平成30年度博士学位論文

ヒト薬剤代謝関連遺伝子にみられる一塩基置換が

タンパク質の機能と構造に及ぼす影響の

生命情報学を用いた研究

お茶の水女子大学大学院

人間文化創成科学研究科 ライフサイエンス専攻

疾患予防科学領域

坂本美佳

平成31年3月

目　次

論文要旨

ヒト薬剤代謝関連遺伝子にみられる一塩基置換が タンパク質の機能と構造に及ぼす影響の生命情報学を用いた研究

坂本美佳

近年, 塩基配列の解析技術が向上し, ヒトゲノムに存在する遺伝子の多様性（バリアント）が数多く明らかになった. ヒトゲノム配列バリアントデータベースgnomADや薬物代謝関連遺伝子データベースPharmVarなどに遺伝子のバリアントの情報が登録されている. 米国国立生物工学情報センター（National Center for Biotechnology Information, NCBI）のデータベースClinVarにはヒトゲノムのバリアントと疾患との関係が記載されているが, ごく最近見つかったバリアントについては疾患との関連についての情報はまだ十分ではない. 個人の全ゲノム配列を対象とする解析, またはゲノムの1~2%に相当しタンパク質の情報をもつ領域である全エキソンを対象とした解析による薬剤の効果や疾患原因を探る検査が行われるようになり, 影響未知のバリアントが多く発見されることも問題となっている.

本研究ではヒトゲノム配列にみられる多様性のうち, 薬剤の輸送と代謝に関連する遺伝子に着目した. 薬物が細胞内で代謝されるとき, 初めに細胞内への薬物の輸送, 薬物が酸化・還元・加水分解をうける第1相反応, 極性基の転移をうける第2相反応, 最後に代謝された薬物の細胞外への排出がおこる. 細胞内への輸送はSLCトランスポーターやチャネルが行い, ヒトシトクロムP450 (P450) は第1相反応に関与し, 多くの転移酵素が第2相反応に関与する. 細胞外への排出は主にABCトランスポーターが行う.

P450には, 多くのアイソザイムが存在し, 薬物代謝のほか生体内のステロイド合成などに関わっている. P450をコードする遺伝子群には多くのバリアントが見つかっている. それらのバリアントの出現頻度に地域差があることが知られている. また, それらのバリアントにより薬物動態の違いが生じ, 薬物の投与量に影響を与える事例があることも知られている.また, P450バリアントについてはこれまで多くの研究が行われてきた。単一の遺伝子についての研究が多く, P450全体を俯瞰して解析したものは少ない. 本研究では, P450遺伝子ミスセンスバリアントの全体像を俯瞰することで, P450遺伝子ミスセンスバリアントに共通してみられる特徴を明らかにすることをめざした. はじめに, 地域の違いによるミスセンスバリアントの発生箇所の偏りを検討した. その結果,

P450遺伝子ミスセンスバリアントのうち3つの地域（東アジア系，アフリカ系，ヨーロッパ系）で出現頻度が異なるバリアントでは，リシンからメチオニンへの置換およびアルギニンからプロリンへの置換は，出現頻度の地域差が存在する割合が高い傾向がみられた．基質認識部位とヘム結合領域にバリアントがおこる割合には，一部の領域を除き，3つの地域での出現頻度の違いによる差がなかった．P450遺伝子ミスセンスバリアントは一定の割合で中立的に起こり，タンパク質の機能に重要なアミノ酸残基を置換する場合，地域の外的環境の違いなどの要因によりミスセンスバリアントの残りやすさが異なることが示唆された．

つぎに，P450遺伝子ミスセンスバリアントのタンパク質機能への影響予測を試みた．既知のタンパク質への影響ありバリアントとの距離の情報およびタンパク質相互作用に関わるアミノ酸残基との距離の情報によるランダムフォレスト法を用いた影響予測モデルを構築した．この影響予測モデルは，既存のミスセンスバリアント影響予測法と比較しても良い成績であった．gnomAD exomeに存在するP450遺伝子ミスセンスバリアントのうち影響未知のものに影響予測モデルを適用したところ，約1/3のミスセンスバリアントがタンパク質への影響ありバリアント候補と予測された．タンパク質の立体構造データから得られるアミノ酸残基間の空間的情報がP450遺伝子ミスセンスバリアントの影響予測に役立つ可能性を示唆した．

さらに，薬物の代謝物の細胞外排出にかかわるタンパク質であるABCトランスポーターについて疾患原因となるバリアントと立体配座の関係を解析した．ABCトランスポーターのアポ型とATP結合型を含む各種立体構造を比較し，膜貫通ドメイン中の分子内回転の中心点となるアミノ酸残基の周りの立体配座変化を見出した．また，この立体配座変化と，病原性バリアントの位置を比較し，バリアントによる立体配座変化の障害がATPの結合と膜表面相互作用を弱めることで疾患を引き起こす可能性を明らかにした．

本研究の成果は，病原性バリアントがどのような機能をもつかを解明するため，立体構造データを利用した新しいアプローチである．そして，遺伝子バリアントがもたらす個人差を明らかにし，疾患原因や薬物の効果などの個人差に対応した疾患予防や個別化医療を推進するための一助となることが期待される．

外国語要旨

Bioinformatic analysis of the effect of single nucleotide polymorphism of human drug metabolism genes on function and conformation of protein

Mika Sakamoto

Recently, the technique for analyzing the genomic sequence has been improved, and we found numerous variation of the human genome and the genes. This study focused on the genomic variation of cytochrome P450 and ABC transporter. Human cytochrome P450 (P450) is an enzyme that is associated with the oxidative metabolism of a large number of xenobiotics and endogenous organic compounds. P450 is known for its diversity, and we found numerous genetic polymorphism in genes coding P450. Genetic polymorphism of P450 has been extensively studied with regard to alteration in enzymatic activity. The distribution of genetic polymorphism of P450 shows significant difference among geological regions. Genetic polymorphism is an important factor that influences drug concentrations and has the potential to predict the optimal dosage of drug in personalized medicine. GnomAD, database of human genetic variant, and PharmVar, a database of pharmacogene variant, contains information of many variant of P450 genes. ClinVar, a database of human genetic variation with information of clinical significance, also contains variant information of of many P450 genes, but there are few information on variants that are identified.

In this study, I focused on genes related to drug transport and metabolism among the diversity found in the human genome sequence. When drugs are metabolized in cells, cytochrome P450 (P450) oxidizes, reduces, hydrolyzes the drug, ABC transporter pumps the metabolite out of the cell.

This study was to clarify the characteristics of genetic variants of P450 genes included in gnomAD exome. First, the relationship between the characteristics of amino acid substitution and the minor allele frequencies in three human population; East Asian, African and Caucasian, was examined. The result showed that there were no significant differences in the missense variant with different allelic frequencies in the three human population. Next, to clarify the three-dimensional structural features of the variation, the amino acid substitution sites by missense variant of P450 genes were mapped on the three-dimensional structures of P450 proteins. Features of the amino acid substitution sites were acquired from the coordinates of

amino acid residue on the three-dimensional structure, and the information of protein-protein interaction related to redox and known pathogenic/drug response mutation. With classification by principal component analysis, the features of missence variants of P450 genes whose clinical significance is known was analyzed. The result of principal component analysis was able to classify the variants by the features computed from the coordinates of three-dimensional structure, such as the distance of known pathogenic/drug response mutation to the target residue and the distance of the residue related to protein-protein interaction to the target residue. Furthermore, the set of features of known pathogenic and drug response mutation of P450 genes computed from the coordinates of three-dimensional structure was used to train classifiers using several algorithms of machine learning, such as logistic regression classifier, support vector machine classifier and random forests classifier. The accuracy of these classifiers were compared with one another and it was found that the random forests model was the best model to classify variants of P450. The pathogenicity and drug responsibility of the missense variants of P450 genes of unknown clinical significance were predicted using the random forests classifier. Approximately one third of missense variant of P450 genes that were unknown for clinical significance were classified as pathogenic or drug responsible. This result may lead that the information of three-dimensional structure of protein is a valid source for predicting the effect of missense variant of P450 genes.

ABC transporter family is a huge group in the transporter membrane proteins and actively transports the substrates using the energy derived from ATP hydrolysis. A variation of a single amino acid in the amino acid sequence of ABC transporter has been known to be linked with certain disease. The mechanism of the onset of the disease by the variation is, however, still unclear. I compared the structures of ABC transporter in apo and ATP-binding forms and found a possible conformation shift around pivot-like residues in the transmembrane domains. When this conformation change in ABC transporter and the location of pathogenic variation were compared, I found a reasonable match between the two, explaining the onset of the disease by the variation.

This study provided a new approach using three-dimensional structure data to clarify the function of the pathogenic variant. And it is expected to clarify individual differences caused by genetic variants and to promote preventive and personalized medicine.

第一章
序論

1-1. はじめに

近年, 次世代シークエンサー（以下, NGS）と呼ばれる高品質かつ一度に大量の塩基配列を解読することができる機器が普及し, NGSを用いた疾患の診断が広く行われるようになった[1]. 本邦においても, 未疾患診断イニシアチブIRUDというプロジェクトにより, 従来の医学的検査では診断のつかない稀な疾患や難病の診断や原因解明のために, NGSによる塩基配列解読が用いられている[2, 3]. よくある病気に罹患しやすいかどうかを知るために Direct-to-consumer 遺伝子検査とよばれる一般消費者向けの遺伝子検査も広く行われるようになった[4]. 一般消費者向けの遺伝子検査の結果と医師の適切なフォローにより, 疾患にかかりにくいように生活習慣を変えることができた例[5]もあり, 今後も遺伝子レベルの検査はますます普及すると考えられる.

疾患の診断に用いられる検査法として, 全エキソーム解析（Whole Exome Sequence Analysis, WES）が用いられている. 全ゲノム配列の約1~2%に相当する, タンパク質の情報をもつ部分を解読する方法である[6]. 全ゲノム解析に比較するとコストが少なく, 診断に広く用いられている[7, 8].

1-2. ヒトゲノム配列の多様性とデータベース

多くの人のゲノム配列が解読されることで, ヒトゲノム塩基配列のバリエーションがみつかっている. 従前はゲノム配列のバリエーションを「多型（polymorphism）」と呼んでいたが, polymorphism には「疾患原因とならない変化」と「集団内に1%以上の頻度で固定された変化」という語義の混乱があり, 米国臨床遺伝・ゲノム学会（American College of Medical Genetics and Genomics, ACMG）の配列多様性解釈についてのガイドライン（ACMGガイドライン）およびHuman Genome Variation Society（HGVS）の勧告ではpolymorphismの代わりに中立的な用語としてvariantを用いるよう提唱されている[9, 10]. 同様に「変異（mutation）」という用語も「突然変異」, 「疾患の原因となるゲノム配列の変化」という語義の混乱があり, ACMGガイドラインおよびHGVSの勧告では中立的な用語としてvariantを用いるよう提唱されている[9, 10]. ACMGガイドライン, HGVSの勧告および日本人類遺伝学会『遺伝学用語の改訂』（2009年11月改訂, http://jshg.jp/about/notice-reference/）を踏まえ, 本論文中ではゲノム配列のバリエーションを「多様性」とし, 「疾患原因との関連がわかっている変化」「疾患原因との関連がわかっている変化」を区別せずに「バリアント」とした.

ゲノム配列の多様性には，交叉および組換えなどの染色体レベルの多様性と，遺伝子（塩基配列）レベルの多様性があるが，本論文では染色体レベルの多様性については言及しない．遺伝子レベルの多様性には1箇所の塩基の多様性（一塩基置換），1塩基以上の挿入及び欠失，数塩基〜100塩基程度の繰り返しの多様性（マイクロサテライト，ミニサテライト），1000塩基以上の繰り返しの多様性（コピー数異常，Copy number variation, CNV）がある[6]．遺伝子領域におこる一塩基置換には，置換前後にアミノ酸配列の変化がおこらないサイレントバリアント（同義置換），置換によってアミノ酸配列が変化するミスセンスバリアント（非同義置換），置換によって終止コドンになるナンセンスバリアント，スプライス部位に置換がおこりスプライシング異常をおこすスプライスサイトバリアントがある[6]．本論文では，遺伝子領域におこる一塩基置換のうち置換によってアミノ酸配列が変化するミスセンスバリアントに注目し，他の種類のバリアントは取り上げていない．

ヒトゲノム配列の多様性を記録している国際的なデータベースとして，1000 Genomes Projectがある．2008年から開始され，2010年にパイロット版[11]が公表されて以来，1,092人のデータ（Phase 1）[12]を経て，最新では26の地域別ヒト集

団の2,054人のデータ（Phase 3）[13, 14]が公開されている．また，エキソームのデータをもとにした国際的なヒトバリアントデータベースとして，米国ブロード研究所のExome Aggregation Consortiam (ExAC) がある．これは，地域別コホートのデータと，疾患特異的なエキソームプロジェクトのデータを集めたデータベースであり，60,706人のデータから構成されている[15]. 2018年現在，このプロジェクトはGenome Aggrigation Database (gnomAD, 「ノマド」と読む) に引き継がれており，ExACよりも多い123,136人のエキソームデータと，15,496人のゲノムデータからなる[15]. 国内のヒトバリアントデータベースには，京都大学のHuman Genetic Variation Database (HGVD) があり，日本人1208人のエキソームデータと，3248人のジェノタイピング結果を基につくられている[16, 17]. また，東北メディカルメガバンクのIntegrative Japanese Genome Variation（iJGVD）は東北地方（宮城県，岩手県）在住のヒトを中心としたゲノムデータを基につくられており，2018年現在は近畿地方，九州地方のコホートデータを加えた3,554人のデータとなっている[18, 19].

本研究では，疾患に対する関連性や薬物代謝への影響に関する情報のデータベースとしてClinVar[20]を利用した. ClinVarは，米国国立生物工学情報センター

（National Center for Biotechnology Information, NCBI）で運営されているバリアントと疾患の関連性についてのデータベースである．バリアントの染色体上の位置と形式，遺伝子名，その疾患関連性についての解釈（clinical significance），情報提供者と根拠となる情報などがまとめられている[20]. ClinVarの遺伝性疾患に関するclinical significanceの内容は，情報提供者がACMGの配列多様性解釈についてのガイドライン（ACMGガイドライン）[9]，培養細胞などを用いた機能解析実験および過去の文献情報を根拠とした解釈に従っている[20]. Clinical significanceの分類には，分類根拠の確かさにより順位付けがあり，病原性バリアントではPathogenic > Pathogenic/Likely pathogenic > Likely pathogenicの順に確かさが低下し，無害なバリアントではBenign > Benign/Likely benign > Likely benignの順に確かさが低下する[9]．また，薬理遺伝学的解析による薬剤に対する反応の多様性については，The Clinical Pharmacogenetics Implementation Consortium (CPIC) の勧告に基づき，薬剤に対する反応の表現型が報告されているバリアントにdrug responseというアノテーションが付与されている[21]．本論文では，遺伝性疾患関連の病原性バリアント（ClinVarのclinical significance でPathogenic , Pathogenic/Likely pathogenic , Likely pathogenicとされているバリアント）と薬理

遺伝学的解析による薬物反応に関連するバリアント（ClinVarのclinical significance でdrug responseとされているバリアント）を「タンパク質へなんらかの影響をあたえるバリアント」として考えた.

1-3. 薬物代謝に関わる遺伝子の多様性

生体に薬物が投与されたとき, 生体内では, 初めに投与部位からの吸収, 次に体内への拡散, 酵素による代謝, 最後に体外への排出がおこる. 薬物代謝とは「生体内の酵素による薬物の化学構造の変化」と定義される[22]. 水溶性の薬物は代謝を受けずにそのまま排出されるが, 脂溶性の薬物は代謝を受けて水溶性となり排出される. 代謝により多くの薬物は不活性化されるが, 薬物によっては代謝により薬効を生じる場合や, 生体に対して毒性を増してしまう場合もある[22].

細胞レベルでみたときは, 初めに細胞内への薬物の輸送, 次に薬物代謝反応, そして細胞外への排出がおこる. 図1-1に肝細胞での薬物代謝の模式図を示した. 薬物の細胞内への輸送に関わるタンパク質としてSLCトランスポーターやチャネルタンパク質がある. 薬物代謝反応は第1相反応と第2相反応に大別される

[22]. 第1相反応は酸化還元反応または加水分解反応であり, 水酸基やカルボキシル基などの極性を有する官能基が薬物に導入される. 第1相反応に関わる主なタンパク質としてシトクロムP450（P450）があげられる[22]. 第2相反応は抱合反応であり, グルクロン酸や硫酸などの極性の大きな官能基が薬物に導入されるため, 薬物は水に溶けやすくなり体外へ排出されやすくなる. 第2相反応に関わるタンパク質としてはグルクロノトランスフェラーゼやスルホトランスフェラーゼなどがあげられる[22].細胞外への排出に関わる主なタンパク質として, ABCトランスポーターがあげられる[22].

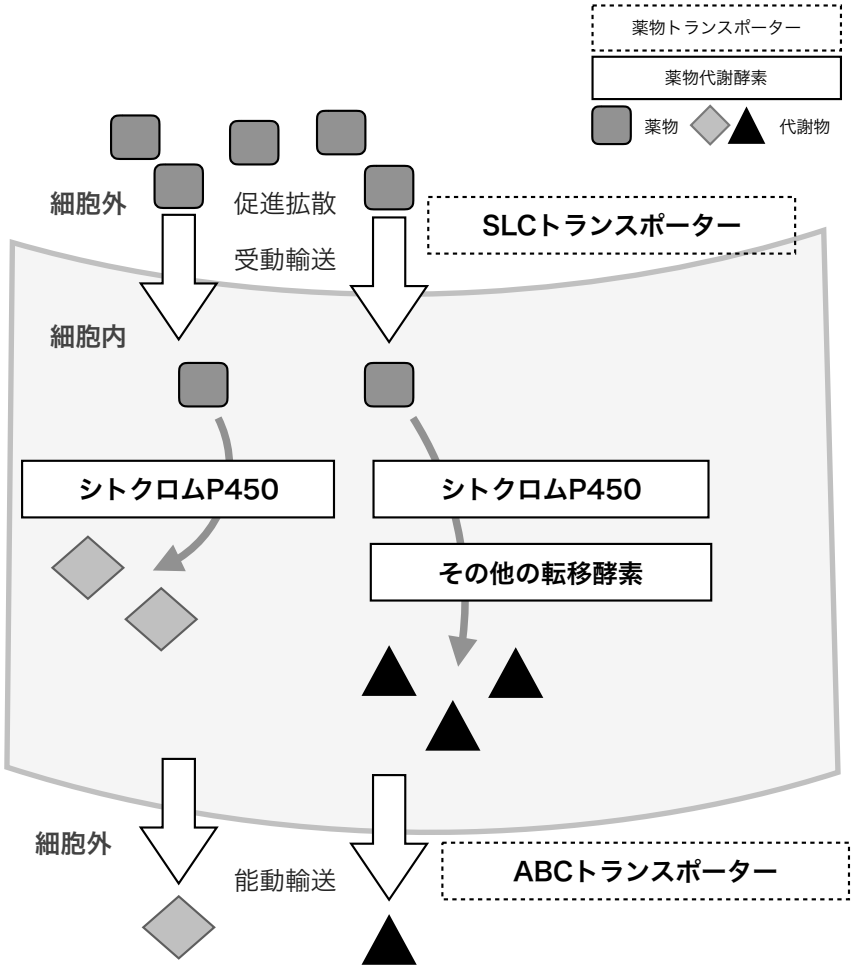

図1-1 肝細胞での薬物代謝（模式図）

肝細胞の内部に薬物が輸送され, 代謝されて細胞外に排出されるまでの流れを模式図として示した.

薬物代謝反応には地域差や個人差が見られる場合があり，その原因の一つとして薬物代謝酵素やトランスポータータンパク質の塩基配列やアミノ酸配列の多様性（バリアント）がある[22, 23]. 地域差や個人差による代謝能力または輸送能力の違いにより，なんらかの疾患の薬物療法を行うときに，正常型酵素をもつヒトで決められた標準的な薬物投与量は，患者のもつ遺伝子多様性によっては過不足を起こす場合もある[24, 25].

近年，次世代シークエンサーの普及によるヒト塩基配列データの蓄積および解析法の進歩により，疾患原因遺伝子を含む多くの遺伝子バリアントが明らかになった[26, 27]. 薬物代謝関連遺伝子にも多くのバリアントがあり，治療薬の副作用予測や投与量調節を目的とした遺伝学的検査が行われるようになった[28–31]. しかし，すべての薬物代謝関連遺伝子バリアントについて疾患に関する感受性が明らかになっているわけではないため[21]，遺伝学的検査の解釈にあたっては，意義不明のバリアントが数多く検出されることが問題となっている[32]. 薬剤投与により誘導される遺伝子[33]については，バリアントと薬物に対する反応性の関連を検証することが難しい[34]. 意義不明のバリアントに対し疾患に関する感受性や薬物に対する反応性を明らかにすることは疾患の原因

解明に役立つだけでなく, 薬物療法の副作用による障害を予防することにも役立つと考えられる[21, 26, 35].

1-4. 本研究の目的と構成

本研究では, 薬物代謝関連遺伝子にみられる, コドンが別のアミノ酸を指定するものに変わるバリアント（ミスセンスバリアント）がタンパク質の機能にどのような影響を与えているのかを探るため, 公共データベースに存在するヒト遺伝子バリアントの情報とバリアントの影響に関する情報を利用したデータ駆動型の解析を行った. Higuchiら（2018）によって, 細胞内への薬物輸送に関わるSLCトランスポーターのミスセンスバリアントと病原性に関する考察はすでに行われており[36], 本研究では薬剤代謝反応にかかわるP450と細胞外への代謝物排出にかかわるABCトランスポーターのミスセンスバリアントに注目した. 第二章では, P450遺伝子のミスセンスバリアントについて地域による出現頻度の偏りとアミノ酸置換の起こりやすい位置などの傾向を明らかにするため, 約12万人のエキソームデータを集めたバリアントデータベースgnomAD exomeに存在するP450遺伝子のミスセンスバリアントを抽出し, 3つの地域(東アジア, ア

フリカおよびヨーロッパ)での出現頻度に顕著な違いがあるものを見出したことを報告した. そして, それらのバリアントから引き起こされるアミノ酸置換のパターンや, 置換が起こった部位の特徴を, 3つの地域での出現頻度に顕著な違いがあるバリアントと出現頻度に顕著な違いがないバリアントで比較した. 第三章では, P450遺伝子のミスセンスバリアントについて, P450の立体構造から得られる情報によるバリアント影響予測を試みた. P450と電子供与体の複合体構造解析データ[37]を基に得られた情報, および既知のタンパク質への影響をもつバリアントとの位置関係を用いた主成分解析を行った. そして, タンパク質への影響をもつバリアントと無害なバリアントを分ける特徴を明らかにした. その結果を踏まえ, 機械学習によるタンパク質への影響をもつバリアントの予測モデルを構築し, 影響未知バリアントの判別を試みた. 第四章では, ABCトランスポーターについて, 結合している分子の違いによるタンパク質分子内での空間的な原子の配置（立体配座）の変化と病原性バリアントの関係を考察した. 第五章では, 本論文の結論として本研究で得られた成果についてまとめ, さらに本研究からつながる将来の研究への展望について述べた.

1-5. 本研究で用いたデータの入手先について

本研究で用いたデータの入手先を表1-1に示した. 遺伝子リストは米国国立生物工学情報センター（National Center for Biotechnology Information, NCBI）のデータベース Gene (https://www.ncbi.nlm.nih.gov/gene/) から得た. 遺伝子バリアントのデータは, 米国マサチューセッツ工科大学・ハーバード大学ブロード研究所（Broad Institute of MIT and Harvard）によるデータベース Genome Aggregation Database （gnomAD, http://gnomad-old.broadinstitute.org）のダウンロードサイト（http://gnomad-old.broadinstitute.org/downloads）からデータファイル (https://storage.googleapis.com/gnomad-public/release/2.0.2/vcf/exomes/gnomad.exomes.r2.0.2.sites.vcf.bgz)を入手した. ClinVar のアノテーションデータは, NCBIのデータベース ClinVar（https://www.ncbi.nlm.nih.gov/clinvar/）のダウンロードサイト（ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/）からデータファイル（variant_summary.txt.gz）を入手した.

表1-1 本研究で用いたデータの入手先

項目	データベース名称	提供元	ウェブサイトURL	データサイズ	確認した日付
遺伝子リスト	Gene	米国国立生物工学情報センター	https://www.ncbi.nlm.nih.gov/gene/	25 KB	20180514
遺伝子バリアント	Genome Aggregation Database（gnomAD)	米国マサチューセッツ工科大学・ハーバード大学ブロード研究所	http://gnomad-old.broadinstitute.org	10.12 GB (圧縮)	20180523
バリアントの病原性・薬物反応性	ClinVar	米国国立生物工学情報センター	https://www.ncbi.nlm.nih.gov/clinvar/	28.7 MB (圧縮)	20180511

1-6. gnomAD exomeについて

本研究では, gnomADデータベースのエキソームデータをバリアントのデータベースとして利用した. gnomAD exomeは，リファレンスゲノムとしてGRCh37(hg19)を利用している[15]. gnomAD exome のVCF形式のデータはバリアント抽出のためのソフトウェアGATK HaplotypeCaller[38] によって作られたものである[15]. VCFとはVariant call formatの略で，ヒトゲノムなどのリファレンス配列に対する塩基配列の多様性を記述するためのフォーマットである[39]. ヘッダーにはVCFを作成したソフトウェア，付加される情報についての説明が記述され，その後に各バリアントが1箇所ごとに記述されている．データ列はバリアントのある1箇所にデータ1行分が対応し, 1列目は染色体, 2列目はバリアン

ト箇所の塩基番号（ただし染色体の塩基配列の始まりを1とする）, 3列目はdbSNPのアクセッション番号（dbSNPに登録されていないバリアントについては「.」と記載）, 4列目はリファレンス配列の塩基, 5列目はバリアントの塩基, 6列目はバリアント抽出の品質, 7列目はバリアント抽出の品質によるフィルタリング結果, 8列目以降にバリアント箇所の情報が記載されている[39]. gnomAD exomeのVCFの場合, 8列目の情報（地域別アレル頻度, 地域別のアレル数, 遺伝子名, トランスクリプト名, タンパク質名, アミノ酸置換など）はVariant Effect Predictor (VEP) [40]によりGENCODE v19[41]のヒトゲノム遺伝子アノテーションが追加されたものである. 追加されたアノテーションは, 8列目の「CSQ=」以降に以下のような順序で記載されている. 1.アレルの塩基, 2. Sequence ontrogy[42]の表記法によるバリアントの種別と, 遺伝子のどの領域にあるかの記述, 3.バリアントによる影響の強さ, 4.Symbol,　5.Ensembl geneのアクセッション番号, 6~7.Ensembl transcriptのアクセッション番号, 8.トランスクリプトの性質, 9.エキソンの位置（該当のエキソン/総エキソン数）, 10. Human Gene Variation Society (HGVS) のバリアントの記載法[10]に基づく塩基置換, 11. HGVSのバリアントの記載法に基づくタンパク質のアミノ酸置換, 12.cDNAの位置, 13.コーディング配

列の位置, 14.タンパク質のアミノ酸配列上の位置, 15.アミノ酸置換, 16.コドン, さらに17番目以降にはSIFT , PolyPhen-2によるタンパク質機能障害の予測結果, 地域別のマイナーアレル頻度などの記載がある. ひとつのバリアント箇所に複数のトランスクリプトが該当する場合はコンマ（,）で区切られて別のトランスクリプトの情報が記載されている. これらのアノテーションを手がかりにして, バリアントが個々の遺伝子に対してどのような変化をもたらすかを知ることができる. gnomAD exomeのデータには疾患に対する関連性や薬物代謝への影響など表現型に関する情報は記載されていないため, 表現型に関する情報はほかのデータベースから得る必要がある.

第二章

ヒトシトクロムP450遺伝子にみられる
ミスセンスバリアントの地域差

概要

ヒトシトクロムP450（P450）遺伝子の多様性に地域による偏りが生じた過程を明らかにすることをめざし，ミスセンスバリアントの出現頻度に地域差が生じやすい位置，アミノ酸置換の型のうちP450で地域差が生じやすい型を調べることで，バリアントの発生と地域分布の差にアプローチすることを試みた．ミスセンスバリアントによるアミノ酸置換のうち一部は，出現頻度の地域差が存在する割合が高い傾向があった．基質認識部位およびヘム結合領域のミスセンスバリアントは地域の違いにかかわらず一定の頻度でおこることが示唆された．したがって，P450のミスセンスバリアントの地域差は遺伝的浮動と民族移動によるボトルネック効果によって生じたことが考えられるが，本研究で扱ったミスセンスバリアントの出現頻度は,時系列を考慮したものではないため，今後ミスセンスバリアントの出現頻度がどのように変化していくかを観察していく必要がある．

2-1. 背景

2-1-1. ヒトのシトクロムP450

P450は薬物代謝の第1相反応（薬物の酸化，還元と加水分解）にかかわる酵素

である[43, 44]. P450の命名については, 1987年に命名法に関する最初の報告がなされて以来,いくつかの改定を経て現在の形式に至っている[44] . 遺伝子のシンボルをCytochromeのCy, P450のPをとってCYPとし, 遺伝子を示す時は斜体, タンパク質を示す時は立体（正体）, つづく数字でファミリーを, その次のアルファベットでサブファミリー, さらに数字でひとつのタンパク質を示すルールで命名されている[44]. ヒトでは57個の遺伝子（偽遺伝子を除く）, 18ファミリー, 41サブファミリーがある[43, 44]. P450タンパク質の場合, アミノ酸配列で40%以上一致すると同一のファミリーに分類され, 55%以上一致すると同一のサブファミリーに分類される[44, 45]. P450のアミノ酸配列の類似度が高いほどP450の基質特異性の傾向はよく対応している[46, 47]. 表2-1にヒトでみられるP450ファミリーについて, 機能と細胞内局在を示した. CYP1は多環芳香族炭化水素, ハロゲン化炭化水素, 複素環式化合物, 芳香族アミンの代謝に関与する. CYP2, CYP3は薬物や環境化学物質の代謝, CYP4は脂肪酸代謝, CYP5はエイコサノイドの一種であるトロンボキサン合成に関与する. CYP7, CYP11, CYP17, CYP21, CYP24, CYP51はコレステロールやステロイド代謝に関与する. CYP8はプロスタサイクリンの合成に関与する[44]. 薬物や外来の化学物質の分解にかかわるCYP2ファ

ミリーと, 内在性の物質であるステロイド代謝に関わるCYP51ファミリーを比較すると, 基質認識部位でのアミノ酸残基保存率は, CYP51ファミリーに比べてCYP2ファミリーで低く, P450には, 基質特異性の厳密さと基質認識部位でのアミノ酸保存性に相関があると考えられている[45, 46].

