ISSN 0006-2979, Biochemistry (Moscow), 2007, Vol. 72, No. 11, pp. 1179-1186. © Pleiades Publishing, Ltd., 2007.
Original Russian Text © I. V. Gainetdinov, T. L. Azhikina, E. D. Sverdlov, 2007, published in Biokhimiya, 2007, Vol. 72, No. 11, pp. 1449-1458.

REVIEW

Use of Short Representative Sequences for Structural and Functional Genomic Studies

I. V. Gainetdinov*, T. L. Azhikina, and E. D. Sverdlov

Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences,
ul. Miklukho-Maklaya 16/10, 117997 Moscow, Russia; fax: (495) 330-6538; E-mail: ildargv@mail.ru

Received May 15, 2007
Revision received July 9, 2007

Abstract—Existing approaches to direct genomic studies are costly and time-consuming. To overcome these problems, a series of tag-based methods utilizing short fragments uniquely representing full-length transcripts/genes from which they originate has been developed. This review summarizes basic principles underlying these methods and their numerous modifications designed for studying transcriptome profiles, searching for unidentified expressed loci, characterization of promoter regions, and high-throughput mapping of various genomic sites, such as hypo- and hypermethylated CpGs, and chromatin-binding and DNase I cleavage sites.

DOI: 10.1134/S0006297907110028

Key words: functional genomics, transcriptome, tag, SAGE

The growing need for simultaneous examination of completed assemblies of genes, transcripts, and protein products of particular organisms led to diversified methodical approaches to this problem [1]. The goal of these approaches is to extract maximum information concerning functional activity of genome loci (gene annotation), their variability, as well as expression profiling and gene regulation.

All methods for massively-parallel sample analysis can be divided into two main groups. The first comprises **methods based on hybridization** including hybridization-on-chip [2, 3] and subtractive hybridization [4, 5]. Common drawbacks of these indirect methods are nonspecificity and kinetic features that do not allow reliable analysis of small-number RNA species in transcriptomes [6].

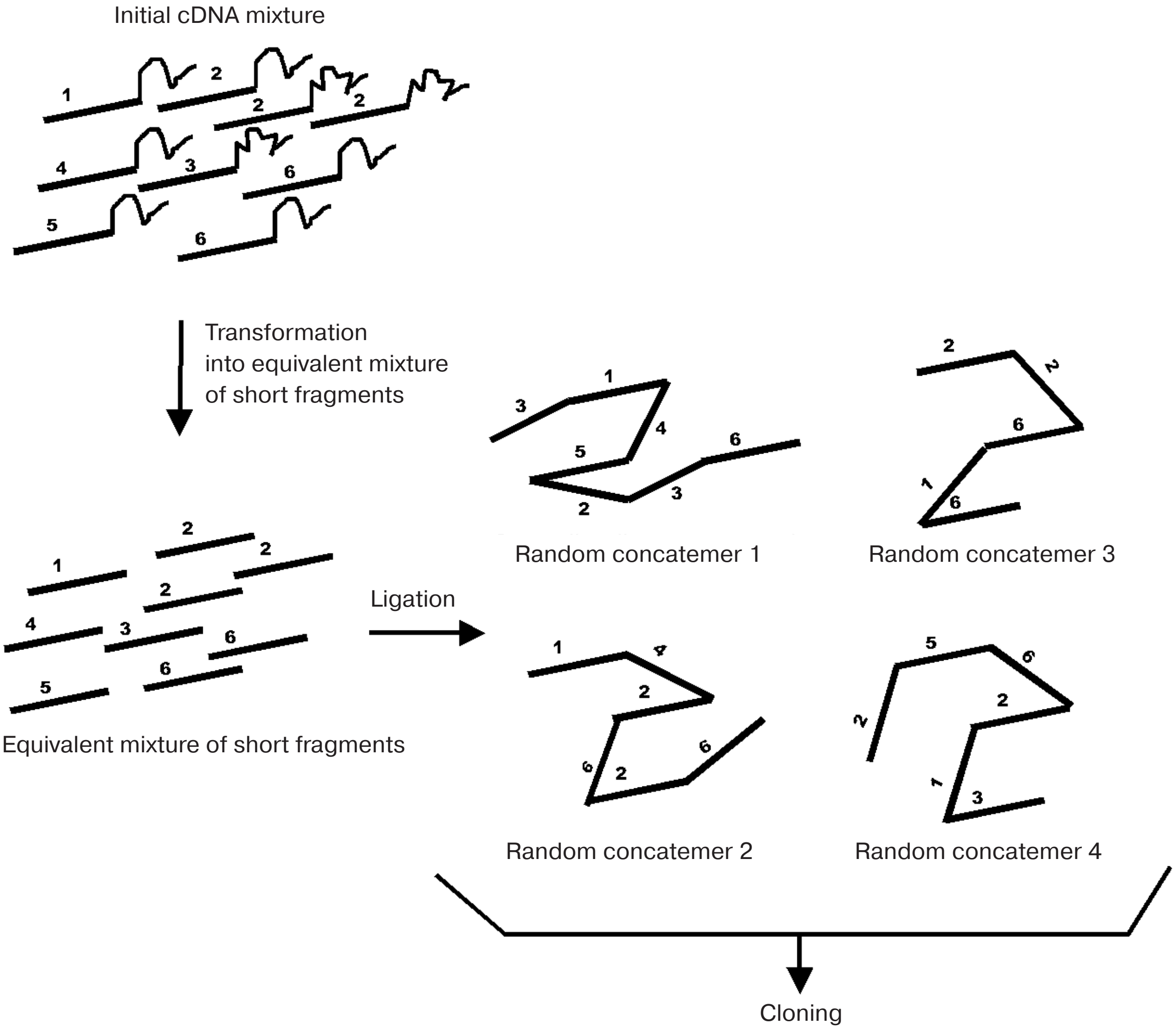
In this aspect, the methods of **direct comparison of genomes and transcriptomes** are theoretically perfect, but this comparison is extremely labor-intensive. Therefore, recent rapid development of new methods that are ever faster and cheaper promises to revolutionize such studies [7]. These methods utilize short sequences (tags) instead of whole genes or their transcripts, which definitely determine the gene/transcript from which they derive. The substitution of whole genes/transcripts by such short representative sequences allows a significantly cheaper sequencing with acquisition of required data on functional characteristics of these genes.

