノ酸に対する最頻出コドンで置換

アミノ酸	コドン	アミノ酸	コドン	アミノ酸	コドン	アミノ酸	コドン	アミノ酸	コドン
A	GCC	C	TGC	D	GAC	E	GAG	F	TTC
G	GGC	H	CAC	I	ATC	K	AAG	L	CTG
M	ATG	N	AAC	P	CCC	Q	CAG	R	AGA
S	AGC	T	ACC	V	CTG	W	TGG	Y	TAC

アラインメントからSAMフォーマットへ



Cigar code

Match/Mismatch	M
Insertion	I
Deletion	D
Intron	N

SAMフォーマット

```
pdb|3WHD|A_1 0 chr12 8670819 255 52M734N152M1065N116M782N145M * 0 0 CATGCA...
```

配列名　参照配列　位置　Cigar code　塩基配列

SAM/BAM変換

SAM -> BAM変換

```
samtools view -Sb pdbbam.sam | samtools sort > pdbbam.bam
```

インデックス作成

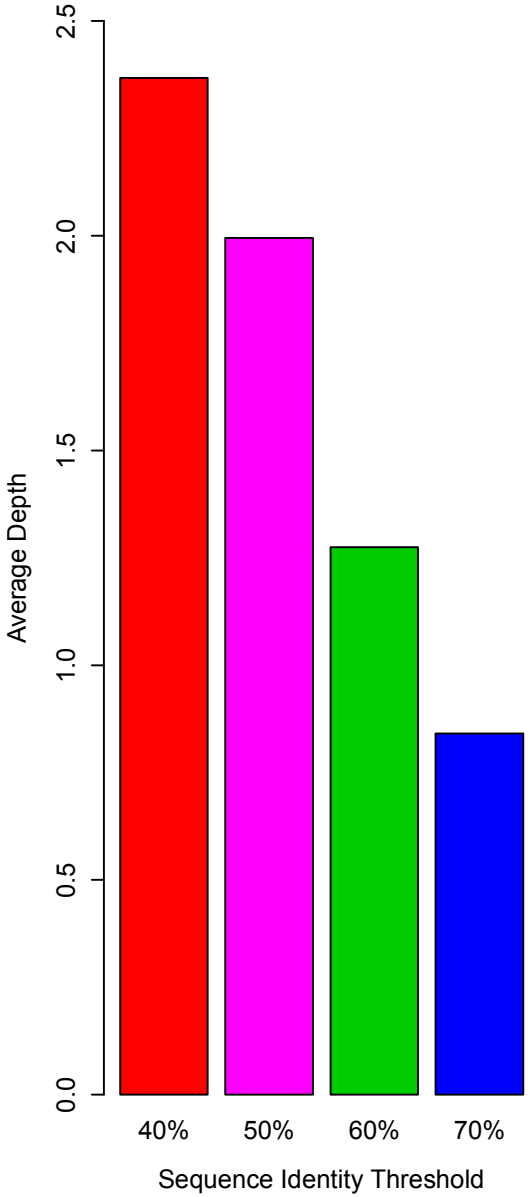
```
samtools index pdbbam.bam
```

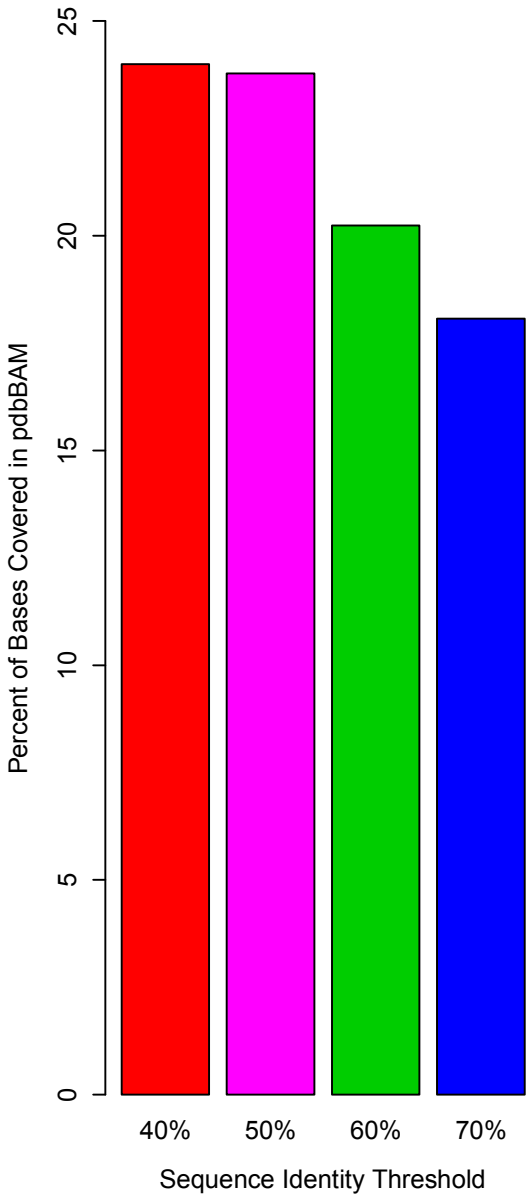
エクソン情報にあるPDBの抽出

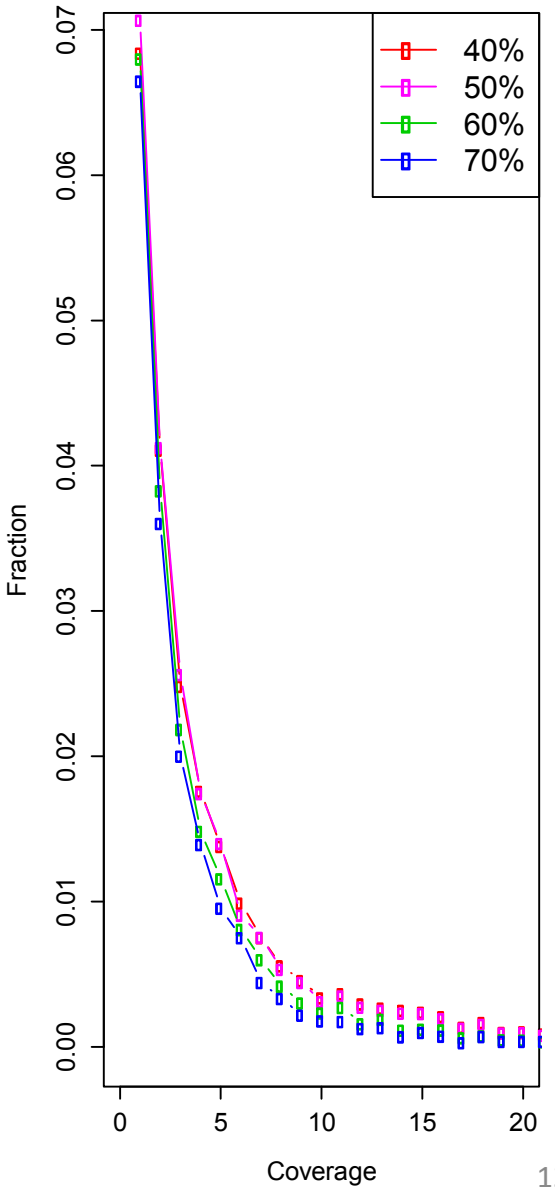
```
bedtools coverage –hist –abam pdbbam.bam –b refFlat.bed >
pdbbam.coverage.txt
```

samtools 0.1.18 Li H et al. Bioinformatics 25(16):2078-9, 2009
BEDtools 2.150.0 Quinlan AR, Hall IM. Bioinformatics 26(6):841-2, 2010