図2-1に肝細胞におけるP450の反応機構の模式図を示す. P450の反応機構には5つのステップがあり, 1) P450への基質の結合, 2) 電子供与体からの電子の供給およびヘムの鉄イオンの還元, 3) 還元されたヘムの鉄イオンと分子状酸素の結合, 4) 電子供与体からの電子の供給および活性酸素の生成, 5) 代謝物および水分子の生成である[48]. 図2-2にP450の電子伝達の模式図を示した. P450タンパク質と電子供与体との相互作用には主に2種の機構があり, 真核生物のミトコンドリアで発見されたFADドメインを含む還元酵素と鉄硫黄タンパク質を介するタイプと, 真核生物のミクロソームにあるFADおよびFMNドメインを含む還元酵素を介するタイプである. 真核生物のミクロソームではシトクロムb_5を介するタイプも存在する[47, 49–51]. 土壌微生物 *Pseudomonas putida* から見つかったP450cam[52]は還元酵素と鉄硫黄タンパク質を介するタイプの電子伝達を行うものであり, 電子供与体Putidaredoxinとの複合体のNMRによる構造解析例が存在

する[37]. また, ヒトP450ではミトコンドリアに局在するCYP11A1と電子供与体Adrenodoxinとの複合体のX線構造解析が行われている[53].

表2-1 ヒトP450の機能と局在部位

ファミリー	機能[44, 54]	細胞内局在[44, 54]
CYP1	多環芳香族炭化水素，ハロゲン化炭化水素，複素環式炭化水素，芳香族アミンの代謝	小胞体
CYP2	薬物や環境化学物質の代謝	小胞体
CYP3	薬物や環境化学物質の代謝	小胞体
CYP4	脂肪酸水酸化	小胞体
CYP5	トロンボキサン合成	小胞体
CYP7	コレステロール水酸化	小胞体
CYP8	プロスタサイクリン合成	小胞体
CYP11	コレステロール,ステロイド水酸化，アルドステロン合成	ミトコンドリア
CYP17	ステロイド水酸化	小胞体
CYP19	アンドロゲン芳香化	小胞体
CYP20	不明	不明
CYP21	ステロイド水酸化	小胞体
CYP24	ビタミンD3水酸化	ミトコンドリア
CYP26	レチノイン酸水酸化	小胞体
CYP27	ステロイド水酸化，ビタミンD3水酸化	ミトコンドリア
CYP46	コレステロール水酸化	小胞体
CYP51	ステロール脱メチル化	小胞体

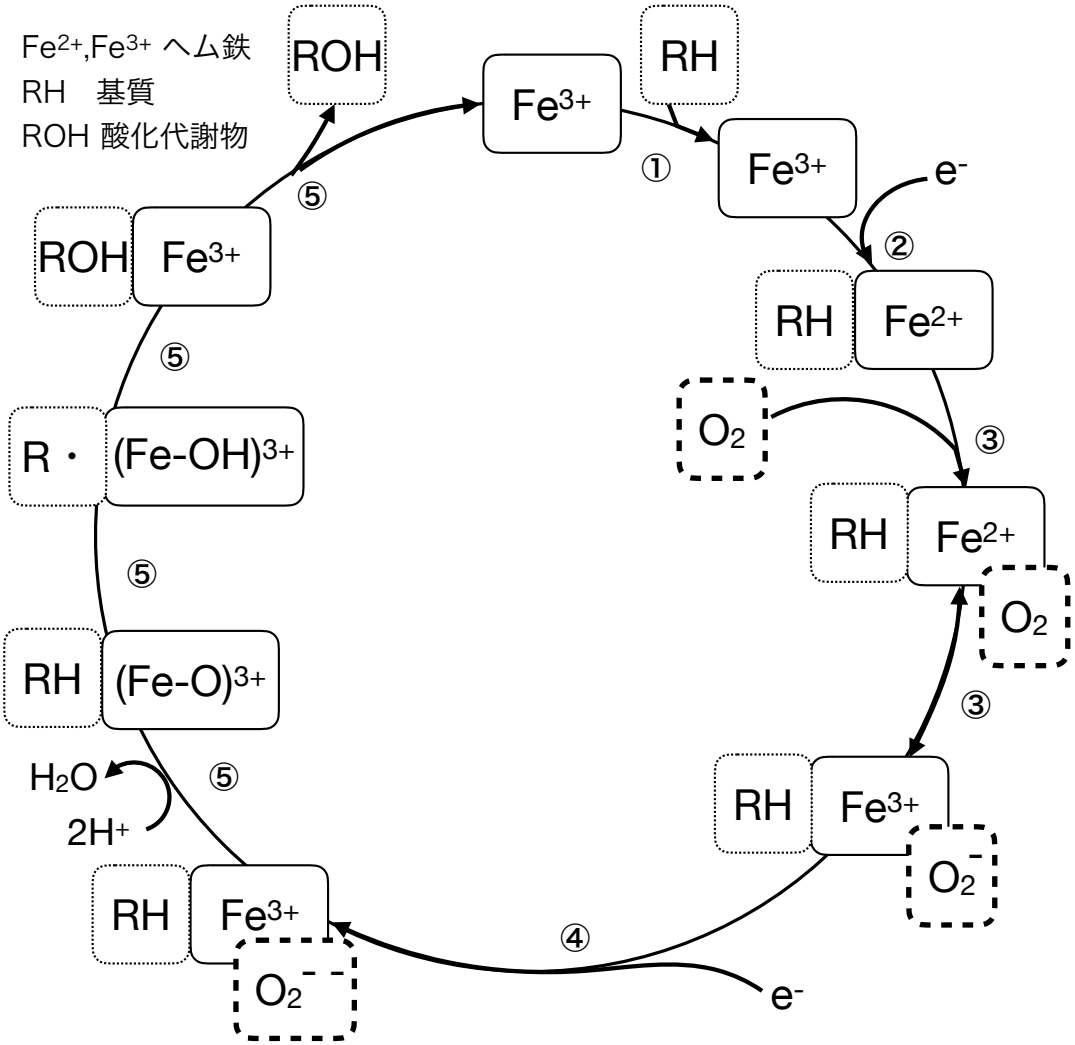

図2-1 肝臓（肝細胞）におけるP450反応機構（模式図）

肝臓（肝細胞）におけるP450の反応機構を模式的に示した．図中の丸囲み数字は反応のステップを表し，①P450への基質の結合，②電子供与体からの電子の供給およびヘムの鉄イオンの還元，③還元されたヘムの鉄イオンと分子状酸素の結合，④電子供与体からの電子の供給および活性酸素の生成，⑤代謝物および水分子の生成である[48].

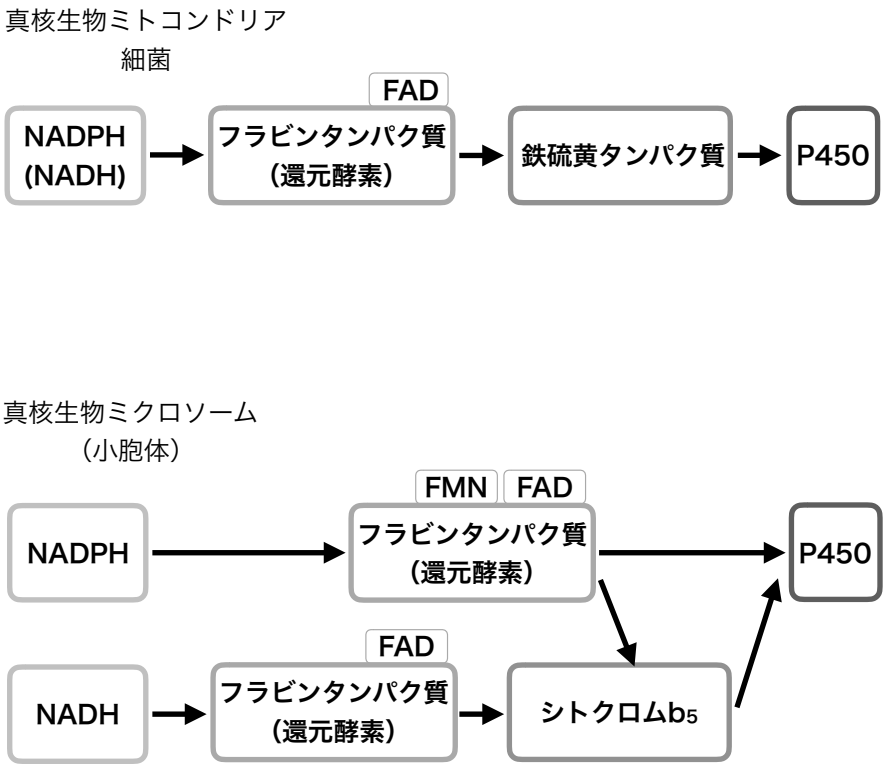

図2-2　P450の電子伝達のしくみ

P450が還元されるときの電子伝達の模式図. P450タンパク質と電子供与体との相互作用には主に2種の機構があり, 真核生物のミトコンドリアで発見されたFADドメインを含む還元酵素と鉄硫黄タンパク質を介するタイプと, 真核生物のミクロソームにあるFADおよびFMNドメインを含む還元酵素を介するタイプである. 真核生物のミクロソームではシトクロム b_5を介するタイプも存在する[47, 49–51].矢印は電子の流れを示す. NADPH：ニコチンアミドアデニンヌクレオチドリン酸, NADH：ニコチンアミドアデニンヌクレオチド, FMN：フラビンモノヌクレオチド, FAD：フラビンアデニンジヌクレオチド.

2-1-2. ヒトP450の多様性

P450には酵素活性の変化を伴う多様性が存在することが知られている．特定の形質に関与するゲノム配列上の位置を座位といい，ある座位におけるゲノム配列の多様性（一塩基置換，欠失，挿入，遺伝子コピー数の多様性を含む）をアレルという．ひとつの相同染色体上のアレルの組み合わせをハプロタイプと呼び, P450各遺伝子に見られるハプロタイプについては, 1999年にThe Human Cytochrome P450 (CYP) Allele Nomenclature Committee が設立され, 運営するデータベースThe Human Cytochrome P450 (CYP) Allele Nomenclature Database [55] にまとめられた．このデータベースは, 2018年現在, The Pharmacogene Consortium の運営するPharmVar（https://pharmvar.org/）に移管され, 遺伝子配列の多様性と医薬品の作用について研究されているP450以外の遺伝子とともにハプロタイプの名称, 塩基置換, アミノ酸置換と機能に関する情報がまとめられている[56]．このデータベースでは, P450のハプロタイプの命名は，アミノ酸置換のないものを遺伝子名の後に*1をつけて表し，データベースへの登録順に*2, *3…と, *のあとの数字で識別する．

P450の酵素活性の変化を伴う多様性の例として, *CYP2D6*と *CYP2C9*の多様性

について述べる. *CYP2D6**10は，アミノ酸配列の34番目のプロリンがセリンに置換されるバリアントと，アミノ酸配列の486番目のセリンがスレオニンに置換されるバリアントをもつハプロタイプである. *CYP2D6**10はブフラノールのヒドロキシ化反応の速度とデキストロメトルファンの脱メチル化反応の速度が通常の酵素活性を持つ型に比べて低下することが知られている[57]. また，*CYP2C9**3は，アミノ酸配列の359番目のイソロイシンがロイシンに置換されるバリアントをもつハプロタイプである. *CYP2C9**3は, *S*-ワルファリンおよびトルブタミドのヒドロキシ化の速度が通常の酵素活性を持つ型に比べて低下することが知られている[58]. P450の多様性の出現頻度には地域差がある[59–67]. Kubotaら（2001）によれば, *CYP2D6**10の出現頻度は日本で38.6％，中国で50.7%であり，他の地域（トルコ，サウジアラビア，ドイツ，エチオピア）の出現頻度に比べて4~26倍高かった[60]. Myrandら（2008）は, *CYP2C9**3は日本（3.5%）韓国（3.5%）に比べ欧米（5.6%）に多いことを報告した[68]. Otaら（2008）が行った日本人1,017人を対象にしたP450の遺伝子型の分布に関する研究では，*CYP2D6**10の日本人での出現頻度はコーカソイドでの出現頻度より多く，*CYP2D6**10アレルをホモ（酵素活性欠如型）またはヘテロ（酵素活性低下型）

でもつ割合は60.5%であった．それゆえ, CYP2D6によって代謝されることで効果を発揮するタモキシフェンを日本で投与する場合は，投与量を欧米より多くする，または別の薬物に替えるなどの調整が必要であることを示した[69].

P450多様性の特徴についての知見は，薬物による副作用の予防や個人の遺伝的特徴に合わせた薬物療法を行うために不可欠である[70]．本研究では，どのような理由でP450の多様性に地域差による偏りが生じたかを解明することをめざし，ミスセンスバリアントの出現頻度に地域差が生じやすい位置，アミノ酸置換の型のうちP450で地域差が生じやすい型に注目し，バリアントの発生と地域分布の差のしくみにアプローチすることを試みた．

2-2. 手法

2-2-1. 地域で出現頻度に差があるP450遺伝子のミスセンスバリアント

図2-3に解析のフローチャートを示す．表2-2に本研究で用いたヒトP450遺伝子を示した．ヒトP450遺伝子のミスセンスバリアントは, gnomAD exomeデータのアノテーションに記載されている情報のうち，バリアント解析ソフトウェアであるGenomeAnalysisToolkit（GATK）[38]のクオリティフィルタリング[71]を

通過していたかどうか, 遺伝子名（Symbol）, バリアントの種類（一塩基置換であり, アミノ酸置換が起こるミスセンスバリアントであること）および表2-2に示すゲノム情報データベースEnsembl[72]のP450タンパク質のアクセッションIDの文字列をgnomAD exomeのデータファイルから検索し, 該当する文字列のある行を抽出することで得た. gnomAD exomeのデータファイルには各地域のミスセンスバリアントの出現数と地域ごとの全サンプル数が記載されていた. その数値を用いて, 東アジア（EAS）, アフリカ（AFR）およびフィンランドを除くヨーロッパ（NFE）のミスセンスバリアントの出現数と地域ごとの全サンプル数の比の差をFisherの正確確率検定法[73, 74]を用いて検定した. 得られたp値に多重比較補正であるBenjamini-Hochberg補正を行なった値を補正後p値とし, 補正後$p<0.05$を統計的有意差ありとした. ただし, EAS, AFR, NFEのバリアントの出現数がすべて0である箇所については操作から除外し検定を行わなかった. Fisherの正確確率検定法で有意差ありとされたバリアントを地域差ありミスセンスバリアント, 有意差がなかったミスセンスバリアントを地域差なしミスセンスバリアントとした. EAS, AFR, NFEのバリアントの出現数がすべて0である箇所にあるミスセンスバリアントは地域差あり, 地域差なしのどちらにも含めていな

いが, 全アミノ酸置換数に含めた.

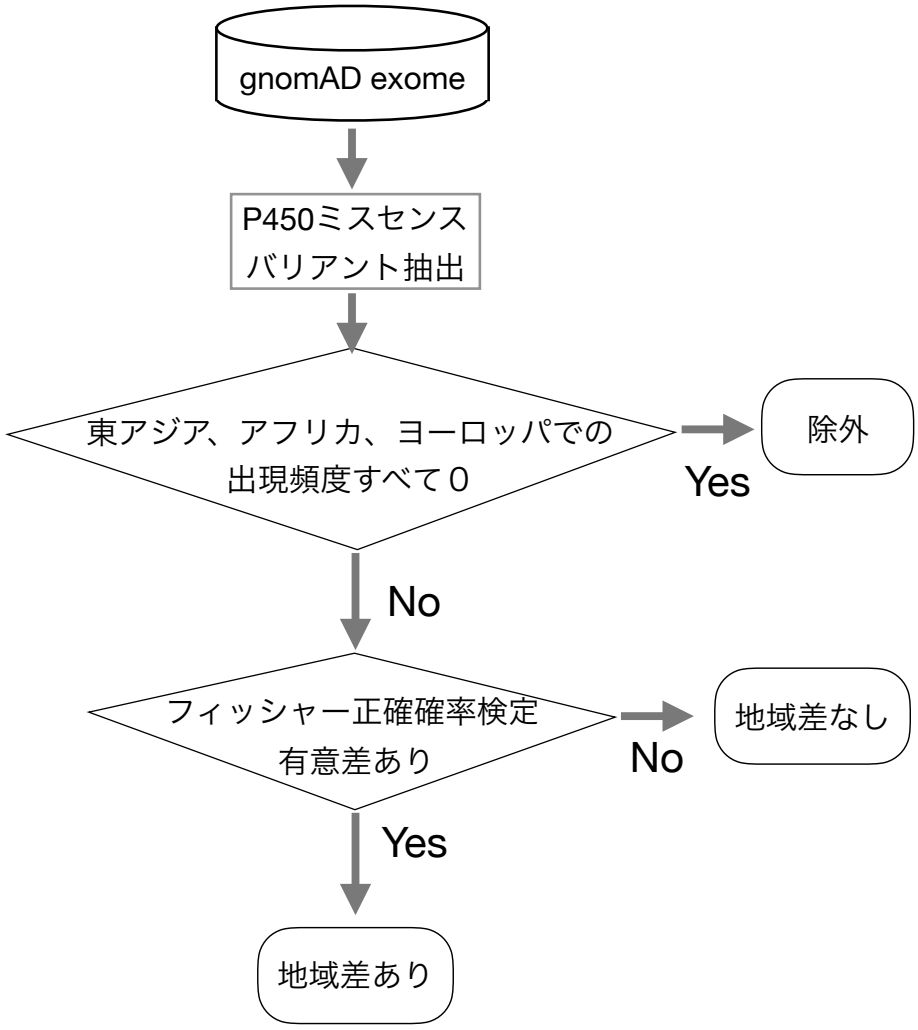

図2-3　第二章の解析フローチャート

表2-2　本研究で用いたヒトP450遺伝子

Symbol	Ensembl protein
CYP1A1	ENSP00000369050.3
CYP1A2	ENSP00000342007.4
CYP1B1	ENSP00000260630.3
CYP2A13	ENSP00000332679.1
CYP2A6	ENSP00000301141.4
CYP2A7	ENSP00000301146.4
CYP2B6	ENSP00000324648.2
CYP2C18	ENSP00000285979.6
CYP2C19	ENSP00000360372.3
CYP2C8	ENSP00000360317.3
CYP2C9	ENSP00000260682.6
CYP2D6	ENSP00000353820.5
CYP2E1	ENSP00000252945.3
CYP2F1	ENSP00000333534.2
CYP2J2	ENSP00000360247.3
CYP2R1	ENSP00000334592.5
CYP2S1	ENSP00000308032.3
CYP2U1	ENSP00000333212.6
CYP2W1	ENSP00000310149.7
CYP3A4	ENSP00000337915.2
CYP3A5	ENSP00000222982.4
CYP3A7	ENSP00000337450.2

CYP3A43	ENSP00000346887.2
CYP4A11	ENSP00000311095.4
CYP4A22	ENSP00000360958.3
CYP4B1	ENSP00000271153.4
CYP4F2	ENSP00000221700.3
CYP4F11	ENSP00000384588.2
CYP4F12	ENSP00000448998.1
CYP4F22	ENSP00000269703.1
CYP4F3	ENSP00000221307.6
CYP4F8	NA[1)]
CYP4V2	ENSP00000368079.4
CYP4X1	ENSP00000360968.3
CYP4Z1	ENSP00000334246.3
CYP7A1	ENSP00000301645.3
CYP7B1	ENSP00000310721.3
CYP8B1	ENSP00000318867.4
CYP11A1	ENSP00000268053.6
CYP11B1	ENSP00000292427.4
CYP11B2	ENSP00000325822.2
CYP17A1	ENSP00000358903.3
CYP19A1	ENSP00000379683.1
CYP20A1	ENSP00000348380.4
CYP21A2	ENSP00000408860.2
CYP24A1	ENSP00000216862.3
CYP26A1	ENSP00000224356.4

CYP26B1	ENSP00000001146.2
CYP26C1	ENSP00000285949.5
CYP27A1	ENSP00000258415.4
CYP27B1	ENSP00000228606.4
CYP27C1	ENSP00000334128.7
CYP39A1	ENSP00000275016.2
CYP46A1	ENSP00000261835.3
CYP51A1	ENSP00000003100.8
PTGIS (CYP8A1)	ENSP00000244043.3
TBXAS1 (CYP5A1)	ENSP00000263552.6

1) Ensemblデータベース上に, *CYP4F8*遺伝子に対応するタンパク質が存在しなかった

2-2-2. ミスセンスバリアントによるアミノ酸残基の置換

P450遺伝子のミスセンスバリアントによるアミノ酸置換型（置換前のアミノ酸-置換後のアミノ酸で示す）は以下のようにして数え上げた．地域差ありミスセンスバリアント，地域差なしミスセンスバリアントの2群について，それぞれ置換前のアミノ酸-置換後のアミノ酸の各組み合わせの出現数を数えた．遺伝子の塩基配列上の1箇所に複数のアミノ酸置換型がある場合は，それらを独立に数えた．アミノ酸置換の全出現数には，EAS, AFR, NFEのバリアントの出現数がすべて0である箇所のアミノ酸置換を含めて数えた．

つぎに，アミノ酸残基を表2-3に示す2つの観点，アミノ酸残基の極性（分類1）とアミノ酸残基の電荷（分類2），に注目して分類し，以下の比較を行った．1)親水性残基から疎水性残基への置換，疎水性残基から親水性残基への置換，疎水性残基から疎水性残基への置換，親水性残基から親水性残基への置換にわけてそれぞれの出現数を数えあげ，全出現数に対する割合を求めた．同様に，地域差ありミスセンスバリアントについても，それぞれのアミノ酸置換のすべての地域差ありミスセンスバリアントの数に対する割合を求めた．地域差なしミスセンスバリアントについてもそれぞれのアミノ酸置換のすべての地域差ありミスセンスバリアントの数に対する割合を求めた．2)酸性残基から酸性残基への置換，

酸性残基から中性への置換, 酸性残基から塩基性への置換, 中性から酸性残基への置換, 中性から中性への置換, 中性から塩基性への置換, 塩基性から酸性残基への置換, 塩基性から中性への置換, 塩基性から塩基性への置換に分け, それぞれの出現数を数えあげ, 全出現数に対する割合を求めた. 同様に, 地域差ありミスセンスバリアントについても, それぞれのアミノ酸置換のすべての地域差ありミスセンスバリアントの数に対する割合を求めた. 地域差なしミスセンスバリアントについてもそれぞれのアミノ酸置換のすべての地域差ありミスセンスバリアントの数に対する割合を求めた.

$$P_{all} = \frac{S_{all}}{A}$$

$$P_1 = \frac{S_1}{Q}$$

$$P_2 = \frac{S_2}{R}$$

P_{all}：ある分類に属するすべてのミスセンスバリアント出現数のすべてのミスセンスバリアントのアミノ酸置換全数に対する割合を, P_1：ある分類に属する地域差ありミスセンスバリアント出現数の地域差ありミスセンスバリアント全数に対する割合を, P_2：ある分類に属する地域差なしミスセンスバリアント出現数の地域差なしミスセンスバリアント全数に対する割合を, A：すべてのミスセンス

バリアントのアミノ酸置換の全数を, Q：地域差ありミスセンスバリアントのアミノ酸置換の全数を, R：地域差なしミスセンスバリアントのアミノ酸置換の全数を, S_{all}：ある分類に属する全出現数を, S_1：ある分類に属する地域差ありミスセンスバリアントの出現数を, S_2：ある分類に属する地域差なしミスセンスバリアントの出現数を意味する. アミノ酸置換型の出現率の差についてはFisherの正確確率検定法を用いて検定した. さらに, アミノ酸置換の全てのタイプ（リファレンス配列によって指定されるアミノ酸20種×バリアントにより置換されるアミノ酸20種）について地域差なしミスセンスバリアントに対する地域差ありミスセンスバリアントのオッズ比を調べた. 以下の式で, あるアミノ酸置換型についての地域差ありミスセンスバリアントと地域差なしミスセンスバリアントのオッズ比を計算した.

$$OR = \frac{(x/A)/\left(1-(x/A)\right)}{(y/A)/\left(1-(y/A)\right)}$$

OR：あるアミノ酸置換型についての地域差ありミスセンスバリアントと地域差なしミスセンスバリアントのオッズ比を, x: あるアミノ酸置換型の地域差ありミスセンスバリアント出現数を, y: あるアミノ酸置換型の地域差なしミスセンスバリアント出現数を, A: あるアミノ酸置換型のすべてのミスセンスバリアン

ト出現数を意味する.

表2-3　アミノ酸残基の分類

アミノ酸残基の分類	アミノ酸残基
分類 1)	
疎水性残基	アラニン，グリシン，イソロイシン，ロイシン，メチオニン，フェニルアラニン，プロリン，トリプトファン，バリン
親水性残基	アルギニン，アスパラギン，アスパラギン酸，システイン，グルタミン，グルタミン酸，ヒスチジン，リシン，セリン，トレオニン，チロシン
分類 2)	
塩基性残基	アルギニン，ヒスチジン，リシン
中性残基	アラニン，グリシン，イソロイシン，ロイシン，メチオニン，フェニルアラニン，プロリン，トリプトファン，バリン，システイン，セリン，トレオニン，チロシン，アスパラギン，グルタミン
酸性残基	アスパラギン酸，グルタミン酸

2-2-3. ミスセンスバリアントの位置

P450のアミノ酸配列は，ゲノム情報データベースEnsembl [72]のアクセッション番号から, Ensembl REST API（https://rest.ensembl.org）を用いて取得した．それぞれのアミノ酸配列を用い，タンパク質立体構造データベースPDB[75]に対して, BLAST[76] による相同性検索を行った．各アミノ酸配列の検索結果第1位の配列がもつ立体構造を「対応するタンパク質の立体構造」とした.

基質認識部位（Substrate Recogunision Site，以下, SRS）は以下のように推定した. CYP2ファミリーについて行われた研究[77]をもとに, P450ファミリーごとにP450タンパク質のアミノ酸配列を用いたMUSCLE v3.8.31[78] による多重配列アラインメントを行い, SRS1~SRS6の位置を推定した．ヘム結合領域の推定は，タンパク質ドメインと機能部位データベースPROSITEである[79]にP450のヘム結合領域の配列として登録されているPS00086（CYTOCHROME_P450）の配列,

[FW]-[SGNH]-x-[GD]-{F}-[RKHPT]-{P}-C-[LIVMFAP]-[GAD]

をP450のアミノ酸配列に対してアラインメントした結果と目視で探すことで行った．すべてのミスセンスバリアント，地域差ありミスセンスバリアント，地域差なしミスセンスバリアントについて，バリアントの位置がSRS1~SRS6あるい

はヘム結合領域領域に存在するもの, SRS1~SRS6とヘム結合領域領域のどちらにも属さないものの出現数を数えあげた. SRS1~SRS6およびヘム結合領域に存在するミスセンスバリアントの出現率について, 地域差ありミスセンスバリアントの出現率と地域差なしミスセンスバリアントの出現率の違いをFisherの正確確率検定法で検定した.

2-3. 結果

2-3-1. 地域で出現頻度に差があるP450遺伝子のミスセンスバリアント

gnomAD exomeのデータからP450遺伝子領域51,932箇所のバリアントを得た. さらにバリアント解析ソフトウェアであるGenomeAnalysisToolkit（GATK）[38]のクオリティフィルタリング[71]を通過したミスセンスバリアントとして14,371箇所のミスセンスバリアントを得た. 図2-4はミスセンスバリアント抽出の概略である. このうち, 3種以上のアリルが1座位で発生している場合は2,335箇所あった. 以下, 1座位で複数のミスセンスバリアントが存在する場合を考慮し, ミスセンスバリアントの数を示す場合は 「〜件」という表記を使う. 東アジア, アフリカ, フィンランドを除くヨーロッパの3つの地域で出現頻度が異な

るミスセンスバリアント（地域差ありミスセンスバリアント）は2071件（1701箇所）であった.

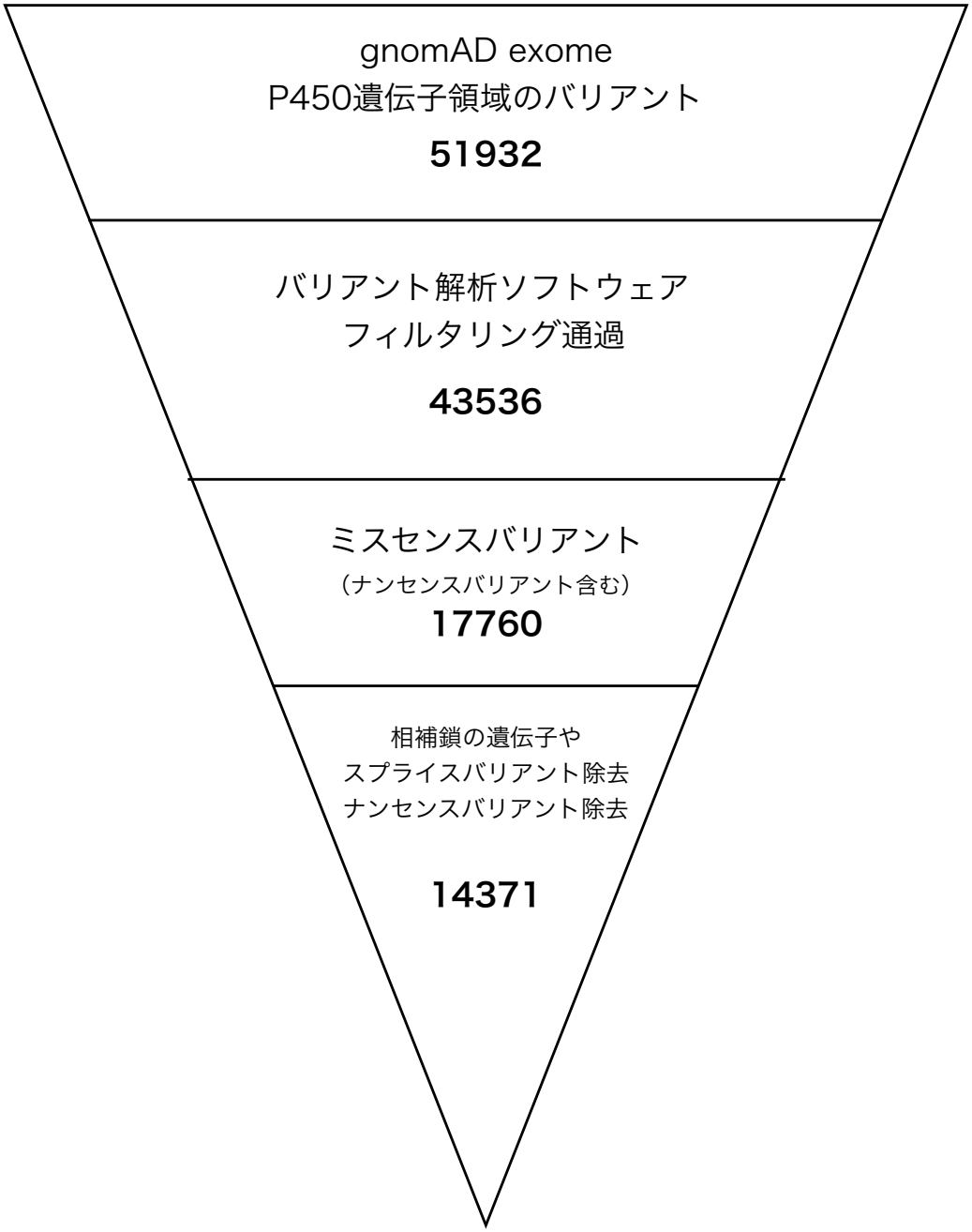

図2-4　P450ミスセンスバリアントの抽出

gnomAD exomeのデータからP450ミスセンスバリアントの抽出を行なった概略を示した．各段階の数値は，その操作によって得られたバリアント箇所の数である．

2-3-2. アミノ酸置換型の比較

アミノ酸置換によるアミノ酸残基の化学的性質の変化を地域差ありミスセンスバリアントと地域差なしミスセンスバリアントで比較した. 親水性残基から疎水性残基への置換および疎水性残基から親水性残基への置換の場合では, 地域差ありミスセンスバリアントと地域差なしミスセンスバリアントで顕著な違いは認められなかった（図2-5左）. 同様に, 疎水性残基残基から疎水性残基残基への置換および親水性残基から親水性残基への置換でも地域差ありミスセンスバリアントと地域差なしミスセンスバリアントで顕著な違いは認められなかった（図2-5左）. 次に酸性残基から酸性残基への置換, 酸性残基から中性への置換, 酸性残基から塩基性への置換, 中性から酸性残基への置換, 中性から中性への置換, 中性から塩基性への置換, 塩基性から酸性残基への置換, 塩基性から中性への置換および塩基性から塩基性への置換を地域差ありミスセンスバリアントと地域差なしミスセンスバリアントで比較した. 酸性残基残基から中性残基への置換および塩基性残基から中性残基への置換では, 地域差ありミスセンスバリアントでの出現率と地域差なしミスセンスバリアントでの出現率に統計的な有意差があった（図2-5右, 酸性残基残基から中性残基への置換 $p = 0.004182$, 塩

基性残基から中性残基への置換 $p = 1.204 \times 10^5$）.

