

# 林崎生体分子機能研究室 Genome Science Laboratory

主任研究員 林 崎 良 英  
HAYASHIZAKI, Yoshihide

ゲノム科学研究室は理研のゲノム科学研究の中核として、理研ゲノム科学総合研究センターと連携しつつ、ヒトや哺乳類のトランスクリプトーム研究を推進してきた。研究から、哺乳類のトランスクリプトームが予想以上に多様で複雑であることが次々と明らかになってきており、今後の研究のために新たな技術の開発が求められる。ゲノム科学研究室は、これまでもさまざまな技術の開発実績をもち、1細胞のトランスクリプトームを解析する技術、タンパク質によるナノ構造体の構築など、現在も革新的な技術の開発を進めている。また、すでに開発した技術についても、高速シーケンサに対応させるための改良などを継続して行っている。新たな技術により解明されるトランスクリプトームの詳細は、さまざまな生命活動や疾患の研究に大きく貢献するものである。

## 1. 一細胞ライブラリー作製法の開発 (Plessy, Carninci; Hensch(学習機能研究グループ))

前年度我々はナノレベルのCAGE法や、全脳組織を使ったDNase hypersensitivity解析を含めた一連の重要な技術開発を報告した。ナノCAGE法は理研BSI (Hensch)とGustincich 教授が率いるSISSA研究所との共同開発によるものである。個別のプロジェクトの詳細は以下の通りである。

### ナノCAGE法の開発と検証

我々はcDNA マイクロアレイおよびCAGE法による神経細胞における遺伝子発現の包括的プロファイリングを目指して、タンパク質mRNA、non-Coding RNAs (ncRNAs)さらにアンチセンスRNAsも含めた全てのRNA転写因子を包括的解析を行った。これら全てのRNAsの発現レベルの観測とそれらのプロモーターの詳細なマップ構築を目指すものである。この研究は転写とプロモーター構造の関連を明らかにし、ニューロンの転写ネットワークの解明につながるものである。

本研究においては、神経細胞群の完全な転写プロファイルの解明を目的としているため、我々は下記のサンプルを用いた。

1. 嗅上皮
2. 中脳ドーパミングループA9とA10

我々が嗅上皮に注目した理由として、過去においてこの組織についての分子生物学研究が大変有益であったこと、例えば単一のニューロンのみならず全組織を対象とした遺伝子発現のいくつかのサンプルを検出できたことなどである。ドーパミンニューロンは、同一の細胞型の異なるサブグループが脳の機能において異なる役割を果たしている点で確立されたモデルである。

我々はまずレーザーキャプチャー法の適用に向けてRNA回収プロトコルを最適化した。Zinc固定剤はRNAを安定に保ちつつ組織の構造および形態を保存することが可能であると同時に、後に容易に精製することが可能であった。また、この固定剤はGFPと同時に利用可能であった。はじめに、レーザー照射により顕微鏡下で500DaのGFP発現細胞を含む嗅上皮細胞切片20個を(SNc)(A9)や中脳の腹側被蓋領域(A10)から切り取り、その切除部位を回収した。その後、 $\mu$ MACS Super Amp Kitを用いてRNAを用いて増幅し、FANTOM2マウスcDNAクローン14000個にハイブリさせた。Axon 4100を用いてスキャンし、得られたデータは規格化した後、R Bioconductorを用いて解析した。

さらに、我々は前年度の技術開発をさらに進歩させ、ナノCAGE法を開発し、これを嗅上皮細胞A9及びA10細胞に適用した。ナノCAGE法はナノスケール遺伝子発現 (Cap Analysis Gene Expression) の頭文字をとったものである。この技術はmRNAもしくはncRNA (non-coding RNA)の5'末端から25ヌクレオチド配列のタグ配列を切り出すものである。これらの短いタグの塩基配列を決定し、マウスゲノム上にin silicoでマッピングすることにより、mRNA及びncRNA転写開始点 (TSS) を同定する。これらのタグを数えることにより全細胞のmRNAs発現を『transcript per million』という単位で測定することが可能になる。言い換えれば、これはプロモーターごとの転写活性をデジタル的に測定するということである。この技術はRT (逆転写酵素) を使ってcDNAを合成する際に起こるcap-switch-reactionに基づいたものである。数ナノグラムのRNAから始まり、RTはRNAのcapサイト(長鎖ncRNAやmRNAなどのRNAポリメラーゼにより合成されたRNAの5'末端に存在する構造)にたどり着き、他のオリゴヌクレオチド (cap-switch oligonucleotide) の転写を開始する。このキャップスイッチオリゴヌクレオチドはcDNAを約25ベースに切断し、タグ (CAGEタグ) を作り出す酵素であるEcoP15 を含有する。この技術の開発は、多数のncRNAsを含むポリアデニル化していないRNAsを分離するためのランダムプライマーにより二量体が生成するため困難であった。とはいえ、ncRNAには遺伝子制御の役割があると思われることから、ncRNAを回収することは必須であった。また、レーザーを用いて回収・精製したニューロンは部分的に分解しており、オリゴ-dTのみで転写開始ができないことが考えられるため、ランダムプライミングも必須であった。キャップスイッチ反応後のcDNAの増幅を目的としたオリジナルのプライマーデザインの開発によってランダムプライマーの使用が可能になった。我々はこの方法を『semi-suppressivePCR』とよんでいる。この方法では短いランダムプライマーによって作られたprimers dimmerのような短分子は増幅されず (抑制されている) 長いcDNAは抑制されずに増幅させることが可能で、数マイクログラムのcDNAを回収して、ナノCAGEタグの調整に供することができる。もう1つの重要な成果として、1回のシーケンスで2000万までのCAGEタグをシーケンスすることができるようになったことがあげられる。1つの細胞が30万から50万のmRNAs/ncRNAsを含むと考えた場合、最も希少なものを観測するには、各遺伝子につき10倍のサンプルをとることが必要になる。1回のシーケンスで百万のタグのシーケンスが可能になれば、トランスクリプトームの包括的なマッピングに利用することができる大量のタグを作成する事が可能となる。