統計情報（GRC.h37のCDS 34,733,472塩基あたり）







個人ゲノムのBAMファイル作成

DDBJ(http://www.ddbj.nig.ac.jp/index-j.html)よりFastqをダウンロード

↓

BWA[1]を用いてhg19(GRC.h37)にマッピング

↓

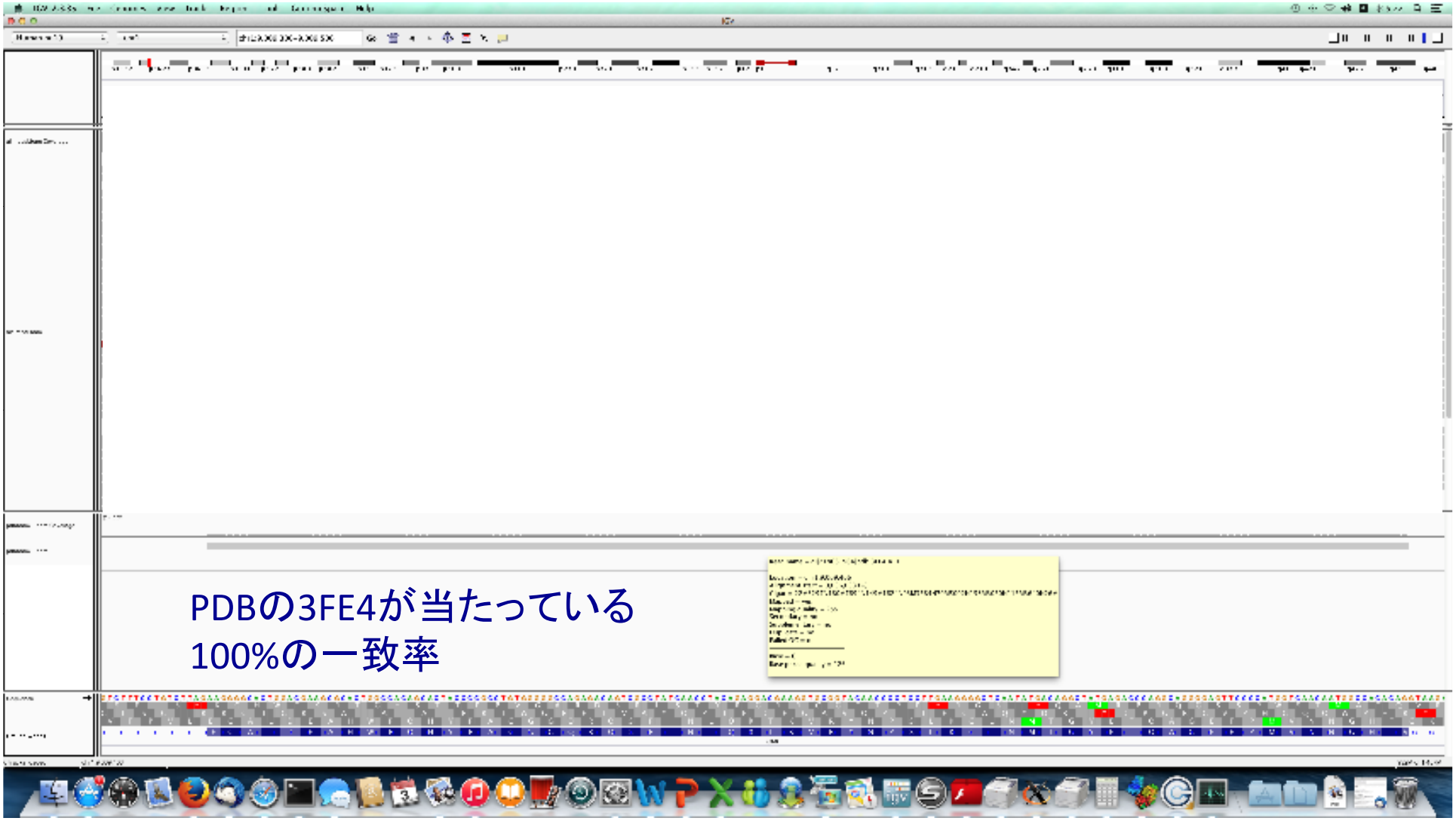
GATK[2]によるリアライメント

↓

BAMファイル

[1]Li H, Durbin R. Bioinformatics 25(24):1754-60, 2009
[2]DePristo MA et al. Nat Genet 43(5):491-8, 2011