アミノ酸置換の全てのタイプ（リファレンス配列によって指定されるアミノ酸20種×バリアントにより置換されるアミノ酸20種）について地域差なしミスセンスバリアントに対する地域差ありミスセンスバリアントのオッズ比を調べた. その結果, 図2-6および表2-4に示したように, 地域差なしミスセンスバリアントに対する地域差ありミスセンスバリアントのオッズ比が大きいアミノ酸置換型は, リシンからメチオニンへの置換, アルギニンからプロリンへの置換であり, これらのアミノ酸置換型は出現頻度の地域差が存在する割合が高い. そのほかのアミノ酸置換型については地域差ありミスセンスバリアントと地域差なしミスセンスバリアントで顕著な違いは認められなかった.

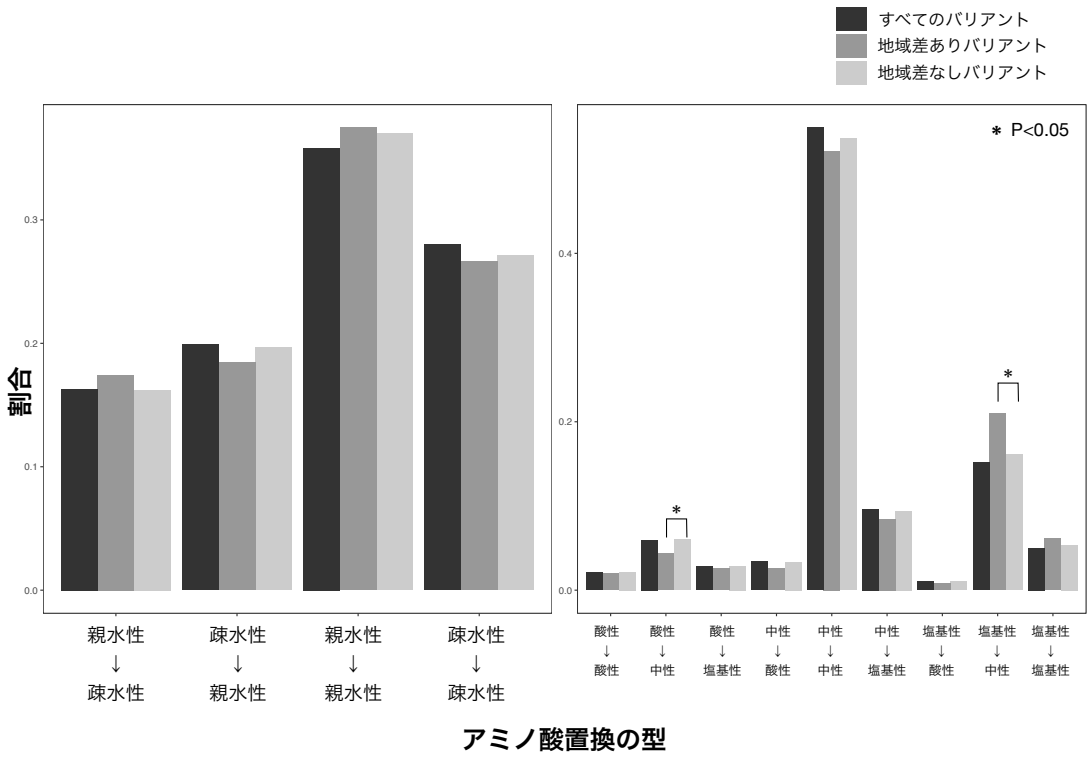

図2-5 P450遺伝子のミスセンスバリアントによるアミノ酸の化学的性質の変化

縦軸はそれぞれのバリアント数に対するアミノ酸置換型の割合. 左図：親水性残基-疎水性残基の変化, アミノ酸残基の極性が変わる置換, 極性が変わらない置換にわけて各分類群の出現数を数え, 全出現数に対する割合を求めた. 地域差あり（なし）ミスセンスバリアントについても, 各分類群の地域差あり（なし）ミスセンスバリアントのすべての地域差あり（なし）ミスセンスバリアントの数に対する割合を求めた. 右図：酸性残基-中性残基-塩基性残基の変化. アミノ酸残基の電荷が変わる置換, 電荷が変わらない置換に分け, 各分類群の出現数を数え, 全出現数に対する割合を求めた. 地域差あり（なし）ミスセンスバリアントについても, 各分類群の地域差あり（なし）ミスセンスバリアントのすべての地域差あり（なし）ミスセンスバリアントの数に対する割合を求めた.アミノ酸置換の全出現数15950件, 地域差ありミスセンスバリアントの全出現数2071件, 地域差ありミスセンスバリアントの全出現数10226件 (1座位に複数種類のバリアントが観察されたものを含む）.

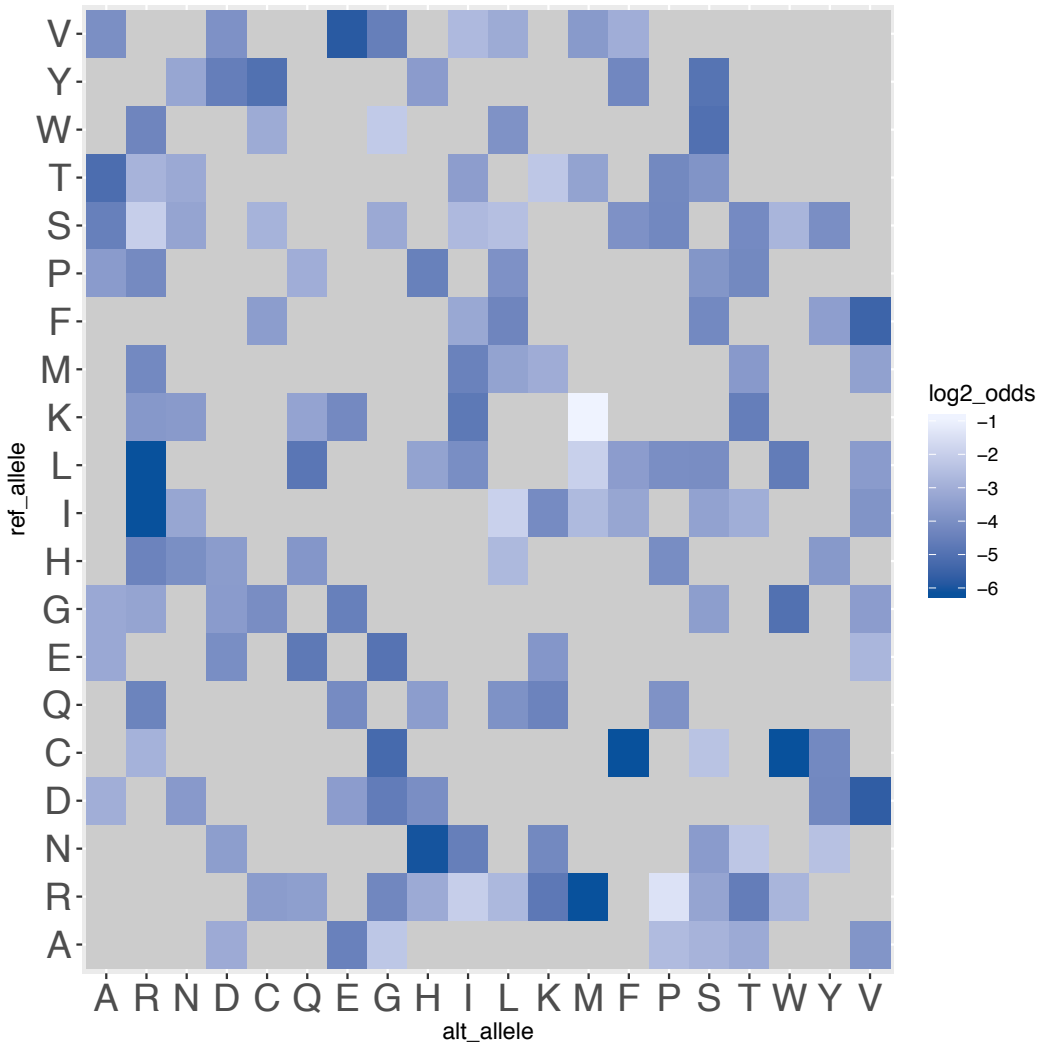

図2-6 アミノ酸置換型別の地域差のおこりやすさ

アルファベットは，一文字表記法でアミノ酸残基の種類を示す．縦軸はヒトゲノムリファレンス配列hg19によって指定されるアミノ酸残基，横軸はミスセンスバリアントにより置換されたあとのアミノ酸残基を示した．色の濃淡はあるアミノ酸置換型についての地域差ありミスセンスバリアントと地域差なしミスセンスバリアントの対数オッズ比（底を2とする）の大小を示し，色が薄くなるほど対数オッズ比の値が大きくなることを示す．対数オッズ比は以下の式で計算した．

$$OR = \frac{(x/A)/\big(1-(x/A)\big)}{(y/A)/\big(1-(y/A)\big)}$$

$$logOR = log_2(OR)$$

OR：あるアミノ酸置換型についての地域差ありミスセンスバリアントと地域差なしミスセンスバリアントのオッズ比, x : あるアミノ酸置換型の地域差ありミスセンスバリアント出現数, y : あるアミノ酸置換型の地域差なしミスセンスバリアント出現数, A : あるアミノ酸置換型のすべてのミスセンスバリアント出現数, *logOR*：あるアミノ酸置換型についての地域差ありミスセンスバリアントと地域差なしミスセンスバリアントの対数オッズ比. 図中の灰色は該当するアミノ酸置換型が存在しないことを示す.

表2-4　出現頻度の地域差が存在する割合が高いアミノ酸置換

置換のタイプ	対数オッズ比*
Lys→Met	-0.95
Arg→Pro	-1.44
Ile→Leu	-1.95
Leu→Met	-1.99

*同じバリアントの型の「地域差なしバリアントに対する地域差ありバリアントのオッズ比」を底を2とする対数で表した値

2-3-3. ミスセンスバリアントの位置

表2-5に, P450遺伝子とEnsemblデータベースのP450タンパク質のアクセッションIDおよび各遺伝子に対応するP450タンパク質の立体構造のPDB IDを示した. この対応は, BLASTによるアミノ酸配列相同性検索によって決定された. P450のアミノ酸配列において推定された基質認識部位およびヘム結合領域を表2-6, 図2-7に示した. 表2-7にミスセンスバリアントの基質認識部位およびヘム結合領域での出現率を示した. SRS4に存在する地域差ありミスセンスバリアントと地域差なしミスセンスバリアントの出現率には統計的有意差があった（P = 0.02534）. SRS1, SRS2, SRS3, SRS5, SRS6およびヘム結合領域に存在する地域差ありミスセンスバリアントと地域差なしミスセンスバリアントの出現率には統計的有意差は認められなかった.

表2-5　ヒトP450遺伝子スーパーファミリーと
バリアントのマッピングに用いた立体構造データ

Symbol	Ensembl protein	PDB
CYP1A1	ENSP00000369050.3	4I8V
CYP1A2	ENSP00000342007.4	2HI4
CYP1B1	ENSP00000260630.3	3PM0
CYP2A13	ENSP00000332679.1	4EJG
CYP2A6	ENSP00000301141.4	1Z11
CYP2A7	ENSP00000301146.4	4EJG
CYP2B6	ENSP00000324648.2	5UAP
CYP2C18	ENSP00000285979.6	5W0C
CYP2C19	ENSP00000360372.3	4GQS
CYP2C8	ENSP00000360317.3	2NNJ
CYP2C9	ENSP00000260682.6	5W0C
CYP2D6	ENSP00000353820.5	4WNW
CYP2E1	ENSP00000252945.3	3KOH
CYP2F1	ENSP00000333534.2	4EJG
CYP2J2	ENSP00000360247.3	3DL9
CYP2R1	ENSP00000334592.5	3DL9
CYP2S1	ENSP00000308032.3	3KW4
CYP2U1	ENSP00000333212.6	3DL9
CYP2W1	ENSP00000310149.7	1NR6
CYP3A4	ENSP00000337915.2	5TE8
CYP3A5	ENSP00000222982.4	5VEU
CYP3A7	ENSP00000337450.2	5TE8

CYP3A43	ENSP00000346887.2	5TE8
CYP4A11	ENSP00000311095.4	5T6Q
CYP4A22	ENSP00000360958.3	5T6Q
CYP4B1	ENSP00000271153.4	5T6Q
CYP4F2	ENSP00000221700.3	5T6Q
CYP4F11	ENSP00000384588.2	5T6Q
CYP4F12	ENSP00000448998.1	5T6Q
CYP4F22	ENSP00000269703.1	5T6Q
CYP4F3	ENSP00000221307.6	5T6Q
CYP4F8	NA[1)]	NA[1)]
CYP4V2	ENSP00000368079.4	5T6Q
CYP4X1	ENSP00000360968.3	5T6Q
CYP4Z1	ENSP00000334246.3	5T6Q
CYP7A1	ENSP00000301645.3	3V8D
CYP7B1	ENSP00000310721.3	3V8D
CYP8B1	ENSP00000318867.4	3B99
CYP11A1	ENSP00000268053.6	3N9Y
CYP11B1	ENSP00000292427.4	4FDH
CYP11B2	ENSP00000325822.2	4FDH
CYP17A1	ENSP00000358903.3	5IRV
CYP19A1	ENSP00000379683.1	5JL9
CYP20A1	ENSP00000348380.4	5VCG
CYP21A2	ENSP00000408860.2	5VBU
CYP24A1	ENSP00000216862.3	3K9V
CYP26A1	ENSP00000224356.4	2VE3

CYP26B1	ENSP00000001146.2	2VE3
CYP26C1	ENSP00000285949.5	2VE3
CYP27A1	ENSP00000258415.4	3K9V
CYP27B1	ENSP00000228606.4	3K9V
CYP27C1	ENSP00000334128.7	3K9V
CYP39A1	ENSP00000275016.2	6AY4
CYP46A1	ENSP00000261835.3	2Q9F
CYP51A1	ENSP00000003100.8	3LD6
PTGIS (CYP8A1)	ENSP00000244043.3	3B99
TBXAS1 (CYP5A1)	ENSP00000263552.6	5TE8

1) CYP4F8はgnomAD exomeのデータにミスセンスバリアントが存在していなかった

P450遺伝子とEnsemblデータベースのP450タンパク質のアクセッションIDおよび各遺伝子に対応するP450タンパク質の立体構造のPDB IDを示した．この対応は, BLASTによるアミノ酸配列相同性検索によって決定された.

表 2-6 基質認識部位とヘム結合領域の推定

Symbol	Ensembl protein	SRS1		SRS2		SRS3		SRS4		SRS5		Heme		SRS6	
		start	*end*	*start*	*end*	*start*	*end*	*start*	*end*	*start*	*end*	*start*	*end*	*start*	*end*
CYP1A1	ENSP00000369050.3	G105	P129	V218	G225	A250	Y259	K306	T324	S379	S389	F450	G459	T491	T497
CYP1A2	ENSP00000342007.4	G107	P131	V220	V227	R252	L261	K306	T324	S379	S389	F451	G460	T492	T498
CYP1B1	ENSP00000260630.3	D116	S138	L225	G232	V257	S269	N319	T337	S392	A402	F463	G472	S506	T510
CYP2A6	ENSP00000301141.4	G100	G121	L202	F209	A237	E245	N290	T308	G363	R373	F432	G441	S474	A481
CYP2A7	ENSP00000301146.4	G100	G121	L202	F209	A237	E245	N290	T308	G363	R373	F432	G441	S474	A481
CYP2A13	ENSP00000332679.1	G100	G121	L202	F209	A237	E245	N290	T308	G363	R373	F432	G441	S474	A481
CYP2B6	ENSP00000324648.2	G97	G118	L199	F206	V234	N242	N287	T305	S360	I370	F429	G438	T471	G478
CYP2C18	ENSP00000285979.6	G96	G117	M198	L205	L233	K241	S286	T304	I359	A369	F428	G437	T470	V477
CYP2C19	ENSP00000360372.3	G96	G117	M198	I205	L233	E241	N286	T304	I359	A369	F428	G437	T470	A477
CYP2C8	ENSP00000360317.3	G96	G117	M198	F205	L233	R241	N286	T304	S359	A369	F428	G437	T470	V477
CYP2C9	ENSP00000260682.6	G96	G117	M198	I205	L233	K241	S286	T304	I359	A369	F428	G437	T470	A477
CYP2D6	ENSP00000353820.5	D100	G125	L206	L213	V240	L248	N294	T312	G367	M377	F436	G445	R474	L484
CYP2E1	ENSP00000252945.3	G99	G119	M200	F207	V235	K243	G288	T306	I361	E371	F430	G439	S472	G479
CYP2F1	ENSP00000333534.2	G97	G118	I199	F206	I234	R242	T287	T305	A360	R370	F429	G438	T471	G478
CYP2J2	ENSP00000360247.3	N110	G131	L212	T219	L247	K255	N300	T318	G373	E383	F441	G450	K482	T488
CYP2R1	ENSP00000334592.5	D108	G130	I211	V218	L246	Y254	N299	N317	C372	A382	F441	G450	K482	T488
CYP2S1	ENSP00000308032.3	G100	G121	V202	L209	L237	A245	N291	T309	L364	T374	F433	G442	K475	F482
CYP2U1	ENSP00000333212.6	D149	G171	L252	L259	L287	T295	Y341	N359	T414	M424	F483	G492	K521	F528
CYP2W1	ENSP00000310149.7	D98	G119	L199	M206	V233	R241	N285	A303	I358	C367	F426	G435	T470	T476
CYP3A4	ENSP00000337915.2	N104	E124	V204	L211	V240	R243	E294	S312	F367	V376	F435	G444	K476	L483
CYP3A5	ENSP00000222982.4	N104	E124	V204	L211	L240	K243	E294	S312	F367	T376	F434	G443	K475	L482
CYP3A7	ENSP00000337450.2	N104	E124	V204	L211	V240	R243	E294	S312	F367	V376	F435	G444	K476	L483

CYP3A43	ENSP00000346887.2	N104	E124	L204	L211	L240	K243	E294	T312	F367	V376	F435	G444	K476	L483
CYP4A11	ENSP00000311095.4	H117	Q137	Y217	L224	T252	D264	D310	S328	Y383	E392	F450	G459	A492	V495
CYP4A22	ENSP00000360958.3	H117	Q137	Y217	L224	T252	D264	D310	S328	Y383	E392	F450	G459	A492	V495
CYP4B1	ENSP00000271153.4	P112	P132	Y211	L218	F246	D258	D304	S322	Y377	Q386	F446	G455	P488	V491
CYP4F2	ENSP00000221700.3	K122	D142	Y221	L228	F256	D268	D317	S335	H392	H401	F461	G470	P502	V505
CYP4F3	ENSP00000221307.6	K122	E142	Y221	L228	F256	D268	D314	S335	H392	C401	F461	G470	P502	V505
CYP4F11	ENSP00000384588.2	M122	D142	Y221	L228	F256	D268	D317	S335	H392	C401	F461	G470	P502	I505
CYP4F12	ENSP00000448998.1	N122	D142	Y221	L228	F256	D268	D317	S335	H392	C401	F461	G470	L502	I505
CYP4F22	ENSP00000269703.1	D130	D150	Y229	L236	F264	T276	D324	S342	Y399	Q408	F468	G477	P510	I513
CYP4V2	ENSP00000368079.4	M123	N141	Y219	M226	H254	N266	D318	A336	F392	S401	F460	G469	E502	I506
CYP4X1	ENSP00000360968.3	Q111	P131	Y212	L219	F247	D259	D305	A323	I378	D387	F447	G456	P488	I492
CYP4Z1	ENSP00000334246.3	A111	S131	Y211	L218	F246	E258	D303	S321	Y376	L385	F445	G454	R487	V490
TBXAS1	ENSP00000263552.6	N110	K131	V211	F218	K247	D250	E331	N349	Y405	E414	F473	G482	Q514	L521
CYP7A1	ENSP00000301645.3	W97	G119	L203	K210	M227	A239	K275	A293	S358	T365	F437	G446	L479	G487
CYP7B1	ENSP00000310721.3	F111	K130	S210	L217	L234	I246	I282	T300	S364	F371	F442	G451	L482	G490
CYP8B1	ENSP00000318867.4	R95	H120	Y182	T189	L220	V229	F275	T290	R351	L359	W433	G442	D476	G483
PTGIS	ENSP00000244043.3	T92	P117	Y177	E184	S221	M230	A276	A291	T352	E360	W434	G443	D477	G484
CYP11A1	ENSP00000268053.6	I123	S144	A230	I237	T262	F274	D315	M333	H388	Y397	F455	G464	T496	I500
CYP11B1	ENSP00000292427.4	L113	G134	S220	L227	V252	F264	A303	F321	Y376	V385	F443	G452	V484	I488
CYP11B2	ENSP00000325822.2	L113	G134	S220	L227	V252	F264	A303	F321	Y376	V385	F443	G452	V484	I488
CYP17A1	ENSP00000358903.3	G95	G118	Q199	I206	K231	N240	H291	S309	R364	H373	F435	G444	L473	V483
CYP19A1	ENSP00000379683.1	S118	N137	I227	A223	L240	K252	N295	V313	Q367	A377	F430	G439	I471	L479
CYP20A1	ENSP00000348380.4	E101	N120	E186	I192	K222	E225	Q265	K283	T340	I349	L401	E411	K443	T448
CYP21A2	ENSP00000408860.2	G91	S114	Y191	L199	R226	D235	H281	N299	R357	R367	F422	G431	L465	I471
CYP24A1	ENSP00000216862.3	I131	G152	A235	I242	V267	F279	E315	N333	T388	T397	F455	G464	L496	L501

CYP26A1	ENSP00000224356.4	S109	D130	F214	L221	G232	H244	A289	S307	N367	V376	F435	G444	K475	T476
CYP26B1	ENSP00000001146.2	S114	G135	Y214	V221	G232	Q244	E285	S303	F365	T374	F434	G443	T475	L476
CYP26C1	ENSP00000285949.5	R114	G135	F214	L221	G232	H244	E285	S303	L380	T389	F452	G461	Q493	T494
CYP27A1	ENSP00000258415.4	M130	G151	T237	I244	F267	F279	E324	N342	Y397	I406	F469	G478	V511	L516
CYP27B1	ENSP00000228606.4	F110	G131	T218	V225	P249	F261	S306	N324	Y381	V390	F448	G457	K490	L495
CYP27C1	ENSP00000334128.7	-	-	T81	L88	P115	F127	E169	F187	F242	V251	F311	G320	K353	L358
CYP39A1	ENSP00000275016.2	L95	K109	H191	V196	K217	E227	N265	P280	K339	V348	F407	A416	L449	V454
CYP46A1	ENSP00000261835.3	Y109	Y131	Q208	V215	Q239	R251	G291	N309	Y364	L373	F430	G439	Q471	T475
CYP51A1	ENSP00000003100.8	A133	N155	Y233	F240	R261	K267	E306	T324	R380	M389	F448	G457	T491	I494

基質認識部位（SRS）はCYP2ファミリーについて行われた研究[77]をもとに, Ensembl protein のアミノ酸配列とPDBから得たアミノ酸配列（表2-5）を用いた多重配列アラインメントにより推定された. ヘム結合領域は, タンパク質ドメイン・機能部位データベースPROSITEにP450のヘム結合領域の配列として登録されているPS00086の配列([FW]-[SGNH]-x-[GD]-{F}-[RKHPT]-{P}-C-[LIVMFAP]-[GAD])から推定された.

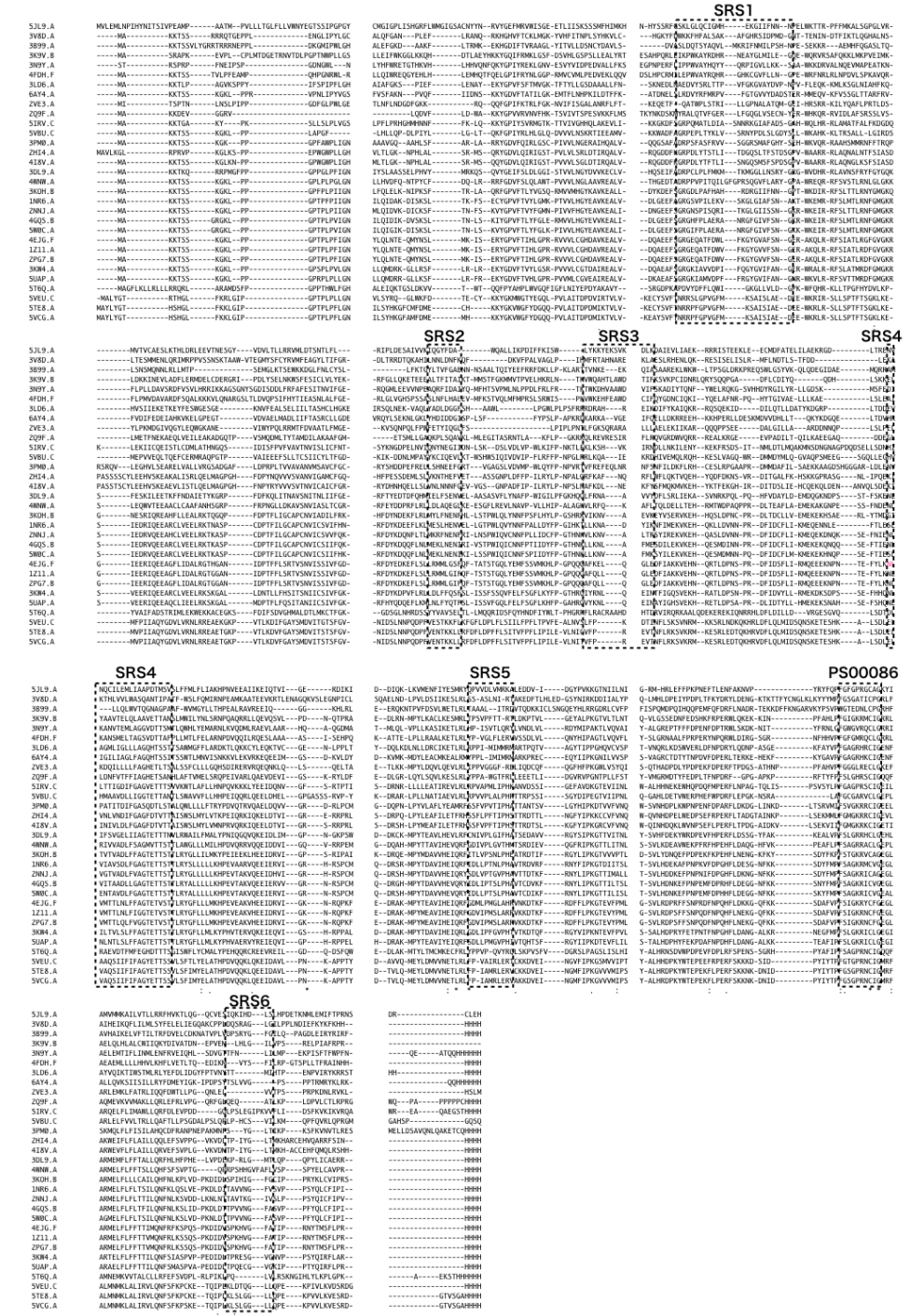

図2-7 基質認識部位とヘム結合領域の推定

破線で囲まれた部分は推定された基質認識部位および推定されたヘム結合領域．基質認

識部位（SRS）はCYP2ファミリーについて行われた研究[77]をもとに, Ensembl protein のアミノ酸配列とPDBから得たアミノ酸配列（表2-5）を用いた多重配列アラインメントにより推定された. ヘム結合領域は, タンパク質ドメイン・機能部位データベースPROSITEにP450のヘム結合領域の配列として登録されているPS00086の配列

([FW]-[SGNH]-x-[GD]-{F}-[RKHPT]-{P}-C-[LIVMFAP]-[GAD])

から推定された.アラインメントの左側にあるIDは, 各P450タンパク質のアミノ酸配列の相同性検索によって選択されたP450タンパク質の立体構造データのPDB IDを示す.

表2-7　P450の機能領域に存在するミスセンスバリアント

領域	すべての出現数 (%)	地域差あり (%)	地域差なし(%)
SRS1	579(3.63)	69(3.33)	382(3.74)
SRS2	251(1.57)	43(2.08)	150(1.47)
SRS3	352(2.21)	45(2.17)	230(2.25)
SRS4	659(4.13)	63(3.04)	421(4.12)
SRS5	396(2.48)	56(2.70)	275(2.69)
SRS6	206(1.29)	21(1.01)	130(1.27)
Heme	324(2.03)	31(1.50)	222(2.17)
その他	13183(82.65)	1743(84.16)	8416(82.30)
合計	15950(100)	2071(100)	10226(100)

基質認識部位（SRS）[77] SRS1~SRS6およびヘム結合領域（タンパク質ドメイン・機能部位データベースPROSITEにP450のヘム結合領域の配列として登録されているPS00086）に存在するミスセンスバリアントの出現数と割合を示した. () 内の数字はそれぞれの合計出現数に対する割合を示す. すべての出現数：P450のすべてのミスセンスバリアントについて各領域の出現数, 地域差あり：地域差ありミスセンスバリアントについて各領域の出現数, 地域差なし：地域差なしミスセンスバリアントについて各領域の出現数.

