特集

SNPsカタログDBシステム構築と活用

Creation and Use of SNPs Catalog DB System

あらまし

ポストシーケンス時代に突入し，遺伝子多型のなかでSNP（Single nucleotide polymorphism＝1塩基多型）が注目されている。ゲノム上に一定の密度で分布したSNPは，疾患関連遺伝子探索や薬剤応答性分野の研究では非常に有用なマーカである。

公的サイトで運営されるSNP関連データベース（DB）には膨大な数のSNP情報が登録されており，その利用方法について様々な試みが行われている。

本稿では，新たな試みとして複数DBに存在するSNP情報を統合するSNPsカタログDBシステムの構築および活用方法について紹介する。

Abstract

The age of post-sequencing has begun, and single nucleotide polymorphism (SNP) has attracted attention in DNA polymorphism research. SNPs, which are found at a constant rate on the genome, are very useful markers for disease-related gene research and drug-effect studies. Huge volumes of SNP information have already been registered in public SNP databases, and many researchers have approached such data in various ways. This paper describes the creation and use of a new SNP Catalog DB system that integrates the SNP information stored in several databases and makes it available for use.

赤坂英俊（あかさか　ひでとし）

（株）FFC医療ライフサイエンス部 所属

現在，製薬企業，研究所向けシステム開発に従事。

ま え が き

SNP（Single nucleotide polymorphism＝1塩基多型）は特定人口集団において配列比較をした場合，1%以上の頻度で認められる配列の違いである（**図-1**）。SNPはゲノム上に一定の密度で分布した最小単位のマーカとして有用であるため，SNPを用い，疾患関連遺伝子探索・特定，蛋白質発現，薬剤感受性，毒性などの研究において遺伝子多型解析(注1)が数多く行われている。現在，インターネット上で閲覧可能な公的サイトに公開されているSNP関連のデータベース（以下，DB）には数百万単位の膨大な数のSNP情報が登録されており，人種の違いや，アレル(注2)（対立遺伝子）頻度の有無の情報が異なったフォーマットにより登録されている。これらのDBから網羅的にデータを抽出し，整理・統合（カタログ化）を実現するシステムの構築は，単に基礎情報となるだけではなく，遺伝子多型解析研究の効率化，スピードアップにつながる有用なツールと言える。

本稿では，SNPsカタログDBシステムを利用した遺伝子多型解析への取組みについて紹介する。(1)－(7)

SNPsカタログDBシステム構築

遺伝子多型解析研究において，対象となる遺伝子（ゲノム）上のどの位置にSNPが存在しているかを知ることが第一に必要となる。再シーケンシングによるSNP探索においても公的サイトにエントリされているSNP位置を目安とし，遺伝子コード領域周辺に絞り込んだシーケンス(注3)を行うのが効率的かつ経済的と言える。

各SNP関連DBにはSNPの位置に関する情報，アレルやアレル頻度，人種といった属性情報が登録されており，それぞれのDBによりデータ収集の方法や条件が異なる。

これらのSNP情報を各DBにまたがり整理・統合を行うことで網羅的なSNP情報を得ることが可能となる。さらにexon（翻訳領域），intorn（非翻訳領域）をSNP情報に重ねることでSNPによる表現型への影響を示すことも可能となる。

SNPsカタログDBシステム構築において利用する主要なDBを**表-1**に示す。

SNP関連DBの中にはそれぞれ，1対1の相互関係を持っている場合がある。そのリンク関係からSNP位置を整理・統合することも可能であるが，二つ以上の複数SNP関連DBをすべて整理・統合するためには，共通の基準が必要となる。

そのため，NCBIのHuman Genome Sequencingに基づくContig配列にそれぞれのSNP位置をマークすることにより統合可能なDBを構築し，Contig配列でのSNP位置を決定する方式とした。

ContigDBは精度的な問題を除けば配列の大きさ，情報量，更新頻度が適当であること（最新性）など利点が多く，共通プラットフォームとして適切であると判断した。

カタログ化の単位は遺伝子で管理することを前提とし，Contig配列上に各SNP関連DB由来のSNPをマッピング(注4)した後，独自のSNP統合番号を付加し，各SNP関連DBとのリンク管理可能なDBを構築した。

SNPsカタログDBシステムの機能は，大きく以下の

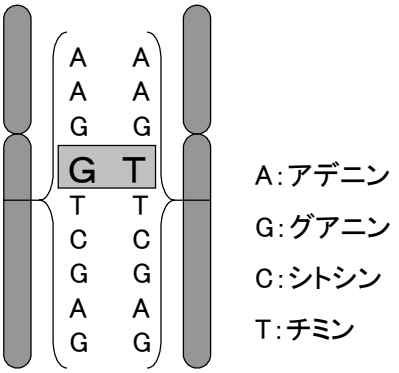

図-1　SNPの例
Fig.1-SNP example.