実際の例 （Carbonic Anhydrase遺伝子）

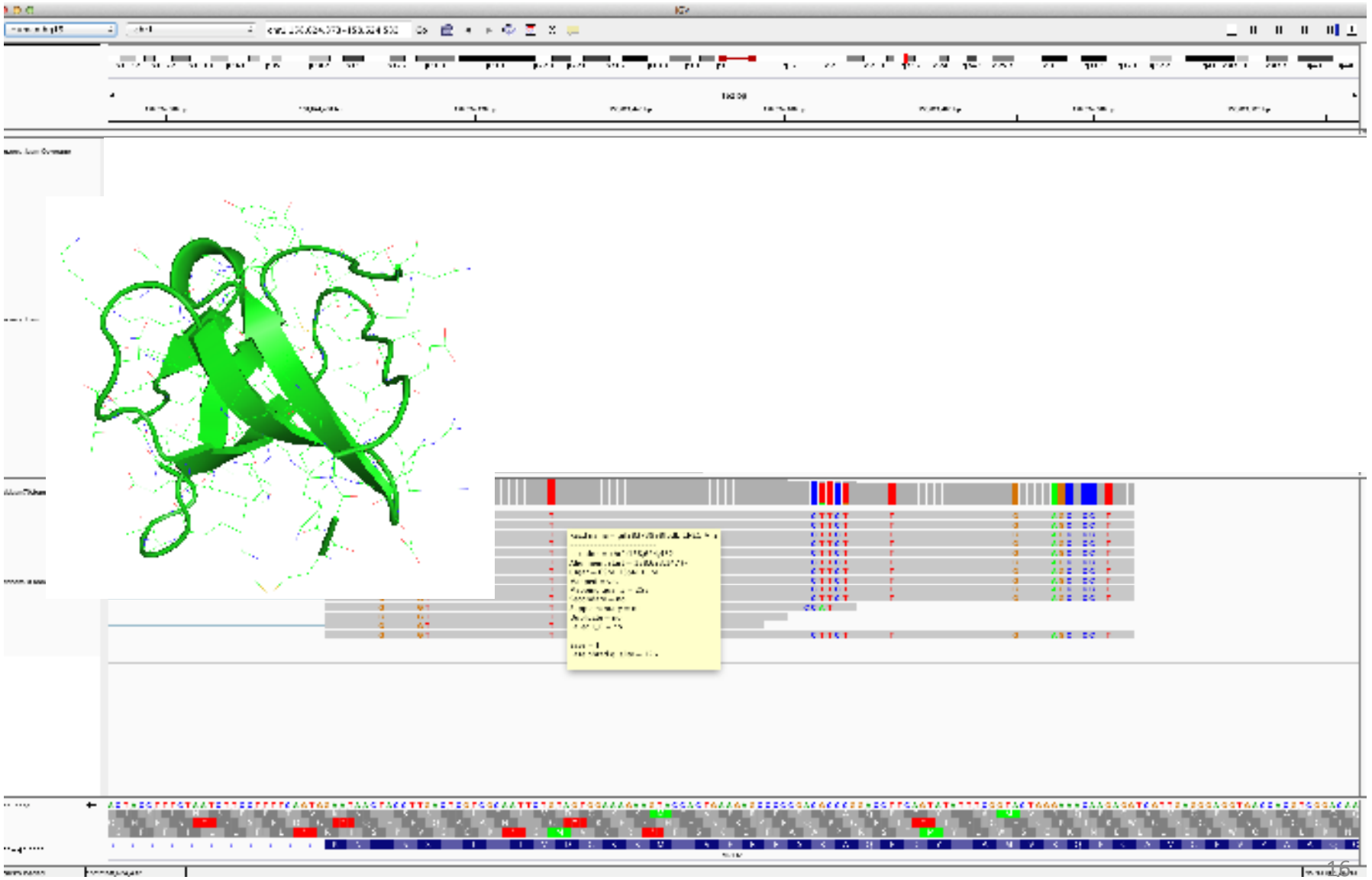