2-4. 考察.

gnomAD exomeに存在するP450遺伝子のミスセンスバリアントを東アジア（EAS），アフリカ（AFR），フィンランドを除くヨーロッパ（NFE）における出現頻度で地域差ありミスセンスバリアントと地域差なしミスセンスバリアントに分け，アミノ酸置換型やバリアントの位置を比較した．アミノ酸残基の電荷の変化に注目すると，酸性残基から中性残基への置換および塩基性残基から中性残基への置換では，地域差ありミスセンスバリアントでの出現率と地域差なしミスセンスバリアントでの出現率に統計的な有意差があった（図2-5右）．アミノ酸置換型を地域差ありミスセンスバリアントと地域差なしミスセンスバリアントで比較すると，リシンがメチオニンに置換する型，アルギニンがプロリンに置換する型は，出現頻度の地域差が存在する割合が高い傾向があった（図2-5，表2-4）．アルギニン残基はP450の基質認識[80–82]とタンパク質間相互作用[83]に重要なアミノ酸残基と考えられているため，それぞれの地域では，P450の機能に重要なアミノ酸残基の置換による影響の大小が生じている可能性がある．また，基質認識部位およびヘム結合領域での地域差ありミスセンスバリアントと地域差なしミスセンスバリアントの出現率には一部を除いて統計的な有意差

がなく，基質認識部位およびヘム結合領域のミスセンスバリアントは地域の違いにかかわらず一定の頻度でおこることが示唆された．

P450のうち，薬物を代謝するタンパク質は，薬物または異物で誘導を受けることが知られている[33]. P450にあるバリアントは，薬物または異物の暴露を受けるまではその影響が現れにくいことが考えられ，そのため，バリアントのほとんどは自然選択に対し有利でも不利でもない中立的な変化であると考えるのが妥当である．このようなバリアントの頻度に地域差がおこる機序として，遺伝的浮動とボトルネック効果によるものが考えられる．遺伝的浮動とは，有限数の生物集団のなかで，遺伝子頻度が偶然に変化していくことである[84]．ボトルネック効果とは，なんらかの原因で生物集団の個体数が激減し，残った個体がさらに増えることで遺伝子頻度が変化することである[85]. Y染色体に注目した研究により，現生人類は何度かボトルネック効果を繰返していることが知られている[86]. P450ミスセンスバリアントの出現頻度の地域差は，偶然による遺伝的浮動と，民族移動による集団の分割と増加により生じているのではないかという仮説をたてることが可能である．しかし，本研究で扱ったP450のミスセンスバリアントの出現頻度は，時系列を考慮した変化ではなく，配列を比較し

た時点でのヒトリファレンスゲノム配列との違いとして記述されたものであり,

将来, P450のミスセンスバリアントの出現頻度がどのように変化していくかを

観察していく必要がある.

第三章
影響未知のヒトシトクロムP450
ミスセンスバリアントの影響予測

概要

前章では，ヒトシトクロムP450（P450）遺伝子の多様性に地域による偏りが生じた過程を明らかにすることをめざし，出現頻度の地域差があるミスセンスバリアントの特徴を解析した．しかし，地域差の生じた過程を明確に説明できるような特徴はみつけられなかった．この章では，P450遺伝子の多様性として，個人レベルの差である「バリアントのタンパク質への影響」に注目した．P450遺伝子のミスセンスバリアントをP450タンパク質の立体構造にマッピングして得られた情報から，機械学習アルゴリズムの一種であるランダムフォレスト[87]によるタンパク質への影響予測モデルを構築した．この予測モデルは既存のミスセンスバリアント影響予測ツールであるSIFT[88, 89]およびPolyPhen-2[90]と比較した．その結果，本研究で構築した予測モデルは，予測成績の評価指標である受信者操作特性（ROC）曲線とROC曲線下面積（AUC）においてSIFTおよびPolyPhen-2を上回った．したがって，P450タンパク質の立体構造にマッピングして得られた情報がタンパク質への影響予測に役立つことが示された．本研究で構築した予測モデルを用いて，P450遺伝子のミスセンスバリアントのうち，影響未知の9641件について，タンパク質への影響予測を試みた．その結果，影響未知

ミスセンスバリアントの約3分の1はタンパク質への影響ありバリアントである可能性を示すことができた.

3-1. 背景

P450遺伝子のバリアントにより, 薬物投与において副作用や予期せぬ効果が見られることが知られている[24]. たとえば, 血栓症の治療に抗凝固剤としてもちいられる*S*-ワルファリンは, CYP2C9の代謝により薬効を生じる. *S*-ワルファリンの投与にあたっては, 過度の血液凝固あるいは過度の出血傾向を防ぐため, 慎重な用量モニタリングが必要とされる. *CYP2C9*遺伝子には酵素活性が低下するハプロタイプ（複数のバリアントの組み合わせ）があり, 酵素活性低下型のヒトに *S*-ワルファリンを投与する場合は通常よりも低用量にしなくてはならない[24]. 個人の遺伝的特徴に基づいた個別化医療を進めるにあたり, 主要な薬物の毒性に関連するP450の多様性の影響を見積もることが求められている[24].

計算科学的なバリアントの影響予測は, アミノ酸配列の保存性を利用したSIFT [88, 89]やアミノ酸配列と立体構造の特徴を利用したPolyPhen-2[90]がよく利用されている. しかし, ACMGガイドラインでは, *in silico*での予測は病原性判

定の根拠としてはもっとも弱いものとされている[9]. *CYP2U1*のミスセンスバリアントの影響を4種の予測ツールSIFT, PolyPhen-2, MutationTaster[91], MutaCYP[92]により予測した結果と，培養細胞で実際に発現させたタンパク質の活性や分光学的性質を比較した研究では, *in silico*の予測と*in vitro*の分析結果は必ずしも一致せず, *in silico*の予測ツールだけでは，ミスセンスバリアントがタンパク質に与える影響を正確に見積もるにはまだ十分ではないことが示された[93]. すべてのタンパク質に対し広く適用できるミスセンスバリアント影響予測ツールだけではなく，個々のタンパク質の性質や機能の特徴に合わせたミスセンスバリアント影響予測ができれば，予測結果の向上が期待できる. CoDPは，遺伝性腫瘍疾患であるLynch症候群の原因遺伝子のひとつである*MSH6*に特化し，ミスセンスバリアントがタンパク質にどのような影響を与えるかを予測するツールである. アミノ酸残基がタンパク質の表面に露出している程度を表す溶媒接触度とアミノ酸の重原子の変化および既存の予測ソフトウェアによる予測結果を統合した手法を用いている. CoDPは, SIFTおよびPolyPhen-2より高い精度での予測を達成した[94]. しかし, P450タンパク質に特化したミスセンスバリアントの影響予測ツールはまだ少ない. DL-ADRは，深層学習を利用したP450

ハプロタイプの影響予測ツールである[95]. MutaCYPは，アミノ酸配列の保存性とタンパク質の溶媒接触度からP450バリアントの影響予測をするツールであるが，タンパク質間相互作用の情報は利用していない[92]．したがって，タンパク質表面にあるバリアントに対しては正しい予測が得られない可能性がある．そこで，本研究では, P450タンパク質と, P450タンパク質に電子を伝達する電子供与体の相互作用（第二章，図2-2）に注目し，バリアントとタンパク質分子の重心との距離，バリアントとアミノ酸残基間の距離という立体構造から得られる特徴を用いたP450遺伝子ミスセンスバリアントのタンパク質への影響予測モデルを構築した．構築した予測モデルの精度を検証し，既存のミスセンスバリアントの影響予測ツールと比較した．そして，構築した予測モデルにより, gnomAD exomeに存在する影響不明のP450遺伝子のミスセンスバリアントのタンパク質への影響の予測を試みた．

3-2. 手法

3-2-1. P450遺伝子のミスセンスバリアント

図3-1に解析全体の流れを示した. P450遺伝子のミスセンスバリアントは，gnomAD exome[15]より得た．ヒトP450のEnsembl protein IDに対応するアミノ酸

配列はEnsembl GRCh37版[72]より得た. 1つの遺伝子名に対して複数のEnsembl protein IDが存在する場合は, 米国国立生物工学情報センター（National Center for Biotechnology Information, NCBI）のリファレンス配列データベースRefSeqに登録されている配列のEnsembl protein IDとした. タンパク質立体構造データバンクProtein Data Bank（PDB）[75]に登録されているタンパク質のアミノ酸配列を問い合わせ先データベースとしたBLAST[76]による検索結果から, P450の各アミノ酸配列に対し最も類似度が高いアミノ酸配列をもつP450タンパク質の立体構造データのPDB IDを得た. さらにP450のアミノ酸配列とP450タンパク質立体構造のアミノ酸配列をMUSCLE v3.8.31[78]を用いて多重配列アラインメントを行い, この多重配列アラインメントにより, P450遺伝子のミスセンスバリアントをP450タンパク質の立体構造にマッピングした. さらに, バリアントのゲノム配列上の位置と塩基置換, アミノ酸置換を手掛かりに, ClinVarデータベースから得た臨床的な重要性（Clinical significance）を各ミスセンスバリアントに結びつけた.

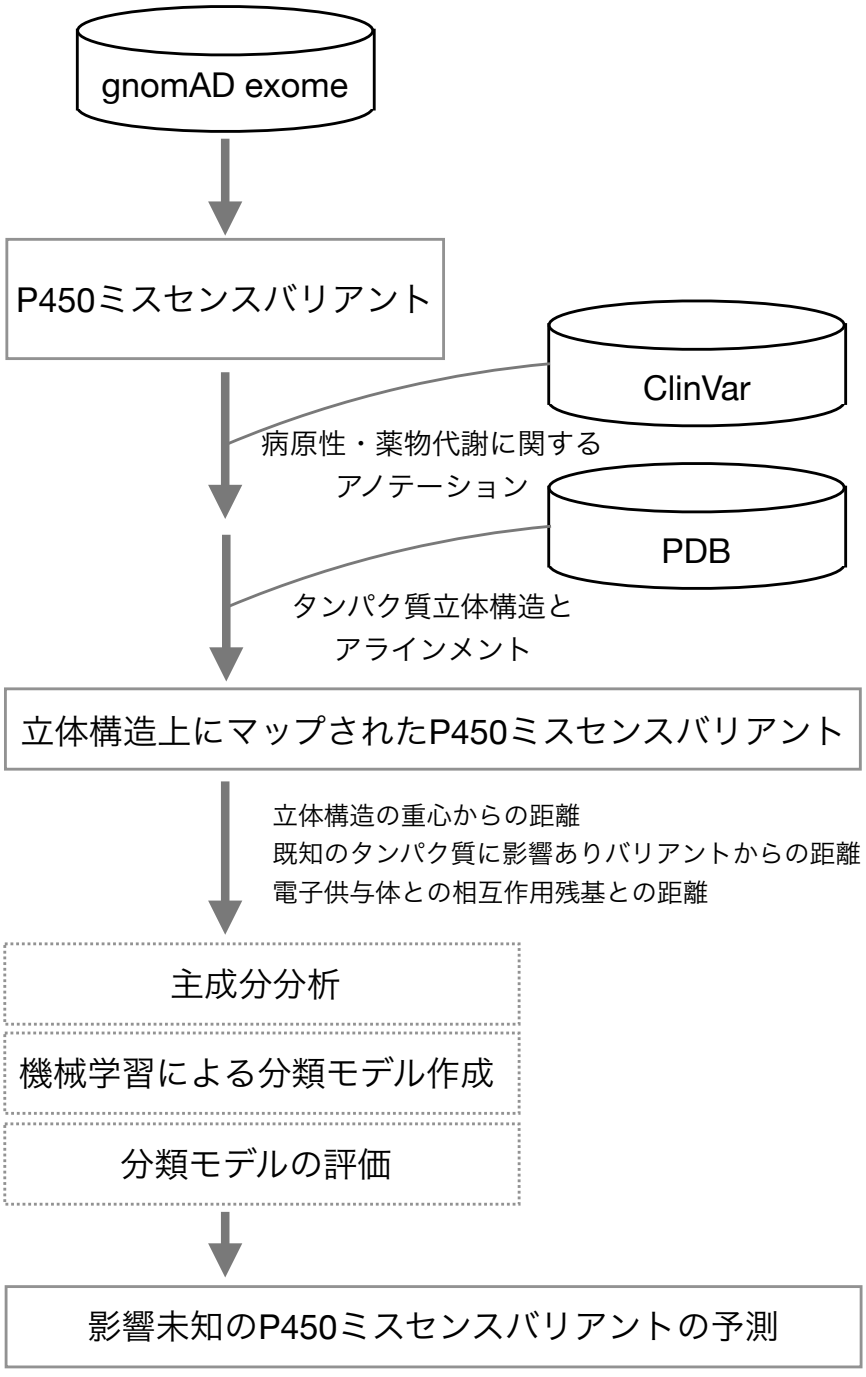
gnomAD exome
P450ミスセンスバリアント
ClinVar
病原性・薬物代謝に関する
アノテーション
PDB
タンパク質立体構造と
アラインメント
立体構造上にマップされたP450ミスセンスバリアント
立体構造の重心からの距離
既知のタンパク質に影響ありバリアントからの距離
電子供与体との相互作用残基との距離
主成分分析
機械学習による分類モデル作成
分類モデルの評価
影響未知のP450ミスセンスバリアントの予測

図3-1 解析の流れ

3-2-2. ミスセンスバリアントによるアミノ酸置換部位の特徴量

P450遺伝子のミスセンスバリアントをP450タンパク質の立体構造にマッピングし，ミスセンスバリアントにより置換されるアミノ酸残基のCα原子の座標から，解析に用いる数値（特徴量）を計算した．各数値は立体構造上の座標間の距離として以下のように計算した．

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 (z_1 - z_2)^2}$$

d：求める距離を，(x1, y1, z1)：ミスセンスバリアントにより置換されるアミノ酸残基のCα原子の座標を，(x2, y2, z2) :距離を求めたい座標を示す．

表3-1と図3-2に各特徴量の概略を示した．特徴量は図3-2に示したA~Cの3つの観点から考えられた．Aは当該バリアントがタンパク質の表面近くにあるか，中心近くにあるかを示す．Bは当該バリアントとタンパク質への影響ありバリアント（既知）に対する3次元的な配置を示す．Cは当該バリアントP450の機能に重要と考えられるタンパク質間相互作用に関与する残基との3次元的な配置を示す．あるミスセンスバリアントについて，以下のように特徴量の値を計算した．A：ミスセンスバリアントと，マッピングされた立体構造の幾何重心との距離を求めた．B：ミスセンスバリアントと，同じ立体構造上にマッピングされたP450の

タンパク質への影響ありバリアントのアミノ酸残基との距離を計算し，そのうち最短のものをタンパク質への影響ありバリアントとの距離とした. C：Putidaredoxinと相互作用する*Pseudomonas putida*由来のP450camのX線結晶構造解析データ（PDB ID : 3W9C）[37] を利用して，電子供与体との相互作用に関与するアミノ酸残基との距離を以下のように計算した．各ミスセンスバリアントをマッピングしたP450タンパク質の立体構造データをP450camの立体構造データに3D構造で重ね合わせ, P450camの14個のアミノ酸残基[37]のCα原子と，ミスセンスバリアントがマップされたアミノ酸残基のCα原子との距離を計算した.

表3-1　バリアントの特徴量

特徴量の名称	定義
mutdist	立体構造の重心と当該バリアントとの距離
mindist	タンパク質に影響ありバリアント（既知）と当該バリアントとの距離のうち最短のもの
a76	P450camの76番目のGluと当該バリアントとの距離
a109	P450camの109番目のArgと当該バリアントとの距離
a112	P450camの112番目のArgと当該バリアントとの距離
a113	P450camの113番目のAlaと当該バリアントとの距離
a116	P450camの116番目のAsnと当該バリアントとの距離
a121	P450camの121番目のMetと当該バリアントとの距離
a122	P450camの122番目のProと当該バリアントとの距離
a125	P450camの125番目のAspと当該バリアントとの距離
a352	P450camの352番目のHisと当該バリアントとの距離
a353	P450camの353番目のGlyと当該バリアントとの距離
a358	P450camの358番目のLeuと当該バリアントとの距離
a356	P450camの356番目のLeuと当該バリアントとの距離
a360	P450camの360番目のGlnと当該バリアントとの距離
a361	P450camの361番目のHisと当該バリアントとの距離

特徴量を識別するために用いた名称と, 特徴量の定義を示した.

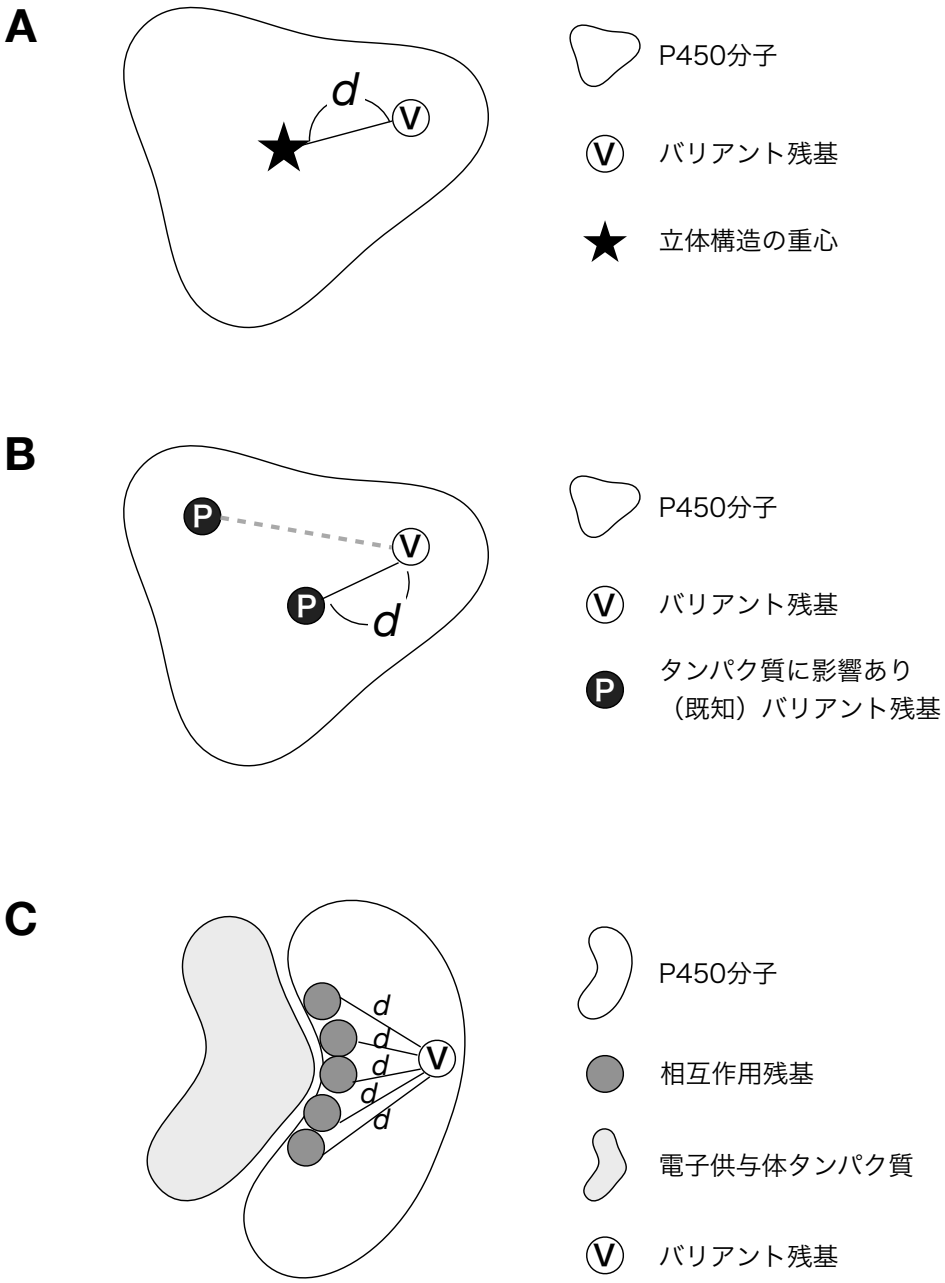

図3-2 特徴量計算の概略

特徴量計算方法の概略を示した. A. 当該バリアントの, P450立体構造の重心からの距離. B. 当該バリアントと既知のタンパク質への影響ありバリアントとの距離. 既知のタンパク質への影響ありバリアントが複数ある場合は, 最短距離とした. C. 当該バリアントと電子供与体との相互作用に関与するアミノ酸残基との距離. d : 求める距離.

3-2-3. 主成分解析および機械学習によるタンパク質への影響予測モデルの構築

P450遺伝子のミスセンスバリアントは, ClinVarデータベースでの臨床的な重要性（Clinical significance）をもとに以下の2群に分けた. 1) タンパク質への影響ありとされているもの（Clinical significanceがPathogenic, Pathogenic/likely pathogenic, Likely pathogenicまたはdrug responseであるもの, 以下patho-drug群とする）, 2) 無害とされているもの（Clinical significanceがBenign, Benign/likely benignまたはlikely benignであるもの, 以下, benign群とする）. patho-drug群は, 遺伝性疾患関連の病原性バリアント（ClinVarのclinical significance でPathogenic , Pathogenic/Likely pathogenic , Likely pathogenicとされているバリアント）と薬理遺伝学的解析による薬物反応に関連するバリアント（ClinVarのclinical significance でdrug responseとされているバリアント）を「タンパク質へなんらかの影響をあたえるバリアント」として捉えることとし, 影響の大小は考慮に入れなかった. patho-drug群, benign群のバリアントのうち, 表3-1の16種の特徴量がすべて揃っているバリアントについて特徴量の主成分解析を行った.

続いて, タンパク質への影響の予測を目的とした, 前記16種の特徴量を用い

た機械学習による予測モデルを構築した．特徴量のデータは，はじめに乱数を発生させ，その乱数に従ってデータをランダムに並べ替えた後，75%の訓練データと25%のテストデータに分割した．乱数発生の開始点（シード）を固定し，データの分割に再現性が得られるようにした．予測モデルの機械学習アルゴリズムは，二値分類に適したアルゴリズムとされている[96]，ロジスティック回帰[97]，サポートベクターマシン[98]およびランダムフォレスト[87]を試し，テストデータによる評価がもっとも高いものを選んだ．モデル構築に必要な係数は，訓練データを分割し，一部で係数の設定テストを行い，残りで検証を行う手法（交差検証）を用いて決定した．各予測モデルの評価は，モデル構築時の訓練データとテストデータによる分類結果をまとめた表（混同行列，表3-2）と評価指標である正確度，適合率，再現率およびf1-値の比較により行なった．正確度（Accuracy)とは正確な分類を行なった数をすべてのサンプルの個数で割った数，適合率（Precision）とは，陽性と分類したものが実際に陽性であった割合，再現率（Recall）とは，実際に陽性であるものが陽性と分類されたものの割合，f1-値とは適合率と再現率の調和平均である[96]．本研究においては陽性＝タンパク質への影響あり，陰性＝無害とした．

表3-2　混同行列

	陽性と予測	陰性と予測
実際に陽性	真陽性（true positive,TP）	偽陰性（false negative, FN）
実際に陰性	偽陽性（false positive, FP）	真陰性（true negative,TN）

陽性・陰性の2つのクラスに分類するときの評価結果を表現する表である.

TP : 真陽性, FP : 偽陽性, TN : 真陰性, FN : 偽陰性とすると,

$$\text{正確度} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{適合率} = \frac{TP}{TP + FP}$$

$$\text{再現率} = \frac{TP}{TP + FN}$$

$$f_1 - \text{値} = 2 \times \frac{\textit{適合率} \times \textit{再現率}}{\textit{適合率} + \textit{再現率}}$$

となる[96]. 各予測モデルの評価は, テストデータによる受信者操作特性（ROC）グラフ下面積（AUC） の比較によっても行った. 予測結果はタンパク質への影響がある確率（0~1の値）として出力されるため, 分類のための閾値が設定される. ROCグラフは横軸に実際は陰性であるものを陽性と分類（benign群をタンパク質への影響ありと分類）した割合（偽陽性率）, 縦軸に実際に陽性であるもの

を陽性と分類（patho-drug群をタンパク質への影響ありと分類）した割合（真陽性率）とし, 分類の閾値を変化させたときの偽陽性率と真陽性率をプロットしたグラフである.

3-2-4. 既存のミスセンスバリアント影響予測ツールとの比較

既存のミスセンスバリアント影響予測ツールSIFT, PolyPhen-2（HumDiv, HumVar）でテストデータの影響予測を行い, ROCグラフとROCグラフ下面積（AUC）の比較を行った. SIFTの結果スコア（SIFT score）は0に近づくほど有害バリアントである可能性が高く[88, 89], PolyPhen-2では結果スコア（probability）が1に近づくほど有害バリアントである可能性が高いため[90], ROCグラフを描くスコアとして, SIFTは1からSIFT scoreを引いた差分, PolyPhen-2ではHumDivのprobabilityおよびHumVarのprobabilityを用いた.

3-3. 結果

3-3-1. 主成分解析

表3-5にP450タンパク質のEnsembl protein IDとP450タンパク質立体構造データ

の対応を示した．表3-1に示した16種の特徴量データ（立体構造の幾何重心からの距離，既知のタンパク質への影響ありバリアントアミノ酸残基からの最短距離, P450camと電子供与体との相互作用に関わるアミノ酸残基からの距離）がすべて揃っているミスセンスバリアントは148件あり，そのうち, patho-drug群は102件, benign群は46件あった．図3-3に, patho-drug群の位置をマゼンタ色の空間充填モデルとして示した．図3-3に示すように, patho-drug群のバリアントによるアミノ酸置換位置はP450の立体構造の全体に散在していた. 16種の特徴量データを用いて主成分解析を行なった．その結果，図3-4に示したように主成分3および主成分4において, patho-drug群とbenign群の分布には偏りがみられた．主成分1と主成分2では明瞭な偏りが認められなかった．各主成分の寄与率は主成分1が0.676，主成分2が0.185，主成分3が0.083，主成分4が0.035，主成分5が0.020であった（図3-5）．図3-6にそれぞれの主成分に対する特徴量の相関の強さを示す値である主成分負荷量を示した．主成分1の主成分負荷量ではすべての特徴量で高く，主成分2の主成分負荷量では電子供与体との相互作用アミノ酸残基（P450camのアミノ酸配列の76番目, 121番目, 122番目, 125番目, 352番目, 353番目, 358番目に相当するアミノ酸残基との距離）が高かった．主成分3では立体構造の幾何重心

からの距離, 既知のタンパク質への影響ありバリアントアミノ酸残基からの最短距離, P450camの109番目, 112番目, 113番目との距離の主成分負荷量が高かった. 主成分4では既知のタンパク質への影響ありバリアントアミノ酸残基からの距離の主成分負荷量が高かった.

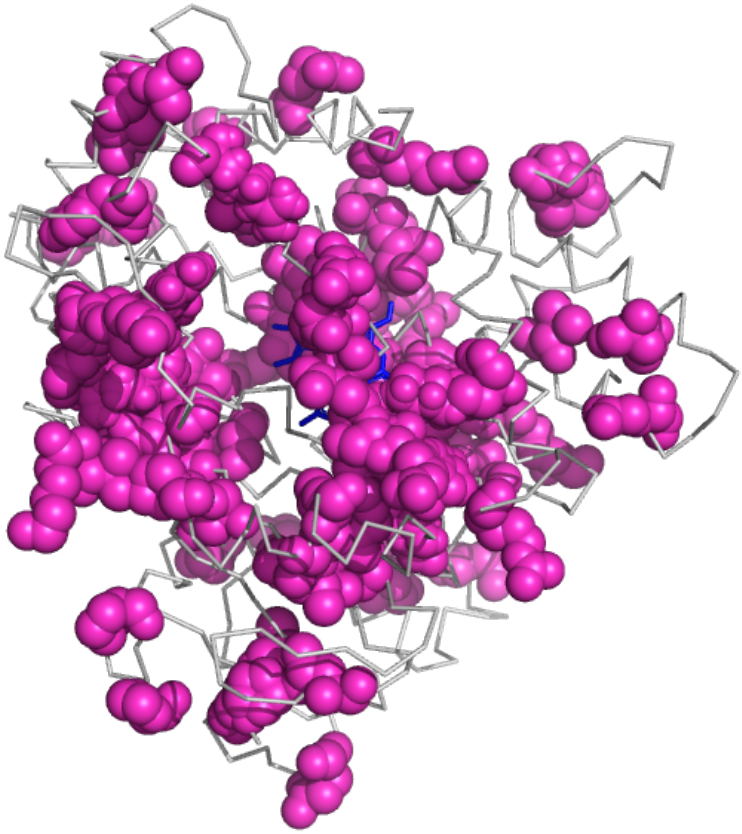

図3-3 ヒトシトクロムP450のタンパク質への影響ありバリアントの位置

ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）102件を立体構造上にマッピングし，立体構造をすべて重ね合わせた状態でマゼンタ色の空間充填モデルとして示した．立体構造の主鎖は灰色の折れ線として表し，ヒトCYP2D6（PDB ID : 4WNW, A鎖）のみを表示した．ほかのP450タンパク質のEnsembl protein IDとP450タンパク質立体構造データの対応は表3-5に示した．青色の折れ線：ヘム分子.

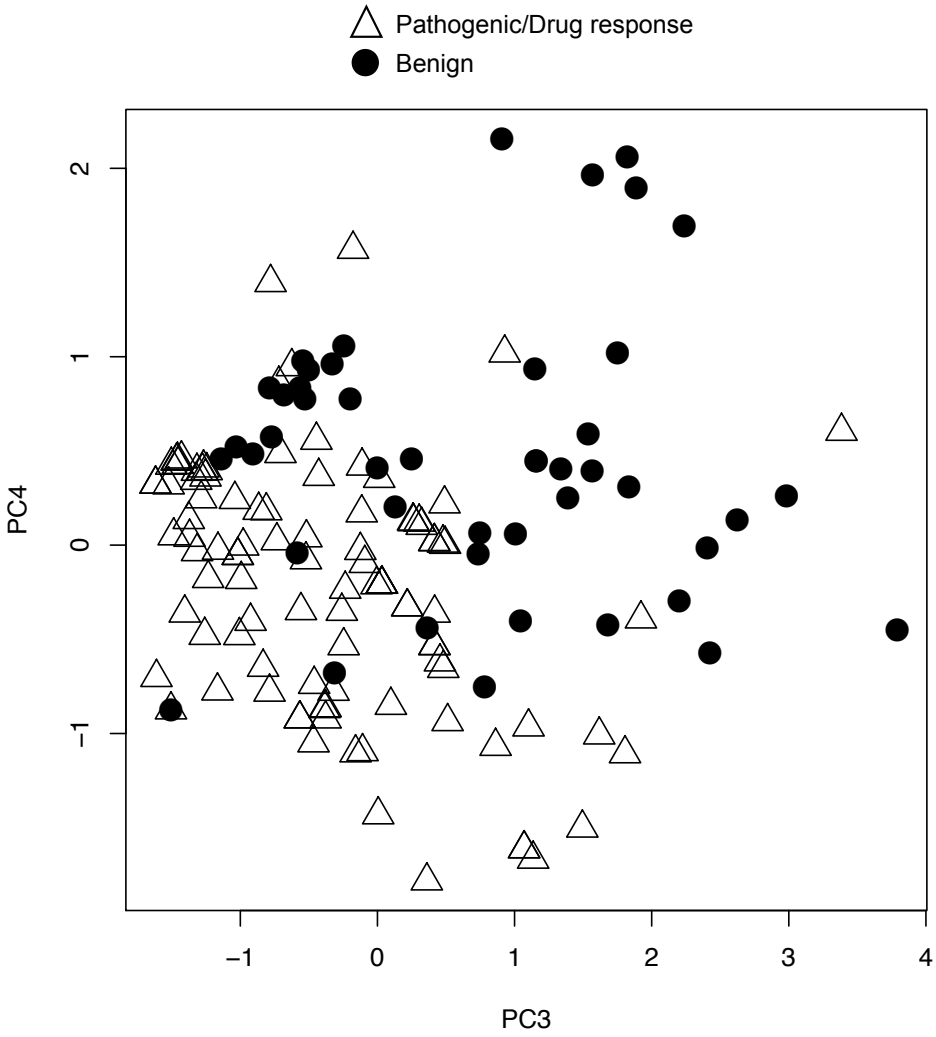

図3-4　ミスセンスバリアントアミノ酸残基の座標情報による主成分解析

ミスセンスバリアント148件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）102件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）46件）の特徴量データの主成分解析のうち第3主成分と第4主成分の散布図. 黒色丸はClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント, 白色三角形はClinVarデータベースのclinical significanceで病原性または薬物反応性（Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアントを示す.

