

ヒト完全長 cDNA クローンの単離とそのバンク化及びそれらを用いた 多型マーカー開発並びに発現情報解析のための方法の確立

I 試験研究の全体計画

1. 研究の趣旨

一般に我々が体質や気質として理解している種々の生理的・肉体的あるいは精神的・神経的な性質には遺伝的な影響があることはよく知られている。さらに典型的な遺伝病は当然のことであるが、糖尿病・高血圧などの生活習慣病、がん、アレルギーなどの多因子性疾患、あるいは種々の感染症を含むすべての疾患の発症や増悪には、遺伝的素因、すなわち、遺伝子産物の質や量の差が様々な程度で関与する。ヒトゲノム解析研究の最も重要な目的は、このような遺伝的要因を解明して、疾患の予防・診断に役立てる、あるいは画期的な治療の開発につなげる点にある。遺伝子の発現量の差（遺伝子発現プロフィール）や遺伝子産物の機能的性質の差（翻訳領域における遺伝子多型や遺伝子のスプライシングの差）を体系的に解析し、これをもとに疾患、あるいは、特異な病態との関連を調べて遺伝子機能を類推していく研究は、わが国の将来の医療を見据えた観点からきわめて重要であることは明白である。21世紀は情報と医療が重要な産業の柱となると予測されているが、医学・生物学研究の基盤整備の遅れは、国益の点からみてもきわめて危機的な状況を招くことは必至であり、国家財産の確保という観点からもここに提案する研究計画の強力な推進が必要である。

本研究は、ヒト遺伝子機能研究、疾患関連遺伝子研究を効率的に進めていく観点から計画されたものであり、(1)研究の基盤材料となる完全長 cDNA クローンバンクの構築、(2)遺伝子発現プロフィールを効率よく検索する方法の確立、(3)疾患遺伝子研究のための基盤情報となる特に遺伝子発現に影響を与える遺伝子多型情報の収集とデータベース化、(4)得られた完全長 cDNA 配列、それらの発現プロフィール、SNP 配列とその多型頻度を統合したデータベースを構築し、インターネットを通じて公表する、の4個の柱から構成される。(1)の完全長 cDNA バンクの構築では、オリゴキャップ法を利用して、多数の臓器と細胞株から抽出した mRNA を材料に効率よく完全長 cDNA を取得する。これらの cDNA クローンを大量にシーケンスすることによってクローンを選別化し、5年間でヒト全遺伝子に関する完全長 cDNA の単離とバンク化を行うことを目的とする。このバンクの構築は、遺伝子の機能解析を速やかに進めていく観点から非常に重要であり、東京大学医科学研究所のヒトゲノム解析センターを通して国内外の研究者に提供する。(2)の遺伝子発現プロフィールを効率よく検索す

る方法の確立は、さまざまな病態や環境の変化に応じて発現量の変化する遺伝子を効率的に解析し遺伝子機能解明に応用するために重要な基盤技術であり、数万種類の cDNA クローンからなる DNA チップ等のマイクロアレー技術を確立する。(3)の遺伝子多型解析は、遺伝子産物の質・量の差と疾患遺伝子との関連を効率よく検索していくために必要不可欠な基盤情報を提供するものであり、疾患遺伝子の解明やそれに基づく画期的治療法や診断法の開発につながることを期待される。多型を見つけるために、SSCP法によるスクリーニングや DNA シーケンス法を利用し、5年間に 4×10^4 個のヒト全遺伝子のそれぞれの、特に遺伝子発現制御領域に対応する複数個の高頻度多型 SNP を同定する。また、これら SNP の日本人を含む複数人種集団内及び集団間での頻度の分散を決定し、これをデータベース化する。(4)の cDNA 及び SNP に関するデータベースの構築は、大量の情報に関するデータの効率的処理、各データ間の統合化に関する情報学的基盤の開発を含み、これによって得られる知識発見支援システム、すなわち有用な医学的・生物学的知識、例えばヒト多型と疾患との関係等の発見を支援する計算技術を確立することを目標とする。

2. 研究内容及び目標

1. 完全長 cDNA ライブラリーの作製とそのバンク化に関する研究（東京大学医科学研究所癌ウイルス研究部）

5年間で約4万種類の遺伝子（ヒト全遺伝子）に関する完全長 cDNA クローンの単離とその配列決定を最終目標とする。

(1) 上記の目標を達成するために、多数のヒト組織（皮下脂肪・内臓脂肪・子宮・骨格筋・大腸・胃・肝臓・膵臓・脾臓・肺・脳の各領域）および、多数の培養細胞（大動脈・冠動脈・胎盤等の血管内皮細胞や種々のがん細胞株）から mRNA を抽出し、これらを材料に、オリゴキャップ法を用いた完全長 cDNA ライブラリーの作製を行う（年間10種類程度の組織もしくは細胞株からのライブラリーの作製と解析を目標とする）。