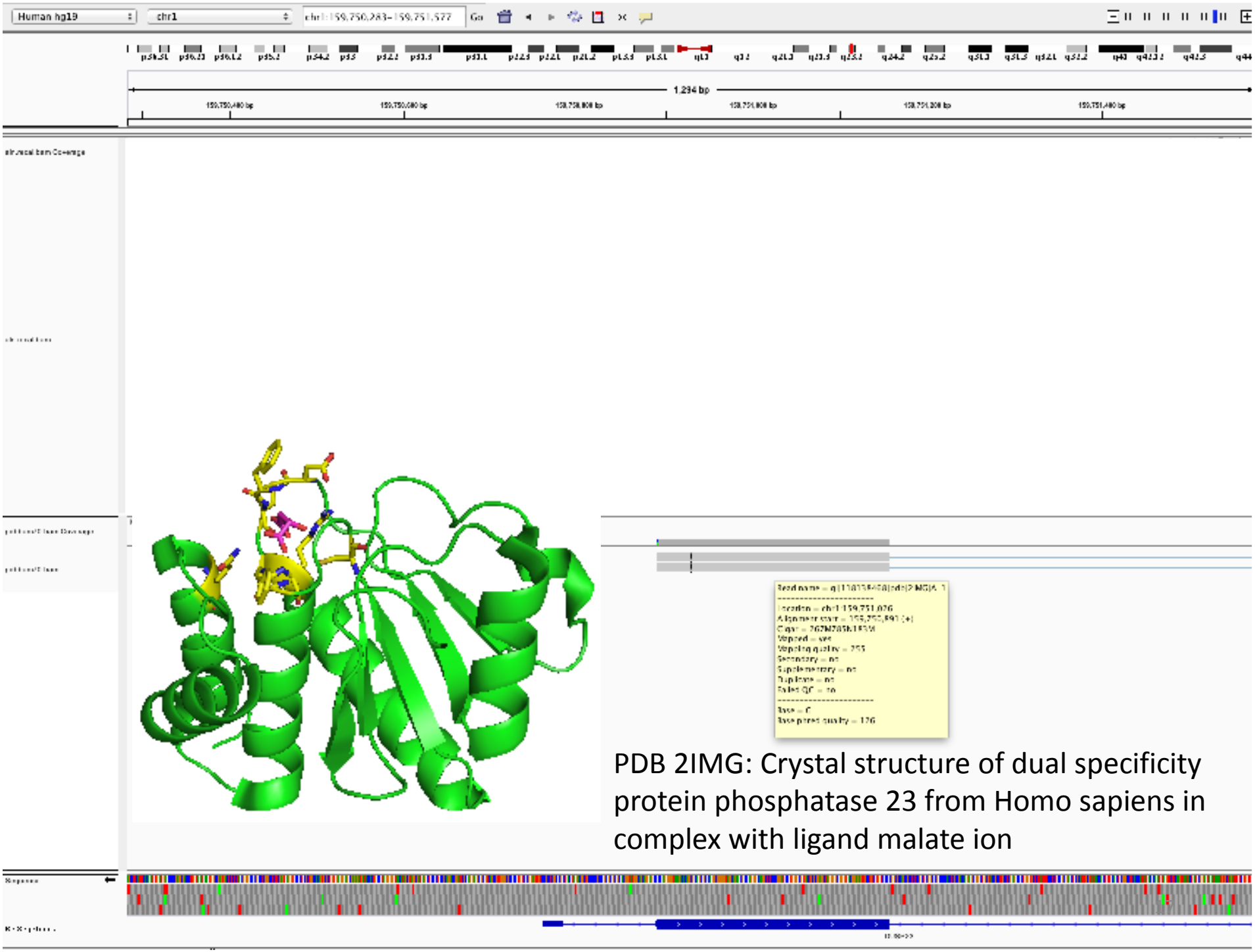
上段：NGSによる個人ゲノム解析結果
中断：pdbBAM70
下段：遺伝子領域と翻訳