この技術開発の後、我々は3つのサンプルについての転写プロファイリングにこのナノCAGE法を幅広く活用し、少ない細胞数で包括的CAGEプロファイリングを成功させた (マウスの嗅神経上皮、A9とA10の catecholaminergicニューロン)。我々は最近これらについてSolexaおよびIllumina Genome Analyzerを使用してより深いシーケンシングを行った。これらのサンプルから

3500万のCAGEタグをマウスゲノム上にマッピングすることに成功した(嗅神経細胞で1500万、A9、A10にはそれぞれ1000万)。これらのCAGEタグは過去のニューロンのトランスクリプトーム解析研究の中で最もdeepなものである(また、単独の細胞型や組織についても)。まだ大量のデータ解析が残っているものの、単独ニューロン転写の大半を把握した。データセットは既に quality control をパスしており、タグについてのデータは順調にゲノム上にマッピングされている。新規の転写の解析・解明が現在進行中であり、また少なくとも2つの論文 (1) 嗅神経上皮をテストケースとした技術、(2) A9対A10のトランスクリプトーム解析について を準備中である。

遺伝子予測では同定できず、CAGEタグによって発見はじめて発見された新規の嗅神経レセプターのプロモーターの特徴は、アンチncRNA転写活性の発見と、多数の遺伝子の3' UTR(3' UTRプロモータ)からの転写活性の発見である。

### DNA超感受性領域検出法の開発

ゲノムDNAには広範囲にわたる転写因子が結合可能な部位が含まれている。しかし、生きた細胞や組織ではこの結合領域のほとんどは使用されることがない。これはDNAがクロマチン化されており、到達不可能であるからである。ヒストンが離れることによってDNAが脱凝縮されている(通常これはアセチル化を含むいくつかのヒストン修飾による)アクティブな部分にのみ到達可能である。

DNase hypersensitivity 技術とは、DNase 処理によって、クロマチンが凝縮解除(open)されている領域特異的に解裂を引き起こす技術である。これらの領域は転写因子が到達できるポジションとであり、また遺伝子調節領域/相当する。DNAの調節領域に切れ目を入れるDNase 処理の後、ゲノムDNAを抽出し、これにII型制限酵素サイトを含むリンカーを結合させる。さらにいくつかの工程の後、II型制限酵素作用させて切断すると、20 ntのタグが生成するので、この塩基配列をSolexaシーケンサを用いて決定する。これをコンピュータを用いてゲノム配列と比較してDNase hypersensitive 部位を同定することにより、一度にサンプルの全制御領域を検出することができる。

技術は既に構築されているが、FANTOM-4/ゲノムネットワークプロジェクトにデータセットを提供するため、THP-1 単球細胞株をサンプルとして、Solexaを用いて250万個のタグを作成および解析し、また、同じ組織からのCAGEタグと比較した。ここで重要なことは、さらに小スケールのサンプル及び脳全体にこのプロトコルを適用したことである。この動機は、ヒストン deacetylase inhibitors (T.Henschとの共同開発) を加えることにより脳の視覚野における可塑性を制御する広い領域を識別する必要性があったことである。この理由から、様々な技術的問題を解決しつつ、脳全体に適用できるプロトコルを開発した。このプロトコルを用いて、2つのDNase hypersensitiveライブラリーを作成した。すなわち、視覚野およびバルプロ酸処理したサンプル(脳の皮質を活性化させることを発見した)である。これらのライブラリーから1400万以上のタグを作成し、3つ以上のタグを含むクラスタにおいてもDNase hypersensitiveの明確なピークが観察された。現在、遺伝子発現と規定領域を関連付けるために、同じタイプのサンプルからのDNase hypersensitiveタグをAffymetrix Gene-Chipを用いて解析している。