(2) 各完全長 cDNA ライブラリーより 2×10^4 個のクローンをマイクロタイタープレートに拾い上げ、これらのクローンより発現量の多い遺伝子を、高密度フィルターを利用したハイブリダイゼーションで除去する。それぞれのライブラリー中で発現量の低い遺伝子について各ライブラリー当たり $1 \times 10^4 \sim 1.5 \times 10^4$ 個のクローンについて部分配列の決定を行う。年間 1×10^5 個クローン（平成10年度） $\sim 2 \times 10^5$ 個クローン（平成11年度以降）の塩基配列を決定す

る。

(3) これら cDNA 部分配列から得られた膨大な情報を解析して、完全長 cDNA クローンの分離同定とバンク化を行う。

2. マイクロアレー解析技術の開発及びそれを応用した病態と遺伝子発現変化との関連性の研究（武田薬品工業(株)開拓第2研究所）

マイクロアレー技術についてマイクロアレーの作製法・定量性・感度などを検討し、定量的かつ高感度の検出法や微量な検体でも検出可能な方法を開発する。これらを見きわめた上で、ヒト主要組織における発現パターンを解析し、データベースを作成する。中間目標としては、まず、技術の確立を目指し、1万種類の cDNA に関する発現パターンデータベースを作成する。これと並行して、マイクロアレー技術を用いて、さまざまな病態の変化や薬剤・物理的刺激により発現が変化する遺伝子の単離を目指す。最終的には、cDNA シークエンス解析チームによって単離された7万種類の cDNA クローンをマイクロアレー技術により解析するシステムの確立を目標とする。

3. 一塩基多型の体系的解析・収集のための研究（肺癌研究会癌化学療法センター、九州大学生体防御医学研究所遺伝情報実験センター）

ヘテロ体出現頻度の高い、すなわち、連鎖解析等に有効な一塩基多型（single nucleotide polymorphism, 以下 SNP）を見つけるために、SSCP 法に基づく多型解析を利用して体系的な解析を行う。また、個々の試料に対する効率的な遺伝子型決定法を確立する。

(1) 集団中のヘテロ体出現頻度を効率よく推定するためには、多数の PCR 産物を SSCP 分析する必要がある。そこで、初年度は PLACE-SSCP 分析の多重化と、キャピラリー電気泳動の多チャンネル化をはかり、解析の大容量化を確立する。

(2) 多チャンネルキャピラリー電気泳動装置による大規模多重 PLACE-SSCP 法を確立させるとともに、これを用いて現在マップされている全遺伝子に対応する STS を対象とする日本人および他の人種集団における多型頻度検索を行う。これには、まず 50-100 人のゲノム DNA を混合して特定の STS について SSCP 分析を行い、多数のゲノム中でのアレルを同時に検出する。これにより、最多アレルの存在比が低い STS をヘテロ体出現頻度の高いものとして選別する。次いで、複数のゲノム中での塩基配列を決定することによりヘテロ体が高頻度に出現する SNP を同定する。

(3) 上記方法を用いてさらに多数の STS について高頻度多型 SNP を検索する。これには蓄積されつつあるゲノム DNA 配列をもとに特に遺伝子発現制御ゲノム領域に PCR プライマーを設定して行う。これにより約 4 万個の遺伝子の個々に対応する複数の高頻度多型 SNP を 10^6 個オーダー

のレベルで獲得する。

(4) 得られた高頻度 SNP を用いて種々の多因子性疾患に関して、患者集団での SNP アレル頻度の偏りを決定し、原因遺伝子の同定と解析を行う。さらに、個々の試料について効率的かつ安価に遺伝子型を決定できる方法を確立する。これには、SNP 遺伝子型決定のための DNA チップのデザイン等を行い、遺伝子型決定法としての有効性を判定する。一方、DNA チップによる決定でその判定精度に問題がある（疑陽性、疑陰性の頻度が高い）場合には、これに代わる多数の SNP を塩基配列レベルで同時に決定する方法の開発を試みる。