Carbonic Anhydrase (CA6) PDB ID 3FE4

$$H_2CO_3 = CO_2 + H_2O.$$

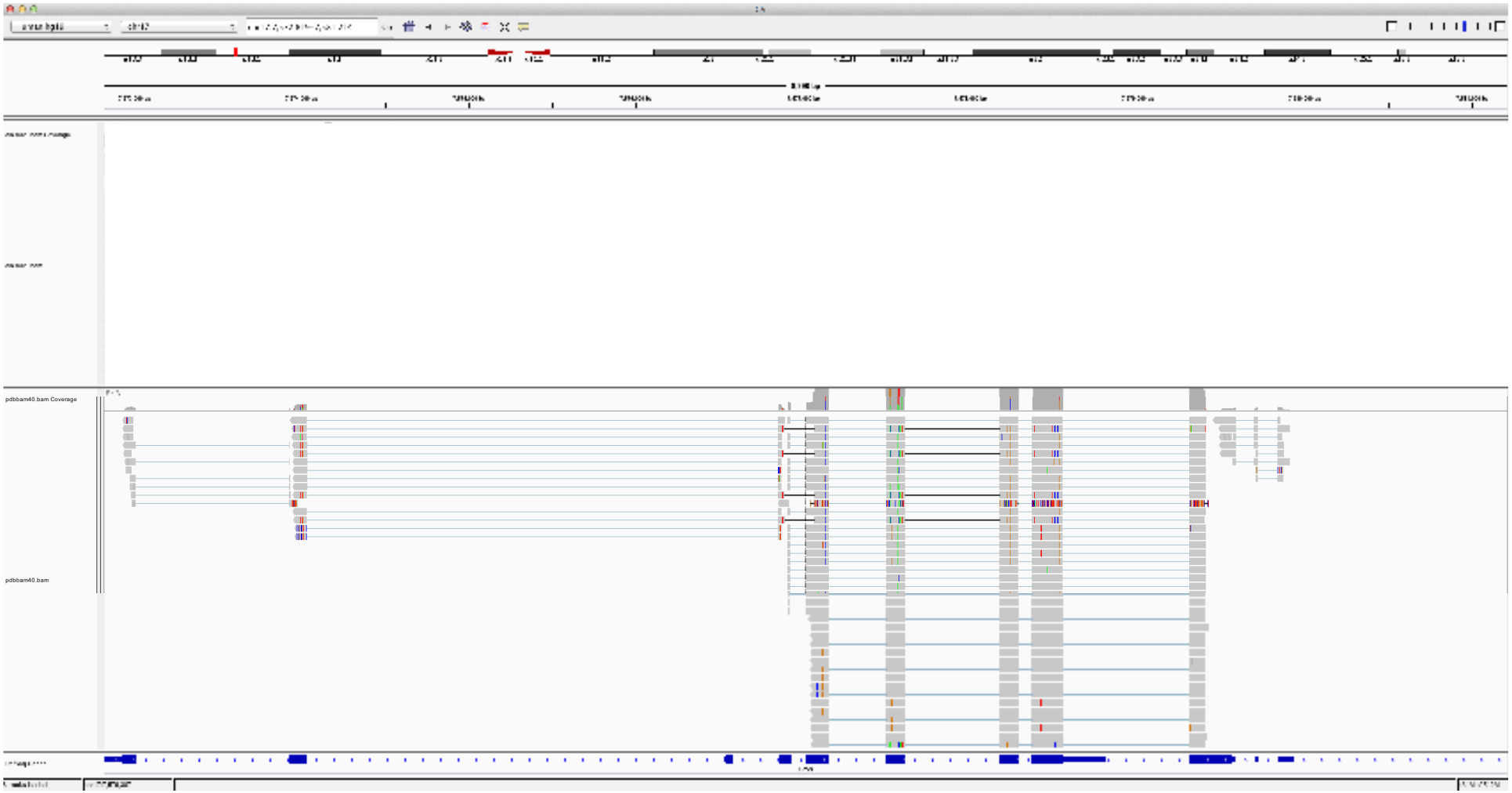
SH3 and cysteine domain-containing protein isoform 1 (SPTA1)





PDB 2IMG: Crystal structure of dual specificity protein phosphatase 23 from Homo sapiens in complex with ligand malate ion