### マイクロアレイと発現解析

過去に理研BSIのT.Henschとの共同開発で、脳の皮質を研究するためのマイクロアレイを開発した。これについては、基盤の性能の概要を論文にまとめ、現在出版のためPLoS Oneに投稿している。

脳の皮質のトランスクリプトームは、領域ごとの差異よりも個人差の方が顕著であることから明らかであるほど均質である。これは、多数の細胞型が存在するため、ひとつの細胞における分布が他の細胞型からのノイズによって均一化されることによると考えられている。同種のトランス遺伝子を発現する細胞の研究は、細胞群がその転写プロファイルごとに識別されていることを示している。我々は、記号化及び非記号化の両方のマウス視覚野の転写を示す冗長性の低い16,209個の完全長cDNAクローンを作成し、別のタグ技術であるCAGE法により、72%のクローンに皮質発現を確認した。このクローンをPCRにより増幅した後、ガラスディスク上に配置して、マイクロアレイを作成し、これを用いてparvbumin-egfpトランスジーンマウスの視覚野から採取した細胞から得られたRNAを解析した。マウスの脳の転写解析に特に適しているアノテーションされたcDNAクローンをを用いてマイクロアレイを作成し、EGFP活性および不活性細胞を比較したところ、30%以上のクローンが異なる発現を呈した。このような資源および情報は、マウスの脳の可塑性についての研究に有用であろう。

## 2. DNAポリメラーゼライブラリーの構築及び核酸関連タンパク質の発現系構築 (伊藤, 河合, 林崎)

さまざまな生物が有する DNA ポリメラーゼを初めとする核酸関連タンパク質には、それぞれにその由来する生物に伴って特有の性質を有することがある。このような特殊な性質は学術及び産業上、非常に有用となることが多くの例で見いだされている。本研究は、理研 BRC 微生物材料開発室と共同で、JCM が有する細菌より DNA ポリメラーゼを初めとする核酸関連タンパク質の遺伝子及びその発現系を構築してライブラリー化することを目的とする。

まず、理研 GSC 遺伝子構造機能研究グループで開発された等温核酸増幅法である SMAP (Smart Amplification Process) 法に用いられている Strand Displacement (SD)活性に着目し、本活性を有する新規 DNA ポリメラーゼのクローン化と発現系構築を行った。これまでに本活性を有する酵素が *Bacillus* 属細菌からクローン化されていることから、本属細菌に注目し、JCM にある好温の細菌をターゲットとしてクローニングを行った。その結果、*B. smithii* 及び *Alicyclobacillus acidocaldarius* subsp. *acidocaldarius* を初めとする複数の細菌種よりそれぞれ DNA ポリメラーゼ I をコードする ORF (*polA* 遺伝子)をクローン化することに成功した。

このうち、*B. smithii* (*Bsm*)及び *A. acidocaldarius* subsp. *acidocaldarius* (*Aac*)由来の *polA* 遺伝子及びコードするアミノ酸配列は新規であった。これらの *polA* 遺伝子より、ラージフラグメント (LF)に相当する部分を別途クローン化し、タンパク質発現系を構築した。これら発現系で得られた DNA ポリメラーゼ LF は、強力な SD 活性を有しており、*Bsm* DNA ポリメラーゼ LF は 55、*Aac* DNA ポリメラーゼは 65 のそれぞれ至適温度であった。さらに、*Aac* DNA ポリメラーゼについては、上記 SMAP 法への応用を図り、これまでに市販されている SD 活性を有する DNA ポリメラーゼの中で最も SMAP 法に適していることが判明した。

なお、両酵素については、既に特許申請済みであり、現在、論文を執筆・投稿中である。また、*Bsm* DNA ポリメラーゼについては、遺伝子工学用酵素メーカーへのライセンス供与を検討中である。

## 3. ナノレゴ (臼井, 伊藤(吹), 伊藤(昌), 河合, 林崎; 鈴木(治) (遺伝子構造機能研究グループ))

生物体は、タンパク質に代表される特異的な結合能力を持つ生体分子を巧みに利用し、外部からの力を借りずに自発的に形成

するナノ構造体、すなわち「自己組織化」する複合体を形成し、様々な生命機能をはたしている。本プロジェクトでは、この特異的相互作用を示すタンパク分子を特異的結合素子という概念で捕らえ、それらの複数個の素子から人工融合タンパク質を設計・作成し、ナノレベルで制御可能な自己組織化を実現する機能性材料（ナノレゴ）の開発を目指した。これまでに、対称性を有するホモ多量体を「骨格素子」、及びヘテロ相互作用分子を「接着素子」とし、それらの2つの素子要素からなる融合タンパク質を作成するナノレゴの基本構築原理を元に、新たなナノ線状構造体へと自己組織化を果たすナノレゴの開発（SOR-PDZ, SOR-Zpep）を達成した。本研究では、このナノレゴ線状構造体をもとに、ナノ構造体の分子制御を達成する技術開発を行った。また、ナノレゴを用いた機能性材料開発の一端として、タンパク質相互作用を架橋点に用いた細胞親和性ハイドロゲルの開発を行った。

