SAGE データベース　(SAGE Database)

橋本真一,東京大学大学院医学系研究科･分子予防医学教室
Department of Molecular Preventive Medicine,
Graduate School of Medicine, The University of Tokyo

サマリー

ヒトゲノム配列のほぼ全てが解明されポストゲノム時代を迎えたが、実質的にはゲノム情報の創薬、診断への利用には多くの克服すべき問題を残している。たとえば、ゲノム配列情報は遺伝子の静的な状態を示し、多くの疾患や体調による動的変化を反映していないこと、またこれまでの包括的遺伝子発現解析は、転写開始点の使われ方を無視した方法であるとともに、検出感度に問題があり、ほんの一部の遺伝子発現変化しか捉えていないことなどである。SAGE(Serial Analysis of Gene Expression)法及びここ数年開発されている SAGE 関連の方法は、トランスクリプトーム解析だけでなく、ゲノムの変化やエピジェネティックな変化を測定するのに非常に有用でありこれらの問題を解消できるツールである。本稿では SAGE 法と SAGE 関連技術の簡単な概要を紹介すると共に今後の有用性を紹介する。

はじめに

SAGE 法は DNA チップとともに包括的遺伝子解析法として 1995 年に開発された(*1*)。SAGE 法は、未知、既知にかかわらず遺伝子の発現を何百万という単位で包括的に調べることが可能な方法であり、発現解析データを数値化でき容易にコンピュータ上で比較ができるところを特徴とする。この方法を用い微生物(*2*)、細胞分化(*3, 4*)、薬物処理細胞／組織(*5*)などの多様なライブラリーでの遺伝子発現が報告されている。また、Johns Hopkins 大学のグループを中心として多種の癌組織／細胞の遺伝子解析が SAGE 法により精力的に進められ、癌特異的な抗原の同定、癌化に関与する遺伝子の検索が行われている(*6*)。解析された転写産物はすでに 1,000 万個以上にのぼっておりそのデータは NCBI (http://www.ncbi.nlm.nih.gov/geo/)をはじめとする多くのサイトから公開されている（表1）。また、SAGE 法を用いた miRNA(*7*)、および微量のサンプルでの解析法(*8*)も開発されており、現在、定量的な包括的遺伝子発現解析として SAGE 法が幅広く利用されている。

SAGE 法の概略

SAGE の基本的な原理（図1）は、1分子の mRNA の名札になる部分をシークエンスし、これらの出現頻度と種類を解析するものである。名札になる部分は遺伝子をコードする cDNA の poly A tail に一番近い制限酵素(NlaIII)部位（4塩基認識制限酵素；理論的には cDNA が平均 256bp に一つの割合で切断される）の下流 10〜11bp である（LongSAGE では 16〜17bp、後に説明）。この 10〜11bp の情報があれば、ほとんどの遺伝子は同定可能である。この 10〜11bp を tag (名札)と呼び、この tag をもとに遺伝子を特定し、同じ tag の個数を数えれば発現量がわかる。cDNA ライブラリーのクローンをランダムにシークエンスすれば同様な結果が得られるが、SAGE の場合 tag を数珠つなぎにして、シークエンスを読むので、EST などの方法と比べてシークエンス効率が 30 倍以上良い。また、キャピラリーシークエンサーなどで解析技術が向上したことによって 1〜2 日でも 5 万個以上の tag を解析することも可能となった。

一方、ゲノム情報および EST の情報解析が進むにつれオリジナルの SAGE 法では特定できない遺伝子が出現することがわかってきた。そこで Saha らは酵素を変えることによって 21bp を特定できる LongSAGE 法を開発した(*9*)。LongSAGE の基本的な原理はオリジナルのものとほとんど同じあり、相違点は制限酵素 MmeI を使うことによって CATG の下流 17 塩基の切断が可能となることである。この結果、CATG と合わせて 21bp を同定できることになり、ゲノムの情報からの遺伝子の同定が可能となる（図1）。また、松村ら(*10*)は、制限酵素認識サイトから 25-27 塩基下流で切断する EcoP15I を用いて SuperSAGE 法を開発した。SuperSAGE 法はタグの配列を長くすることによって遺伝子を特定しやすくしている。この情報を利用することでオリジナル SAGE では特定ができなかった遺伝子を明らかにできるだけでなく、コンピュータソフトでは予想されていない発現遺伝子の発見や、遺伝子の発現様式を知ることが可能となる。