The problem in analyzing of such fragments is to find the least expensive way for their sequencing without isolation of each of them from the mixture. To solve this problem, an elegant method was offered (Scheme 1) based on the cross-ligation of representative fragments to produce their arbitrary heteropolymers (concatemers). Integration of many short nucleotide segments into a concatemer allows simultaneous sequencing of many short fragments. Concatenation is of fundamental importance for lowering costs and time of experiment because it allows simultaneous exploration of 10-50 times more individual genes/transcripts.

Following the cloning and subsequent sequencing of concatemers, a computer-based extraction of individual representative fragments and finding coincidences with nucleotide sequences of the genes/transcripts, which these fragments represent, should be carried out. Simultaneous preparation of a multitude of fragments reveals both qualitative composition of transcriptome/genome and quantitative balance between its individual components. Representation of a particular transcript/gene in a general mixture will directly determine the

Abbreviations: bp) base pair(s); RFLP) restriction fragment length polymorphism.

* To whom correspondence should be addressed.



Utilization of short representative fragments (demonstrated with cDNA). The initial mixture is treated in such a way as to prepare representative fragments uniquely representing each cDNA. Then, to lower costs of sequencing, these fragments are ligated with each other to obtain heteropolymers (concatemers) that are further cloned. The sequencing of concatemers and extracting information on the representation of each of the representative fragments enables qualitative and quantitative evaluation of cDNA in the initial mixture. 1-6) Fragments differing in nucleotide sequence

Scheme 1

number of representative fragments obtained from the experiment.

This kind of idea allows direct comparison of sequences without hybridization and avoidance of hybridization-derived bias for estimation of representation of the material in the initial genomes/transcriptomes.

SAGE AND ITS MODIFICATIONS USED FOR TRANSCRIPTOME CHARACTERIZATION

A method based on this principle was first applied by Velculescu et al. in 1995 [8]. It was named SAGE (serial analysis of gene expression) and used for studying of gene transcription profile. The approach is based on substitution of a cDNA set representing the transcriptome of a particular sample by an equivalent set of representative 14-nucleotide sequences, each derived from the unique site of distinct cDNA (Scheme 2). A 14-nucleotide sequence, if it is not a fragment of a repetitive element of the transcriptome, with high probability occurs in it only once, namely, in that cDNA from which it derives. So, detection of distinct sequence in the mixture of prepared oligomers, first, is indicative of the presence of the initial cDNA in the transcriptome and, second, allows quantitative evaluation of this cDNA in the initial mixture.