#### (1) SOR ナノレゴを用いたナノ構造体の分子制御

ホモ4量体を形成する超好熱古細菌 *Pyrococcus* 由来 Superoxide reductase (SOR) を骨格素子とし、可逆的ヘテロ相互作用を示すマウス由来 PDZ domain (PDZ) 又は PDZ-binding peptide (Zpep) を接着素子に用いたナノレゴ SOR-PDZ 及び SOR-Zpep は、溶液中での混合の結果、線状構造化を果たすことが TEM 観察において明らかとなった。また、MD 計算を用いた SOR ナノレゴ線状構造体のシミュレーションと TEM 観察結果との比較から、SOR ナノレゴに存在する4個の接着素子のうち2個が隣接するナノレゴとの会合に使われる多点結合を介し、線状構造化を果たしていることが示唆された。我々は、接着素子間の可逆的相互作用を制御するため、接着素子に Cys 残基を導入した改変型ナノレゴ SOR-PDZ<sub>R34C</sub> 及び SOR-Zpep<sub>W3C</sub> を構築し、接着素子間でのジスルフィド架橋形成に伴う相互作用の不可逆化を行った。この2種の改変型ナノレゴを用いることで、基板上に固定されたナノレゴ分子を基点として、フローセルによる2種のナノレゴの交互添加により、ナノレゴ分子単位での線状構造体の伸長制御を達成できた。加えて、SOR ナノレゴにおける接着素子を Ca<sup>2+</sup> に依存して安定な相互作用を示す *Clostridium thermocellum* 由来 Cohesin/Dockerin 複合体に置換した結果、ナノレゴ間の会合において単独の接着素子のみで安定した結合を示し、線状構造体とは異なる分岐を有した高分子化が起こることを見いだした。これらのことから、骨格素子による接着素子の配向性のデザインとともにナノレゴ間の接着素子の会合状態を変化させることによって、様々な自己組織化ナノ構造体の制御が可能になるものと判断された。

#### (2) CutA ナノレゴを用いた細胞親和性ハイドロゲルの開発

ナノレゴ CutA - Tip1 (骨格素子: *Pyrococcus* 由来 CutA、接着素子: マウス由来 PDZ ドメインを有する TIP-1) は、非常に安定なホモ3量体骨格構造を保ちながら PDZ 認識ペプチドと相互作用する。そこで、この CutA ナノレゴと PDZ 認識ペプチドを末端に有する4アーム型ポリエチレングリコール (PDZ-peptide-PEG) を作成し、生理学的条件下で混合すると自発的にハイドロゲルを形成することができた。さらに細胞接着サイトを持たせるため、CutA タンパク質のループ部分3箇所細胞接着モチーフである RGD ペプチドを組み込ませた変異体 CutA(RGD)-TIP1 を作成した。CutA-TIP1 への軟骨細胞の接着は、ほとんど観察できなかったが、CutA(RGD)-TIP1 では、細胞の接着・伸展が確認できた。そこで CutA(RGD)-TIP1 と PDZ-peptide-PEG と軟骨細胞の3種を混合させたところ、細胞が3次的に分散するゲルの形成に成功した。また包埋された細胞は正常軟骨と同様な円型を示し、長時間培養でも細胞が生存していた。以上のことから、遺伝子工学技術と合成高分子の技術を組み合わせ、全く新しい概念と設計思想による細胞を包埋できるゲル形成技術のプロトタイプを作成した。再生医療のスカフォールドとして有用と考えられる。

### 4. 次世代シーケンサを用いた研究 (前田, Carninci, 林崎; 鈴木 (治) (遺伝子構造機能研究グループ))

近年の次世代シーケンサ出現に伴い、DNA シーケンスの世界は劇的な変化を遂げている。次世代シーケンサにおいては、従来のサンガー法が必要となっていた、クローニング、コロニーピッキングなど煩雑な実験ステップを全て省略することができ、出発物質としてマイクログラムオーダーの二本鎖 DNA が入手できれば、そのままシーケンスを行うことが可能である。1回のシーケンスランにより、1000万塩基を超える配列を決定する事が可能であり、コスト、時間面からもゲノムワイドな解析には欠かす事のできないシーケンス技術である。また、次世代シーケンサの極めて高いスループットを利用すれば、一人のオペレーターが細菌のドラフトゲノムを数日間決定することも可能である。現在、次世代シーケンサの能力を最大限に活かした研究が求められていると言える。本プロジェクトでは、従来の CAGE 法を改良し次世代シーケンサに対応させ、転写ネットワーク描出に向けたデータ生産を行った。また、大阪大学微生物病研究所と共同で、次世代シーケンサを用いた迅速な病原微生物同定システム構築に向けた研究を行った。

