シンポジウム

マウスゲノムエンサイクロペディア

林﨑　良英
（理化学研究所ゲノム科学総合研究センター）

座長　松下　祥（埼玉医科大学免疫学）

それでは，林﨑良英先生をご紹介させていただきます．私は本学免疫学の松下でございます．

林﨑先生は，現在，理化学研究所ゲノム科学総合研究センター遺伝子構造機能研究グループの，プロジェクトチームリーダーでいらっしゃいます．先生は82年に阪大の医学部のご卒業，86年に大学院終了，92年から理科研のライフサイエンス筑波研究センター・ジーンバンク室研究員，ヒトゲノムプロジェクトの推進室，ゲノム機能解析研究グループ・プロジェクトリーダーを経て，98年から現職に就いていらっしゃいます．

先生は賞を各種，受賞なさっておりますが，98年には東京テクノフォーラム賞ゴールドメダル，遺伝子辞書の作成，今日お話しいただくタイトルと非常に近いものですが，2001年にはつくば賞を受賞しておられます．

本日は「マウスゲノムエンサイクロペディア」という演題でお話をいただきます．どんな百科事典ができつつあるのか，私も非常に楽しみに聞きたいと思っております．林﨑先生，よろしくお願いいたします．

ご紹介ありがとうございます．まず，埼玉医科大学のゲノム医学研究センターの開設，おめでとうございます．今日私がお話しするのは，このところずっと私はこのタイトルでお話ししていますが，実は内容が年度ごとに進行して変わっております．

「マウスの遺伝子の辞書」というタイトルで1995年から，日本の中でゲノム科学を再建せよということを，理研の当時の理事長であった小田稔さんに言われまして，何をやったらいいかを考えて，ずっとやってきたのがこの仕事です．

当時，とにかく再建するために，まずそこにおられます村松先生が理化学研究所の顧問になられ，それで今のようなプロジェクトをとにかく実施することが可能になったわけです．

今日はスライドを全部英語で書いてあり，訳しながら話をしますので少し遅くなります．後半はひょっとしたら飛ばすかもしれません．とにかく，どういうものができて，どういう使い方ができるかがわかればよろしいかと思います．

私たちが最初狙ったのは1995年と申し上げました．その95年には，世の中のゲノム科学はどういう方向になっていたかというと，アメリカ合衆国はヒトゲノムのシーケンスを実行することを決意した年であります．

そのほか，1990年代初頭から，製薬会社にサポートされたアメリカのベンチャーによって，EST（Expression Sequence Tag）という，RNAになったその一部分のかけらのcDNAを山のようにシーケンスするというプロジェクトが，すでにもう終わっていました．さて，私たちは何をしたらよいか，もうやるものはないのではないかと思ったのですが，こういうものをやることにしました．Full-length cDNAをやろう．「完全長cDNA」といいます．

それはRNAの端から端までの全長を含むものです．なぜこの完全長cDNAが重要なのかといいますと，1つは，RNAの完全なかたちがわかる．タンパクの全体のかたちがわかる．もう1つは，これは全長を含んでいますので，直接それでタンパク質を発現できます．かけらのDNAをコンピュータ上でいくらつないでも，それは物質としてつながっていませんので，実際にタンパクを作ることはできないわけです．ですから，タンパクを作ることにも重要である．

それから完全長cDNAは，染色体のDNAと比較することによってプロモーター，要するに転写を制御している領域がわかる．もう1つは，ゲノム・プロジェクトというのは，先程，村松先生が言われましたが，ショットガン・シーケンスというやり方があります．そのショットガン・シーケンスをつないでいくのに，このDNA，cDNAそのものが，それをつなぐのに有用な情報を与えます．そういうことで，これをやろうということになりました．

生体内に存在する全部の遺伝子を，まるごとフルレングスのかたちで取ってきたものを集めてバンクを作って，その構造を決定しようというのが私たちの目的であったわけです．

それをやるためには，まず当時何もなかったので，技術から作らなければいけないということで，完全長cDNAを作る．それからハイスピードのシーケンシングをするシステムを作る．それを用いてマウスゲノムエンサイクロペディアを作ることにしました．（図1）

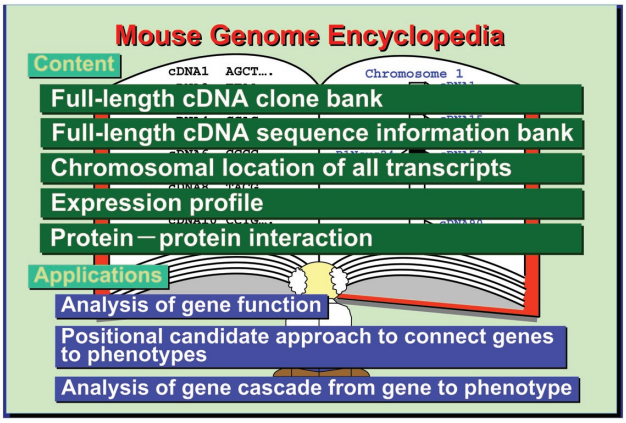
我々のこのマウスゲノムエンサイクロペディアは，全部の遺伝子をかき集めるというのですが，全部の遺伝子の内容というこの辞書には何が載っているか．それは完全長cDNAのクローンのバンクがあります．もう1つは，完全長cDNAのクローンのバンクを全遺伝子について集める．それから，それのシーケンスを全部決定する．

その完全長のcDNAとは，mRNA（転写単位）のことをいいますが，転写単位が染色体のどこにあるかを決定する．もう1つは発現タンパクになる．すなわち，その遺伝子がいつどこでタンパクになっているのか．mRNAになってタンパクになっているのかを見るための，発現プロファイルを明らかにする．

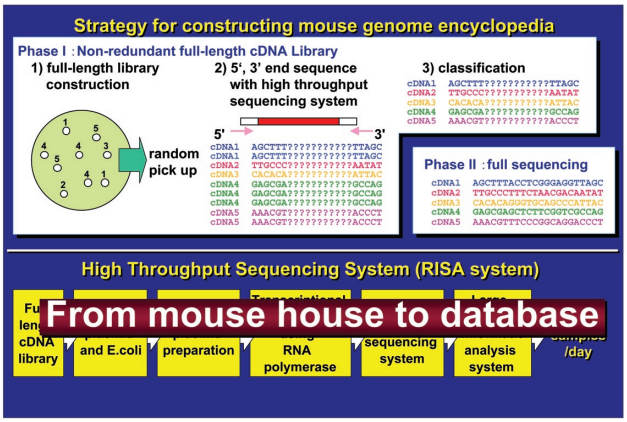
もう1つは，タンパクは何もそれ自身で機能するというのではなく，互いに相互作用しております．あるタンパクがあるタンパクに情報を伝えるためには，タンパクとタンパクの相互作用を明らかにしなければいけない．そういうことで，全遺伝子のどのタンパクとどのタンパクが相互作用するかというのを，ラフでもいいから，とにかく作ろうということにしました．これによって遺伝子の機能がわかります．

あと，疾患の遺伝子を見つけるためのアプローチの方法ですが，染色体の位置情報から見るやり方で，positional candidate approachというやり方があるのです．そのやり方は，フェノタイプ（phenotype：表現型），つまり病気と遺伝子を関連付けするために，非常にいい役割をするであろう．もう1つ，遺伝子がつぶれると，なぜ病気が出るかということのパスウェイが，これをやってわかるであろうということです．

どういう戦略を取ったか．まず，完全長cDNAのテクノロジーを作って，高品質の完全長cDNAを作る技術を作って，プレートの上にクローンをまきます．大腸菌のクローンをランダムにピックアップして，両端からシーケンスを組みます．そうすると同じクローンが2回出てきたり3回出てきたりします．それを除いて重複（redundancy）のないcDNAを作ります．（図2）



