ライフサイエンスデータベース統合推進事業
統合化推進プログラム(統合データ解析トライアル)
研究開発課題
「PDBj タンパク質をゲノムにマップした pdbBAM の作成」

研究開発終了報告書

研究開発期間:平成26年9月～平成27年2月

研究代表者:城田　松之
(東北大学大学院医学系研究科　助教)

§1 研究開発の概要

次世代シークエンサの発展に伴い大規模なゲノムシークエンスが多数行われ，その結果がゲノムブラウザ上で様々なアノテーションデータと統合して表示・解釈されている．しかし，タンパク質の構造・機能情報はアミノ酸配列をベースとしているためにゲノム配列に紐づけて表示するツールが不足していた．この問題に対して，本研究開発では，PDBj に含まれる立体構造情報をゲノムブラウザから網羅的かつ俯瞰的に参照できるデータリソースとして pdbBAM を作成した．これは PDBj に登録された全タンパク質のアミノ酸配列をヒトタンパク質配列について相同性検索を行い，対応がとれたものについてmRNA配列を作成し，ヒトゲノムDNA配列にマッピングしゲノムブラウザで表示可能なBAM形式として統合したものである．pdbBAMをゲノムブラウザで読み込むことで，ゲノム上のどの遺伝子のどのドメインがどの程度の頻度で構造解析されているか，また，ある遺伝子に対してヒトのタンパク質そのものの構造が決まっているのか，あるいは相同なタンパク質の構造が決まっているのかを可視化することができた．これにより，個人ゲノム解析結果と構造生物学の成果の橋渡しをすることができ，将来的には個別化医療や創薬に応用が期待される．

§2 研究開発のねらい

次世代シークエンサ（NGS）による塩基配列決定の高速化と低コスト化により，近年 1000 人ゲノムプロジェクトをはじめとしてヒト集団内の多数の個人のゲノム解析が行われ，個人個人が持つ変異が多数同定されてきた．また，ゲノム解析だけでなく，遺伝子発現，タンパク質ーDNA 間相互作用，DNA メチル化の解析においても次世代シークエンサは現在の解析の主流となっている．これらの解析結果はヒトゲノムへの機能注釈としてゲノムブラウザを用いて統合的に表示・解釈されている．

NGS による研究が進む一方で，タンパク質科学の分野では個々のタンパク質の機能についての情報が蓄積してきた．その中でも蛋白質構造データバンク(PDB)に登録された立体構造はゲノム配列上の変異の影響の評価や予測においても頻繁に用いられてきた．ゲノム上の非同義変異がタンパク質の立体構造において内部に埋もれているか表面に露出しているか，また，酵素の活性部位からの空間的距離や他のタンパク質との相互作用面にあるかどうかなどの情報は立体構造なくしては得られないからである．しかし，ゲノム上のある一塩基変異が与えられた時に，それをタンパク質のアミノ酸変異に変換することは比較的容易であるが，タンパク質立体構造のどこの残基が変異するかを調べることはやや面倒な手続きが必要である．なぜならば，一つのタンパク質について複数の条件や変異体で構造解析されることもあれば，相同タンパク質の構造しか解かれていない，または全く類似のタンパク質の構造が存在しないこともある，というようにタンパク質の立体構造情報のあり方は非常に複雑だからである．ゲノム上の一塩基変異に対して立体構造を通して得られる情報を利用しやすくするためには， PDB 全体を見渡してどの程度の配列一致度の構造がどの程度あるのか，といった情報をゲノムブラウザ上で網羅的に表示することが必要である．

現在，個人ゲノム解析の結果として個人個人が持つ変異が多数同定されており，これらの変異の中には集団内でのアリル頻度が少なく，生物学的影響が不明なものが多く存在する．このような状況の中で立体構造をもとにゲノム上の非同義変異の生物学的影響を検討する手法の利用される頻度とその重要性はますます高まってくるであろう．そのためにゲノム科学とタンパク質立体構造の間の溝を埋めるツールが必要とされている．

そこで，本研究ではヒトゲノムの遺伝子コード領域上にPDBに登録された立体構造がどの程度存在するか，およびその構造の配列相同性についてゲノムブラウザを用いて網羅的に俯瞰するためのツールを開発する．実装として，NGS のショートリードのゲノム上へのマッピング結果を表示する際に使われる sequence alignment/mapping (SAM) および binary SAM (BAM)フォーマットを用いる．PDB のタンパク質のうち，ヒトタンパク質に相同性のあるものについてアミノ酸配列を mRNA 配列に変換し，次世代シークエンサから得られるショートリードのようにゲノムへのアラインメント情報を計算する．これらをPDB 全体について統合しBAMフォーマットに変換する．この結果をpdbBAMファ