P53領域

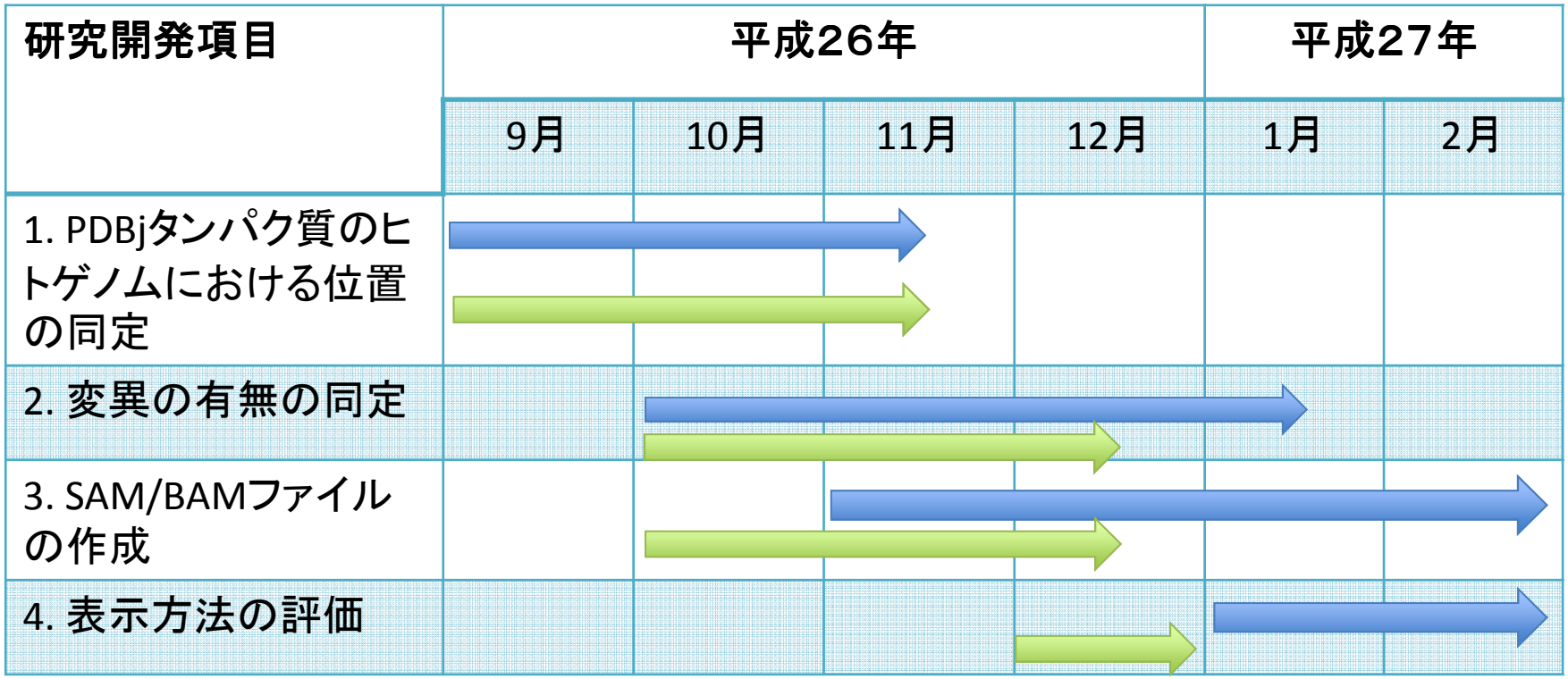


上段：NGSによる個人ゲノム解析結果
中断：pdbBAM70
下段：遺伝子領域と翻訳

現在の進捗まとめ

- PDBのタンパク質を配列相同性を用いてヒトゲノムにマッピングしてpdbBAMを作成
- 配列一致度について40%~70%までで作成
- ヒトゲノム上のコード領域におけるカバー率と平均深度は
  - カバー率：19%（配列一致度70%）~ 24%（同40%）
  - 平均深度：0.8（70%）~ 2.4（40%）
- 個人ゲノム解析結果とあわせてゲノム上のどこに，どの程度の数と類似度の立体構造情報があるかを表示できる

進捗状況


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
予定
実績

今後の課題

- 構造生物学的観点から
  - 構造情報を見たい（アミノ酸の埋もれ度など）
  - PDB IDだけでなく変異箇所の残基番号情報
  - ゲノムブラウザから直接構造を見られるツール
- ゲノム科学の観点から
  - dbSNPなどとの連携
  - マウスなど他の種への応用
- データ更新について
  - pdbBAMの作成は約6時間程度（ほぼBLAST検索）
  - PDBの更新に伴って月単位で作り直すことは容易