（図1）



（図2）

このcDNAの代表選手を選んで全長を決めるというのが，その次のフェーズです．端からこのシーケンスを全部出す．しかも，ほとんど全部の遺伝子をカバーしようと思って，しかも2～3年でそれをカバーするためには，1日4万個のサンプルを処理する操作が必要です．そのために，こんなシステムがなかったの

で，自分たちで作ろうと．今でこそゲノム科学のいろいろな装置がありますが，当時はありませんでしたので，それを自分たちで作ろうということです．

これはマウスを使ってやろうということで，マウスからデータベースを結ぶパイプラインとして，「RIKEN Integrated sequence analyser research sysytem（RISA system）」を作りました．

まず，完全長を得るための完全長cDNAをmRNAから逆転写するためのテクノロジーですが，これをelongation methodということで，伸長反応で伸ばすわけです．こういうテクノロジーを開発しました．（図3）

これはどんなテクノロジーかといいますと，普通，逆転写酵素はステムループ，RNAの二次構造のところで止まります．これは電話線がねじれているような格好です．そういうところで止まるわけです．これがあるから止まるので，これを止まらないようにするためにどうしたらいいか．

温度を上げればいい．分子運動が大きくなっていくのですが，温度を上げると逆転写酵素が失活します．そこで我々が見つけた発見ですが，トレハロース（trehalose）という二糖類を入れると，高温でも逆転写酵素が活性化して，端まで行くことを発見しました．

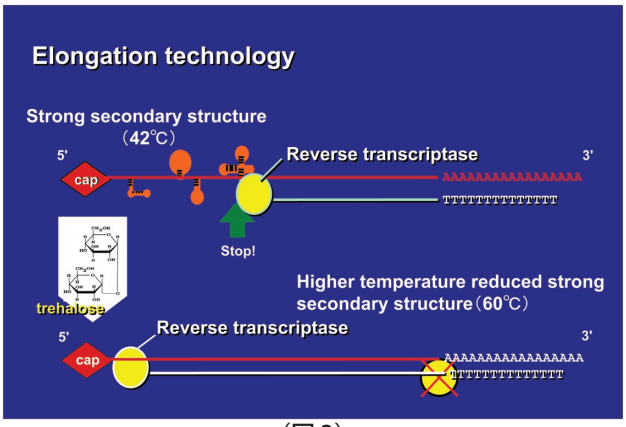
何でそんなことが見つかったかということですが，やけども卵焼きもそうですが，タンパクは熱をかけると変性します．高次構造が崩れるわけです．細胞の中でこういうものが起きますと，ものすごく毒性が強いので，それを元に戻す機能を持つシャペロニン（chaperonine）という物質があります．そのシャペロニンという物質が世の中で知られているのですが，作業仮説として，こういうふうに折りたたみを元に戻すような物質が，反応液の中にある．ひょっとしたら，酵素としますと，酵素を守ってくれるかもしれないというので，そういうものを探したわけです．（図4）

こんな文献がありました．トレハロースはこんな物質ですが，これが酵母菌にヒートショック，酵母菌をやけどさせます．そうするとトレハロースをいっぱい作るようになります．なぜだろうと思ったのです．

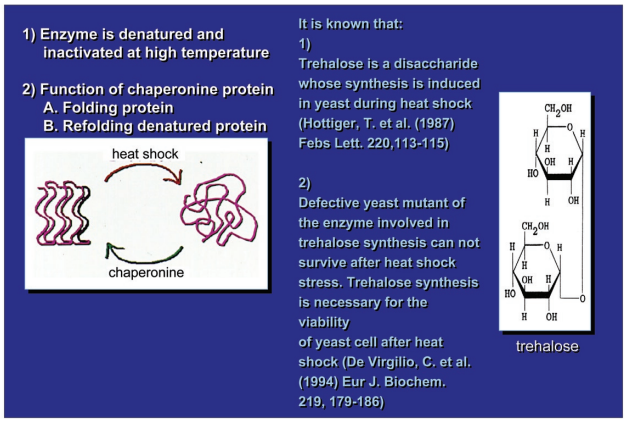
もう1つ，このトレハロースの合成型酵素が欠損したmutantにheat shockをかけると，全然生き返ってこないで，そのまま死んでしまう．heat shockに非常に弱いわけです．

そういうことから，シャペロニンを「物質ではない」と考えて実験しました．実際，5 KbのRNAを鋳型にして，DNAを合成することをしたのですが，正常の反応でやりますと5 Kbのバンドは見えますが，こういうパーシャルなcDNAがいっぱいになります．ところが60℃にしますと酵素が失活しますのでバンドがなくなります．ところがトレハロースを入れると，酵素は失活しないで，ちゃんと5 Kbのバンドが見えるけれども，パーシャルのcDNAがなくなるということで，非常に効果的である．（図5）

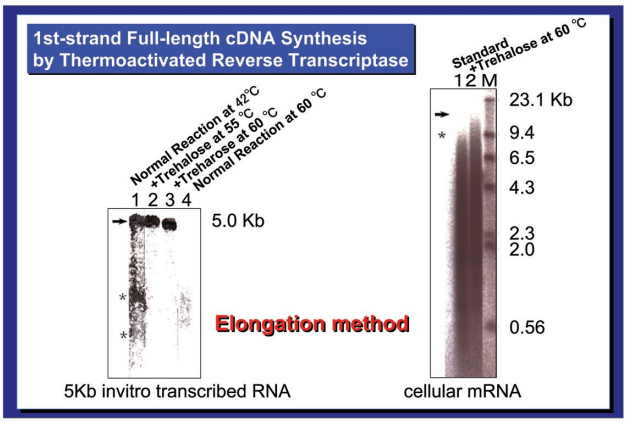
しかも，こういうものを入れますと，60℃でこういう反応をする．トレハロースを入れますと，これはcellular mRNAで，一番長いのは，普通の条件としては9 Kbぐらいですが，16 Kbぐらいまで伸びるようになるということで，非常に効果的であることがわかりました．



（図3）



（図4）



（図5）

これはセレクションする方法ですが，今日は時間の都合上，これを省きます．このような方法をいろいろ作った結果，我々のクオリティが非常に上がってまいりました．

特に最近，cDNAというのは，今までどれだけ頑張っても5 Kbか4 Kbが取れればいいところだったのですが，最近我々の新たなベクターを作りますと，平均差長が9とか8，一番長いのは十数Kbのクローンが取れます．ジストロフィーという大きな遺伝子が一発で取れます．

何でこんなことができるようになったかというと，新しいベクターです．何が新しいかというと，これは基本的にBACという非常に長大な遺伝子を保持するようなものを，cDNAに使いますと，長い遺伝子でもできるということで，こういうものが出来上がっています．この辺は飛ばします．（図 6）

最終的に，我々の完全長cDNAの長さですが，全体の長さでここに9 Kb以上というのがありますが，5’エンドの完全長率は非常によい成績を得ています．少なくともここでの完全長の定義はCap siteまでということではなくて，タンパクの開始codonを含んでいるパーセンテージを上げました．こういうふうに非常に高率に，タンパクの全長をコートするようなcDNAが得られてきたわけです．（図 7）

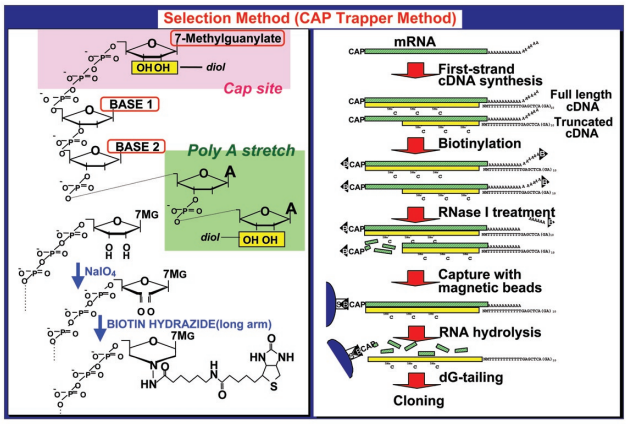
我々のcDNAのサマリーですが，普通に作ると大体2～2.5 Kbぐらいの平均差長があります．長いところを核として，非常に長いところだけ集めてきますと3.5 Kbぐらいです．一番長いベクターを使いますと5～16 Kbぐらいのものが取れてきます．完全長率は，普通，ライブラリを作りますと90～95％が完全長のcDNAで，少なくともタンパクの開始codonであるA・T・Gを含みます．（図 8）