SAGE 関連技術

最初の SAGE プロトコールは多くの研究者によってさまざまに改変されている。

1) MPSS(Massively Parallel Signature Sequencing)

Brenner ら(*11*)は、マイクロビーズの表面でタグの配列決定することによりデータ収集を加速する MPSS（大規模並列処理特徴配列決定）法を開発した。MPSS 法は Megaclone 法により polyA(+)RNA から調製した cDNA を個々にマイクロビーズに固相化した後、フローセルの中に充填し、制限酵素で処理、配列認識用アダプターのライゲーションおよび蛍光プローブのハイブリダイゼション操作を繰り返し、蛍光をそのつ

どCCDカメラで撮影してフローセル内のすべてのマイクロビーズ上のcDNAを同時にシーケンスする。この方法により一回のMPSSで20-30万個の配列を得ることが出来る。しかしながらHeneらがLongSAGE法とMPSS法を比較した結果、MPSS法はシークエンスエラーが多いため正確な遺伝子発現が観察出来ないと報告している(*12*)。

2) 5'SAGE

オリジナルSAGE法は3'側の特定の断片を用いて遺伝子を特定するものであり最長27bpの断片を用いることで発現遺伝子のゲノム上の位置を決定することができる。しかしながら5'端の情報は正確でないものが多く、これらのデータから得た遺伝子の機能を明らかにする上で問題となることが明らかになった。それらの問題を克服し、さらに詳細な解析を行う為、転写開始点及び遺伝子発現頻度を観察できる5'-end SAGE (5'SAGE)法(図2)が我々から(*13*)、また同様の手法であるCAGE法が理化学研究所により開発された(*14*)。ヒトゲノムプロジェクトでは遺伝子がEST、完全長cDNAおよび計算上の分析(Genscan、FGENESおよび他のプログラム)から得られた情報に基づいて推定されている。しかしながら、計算上の分析には限界があり、推定された遺伝子が現実に発現したという決定的な証拠はない。mRNA転写開始点の検索は、遺伝子の完全長cDNAの単離だけでなくプロモータ領域の分析において不可欠であり、5'SAGE法は転写開始点を特定するのに非常に有用な方法である。実際にメダカゲノムの遺伝子発現領域として約110万個の5'SAGEタグを解析することによって20,141遺伝子領域が同定された。(*15*)。さらに、カイコ、ハエ、ヒメツリガネゴケなどのゲノム中の遺伝子領域決定にも使用されている。

一方、転写産物の中にはmiRNAを始めとしてタンパク質をコードしていないnon-coding RNAが数多く報告されている。このような転写産物の一部はRNA polymeraseIIによって転写されcap構造やpoly(A)配列が付加されていることから5'SAGE法を利用することで未知のnon-coding RNAの同定が可能であり(*16*)、最近明らかになってきた複雑な遺伝子発現制御機構を理解する上でも非常に有用と思われる。以上の点から5'SAGE法により様々な生物の細胞および組織の詳細なトランスクリプトーム解析が促進されると予想される。この5'SAGEのデータはhttp://5sage.gi.k.u-tokyo.ac.jp/にて公開している(*17*)。

3) GIS (gene identification signature)

上で示した遺伝子の5'endならびにオリジナルSAGEのように3'領域の情報だけでは個々の転写産物の全体像すなわち何処から始まり何処で終わるのかを示してはいない。そこでRuanらは5'-と3'-末端のペアをタグ(paired-end ditag (PET))としてシ

ークエンスする方法、GIS を開発した(*18*)。従って、GIS 法は、異なった長さの転写産物の発現を明らかにすることに対して有用である。実際に幾つかのスプライシングフォームや転写産物同士がフュージョンしたものが報告されている。しかしながらベクターに遺伝子を挿入する操作があるため転写産物の長さに左右され正確な発現頻度を決めるのには適していない。

一方、彼らはこの方法を用いて転写産物だけでなく p53 の結合サイトをクロマチン免疫沈降し、フラグメントの両端をシークエンスすることにより詳細に解析している(*19*)。

4)デジタル核型分析法(Digital Karyotyping)