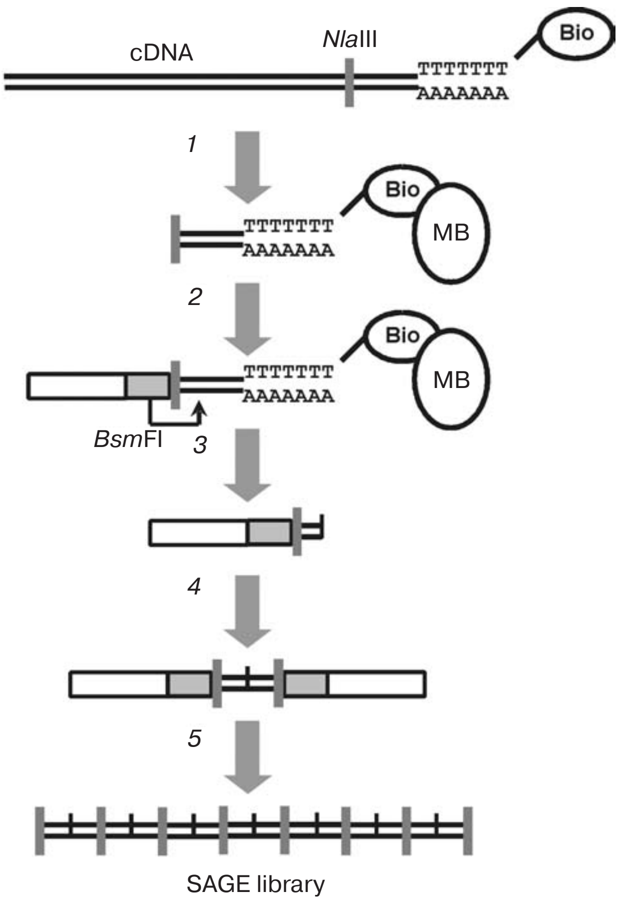
The problem of preparation of such sequences could only be solved with the discovery of specific type IIS endonucleases, which hydrolyze DNA at a strict distance from the recognition site. The recognition site-containing oligonucleotide adapters are used for its addition to cDNA sequence. Ligation is conducted so that the introduced site is localized near a unique cDNA sequence. To do this, cDNAs were co-synthesized with biotinylated oligothymidine adapters and hydrolyzed with *Nla*III endonuclease into short fragments. Then, 3′-terminal fragments were isolated on streptavidin-coated magnetic beads. This procedure resulted in a set of 3′-terminal cDNA fragments bounded by the *Nla*III recognition site on their 5′-ends. Following ligation with oligonucleotide adapter containing the recognition site for type IIS endonuclease (*Bsm*FI, GGGAC$(N)_{10}\downarrow$) and restriction with this enzyme, a set of 14-bp fragments representing unique cDNA sites was obtained.

Unlike routine experiments with microchips, in which the number of analyzed transcripts is restricted to the number of probes placed on a substrate and specific to already known genes and reading frames, SAGE enables detection of all transcripts, thus allowing more reliable quantitative data, because this method depends on a number directly characterizing the transcript representation, rather than on readout (for example, fluorescence intensity).

Nonetheless, the abundance of homologous and repeated sequences in genomes and transcriptomes means that several loci can contain representative fragments with the same primary structure. This prevents their individual estimation. Thus, the mapping of short 14-bp sequences is often troublesome. One solution for this problem is utilization of fragments longer than 14 bp. This can be achieved by using other type IIS endonucleases. Use of *Mme*I endonuclease (TCCRAC$(N)_{20}\downarrow$) instead of *Bsm*FI in the modified method LongSAGE lengthens the representative fragments from 14 to 21 bp and thereby reduces their mapping ambiguity by more than an order of magnitude (the theoretical maximum number of transcribed loci corresponding to one representative fragment is decreased from 279 to 15) [9].

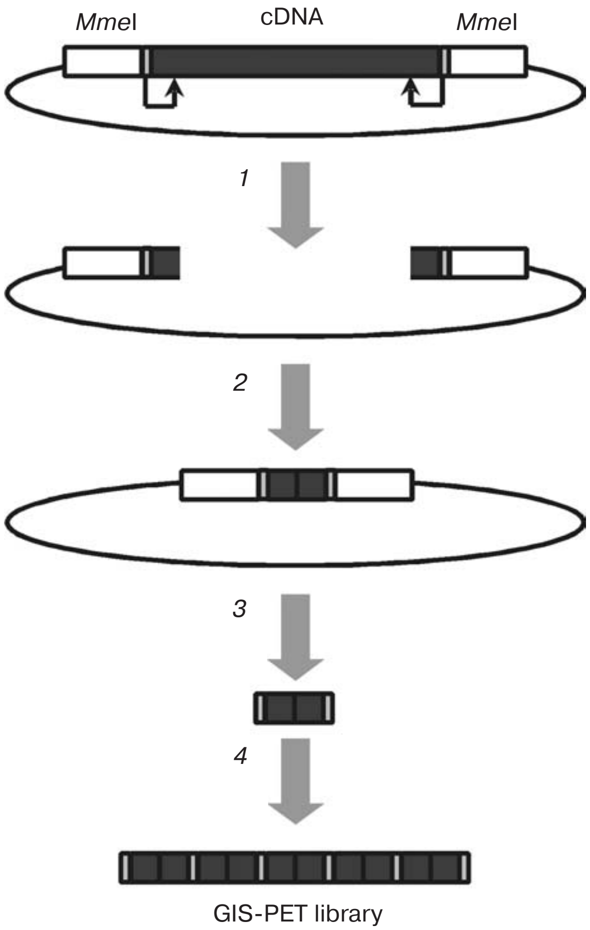
A detailed comparison between the data of LongSAGE and SAGE made by several research groups demonstrated that both datasets can virtually equally represent the initial cDNA sampling. In particular, the set of 21-bp fragments from LongSAGE, which were computationally reduced to 14-bp ones, was found to be virtually equal to the real one obtained from the standard SAGE experiment. However, only 4-7% of ambiguously mapped 14-bp sequences can be univocally linked to the represented cDNA due to use of LongSAGE. Nonetheless, despite reports on better effectiveness of standard SAGE in studies on differential expression of various transcripts, the data from LongSAGE can be considered as more reliable. This is due to increased mapping accuracy of 21-bp representative sequences [9].