それから，今から申し上げますが，どんどんデータベース，バンクを作っていきますと，もしパーシャルcDNA，不完全長でも，新しい遺伝子なら，バンクの中に入れますから少し妥協するということで，70％ぐらいの効果があります．

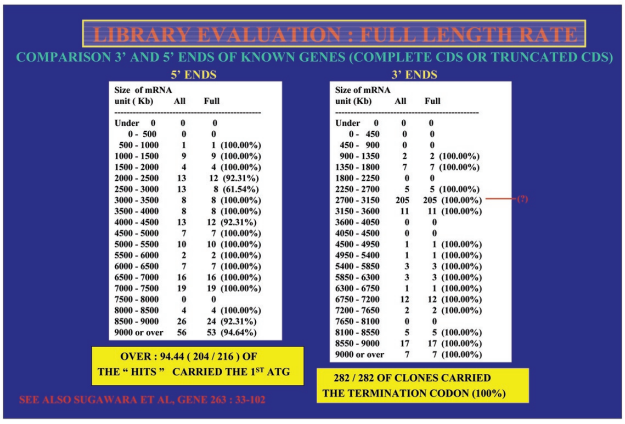
さて，実はそれを実際2～3年でやろうと思ったら，1日4万個のシーケンスを処理しなければいけない．これはクレージーだと言われたのですが，ともかく技術がなかったら作らなければならないということで，必死になって装置類を作ってきました．

これはプラスミド・プレパレーターです．普通，プラスミドを調整するのは手でやります．あれは1日100～200個やれば，翌日しんどくてできないのですが，この機械は全自動で4万個取ってくれます．非常に効率のよい機械です．（図 9, 10）

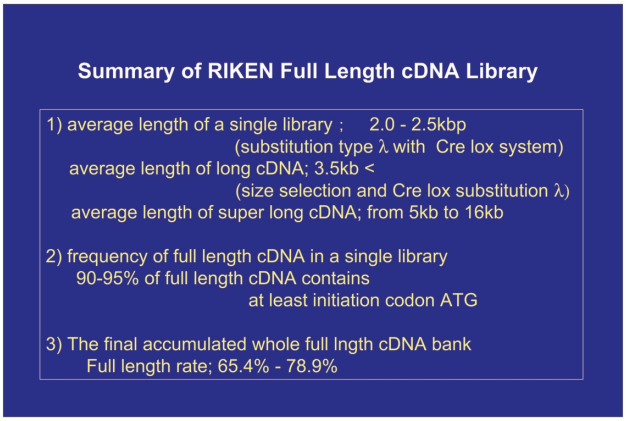
それから，PCRがたくさんできるようなものを作りましたし，最終的にシーケンサーも作りました．このシーケンサーは16×24の384のフォーマットのプレート，穴が384開いています．そこから，こういうガラス細管の中に直接インジェクションして電気泳動し，最終的にシーケンスを決定するような機械も作りました．シーケンサーそのものです．これは新技術事業団の生命活動のプログラム，村松先生が総括されていますが，その資金を得て，こういうものを開発してきたわけです．（図 11）



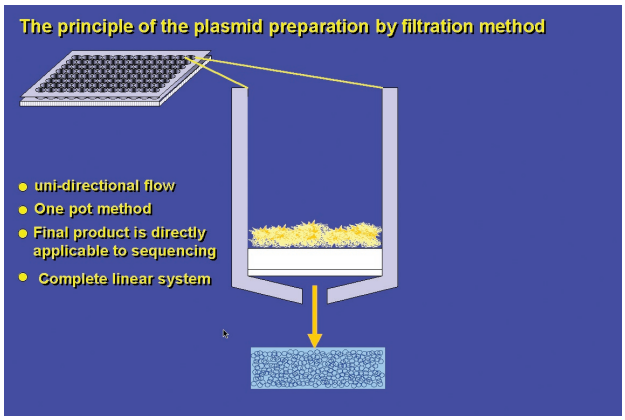
（図 6）



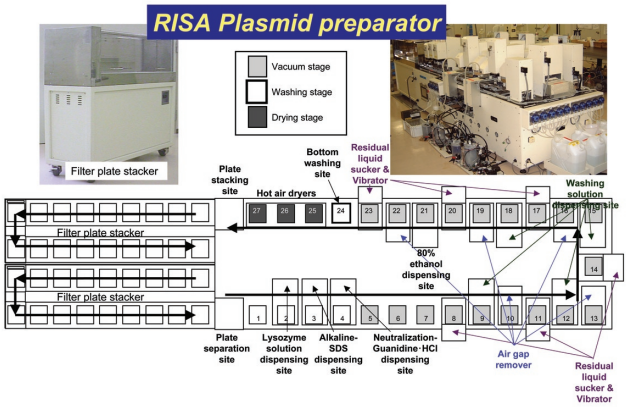
（図 7）



（図 8）



（図 9）



（図 10）

（図 11）

最終的にこういうパターンが出ます．これはおもしろい話ですので，念のために説明します．我々はプラスミドを4万個も取ることをなぜできるかというと，私は元は医者なのですが，こういう機械的なシステム化についてかなり勉強しました．

それは，量産しようと思ったら，液を入れるとき，試薬を入れるのは全部1方向（uni-directional flow）でなくてはいけない．もう1つは，絶対チューブへ移してはいけない．one-pot methodである．最後に得たものが，直接その次の反応に使える．それができると，自動車工場のようにラインの上に乗せていける．こういうものが必要です．

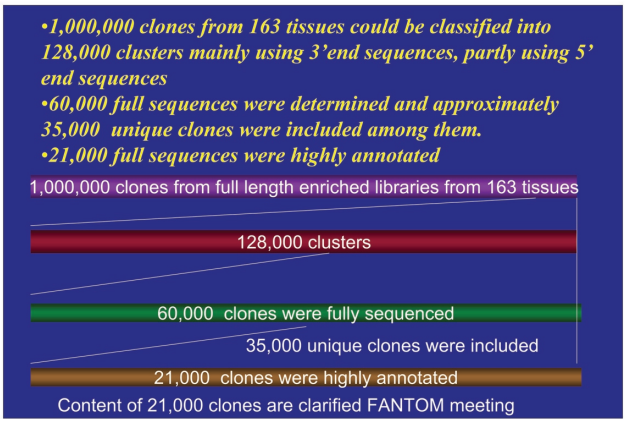
それで，まず大腸菌の培養液をやりますと，ここはガラスのフィルターがあります．membrane filterがありますが，リゾチウムを入れて，アルカリSDSを入れると大腸菌が壊れます．中和・吸着溶液を入れるとこのようになり，吸着するとプラスミドがガラス板にくっつきます．あとは洗浄液を入れて洗います．そのあと溶出溶液を入れて溶出しますと，最終的にプラスミド溶液が得られます．これはシーケンスそのものに使えます．

この図は，1方向から入れて1方向，uni-directional flowです．こういうチューブを使ってやるのですが，ライン化を進めるうえで非常に量産に向いているということです．参考までに付け加えました．

そういうことをしてラインができてきましたので，これで実際にライブラリを作り始めようと．ここにはマウスのいろいろな組織が書いてあります．163個の異なるステージ，異なる組織から取ったものの表です．こういうシーケンサーを使って，まず両端から読んで，どんどんシーケンスをしていきます．