イルとして公開し，ゲノムブラウザで読み込んだ時に各遺伝子位置に対応する立体構造情報全体を俯瞰することができるようにする．

本研究では主に以下の3つのデータベースを用いて開発を行う．

1) 日本蛋白質構造データバンク(PDBj)に登録された立体構造解析されたタンパク質のアミノ酸配列
2) NCBI RefSeq に登録された mRNA とタンパク質の対応する配列情報
3) ヒトゲノム配列，および RefSeq mRNA のゲノム上の位置

§3 研究開発計画

(1) 当初の研究開発計画

研究開発時の計画は以下の4点である．

1. PDBj タンパク質のヒトゲノムにおける位置の同定
   PDBj のタンパク質と NCBI RefSeq におけるヒトのタンパク質の間で相同性検索を行い，アミノ酸配列アラインメントを作成する．RefSeq のタンパク質配列は mRNA 配列と対応しており，その mRNA 配列がヒトゲノム上にマップされていれば，PDBj タンパク質のゲノム上の位置を決定することができる．
2. 変異の有無の同定
   1 の段階で，PDBj のタンパク質は A．ヒトのタンパク質そのもの，B．ヒト以外のタンパク質またはヒトのパラログ，C．RefSeq で同定できないタンパク質に分類できる．この時，B，C においてはゲノム上の DNA 配列と mRNA 配列の間に変異が見られるので，この変異情報を取り込む．
3. SAM/BAM ファイルの作成
   1, 2 で作られたゲノム上への PDBj タンパク質配列（および相当する mRNA 配列）のアラインメント・変異情報を SAM および BAM フォーマットに変換する．これを公開することで，PDBj タンパク質のヒトゲノムへのマッピング情報をゲノムブラウザで見ることを可能にする．
4. 表示方法の評価
   実際に作成された BAM ファイルと個人ゲノム解析の結果をゲノムブラウザを用いて表示させてどのような表示方法が良いかを評価，検討する．

(2) 新たに追加・修正など変更した研究開発計画

5. PDBj タンパク質とヒトタンパク質間の相同性のしきい値の検討
   1 で PDB タンパク質のヒトのタンパク質の対応付けをする時に配列相同性のしきい値を設定する必要がある．しきい値が高ければタンパク質間の構造の類似度が高いことが期待されるが，立体構造と関連づけられるヒトのタンパク質数は減少する．しきい値を下げればカバー率は上がるが，構造が類似していないタンパク質を誤って割り当てる可能性がある．そこで，相同性のしきい値を変えることがどのようにカバー率などに影響するかの評価を行う．
6. PDBj アミノ酸配列とヒト mRNA 配列の対応の問題
   計画段階 2 では PDB タンパク質を A ヒトのタンパク質，B RefSeq に含まれるヒト以外のタンパク質またはヒトのパラログタンパク質，C RefSeq に含まれないタンパク質に分けることを想定していた．しかし，実際に研究を進めていく上で，A と B の違いはあまり明確ではないこと，また，C は mRNA とタンパク質の対応付けを得ることが難しいことがわかった．そこで，PDBj タンパク質についてヒトタンパク質に相同性検索を行い，アミノ酸配列アラインメントから mRNA に変換する方法を採用した．

§４ 研究開発成果

本研究では PDBj のタンパク質を網羅的にゲノムにマップした pdbBAM を作成した．PDBj タンパク質とヒトタンパク質のアミノ酸配列の相同性のしきい値としては配列一致度 30%から 70%まで 10%ごとに 5 つの BAM ファイルを作成してそれぞれ pdbBAM_30.bam のように名付けた．また，これらのファイルをゲノムブラウザで表示する際に必要なインデックスファイル（pdbBAM_30.bam.bai 等）も作成した．pdbBAM を利用するためには対応する.bam および.bam.bai ファイルを同じディレクトリに置き，IGV などのゲノムブラウザで.bam ファイルをインポートすればよい．

網羅的に PDBj のタンパク質をゲノムにマップした結果，配列一致度のしきい値が 70%と 30%の場合，ヒトの全タンパク質において少なくとも一つの PDBj タンパク質によってカバーされる残基の割合はそれぞれ 18%，24%であり，コーディング領域の pdbBAM による平均深度はそれぞれ 0.8，2.4 であった．このことから，ヒトタンパク質において配列相同性を用いて PDBj の立体構造情報を利用できる領域は約 20%〜25%であり，そのうちの大部分（18%）は配列一致度 70%と高い一致度であることがわかる．これはヒトのタンパク質は最も重要な構造解析のターゲットであり，多くのヒトタンパク質そのものが構造決定されていることを反映していると考えられる．一方，配列相同性のしきい値を下げるとカバー率よりもむしろ平均深度が増加するため，ヒトのタンパク質のホモログも数多く構造解析されており，種間での立体構造の比較検討などに役立つことが期待される．