次世代シーケンサでは、サンガー法とは異なる試料調製が必要である。1) シーケンスしたい二本鎖 DNA の両端にシーケンサ特有の配列を付加する必要がある。2) シーケンサの効率を最大限に引き出すため、調製する DNA 長に制約がある。などである。そのため、PCR 条件の変更、コンカテネーション条件の変更を行い、従来の CAGE 法を改良した。また、細胞分化の動的転写ネットワーク描出に向け、5塩基の DNA タグを導入し、各時間での発現パターンを経時的に観察できるように改良した。そして、RNA を抽出した各時間に対し、100万 CAGE タグ以上となるようにデータ生産を行った。

大阪大学微生物病研究所との共同研究においては、まず最初に、既に単離された病原微生物からゲノム DNA を抽出し (RNA ウイルスの場合には、逆転写反応により cDNA 化) 次世代シーケンサによるゲノム配列解読を行った。その後、臨床サンプルを利用し、臨床検体という限られたサンプル量からの病原微生物同定に向けた解析を行なった。

## The summary of the research annual report

Genome Science Laboratory (GSL) has been leading the genome research in RIKEN and elucidating the transcriptome of human and mammals collaborating with closely with Genomic Science Center, RIKEN, in which amazing transcriptome world of mammals was explored. Mammalian transcriptome is so complex more than expected so far, and still needs novel innovative technologies for further study. GSL is conducting developments of various technologies for new era of further transcriptome research, for examples, single cell technology and new enzymatic tool for DNA amplification. Transcriptome research promoted with new technologies could contribute analysis of complex biochemical events underlying life and diseases.

### 1. Development technology for single cell library construction

## Summary

In the previous year, we have completed a series of important technology development, which include the Nano-Scale Cap-Analysis Gene Expression (nano-CAGE) technology, and the DNase hypersensitivity study using whole brain tissues. The nano-CAGE has been developed in collaboration with the RIKEN BSI (Hensch) and the SISSA laboratory of Prof. S. Gustincich (Trieste, Italy).

Details for each project are below.

### Nano-CAGE technology expression and validation

We aimed at the comprehensive expression profiling of homogeneous neuronal cell types with cDNA microarrays and CAGE protocol, in order to comprehensively identify all the RNA transcripts including protein coding mRNAs; non-coding RNAs (ncRNAs) as well as antisense RNAs. For all of these RNAs, we aimed at measuring their expression level and to finely map their promoters.

This paves the way for the description of the transcriptional network of single neurons, connecting transcription to promoter structures.

Since we aimed to describe the complete transcriptional profiles of a neuronal cell population, we chose as representative cases the following neuronal centers:

1. Olfactory epithelia
2. mesencephalic dopaminergic cell groups A9 and A10.

We focused our analysis on the olfactory epithelia since molecular biology approaches to this tissue have been very fruitful in the past providing several examples of gene expression analysis of the entire tissue as well as of single neurons.

Dopaminergic neurons provided a well-characterized cell model system where different subgroups of the same cell types were organized anatomically and were having different roles in brain function.

We first optimized the harvesting protocols for Laser Capture Microdissection. Zinc fixatives showed the ability to preserve the anatomy and morphology of the tissue while keeping the RNA safe and ready to be efficiently purified later. This fixative was also compatible with GFP staining. 20 sections of Olfactory Epithelia as well as 500 DA GFP- expressing cells were excised by laser microdissection and collected by laser pressure catapulting (LPC) from the Substantia Nigra compacta (SNc) (A9) or from the Ventral Tegmental Area (A10) of the mesencephalic sections. RNA was then amplified using the  $\mu$ MACS SuperAmp Kit and hybridized on our *in house* cDNA platform which contains 14000 clones in triplicates from the FANTOM2 mouse cDNA clone collection. Scans were made with the Axon 4100 scanner and the GenePix version 5.0 software. Data were normalized and analyzed with the R Bioconductor package.

We then achieved a further technological breakthrough from the previous year developing the nano-CAGE technology and applying it to olfactory epithelia, A9 and A10 cells.

Nano-CAGE stands for nano-scale cap analysis gene expression. This technology allows preparing short tags of 25 nucleotides, which correspond to the 5' initial part of the mRNA or ncRNA (non-coding RNAs). These short tags are sequenced and *in silico* mapped on the mouse genome, where they identify the mRNA/ncRNA transcriptions starting sites (TSS). By counting these tags, the expression of all cellular mRNAs can be measured as "transcript per million". In other words, this is a digital measurement of the transcriptional activity for each of the hundred of thousands cellular RNAs at each specific promoter.