これは今のシーケンスの結果です．163 tissueから100万クローンを取りました．これはノーマライゼーションという手法を使っていますので，これは普通に取ったライブラリでは，大体1億クローンぐらいから取った分に該当します．ものすごい量です．それを，両端のシーケンスを決めて，同じものは除きますので，12万8000ぐらいのものに落ち着きます．（図 12）

ところが，これぐらいのグループに分類しても，やはりまだredundancyがあります．そこで，しょうがないから全長を決めていきます．現在6万個のクローンの全長シーケンスが決定しておりますが，これはまだredundancyが入っています．同じものが2個以上含まれているものがあります．3万5千個，ユニークなシーケンスが取れてきました．



（図 12）

そのうち，この6万個に3万5千種類の遺伝子が入っております．しかも2万1000クローンに関しては，注釈づけまで行いました．今，世の中ではヒトゲノムプロジェクトが出ましたが，あれから見ますと，こういう高等生物の遺伝子の数は約3～4万といわれています．この中に，かなりの部分は含まれているといえます．

私どもはそういうものを注釈づけて，目でシーケンスを見ても何かわかりませんので，注釈づけしていきます．注釈づけをしていくために，例えば全く今まで知られていた遺伝子は知られていた遺伝子とする．その知られている遺伝子によく似た遺伝子，マウスの遺伝子によく似た遺伝子，これはsimilar-IIとか，定義づけをしてどんどんやっていきます．（図 13）

そしてヒトには知られているけれども，マウスに知られていないものとか，モチーフをまた見てみる．例えば転写調節因子が含まれるzinc finger domainがあるものはこれだけとか，そういうモチーフを見る．それから，タンパクはコードしているけれども，絶対に何かがわからないというようなものも入っています．

こういうものをどんどんコンピュータで分類して，あとは目で確認していきます．そのためには専門家を呼ばなければいけないというので，去年の夏に60人ぐらい，世界各国からこういうものに興味がある研究者を呼んで，ミーティングをしました．そしてこういうデータベースを作っていったわけです．これは何に似ていますとか，これはこういうものです，などと書いてあります．

これを分類します．あるものはcell division，分裂に関係あるものとか，energy metabolism，エネルギー代謝に関係あるとか，いろいろ書いてあります．

びっくりすることに，機能のわからない新しい遺伝子が山のようにあります．これはいったい何をやっているのだろうかというのが，今後の課題になってくるのです．

それをもう少し細かく分類したものがこれです．

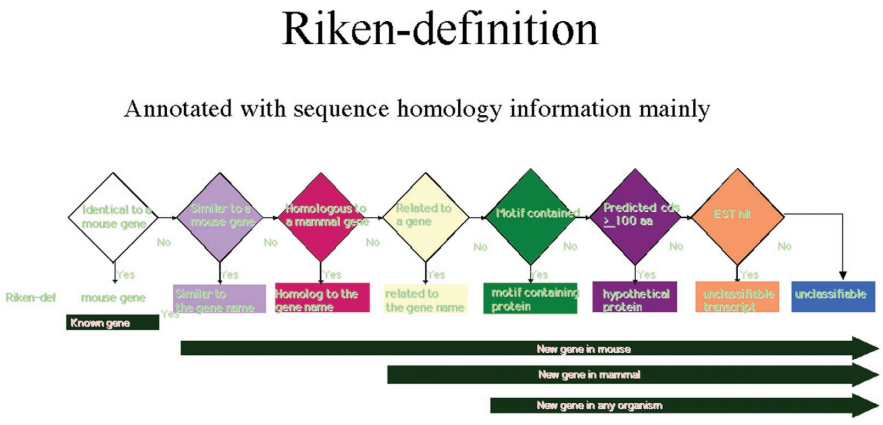
現在2万1千，これは去年終了しました．現在，‘typhoon set’とありますが，FANTOMというのはそのミーティング名の略です．Functional ANnotation of Mouse cDNA．annotationというのは注釈づけです．シーケンスに，一個一個目で見て（curation），これは何かという名前を付けていく．名前を付ける操作をするものを，我々のこういう国際的なコンソーシアム，FANTOM Consortiumと呼ぶようになったのですが，FANTOM1というのがその名前で，最初の2万1000個を終わらせました．

それが‘typhoon set’となり，これは今年の10月で4万5千個になります．たぶん‘cherry-blossom set’が来年4～7月ぐらいまでの間に12万8千，全部，シーケンスを決定していきたいと思っています．要するにこの全部の中に5万個のユニークな遺伝子がたぶん入っているだろう．そのうちの3万～3万5千はタンパクをコードしているユニークなジーンであろうと思われます．

いったんこういうシーケンスが決まりますと，ゲノムの配列がわかっていますので，コンピュータで比較しますと，先程も村松先生のお話に出てきましたが，hybridization in silicoで，コンピュータの中で考えて，ゲノムのどこにcDNAがあるかというのを見ていく操作をします．

現在，6万個のうちの3万8千個ぐらいがゲノムの上に張り付いているのですが，ときには1個のcDNAが，いろいろなところに張り付きます．これはよく似た相同的なファミリーがあちこちにあるということです．ときには1個の遺伝子から複数のcDNAが出てきます．これはalternative splicingといいます．exon（構造配列）のつながり方が違います（図 14）．

現在，我々のバンクはどの程度を含んでいるかというと，アメリカのNIHの中にあるNCBIというバイ



（図 13）

オインフォマティクス・センターがありますが，そこで1個ずつ遺伝子を目で見て，これは絶対遺伝子だというのを，セットを作っています．そのセットのことをレフセック（Ref-Seq:Reference Sequence）と呼んでいます．そのレフセックで今現在，絶対これは遺伝子だぞというのが，7千個報告されているのです．その7千個のうち大体6400個を，我々がランダムに取ったものがカバーしています．それ以外に1万8千個あるのですが，この個数から見ると，我々のバンクは全遺伝子の9割をカバーしているのではないかと考えます．

その中から，いろいろなタンパクのシーケンスがあります．いろいろなモチーフがあります．新しいモチーフが，実は見つかってきたりしています．例えば新しいタイプのロイシンジッパーとか，今日は全部挙げませんが，いろいろなものが見つかっています．

こういうことを実行しますと，世の中には2つのゲノム・プロジェクトがあります．DNAからRNAになってタンパクになります．このDNAのシーケンスを決める．これがゲノム・プロジェクトです．RNAの配列，クローンを集めて配列を決めるのがcDNAプロジェクトです．（図 15）

ゲノム・プロジェクトの方は21世紀になった今年の2月15日にヒューマンゲノムのドラフトが出ました．cDNA・プロジェクトの方は，今度我々がこれより1週間前，2月8日号に，我々の最初の2万1千個のセットを報告しました．

おもしろいことに，この2つは関係があります．なぜ関係があるかというと，実はこれは我々が強調するのですが，その中に，マウスのcDNAというのは，なぜマウスをやったか．ヒトの病気，医学に貢献がない場合は，あまりおもしろくない．ところがマウスは非常に重要です．ヒトでできない医学上の研究のほぼすべてを，網羅できます．

ヒトゲノムのゲノム配列だけでは，どこが遺伝子かわからない．そこで我々が取った遺伝子そのものの配列を見て，ヒト遺伝子のゲノムの中から，どこが遺伝子であったのかというのを想定する．ヒトゲノムの表のこの部分のくだりは私たちが書いた部分ですが，こういうところが非常に使えるということで，cDNAは非常に重要です．

さて，とにかく今，一生懸命集めているということですが，次は，集めてもその遺伝子が，いつどこでタンパクになっているか，mRNAになっているかを知るべきであると思います．そのためにマイクロアレイを使います．このマイクロアレイは，たぶん次の油谷先生が言ってくれるので，今日，僕は原理のスライドを書いてきていません．これは見てもらったらわかりますが，スライドガラスの上に，細かいDNAの点が打ってあります．ここの中に22,000個の点が打ってあります．（図 16）