pdbBAM では PDBj タンパク質は mRNA 配列に変換されてからゲノム上に 1 本のリードとしてマップされる．RefSeq のヒトタンパク質とのアラインメントにおいて一致するアミノ酸には RefSeq mRNA における対応するコドンを割当て，一致しないアミノ酸についてはヒトのタンパク質内での各アミノ酸に対応する再頻出コドンを割り当てた．IGV などのゲノムブラウザで pdbBAM を表示させると，1 つ 1 つの PDBj タンパク質は灰色のボックスとして表示され，PDBj タンパク質の mRNA 配列とゲノム配列の間で塩基の不一致があるとその位置は塩基に対応して着色して表示される．PDBj タンパク質と RefSeq タンパク質の間でアミノ酸変異がある場合はコドンのうち少なくとも一つの塩基が着色して表示される．つまり，各 PDBj タンパク質に対応するボックスの着色の程度により， PDBj タンパク質がゲノム上のタンパク質にどの程度の配列一致度であるかがわかる仕様になっている．このことは，相同タンパク質の立体構造をもとにゲノム上の非同義変異の影響を議論する際に構造の信頼性を評価する上で有用な指標となる．また，同じ遺伝子領域に対応する PDBj タンパク質は pdbBAM のトラックの中で縦に並ぶようになっており，これにより同じタンパク質がどの程度冗長に構造解析されているかを把握することができる．一つの PDBj タンパク質が複数のエクソンにまたがってゲノム上にマップされている場合は，イントロンは細く表示される．遺伝子領域に表示された PDBj タンパク質の範囲を見ることで，タンパク質のどのドメインが構造決定されているかを把握することができる．また，各 PDBj タンパク質の PDB ID とチェイン ID はブラウザ上でリードにマウスをポイントすることでポップアップ表示させることができる．以上のように，pdbBAM をゲノムブラウザで表示することで遺伝子領域に該当する PDBj 立体構造の数や配列一致度，ドメイン構造，そして ID といった情報を簡単に取得することができる．ゲノムブラウザでタンパク質の立体構造情報のある領域を俯瞰することが可能になったともいえる．これにより，ヒトゲノム解析の研究者が立体構造情報を扱うことがより容易になり，構造をもとにした機能の推定に役立つと考えられる．

本研究の成果は以下の URL からダウンロード可能である．また，ツールの説明や利用方法についても同サイトからアクセス可能である．
http://aoba.hgc.jp/pdbbam/

1） 統合化推進プログラムで統合化されたデータベースの活用法
本研究では PDBj ブラウザから統合化推進プログラムで統合化されたデータベースのうち Protein Data Bank (PDB)のデータの活用を推進するものである．
2） 有用な知見の発見
pdbBAM を用いることで，ゲノムブラウザ上で既存の立体構造情報を俯瞰することができる．現在ヒトゲノムの変異が見つかる数はますます増加している．これらの情報と立体構造をあわせて有用な

知見の発見に結びつくことが期待される.
3) 汎用性
本研究の成果はゲノム解析だけでなく RNA-seq や ChIP-seq といった他のハイスループットシークエンシングの結果ともあわせて表示することが可能である. この意味で, 様々なシークエンシング技術との融合による新しい知識発見が期待される

§5 研究開発計画に対する達成状況と将来展望

(1)達成状況

当初の開発計画に対する研究達成状況

1. PDBj タンパク質のヒトゲノムにおける位置の同定
   PDBj に登録された全てのタンパク質についてヒトのタンパク質との相同性検索を行い, そのアラインメントと mRNA のゲノム上の位置に基づいて PDBj タンパク質をゲノム上にマッピングすることができた.
2. 変異の有無の同定
   PDBj タンパク質とヒトのタンパク質の間にアミノ酸変異があるケースとないケースどちらについてもゲノム上にマッピングすることができたとともに, 変異を同定することも可能である.
3. SAM/BAM ファイルの作成
   PDBjタンパク質とヒトタンパク質の間のアミノ酸配列アラインメントに基づいてヒト mRNA 配列を修正し, ヒトゲノムに対する PDBj タンパク質の領域と変異, 挿入, 欠失を SAM/BAM フォーマットで表示した.
4. 表示方法の評価
   pdbBAM をゲノムブラウザ IGV で見ることにより表示方法の評価を行った.