4. ヒト多型 DNA データベースの構築及び有用情報発見・抽出アルゴリズムの研究（東京大学医科学研究所附属ヒトゲノム解析センター）

cDNA マイクロアレー等を利用した実験結果から得られる大量のヒト遺伝子発現情報やヒト多型に関するデータを効率よく格納する方式、および、それらのデータと DNA 配列データ等との統合化に関する研究を行うことによって、それらから導き出される知識発見支援システムの開発研究を行う。また、並行して、そのようにして作られたデータベースから有用な医学的・生物学的知識、例えばヒト多型と疾患との関係などの発見を支援する計算技術の研究を行う。

3. 年次計画

1. 完全長 cDNA ライブラリーの作製とそのバンク化に関する研究

年間 10~15 種類ずつの cDNA ライブラリーを作製し、それぞれから 10,000~20,000 クローンについて、計年間 100,000 クローン（平成 10 年度）~200,000 クローン（平成 11 年度以降）の塩基配列を決定し、これら cDNA 部分配列から得られた膨大な情報を解析して、

第 I 期では約 2 万種類（全遺伝子の 5 割）の cDNA バンクの作製を目標とする。

第 II 期ではさらに完全長 cDNA クローンを単離し、全遺伝子に対応した完全長鎖 cDNA を構築し、その全塩基配列を決定するとともに、この情報を用いた遺伝子発現制御ゲノム領域を同定することを目標とする。

2. マイクロアレー解析技術の開発及びそれを応用した病態と遺伝子発現変化との関連性の研究

第 I 期の目標はマイクロアレーの作製法・定量性・感度などを検討し、定量的かつ高感度の検出法や微量な検体でも検出可能な方法を開発し、ヒト主要組織における 1 万種類の cDNA に関する発現パターンデータベースを作成する体制を整えることにある。

第 II 期については、前期に研究が完了したのものとしてこれを継続しない。

3. 一塩基多型の体系的解析・収集のための研究

第 I 期の目標は PLACE-SSCP 分析の多重化・キャピラ

リー電気泳動の多チャンネル化・解析の大容量化を推進し、多数の全STSを対象とする主に日本人集団における多型検索による大量のSNPの獲得のための技術基盤を確立することにある。

第Ⅱ期の目標は上記技術をさらに大規模化し、決定されたゲノムDNA配列を参考にして、10⁵個オーダーの特に遺伝子発現制御ゲノム領域に存在する高頻度多型SNPを収集することにある。また高頻度SNPを個々の試料について効率的かつ安価に遺伝子型を決定できる方法をも確立する。

4. ヒト多型DNAデータベースの構築及び有用情報発見・抽出アルゴリズムの研究

第Ⅰ期の目標は、ヒト多型DNAデータベースの構築を主な目標とする。

この多型データベースはゲノムの配列データのうち人種や個人によって異なる部分のデータを網羅的に納めたものである。(本プロジェクトでは日本人に特有な多型データベースの構築を目指している。)具体的には、それらの情報を実験装置からでてくるデータから効率よく作成する方法を開発したり、それらの情報を有効にかつ効率よく活用

するには、データベースにどのような形で表現し、どのようなフォーマットで格納すべきかなどの研究を行う。なお、これらの多型情報や発現情報は、ゲノム配列をはじめとする他の生物学医学情報と関連づけてはじめて意味を持つ。そこで、多様なデータベースとの統合化の研究を並行して行う。また、関連づけるべき情報をインターネットから自動的に収集してくるエージェントシステムについても併せて開発する。

第Ⅱ期についてはcDNAマイクロアレー等を利用した実験結果から得られる大量のヒト遺伝子発現情報を効率よく格納する方式、および、それらのデータとDNA配列データ等との統合化に関する研究を行うことによって、それらから導き出される知識発見支援システムの開発研究を行う。具体的には、膨大な発現情報データベースに格納されている発現パターンを分類したり、それをもとにデータベースから特徴的な発現パターンを抽出したりするような計算機ソフトウェアを開発する。また、多型データと薬剤の効果との関連や遺伝子発現パターンと疾患との関係などの医学生物学的知識情報をデータベースから自動的に発見するためのシステムを開発する。

研究項目	10年度	11年度	12年度	13年度	14年度
1. 完全長cDNAライブラリーの作製とそのバンク化に関する研究	← cDNAライブラリーの作製 →				
			← cDNAクローンの大量シーケンス →		
			← シーケンスデータの解析 →		
				← バンク化 →	
2. マイクロアレー解析技術の開発及びそれを応用した病態と遺伝子発現変化の関連性の研究		← 技術の確立 →			
			← 発現情報の収集と応用 →		
3. 一塩基多型の体系的解析・収集のための研究	← PLACE-SSCP法の確立 →				
			← STS多型のスクリーニング →		
4. ヒト多型DNAデータベースの構築及び有用情報発見・抽出アルゴリズムの研究			← 高速遺伝子型決定法の確立 →		
			← ヒト多型DNAデータベースの構築 →		
			← 有用情報発見・抽出アルゴリズム →		
所要経費(合計)	255百万円	324百万円	299百万円	267百万円	