この22,000個の点を打つために，こういう機械を作りました．この機械の先の直径が100 μ ですが，その中に異なるDNAを打っていくのです．1つのアームの先端がこういうふうに48ピンあり，1ストローク0.5秒で打つような機械ですので，1秒間に192点打ちます．

こういうものを20 K・2万種類の遺伝子がいつどこで発現しているかを，どんどん見ていきました．今，40 Kに伸ばしているところですが，brain , liver , kidney , lung , 脾臓，心臓，こういうもので，いつどこでどの遺伝子が発現しているかを見るデータベースを，どんどん作っていきます．現在のところ2万遺伝子が50個の種類のtissueで発現しているというデータベースを作りました．これを4万遺伝子に増やしつつあります．（図 17）

さて，タンパクが，次にタンパクとどう相互作用しますかというインタラクションをするデータベースも必要である．タンパク－タンパク・インタラクションはどういうことか．

これはあるcDNA , 10万個あるのか, 3万個以上は絶対ありますが，数万個あるのでしょう．そういうものはどれとどれが相互作用するかを，全部スクリーニングしていく．このようなタンパク・インタラクションのスクリーニングは，いろいろなところで役に立ちます．

基本的に，こういう完全長cDNAの一番重要なポイントは何かというと，タンパクを作ることができるわけです．それで，こういうスクリーニングのできるシステムを作りました．mammalian cell hybrid とありますが，これはスクリーニングシステムの原理です．

今日は詳細を省きますが, 100万組を1日にスクリーニングできるようなシステムです．（図 18）

こういうものを作って，どのタンパクとどのタンパクが相互作用するかをずっと調べて，それをこういうかたちのデータベースにしてあります．これはどういうことかというと，この黄色の四角は1つずつの遺伝子で，タンパクです．このタンパクは，例えば太い矢印で，このタンパクは強くこのタンパクに相互作用するということが，ずっと示されています．これがデータベースのかたちで入っています．

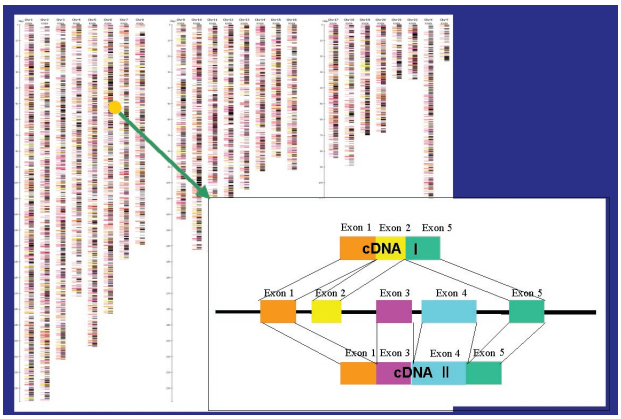
さて，こういうものを全部作っていくと，いったい何ができるかを少しだけお話ししたいと思います．

まず我々，医学領域の研究は，おそらくこれは堀川先生が話されるのではないかと思いますが，糖尿病の原因遺伝子は何かというのを調べます．それを調べるのは遺伝学です．GeneticsとGenomicsを使います．

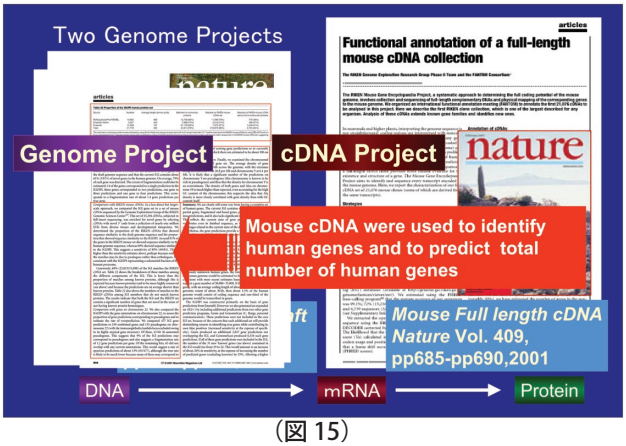
positional candidate cloningとは何かといいますと，親から子に病気が伝わると，その病気と一緒に伝わるゲノムの領域はどこかを探します．そうすると，その位置にその病気の原因遺伝子があるでしょうということで，それをずっと詰めていくやり方です．（図 19）

それで病気の原因遺伝子がつかまります．しかし，それで何がわかったか．例えばこのタンパクは，この病気の原因遺伝子だと一発でわかる場合もありますが，わからないことがいっぱいあります．それを解明するためには，パスウェイを調べなければいけない．この遺伝子のパスウェイはこういうパスウェイで，最終的にこういう病気が生じるというのを，ずっと転写のパスウェイを見ていく．この遺伝子がその次の遺伝子を活性化したり，抑えたりします．それがcascade（滝）のように流れていくわけです．こういうものがどういうルートをたどっているかを，見なければいけないことになります．

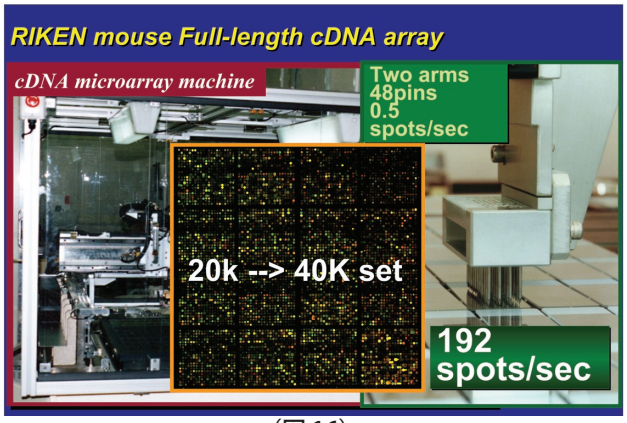
その一項に注目してみますと，これも非常に簡素化された図ですが，あるタンパクとタンパク，例えばこのタンパクと別のタンパクが相互作用して，何かのコ