新たに追加, 修正した開発計画に対する研究達成状況

5. PDBj タンパク質とヒトタンパク質間の相同性のしきい値の問題
   本研究では配列一致度が 30%から 70%までしきい値を変化させることで, ヒトタンパク質における PDBj タンパク質のカバー率を比較した. ヒトタンパク質の立体構造でのカバー率はしきい値を 70%から 30%に下げることで約 18%から 24%に増加した.
6. PDB アミノ酸配列とヒト mRNA 配列の対応の問題
   PDBj タンパク質を分類なしにヒトタンパク質に割当て, mRNA に変換するために疑似 mRNA 配列を作ることにした. まずヒト mRNA 配列を出発点として, PDB タンパク質と RefSeq のヒトタンパク質の間でアミノ酸配列の違いがあればコドンごとに置換するという方法を選択した.

(2)ツール等の将来展望

本ツールを開発することで, ゲノムブラウザ上でタンパク質の構造や機能についての情報を知りたいというニーズが比較的多いことがわかった. タンパク質の天然変性領域, アミノ酸残基の構造上の内外, リン酸化, ユビキチン化などの修飾部位, リガンドやタンパク質との相互作用部位などの残基レベルの情報をゲノムブラウザから見たいという要望を多く受けている. 本研究の発展方向の一つとしてそのようなタンパク質のアミノ酸残基レベルの情報をゲノム情報とともに可視化することがあげられる. UniProt などのタンパク質データベースに登録されている活性部位などの情報, 天然変性領域の予測, さらに PDBj 立体構造に基づくアミノ酸残基の埋もれ度, 溶媒接触表面積, 二次構造, リガンドや他のタンパク質との相互作用残基情報など, さまざまな情報をゲノムブラウザ上で利用可能にする方法を検討中である.

同時に, ヒト以外の生物種のゲノムについても同様に PDBj タンパク質のマッピングと BAM ファイル作成を行うことができる. 特にマウスを扱う研究者からは mm9, mm10 といったゲノムへの拡張の希望があった. そのため, マウス, ラット, ゼブラフィッシュ, ショウジョウバエ, 線虫, 酵母などのモデル生物を対象に, 各種の pdbBAM を作成することはそれぞれの生物種の研究者にとって有

益であろう. この先の発展性としてはゲノム解析された全生物種に拡張するという展望が考えられる. もちろん, 前述の立体構造以外のタンパク質の構造や機能についての情報もこれらの生物種に拡張することが可能である.

網羅的に立体構造を取り扱うことで, ヒトの非同義変異のうち, PDBj で変異体が解析されているものを集めた解析も可能となる. 現在, PDBj に集められた立体構造が持つ変異は構造解析の目的で導入されたものが多い. しかし, 個人ゲノム解析が進んだ現在では, 天然に存在する変異の構造・機能への影響が興味を集めている. 今後天然の残基変異の構造解析が進んでくれば, dbSNP などのヒトの変異データベースと PDBj の間の対応付けも必要になるかもしれない.

また, 今回はゲノム解析に立体構造情報を適用するという方向での利用可能性を向上させるツールを作成したが, その逆向きの流れ, つまり大規模ゲノム解析の結果を構造生物学に還流する研究も今後必要となってくると考えている. ゲノム解析によって個々のタンパク質の変異の有無, 頻度などが解明されており, 生物学的・医学的に重要なタンパク質変異も見つかっている. これらの変異の構造・機能上の影響を解明するためのタンパク質科学研究も今後重要となってくる.

§6 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○城田松之	東北大学大学院医学形研究科	助教	pdbBAM の作成, 評価, 改良	H26.9-H27.2
水谷卓郎	東北大学大学院情報科学研究科	修士 1 年	pdbBAM のテスト利用と評価	H26.11-H27.2

§7 成果発表等

(1)原著論文発表 (国内(和文)誌 0 件、国際 (欧文) 誌 0 件)

(2)その他の著作物(総説、書籍など)

なし

(3)国際学会発表及び主要な国内学会発表

① 招待講演 (国内会議 0 件、国際会議 0 件)
② 口頭発表 (国内会議 0 件、国際会議 0 件)
③ ポスター発表 (国内会議 0 件、国際会議 0 件)
④

