

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方法の確立

研究代表者：林 健志（九州大学生体防御医学研究所）

I 試験研究の全体計画

1. 研究の趣旨

体質や気質として理解されている種々の生理的・肉体的或いは精神的な性質には遺伝的な影響がある。さらに典型的な遺伝病以外にも、糖尿病・高血圧、アレルギーなどの生活習慣病、がん、あるいは種々の感染症を含むすべての疾患の発症やその感受性、増悪には、多数の遺伝子が関与した遺伝的素因がある、すなわち、これらは多因子性疾患であると理解されている。これらの疾患、体質には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、このような多因子性の量的遺伝的要因を解明して、疾患の予防・診断に役立て、さらに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と個々のクローンの配列情報の網羅的収集、(2)疾患遺伝子研究のための基盤情報となる遺伝子発現に影響を与える遺伝子多型、特に SNP 情報の収集とデータベース化、(3)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の構築とそのインターネット上での公表の3本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 の単離とバンク化及びそれらの配列決定を行う。また得られた情報をもとに遺伝子発現調節ゲノム領域を決定する。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)の遺伝子多型解析は、ゲノム全体、特に遺伝子領域とその発現調節領域に存在する配列の個人差（多型）を日本人を含む人集団を対象として検索し、これらのアレル頻度情報とともにデータベース化する。これにより、遺伝子産物の質的あるいは量的差異に関与する多型と疾患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。この研究によって疾患遺伝子の解明やそれに基づく画期的な治療法や診断法の開発につながる事が期待される。また多型を効率よく解析するために、SSCP 法、DNA シーケンス法を統合的に利用した多型、特に SNP の量的解析技術を確立し、これを広く公開することにより、ゲノム多様性を基礎とした種々の研究に資する解析技術基盤を提供する。(3)の cDNA 及び SNP に関する統合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、多型情報を利用した新たな遺伝学的解析技術の開発を含み、これによって知識発見支援システム、即ち有用な医学的・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。

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・ヒトゲノム解析センター）

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、および得られた情報をもとに、ゲノム上の遺伝子発現調節領域を全ての遺伝子に関して決定することを最終目標とする。

(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。

(2) 完全長 cDNA の各ライブラリーあたり 10^4 個オーダーのクローンをマイクロタイタープレートに拾い上げ、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法で除去する。これによってライブラリー中での出現頻度を平準化し、各クローンについて部分配列の決定を行う。年間 10^5 個オーダーのクローン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、重複配列を除き、独立した完全長 cDNA クローンの分離同定とバンク化、および全配列決定を行う。

(4) 得られた cDNA 5' 配列情報と一致するゲノム領域、すなわち遺伝子転写開始点を同定し、その周辺配列を遺伝子発現調節領域としてデータベース化するとともに、その機能解析を行う。

2. 一塩基多型の体系的解析・収集のための研究（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）を効率よく検出し、かつそれらの集団中でのアレル頻度を正確に決定するために、SSCP 法に基づく多型解析を主軸とした体系的な解析方法を確立する。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) SNP を効率よく検出し、且つそれらの集団中でのアレル頻度を正確かつ効率よく決定するために、多数の PCR 産物を SSCP 分析する技術開発を行う。すなわち 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、これを用いて疾患関連を含む多数の遺伝子に対応する STS を対象とする日本人及び他の人種集団における多型頻度検索を行う。このために多数のゲノム DNA をプールして特定の STS について SSCP 分析を行い、試料採取集団中での SNP アレルを検出すると同時にこれらを正確に定量する。さらに個々の DNA の STS 配列を決定して SNP を確定する解析システムを構築する。

(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 を多数獲得するとともに、それらの量的情報、すなわちアレル頻度を決定し、データベース化して公開する。

(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、多数の SNP を塩基配列レベルで同時に決定する新たな方法を開発する。

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（東京大学医科学研究所・ヒトゲノム解析センター）

上記研究 1 及び 2 の情報学的支援を行うとともに以下の開発研究を行う。

(1) 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。

(2) 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) また関連解析における偽陽性シグナルを減らすことを目的として、隣接する複数個の SNPs に関するハプロタイプ頻度および連鎖不平衡の大きさの推定、および罹患者の集団で健常者に比べて有意に高い頻度で存在するようなハプロタイプの検出を行うプログラムを開発する。

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 バンクを構築し、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。

2. 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、決定されつつあるゲノム DNA 配列を参考にして、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 を大規模収集しデータベース化して公開する。またこの研究遂行によって開発された解析技術を公開し、SNP を用いた遺伝解析の促進を図る。また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の分類と、それをもとにしたデータベースから特徴的な発現パターンを抽出する計算機ソフトウェアを開発する。また、関連解析の理論的基礎と実践における連鎖不平衡領域の決定を行い、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	213百万円

4. 平成14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、全遺伝子の発現制御ゲノム領域を決定しデータベースとしてインターネット上に公開する。

2. 一塩基多型の体系的解析・収集のための研究

多チャンネルキャピラリー電気泳動装置を用いた PLACE-SSCP 法の実験系が順調に確立され、ヘテロ体出現頻度が

高い STS を大規模に収集し、SNP 化する作業が進行しつつあり、また得られたデータのインターネット上での公開も進行しつつあるので、これをさらに拡大して行う。また、確立した解析システムの公開を行う。このために、SNP データ統合解析システム、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、定量が広く行えるようにする。また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とともにこれらをインターネット上に公開する。これには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。

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

全長 cDNA クローンバンクのシーケンシング及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。また、ヒト遺伝子発現情報データベースを構築する。さらに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 14 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・ヒトゲノム解析センター	菅野純夫
2. 一塩基多型の体系的解析・収集のための研究	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	九州大学生体防御医学研究所	小樋孝雄

(注：◎は研究代表者)

III 研究推進委員会

委員	所 属
[プロジェクト内委員]	
◎林健志	九州大学 生体防御医学研究所遺伝情報実験センター長
菅野純夫	東京大学 医科学研究所ヒトゲノム解析センター 助教授
高木利久	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	東京大学 医科学研究所ヒト疾患モデル研究センター長 教授
鶴尾隆	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
寺田雅昭	厚生労働省 国立がんセンター研究所 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	東京大学 医科学研究所ヒトゲノム解析センター 助教授
高木利久	東京大学 医科学研究所ヒトゲノム解析センター 教授
林健志	九州大学 生体防御医学研究所遺伝情報実験センター長