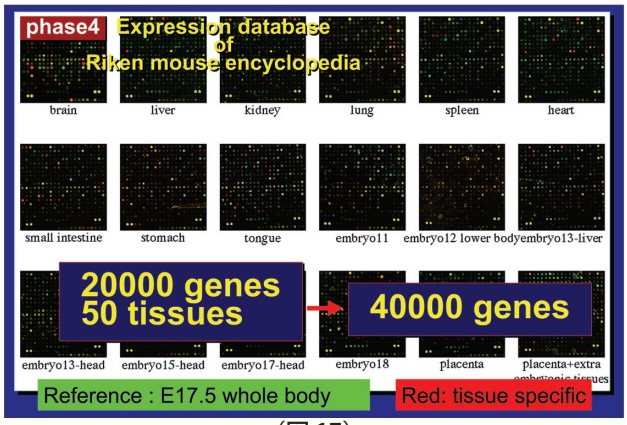
（図 14）



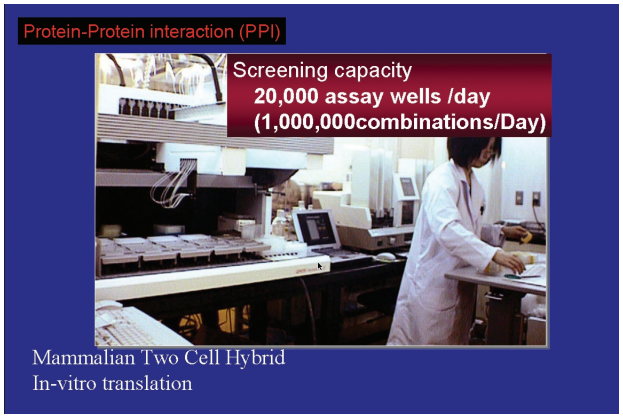
（図 15）



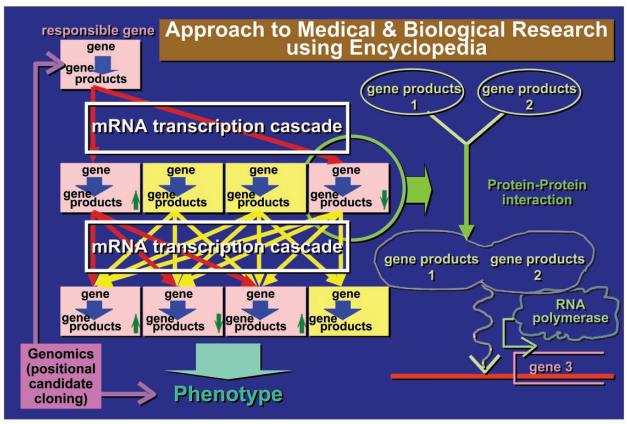
（図 16）



（図 17）



（図 18）



（図 19）

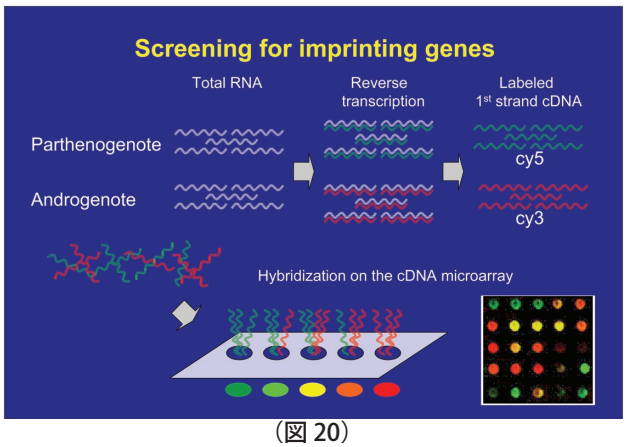
ンプレックスを形成して，その次の遺伝子を活性化する．RNAポリメラーゼを活性化するということで，こういう流れができるわけです．こういう流れを見ていくことが，非常に重要なことです．

遺伝子を同定していくために，例えばどういう同定方法がありますか．ヒトの遺伝子のマップ，病気のマップがあります．病気と一緒に伝わるゲノムの領域はどこかを調べて，病気の遺伝子が，染色体のどこにあるかが記述してあるマップがあります．逆に，完全長cDNAで（我々はcDNAを山のように取っていますから），ゲノムの配列が出てくると，どこにあるかが全部わかります．これだけでも結構，遺伝子がつかまってきますが，これプラスまた別の情報を入れると，もっとよくつかまるのです．

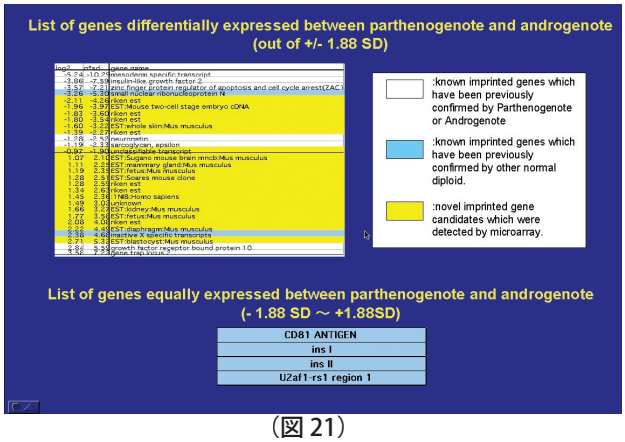
実例を言います．インプリント（imprint）という現象があります．このインプリントは刷り込み現象といいますが，例えば父親と母親の遺伝情報，卵子と精子から来る遺伝情報は均一ではないです．お父さんから来たゲノムの染色体からだけRNAができて，お母さんから来た場合はできないとか，その逆という現象があります．現に，お父さんから病気の原因遺伝子を受け継いだときだけ，発症するような病気がたくさんあります．逆に言いますと，転写がインプリントしている．お父さんからだけRNAがありますよというマップがあれば，これとこれを比べると，候補遺伝子（candidate gene）を挙げることができます．（図 20）

これでうまくいった実例を1つ言います．インプリントの遺伝子をまとめて取ってこようと思ったらどうしたらいいか．単為発生胚（parthenogenote）と雄核発生胚（androgenote）ありますが，パルセノゲノムというのは，単為発生胚です．この単為発生胚というのは，マウスなどで母親の雌性前核だけを媒体にして人工的に作る胚です．それは父親のゲノムが入っておらず，母親のゲノムしか入っていません．逆にアンドロゲノムというのは父親の遺伝子しか入っていません．この2つを用いて，こういう胚からRNAを取ってきます．このRNAを取ってきて，片方を緑の色素，片方を赤の色素でラベルして，標識をしておいて，先程，マイクロアレイ（各種の遺伝子が載っている）に配列しますと，この2つは競合して，母親からだけ発現している遺伝子が緑に，父親から発現している遺伝子は赤に染まるはずです．これで実際にやってみますと，いっぱいインプリント・ジーンが取れてきました．（図 21）

おもしろいことに，我々は実はもうシーケンスを全部やっていますので，ゲノムの中の染色体のどこに何があるか．横の1本のバーがcDNAの位置を示してい



（図 20）



（図 21）

ます．これで3万5千個，ほとんど全部がマップされていますので，この中のどれかという話です．

実はおもしろい1個の遺伝子が見つかりました．これはZAC1という遺伝子ですが，これはzinc finger domain（亜鉛をcurateするかたちの構造）を持った，あるタンパクです．このZACタンパクは，もともと肺癌や卵巣癌を抑える癌抑制遺伝子として取られてきた遺伝子です．現にこのZAC1が，今のやり方でスクリーニングしたらインプリントしていることがわかりました．父親からだけ発現します．（図 22）

おもしろいことに，このZAC1は染色体の69-24にマップされているのですが，ちょうどそこに新生児一過性糖尿病の遺伝子がマップされました．この新生児一過性糖尿病は，父親から遺伝子を引き継いだときだけ病気を発症します．（図 23）

そういうものを実際に見てみますと，例えばこれはtrisomyですが，こういうふうに父親からの領域が2倍になると，病気が発症する．すなわち父親からだけ発現している遺伝子ですから，発現量が増えるという現象のときに，この病気が出てくることになります．

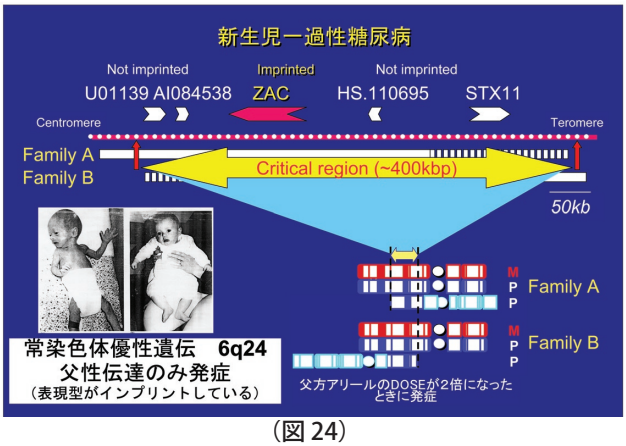
実際，こういうブレイクポイントを，染色体のトランスロケーションの領域がいろいろな患者さんがいま

すので，つぶさに調べると，ここからここまでの領域の中，400 Kbの中に入っていることがわかります．その中に実はZAC1という遺伝子があって，インプリントしていたわけです．（図 24）