This technology is based on a cap-switch reaction, which takes place when the reverse transcriptase (RT) is used to synthesize cDNA. Starting from few nanograms of RNA, the RT reaches the RNA cap site (the structure at the 5' end of the RNA produced by RNA polymerase II, which includes all mRNAs and the majority of long ncRNAs) and "switches" the transcription onto another oligonucleotide (the cap-switch oligonucleotide). This cap-switch oligonucleotide contains the EcoP15I, an enzyme which cleaves about 25 bases into the cDNA and produces a short tag (the CAGE tag). It has to be noticed that the development of the method has been difficult because of the formation of dimers using random primers, to isolate non-polyadenylated RNAs, which include many ncRNAs. The inclusion of ncRNA has been in any case essential because they seem to have a regulatory role. Random priming is also essential, for the protocol to be used with laser capture purified neurons, which may be partially fragmented and are not amenable to be primed with oligo-dT alone.

The usage of the random primer was possible because of the development of an original design to amplify the obtained cDNA after the cap-switch reaction. We call this method "semi-suppressive PCR", by which short molecules like primers dimers created by short random-primer adaptors are not amplified (suppressed), and longer cDNA are not repressed and can be amplified to obtain micrograms of cDNA, later used for the nano-CAGE tags preparation.

Another breakthrough was to make this protocol compatible with the Illumina/Solexa sequencing instrument, which allows to sequence up to 20 million CAGE tags per each sequencing run. If we consider that a cell may have >3-500,000 mRNAs/ncRNAs, it is important to sample at least 10 fold each transcript, including the rarest ones. As the instrument allows the sequence of million tags per each run, this produces a very large amount of tags which can then be used for comprehensive mapping of the transcriptome and promoters starting site usage.

After the development of the technology, we have extensively used the nano-CAGE for the transcriptional profiling of 3 samples, allowing for the first time a comprehensive CAGE profiling of a reduced amount of cells: the mouse olfactory epithelia and the A9 and A10 catecholaminergic neurons;. For each of these, we have recently achieved deep sequencing with the Solexa/Illumina Genome Analyzer. We have achieved from these samples 35 millions of CAGE tags, which we

were able to map onto the mouse genome; 15 millions tags for the olfactory epithelia, and ~10 millions each for the A9 and A10. These CAGE tags represent the deepest survey of transcriptome of neurons (and in any single cell type or tissue) done so far. Although there is a very large amount of data still to analyze, we have captured a very large part of the transcriptome of single neurons. The datasets have passed the quality controls and the data refers only to tags that are successfully mapped onto the genome. Analysis and validation of the novel transcripts are in progress as well as the preparation of at least 2 papers, regarding (1) The technology, using the olfactory epithelia as a test case; (2) the transcriptome of the A9 versus A10.

Findings include the characterization of novel promoters of olfactory receptors, what are not identified by gene prediction but only by CAGE tags; identification of antisense ncRNA transcriptional activity; and identification of transcriptional activity from the 3' UTR of many genes (3' UTR promoters).

As reported in the previous year, we have progressed with the analysis of the CAGE tags in the hippocampus, including RACE validation to verify that the rare transcription starting sites identify by deep-CAGE (2 million tags of whole tissue CAGE) are real. Validation rate exceed 80% in the first round and we are chasing the remaining transcriptional starting sites for further validation. We are preparing this paper as a resource for the identification of tissue specific promoters in the hippocampus to be submitted to *Neuron* as a methodological/resource paper.

### **Development of DNA hypersensitive sites**

The genomic DNA contains a very large amount of potential binding sites for transcription factors. However, in the living cells and organisms, most of these potential binding sites are never used. This is due to the fact that the DNA is not accessible, as it is packed in chromatin. Only actively used regulatory regions are accessible, either because the chromatin is decondensed (usually due to several histone modifications, including acetylation), to the extent that it is devoid of its histones in the most actively transcribed regions. DNase hypersensitivity is a technology by which the cell nuclei are treated by DNase I, which under appropriate conditions introduces breaks in the region where the chromatin is de-condensed (open). These sites overlap to the positions, which are accessible to transcription factors and correspond to regulatory regions. After the cleavage of the nuclei with DNase I, which introduces nicks to the DNA controlling regions, the genomic DNA is gently extracted and ligated with linkers, which contains a ClassIIs restriction enzyme cutter. After further manipulation, cutting with such class II enzyme produce a 20 nt tag, which is purified and sequenced with the Solexa sequences (as for the nano-CAGE technology above). Computational alignment with the genome identify the DNase hypersensitive sites, which identify all the regulatory regions in a given sample at once.

Although the technology was previously established, we have adapted for the Solexa instrument and tested it for the THP-1 monocyte cell line, providing a dataset for the Fantom-4/Genome Network project. We have produced 2.5 millions tags, which are being analysed and connected with the CAGE tags from the same tissue. More importantly, we have also adapted the protocol for small scale sample and whole brain. This was motivated by our need to identify on the broad scale the regions that are controlling the reactivation of brain plasticity in the visual cortex, by adding histone deacetylase inhibitors (in collaboration with T. Hensch). For this reason, we have developed a protocol that works on whole brain, solving various technical issues. The protocol was used to produce two DNase hypersensitive libraries, containing a control visual cortex and a sample treated with Valproic acids, a histone deacetylase inhibitor, which we found to reactivate the brain plasticity. From these libraries, we have produced more than 14 million tags; we see clear peaks of DNase hypersensitive tags even with cluster including more than 3 tags. We are currently connecting the DNase hypersensitive sites with Affymetrix Gene-Chip expression analysis from the same type of samples, to correlate gene expression and regulatory regions.