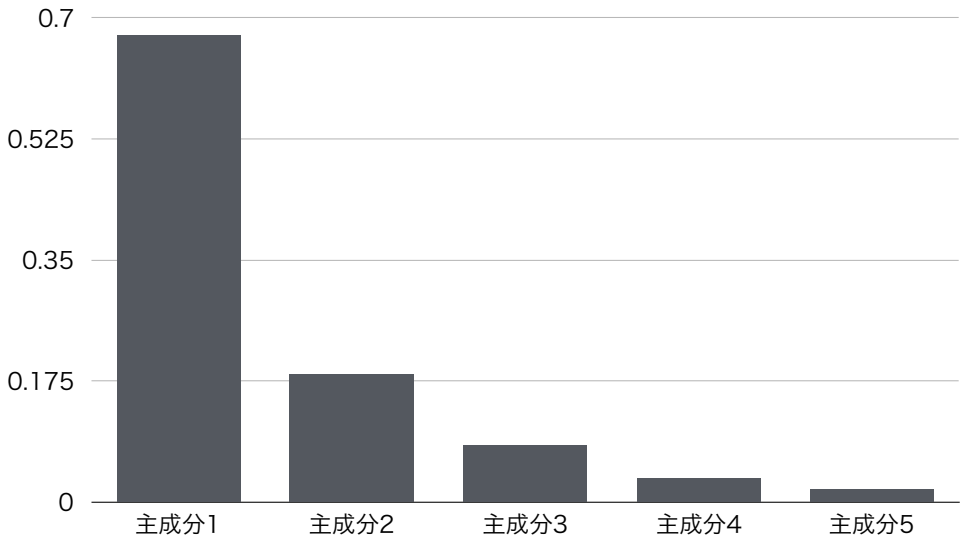

図3-5　主成分の寄与率

ミスセンスバリアント148件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）102件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）46件）の特徴量データの主成分解析（図3-4）の寄与率を主成分1から主成分5まで示した. 縦軸は寄与率を示す.

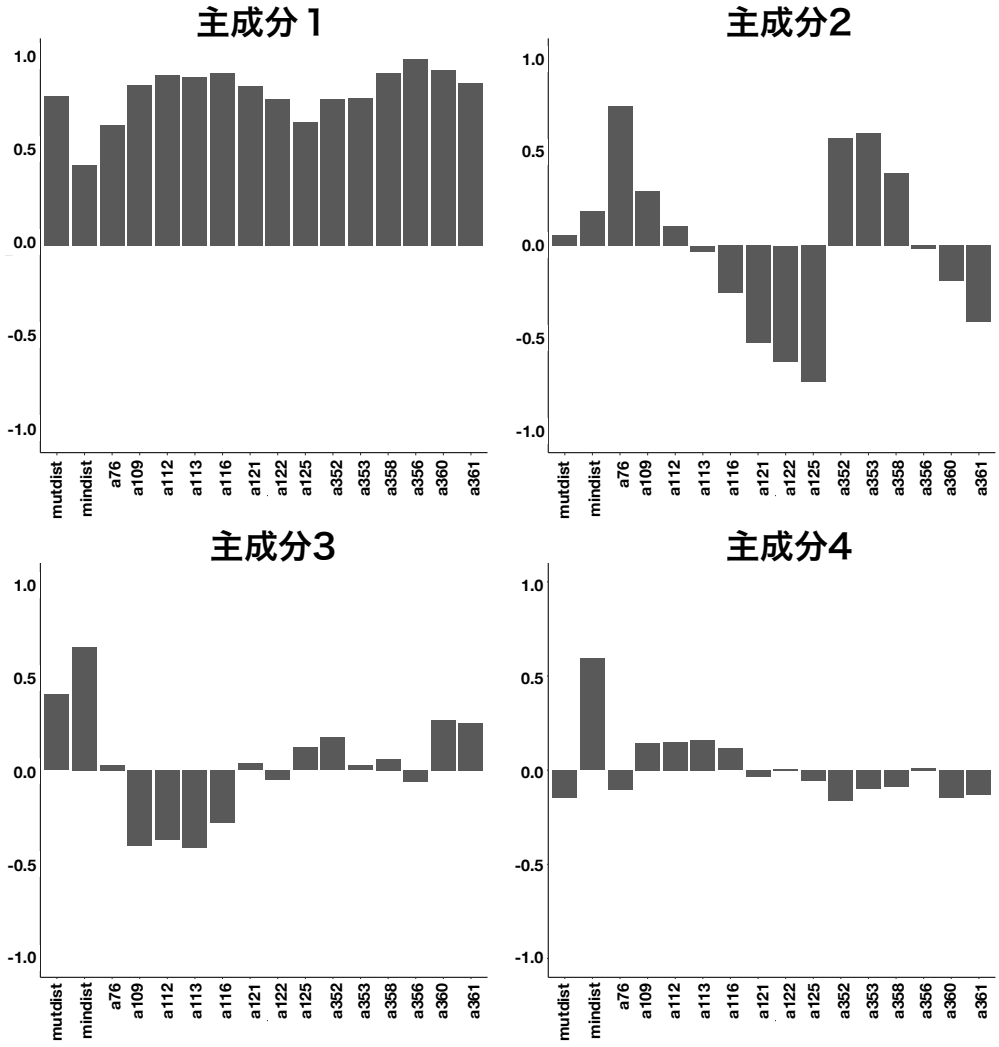

図3-6　主成分解析の主成分負荷量

ミスセンスバリアント148件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）102件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）46件）の特徴量データの主成分解析（図3-4）の主成分負荷量を主成分1から主成分4まで示した. 主成分負荷量とは, それぞれの主成分に対する特徴量の相関の強さを示す値である. PC1 : 主成分1の主成分負荷量, PC2 : 主成分2の主成分負荷量, PC3 : 主成分3の主成分負荷量, PC4 : 主成分4の主成分負荷量. 縦軸は主成分負荷量の大き

さであり, 絶対値が大きいほどその主成分との相関が強い. 横軸は各特徴量を示す.
mutdist：バリアントアミノ酸残基C_α原子と対応するタンパク質立体構造の幾何重心との距離, mindist：バリアントアミノ酸残基Cα原子と対応するタンパク質立体構造を共有する遺伝子上の既知のタンパク質への影響ありバリアントアミノ酸残基Cα原子との距離の最小値, a76 ~ a361：p450cam（PDB id：3W9C）のPutidaredoxinとの相互作用[37]に関わるアミノ酸残基Cα原子との距離. うち, a76, a112, a125は静電相互作用, a109, a113~a122, a352~a361はファンデルワールス力による相互作用[37].

3-3-2. タンパク質への影響予測モデルの構築と評価

表3-1の特徴量を用いて機械学習によるタンパク質への影響予測モデルを構築した． 表3-3に, ロジスティック回帰, サポートベクターマシン,ランダムフォレストによる予測モデルの評価を示した. 各予測モデルの混同行列は表3-4に示した. ロジスティック回帰での特徴量の寄与度は, 重心からの距離, 既知のタンパク質への影響ありバリアントとの距離, 重ね合わせした立体構造上のP450camの76番目のグルタミン酸, 122番目のプロリン, 125番目のアスパラギン酸との距離について高かった（図3-7）. また, ランダムフォレストでは, 既知のタンパク質への影響ありバリアントとの距離, 重ね合わせした立体構造上のP450camの352番目のヒスチジンとの距離について高かった（図3-8）.

表3-3　機械学習による予測モデルの評価

アルゴリズム	正確度 (訓練)	正確度 (テスト)	適合率	再現率	f1-値	AUC
ロジスティック回帰	0.910	0.946	0.960	0.960	0.960	0.962
サポートベクターマシン（線形）	0.910	0.892	0.890	0.960	0.930	0.944
サポートベクターマシン (RBF)	0.910	0.946	0.960	0.960	0.960	0.951
ランダムフォレスト	0.982	0.919	0.930	0.960	0.940	0.970

テストデータ37件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）26件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）11件）を分類した時の評価指標を示した. 正確度 : 正確な分類を行なった数をすべてのサンプルの個数で割った数, 適合率 : 病原性または薬物反応性と分類したものが実際に病原性または薬物反応性とされているバリアントであった割合, 再現率 : 実際に病原性または薬物反応性とされているバリアントが病原性または薬物反応性と分類されたものの割合, f1-値：適合率と再現率の調和平均, AUC：ROC曲線下面積 (area under the ROC curve). 適合率, 再現率, f1-値はタンパク質への影響ありとされているバリアント（patho-drug群）の値を示した.

表3-4　予測モデルの混同行列

アルゴリズム	真陽性	真陰性	偽陽性	偽陰性
ロジスティック回帰	25	10	1	1
サポートベクターマシン（線形）	25	8	3	1
サポートベクターマシン (RBF)	25	10	1	1
ランダムフォレスト	25	9	2	1

テストデータ37件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）26件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）11件）を分類した時の混同行列（confusion matrix）. 真陽性 : タンパク質への影響ありとされているバリアントを正しくタンパク質への影響ありと分類した数, 真陰性 : 無害とされたバリアントを正しく無害と分類した数, 偽陽性 : タンパク質への影響ありとされているバリアントを無害と分類した数, 偽陰性 : 無害とされたバリアントをタンパク質への影響ありと分類した数.

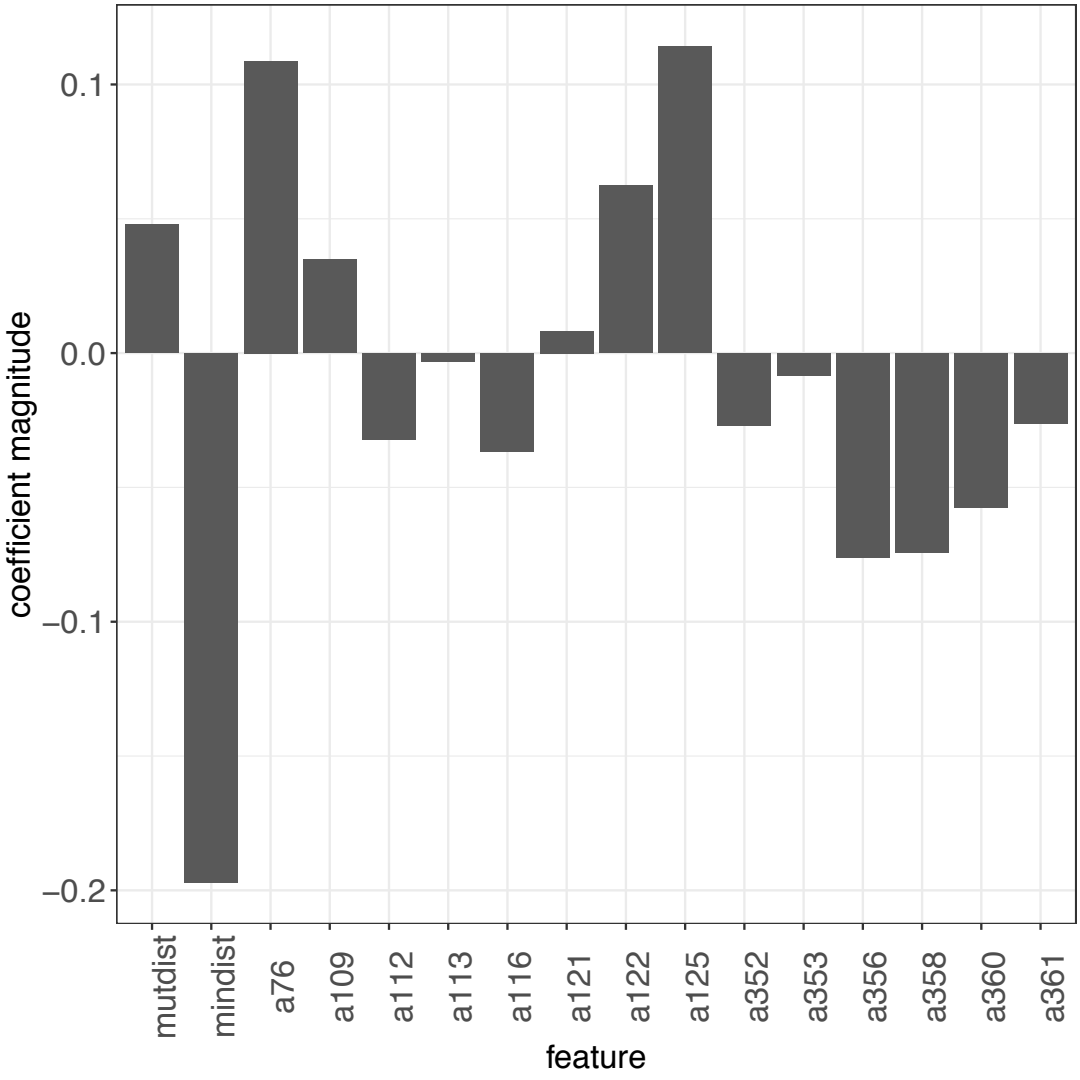

図3-7 ロジスティック回帰に用いた特徴点の寄与

訓練データ111件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）76件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）35件）による分類モデル構築における特徴点の寄与の大きさを示した．縦軸は特徴点の分類に対する寄与度を表し，絶対値が大きいほど分類に対する寄与が大きい．横軸は各特徴量を示した. mutdist：バリアントアミノ酸残基C$_\alpha$原子と対応するタンパク質立体構造の幾何重心との距離, mindist：バリアントアミノ酸残基Cα原子と対応するタンパク質立体構造を共有する遺伝子上の既知のタンパク質への影響ありバリアントアミノ酸残基Cα原

子との距離の最小値, a76 ~ a361：p450cam（PDB id：3W9C）のPutidaredoxinとの相互作用[37]に関わるアミノ酸残基（計14残基）のCα原子との距離. うち, a76, a112, a125は静電相互作用, a109, a113~a122, a352~a361はファンデルワールス力による相互作用[37].

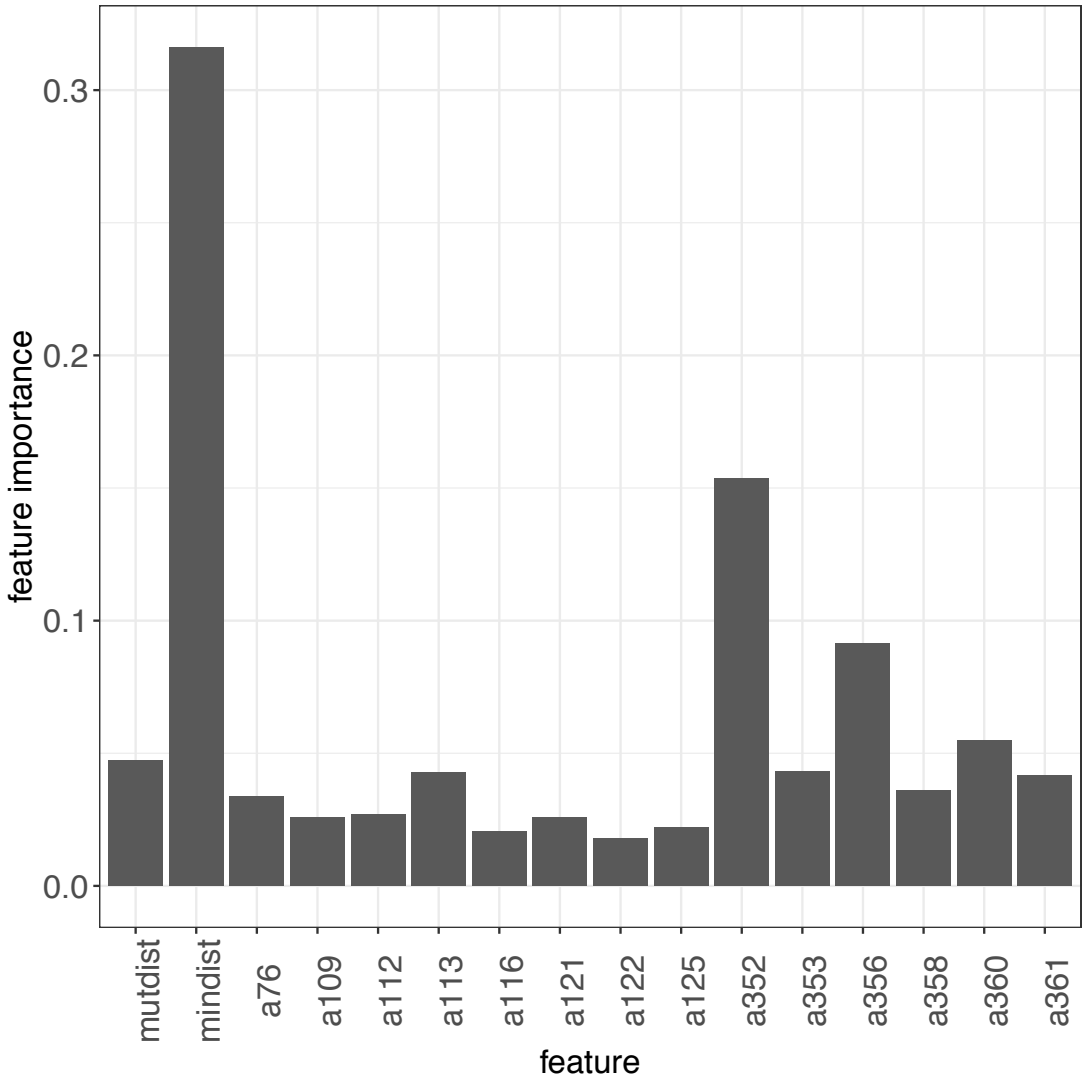

図3-8 ランダムフォレストに用いた特徴点の寄与

訓練データ111件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）76件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）35件）による分類モデル構築における特徴点の寄与の大きさを示した. 縦軸は特徴点の分類に対する寄与度を表し, 絶対値が大きいほど分類に対する寄与が大きい. 縦軸は各特徴量を示した. mutdist：バリアントアミノ酸残基C_α原子と対応するタンパク質立体構造の幾何重心との距離, mindist：バリアントアミノ酸残基Cα原子と対応するタンパク質立体構造を共有する遺伝子上の既知のタンパク質への影響ありバリアントアミノ酸残基C_α原

子との距離の最小値, a76 ~ a361：p450cam（PDB id：3W9C）のPutidaredoxinとの相互作用[37]に関わるアミノ酸残基（計14残基）のC_{α}原子との距離. うち, a76, a112, a125は静電相互作用, a109, a113~a122, a352~a361はファンデルワールス力による相互作用[37].

3-3-3. 既存のミスセンスバリアント影響予測ツールとの比較

同じテストデータを既存のミスセンスバリアント影響予測ツールであるSIFTおよびPolyPhen-2に適用し予測結果を得た．ランダムフォレストによる予測結果をSIFTおよびPolyPhen-2による予測結果と比較すると, ROCグラフでは，ランダムフォレストがもっとも左上に近づき, SIFTおよびPolyPhen-2による予測より有効性が高いことが示された（図3-9）．表3-5にランダムフォレスト, SIFTおよびPolyPhen-2のROCグラフ下面積（AUC）を示した．今回構築したランダムフォレストによる予測モデルは, SIFTおよびPolyPhen-2よりROC曲線下面積でも上回った.

3-3-4. 影響未知ミスセンスバリアントの影響予測

ランダムフォレストによる予測モデルを用いて影響未知ミスセンスバリアントの影響予測を行なった．その結果，影響未知であったミスセンスバリアント9641件のうち, 3578件がタンパク質への影響ありバリアント候補と予測された.

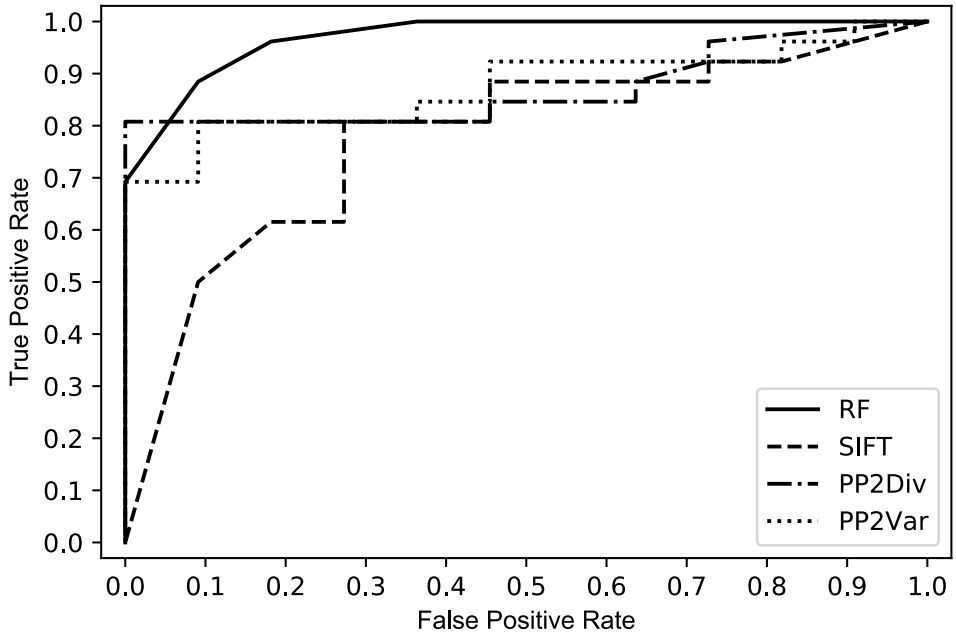

図3-9　既存のバリアント影響予測ツールとの比較

テストデータ37件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）26件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）11件）の分類結果をランダムフォレストによる予測モデルと既存の予測ツールであるSIFT[88, 89], PolyPhen-2[90]で比較した．横軸は無害とされているバリアント（benign群）を「タンパク質に影響あり」と分類した割合（偽陽性率），縦軸はタンパク質に影響ありとされているバリアント（patho-drug群）を「タンパク質に影響あり」と分類した割合（真陽性率），分類の閾値を変化させたときの偽陽性率と真陽性率をプロットした．ランダムフォレストのAUC=0.970. RF:ランダムフォレスト, PP2div : PolyPhen-2 HumDiv, PP2Var : PolyPhen-2 HumVar.

.

表3-5　予測モデルのROCグラフ下面積の比較

予測手法	AUC
ランダムフォレスト	0.970
SIFT	0.769
PolyPhen-2(HimDiv)	0.871
PolyPhen-2(HimVar)	0.874

テストデータ37件（ClinVarデータベースのclinical significanceでタンパク質に何らかの影響がある（病原性または薬物反応性, Pathogenic, Pathogenic/Likely pathogenic, Likely pathogenic, drug response）とされているバリアント（patho-drug群）26件, ClinVarデータベースのclinical significanceで無害（Benign, Benign/Likely benign, Likely benign）とされているバリアント（benign群）11件）の分類結果をランダムフォレストによる予測モデルと既存の予測ツールであるSIFT[88, 89], PolyPhen-2[90]で比較したときのROCグラフ下面積を示した. AUC：ROCグラフ下面積.

3-4. 考察

P450遺伝子のミスセンスバリアントをP450タンパク質の立体構造にマッピングし，アミノ酸置換部位の座標から特徴量データが得られた．これらの特徴量を用いた主成分分析では, ClinVarデータベースに示された臨床的な重要性（Clinical significance）でPathogenic, Pathogenic/Likely pathogenic, Likely pathogenicまたはdrug responseである群の特徴量と, Benign, Benign/Likely benignまたはLikely benignである群の特徴量の分布に偏りがあることがわかった．特徴量データを学習データとして，機械学習によるタンパク質への影響をもつバリアントの予測モデルを構築し，良好な分類結果を得ることができた．ロジスティック回帰において電子供与体との相互作用に関わるグルタミン酸残基との距離はタンパク質への影響ありバリアントの分類に寄与していた. P450camの76番目のグルタミン酸がPutidaredoxinの66番目のアルギニンと弱いイオン性の相互作用である塩橋を作り, P450camとPutidaredoxinの複合体の安定化に働くこと[37]は，この特徴量がP450タンパク質の機能に重要であることの裏付けとなっている．

今回構築したランダムフォレストによる予測モデルは, SIFTおよびPolyPhen-2よりROC曲線下面積で上回った．これはP450のミスセンスバリアント影響予測

では, P450の特徴を学習させた予測モデルが有効であることを示している. 特定のタンパク質に最適化した予測モデルの有効性はすでに知られている[94]が, P450のような類似した配列と構造を持つタンパク質においても, なんらかの共通した特徴でミスセンスバリアントのタンパク質への影響が測れることが示唆された. 今回構築した予測モデルは, 今後あらたにClinVarに追加されるであろうP450のデータや, 将来得られるであろうアミノ酸置換と酵素活性の関連性についての知見により検証することが可能である.

ただし, 今回構築した予測モデルを適用できたのは, 立体構造上にマッピングできるミスセンスバリアントに限られた. 立体構造データに座標が示されていないアミノ酸残基の場合は, 特徴量の計算ができないため予測モデルを適用することができなかった. また, microRNAや調節因子による転写制御[33, 99, 100]などタンパク質の量に関する影響のデータは, 今回の予測モデルの構築には用いていなかった.

本研究により, P450の影響未知ミスセンスバリアントの中には多くのタンパク質への影響をもつバリアントが含まれている可能性も明らかになった. このなかには, なんらかの薬物代謝に影響しうるものが含まれているかもしれない.

P450のタンパク質間相互作用の情報は, 影響未知のミスセンスバリアントがタンパク質に与える影響を見積もるための大きな助けとなった. P450の構造が互いに似ていることで[45, 101], 近縁種ではないタンパク質の相互作用の情報でも利用することが可能であった. 今後, タンパク質の立体構造データを利用したバリアントアノテーションがさらに発展することが期待される.

第四章
ABCトランスポータータンパク質の
立体配座変化と病原性バリアントの関係

概要

第二章および第三章ではヒトシトクロムP450遺伝子のミスセンスバリアントの地域差と個人差について解析した．この章では, ABCトランスポーターに着目し, ABCトランスポーターの立体配座変化と病原性バリアントについて解析した.

細胞膜を介した小分子の輸送は, 細胞活動の維持のため, 非常に重要な生物学的機能である. ABCトランスポーターは細胞膜やオルガネラの膜に存在する膜輸送タンパク質であり, ATPの加水分解に由来するエネルギーを利用して, リガンドを能動輸送する．ヒトでは48種のABCトランスポーター遺伝子がコードされている. ABCトランスポーターの塩基配列にあるミスセンスバリアントはある種の疾患に関わっている場合がある．しかし, 疾患に至る機序はまだ明らかになっていない．近年の膜タンパク質構造決定法の進歩により, ABCトランスポーターの立体構造決定が可能となり, データが蓄積されることで, アミノ酸残基のバリアントを原因とする疾患機序解明の手がかりを得られる可能性があらわれた．この章では, ABCトランスポーターのタンパク質部分であるアポ型とATP結合型を含む各種立体構造を比較し, 膜貫通ドメイン中の分子内回転の中

心点となるアミノ酸残基の周りでおこりうる立体構造の変化を見出した．また，この立体構造の変化と，病原性バリアントの位置を比較することで，バリアントによっておこる疾患についての合理的な説明として，分子内回転の障害がATPの結合と膜表面相互作用を弱める可能性を明らかにした．これらの知見により，ABCトランスポーター遺伝子のバリアントについての新しい解釈とバリアントの影響解析における新しいアプローチがひらかれた．

4-1. 背景

細胞膜を介した小分子の輸送は，細胞機能の維持のための物質の取り込みと排出に関わる非常に重要な機能である[102]．チャネル，溶質輸送体（SLC）トランスポーター，ATP結合カセット（ABC）トランスポーターなどの膜タンパク質によって，それらの輸送が行われている[103]．チャネルタンパク質は拡散によってイオンや水分子を受動輸送し[104]，SLCトランスポーターはATPの加水分解をともなわずに，立体構造の変化を伴った小分子の輸送を行い[105]，ABCトランスポーターはATPの加水分解によるエネルギーを利用して，基質を能動輸送する[106]．ABCトランスポータータンパク質ファミリーはヌクレオチド結合ド

メイン（NBD）と呼ばれるドメイン構造を共通してもつ. NBDにATPが結合し, ATPをADPに加水分解するドメインとして知られている[107]. 配列類似性の高さにより, ABCトランスポーターは7つのサブファミリー（ABCA, ABCB, ABCC, ABCD, ABCE, ABCF, ABCG）に分類される（表4-1）. それらのうち, ABCA, ABCB, ABCC, ABCD, ABCGには類似した膜貫通ドメイン（TMD）が存在するが, ABCEとABCFにはTMDが存在せず, 膜を介した輸送には関与していない[108]. サブファミリーごとのドメイン構成は互いに異なり, ABCトランスポーターファミリーの進化は複雑な道筋を辿ったと考えられる[109].

ヒトABCトランスポーターには48種の遺伝子が存在する[110]. ABCトランスポーターはいくつかの遺伝性疾患の原因としても知られており, ClinVarデータベース[111]には, ABCトランスポーターの塩基配列のバリアントと疾患との関係が記述されている.

ABCトランスポーターは, 抗がん剤をはじめとする多くの薬物を細胞外へと排出してしまうため, 薬物療法の障害となることも知られている[112]. それゆえ, 疾患原因となる分子レベル・原子レベルの機能を理解するため, 機能発現および機能不全の機序と疾患の関係についての研究には高い需要がある[113]. し

かし，すべての疾患とバリアントを結びつける機構の詳細が明らかになっているわけではない．その要因として，膜タンパク質の立体構造を得ることが難しいことが挙げられる[114]．近年，タンパク質構造決定法，特に膜タンパク質の構造決定法が進歩し，ABCトランスポーターの立体構造決定が電子顕微鏡により可能となった[115]．ABCトランスポーターを含む多くのタンパク質立体構造の電子顕微鏡データがProtein Data Bank（PDB）[116]に登録されている．様々な条件（apo型，ATP結合型，阻害剤結合型）のABCトランスポーターの立体構造が決定され，異なる条件下の立体構造の比較が可能となった．この章では，ABCトランスポーターの異なる条件下の立体構造を比較し，ATPまたは阻害剤との結合による構造変化について考察した．構造変化や生物学的機能の鍵となるアミノ酸残基の位置を既知のバリアントの位置と比較し，バリアントと疾患発生との関連性を解くことを試みた．

4-2. 手法

4-2-1. ヒトABCトランスポーター

ヒトABCトランスポーター遺伝子およびその転写産物の塩基配列データ，ア

ミノ酸配列データ，染色体上の位置はEnsemblデータベース[72]とUniprotデータベース[117]より取得した．染色体上の位置に基づき，バリアントの情報はgnomADデータベース[15]から抽出した．バリアントのデータはClinVarの疾患に関するデータと同期している．ABCトランスポーターの立体構造はProtein Data Bank （PDB）[75]から抽出した．すべてのABCトランスポーターの登録情報を取得するため，Ensemblから得たABCトランスポーターのアミノ酸配列のすべてを用いてBLAST[76]によるホモロジー検索を実施した．

4-2-2. ATP結合残基の検出

ATPに結合する残基はKobayashiらの方法[118]で検出した．その方法は以下の通りである．ATP結合ABCトランスポーターの各原子の溶媒接触表面積を自作のプログラムで計算した．ATPをタンパク質の座標から消去し，再び溶媒接触表面積を計算した．2つの異なる溶媒接触表面積の値を持つ原子をATP結合原子とし，その原子を含む残基をATP結合残基とした．

4-2-3. 差分地図の計算

異なる構造をもつ2種のABCトランスポーターを比較するため, 差分地図を作成した. 差分地図とは, 異なる2種のABCトランスポーターの距離地図の差分である. 距離地図とは, タンパク質のアミノ酸配列に番号をつけ, その数字をグラフの縦軸と横軸に並べ, 各数字の交点にそのアミノ酸残基間の距離を記入したグラフである[119]. アミノ酸残基 iとアミノ酸残基 j の間の距離を d_{ij} とする. d_{ij} は2個のCα原子の距離として定義され, この値を三角形の内側の i 側と j 側の交点に示す. ABCトランスポーターの立体配座Bの距離地図から立体配座Aの距離地図を引いた差分を以下のように計算した.

$$\Delta D_{ij}^{B-A} = d_{ij}^{B} - d_{ij}^{A}$$

異なる条件のABCトランスポーターにおいてアミノ酸残基 i と j の対応は, アミノ酸配列のアラインメントに基づいている. 比較には,同じアミノ酸配列をもつ立体配座を用いるべきであるが, 異なる種から得られたタンパク質の配列には多様性があるため, 配列のアラインメントが必要であり, できた地図には空隙（ギャップ）が生じるかもしれない. ある程度の対応が得られれば, ヒト以外の種（マウス, ゼブラフィッシュ）の立体構造を用いた. ΔDの範囲は-10 Åから10 Åとし, 負の値を青色から白色のグラデーションで, 正の値を赤色から白

色のグラデーションで描いた.