トランスクリプトーム解析に加えて、ＤＮＡコピー量、DNA メチル化またはヒストン修飾の解析法も SAGE 法を基に開発されている。基本的には全てのこれらの方法は SAGE 法と類似し、2 つの概念に基づいている。1) 短い配列(19-27bp)のタグはゲノムの特定の部位から由来する。2) そして、これらのタグ数とゲノム上の位置により充分な解析ができる。

デジタル核型分析法 (DK) (*20*)は、ゲノムスケールでのＤＮＡコピー量を定量的に解析する技術である (図3) 。また、Polyak らは DNA のメチル化を定量的に測定する方法として DK を改良し MSDK 法(*21*)を開発した。MSDK 法はメチル化感受性の酵素を用いライブラリー間における酵素切断の差異をタグの出現頻度によって測定する方法である。また、Zhao らはクロマチン免疫沈降法と SAGE 法を合わせて GMAT(genome-wide mapping technique)法を開発した(*22*)。これらの方法は、遺伝子発現解析だけでなく、エピジェネティクな変化を全ゲノムレベルで観察することが非常に有用な方法である。現に以下で述べる次世代シークエンサーを使うことによって非常に威力を発揮する(*23*)。

次世代シークエンサーと SAGE 法

近年、数千～数万の発現遺伝子が包括的に解析されているが、特定の細胞や組織における全ての転写産物が明らかになっているわけではない。なぜならば、DNA マイクロアレイの検出限界や SAGE 法にかかるコスト／時間ではゲノムワイドな解析に際し限界があるため低頻度で発現している遺伝子は見落とされることになる。このことにより、特定の細胞、組織同士を比較する時に両者での差が非常に曖昧で再現性のないデータが数多く報告されるという結果になる。

そこに、登場したのが大規模並列処理配列決定法による超高速 DNA シークエンサーである。これらの次世代型シークエンサーは低コストで高速に大量の塩基配列を決定

できることからＤＮＡ塩基配列解析に革命をもたらしたと言っても過言でない。これらは、大量の DNA 断片（20-35 塩基）を高速に塩基配列決定する方法であり、まさに SAGE や SAGE 関連法の研究をさらに加速させるツールである。現在、短い配列を大規模に読む機械として主に Solexa/Illumina 社と SOLiD(Sequencing by oligonucleotide ligation and detection)/ Appliedbiosystems 社がある。Solexa シークエンシングは、まず両端に特異的な配列を付けた DNA をフローセルと呼ばれる基盤に付着させ、この基盤上で DNA を増幅する。可逆化ターミネーターを用いた Sequence-by-Synthesis 法を採用し、クラスター内のテンプレート DNA 塩基配列を読み取る。最近、この方法を用いた、ゲノム全体のヒストンのメチル化部位と遺伝子発現領域の関係の解析、ならびに転写因子の DNA への結合サイトの解析が報告された(*23-25*)。一方、SOLiD システムは、段階的連結反応と呼ばれる独自の技術を利用し、さらに 2-塩基コード化（塩基配列決定の間、エラーに対して二回それぞれの塩基を調べる機構）を特徴とし DNA 塩基配列を読み取る。Solexa/SOLiD システムともに一回のランで 1 ギガベース以上解読することが出来る。

我々は次世代の遺伝子配列解析技術である SOLiD シークエンス法と 5’SAGE 法を組合せ、数百万の転写開始点を含んだ遺伝子発現情報を定量的に観察できる方法(5'end-SOLiD 法)を開発した（投稿中）。脱メチル化剤(5-aza-2’deoxycytidine)及び脱アセチル化阻害剤(trichostatin A)処理した大腸癌細胞株 HT-29 から mRNA を単離し、5'end-SOLiD 法により転写開始点及び発現量について検討した。25 ベースの 5’-end タグをシークエンスし、その中からゲノムにヒットした約 4,000 万個のタグを解析した。解析したタグの 64%が RefSeq cDNA に関連した領域にマッチし、この遺伝子の頻度を、5’SAGE 法にて従来のサンガー法で解析した数十万タグのものと比較したところ 100-1,000 倍感度が高かった。さらに細胞1個あたりの検出感度を測定したところ 0.05 コピー程度という非常に低い発現頻度レベルまでの観測が可能であった。このように、次世代シークエンサーを使用することによって今まで検出できなかった遺伝子を1分子に近いレベルまで検出できると考えられる。