(4)知財出願

①国内出願 (0 件)
②海外出願 (0 件)
③その他の知的財産権

なし

(5)受賞・報道等

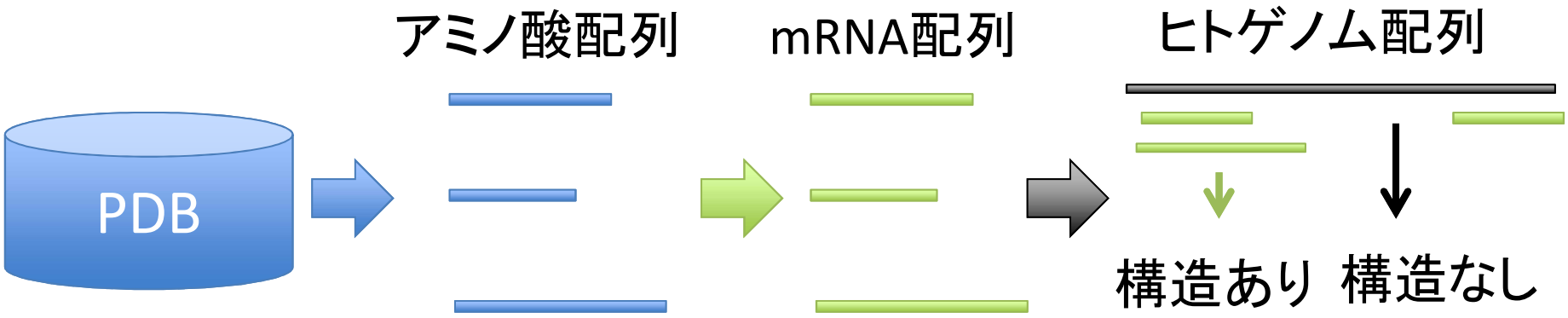
なし

§8 自己評価

本研究開発では開始時に設定した，PDBj タンパク質全体をヒトゲノムにマップしてゲノムブラウザから閲覧できるようなツールを作るという目的は達成することができた．また，そのツールを個人ゲノム解析の結果とあわせて評価することで，ゲノム解析の研究者にとって有用なツールとなりうることが示唆された．さらに，今後タンパク質の機能に関わる情報をゲノムブラウザ上に表示していくことが求められているということがわかった．pdbBAM は一般的なゲノムブラウザで閲覧することができ，ゲノム上の変異から立体構造上のアミノ酸変異にいたるステップを減らすことに貢献できると考える．現在，大規模なゲノム解析・エクソーム解析プロジェクトの成果として，数万～数十万の非同義変異がヒト集団において見つかっている．これらの変異がどの程度生物学的影響があり表現型に影響を及ぼすかはほとんどわかっていない．タンパク質構造とこれらの変異情報を網羅的に統合することで，生物学，医学，健康科学において重要な発見につながることが期待される．

pdbBAMとは

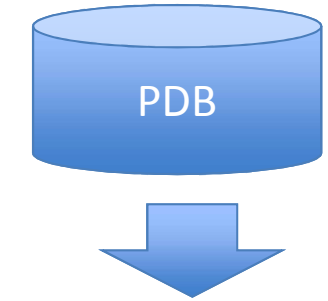
- PDBjに含まれるタンパク質のアミノ酸配列をmRNA配列を介してヒトゲノムにマップしてBAM形式としたもの
- PDB全体をゲノムに貼付ける
- ゲノムのどこの遺伝子が構造解析されているかを一目でわかるようにする