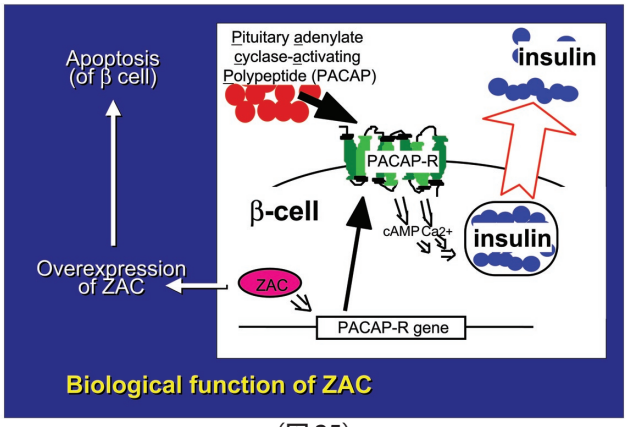
現に，これが病気の原因遺伝子です．なぜなら，Pituitary adenylate cyclase-activating Polyopeptide (PACAP) のレセプターがあります．このレセプターにシグナルが行くと，すい臓のベータ細胞からインシュリンを分泌することを刺激するシグナルになりますが，このレセプターの受容体の遺伝子を調節しているのがZAC1です．また，ZAC1がover expressionしますとapoptosisが生じます．こういうことで，糖尿病が発症しているのだろうということがわかったわけです．

このようなアプローチは，これは全部，こういう表現系として，病気がインプリントしているものです．こういうものを全部使いますと，同じようなやり口で取れていくのではないかと考えます．（図 25）

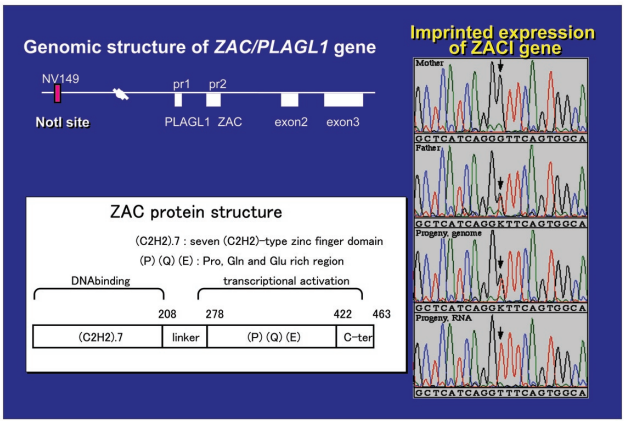
このようなデータを，人間が頭で考えてもしょうがないので，コンピュータに考えさせますので，データベースを作ります．このデータベースのマネジメントシステムを作ります．ジェノマッパーといいます．（図 26-29）



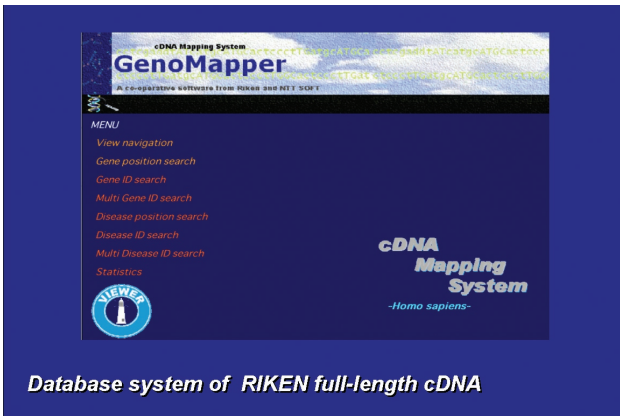
（図 24）



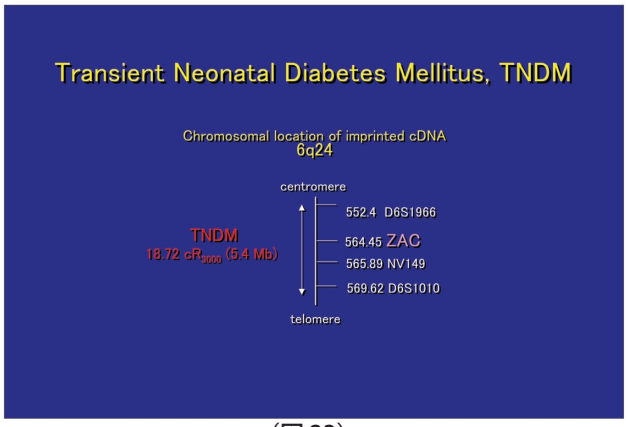
（図 25）



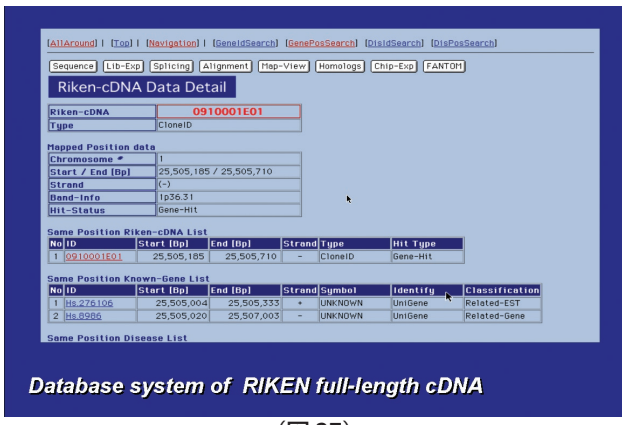
（図 22）



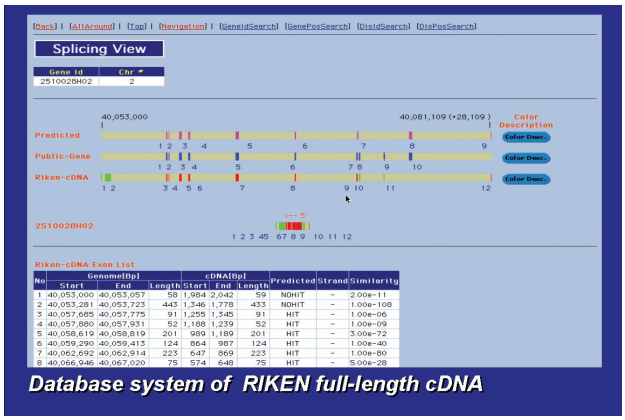
（図 26）



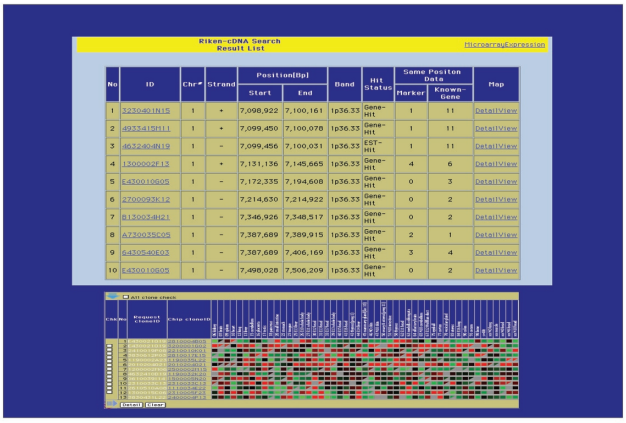
（図 23）



（図 27）



（図 28）



（図 29）

このジェノマッパーでは，クローンのIDをここに，染色体は何番で，染色体の何ベース目から何ベース目にマップされています．ストランドのプラスマイナスでは，マイナスですと．リージョンは染色体19のどこですと．こういうのをざっと入れていくわけです．

そうすると，ゲノムのA・G・C・T，遺伝子の予測プログラムで予測された配列（exon）が，こことここことここと教えてくれるのですが，実はコンピュータはしょせんコンピュータです．実際，RNAを取ったわけではないですから，実際にRNAを取ってできません．ですから，理研のcDNAを使いますと，こんなところに別のexonがあったり，こんなところに余分なexonがあったりすることがわかります．

実際，今のようなpositional candidate cloningをやるためには，染色体の番号とフランキングマーカーを入れますと，どんな遺伝子がそこにあるかを教えてくれますし，そのときの発現情報も教えてくれます．