4-2-4. 差分プロット

差分地図の次元は以下のようにして削減することができる.

$$\Delta W_i^{B-A} = < \sum_{j=1}^{N} \left| \Delta D_{ij}^{B-A} \right| >$$

ここで, N は アミノ酸残基 i の ΔDの数である. ΔW_i が0であれば, 立体配座の変化が起きてもアミノ酸残基 i のタンパク質内での位置は変わらない. このとき, アミノ酸残基 i はタンパク質分子内ダイナミクスにおける構造変化（おそらくは回転）の軸の候補となる. ΔW_i が極小値をとるとき, アミノ酸残基 i は立体配座変化の中心点候補となる.

表4-1 ヒトABCトランスポーター

ファミリー	遺伝子名	機能	3D構造（PDB ID）
ABCA	ABCA1	スルホニル尿素感受性陰イオン輸送	5xja (アポ型)
	ABCA2	トランスポーター（推定）	
	ABCA3	コレステロール輸送（推定）	
	ABCA4	内向きのレチノイド輸送	
	ABCA5	オートリソソーム	
	ABCA6	マクロファージ脂質恒常性	
	ABCA7	マクロファージ食作用	
	ABCA8	親油性薬物トランスポーター	
	ABCA9	単球分化	
	ABCA10	マクロファージ脂質恒常性	
	ABCA12	脂質恒常性	
	ABCA13	脂質恒常性	
ABCB	ABCB1	多剤耐性細胞における排出ポンプ	5ko2 (アポ型) 5och (ADP結合型) 6c0v (ATP 結合型) 4xwk (阻害剤結合型)
	TAP1	細胞質からERへの抗原輸送体	
	TAP2	細胞質からERへの抗原輸送体	
	ABCB4	リン脂質流出トランスロケーター	
	ABCB5	薬物排出トランスポーター	
	ABCB6	ミトコンドリアへのヘムとポルフィリンの取り込み	
	ABCB7	ミトコンドリアからサイトゾルへのヘム輸送	
	ABCB8	ミトコンドリアの多剤耐性	
	ABCB9	低親和性ペプチドトランスポーター	
	ABCB10	ヘム生合成のためのミトコンドリア輸送	
	ABCB11	胆汁酸塩の分泌	
ABCC	ABCC1	細胞質からの有機陰イオンおよび薬物の排出	5wua (アポ型) 5ywc (ADP結合型) 5w81 (ATP結合型)
	ABCC2	有機陰イオンの肝胆汁中排泄	
	ABCC3	陰イオンの腸排泄	
	ABCC4	有機陰イオンポンプ	
	ABCC5	多重特異性有機陰イオンポンプ	
	ABCC6	細胞内小器官への薬物の輸送	
	CFTR	上皮イオンチャネル	
	ABCC8	ATP感受性カリウムチャンネル	
	ABCC9	ATP感受性カリウムチャンネル	
	ABCC10	親油性アニオン排出	
	ABCC11	cAMPおよびcGMPの細胞からの排出を促進	
	ABCC12	トランスポーター（推定）	
ABCD	ABCD1	トランスポーター（推定）	
	ABCD2	トランスポーター（推定）	
	ABCD3	分岐鎖脂肪酸トランスポーター	
	ABCD4	ビタミンDの細胞内処理	
ABCE	ABCE1	RNase L阻害剤（TMDなし）	
ABCF	ABCF1	mRNA翻訳開始（TMDなし）	
	ABCF2	機能不明（TMDなし）	
	ABCF3	フラビウイルスに対する抗ウイルス効果（TMDなし）	
ABCG	ABCG1	マクロファージ脂質恒常性	5do7 (アポ型) 6hbu (ATP結合型)
	ABCG2	尿酸の排出	
	ABCG4	マクロファージ脂質恒常性	
	ABCG5	ステロール輸送	
	ABCG8	ステロール輸送	

（M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019).Table 1より改変・転載）

4-3. 結果と考察

4-3-1. ヒトABCトランスポーター

表4-1にデータベース検索結果の概要をまとめた. ヒトゲノム中には48種のABCトランスポーター遺伝子が存在し, 7種のサブファミリーに分類された[108]. ABCEとABCFには膜貫通領域がなく, 他のサブファミリーとは大きく異なっていた. そのため, ABCEとABCFは解析の対象としなかった. サブファミリーABCAとABCCは2個のTMDと2個のNBDをもち, 単量体としてはたらくのに対し, ABCDとABCGは, TMDとNBDを1つずつもつサブユニットからなるホモ二量体やヘテロ二量体として機能していた. ABCBには単量体型トランスポーターと二量体型トランスポーターの両方があった. 各ドメインが並ぶ順番は, ABCGではNBD-TMDの順だが, ABCG以外では, TMD-NBD-TMD-NBDの順であった.本項が書かれた時点では, ABCAサブファミリーでは1種類の立体構造, ABCB, ABCC, ABCGサブファミリーでは複数種類の立体構造があったが, ABCDサブファミリーの立体構造はまだ存在していなかった.

4-3-2. ABCトランスポーターの無害なバリアントと病原性バリアント

ヒトABCトランスポーターのミスセンスバリアント30,384件の病原性に関する情報はClinVarデータベースから取得した. これらのバリアントのうち, 201件は無害（Benign）, 407件は病原性（Pathogenic）であった. 表4-2に, 無害なバリアントと病原性バリアントにそれぞれよく見られるバリアントの型を, 多いものから順に示した. 無害なバリアントと病原性バリアントの間の違いとして, 病原性バリアントでアルギニンが別のアミノ酸に置換される場合がよくみられた. 無害なバリアントのうち16%, 病原性バリアントのうち28%が, アルギニンから別のアミノ酸への置換であった. 表4-2のlog2-oddsは, 同じバリアントの型の「無害なバリアントの割合に対する病原性バリアントの割合」を底を2とする対数であらわした値である.アルギニンからグルタミンへの置換の場合を除くと, ほとんどの置換で病原性バリアントが無害なバリアントの2倍の頻度で生じていたことがlog2-oddsの値からわかった. アルギニンから別のアミノ酸への変化は, ABCトランスポーターの機能になんらかの影響を与えていると考えられる. SLCトランスポーターにも同様の傾向があり, 膜表面近くにアルギニンが置換される型の病原性バリアントがよくみられる[36]. ABCトランスポーターのアルギニン残基は膜表面以外にも存在しているが, アルギニン残基の置換が疾患の

原因になるという点では, SLCトランスポーターのもつ傾向とよく似ていた.

表4-2　ABCトランスポーターのバリアントのタイプ

無害			病原性			
置換	数	%	variation	数	%	log2-odds*
Val-Ile	13	6.5	Arg-Trp	23	5.7	0.7
Arg-Gln	13	6.5	Arg-His	22	5.4	1.44
Ala-Thr	8	4	Arg-Gln	22	5.4	-0.25
Arg-Trp	7	3.5	Arg-Cys	20	4.9	1.3
Ile-Val	7	3.5	Gly-Arg	15	3.7	1.3
Ala-Val	6	3	Glu-Lys	15	3.7	0.89
Ile-Met	5	2.5	Leu-Pro	11	2.7	2.44
Val-Met	5	2.5	Arg-Gly	10	2.5	1.3
Val-Ala	5	2.5	Ara-Val	10	2.5	-0.29
Arg-Leu	4	2.0	Arg-Pro	9	2.2	2.15
Arg-Cys	4	2.0	Trp-Cys	9	2.2	2.15
Arg-His	4	2.0	Arg-Leu	9	2.2	0.15
Glu-Lys	4	2.0				
Leu-Phe	4	2.0				
Pro-Leu	4	2.0				
Thr-Met	4	2.0				
Val-Leu	4	2.0				

*　同じバリアントの型の「無害なバリアントの割合に対する病原性バリアントの割合」を底を2とする対数であらわした値.（M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019).Table 2より改変・転載）

4-3-3. ABCAにおけるバリアントの局在

ヒトABCA遺伝子に存在するミスセンスバリアントをABCAタンパク質の立体構造（PDB ID: 5XJY）上にマップした（図4-1）. 赤色で示した残基は病原性バリアント, 青色で示した残基は無害なバリアントであることを示している. 病原性バリアントは, TMD2ドメインを除く領域に存在し, 無害なバリアントは, 2個のNBDドメインに集中しているように見えた. すべてのバリアントをマッピングしても,バリアントによる疾患原因につながるメカニズムを知る手がかりとなるような, 特異的な傾向を見出すことはできなかった.

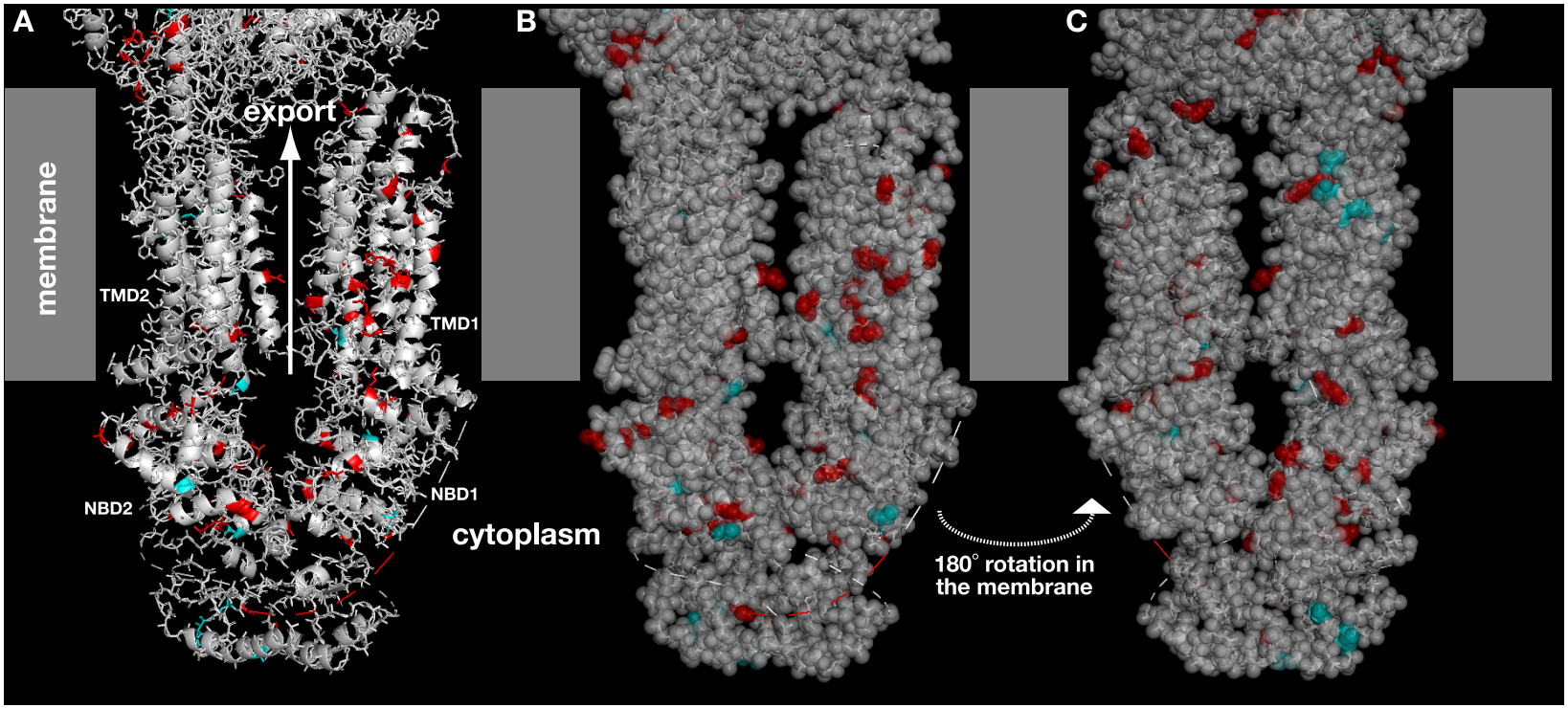

図4-1 ヒトABCAタンパク質のミスセンスバリアント

立体構造（PDB ID: 5XJY）上にマップされたヒトABCAタンパク質のミスセンスバリアント．赤色で示した残基は病原性バリアント，青色で示した残基は無害なバリアント．（(M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019).Figure 1より転載）

4-3-4. ABCB上の立体配座変化とバリアントの位置

ABCBでは多くの立体構造が決定されているが, 高解像度で, ヒトまたは近縁種の生物に由来する, それぞれ異なる条件下の4種類の立体構造を選択した（表4-1）. 選択した立体構造のうち, 図4-2にマウスP糖タンパク質の立体構造（PDB ID : 5KO2）を示す. ABCBの全体構造はABCAの全体構造と同様であるが, 図4-2に示される立体構造は2個のNBDが離れている状態である. なお, 図4-1に示したABCAの立体構造ではヌクレオチド結合ドメインが接触している状態である. 図4-3Aは, ABCBのうち阻害剤（PDB ID：4XWK）と結合したマウスP糖タンパク質（ABCB4）の立体構造をマウスP糖タンパク質（ABCB4）だけの構造（PDB ID：5KO2）と比較した差分地図である. 立体構造データには欠けている残基があるため, 軸の番号は不規則である. この差分地図は, 4XWKと5KO2に対して描かれた2つの距離地図の差分である. 両方の座標データは同じタンパク質に由来するため, 2つの距離地図は完全に対応していることが理想的である. しかし, 立体構造データには欠けている残基があり, 実際に描かれた差分地図にはギャップが生じた. 図4-3Aの場合, 最初の2個の残基（残基番号1および2）および中央の2個の残基（残基番号600および601）が4XWKでは欠けていたので, 対応する位置に2箇所のギャップがある. 本研究では立体構造中の残基の番号付けは

Cα原子を基準とした. N末端側から1番目のCα原子を有する残基は残基番号1, 2番目のCα原子を有する残基は残基番号2とし，その後も同様に番号が付けられた.

図4-3Bにみられる顕著なパターンのうち, 1つ目はN末端側ドメイン（TMD1 + NBD1）とC末端側ドメイン（TMD2 + NBD2）を分ける残基番号600付近の境界である. N末端の2個のドメインは一緒に, C末端の2個のドメインに反するように動いていた．残基番号600付近の対角軸上の四角形の領域はほぼ赤色であり，ドメインの組が阻害剤が結合することで離れていることを示した．図4-2では，この動きは, TMD1 + NBD1とTMD2 + NBD2との間の横方向の間隔に対応する．しかし, 2組のドメインは細胞外領域で結合しているため, 離れる動きは細胞外領域に位置する中心点を持つ回転運動と推測された.

2つ目の顕著なパターンは，以下の各ドメインでみられた. TMD 1 + NBD 1とTMD 2 + NBD 2は，差分地図の色のパターンに基づいて3つの領域に分けることができる. TMDの前半, TMDの中央，およびTMDの後半とNBDである．図4-2にTMDの分割を色分けして示した. TMDの前半をTMD-iとし赤系の色で, TMDの中央をTMD-iiとし青系の色で, TMDの後半をTMD-iiiとし黄色系の色とした．図

4-3AのTMD + NBDの赤色の帯として確認できるように, TMD-iiはTMD-iとTMD-iiiから遠ざかるように動き, NBDまでの距離を縮めた. P糖タンパク質に阻害剤が結合することで, TMD-iiはTMD-iとTMD-iiiの束から絞り出されるようなかたちで, NBDに近づいた. TMD-iiには細胞質側に短いヘリックスがあり，それは膜表面とほぼ平行で, TMDとNBDとの間のインターフェースになっている. このヘリックスはカップリングヘリックスと呼ばれている[120]. カップリングヘリックスは各TMDに2箇所ずつ存在し，このカップリングヘリックスは2番目のものである. その位置は，差分地図の対角軸上の黒色の四角形で示されている. TMD-iiは, NBDの立体配座変化をTMDに伝達する役割を果たしていた.

図4-3Aでアポ型P糖タンパク質と阻害剤結合型P糖タンパク質の比較によって観察された立体配座変化は，図4-3Bでアポ型とADP結合型（ヒトABCB8, PDB ID：5och）の比較でも観察されたが，立体配座変化の向きは反対であった. ヒトABCB8はホモ二量体のトランスポーターであり，それぞれのサブユニットは各1個のTMDとNBDから構成される. 図4-3Bの差分地図は, 2つのサブユニットの座標ファイルを連結したものを，アポ型マウスP糖タンパク質と比較することで描いた. 全体の配列一致度は約32％であった. 差分地図の色がついた領域の位

置は，図4-3Aの色がついた領域とほぼ同じであった．色がついた領域がほぼ同じであったことは，立体配座変化の方向が図4-3Aと図4-3Bでは逆であることを意味している. ADPが結合しているときトランスポーターは閉じていた，すなわち, TMD1 + NBD1とTMD2 + NBD2との間の距離はより近く，そしてカップリングヘリックスはNBDに対してより近くにあった．アポ型およびATP結合型（ヒトABCB1, PDB ID：6c0v）の間の比較においても同様な立体配座変化が観察された（図4-3C）．全体の配列同一性は約89％であった．図4-3Bおよび図4-3Cの色のついた領域の位置とその色はほぼ同じであった. ADPまたはATP結合型では，ADPまたはATPがNBDに結合するため, NBDの立体配座変化がTMDに伝達され，トランスポーターは閉じた構造をとるはずである．したがって, TMD-iiには情報伝達の役割があり，カップリングヘリックスはNBDとTMDのインターフェースとして働いていると思われる．

図4-3に示したすべての比較で，差分地図によって以下のような立体配座変化が明らかになった. 1) TMD1 + NBD1とTMD2 + NBD2との間の距離が増減する．2) TMD-iiとNBDとの間の距離が増減する. ABCBにおける立体配座変化は，さきに述べたように回転運動であると推測された．この推測を裏付けるため，さら

に差分プロットを描いた（図4-4A）．手法の節（4-3-4）で説明したように，差分プロットは残基番号に対してプロットされた各残基の差分の絶対値の平均である．図4-4Aには，TMDの差分プロットがNBDの差分プロットより低い値であり，NBDの立体配座変化がTMDの変化より大きいことが示されている．各TMDは，1.0Å付近に3箇所の極小値をもっていた．差分の絶対値の平均値が0であるとき，全体の配置が変化しても当該残基の相対位置は変化せず，立体配座変化の中心点となっている可能性がある．実際のタンパク質ではそのような理想的な状況は期待できないが，値が0に近いとき，当該残基は立体配座変化の中心点となっている可能性があり，低い値は立体配座変化が回転であることを示唆している．TMDに6箇所の低い値があることは，ABCBの立体配座変化が，開閉する蝶番のような運動であることを示している．立体構造中のこれら6個の残基の位置は図4-4Bに示した．図4-4Bは，図4-2の裏側を示したものである．図4-4Bの上側にある6個の残基は，図4-4AのTMDドメインにある6箇所の極小値に対応する．NBDのN末端側には，3.0Å前後の値をもつ一対の極小値がある．これらの残基は，局所的な立体配座変化の中心点候補となる．局所的な立体配座変化の中心点は，全体の位置としては大きく変化しているが，近くの残基に対してはわずかな変

化であろう．立体構造上，それらはNBDがTMDに接する部分に位置していた（図4-4B）．これらの残基は, TMDとNBDとの間の蝶番を開閉するような運動の中心点に位置していた．カップリングヘリックス（図4-4Aの黒い四角形）は, TMDのなかで最大値を示した. C末端のNBDの立体配座変化はN末端のNBDの立体配座変化より大きかった．この傾向はABCBに特有であるが，原因を特定できなかったため，さらに調べる必要がある．残基の平均差分値は残基の接触度との相関性があると推測されたが，両者に相関は見いだせなかった（相関係数0.15）．

バリアントによるアミノ酸置換のあった位置を図4-3Aから図4-3Cおよび図4-4Aにプロットした．赤色の点は病原性バリアントの場所を示し，青色の点は無害なバリアントの場所を示す．青色の点はタンパク質全体に位置しているが，赤色の点は主にTMD-ii（カップリングヘリックスの近く）またはNBDのいずれかにあった. TMD-iiにあるものは, NBDとTMDの間の情報伝達を妨げる可能性がある. NBDにあるものは, ATP結合残基（図4-3A, 4-3B, 4-3Cの対角線上の黒点）の近くにあった. ATP結合残基のバリアントは, ATP結合を阻害し，輸送体の機能障害をもたらすかもしれない．障害の程度が大きければ輸送体の機能は失われ，バリアントは致命的な影響を与えることになる．しかし, ATP結合は多くの

残基を含む集合的な機能であり（図中には53個の黒点が存在する）バリアントの影響は穏やかであると考えられ, バリアントが1箇所であればトランスポーターの機能をわずかに損なう程度の限定的な効果である可能性が高い. 他の領域では, バリアントは立体配座変化の中心点の近く, または膜との界面近くで見られた. これらのバリアントは, 3.0Å未満の平均差分値をもっていた（図4-4A）. そのうち3箇所は平均差分値が1.0Å付近であり, TMDの蝶番様運動の中心点となっていると考えられる. 中心点のわずかな障害は, 蝶番の動きに影響を与え, トランスポータータンパク質の立体配座変化の効力を低下させる可能性がある. 残りの4箇所は, 膜の界面近くにあった（図4-2）. これらのバリアントはタンパク質と膜の間の安定した相互作用を損なう作用をもつかもしれない.

Furutaら（2014）は大腸菌のABCトランスポーターMsbAを解析し, 第2カップリングヘリックスがNBDの点頭（うなずき）様の動きに重要であることを実験と分子動力学シミュレーションによって見出した[121]. 図4-3AにCH1とCH2としてこれらのカップリングヘリックスを示した. 本研究でカップリングヘリックスと呼んでいるのは第2カップリングヘリックス（CH2）のことである. 図4-4Aで, 第1カップリングヘリックスに比べ第2カップリングヘリックスの立体配座

変化が大きく, 2箇所の第2カップリングヘリックスのひとつに病原性バリアントが存在した. この結果は, Furutaらの知見とよく対応していた.

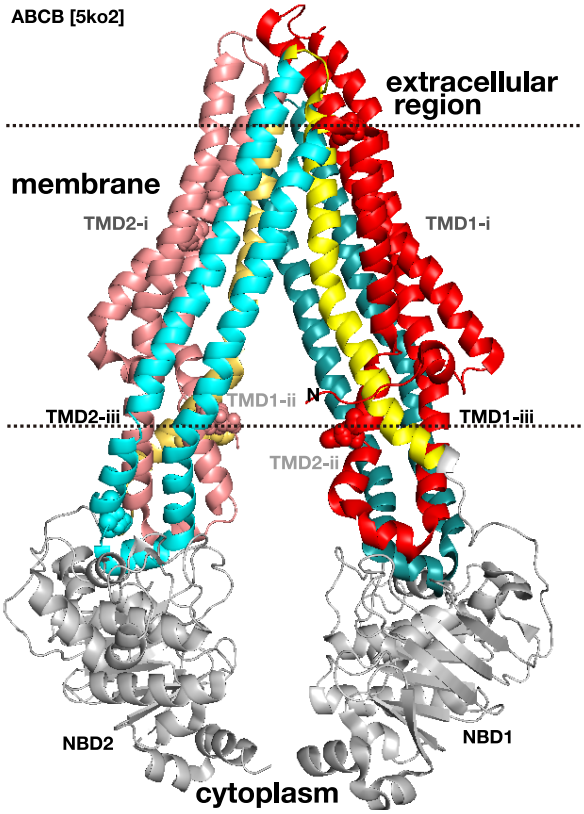

図4-2 ABCBタンパク質の立体構造

ABCBタンパク質の立体構造（PDB ID : 5KO2）. 図中の上側の破線は細胞膜と細胞外領域の境界, 破線で挟まれた領域は細胞膜, 下側の破線は細胞膜と細胞質の境界を示す. TMD1-i, TMD1-ii, TMD1-iii, TMD2-i, TMD2-ii, TMD2-iii ; 膜貫通ドメイン. 膜貫通ドメインの色分けは, 差分地図（図4-3）の色の境界に基づき, TMDの前半をTMD-iとし赤系の色で, TMDの中央をTMD-iiとし青系の色で, TMDの後半をTMD-iiiとし黄色系の色で示した. NBD1, NBD2 ; ヌクレオチド結合ドメイン. （M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019). Figure 2Aより転載）

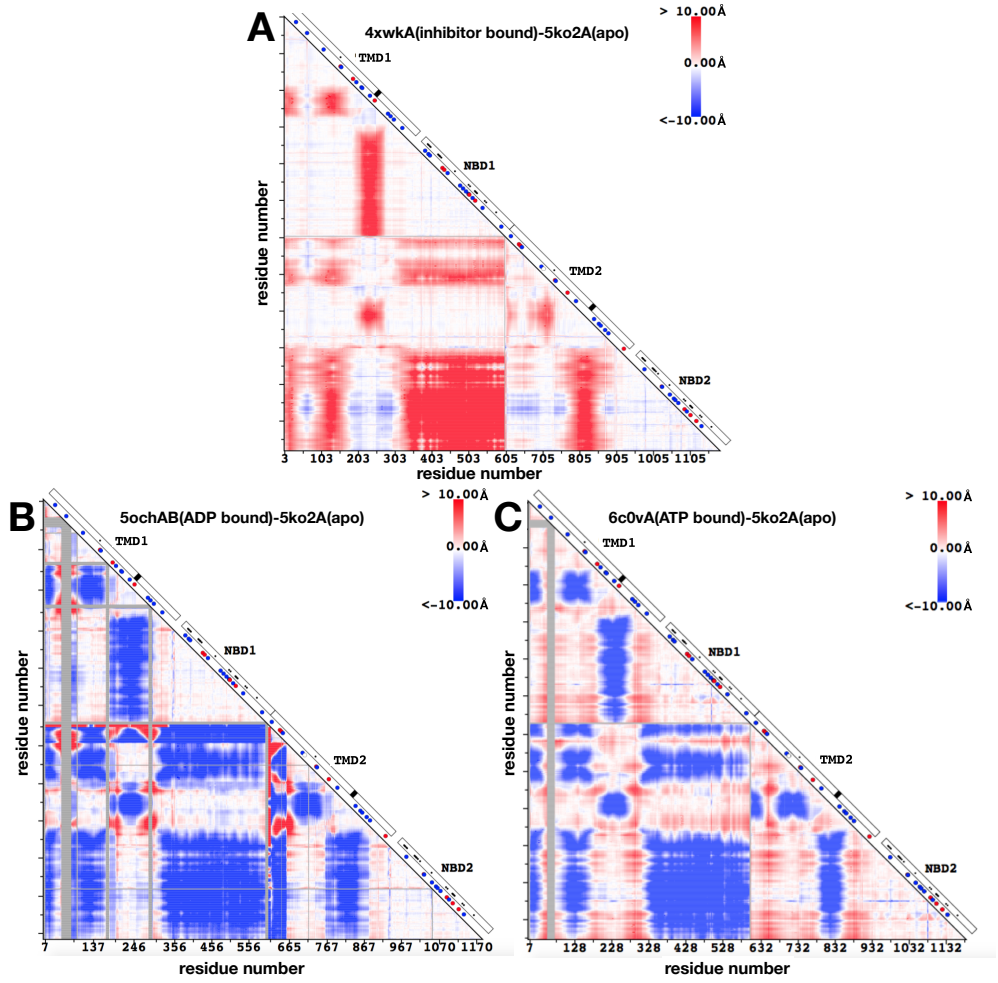

図4-3 ABCBタンパク質の差分地図

A. 5KO2（A鎖, apo型）と4XWK（A鎖，インヒビター結合型）の差分地図. B. 5KO2（A鎖, apo型）と5OCH（A鎖, ADP結合型）の差分地図. C. 5KO2（A鎖, apo型）と6C0V（A鎖, ATP結合型）の差分地図．色は差分の大小を示し，赤色が濃くなるほど対応する残基間の距離が大きく，青色が濃くなるほど対応する残基間の距離が小さい．対角線上の青丸は無害なバリアント，赤丸は病原性バリアントが存在する位置を示す．対角線上の白色四角形はドメイン構造を示す. TMD：膜貫通ドメイン, NBD：ヌクレオチド結合ドメイン．（M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019). Figure 2B~2Dより転載）