おわりに

今後、今まで培われてきた SAGE 法の技術がこうした次世代の配列解読技術と合わさって、ゲノムワイドな研究を加速させると考えられる。特に今まで曖昧にされてきた低頻度の発現遺伝子、加えて今まで観察出来なかったゲノムワイドでのエピジェネティックな変化を正確に測定することは複雑な発生、癌化、免疫システムを理解する上

で非常に重要である。このような技術革新が今後さらに生命科学分野に大きなインパクトを与えると考えられる。

参考文献

1. Velculescu, V. E. *et al.*, *Science* **270**, 484-487 (1995).
2. Velculescu, V. E. *et al.*, *Cell* **88**, 243-251 (1997).
3. Hashimoto, S. *et al.*, *Blood* **94**, 837-844 (1999).
4. Hashimoto, S. *et al.*, *Blood* **101**, 3509-3513 (2003).
5. Inadera, H. *et al.*, *Biochem Biophys Res Commun* **275**, 108-114 (2000).
6. Velculescu, V. E. *et al.*, *Nat Genet* **23**, 387-388. (1999).
7. Cummins, J. M. *et al.*, *Proc Natl Acad Sci U S A* **103**, 3687-3692 (2006).
8. Datson, N. A. *et al.*, *Nucleic Acids Res* **27**, 1300-1307 (1999).
9. Saha, S. *et al.*, *Nat Biotechnol* **20**, 508-512. (2002).
10. Matsumura, H. *et al.*, *Proc Natl Acad Sci U S A* **100**, 15718-15723 (2003).
11. Brenner, S. *et al.*, *Nat Biotechnol* **18**, 630-634 (2000).
12. Hene, L. *et al.*, *BMC Genomics* **8**, 333 (2007).
13. Hashimoto, S. *et al.*, *Nat Biotechnol* **22**, 1146-1149 (2004).
14. Shiraki, T. *et al.*, *Proc Natl Acad Sci U S A* **100**, 15776-15781 (2003).
15. Kasahara, M. *et al.*, *Nature* **447**, 714-719 (2007).
16. Katayama, S. *et al.*, *Science* **309**, 1564-1566 (2005).
17. Kasai, Y. *et al.*, *Nucleic Acids Res* **33**, D550-552 (2005).
18. Ng, P. *et al.*, *Nat Methods* **2**, 105-111 (2005).
19. Wei, C. L. *et al.*, *Cell* **124**, 207-219 (2006).
20. Wang, T. L. *et al.*, *Proc Natl Acad Sci U S A* **99**, 16156-16161 (2002).
21. Hu, M. *et al.*, *Nat Genet* **37**, 899-905 (2005).
22. Roh, T. Y. *et al.*, *Nat Biotechnol* **22**, 1013-1016 (2004).
23. Barski, A. *et al.*, *Cell* **129**, 823-837 (2007).
24. Johnson, D. S. *et al.*, *Science* **316**, 1497-1502 (2007).
25. Robertson, G. *et al.*, *Nat Methods* **4**, 651-657 (2007).