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



74,093 sequences
(SEQRES)
@Dec. 2nd

BLAST with default parameters

配列一致度が一定以上のものを選択（30%, 40%, 50%, 60%, 70%）
配列アラインメントから変異・挿入・欠失箇所を同定

38,095 sequences
19,239 genes

ID mapping by RefSeq

アミノ酸が
保存 → mRNAのコドン
変異 → ヒトでの最頻出コドン

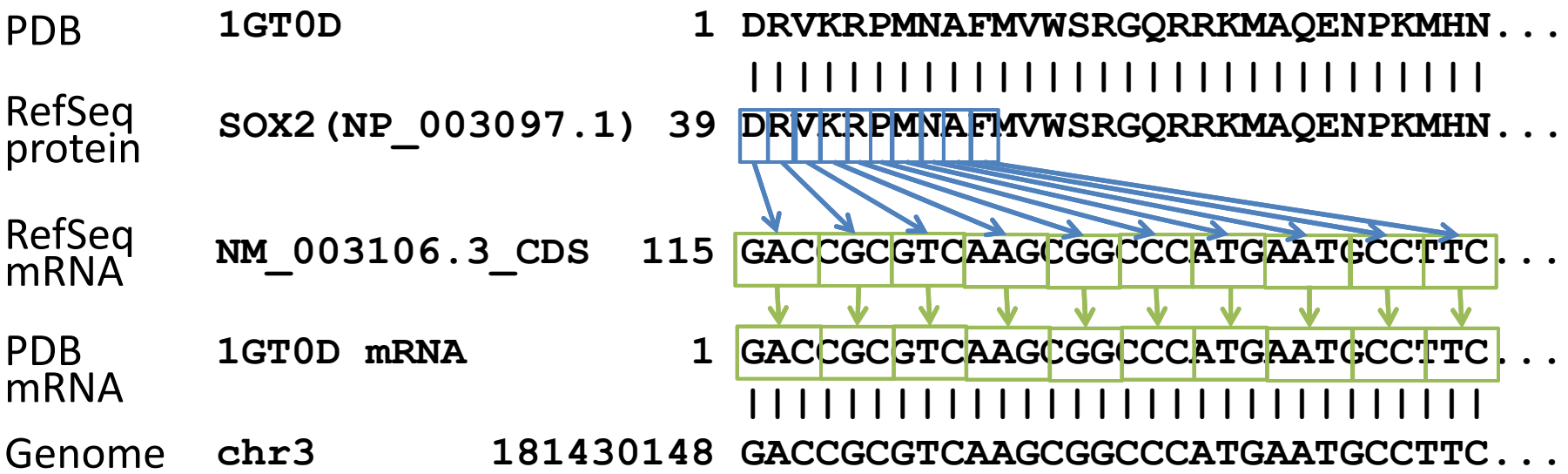
38,095 sequences
19,239 genes

refFlat.txt

ゲノムの対応領域の抽出
・染色体，残基位置
・コード領域
・エクソン・イントロン構造
・挿入欠失

GRC.h37
GRC.h38

PDB配列からゲノムへのアラインメント
～完全一致の場合～