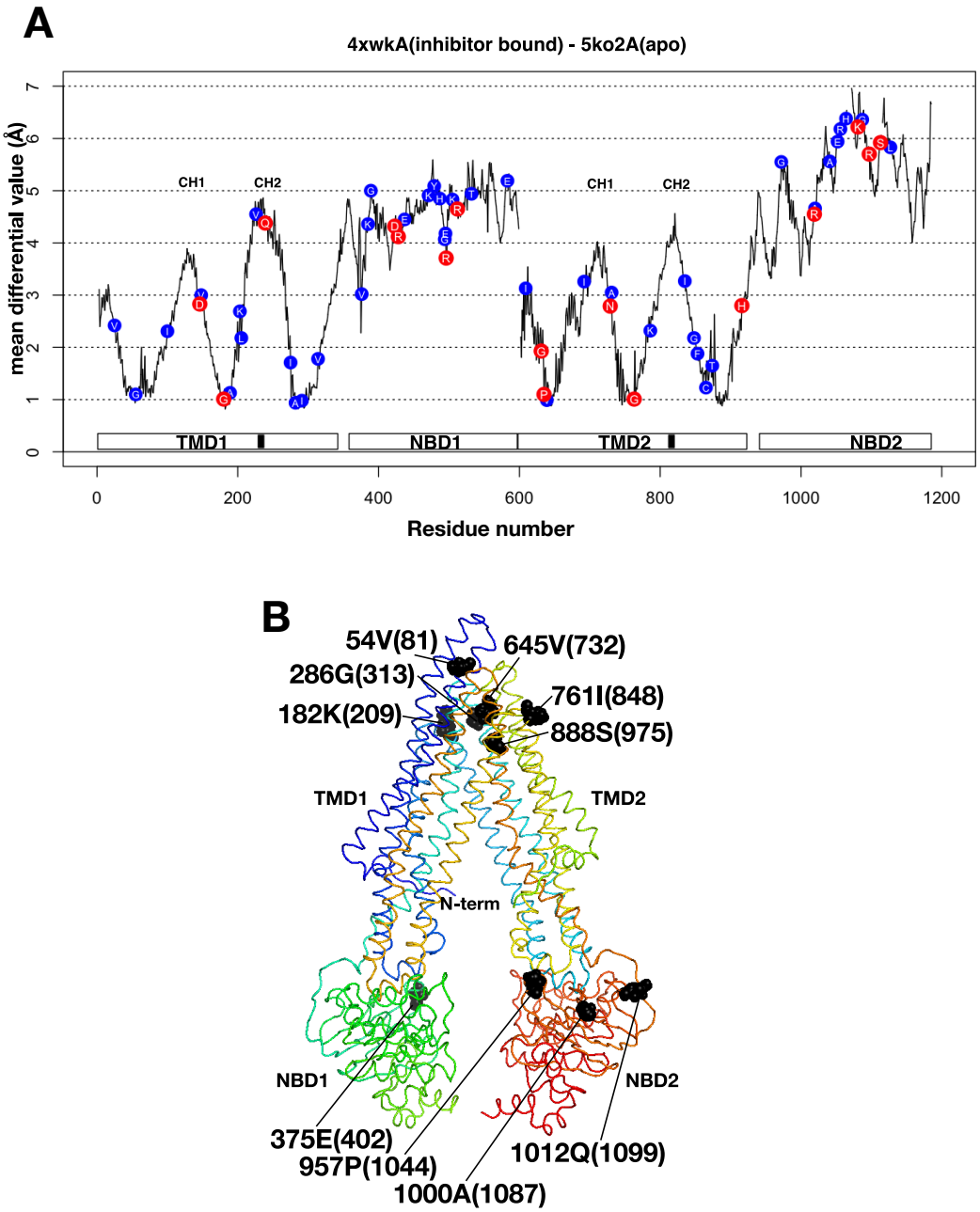

図4-4 ABCBタンパク質のバリアントの影響

A. 5KO2（A鎖, apo型）と4XWK（A鎖, インヒビター結合型）の差分地図から，アミノ酸残基ごとの差分の平均値をプロットしたもの. 縦軸は各アミノ酸残基における2種類の立体構造の差分の平均値, 横軸はアミノ酸残基番号を示す. 青丸は無害なバリアント，赤丸は病原

性バリアント，丸内のアルファベットはアミノ酸の1文字記号．横軸の白四角形はドメインを示す. TMD：膜貫通ドメイン, NBD：ヌクレオチド結合ドメイン, CH1：第1カップリングヘリックス, CH2：第2カップリングヘリックス. B. ABCBタンパク質の立体構造.図4-2の裏側に相当する．黒色で示す残基は，図4-4Aの極小値に対応する. (M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019).Figure 2E,2Fより転載）

4-3-5. ABCC の立体配座変化とバリアントの位置

ABCCの立体配座変化について，マウススルホニル尿素受容体SUR1のアポ型（PDB ID：5WUA）とADP結合型（PDB ID：5YWC），アポ型マウスSUR1とATP結合型のゼブラフィッシュ嚢胞性線維症膜貫通コンダクタンス調節因子（以下CFTR）（PDB ID：5W81）との間の立体配座変化を解析した．これら3種類の立体構造はすべて電子顕微鏡によって決定されたものである．差分地図と差分プロットを図4-5に示す．差分地図の色のついた領域と色の濃淡（図4-5Aと図4-5B）は, ABCBの差分地図（図4-3Bと図4-3C）とよく似ていた. ABCCにおいて見出された立体配座変化は, 1) TMD1 + NBD1とTMD2 + NBD2との間の距離の減少，および 2)カップリングヘリックスとNBDとの間の距離の減少であった. SUR1はN末端側ドメインを余分にもち, TMD1 + NBD1に付随して動く傾向があった．両方の差分プロット（図4-6Aおよび図4-6B）は, ABCBの差分プロット（図4-4A）と似た形状をしていた．しかし, ABCCにおける2個のNBDの平均差分値は, ABCBにおける平均差分値ほど違いがみられなかった．違いの原因は今後の研究で明らかになると思われる．

ABCCでは, ABCC6およびCFTR（ABCC7）で，ヒトゲノム上に膨大な数のバリアントが同定されており，そのほとんどは病原性である[113]．この章で

gnomADから抽出したミスセンスバリアントのうちABCC6は2,300件, CFTRは2,200件であった. これらのバリアントを差分地図および差分プロットにマッピングしたところ, ABCBで同定された領域を含むあらゆる場所に病原性バリアントが現れた（図4-5および図4-6）. ABCC上に病原性バリアントが広く分布していることは, ABCCが機能を発揮する機序がABCBの機序とはまったく異なることを示している. ABCCの配列多様性が許容されないことは, 多くのタンパク質間相互作用があることを示唆している. 実際に, SUR1はカリウムチャンネルと相互作用し, SUR1はチャンネルを囲むようにして閉じた立体配座を安定させていることが知られている[122].

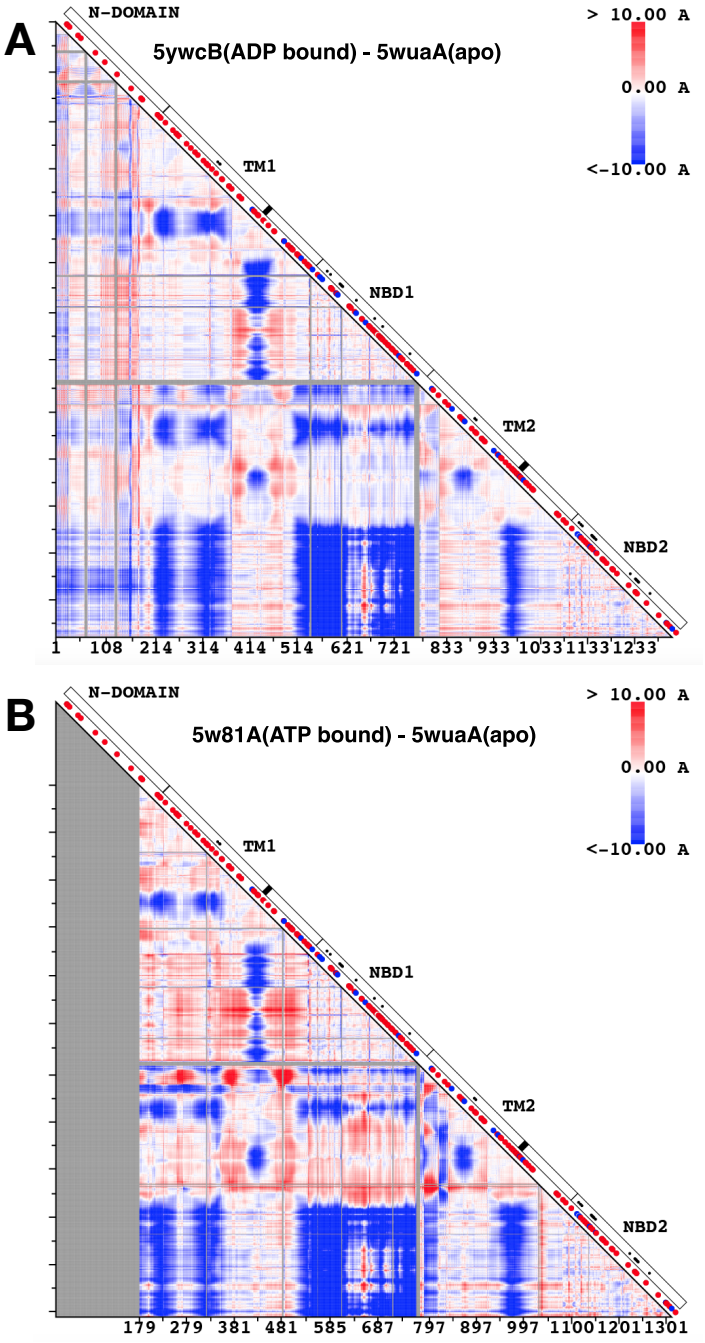

図4-5　ABCCタンパク質の差分地図

A. 5WUA（A鎖, apo型）と5YWC（B鎖, ATP結合型）の差分地図, B. 5WUA（A鎖, apo型）と5W61（A鎖, ATP結合型）の差分地図, D. 5WUA（A鎖, apo型）と5W61（A鎖, ATP結合型）の差分地図からアミノ酸残基ごとの差分の平均値をプロットしたもの. 差分地図（A, C）で

は対角線上の青丸は無害なバリアント, 赤丸は病原性バリアント, 白丸はconfrictとされたバリアント. 対角線上の白色四角形はドメインを示し, TM : 膜貫通ドメイン, NBD : ヌクレオチド結合ドメイン. 差分プロット（B, D）では青丸は無害なバリアント（Benign）, 赤丸は病原性バリアント（Pathogenic）, 丸内のアルファベットはアミノ酸の1文字記号. 対角線の白四角形はドメイン構造を示す. (M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019).Figure 3A,Cより転載）

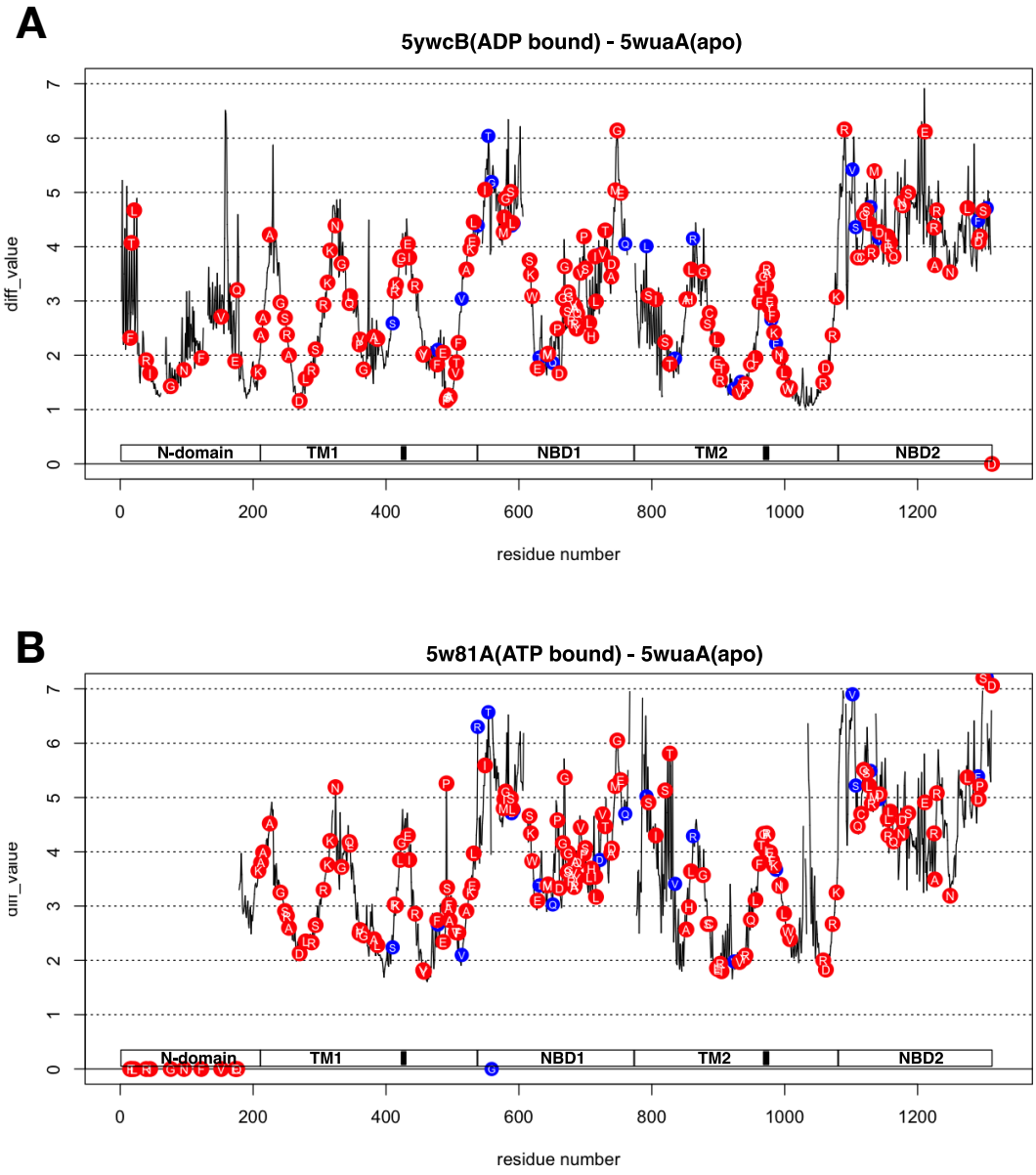

図4-6　ABCCタンパク質の差分プロット

A. 5WUA（A鎖, apo型）と5YWC（B鎖, ATP結合型）の差分地図からアミノ酸残基ごとの差分の平均値をプロットした. B. 5WUA（A鎖, apo型）と5W61（A鎖, ATP結合型）の差分地図からアミノ酸残基ごとの差分の平均値をプロットした. TM : 膜貫通ドメイン, NBD : ヌクレオチド結合ドメイン. 青丸は無害なバリアント（Benign）, 赤丸は病原性バリアント（Pathogenic）, 丸内のアルファベットはアミノ酸の1文字記号. 横軸の白四角形はドメイン

構造を示す. (M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019).Figure 3B,Dより転載）

4-3-6. ABCGの立体配座変化とバリアントの位置

ABCGの構造は，他のABCサブファミリーとは異なっている. ABCGは二量体タンパク質であり，各サブユニットは他のABCサブファミリータンパク質の2分の1に対応する．また, ABCGのドメインが並ぶ順番は他のABCサブファミリーとは異なる. ABCGはNBD-TMDと並び，他のABCタンパク質はTMD-NBDと並んでいる．これらの違いは，他のABCサブファミリーとABCGの静的立体配座および動的立体配座の違いとして反映されている．図4-7はヒトのステロールトランスポーターABCG5 / ABCG8の静的構造を示す. TMDは，他のABCトランスポーターのTMD（図4-2）と異なり，比較的平行な状態で膜に埋め込まれている．さらに, TMDのヘリックスは短い. ABCGの立体配座変化を，アポ型のヒトのステロールトランスポーターABCG5 / ABCG8ヘテロ二量体（PDB ID : 5DO7）とATP結合型のヒトABCG2（PDB ID : 6HBU）との間で解析した．配列一致度は約27%であった．各タンパク質のサブユニットを連結し，サブユニット位置の変化も差分地図内でみられるようにした. ABCGと他のABCトランスポーターとの間の立体構造の違いからもわかるように, ABCGの差分地図の着色領域および色の濃淡は他のABCトランスポーターとはまったく異なっていた（図4-8）. 2つのサブ

ユニットの接合部周辺の領域は，主に青色であり, ATP結合の際に2つのサブユニット間の距離が狭くなったことを意味する．他のABCトランスポーターのカップリングヘリックスに相当する領域は，図4-8において黒色の四角形として示した．この領域はLeeらによってCpHと名付けられている[123]. CpHには特徴的な立体配座変化は観察されなかった．その代わり，特徴的な立体配座変化がABCG5のNBDとTMDの接合部付近に見出された．この領域はドメインを連結するヘリックスであり，ヘリックスは膜の表面に平行に位置する．この領域はLeeらによってCnHと名付けられている[123]. CpHとCnHの平均差分値は高い値を示していた（図4-9）．これは, Leeらによって示されたように[123], CpHとCnHがABCGのカップリングヘリックスとして機能する可能性が高いことを示唆している. Ferreira R.J.らは分子動力学計算を行い，膜貫通ヘリックス2と3の間のループがカップリングヘリックスの機能をもっているのではないかということを示した（彼らはそのループを「カップリングループ」と命名した）[124]. 本研究では膜貫通ヘリックス2と3の間のループは，カップリングヘリックスに特徴的な立体配座変化を示さなかった．いずれの場合も, ABCGはNBDとTMDの間で異なる情報伝達経路をもっているようである.

ABCGの差分プロットを図4-9に示す．他のサブファミリーの差分プロットとは全く異なり，最小値は約1.0Åに達しなかったので, ATPが結合したときのABCGの立体配座変化における中心点となる残基は存在しなかった．また, NBDとTMDのあいだに明確な違いはなかった．しかし, 5箇所の病原性バリアントはプロットの極小値付近に位置していた．これらの位置は，図4-7に黄色の空間充填モデルとして描かれている. 5箇所の病原性バリアントはいずれも細胞外領域側の膜貫通ヘリックスの先端にある．この結果は，他のABCトランスポーターと同様に, ABCGのTMDがATP結合の際に蝶番様運動を受けている可能性を示唆しているが，その動きは最大でもTMDが存在する範囲に限定されている．中心点となる残基またはその付近のバリアントによる弱い影響がABCGの活性の低下をもたらし，疾患発症につながるのではないかと考えられる. Ferreira R.J.ら[124] はABCGタンパク質の分子動力学シミュレーションを行いポンプの機構について調べ，主にATPが結合したNBDの動きを分析した．彼らの論文によれば[124]，細胞外領域（本研究では蝶番様運動の中心点となる残基と病原性バリアントの存在を確認した）で，明らかな立体配座変化を示していた．図4-7に，平均差分値が4.0Åより大きい値をもつ病原性バリアントを，黒色の空間充填モデ

ルで示した．これらの病原性バリアントは, NBDの内部, または膜の細胞質との界面の近くにある．特に, 膜表面でアルギニンを他のアミノ酸に置換する2箇所のバリアントは, TMDと膜の間の相互作用を不安定にし, それによって膜中のABCGの位置または方向を不安定にする可能性がある[36]．図4-9は, より低い平均差分値, 特に4.0Å未満の値をもつ病原性バリアントの偏った分布を示した（Fisherの正確確率検定法で$P=5.0\times10^{-3}$）．このことは, 病原性バリアントがABCGの局所的な動きの中心点付近に存在する傾向があることを示唆していた.

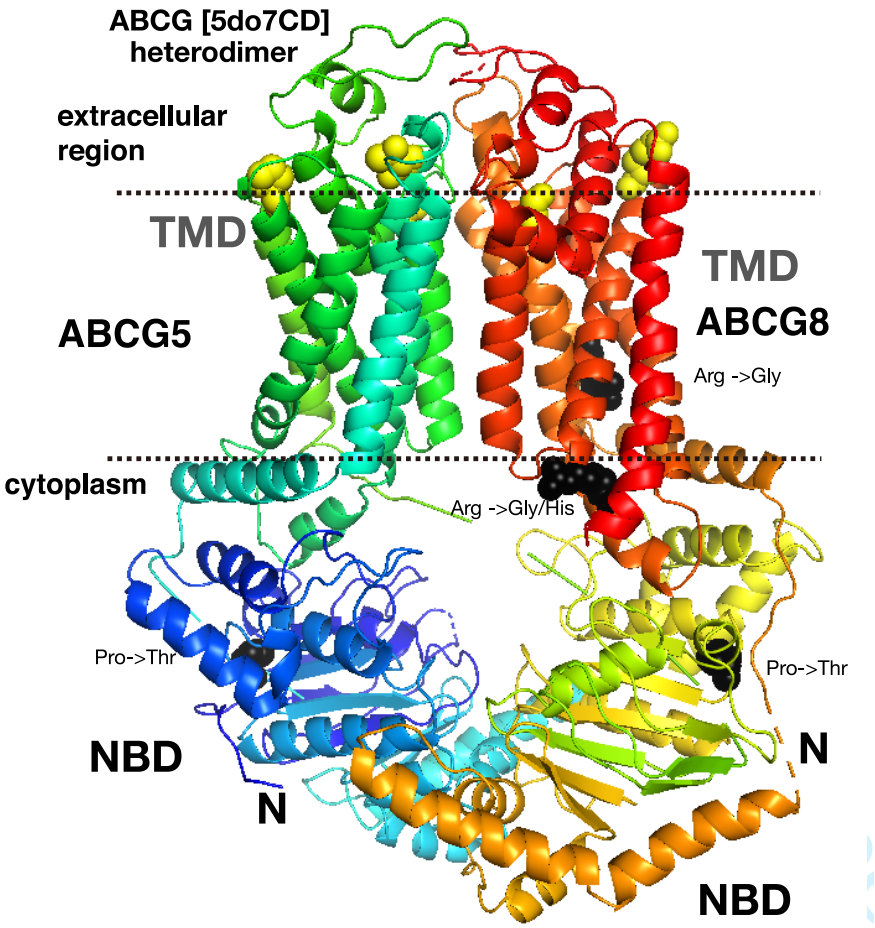

図4-7　ABCGタンパク質の立体構造

A. ABCGタンパク質の立体構造.（M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019). Figure 4Aより転載）

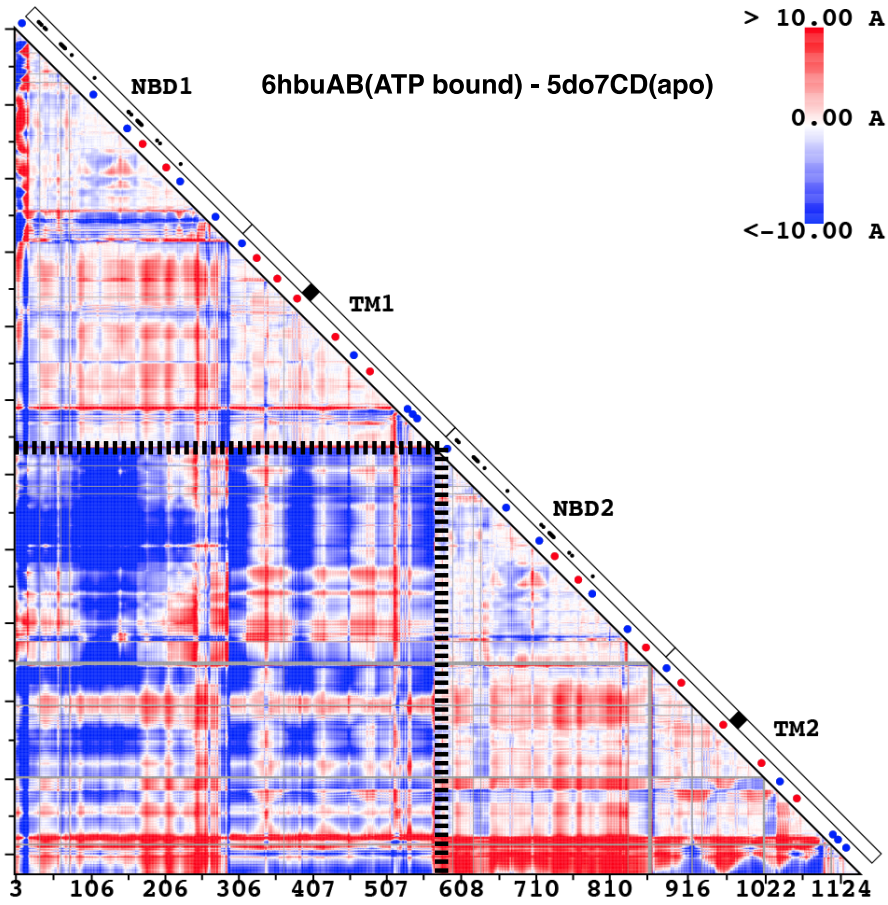

図4-8　ABCGタンパク質の差分地図

5DO7（C, D鎖, apo型）と6HBU（A, B鎖, ATP結合型）の差分地図. 青丸は無害なバリアント, 赤丸は病原性バリアント, TM : 膜貫通ドメイン, NBD : ヌクレオチド結合ドメイン, 対角線の白四角形はドメイン構造を示す.（M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019). Figure 4Bより転載）

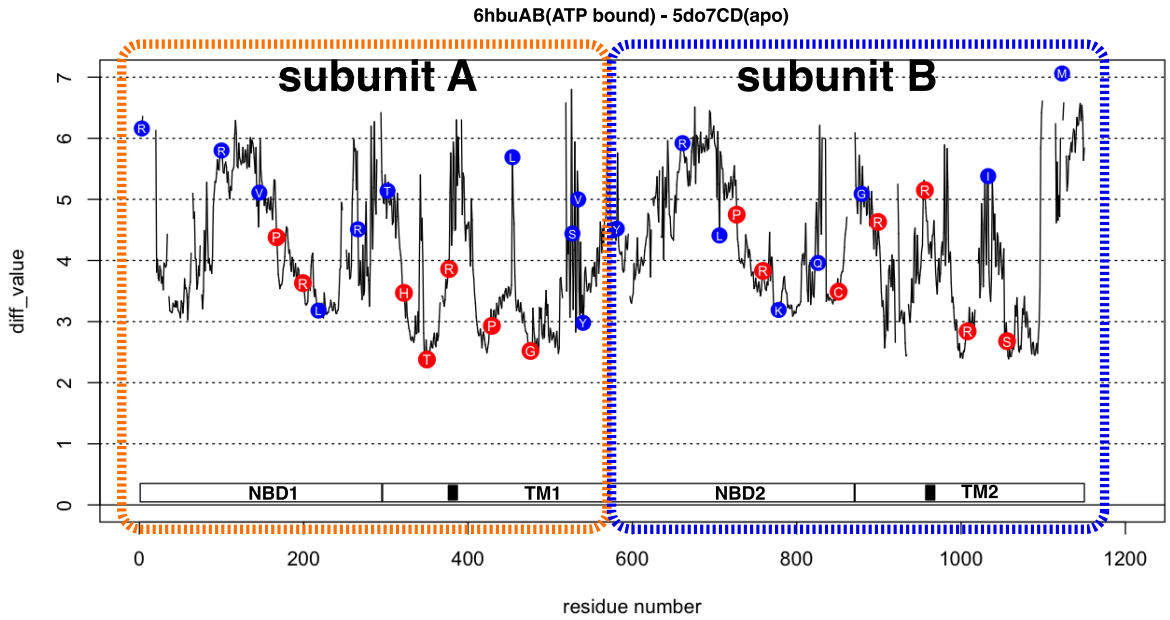

図4-9　ABCGタンパク質の差分プロット

5DO7（C, D鎖, apo型）と6HBU（A, B鎖, ATP結合型）の差分地図から，アミノ酸残基ごとの差分の平均値をプロットしたもの. TM : 膜貫通ドメイン, NBD : ヌクレオチド結合ドメイン，青丸は無害なバリアント，赤丸は病原性バリアント，丸内のアルファベットはアミノ酸の1文字記号. 横軸の白四角形はドメイン構造を示す.（M. Sakamoto, H. Suzuki, K. Yura, Relationship between conformation shift and disease related variation sites in ATP-binding cassette transporter proteins. *Biophys. Physicobiology*. **16**, 68–79 (2019). Figure 4より改変・転載）

4-5. 結論

この章では, Protein Data Bank（PDB）に登録されている異なる立体構造を比較することによって, ABCトランスポーターにおける, ATP結合によってもたらされる立体配座変化について, 差分地図および差分プロットを使用して立体配座変化の中心点となる残基をいくつか見出した. 中心点となる残基の位置はABCトランスポーターの各サブファミリーにおいてそれぞれ異なるが, それらは細胞外領域に位置するものが多かった. カップリングヘリックスの重要性は以前から多くの研究で報告されており, 本研究でもカップリングヘリックスの立体配座変化を確認した.

類似の研究が分子動力学シミュレーションを用いて展開されてきた[125–129]. 分子動力学シミュレーションは, タンパク質の原子運動を解明するための強力な方法であるが, ABCトランスポーター分子のサイズが非常に大きいため, 十分な空間の探索が難しかった. Condic-Jurkic K.ら（2018）は, ABCトランスポーターを含む大きなタンパク質の分子動力学シミュレーションの結果に疑問を投げかけている[130]. 本研究の分析は, 分子動力学シミュレーションがもつ不確実性に対処するための代替手段となり得る.

無害なバリアントと病原性バリアントをバリアントの位置から区別することは依然として困難な課題と考えられているが, バリアントがタンパク質機能に及ぼす影響を正確に予測することで問題が解決するであろう. 本研究では, 立体配座変化と変異の位置の情報を組み合わせて, 分子内回転の中心点となる残基のアミノ酸置換が病原性をもたらす傾向があることを見出した. この傾向はABCBにおいて顕著であり, ABCGにおいては統計的に有意な差があるという程度であるが, ABCCではタンパク質全体に散在する病原性バリアントの分布のため, 適用することは難しい. ATP結合部位および膜の表面に近い部位の情報に加えて, ABCトランスポーターの異なる立体構造の比較によって得られる分子内回転の中心点となる残基の情報は, ヒトゲノム配列にコードされているABCトランスポーター遺伝子に見られるバリアントのアノテーションの改善に役立つことが期待される.

第五章
総括

遺伝子バリアント解析における本研究の位置付けと成果

薬物代謝に関連する遺伝子として, P450バリアントについてはこれまで多くの研究が行われてきた. しかし単一の遺伝子や数個の遺伝子についての研究が多く, P450全体を俯瞰して解析したものは少ない. 本研究は, P450にみられるミスセンスバリアントの全体像を俯瞰し, そのタンパク質の性質における特徴を明らかにしようと試みたものである. また, 薬物などの小分子の輸送に関与するABCトランスポーターにも着目し, アミノ酸配列上のバリアントと立体構造上の特徴から, 病原性バリアントと立体構造の変化についても考察した.

遺伝子のなかにはバリアントが起こりやすい場所があることが知られている. ヒトを含む多くの生物のP450遺伝子バリアントの起こりやすい場所を, それまでに出版された論文からテキストマイニングの手法を用いて探索した研究では, 基質認識部位（SRS）やヘム結合領域にあるバリアントについての報告が多いことがわかった[131]. しかし, 地域によるバリアントの出現頻度の違いは, 遺伝子の構造によるものだけでなく, 遺伝的浮動や民族移動によるボトルネック効果などの結果であることも考えられる. 本研究において, ミスセンスバリアントによるアミノ酸置換のうち一部のものは, 出現頻度の地域差が存在する割合

が地域差が存在しないものにくらべて高い傾向があり，基質認識部位およびヘム結合領域のミスセンスバリアントは地域の違いにかかわらず一定の頻度でおこることが示唆され，遺伝的浮動や民族移動によるボトルネック効果などの結果として地域差が生じたことが考えられた．これは，P450と同じくゲノム内に多くの類似した遺伝子をもつ，HLA遺伝子について遺伝子頻度の違いやハプロタイプの頻度の違いが生じた要因としてあげられたものと同様であり[132]，ゲノム内に多くの類似した遺伝子をもつ多重遺伝子族の進化を考えるうえで，裏付けのひとつになっている．しかし，P450を基質特異性から内在性の生理活性物質型と異物の解毒型の2つの群の進化を比較するとそれぞれの進化速度が異なるという研究もあり[133]，進化という観点でP450を一括して論じるには本研究によって得られた結果では十分ではない可能性もある．

本研究では，タンパク質の立体構造およびタンパク質間相互作用から得られる空間的情報を加えたP450遺伝子バリアントのタンパク質への影響予測の可能性を示した．本研究で用いた手法をP450タンパク質だけではなく他のタンパク質にも応用できれば，これまで臨床的意義が未定義あるいは不明であった遺伝子バリアントにタンパク質の機能への影響に関する情報を追加することが可能

となる.他のタンパク質に本研究の手法を応用するには, 塩基置換またはアミノ酸置換の箇所をタンパク質の立体構造にマッピングすることが不可欠である. 本研究ではBLASTで相同性検索したのちにMUSCLEで配列アラインメントを行った結果を用いたが, 同様な結果を得られるツールとしてPDBjのHOMCOS（http://homcos.pdbj.org/）[134, 135]がある. HOMCOSの「結合分子の検索・タンパク質に対する結合分子の検索」では, アミノ酸配列を出発点として, BLASTによる相同性検索により適合する立体構造を推定し, 各アミノ酸残基の生物間の保存性や溶媒接触度を得ることができる. 今後, これらの情報に追加して, 既知のタンパク質への影響ありバリアントやタンパク質間相互作用から得た情報を反映させた予測モデルを作ることができれば, さまざまなタンパク質のバリアント影響予測に役立つと考えられる.