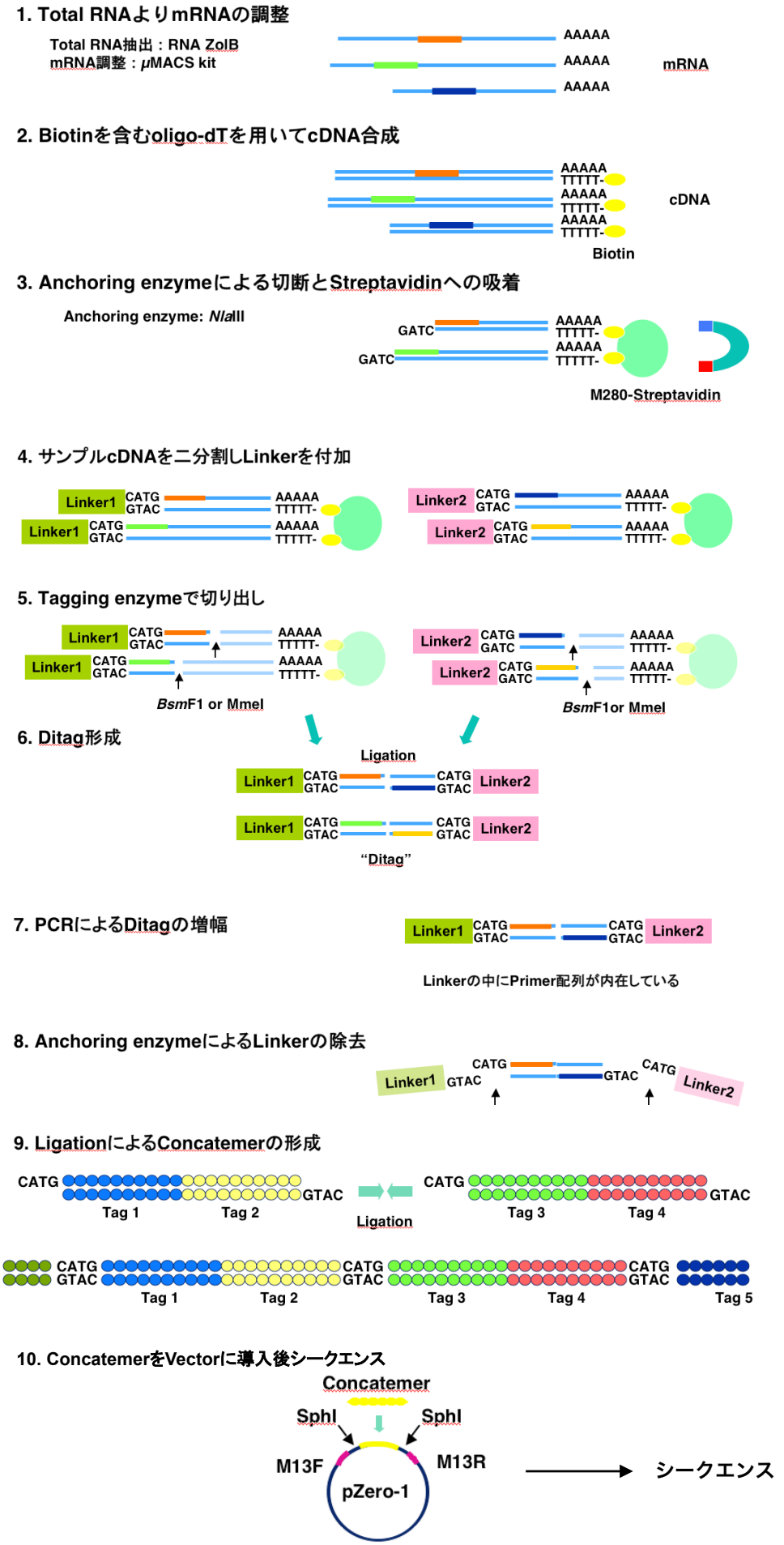
1. Total RNAよりmRNAの調整
Total RNA抽出：RNA ZolB
mRNA調整：μMACS kit
AAAAA
mRNA
2. Biotinを含むoligo-dTを用いてcDNA合成
TTTTT-
cDNA
Biotin
3. Anchoring enzymeによる切断とStreptavidinへの吸着
Anchoring enzyme: NlaIII
GATC
M280-Streptavidin
4. サンプルcDNAを二分割しLinkerを付加
Linker1
Linker2
CATG
GTAC
5. Tagging enzymeで切り出し
BsmF1 or MmeI
6. Ditag形成
Ligation
"Ditag"
7. PCRによるDitagの増幅
Linkerの中にPrimer配列が内在している
8. Anchoring enzymeによるLinkerの除去
9. Ligationによるconcatemerの形成
Tag 1
Tag 2
Tag 3
Tag 4
Tag 5
10. ConcatemerをVectorに導入後シークエンス
Concatemer
SphI
M13F
M13R
pZero-1
シークエンス