Quite recently the type III endonuclease *Eco*P15I (CAGCAG$(N)_{25}\downarrow$) was introduced allowing production of 26-bp tags (SuperSAGE) [10, 11]. These long fragments can also be used as oligonucleotides on microchips. This combines advantages of two platforms: repeated use of microchips and analysis of entire transcriptome rather than its previously annotated part, because oligonucleotides derive directly from mRNA, not from previous synthesis. However, one has to remember that elongation of representative fragments leads to decrease in their population on sequencing of one concatemer. By this, it means that LongSAGE and SuperSAGE, as compared with SAGE, require 1.5 and 2 times more sequencing reactions, respectively, for production of an equal number of tags. In turn, decreased sequencing effectiveness results in the increase in expenditure in comparison with that of the initial protocol.



Preparation of short representative fragments from full-length cDNAs according to the SAGE protocol. *1*) Digestion of double-stranded cDNA with the tagging restrictase *Nla*III followed by isolation of biotinylated (Bio) 3′-fragments using streptavidin-coated magnetic beads (MB). *2*) Ligation of the fragments with the oligonucleotide adapter containing the recognition site of type IIS endonuclease (*Bsm*FI). *3*) Hydrolysis with *Bsm*FI to form 14-bp fragments. *4*) Subsequent ligation to form dimers of representative fragments. *5*) Removal of adapters by hydrolysis and ligation of the dimers to form concatemers

Scheme 2



The GIS-PET protocol. *1*) Full-length cDNA inserted in a special vector containing the *Mme*I recognition site is subjected to hydrolysis with this enzyme. *2*) Ligation of hydrolysis products leads to formation of a dimer composed of 5′- and 3′-terminal parts of the transcript. *3*, *4*) Dimers are released and concatenated with subsequent cloning

Scheme 3

Thus, the choice between SAGE and LongSAGE depends on tasks of analysis. One of such problems is accurate annotation of expressing genome areas. SAGE is not suitable for its solution, because it does not provide information on position of transcription initiation and termination sites. However, the principle of extraction of short representative fragments followed by their concatenation can be successfully utilized for cutting the cost and time of experiment, as it was realized by Ng and coworkers, who developed the method GIS (gene identification signature) [12, 13]. GIS combines capabilities of SAGE and its numerous modifications by simultaneous preparation of 5′- and 3′-terminal mRNA segments, thus allowing mapping to genome sequences to demarcate the transcription initiation and termination boundaries (Scheme 3). Insertion of full-length cDNA into a vector containing the recognition sequence of the type IIS endonuclease (*Mme*I) in ligation sites allows preparation (after hydrolysis with this enzyme) of the representative tags of both ends of the transcript. The subsequent self-circularization of the vector leads to formation of sequences (PET, paired-end ditag) containing both 5′- and 3′-terminal tags, and subsequent concatenation of PET significantly improves sequencing efficiency and cost-effectiveness of analysis. Thus, the given method enables not only acquisition of the data on transcript representation, but also determination of its initiation and termination sites. This allows annotation of previously unidentified mRNAs.

GIS analysis of mouse embryonic stem cells identified hundreds of previously uncharacterized mRNAs, including several intergenically spliced transcripts [14]. Further, software was developed for automation of concatemer sequencing data procession, removal of possible artifact sequences, mapping of found sequences to the genome, and database management. One of the important features of this software is the possibility of data processing via the Internet.

OTHER APPROACHES BASED ON USE OF SHORT REPRESENTATIVE FRAGMENTS (454 SEQUENCING, MS-PET, MPSS)

A novel method for primary DNA structure analysis, which has been developed in 454 Life Sciences (USA) and named 454 sequencing^{TM}, revolutionizes tag-based studies of transcriptomes [15, 16]. This sequencing method significantly outperforms the traditional Sanger sequencing (based on the state-of-the-art solutions) in number of simultaneously analyzed samples (~300,000 and ~400, respectively), but at a cost of substantially shorter reads (~100 and ~500 bp, respectively).

According to the original protocol, the DNA fragments, 300-500 bp in length, are ligated with oligonucleotide adapters, and non-biotinylated single-stranded chains were separated by biotin–avidin interaction, bound with excess of DNA-affinity beads, and encapsulated into droplets of water-in-oil emulsion. Then PCR with primers identical to the initial adapters is conducted in these droplets, or “microfines”, which serve as peculiar microreactors, to produce a set of amplified sequences, each associated with an individual bead. The single-stranded DNA-carrying beads are deposited into wells of a fiber-optic slide. Each well is 44 μm in diameter and contains no more than one bead. Subsequent sequencing is conducted in these wells combined in a flow system: nucleoside triphosphates are delivered in determinate order one after another, and added polymerase complements the primer identical to the initial oligonucleotide adapter. This reaction is accompanied by release of pyrophosphate with chemical modification of luciferin into its oxidized form (pyrosequencing) and light emission. Detection of luminescence allows simultaneous short-read (~100 bp) of primary structures of hundreds of thousands of DNA fragments.