病気の原因遺伝子をつかまえるためには，3つの重要な情報があります．1つは，染色体異常の位置情報，発現情報，タンパクのインタラクションの情報です．こういうものを使っていきますと，どんどん後方遺伝子を縮めていって，最後に病気の転・変異（SNP），1塩基の置換を見ることによって，その病気の原因遺伝子が同定されます．

2～3ここで実例を挙げますが，skin cancer（皮膚癌）の感受性の遺伝子が，遺伝学的にマップされていたのです．これは90年代にこんなマップが出てきたのですが，ここから全然進展していないのです．

というのは，これは広大な領域で，これはいったいどこの遺伝子かわからないということであったのですが，実際，我々のこういうやり方で，遺伝子がほぼ3万5000個マップされています．ですから，そこの領域にあるもののsourceでいきますと，7000個に絞り込むことができる．皮膚癌の感受性遺伝子は皮膚で発現しているということで，皮膚で発現している情報を入れますと，1400に縮められる．

また，これは少し難しいのですが，タンパクのインタラクションをしているものが，例えば先程言いました，病気の原因遺伝子に両方ともたまたまマップされた場合は，この両方ともcandidate geneである．こんな性質を利用しますと．この3つの場合もそうですが飛ばしますと，おもしろいことがわかります．

これは染色体を1番から順番に回したのですが，この赤のバーのところが感受性の遺伝子の領域です．たまたまここにある遺伝子で，皮膚で発現していて，なおかつタンパクの相互作用するものどうしはどれでしょうというのをコンピュータが教えてくれ，こういうものがcandidate geneとして挙がった．6個に絞り込めたということになります．

結局これはシーケンスすると，皮膚癌に罹りやすさを支配する遺伝子は，実はapoptosisに関係する遺伝子だったのですが，こういうこともどんどんわかってくることになります．

最後のまとめですが，我々はテクノロジーを開発してエンサイクロペディアを作りました．これはフェノタイプと遺伝子を結びつけるのに非常に役に立ちます．エクスプレッション・パターンもそれを絞り込むのに役に立ちますし，この中で分泌タンパクがあると，そのままプロテインのタンパク製剤になります．不完全長cDNAは，そのタンパクの三次元構造から，例えば創薬などに用いられるように，三次元構造はX-ray crystallographyとNMRを使いますが，それのタンパクを調達する材料になります．また，タンパクのインタラクションを，例えば阻害するものを見つけてみたりすると，これはdrug designになります．（図 30）

こういうことをやろうと思ったら，やはり非常に多くの人たちが関与しなければいけません．そこで，これは理研の我々のグループで，これは機械を作って

くれる技術部です．これは動物の細胞を取ってくれる化学合成屋さんです．これは病院関係です．企業群が先程の機械を作ってくれました．マイクロアレイをどんどんするコンソーシアム，それからFANTOM Consortium，これは一個一個目で見てやるannotationです．（図31）

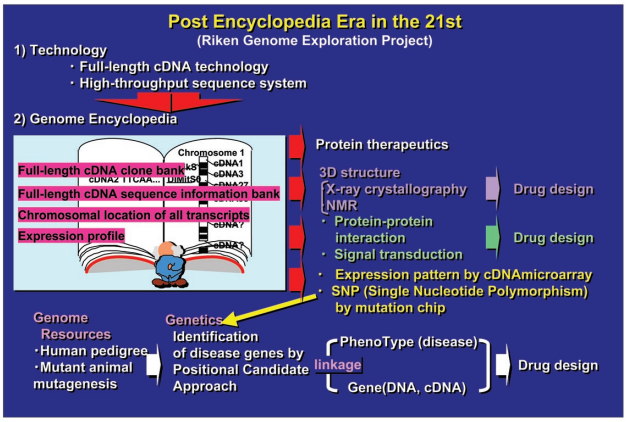
それから，うちの所長の和田先生と，顧問をしていただいている村松先生がアドバイスをしてくださいました．

もう1つ重要なのは，他とよく連携するためには，国立遺伝研のDDBJの五条堀孝先生が，非常に我々のデータをリリースするために，ものすごく手伝ってくれました．それから，世界のみんなが寄り集まって，FANTOMというコンソーシアムを作りました．（図32）

このように，ゲノム・プロジェクトを始めるときには，「ゲノムはアメリカに取られたか」と思いましたが，とにかく「完全長cDNAだけは日本のお家芸にしてやるぞ」と思ってここまで一応もってきたのですが，現在のところ，世界で一番大きなtranscriptomeです．

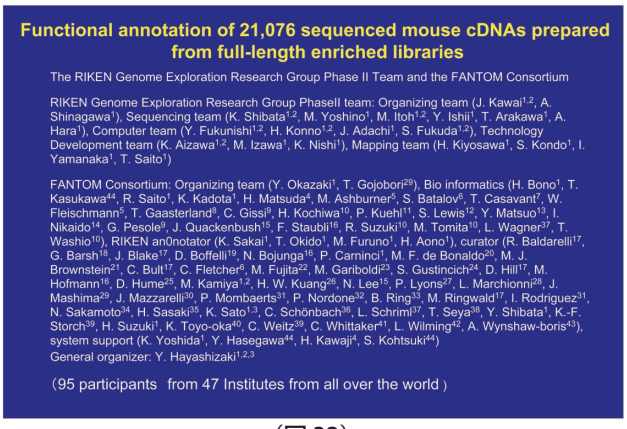
我々が作っただけではだめで，皆さんに使ってもらわないといけないので，みんなにディストリビューションするように今できています．注文してくだされればディストリビューションする会社もありますので，これをお使いになることはできます．

どうもご清聴ありがとうございました．



（図30）

（図31）



（図32）

座長：大変おもしろいお話をいただきましてありがとうございました．せっかくの機会ですので，スペシフィックな質問がありましたら1～2題受けたいと思いますが．よろしいですか．

先生，4万セットのうち，現時点では2万セットですね．仮にcommercially availableとして，一般研究者が手に入れるとして，どれくらいで手に入るのですか．

林﨑：会社に聞いてみないとわからないのです．私は会社のエージェンシーでないのであれなのですが，企業向けと学術向けで完全に分けてあると思います．2万個でいくらぐらいだったでしょうか．100万円と言ったと思いますが．

座長：1セット100万円ですね．

林﨑：1セット，全部のセットです．あれは2万数千あったと思います．

座長：ございませんでしょうか．どうぞ．

参加者：どれくらいの数のマウス特有の遺伝子がありますか？

林﨑：それは非常におもしろい質問ですが，結構あります．結構というのはどの程度かという質問なのですが，実はマウスの中で，ヒトにマップできないcDNAで，なおかつオープン・リーディング・フレームがはっきりとある，というのがあります．個数としては何十ではないですか．もう1桁か2桁上です．

座長：ほかにございませんでしょうか．

参加者：細胞内局在を研究している人たちはいるのですか．

林﨑：それは細胞内局在でやっているグループがあります．我々もこれを使って，*in situ* hybridizationをバーッとやってみたり，そういうのは1つはアメリカ

にMGCというプログラムがあって，特に脳でBMAPというプロジェクトをやっています．脳で*in situ* hybridizationをずっとやって，細胞内の局在を見たり，オーストラリアにも，やっているグループがあります．

参加者：そちらでは，局在の研究を進める可能性はないのですか．

林﨑：うちの研究室の中では，細胞内の局在というより，私たちはどういう順番でやるかというと，全部やるのはものすごく膨大なので，例えばある特定の病気，ある特定の何かというときに，少なくとも同じ細胞で発現しているものどうしを比べる，という戦略を取っています．特に病気にフォーカスしたものであれば，それどうしで先にやる．そうするとパネルのあるもののうちで，虫食い状に埋まっていく，そんなかたちです．

座長：それでは時間も押しておりますので，予定どおりここで10分，ブレイクをいただき，休憩させていただきます．

林﨑先生，どうもありがとうございました．