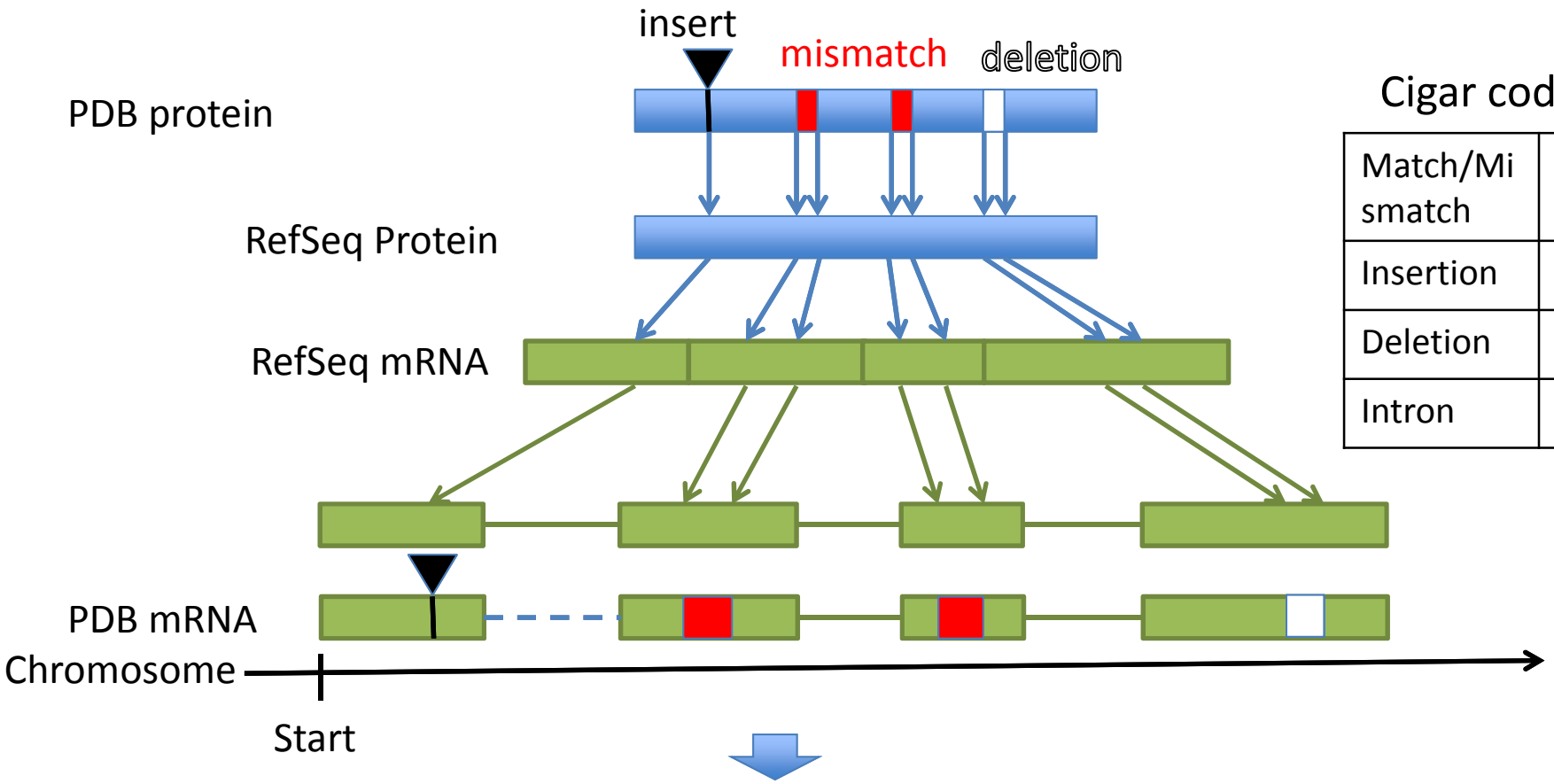
mRNAと一致する場合はmRNAのコドンを利用

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

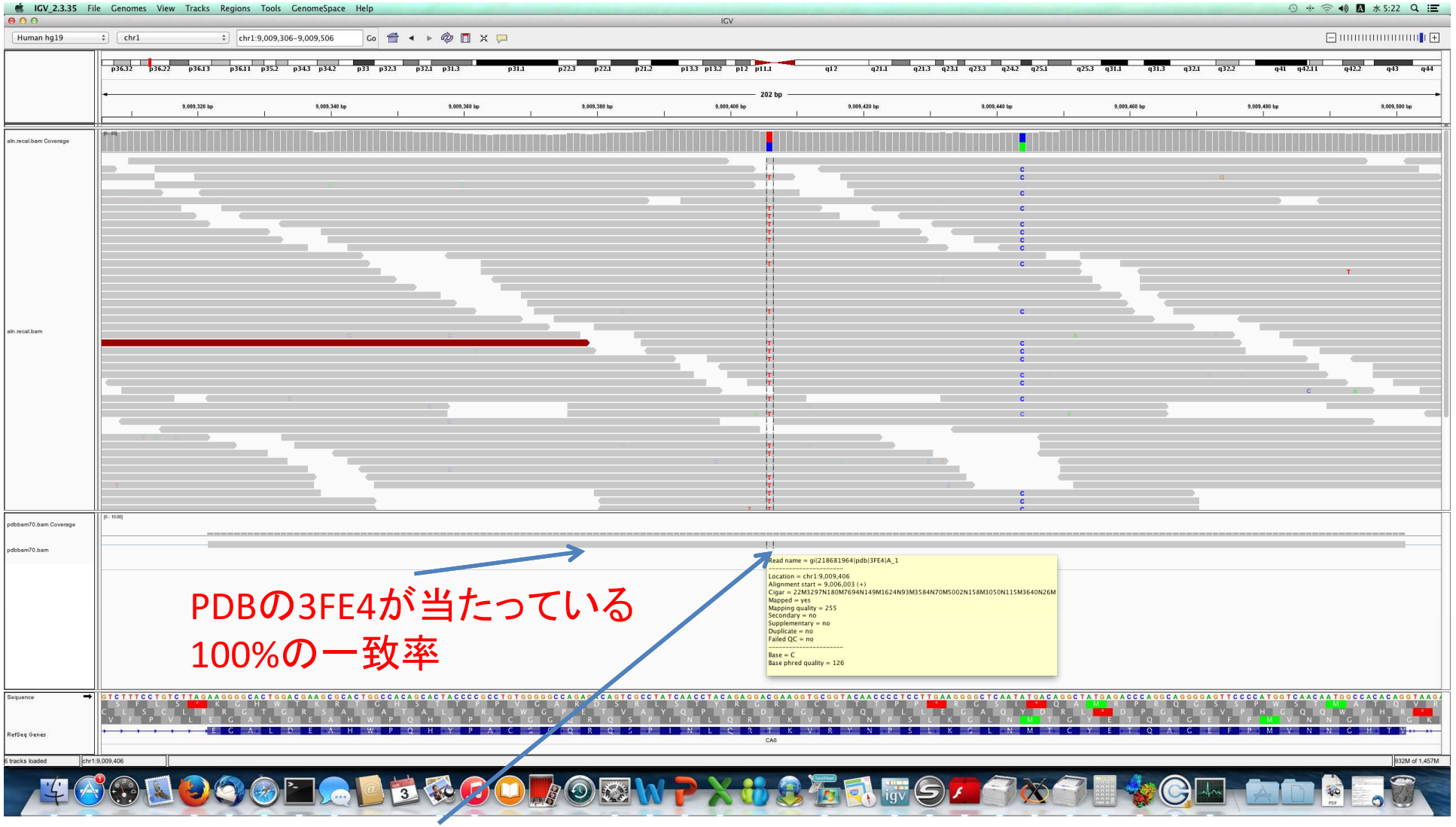
```
pdb|3WHD|A_1 0 chr12 8670819 255 52M734N152M1065N116M782N145M * 0 0 CATGCA...
```

配列名　参照配列　位置　Cigar code　塩基配列

pdbBAMの利用法

1. タンパク質相同性検索における配列一致度のしきい値が30%から70%に対応した5種類のファイルがあるので，適当なファイルを選択
2. pdbBAM_**.bam, pdbBAM_**.bam.baiファイルを同じディレクトリに置く（**はしきい値）
3. IGVなどのゲノムブラウザで.bamファイルを開く
4. 目的のゲノム領域，ゲノム領域などをブラウジングする．BAMのリード情報はブラウザを拡大すると表示される