表-1　SNPsカタログDB作成で利用する主要な公共DB

（1）SNP関連DB

DB名	データ提供機関	備考
dbSNP	NCBI（NIH）	広義のSNP収集
HGVBase	Karolinska Inste（EBI）	人種，アレル頻度含む
TSC	The SNP Consortium	人種別情報含む

（2）遺伝子関連DB

DB名	データ提供機関	備考
LocusLink	NCBI（NIH）	GeneSymbol
Contig	NCBI（NIH）	配列情報，Feature情報
GenBank	NCBI（NIH）	配列情報
RefSeq	NCBI（NIH）	遺伝子情報
DBTSS	東大医科研	ヒト完全長cDNA転写開始

（注1）遺伝情報の違い（SNP）から遺伝子多型と表現型の関係を解析する。
（注2）染色体の特定位置（遺伝子座）を占める遺伝子が複数存在する場合の各遺伝子の頻度。
（注3）DNA塩基配列やアミノ酸塩基配列を決定すること。
（注4）遺伝子上でのSNPの位置情報を決定すること。

四つに分けられる（**図**-2）。

（1） データ収集機能

公開されているFTPサイトからSNP情報および遺伝子情報を自動取得し，それぞれのDBごとに展開する。

（2） データ抽出・加工・格納機能

データ収集機能により獲得したSNP情報および遺伝子情報からカタログ化で必要となるシーケンスアクセッション番号情報，アレル情報を抽出・加工後DBへ格納する（配列情報整理のためのIndex作成処理を含む）。

（3） カタログ機能

格納されたデータを用い，遺伝子に整理・統合されたSNPsカタログを作成し，統合SNP情報をDBに格納する。

（4） 統合SNPマップ表示機能

SNPsカタログDBに登録された遺伝子をブラウザにより検索・表示する仕組みを構築する。

カタログ機能および統合SNPマップ表示機能の詳細を以下に示す。

SNPsカタログ化の実施

SNPsカタログ化は，遺伝子単位にSNP位置情報を網羅的に集め，共通プラットフォームであるContig配列に合わせるアラインメント（Blast）を実施し，統合SNP位置の決定およびアノテーション情報を付加する方法で行った。

カタログ対象となる遺伝子は，LocusLinkにエントリされかつContigにもエントリされている遺伝子とし，

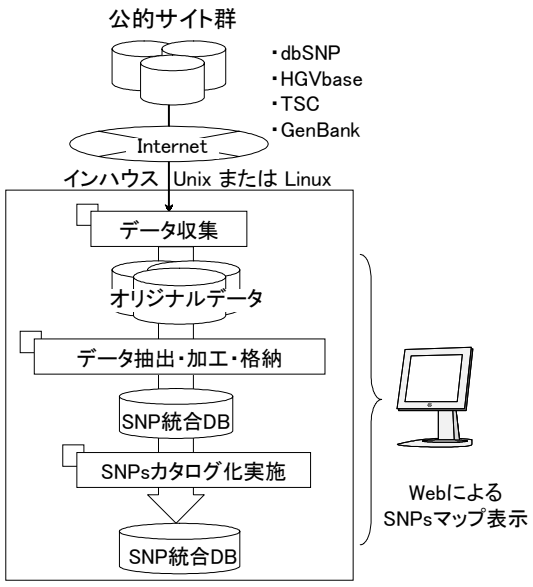

図-2 SNPsカタログDBシステムの概要
Fig.2-Outline of SNPs catalog DB system.

2002年6月時点で約13,000遺伝子に対してカタログ化を実施した。

大まかなカタログ化の実施手順は，以下のとおりである。

（1） 遺伝子基本情報獲得

LocusLinkからTarget遺伝子のSymbol名，染色体番号を獲得する。

（2） 遺伝子Feature情報獲得

ContigからSymbol名で定義されるTarget遺伝子領域，mRNA領域，CDS領域情報を獲得する。

（3） SNP情報獲得

各SNP関連DB内のTarget遺伝子領域内に存在するSNPを抽出し配列と位置情報を獲得する。

（4） Blastによるアラインメント実施

（3）で抽出した配列でBlast用DBを作成する。クエリとしてContig配列のTarget遺伝子領域±5000（プロモータ領域と仮定）の領域を抽出しBlast(注5)を実行する。

（5） SNP統合番号採番

Blast結果から一致したSNP位置をマーキングし，統合SNP位置をContig配列上で決定しDBに格納する。

（6） SNPの分類とアミノ酸変異情報付加

各統合SNP番号で管理されるSNPの分類（sSNP，cSNP，iSNP，gSNP）を実施する。SNPが翻訳領域にある場合，アミノ酸変異があるかを統合SNP番号で実施する。