These characteristics make the offered sequencing method ideal in approaches based on use of short representative sequences, because shortening of read-length in each reaction is significantly compensated by the number of reactions. In this case, concatenation is not necessary because of the multiplicity of simultaneous sequencing reactions.

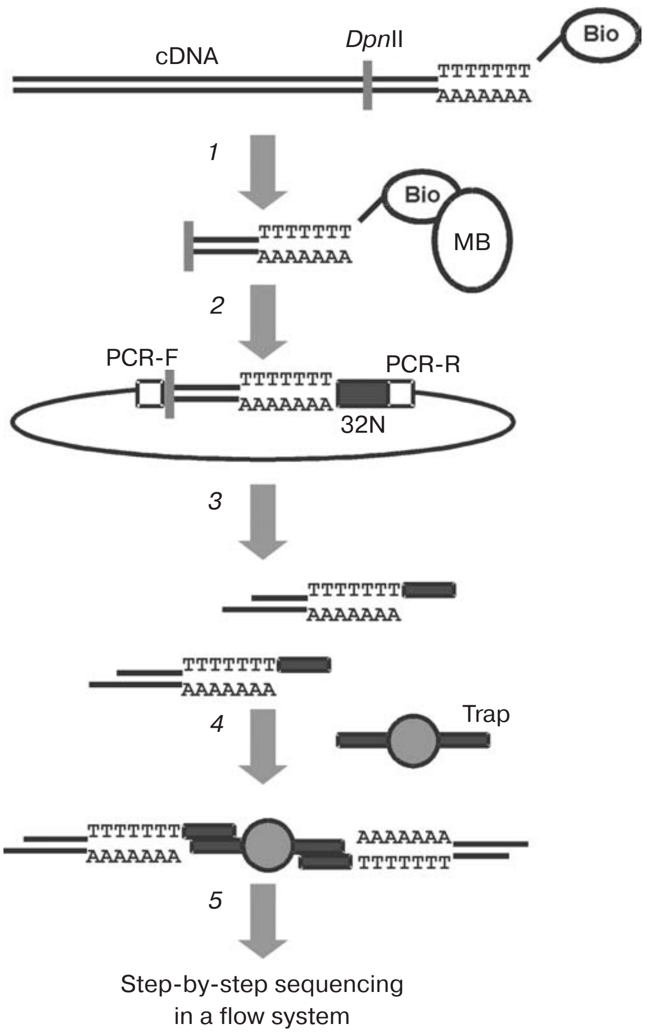
Thus direct cDNA sequencing with the use of technology promoted by 454 Life Sciences enables preparation of libraries containing up to ~300,000 5′-terminal transcriptional fragments, ~100 bp in length, which adequately characterize the transcriptome analyzed. In particular, use of this protocol for investigation of the *Medicago truncatula* transcriptome revealed many new mRNAs [17].

This sequencing principle was used in the development of a GIS modification named MS-PET: short paired-end ditags (PET) formed after ligation and containing 5′- and 3′-terminal tag pairs were subjected to direct sequencing (omitting concatenation) according to the 454-sequencing^{TM} method. An additional stage of PET dimerization allows doubling of MS-PET efficiency. Eventually, the total number of produced PETs increases 100-fold in comparison with their number in GIS [16].

The MS-PET conception can be used in virtually all fields associated with routine mapping of functional loci to genome, including identification of epigenetic elements (ChIP-PET, see below) and genome rearrangements.

The principle that a transcriptome can be represented as a set of tags has become yet more popular with development of the MPSS (massively parallel signature sequencing) strategy [18, 19]. Acquisition of 17-bp representative sequence, which flanks the 3′-terminal hydrolytic site of *Dpn*I, is achieved by direct sequencing of this fragment on cDNA, rather than by utilization of type IIS endonuclease, and without additional concatenation. The sequencing occurs through hybridization of 4-bp "decoding" adapters containing a unique fluorescence label with single-stranded end of mRNA and subsequent hydrolysis releasing the next four nucleotides. So, four such stages are enough for determination of 17-bp structure (including the nucleotide of the *Dpn*I recognition site).

Another key principle is cDNA amplification foregoing the sequencing (Scheme 4). Right after reverse transcription, the cDNAs prepared are ligated with degenerate adapter (random 32-bp sequence). The aggregate diversity of adapter sequences (16.8 million variants) significantly exceeds that of a common transcriptome. This makes us feel certain that every transcript gets a unique label (adapter). Then the ligation product is amplified and hybridized with microbeads, each carrying covalently bound fragments complementary to one of the initial adapters. Following step-by-step determination of primary structure is conducted in a specially designed flow system, which sequentially operates with each bead. This avoids distortion of data on representation of each individual mRNA and provides reliable information on complexity of transcriptome including that on representation of low-copy transcripts.