### **Microarray and expression analysis**

In the collaboration with T. Hensch at RIKEN BSI, we have previously developed a microarray resource to study brain plasticity. In this time, we have summarized the performance of the platform in a paper, currently submitted for publication to *PLoS One*.

Essentially, the transcriptome of the brain cortex is remarkably homogeneous, with variations being stronger between individuals than between areas. It is thought that due to the presence of many different cell types, differences from within one cell population will be averaged with the noise from others. Studies of sorted cells expressing the same transgene have shown that cell populations can be distinguished according to their transcriptional profile. We have prepared a low-redundancy set of 16,209 full length cDNA clones which represents the transcriptome of the mouse visual cortex in its coding and non-coding aspects. Using an independent tag-based approach, CAGE, we confirmed the cortical expression of 72 % of the clones. Clones were amplified by PCR and spotted on glass slides, and we interrogated the microarrays with the RNA from flow-sorted fluorescent cells from the visual cortices of *parvalbumin-egfp* transgenic mice. We provide an annotated cDNA clone collection which is particularly suitable for transcriptomic analysis in the mouse brain. Using it on microarrays, we compared the transcriptome of EGFP positive and negative cells in a *parvalbumin-egfp* transgenic background and showed that more than 30 % of clones are differentially expressed.

This kind of resource will be useful to study the brain plasticity function in the mouse.

## **2. DNA polymerase library construction**

Some proteins and enzymes related to nucleic acids have special features caused by the original organisms. Such features are sometimes applicable in scientific and industrial field. This study is to construct the library of nucleic acid related enzymes and proteins, such as DNA polymerase, which consists of genes and its expression system. This study

is collaboration with Microbe division, RIKEN BRC.

First, we focused on the strand displacement (SD) activity, which employed in the isothermal nucleic acid amplification method, SMAP (SMart Amplification Process), developed in Genome exploration research group, RIKEN GSC. The known DNA polymerases which has this activity was derived from *Bacillus* genus and its bacteriophage. Therefore, we tried to clone from thermophilic *Bacillus* species. Finally, we succeeded to clone some DNA polymerase I genes (*polA*) from *B. smithii*, *Alicyclobacillus acidocaldarius* subsp. *acidocaldarius*, and other species.

Among them, *polA* genes and their amino acid sequences from *B. smithii* (*Bsm*) and *A. acidocaldarius* subsp. *acidocaldarius* (*Aac*) were novel. We subcloned their portion of large fragment (LF) into expression vector. The expressed proteins showed strong SD activity, and the optimal temperatures were 55°C and 65°C for *Bsm* and *Aac*, respectively. Further, *Aac* DNA polymerase LF was applied to SMAP, indicated that *Aac* DNA polymerase LF is most optimum for SMAP among commercially available DNA polymerases.

Both enzymes are already filed, and we are now preparing and submitting the papers. *Bsm* DNA polymerase LF will be also licensed to genetic engineering enzyme manufacturers.

### 3. Nanolego Project

Organisms achieve the cellular biological processes using large number of specific binding molecules typified by the protein, which are capable of forming the nano-scale complex by self-assembly. In this project, we conceive of the binding domain as nano-scale chip (Nanolego element) and design their fusion proteins in the purpose of developing new functional molecules (Nanolego) that self-assembled in order and in a controllable way. Previously, we achieved the developing of SOR Nanolegos self-assembling to the filamentous structure, SOR-PDZ and SOR-Zpep, which were produced by the fundamental concept in Nanolego design – the construction of fusion proteins consisting of both symmetric homo-oligomer (structural element) and hetero-interacting pair (binding element). In this study, we aimed to develop the techniques of molecular control for nano-architecture based on the nano-filament of SOR Nanolegos. Furthermore, as an instance of the functional application with Nanolegos, we tried to create the cytophilic hydrogel using protein-protein interaction as cross-linkage of the gel.

(1) Molecular control of nano-structure with SOR nanolego.