網羅的にSNPを見つけるために，各SNP関連DBの特徴にあわせ，SNP抽出の方法を変えることで対応した。Blastによるアラインメント結果は，精度の高いSNP位置情報を得ることができるとともに，SNP探索を行う上でのfragment決定時に有用な情報である。

SNPsマップ表示の特徴

SNPsカタログDBシステムのマップ表示機能はSNPの位置情報，属性情報を統合的に表示するツールである。大きな特徴として，以下の3点が挙げられる。

（1） 統合SNP情報表示

画面上部のマップ画面では，遺伝子Feature情報および統合SNPがマークされる。マークの種類によりどのSNP関連DBが由来であるか一目で分かる（**図**-3）。

mRNA領域とCDS領域を同時にマップ表示しているためSNPの種類を視覚的に表現することが可能となっ

（注5） NCBIで開発された相同性（ホモロジー）検索ツール。

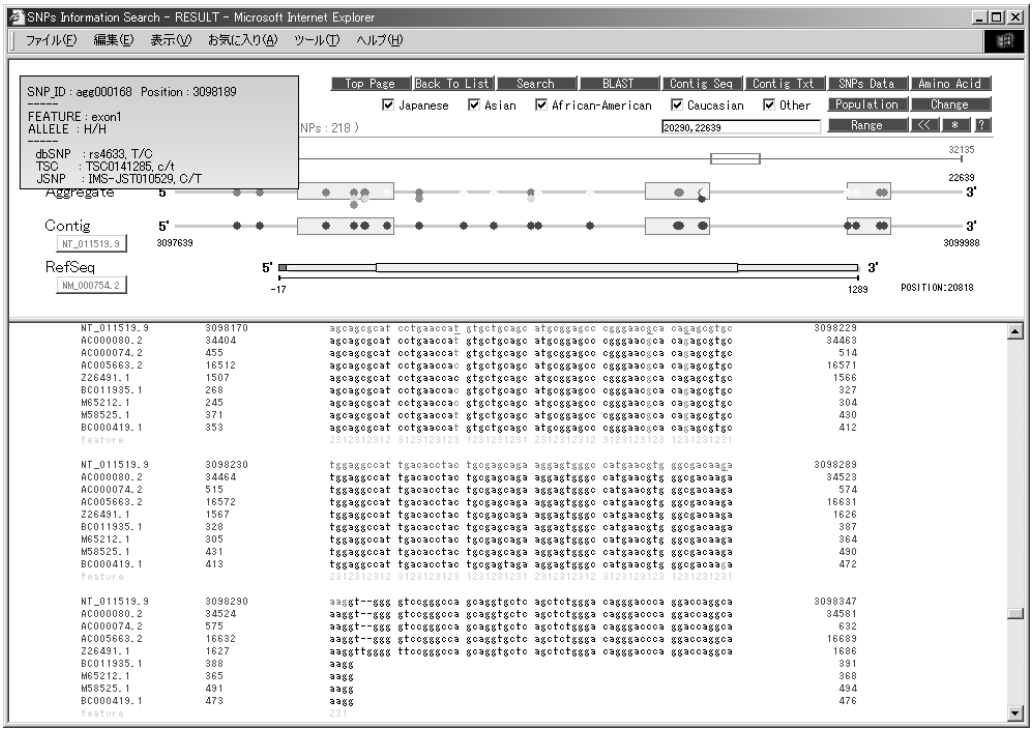

図-3　SNPsマップビューアの例（下段はBlast結果）
Fig.3-Gene Diagram with SNPs (Lower: Blast Result).

ている。

また，SNPマーク位置にマウスポイントを重ねると画面左上に小ウィンドウを表示し，Feature，アレル，SNPIDの参照を可能とする。

画面下部は，Blast結果表示，BlastからContig配列を抜き出したContig配列表示およびSNPテーブルを表示する。Blast結果表示ではSNP位置が認められる塩基を強調表示し，統合SNP番号と対応した表示を可能とする。

SNPテーブルでは，統合されたSNP関連DBを項目配置し，各DBとの関連を見ることができる。

(2)　人種・DBの表示切替え

人種によるSNPの違いを表現するため，SNPマークを，人種を選択し表示する機能を設けた。各SNP関連DBでは正確に人種ごとにSNPの違いが整理され採取，管理されたものがないのが現状である。

今回のカタログ化ではTSCのアレル頻度プロジェクトで実施された約78,000 SNPに対して登録された人種情報に基づき人種の表示切替えを可能とした。

今後，人種情報やアレル頻度が更に公開されることを考慮し，取込みが容易な仕組み，表現方法を構築した。

(3)　アミノ酸変異表示

CDS領域内にSNPが存在した場合は，アレルの違いによりアミノ酸へどのような影響を及ぼすか，変異の有無を統合SNP番号で置換したアミノ酸配列を作成し表示を試みた。