図1、 SAGE法の概略

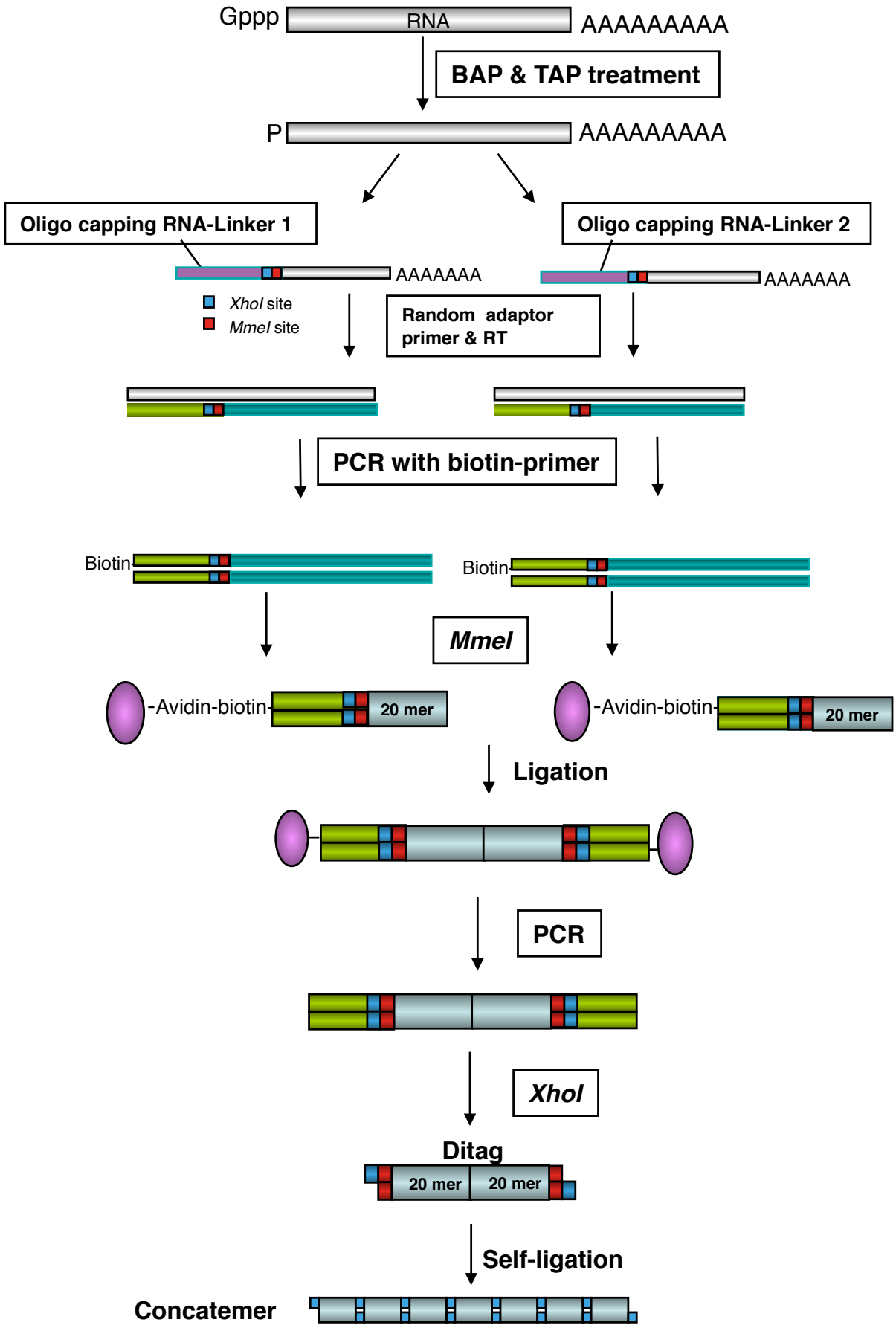

図2、5’SAGE法の概略

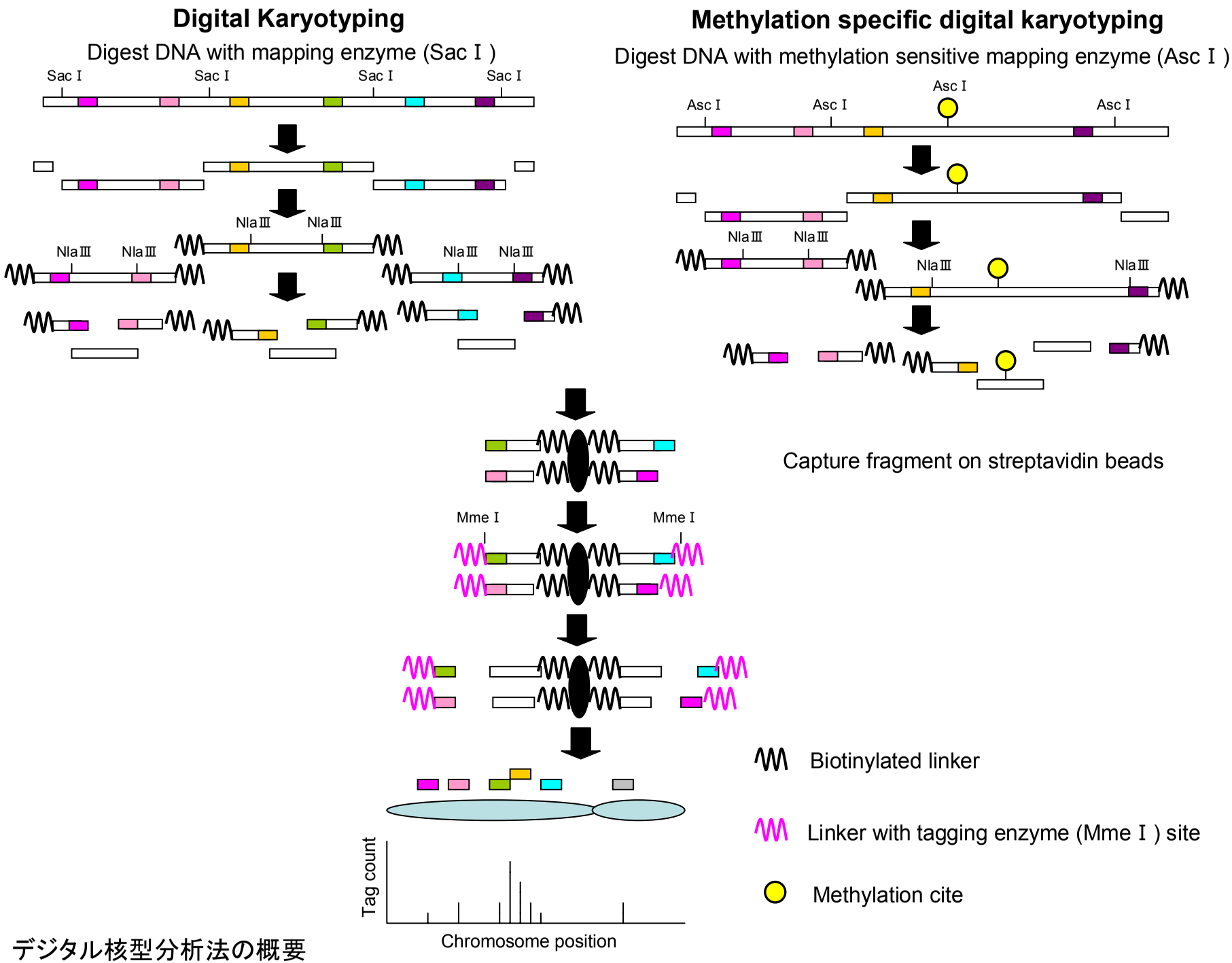

図3　デジタル核型分析法の概要

表1　公開されているSAGE関連のデータベース

データベース名	URL	概要
SAGEnet	http://www.sagenet.org	SAGEを開発したLabのホームページ、大腸癌のデータやプロトコールなどがある。
CGAP	http://cgap.nci.nih.gov	正常細胞、癌細胞を統合的に理解するためのデータベース、遺伝子発現, SNP, RNAi, Chromosome Aberrationsが集まっている。
NCBI SAGEmap	http://www.ncbi.nlm.nih.gov/sage	SAGE tagのデータベースであったがGEOと統合された。
Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo	DNA arrayとSAGEの遺伝子発現情報を集めた統合データベース
SAGE genie	http://cgap.nci.nih.gov/SAGE/	10または17baseのtagのデータベースでありSAGE tagのSNPのデータも含んでいる。ライブラリー同士の比較、また各研究者が単離したtagのゲノム上へのマッピングも可能。
Digital Karyotyping	http://cgap-stage.nci.nih.gov/SAGE/DKViewHome	SAGEベースの技術による解析されたゲノムの増幅と染色体欠失の配列データベース
MD Anderson SAGE site	http://sciencepark.mdanderson.org/ggeg/default.html	ヒト癌細胞とマウスの癌モデルの遺伝子発現データベース
5'SAGE	http://5sage.gi.k.u-tokyo.ac.jp/	5' end 情報と遺伝子発現のデータベース
CAGE	http://gerg01.gsc.riken.jp/cage/	5' end 情報と遺伝子発現のデータベース
Blood SAGE	http://bloodsage.gi.k.u-tokyo.ac.jp/	ヒト血液系における遺伝子の SAGE tag 解析サーバー