In the mixed solution of two nanolegos (SOR-PDZ and SOR-Zpep) consisting of a superoxide reductase (SOR) from the *Pyrococcus horikoshii* as the structural element and a mouse PDZ domain (PDZ) and a PDZ-binding peptide (Zpep) as the binding elements, the filamentous structures were observed by TEM. The results of TEM and MD simulation for the self-assembling form of SOR Nanolegos suggested that the Nanolegos use two binding elements for association each other. To control reversible interaction between the binding elements, we constructed the further engineered SOR Nanolegos, SOR-PDZ<sub>R34C</sub> and SOR-Zpep<sub>W3C</sub>, which are introduced a cysteine (Cys) residue into each binding element of the Nanolegos by site-directed mutagenesis, where the reversible interaction is expected to be covalently linked by the formation of a disulfide bond between Cys-residues at each binding element. Using these Cys-introduced Nanolegos conducted to regulate the extension of the nano-filaments from the Nanolego fixed on the surface by alternate loading of two Nanolegos through a flow cell. As another approach of conversion of the Nanolego for control of self-assembly, we constructed SOR Nanolegos with Ca<sup>2+</sup>-dependence Cohesin/Dockerin complex of *Clostridium thermocellum*. We found the branched structure in the mixture of these Nanolego as opposed to the filamentous structure consisting of SOR-PDZ and -Zpep. The result of MD simulation predicted that this branched formation arose by using single interaction of stable Cohesin/Dockerin association for association next Nanolegos. These evidences in this study suggested that not only orientation of binding site by structural element but also alteration of associating condition of binding elements would enable us to create several type of self-assembling nano-structure.

(2) Development of the cytophilic hydrogel using the CutA Nanolego.

In this study, we developed a novel method of self-assembling hydrogel formation via biospecific interaction between genetically engineered protein, CutA-TIP1 (Tax-interacting protein 1 [TIP1] having PDZ domain was fused at each end of triangular-shaped trimeric CutA), and PDZ-domain-recognition peptide which is covalently bound to each terminal ends of 4-arm poly(ethylene glycol) (PDZ-peptide-PEG). Based on molecular dynamic simulation, genetic manipulation enabled to prepare cell adhesive RGD tripeptidyl sequence at loop region of CutA [CutA(RGD)-TIP1]. Spontaneous viscoelastic gel was formed when stoichiometrically mixed of CutA-TIP1 or CutA-(RGD)-TIP1 with PDZ-peptide-PEG in the presence of chondrocytes in buffer solutions. Round-shaped single cell or its multicellular aggregates were entrapped and resided in a gel without any cellular impairment. Dynamic viscoelasticity measurement showed shear rate-dependent reversible phase transformation: at low shear rate, spontaneous viscoelastic gel was formed, but gel was transformed to sol at high shear rate. Potential application as injectable cartilage tissue was proposed.

### 4. Studies by using next generation DNA sequencer

The recent advent of next generation DNA sequencers has been inducing a drastic change on DNA sequencing. The next generation sequencers enable to continue sequencing by omitting time consuming procedures such as cloning and colony picking which are required for conventional Sanger sequencing as well as by managing to obtain double-stranded DNA itself as initial materials for sequencing. In addition, one instrumental run is capable of determining more than 10 million bases, and that the next generation sequencers are indispensable techniques for genome-wide analyses in terms of cost and time. Furthermore, the capacity of next generation sequencers allows single operator to determine a draft sequence of a bacterial genome within a week. It has become more important that we design research projects to tap the full potential of next generation sequencers. To address this, we have improved our CAGE method to adapt to a next generation sequencer and generated CAGE tags in order to delineate a dynamic transcriptional network. Additionally, we have established a rapid pathogen identification system based on next generation sequencers under collaboration with Research Institute for Microbial Diseases, Osaka University.

Sample preparation for next generation sequencers is different from that for Sanger sequencing technology on several points. For example: 1) double-stranded DNAs, which are to be sequenced, should be flanked with sequencer-specific adaptors at the both ends. 2) the length of double-stranded DNAs should be within proper range to fulfill the potential of sequencer, etc. With dealing with these points, we have adapted our CAGE method to a next generation sequencer by modifying PCR conditions and concatenation conditions. Moreover, we have introduced 5-digit DNA tags (DNA barcodes) to identify RNA sources for monitoring dynamic changes in gene expression with aiming to delineate transcriptional network on cell differentiation. More than one million CAGE tags for each time point were generated with our barcoding method.

In collaboration with Research Institute for Microbial Diseases, Osaka University, we first extracted genomic DNAs from cultured microbial pathogens and determined the sequences. In the case of RNA virus, reverse transcribed genomes were subjected to sequencing. And then, we step forward to establish the rapid identification system with clinical samples.

#### *Staff*

##### *Head*

Dr. Yoshihide HAYASHIZAKI

##### *Members*

Dr. Jun KAWAI

Dr. Piero CARNINCI

Dr. Masayoshi ITOH

Dr. Norihiro MAEDA

Dr. Kazuhiro SHIBATA

Dr. Charles PLESSY

Ms. Harumi URUMA

Ms. Kayoko GOTO

Ms. Yoko KAWASAKI

##### *in collaboration with*

Dr. Harukazu SUZUKI

Dr. Hiroshi ODA

Dr. Takeshi HANAMI

Dr. Rieko OYAMA

##### *Visiting members*

Dr. Kengo Usui

Dr. Fuyu ITO

Dr. Yasumasa MITANI

Mr. Yuki Kawai