実際の例

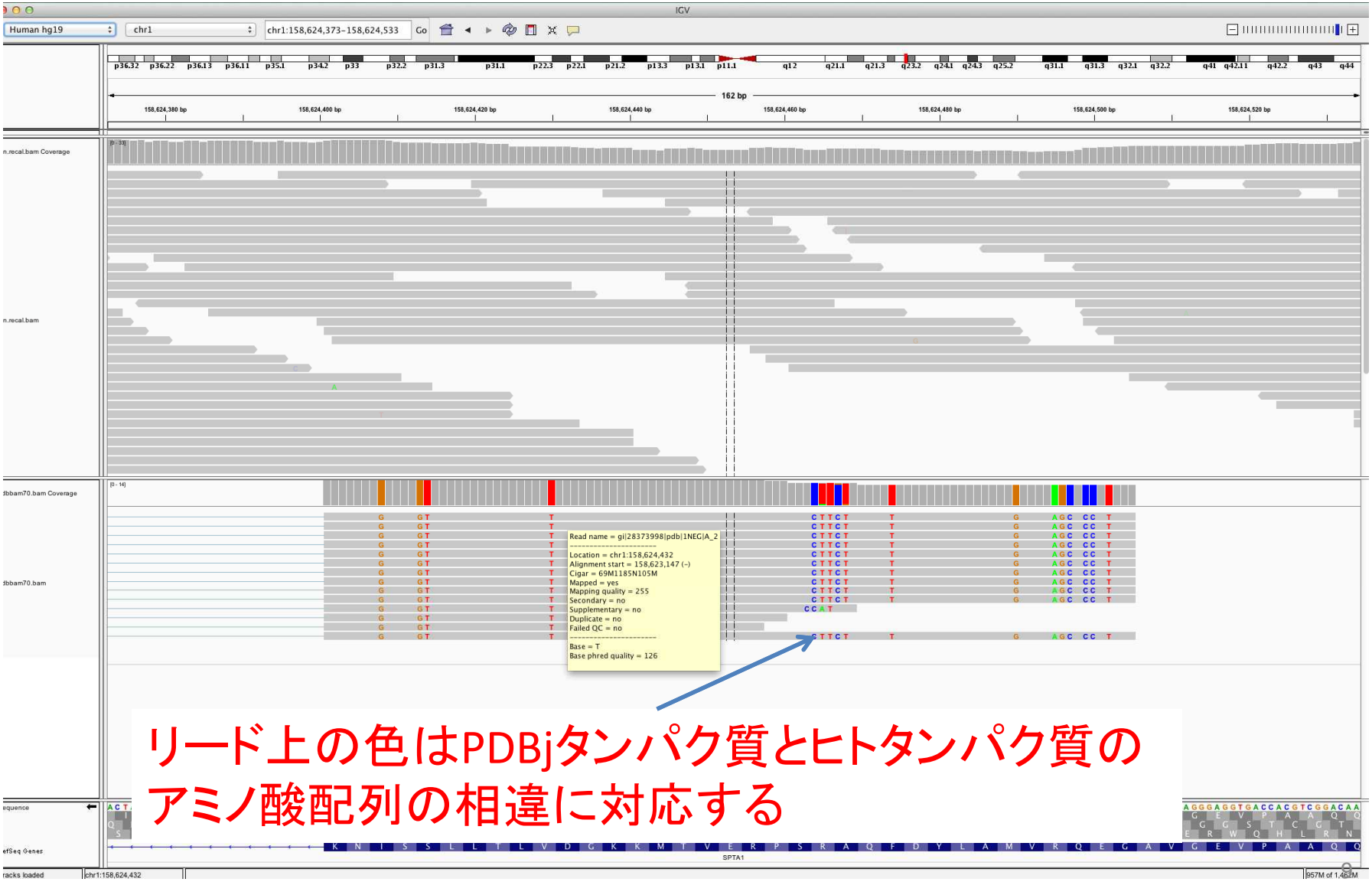


PDBの3FE4が当たっている
100%の一致率

リードにマウスをポイントすると
PDB IDなどの情報が表示される

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

変異を含んだ部位


IGV
Human hg19
chr1
chr1:158,624,373-158,624,533
Go
162 bp
Read name = gi|28373998|pdb|1NEG|A_2
Location = chr1:158,624,432
Alignment start = 158,623,147 (-)
Cigar = 69M1185N105M
Mapped = yes
Mapping quality = 255
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no
Base = T
Base phred quality = 126
リード上の色はPDBjタンパク質とヒトタンパク質の
アミノ酸配列の相違に対応する

P53領域

DNA結合ドメインが最も多く構造解析されている