MPSS protocol. *1*) Biotinylated (Bio) cDNA is hydrolyzed at the *Dpn*II recognition site, and 3′-fragments are isolated using streptavidin-coated magnetic beads (MB). *2*) These fragments are ligated into a special plasmid carrying the random 32-bp adapter (32N) and two sequences identical to primers used in subsequent PCR (PCR-F and PCR-R). *3*) The amplification product is hydrolyzed with endonuclease to obtain single-stranded outstanding termini. *4*) The product thus prepared is hybridized with microbeads (trap), each of them carrying covalently-bound fragments, which are complementary to one of the initial 32N adapters. The final product is microbead with ~100,000 covalently-bound identical molecules. *5*) Each of them is separately sequenced in a flow system. This avoids data bias upon examination of even small amounts of represented transcripts

Scheme 4

Utilization of MPSS for exploration of the *Arabidopsis thaliana* transcriptome revealed 6700 new transcripts in addition to 18,000 sense and ~7000 antisense transcripts described earlier. Similar results were obtained from exploration of transcriptomes of other simple organisms, such as *Alexandrium fundyense* [20-22]. Human gene expression maps created using MPSS for 32 normal tissues have updated the information on transcription of human genes. Evaluation of the data has confirmed a hypothesis that most difference between cells is due to differential transcription of a small number of genes [23]. Not any less impressive are the results of the search for genes specifically expressed in cancer cells [24].

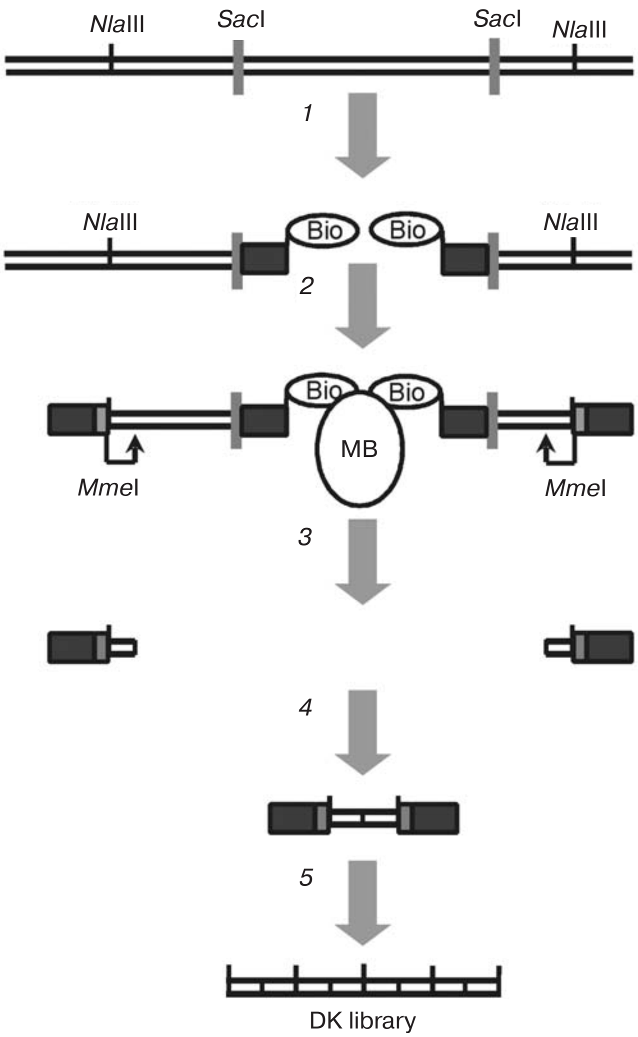
SHORT REPRESENTATIVE FRAGMENTS-BASED MAPPING TECHNOLOGIES FOR OTHER GENOME FUNCTIONAL ELEMENTS

The SAGE method is applicable not only for the study of transcriptomes, but also any functional genome sequences or those important for diagnostics, such as RFLP (restriction fragment length polymorphism) sites, hypo- and hypermethylated regions, as well as sites of chromatin–protein binding. As in the case of SAGE, it is necessary to ligate an oligonucleotide adapter containing endonuclease IIS restriction site to these important functional loci for the formation of short representative fragments. Subsequent hydrolysis by this enzyme leads to the release of short segments flanking and representing each of these loci. Random concatenation of the resulting representative fragments, cloning, and sequencing of concatemers enable mapping to genome of each functional locus under study. Obviously, the stage of concatenation is also important in the given case for substantial saving of time and cost cutting.

At present, the Digital Karyotyping approach is applied for the study of insertion–deletion polymorphism (Scheme 5) [25, 26]. Genome DNA hydrolyzed by one of the widely used in RFLP analysis endonucleases is used as an initial sample. Subsequent ligation of adapters containing the *Mme*I recognition site to genome fragments enables preparation of 21-bp tags used for further concatenation. Comparison of representation of a given tag after sequencing of concatemers provides information on amplification or deletion of the corresponding genome locus. The utilization of methyl-sensitive enzyme as the initial endonuclease makes it possible to study the methylation extent of recognition sites for this enzyme within the whole genome (MSDK, methylation-specific digital karyotyping) [27].

One modification of the GIS method named ChIP-PET (chromatin immunoprecipitation-PET) is designed for mapping of genome loci saturated with sites of transcription factors or other proteins revealed via immunoprecipitation of chromatin [28].