ABCトランスポータータンパク質では, 条件の異なる立体構造を比較することにより, ATP結合によってもたらされる立体配座変化について検討し, 立体配座を変化させる分子内自由回転の軸となる複数の残基を見出した. そして, 立体配座変化と変異の位置の情報を組み合わせて, 分子内回転の中心点となる残基のアミノ酸置換がタンパク質への影響をあたえる傾向があることを見出した.

ABCトランスポーターのように分子のサイズが非常に大きいタンパク質は，分子動力学シミュレーションで十分な位相空間の探索が困難とされているが，本研究で行われた分析手法は，分子動力学シミュレーションが持つあいまいさに対処するための代替手段となり得ることを示した．

遺伝子バリアント解析における展望

本研究において，P450タンパク質の立体構造の情報を用いた機械学習による予測モデルが精度の高い分類を提供することができたことから，タンパク質の立体構造から得られる情報，例えばタンパク質の立体構造から得られるアミノ酸残基間，またはタンパク質と別のタンパク質との空間的情報は，バリアントとタンパク質への影響との関連性を見出すことについて有用であることが示唆された．また，立体構造データの比較によりバリアントの影響をタンパク質分子内のダイナミクスから解析する試みは，計算コストをかけなくても擬似的な分子動力学的解析が可能であることを明らかにした．上記の予測モデルとの組み合わせで，立体構造データを用いたタンパク質のバリアント影響予測をさらに改善させることが期待される．

NGSのコストが下がる[7]とともに，疾患の診断にゲノム解析やエキソーム解析が多く行われるようになったが[1]，疾患原因となるバリアント以外に意義不明のバリアント（Variants of Uncertain Significance, VUS）も多く見つかり，その取り扱いについて臨床では報告義務の要不要，膨大なバリアントの解釈など様々な問題を抱えている[32, 136]．検査自体のコストが下がって気軽に遺伝子検査ができるようになっても，膨大なデータの処理と解釈に多大なコストがかかる可能性がある．バリアントデータの解釈については疾患や薬剤代謝を熟知した専門家による実証が必要であるが，急増するデータ量に追いつくかどうかは未知である．しかし，生命情報学的手法は，これまでに蓄積された『専門家による知識と実証』をもとに，この問題を解決することが可能である．本研究のような，バリアントの意味づけについての新しい視点と手法は，意義不明とされたバリアントの疾患への関連への理解をさらに進めることができる．個人向け遺伝子検査についても，その結果と医師の適切なフォローにより疾患にかかりにくい予防的な生活習慣に変える人が多いという報告がある[5]．解釈可能なバリアントが増えることで，個人向け遺伝子検査内容がさらに充実し，疾患の予防につながる．疾患の原因解明以外でも，遺伝子バリアントが関連する事項と

して，薬物の効果や病状などの個人の多様性に基づいた医療を行う個別化医療（precision medicine）が注目されている[28]．遺伝子の多様性は，疾患に関する個人差や薬剤効果の個人差が生じる大きな原因である．バリアントの解釈が進み，個人向け遺伝子検査が充実することは，疾患のかかりやすさを知るだけでなく，治療薬の副作用予測や投与量調節にも役立つ．そして個別化医療の内容がよりきめ細やかになることが期待される．

用語説明

本論文では遺伝子配列の多様性を表す用語を，日本人類遺伝学会『遺伝学用語の改訂』（2009年11月改訂, http://jshg.jp/about/notice-reference/），米国臨床遺伝・ゲノム学会（American College of Medical Genetics and Genomics, ACMG）の配列多様性解釈についてのガイドライン[9]に準じて用いた.

第一章

次世代シークエンサー（NGS）

高品質かつ一度に大量の塩基配列を解読することができる機器

全エキソーム解析

全ゲノム配列の約1~2%に相当する，タンパク質の情報をもつ部分を解読する方法

VCF

variant call formatの略．ヒトゲノムなどのリファレンス配列に対する塩基配列の多様性を記述するためのフォーマット.

P450

シトクロムP450

SLCトランスポーター

ソリュートキャリアートランスポーター

ABCトランスポーター

ATP結合カセットトランスポーター

第二章

アレル

1つの座位につき，遺伝子の種類が複数存在する場合のひとつひとつのこと.
1つの塩基に複数の種類がある場合も指す.

座位
ゲノム上におけるDNA配列（遺伝子および遺伝子以外）の占める位置

ハプロタイプ
ひとつの相同染色体上のアレルの組み合わせ.

バリアント
塩基配列またはアミノ酸配列に, リファレンス配列と異なる配列をもつ遺伝子またはタンパク質.

ミスセンスバリアント
コドンが別のアミノ酸を指定するものに変わるバリアント.

遺伝的浮動
有限数の生物集団のなかで, 遺伝子頻度が偶然に変化していくこと.

ボトルネック効果
なんらかの原因で生物集団の個体数が激減し, 残った個体がさらに増えることで遺伝子頻度が変化すること.

第三章
主成分負荷量
主成分解析において, それぞれの主成分に対する特徴量の相関の強さを示す値.

正確度（Accuracy)
正確な分類を行なった数をすべてのサンプルの個数で割った数.

適合率（Precision）
陽性と分類したものが実際に陽性であった割合.

再現率（Recall）
実際に陽性であるものが陽性と分類されたものの割合.

f1-値
適合率と再現率の調和平均.

受信者操作特性（ROC）グラフ
横軸を偽陽性率（実際に陰性であるものを陽性と分類したものの割合）, 縦軸を真陽性率（実際に陽性であるものが陽性と分類されたものの割合）とし, 分類の閾値を変化させたときの偽陽性率と真陽性率をプロットしたグラフ.

受信者操作特性（ROC）グラフ下面積（AUC）
受信者操作特性（ROC）グラフ下領域の面積. この値が大きいほど良いモデルといえる.

第四章
立体配座
原子間の単結合の回転などにより, 相互に変換可能な空間的な原子の配置.

TMD
膜貫通ドメイン

NBD
ヌクレオチド結合ドメイン

距離地図
タンパク質のアミノ酸配列に番号をつけ, その数字をグラフの縦軸と横軸に並べる. アミノ酸残基 iとアミノ酸残基 j の間の距離を2個のCα原子の距離として定義し, d_{ij} とする.この値を三角形の内側の i 側と j 側の交点に示したグラフ.

差分地図
異なる2種のタンパク質立体構造の距離地図の差分として描かれるグラフ. アミノ酸残基 iとアミノ酸残基 j の間の距離を2個のCα原子の距離として定義し, d_{ij} としたとき, タンパク質立体構造Bの距離地図から立体構造Aの距離地図を引い

た差分は以下の式で表される．

$$\Delta D_{ij}^{B-A} = d_{ij}^{B} - d_{ij}^{A}$$

参考文献

1. A. Fernandez-Marmiesse, S. Gouveia, M. L. Couce, NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. *Curr. Med. Chem.* **25**, 404–432 (2018).
2. T. Adachi *et al.*, Japan's initiative on rare and undiagnosed diseases (IRUD): Towards an end to the diagnostic odyssey. *Eur. J. Hum. Genet.* **25**, 1025–1028 (2017).
3. T. Adachi *et al.*, Survey on patients with undiagnosed diseases in Japan: potential patient numbers benefiting from Japan's initiative on rare and undiagnosed diseases (IRUD). *Orphanet J. Rare Dis.* **13**, 208 (2018).
4. 鎌谷直之，個人ゲノムが開く健康と医療の未来．痛風と核酸代謝**. 39**, 172 (2015).
5. M. Hayashi, A. Watanabe, M. Muramatsu, N. Yamashita, Effectiveness of personal genomic testing for disease-prevention behavior when combined with careful consultation with a physician: A preliminary study. *BMC Res. Notes*. **11**, 1–6 (2018).
6. 渡邉淳，診療・研究にダイレクトにつながる遺伝医学 (羊土社, 2017; http://ci.nii.ac.jp/ncid/BB23476645).
7. K. J. M. Van Nimwegen *et al.*, Is the $1000 genome as near as we think? A cost analysis of next-generation sequencing. *Clin. Chem.* **62**, 1458–1464 (2016).
8. Y. Yang *et al.*, Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
9. S. Richards *et al.*, Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
10. J. T. den Dunnen *et al.*, HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum. Mutat.* **37**, 564–569 (2016).
11. R. M. Durbin *et al.*, A map of human genome variation from population-scale sequencing. *Nature*. **467**, 1061–73 (2010).
12. G. A. McVean *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature*. **491**, 56–65 (2012).

13. A. Auton *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).
14. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature*. **526**, 75–81 (2015).
15. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. **536**, 285–291 (2016).
16. M. Narahara *et al.*, Large-scale East-Asian eQTL mapping reveals novel candidate genes for LD mapping and the genomic landscape of transcriptional effects of sequence variants. *PLoS One*. **9** (2014), doi:10.1371/journal.pone.0100924.
17. K. Higasa *et al.*, Human genetic variation database, a reference database of genetic variations in the Japanese population. *J. Hum. Genet.* **61**, 547–553 (2016).
18. M. Nagasaki *et al.*, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).
19. Y. Yamaguchi-Kabata *et al.*, iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing. *Hum. Genome Var.* **2**, 15050 (2015).
20. M. J. Landrum *et al.*, ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
21. K. E. Caudle *et al.*, Evidence and resources to implement pharmacogenetic knowledge for precision medicine. *Am. J. Heal. Pharm.* **73**, 1977–1985 (2016).
22. 加藤隆一，山添康，横井毅，薬物代謝学：医療薬学・医薬品開発の基礎として（東京化学同人，第3版., 2010).
23. 佐藤洋美，上野光一，薬物代謝における性差．ファルマシア. **47** (2011), pp. 218–222.
24. A. N. Dong, B. H. Tan, Y. Pan, C. E. Ong, Cytochrome P450 genotype-guided drug therapies: An update on current states. *Clin. Exp. Pharmacol. Physiol.*, 1–11 (2018).
25. T. KIRINO *et al.*, A Case Report of a Phenytoin Toxic Stroke Patient with Genetic CYP2C19 Polymorphism. *Japanese J. Rehabil. Med.* **45**, 617–622 (2008).

26. K. R. Pandey, N. Maden, B. Poudel, S. Pradhananga, A. K. Sharma, The Curation of Genetic Variants: Difficulties and Possible Solutions. *Genomics, Proteomics Bioinforma.* **10**, 317–325 (2012).
27. D. Salgado, M. I. Bellgard, J. P. Desvignes, C. Béroud, How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era. *Hum. Mutat.* **37**, 1272–1282 (2016).
28. 渡邉淳，ファーマコゲノミクス（PGx）とオーダーメイド医療．日本医科大学医学会雑誌．**8**, 9–17 (2012).
29. 日本臨床検査医学会 日本人類遺伝学会，日本臨床検査標準協議会，ファーマコゲノミクス検査の運用指針 (2012), (available at http://jshg.jp/wp-content/uploads/2017/08/120702PGx.pdf).
30. T. M. Ko, C. S. Wong, J. Y. Wu, Y. T. Chen, Pharmacogenomics for personalized pain medicine. *Acta Anaesthesiol. Taiwanica*. **54**, 24–30 (2016).
31. M. Arbitrio *et al.*, Pharmacogenomic Profiling of ADME Gene Variants: Current Challenges and Validation Perspectives. *High-throughput*. **7**, 1–12 (2018).
32. 倉橋浩樹，ゲノム医療の現状と遺伝カウンセリング．日本血栓止血学会誌．**28**, 9–15 (2017).
33. 三村純正，藤井義明，薬物異物と転写制御．蛋白質核酸酵素．**48**, 2261–2266 (2003).
34. F. Stevison *et al.*, Does in vitro CYP down-regulation translate to in vivo drug-drug interactions? Preclinical and clinical studies with 13- *cis* -retinoic acid. *Clin. Transl. Sci.* (2019), doi:10.1111/cts.12616.
35. T. A. Peterson, E. Doughty, M. G. Kann, Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. *J. Mol. Biol.* **425**, 4047–4063 (2013).
36. A. Higuchi, N. Nonaka, K. Yura, iMusta4SLC: Database for the structural property and variations of solute carrier transporters. *Biophys. Physicobiology*. **15**, 94–103 (2018).
37. Y. Hiruma *et al.*, The structure of the cytochrome P450cam-putidaredoxin complex determined by paramagnetic NMR spectroscopy and crystallography. *J. Mol. Biol.* **425**, 4353–4365 (2013).

38. A. McKenna *et al.*, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics*. **27**, 2156–2158 (2011).
40. W. McLaren *et al.*, The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).
41. A. Frankish *et al.*, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, 766–773 (2018).
42. K. Eilbeck *et al.*, The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
43. D. F. Lewis, 57 varieties: the human cytochromes P450. *Pharmacogenomics*. **5**, 305–318 (2004).
44. D. R. Nelson *et al.*, P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*. **6**, 1–42 (1996).
45. 吉田雄三，後藤修, Diversozyme P450の進化と多様性．化学と生物. **36** (1998), pp. 393–398.
46. 今井嘉郎, P450の基質認識:多様性の構造的根拠．化学と生物. **36** (1998), pp. 530–533.
47. 大村恒雄，石村巽，藤井義明，講談社サイエンティフィク, P450の分子生物学 (講談社, 2003; http://ci.nii.ac.jp/ncid/BA64085685).
48. 高橋芳樹，鎌滝哲也，肝薬物代謝の最近の進歩 3.チトクロームP450．肝臓. **42**, 288–296 (2001).
49. D. F. V. Lewis, P. Hlavica, Interactions between redox partners in various cytochrome P450 systems: Functional and structural aspects. *Biochim. Biophys. Acta - Bioenerg.* **1460**, 353–374 (2000).
50. T. Omura, Forty years of cytochrome P450. *Biochem. Biophys. Res. Commun.* **266**, 690–698 (1999).
51. 武森重樹，小南思郎，チトクロムP-450 (東京大学出版会, 1990; http://ci.nii.ac.jp/ncid/BN05250973), *UP biology*.
52. H. Koga *et al.*, Essential role of the Arg112 residue of cytochrome P450cam for electron transfer from reduced putidaredoxin. *FEBS Lett.* **331**, 109–113 (1993).

53. N. Strushkevich *et al.*, Structural basis for pregnenolone biosynthesis by the mitochondrial monooxygenase system. *Proc. Natl. Acad. Sci.* **108**, 10139–10143 (2011).
54. P. R. Ortiz de Montellano, Ed., *Cytochrome P450* (Springer US, Boston, MA, 2005; http://link.springer.com/10.1007/b139087).
55. S. C. Sim, M. Ingelman-Sundberg, The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Hum. Genomics* (2010), doi:10.1186/1479-7364-4-4-278.
56. A. Gaedigk *et al.*, The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin. Pharmacol. Ther.* **103**, 399–401 (2018).
57. K. Sakuyama *et al.*, Functional Characterization of 17 CYP2D6 Allelic Variants (CYP2D6 . 2 , 10 , 14A – B , 18 , 27 , 36 , 39 , 47 – 51 , 53 – 55 , and 57). *Pharmacology*. **36**, 2460–2467 (2008).
58. Y. Niinuma *et al.*, Functional characterization of 32 CYP2C9 allelic variants. *Pharmacogenomics J.* **14**, 107–114 (2014).
59. M. G. Scordo *et al.*, Genetic polymorphism of cytochrome P450 2C9 in a Caucasian and a black African population. *Br. J. Clin. Pharmacol.* **52**, 447–450 (2001).
60. T. Kubota, K. Chiba, T. Iga, Frequency Distribution of CYP2C19 ,CYP2D6 and CYP2C9 Mutant-alleles in Several and Different Populations. *Xenobio. Metabol. and Dispos.* **16**, 69–74 (2001).
61. T. Kubota, Y. Yamaura, N. Ohkawa, H. Hara, K. Chiba, Frequencies of CYP2D6 mutant alleles in a normal Japanese population and metabolic activity of dextromethorphan O-demethylation in different CYP2D6 genotypes. *Br. J. Clin. Pharmacol.* **50**, 31–34 (2000).
62. A. Ishiguro, T. Kubota, H. Sasaki, Y. Yamada, T. Iga, Common mutant alleles of CYP2D6 causing the defect of CYP2D6 enzyme activity in a Japanese population [2]. *Br. J. Clin. Pharmacol.* **55**, 414–415 (2003).
63. S. Bernard, K. A. Neville, A. T. Nguyen, D. A. Flockhart, Interethnic differences in genetic polymorphisms of CYP2D6 in the U.S. population: clinical implications. *Oncologist*. **11**, 126–35 (2006).

64. A. Bhathena *et al.*, Frequency of the frame-shifting CYP2D7 138delT polymorphism in a large, ethnically diverse sample population. *Drug Metab. Dispos.* **35**, 1251–1253 (2007).
65. N. Liu *et al.*, Influence of common and rare genetic variation on warfarin dose among African-Americans and European-Americans using the exome array. *Pharmacogenomics*. **18**, 1059–1073 (2017).
66. D. K. Rao *et al.*, Distribution of CYP2C8 and CYP2C9 amino acid substitution alleles in South Indian diabetes patients: A genotypic and computational protein phenotype study. *Clin. Exp. Pharmacol. Physiol.* **44**, 1171–1179 (2017).
67. V. Krasniqi *et al.*, Genetic polymorphisms of CYP2C9, CYP2C19, and CYP3A5 in Kosovar population. *Arh. Hig. Rada Toksikol.* **68**, 180–184 (2017).
68. S. P. Myrand *et al.*, Pharmacokinetics/genotype associations for major cytochrome P450 enzymes in native and first- and third-generation Japanese populations: Comparison with Korean, Chinese, and Caucasian populations. *Clin. Pharmacol. Ther.* **84**, 347–361 (2008).
69. T. Ota *et al.*, Combination analysis in genetic polymorphisms of drug-metabolizing enzymes CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A5 in the Japanese population. *Int. J. Med. Sci.* **12**, 78–82 (2015).
70. S. Ozawa, Drug-Drug Interactions with Consideration of Pharmacogenetics. *Yakugaku Zasshi*. **138**, 365–371 (2018).
71. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
72. D. R. Zerbino *et al.*, Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
73. R. A. Fisher, On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *J. R. Stat. Soc.* **85**, 87 (1922).
74. A. Agresti, A Survey of Exact Inference for Contingency Tables. *Stat. Sci.* **7**, 131–153 (1992).
75. A. R. Kinjo *et al.*, Protein Data Bank Japan (PDBj): Updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res.* **45**, D282–D288 (2017).
76. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).

77. O. Gotoh, Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J. Biol. Chem.* **267**, 83–90 (1992).
78. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
79. C. J. Sigrist *et al.*, New and continuing developments at PROSITE. TL - 41. *Nucleic Acids Res.* **41 VN-r**, 7 (2013).
80. D. S. Lee *et al.*, Substrate recognition and molecular mechanism of fatty acid hydroxylation by cytochrome P450 from Bacillus subtilis: Crystallographic, spectroscopic, and mutational studies. *J. Biol. Chem.* **278**, 9761–9767 (2003).
81. J. Catalano, K. Sadre-Bazzaz, G. A. Amodeo, L. Tong, A. McDermott, Structural Evidence: A Single Charged Residue Affects Substrate Binding in Cytochrome P450 BM-3. *Biochemistry*. **52**, 6807–6815 (2013).
82. P. Lafite *et al.*, Role of arginine 117 in substrate recognition by human cytochrome P450 2J2. *Int. J. Mol. Sci.* **19** (2018), doi:10.3390/ijms19072066.
83. S. Ahuja *et al.*, A model of the membrane-bound cytochrome b5-cytochrome P450 complex from NMR and mutagenesis data. *J. Biol. Chem.* **288**, 22080–22095 (2013).
84. M. KIMURA, The neutral theory of molecular evolution: A review of recent evidence. *Japanese J. Genet.* **66**, 367–386 (1991).
85. M. Nei, Bottlenecks, genetic polymorphism and speciation. *Genetics*. **170**, 1–4 (2005).
86. M. A. Jobling, C. Tyler-smith, Human Y-chromosome variation in the genome-sequencing era. *Nat. Publ. Gr.*, doi:10.1038/nrg.2017.36.
87. L. Breiman, Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
88. N.-L. Sim *et al.*, SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
89. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
90. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods*. **7**, 248–249 (2010).
91. J. M. Schwarz, D. N. Cooper, M. Schuelke, D. Seelow, MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*. **11**, 361 (2014).

92. K. Fechter, A. Porollo, MutaCYP: Classification of missense mutations in human cytochromes P450. *BMC Med. Genomics*. **7**, 1–9 (2014).
93. C. M. Durand *et al.*, CYP2U1 activity is altered by missense mutations in hereditary spastic paraplegia 56. *Hum. Mutat.* **39**, 140–151 (2018).
94. H. Terui, K. Akagi, H. Kawame, K. Yura, CoDP: predicting the impact of unclassified genetic variants in MSH6 by the combination of different properties of the protein. *J. Biomed. Sci.* **20**, 25 (2013).
95. Z. Liang, J. X. Huang, X. Zeng, G. Zhang, DL-ADR: A novel deep learning model for classifying genomic variants into adverse drug reactions. *BMC Med. Genomics*. **9** (2016), doi:10.1186/s12920-016-0207-4.
96. A. C. Müller, S. Guido, 中田秀基, Pythonではじめる機械学習：scikit-learnで学ぶ特徴量エンジニアリングと機械学習の基礎 (オライリー・ジャパン, オーム社 (発売), 2017; http://ci.nii.ac.jp/ncid/BB23712169).
97. D. R. Cox, The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. Ser. B*. **20**, 215–242 (1958).
98. V. Vapnik, Pattern recognition using generalized portrait method. *Autom. Remote Control*. **24**, 774–780 (1963).
99. M. Nakajima, Role of MicroRNAs in the Regulation of Cytochrome P450s and Transcriptional Factors. *YAKUGAKU ZASSHI*. **132**, 107–116 (2012).
100. 吉成浩一，薬物代謝酵素がかかわる薬物相互作用．ファルマシア. **50**, 654–658 (2014).
101. 今井嘉郎，鎌滝哲也, P450ーその多様な機能と応用ー．蛋白質核酸酵素. **43**, 203–215 (1998).
102. K. Hollenstein, R. J. Dawson, K. P. Locher, Structure and mechanism of ABC transporter proteins. *Curr. Opin. Struct. Biol.* **17**, 412–418 (2007).
103. V. Vasiliou, K. Vasiliou, D. W. Nebert, Human ATP-binding cassette (ABC) transporter family. *Hum. Genomics*. **3**, 281–290 (2009).
104. E. Gouaux, R. Mackinnon, Principles of selective ion transport in channels and pumps. *Science*. **310**, 1461–5 (2005).
105. A. Schlessinger, N. Khuri, K. M. Giacomini, A. Sali, Molecular modeling and ligand docking for solute carrier (SLC) transporters. *Curr. Top. Med. Chem.* **13**, 843–56 (2013).

106. K. J. Tanaka, S. Song, K. Mason, H. W. Pinkett, Selective substrate uptake: The role of ATP-binding cassette (ABC) importers in pathogenesis. *Biochim. Biophys. Acta - Biomembr.* **1860**, 868–877 (2018).
107. M. Dean, T. Annilo, Evolution of the Atp-Binding Cassette (Abc) Transporter Superfamily in Vertebrates. *Annu. Rev. Genomics Hum. Genet.* **6**, 123–142 (2005).
108. H. Glavinas, P. Krajcsi, J. Cserepes, B. Sarkadi, The role of ABC transporters in drug resistance, metabolism and toxicity. *Curr. Drug Deliv.* **1**, 27–42 (2004).
109. J. Xiong, J. Feng, D. Yuan, J. Zhou, W. Miao, Tracing the structural evolution of eukaryotic ATP binding cassette transporter superfamily. *Sci. Rep.* **5**, 1–15 (2015).
110. K. P. Locher, Mechanistic diversity in ATP-binding cassette (ABC) transporters. *Nat. Struct. &Amp; Mol. Biol.* **23**, 487 (2016).
111. M. J. Landrum *et al.*, ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, 980–985 (2014).
112. C. F. Higgins, Multiple molecular mechanisms for multidrug resistance transporters. *Nature*. **446**, 749–757 (2007).
113. F. L. Theodoulou, I. D. Kerr, ABC transporter research: going strong 40 years on. *Biochem. Soc. Trans.* **43**, 1033–1040 (2015).
114. 岩田想, 膜蛋白質の結晶構造解析に将来はあるのか. 蛋白質核酸酵素. **50**, 197–206 (2005).
115. J. Kim *et al.*, Subnanometre-resolution electron cryomicroscopy structure of a heterodimeric ABC exporter. *Nature*. **517**, 396–400 (2015).
116. H. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
117. A. Bateman *et al.*, UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
118. E. Kobayashi, K. Yura, Y. Nagai, Distinct Conformation of ATP Molecule in Solution and on Protein. *Biophysics (Oxf)*. **9**, 1–12 (2013).
119. K. Nishikawa, T. Ooi, Y. Isogai, N. Saitô, Tertiary Structure of Proteins. I. Representation and Computation of the Conformations. *J. Phys. Soc. Japan*. **32**, 1331–1337 (1972).

120. R. J. P. Dawson, K. P. Locher, Structure of a bacterial multidrug ABC transporter. *Nature*. **443**, 180–185 (2006).
121. T. Furuta, T. Yamaguchi, H. Kato, M. Sakurai, Analysis of the Structural and Functional Roles of Coupling Helices in the ATP-Binding Cassette Transporter MsbA through Enzyme Assays and Molecular Dynamics Simulations. *Biochemistry*. **53**, 4261–4272 (2014).
122. N. Li *et al.*, Structure of a Pancreatic ATP-Sensitive Potassium Channel. *Cell*. **168**, 101–110.e10 (2017).
123. J. Y. Lee *et al.*, Crystal structure of the human sterol transporter ABCG5/ABCG8. *Nature*. **533**, 561–564 (2016).
124. R. J. Ferreira, C. A. Bonito, M. N. D. S. Cordeiro, M. J. U. Ferreira, D. J. V. A. Dos Santos, Structure-function relationships in ABCG2: Insights from molecular dynamics simulations and molecular docking studies. *Sci. Rep.* **7**, 1–17 (2017).
125. P. M. Jones, A. M. George, Mechanism of ABC transporters: A molecular dynamics simulation of a well characterized nucleotide-binding subunit. *Proc. Natl. Acad. Sci.* **99**, 12639–12644 (2002).
126. P. M. Jones, A. M. George, Nucleotide-dependent allostery within the ABC transporter ATP-binding cassette: A computational study of the MJ0796 dimer. *J. Biol. Chem.* **282**, 22793–22803 (2007).
127. J. M. Damas, A. S. F. Oliveira, A. M. Baptista, C. M. Soares, Structural consequences of ATP hydrolysis on the ABC transporter NBD dimer: Molecular dynamics studies of HlyB. *Protein Sci.* **20**, 1220–1230 (2011).
128. P. M. Jones, A. M. George, Molecular-Dynamics Simulations of the ATP/apo State of a Multidrug ATP-Binding Cassette Transporter Provide a Structural and Mechanistic Basis for the Asymmetric Occluded State. *Biophys. J.* **100**, 3025–3034 (2011).
129. J.-F. St-Pierre, A. Bunker, T. Róg, M. Karttunen, N. Mousseau, Molecular Dynamics Simulations of the Bacterial ABC Transporter SAV1866 in the Closed Form. *J. Phys. Chem. B*. **116**, 2934–2942 (2012).
130. K. Condic-Jurkic, N. Subramanian, A. E. Mark, M. L. O'Mara, The reliability of molecular dynamics simulations of the multidrug transporter P-glycoprotein in a membrane environment. *PLoS One*. **13**, 1–24 (2018).

131. Ł. Gricman, C. Vogel, J. Pleiss, Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining. *Proteins Struct. Funct. Bioinforma.* **83**, 1593–1603 (2015).
132. 今西規, HLA遺伝子の多様性とヒトの進化. 日本組織適合性学会誌. **1**, 130–134 (1994).
133. A. Kawashima, Y. Satta, Substrate-dependent evolution of cytochrome P450: Rapid turnover of the detoxification-type and conservation of the biosynthesis-type. *PLoS One*. **9** (2014), doi:10.1371/journal.pone.0100059.
134. N. Fukuhara, T. Kawabata, HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res.* **36**, 185–189 (2008).
135. T. Kawabata, HOMCOS: an updated server to search and model complex 3D structures. *J. Struct. Funct. Genomics*. **17**, 83–99 (2016).
136. D. F. Vears, K. Sénécal, P. Borry, Reporting practices for variants of uncertain significance from next generation sequencing technologies. *Eur. J. Med. Genet.* **60**, 553–558 (2017).

謝辞

本研究をまとめるにあたりご指導と温かいご支援を賜りました, 小林哲幸教授に深く感謝いたします. 学位論文の審査委員としてご助言を賜わりました, 千葉和義教授, 三宅秀彦教授, 近藤るみ准教授に深く感謝いたします. 様々な学びの場を賜わりました, 学際生命科学東京コンソーシアムの皆様に深く感謝いたします. データの提供元Genome Aggregation Database (gnomAD), ならびにこのデータベースにexomeデータとゲノム変異データを提供してくださいましたグループの皆様（http://gnomad.broadinstitute.org/about）に感謝いたします.

共同研究者の早稲田大学理工学術院先進理工学部生命医科学科鈴木博文博士に御礼を申し上げます. 3次元情報によるタンパク質の機能異常予測についてのご助言ならびにPythonプログラミングの手ほどきをしてくださいました, 国立成育医療研究センター研究所メディカルゲノムセンター 青砥早希博士, 機械学習を学ぶ機会を与えてくださいました, 国立成育医療研究センター研究所システム発生・再生医学研究部組織工学研究室長 岡村浩司博士, 国立成育医療研究センター研究所メディカルゲノムセンター 瓜生英尚博士に御礼を申し上げます. ヒトバリアントデータベースの使い方や成り立ち, バリアントの分類についてご教示くださいました, 国立成育医療研究センター研究所メディカルゲノムセンター 岸本洋子博士に御礼を申し上げます. 臨床におけるエクソーム解析や人類遺伝学についての知識を得る機会を与えてくださいました, 国立成育医療研究センター研究所メディカルゲノムセンターの皆様, ならびに国立成育医療研究センター研究所周産期病態研究部の皆様に御礼を申し上げます. 研究全般に対するご助言と励まし, Rおよびシェルスクリプトによるデータ解析全般についての知識と技術を得る機会を与えてくださいました, 株式会社日本バイオデータ 緒方法親様, 椎名晃久様, 鈴木彦有様, 小西省吾様, 松田朋子様に御礼を申し上げます. 研究室セミナーや日常のディスカッションなど, 共に研究室生活を過ごしてきた, お茶の水女子大学理学部生物学科生命情報学研究室の皆様に御礼申し上げます.

最後に, 私の家族に感謝します. 皆様が支えてくれたおかげで, 研究を進め, まとめることができました. ほんとうにありがとうございました.