SNPsカタログDBシステムの応用

SNPsカタログを利用した応用方法として，独自にSNP探索を行う場合を例に説明する。

まず，SNPsカタログDBシステムの表示機能を用いて，Targetとなる遺伝子領域周辺で実際にSNPタイピング(注6)を行うFragmentを決定する。複数のサンプルから決定したFragmentに対してシーケンス処理を行う。

波形データからBaseCallingを行った後，複数サンプルから成るFragmentごとの配列データとして出力可能となる。

出力された配列データをもとに公的サイトと同様の情報を持つ独自シーケンスSNP情報を以下の手順で作成する。

(1)　Fragment標準配列化

Fragment単位で複数サンプルから一つの標準配列を決定する。

(2)　SNP位置情報

アラインメントを行い，Fragment標準配列上のSNP位置情報を作成する。

(注6)　SNPが存在する位置の塩基を判別する方法。

(3)　アレル・アレル頻度情報

SNPのアレル情報を作成・アレル頻度を算出する。

これらのデータを加工し，SNPsカタログDBシステムへ取り込む。

以降の工程は公的DBを用いたSNPsカタログ化と同様の手順となり，公的サイトにあるSNPs関連DBと新たにSNP探索でタイピングし発見したSNPを統合した，独自のSNPsカタログを作成することが可能となる。

SNPマップ上では，新たに発見されたSNPなのか既に公的サイトで報告されているSNPであるのか，またアレルの違いや，人種による違いはあるのかといった比較が容易に可能となる。

最新データ更新の対応

SNPsカタログDB構築で利用される公的サイトデータは定期的に更新される。そのため，SNPsカタログDBシステムにおいても定期的にDBを更新し，最新のデータをもとにSNPsカタログDBを再構築する必要がある。

SNPsカタログDBシステムでは，更新によりデータが置き換わることに対応すべく，公的データ，カタログデータの世代管理を行い過去のデータを保持したまま，最新データを更新・格納する方式とした。

また，新たにSNP関連DBやそのほかのDB取込みに対応するため，データ更新機能を各DB単位に更新可能な仕組みを構築し，データ追加への柔軟な対応を可能としている。

今後の方向と課題

遺伝子多型解析の一つの方向として，連鎖不平衡，ハプロタイプ解析が盛んに行われており，いくつかの結果が示されつつある。これらは単独のSNPに注目するのではなく，いくつかのSNPの組合せにより表現型との関連を説明しようとする解析手法である。

この解析では，Case/Control群それぞれのSNPタイピングを行い比較する方法がとられる。この方法で比較を行うためには，Case/Control群のサンプルデータ管理，SNP位置情報管理，解析用データ加工，解析結果加工・格納といった一連の機能が必要となる。これらに対応すべくSNPsカタログDBシステムをトータルシステムの中核として，データサンプル管理，解析用データ加工，解析結果の格納，結果表示機能をシームレスに構築，運用していく仕組み作りを進めている。また，ハプロタイプ解析は遺伝子領域周辺からゲノムワイド（染色体全体）での解析ニーズがあることや，ハプロタイプ同士での関連解析を考慮する必要もある。

上記解析では，疾患関遺伝子探索，薬剤感受性の分野で広く解析が行われているが，今後，遺伝子検査・診断，毒性試験の分野で研究が進んでいくものと考えられ，臨床の現場や非臨床分野での対応を視野に入れシステム整備を進めていきたい。

む　す　び

SNPsカタログ化で利用したゲノム関連のDBは，ニーズに応じ提供内容の追加・変更が頻繁に行われている。

提供されるデータ種，データ量も年々増加の一途をたどり，利用する側でデータの性質や精度を見極めたうえで選択することが迫られる。一方で提供されるデータの最新性を要求される現場も多く，最新性を保った上で精度の高いデータを利用できる（選択できる）環境を構築，維持していくことが重要になると言える。今回構築したSNPsカタログDBシステムでは，上記要望に対応していくための一つの解決方法であると考えられる。

ポストシーケンスの時代に入り，バイオインフォマティクスの重要性は今後更に大きくなると考えられる。今後も顧客の現場の声に耳を傾け，的確なソリューションを提案し，顧客に満足いただける製品，サービスを提供していきたい。

参 考 文 献

(1)　http://www.ncbi.nih.gov/SNP/
(2)　http://elmo.ims.u-tokyo.ac.jp/dbtss/
(3)　http://hgvbase.cgb.ki.se/
(4)　http://snp.cshl.org/
(5)　松原謙一ほか：SNP遺伝子多型の戦略．2000，中山書店.
(6)　鎌谷直之：ポストゲノム時代の遺伝統計学．2001，羊土社.
(7)　（財）ユーマンサイエンス財団調査報告書：ゲノム医療・創薬におけるインフォマティクスの動向．2002.4.