Certain difficulties can be met when studying genome methylation: the general number of CpG-dinucleotides in a genome is great, incommensurably higher than the number of all possible variants of transcripts. Thus, the volume of initial sampling of sequences under study becomes too great leading finally to insufficient length of representative fragments for their univocal mapping along the whole genome. In connection with this, a principal necessity arises for artificial decrease in the initial set of methylation sites under study. The RIDGES method (rapid identification of genomic DNA splits) was developed for the study of prolonged hypomethylated genome regions [29]. It comprises ligation of adapter



Protocol of Digital Karyotyping (DK). *1*) Isolation of genome DNA is followed by its hydrolysis with RFLP endonuclease (*Sac*I) and ligation to biotinylated adapter (Bio). *2*) The ligated fragments are extracted using streptavidin-coated magnetic beads (MB), then the fragments are subjected to enzymatic hydrolysis with *Nla*III splitting DNA into short fragments, and ligation with an adapter containing the recognition site for type IIS endonuclease (*Mme*I). After the hydrolysis with *Mme*I (*3*) and formation of ditags (*4*), they are released from the adapter sequences and ligated to form concatemers (*5*)

Scheme 5

sequences containing the *Mme*I recognition site to the termini of fragments formed after the hydrolysis of the genome DNA with methyl-sensitive endonuclease. Subsequent stages coincide with those in the SAGE method and enable production of concatemers composed of representative tags neighbored with non-methylated CpG-dinucleotides. Preliminary extraction of sequences belonging to anticipatorily determined genome region (from hundreds of thousands to millions of base pairs) from the hydrolysis products of genome DNA comprises the main characteristic feature of the RIDGES method. This is realized via use of the coincidence cloning principle: the genome DNA under study and a preliminarily cloned limited genome region, both hydrolyzed by methyl-sensitive endonuclease, are used as two samples [30]. This enables focusing attention of investigators on the loci of interest.

Methylation-specific primer (MSP)-SAGE is another method utilizing the SAGE principle for the study of methylated sequences. In this case, representative genome tags are derived from a special methyl-specific primer by its completing enabling production of a genome segment containing methylated CpG-dinucleotide. Then 17-bp sequences, which are subsequently concatenated, are released from these fragments by their hydrolysis with type IIS endonuclease. As in SAGE, the sequence of concatemers gives information on the localization of CpG-sites and on the extent of their methylation, which is proportional to the relative portion of corresponding tags in the obtained nucleotide sequences [31].

Finally, the application of type IIB endonucleases hydrolyzing DNA from both 5′- and 3′-termini from the recognition site was proposed for the study of genomes [32]. The obtained fragments of 21-33 bp in length are usually purified in gel and cloned, and, following purification of the plasmid DNA, these fragments are cut out by corresponding restrictase and concatenated. The average distance between neighboring recognition sites for this type endonucleases varies from 500 to 8000 bp. This enables analysis of insertion–deletion and single-nucleotide genome polymorphism. The utilization of methyl-sensitive enzymes also broadens the field of application of type IIB endonucleases.

In view of the great variety of polygenome methods, the selection of a distinct approach plays the key role. As mentioned above, hybridization approaches have limitations associated with characteristic features of the hybridization process (for instance, do not give information on low-represented mRNAs).

In connection with this, the methods of direct genome and transcriptome sequence determination are preferred, which can be significantly simplified by using short fragments unambiguously representing them, instead of full-length sequences (genome or mRNA). Using these representative sequences or tags, one can produce desired functional information at substantially lowered cost for sequence determinations.

It is impossible to choose one or several most effective protocols, because each of them has its own virtues and shortcomings [33]. The elevation of reliability of the results requires a confirmation by alternative methods and integration of data obtained by different methods [34-36]. Approaches based on use of representative tags are applicable for study of transcriptomes, as well as any sites of specific cleavage of a genome: recognition sites for methyl-sensitive (hypo- and hypermethylated regions) or methyl-insensitive endonucleases (RFLP sites), sites of chromatin–protein binding, sites of DNase I cleavage, as well as sites that can be cleft under the action of chemical or physical factors. This principle can be integrated into any technique, when a necessity arises, for routine determination of multiple functionally important loci along the genome length.

This study was supported by the Program supporting the Scientific Schools (project 2006.2003.4) and Programs on Molecular and Cellular Biology of the Presidium of the Russian Academy of Sciences.

REFERENCES

1. Marti, J., Piquemal, D., Manchon, L., and Commes, T. (2002) *J. Soc. Biol.*, **196**, 303-307.
2. Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999) *Nat. Genet.*, **21**, 10-14.
3. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999) *Nat. Genet.*, **21**, 20-24.
4. Ermolaeva, O. D., and Sverdlov, E. D. (1996) *Genet. Anal.*, **13**, 49-58.
5. Diatchenko, L., Lau, Y. F., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E. D., et al. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 6025-6030.
6. Qin, L. X., Beyer, R. P., Hudson, F. N., Linford, N. J., Morris, D. E., and Kerr, K. F. (2006) *BMC Bioinform.*, **7**, 23.
7. Harbers, M., and Carninci, P. (2005) *Nat. Meth.*, **2**, 495-502.
8. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995) *Science*, **270**, 484-487.
9. Li, Y. J., Xu, P., Qin, X., Schmechel, D. E., Hulette, C. M., Haines, J. L., Pericak-Vance, M. A., and Gilbert, J. R. (2006) *BMC Bioinform.*, **7**, 504.
10. Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D. H., and Terauchi, R. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 15718-15723.
11. Matsumura, H., Bin Nasir, K. H., Yoshida, K., Ito, A., Kahl, G., Kruger, D. H., and Terauchi, R. (2006) *Nat. Meth.*, **3**, 469-474.
12. Peters, B. A., and Velculescu, V. E. (2005) *Nat. Meth.*, **2**, 93-94.
13. Ng, P., Wei, C. L., Sung, W. K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., et al. (2005) *Nat. Meth.*, **2**, 105-111.