4. 平成13年度における実施内容と達成目標

1. 完全長 cDNA ライブラリーの作製とそのバンク化に関する研究

前年に引き続き、多数のヒト正常組織より mRNA を抽出し、これらを材料に、オリゴキャッピング法を利用した完全長 cDNA ライブラリーを作成する。これらの完全長 cDNA ライブラリーより、各クローンの 5'-末端塩基配列決定を行い、独立した cDNA クローンを選別する。さらにそれらの全塩基配列を決定する。また得られた配列情報をゲノム配列と照合することにより、各遺伝子の発現制御ゲノム領域を決定する。

2. マイクロアレー解析技術の開発及びそれを応用した病態と遺伝子発現変化との関連性の研究

第Ⅰ期における成果をもって本研究における役割を終了したもとして、第Ⅱ期にはこれを継続しない。

3. 一塩基多型の体系的解析・収集のための研究

前年度の多チャンネルキャピラリー電気泳動装置導入と、これを用いた PLACE-SSCP 法の実験系が順調に確立されつつあるので、この多チャンネル PLACE-SSCP 法を用いて、ヘテロ体出現頻度が高い STS を大規模に収集し、SNP 化する作業を拡大する。これには、前期に開発を開始した SNP データ統合システム、すなわち PLACE-SSCP 実験

とシーケンシング実験から得られるデータを統合し、SNP を判定してこれを実験データとともにデータベース化するシステムをさらに改良して完全なものとし、これによって大規模高頻度 SNP 検索、特に遺伝子発現制御ゲノム領域における SNP の検索をさらに拡張して実施する。また、得られた各種ヒト集団での遺伝子発現制御 SNP 頻度情報とそれにいたる実験データ、解析過程を含む本研究での結果を可視化した、透明性の高い公開用データベースを構築する。さらに本研究の各過程で開発された種々の解析用ソフトウェアを公開し、広く世界で利用可能な体制を整える。

4. ヒト多型 DNA データベースの構築及び有用情報発見・抽出アルゴリズムの研究

(1) 全長鎖 cDNA クローンバンクのシーケンシングおよびアノテーションを支援する計算機システムをさらに充実し、これによりシステムの評価改良を図るとともに、実際の研究現場で使用する。

(2) ヒト遺伝子発現情報データベースを構築する。

(3) 発現データベースと多型データベースを統合した上位データベースを作成し、多因子性遺伝形質の遺伝的背景を解析するシステムを開発する。

II 平成13年度における実施体制

研究項目	担当機関	研究担当者
1. 完全長 cDNA ライブラリーの作製とそのバンク化に関する研究	東京大学医科学研究所癌ウイルス研究部	菅野 純夫
2. 一塩基多型の体系的解析・収集のための研究	(財)癌研究会癌化学療法センター、九州大学生体防御医学研究所遺伝情報実験センター	林 健志
3. ヒト多型 DNA データベースの構築及び有用情報発見・抽出アルゴリズムの研究	東京大学医科学研究所ヒトゲノム解析センター	高木 利久
4. 研究管理	(財)癌研究会	向井 孝始

Ⅲ 研究推進委員会

委 員	所 属
[プロジェクト内委員] ○林 健 志 菅 野 純 夫 高 木 利 久	(財)癌研究会 癌化学療法センター研究員, 九州大学生体防御医学研究所遺伝情報実験センター教授 東京大学 医科学研究所癌ウイルス研究部助教授 東京大学 医科学研究所ヒトゲノム解析センター教授
[プロジェクト外委員] 勝 木 元 也 鶴 尾 隆 寺 田 雅 昭	東京大学 医科学研究所ヒト疾患モデル研究センター長 東京大学 分子細胞生物学研究所分子生物活性研究分野教授 厚生労働省 国立がんセンター研究所長

(注：○は研究推進委員長)

Ⅳ 研究連絡会議

委 員	所 属
菅 野 純 夫 高 木 利 久 林 健 志	東京大学 医科学研究所癌ウイルス研究部助教授 東京大学 医科学研究所ヒトゲノム解析センター教授 (財)癌研究会 癌化学療法センター研究員, 九州大学生体防御医学研究所遺伝情報実験センター教授