14. Chiu, K. P., Wong, C. H., Chen, Q., Ariyaratne, P., Ooi, H. S., Wei, C. L., Sung, W. K., and Ruan, Y. (2006) *BMC Bioinform.*, **7**, 390.

15. Dunn, J. J., McCorkle, S. R., Everett, L., and Anderson, C. W. (2007) *Genet. Eng. (N. Y.)*, **28**, 159-173.

16. Ng, P., Tan, J. J., Ooi, H. S., Lee, Y. L., Chiu, K. P., Fullwood, M. J., Srinivasan, K. G., Perbost, C., Du, L., Sung, W. K., et al. (2006) *Nucleic Acids Res.*, **34**, e84.

17. Cheung, F., Haas, B. J., Goldberg, S. M., May, G. D., Xiao, Y., and Town, C. D. (2006) *BMC Genomics*, **7**, 272.

18. Reinartz, J., Bruyns, E., Lin, J. Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik, R. (2002) *Brief Funct. Genom. Proteom.*, **1**, 95-104.

19. Zhou, D., Rao, M. S., Walker, R., Khrebtukova, I., Haudenschild, C. D., Miura, T., Decola, S., Vermaas, E., Moon, K., and Vasicek, T. J. (2006) *Meth. Mol. Biol.*, **331**, 285-311.

20. Dean, J. F. (2004) *Nat. Biotechnol.*, **22**, 961-962.

21. Meyers, B. C., Tej, S. S., Vu, T. H., Haudenschild, C. D., Agrawal, V., Edberg, S. B., Ghazal, H., and Decola, S. (2004) *Genome Res.*, **14**, 1641-1653.

22. Meyers, B. C., Vu, T. H., Tej, S. S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., and Haudenschild, C. D. (2004) *Nat. Biotechnol.*, **22**, 1006-1011.

23. Erdner, D. L., and Anderson, D. M. (2006) *BMC Genom.*, **7**, 88.

24. Chen, Y. T., Scanlan, M. J., Venditti, C. A., Chua, R., Theiler, G., Stevenson, B. J., Iseli, C., Gure, A. O., Vasicek, T., Strausberg, R. L., et al. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 7940-7945.

25. Dunn, J. J., McCorkle, S. R., Praissman, L. A., Hind, G., van der Lelie, D., Bahou, W. F., Gnatenko, D. V., and Krause, M. K. (2002) *Genome Res.*, **12**, 1756-1765.

26. Wang, T. L., Maierhofer, C., Speicher, M. R., Lengauer, C., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 16156-16161.

27. Hu, M., Yao, J., and Polyak, K. (2006) *Nat. Protoc.*, **1**, 1406-1411.

28. Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., et al. (2006) *Cell*, **124**, 207-219.

29. Azhikina, T., Gainetdinov, I., Skvortsova, Y., and Sverdlov, E. (2006) *Mol. Genet. Genom.*, **275**, 615-622.

30. Azhikina, T., Gainetdinov, I., Skvortsova, Y., Batrak, A., Dmitrieva, N., and Sverdlov, E. (2004) *Mol. Genet. Genom.*, **271**, 22-32.

31. Wang, X., Zhang, C., Zhang, L., and Xu, S. (2006) *Biochem. Biophys. Res. Commun.*, **341**, 749-754.

32. Tengs, T., LaFramboise, T., Den, R. B., Hayes, D. N., Zhang, J., DebRoy, S., Gentleman, R. C., O'Neill, K., Birren, B., and Meyerson, M. (2004) *Nucleic Acids Res.*, **32**, e121.

33. Van Ruissen, F., Ruijter, J. M., Schaaf, G. J., Asgharnegad, L., Zwijnenburg, D. A., Kool, M., and Baas, F. (2005) *BMC Genom.*, **6**, 91.

34. Wang, S. M. (2007) *Trends Genet.*, **23**, 42-50.

35. Gowda, M., Venu, R. C., Raghupathy, M. B., Nobuta, K., Li, H., Wing, R., Stahlberg, E., Couglan, S., Haudenschild, C. D., Dean, R., et al. (2006) *BMC Genom.*, **7**, 310.

36. Oudes, A. J., Roach, J. C., Walashek, L. S., Eichner, L. J., True, L. D., Vessella, R. L., and Liu, A. Y. (2005) *BMC Cancer*, **5**, 86.