European Journal of Human Genetics (1999) 7, 510–512

http://www.stockton-press.co.uk/ejhg

MEETING REPORT

A tale of tags: report on a HUGO/EU SAGE Workshop, 29 January–1 February 1999, Hilversum, The Netherlands

Frank Baas[1] and Henk F. Tabak[2]

[1]*Department of Neurology, Academic Medical Center, 1105 AZ Amsterdam, The Netherlands*
[2]*Department of Biochemistry, Academic Medical Center, 1105 AZ Amsterdam, The Netherlands*

Four years after the publication of a new genome-wide technique to measure mRNA levels, serial analysis of gene expression (SAGE), by Victor Velculescu and co-workers, the first international SAGE workshop was held in Hilversum, The Netherlands. The initiative for this workshop came from the Human Genome Organisation, HUGO, with a BIOMED grant, number BMH4-CT97-2031. The aim of the workshop was to exchange experience between laboratories which are now actively putting SAGE into practice. The workshop, which hosted 50 participants, was doubly oversubscribed, showing the wide interest in SAGE. Participants from all over the world, for example Argentina, Australia, USA and many European countries, visited the workshop. Both academic and industrial groups were present and exchanged experiences in a friendly atmosphere. The fact that all major laboratories involved in SAGE were present showed that this workshop was essential. The programme consisted of three main topics: ongoing SAGE programmes and model systems: statistical analysis of the data and user friendly interfaces; and comparison of SAGE data with other genome-wide techniques.

The keynote lecture was presented by Victor Velculescu (John Hopkins Oncology Center, Baltimore, USA) on transcription analysis using SAGE. In his overview, Velculescu showed SAGE analysis of gene expression patterns in yeast and mammalian cells. Since the entire sequence of the yeast genome is known, application of SAGE should be straightforward, since all tags can easily be linked to the corresponding gene. However, their analysis of over 60 000 SAGE tags from *S. cerevisiae* showed that the yeast transcriptome contains in addition to 6200 ORFs longer than 100 aminoacids several hundreds of thus far unidentified genes. The majority of these genes have very short reading frames, so called NORFs, and are therefore not detected by the first round of computer analysis of the yeast genome sequence. This clearly shows that even when a complete genome sequence is known one should be prepared for surprises and that SAGE can be used for gene discovery. Subsequent application of SAGE to look at complex differences between normal and cancer cells yielded long lists of altered mRNA levels that occur in intestinal cancers. Experiments to identify targets of P53 and APC tumour suppressor genes have identified potentially new targets of these tumour suppressor genes and will furnish more insight into the pathophysiology of cancer. Another application of SAGE, which is also used by other groups, is the search for tumour markers. Identification of highly expressed secreted proteins that are tumour-cell specific might be useful as markers that can easily be identified in body fluids.

DNA micro-array analysis of gene expression is another way of looking for differences in gene expression between two samples. Jörg Hoheisel (DKFZ, Heidelberg, Germany) gave a good overview of the current technologies for DNA micro arrays as they are being used by DKFZ for the yeast genome project (Eurofan II). It is clear that some technical problems still need to be resolved before DNA micro arrays can be implemented in every laboratory. Currently, the fabrication of DNA micro arrays is complex, especially when they are used for quantification. From the data that was shown by Hoheisel it is clear that micro arrays can be used as a rapid screening tool, provided that the equipment is available. For genomes such as yeast, for

which the entire sequence is known all open reading frames (ORFs) can be obtained as clones or PCR primer sets. A micro array yields more information in a shorter period then SAGE, since all yeast genes can be put on a single micro array. For mammalian genomes, however, there is still some way to go before full genome covering microarrays can be constructed.

The sessions on SAGE programmes showed that SAGE has grown to maturity. Many groups have applied SAGE to a variety of model systems, including the analysis of various types of cancer and the analysis of defined cell lines derived from various diseases. Arnoud Kal (Department of Biochemistry, AMC, Amsterdam) presented a detailed analysis of the yeast transcriptome. Possible pitfalls of SAGE can be identified in yeast. The occurrence of two tags from one ORF and the low expression level of regulating genes under conditions where major metabolic changes are expected are two of these expected problems and are in fact found. The largest mammalian SAGE programme is currently undertaken by the Cancer Genome Anatomy Project (CGAP). Greg Riggins (Duke University Medical Center, Durham, USA) presented the CGAP programme which is aimed at determining expression patterns of brain tumours using SAGE. Over 250 000 SAGE tags have already been determined from gliomas and normal brain (white matter) SAGE libraries. All this data is available through the Internet and will serve as a reference database for cancer researchers. In the near future the CGAP programme expects to release transcription profiles of many other brain tumours. Currently the data generated in the CGAP project is being checked by northern blot analysis to confirm the SAGE results. Initial experiments suggest a good correlation between SAGE data and expression patterns determined by northern blot analysis.

Two groups from the Academic Medical Center, Amsterdam, The Netherlands (Baas, Versteeg) presented their analysis of pediatric tumours of the nervous system. Even by analysing 10 000 tags per sample, several genes that were highly expressed in the tumours compared to their control tissues have already been identified. This shows that SAGE can also be used as a rapid differential screening procedure. Northern blot analysis confirmed the SAGE data. Hamish Scott (University of Geneva, Switzerland) showed the analysis of amniocyte cells derived from trisomie 21 and normal foetuses. Analysis of over 50 000 tags from both samples showed considerable differences in expression pattern of at least 50 transcripts.

Several presentations from Genzyme (Molecular Oncology, Framingham, MA, USA, which owns the commercial rights of SAGE) showed that industry is also investing heavily in SAGE analysis. Projects on autosomal dominant polycystic kidney disease on small cell lung cancer and osteomalacia are now used to identify disease pathways.

Not only human samples are analysed by SAGE. Elliot Margulies (University of Michigan Medical School, USA) and George Trendelenburg (Humboldt-University Berlin, Germany) presented data which showed that SAGE can also be used on murine tissues. There are already sufficient cDNA sequences from the mouse in Genbank to allow the identification of mRNA sequences by their SAGE tags.

Statistical analysis was another main topic. The enormous amount of data that is generated in every experiment requires sophisticated statistic analysis. Several approaches for the comparison of SAGE libraries were discussed and a comparison of published procedures showed that even though different approaches were used, the results are similar (Ruiter, AMC, Amsterdam). A so far unsolved problem is the comparison of multiple libraries. This is clearly an area that needs to be explored. Different user interfaces for SAGE programmes were also presented and the main difficulty in all is the identification of the proper mRNA sequence which belongs to a specific tag. The enormous amount of data in the EST section of Genbank makes it impossible to do manual searches. Computer algorithms have been developed to identify possible SAGE tags based on their location in the 3' end of an mRNA, close to the anchoring enzyme recognition site. Alex Lash (NCBI, Bethesda, USA) presented the interface and new software which will be used in the CGAP programme. This algorithm not only looks at the position of a sequence in the EST but also incorporates some corrections for potential sequencing errors. USAGE, a Unix based program, developed by van Kampen (AMC, Amsterdam) uses similar approaches and is also directly linked to the Unigene section of Genbank. Whereas the NCBI programme is more directed to the end user, SAGE labs might be more interested in USAGE. All the different approaches result in a series of tags with potentially related cDNAs and EST sequences, but none of them have been confirmed in experiments yet. Examples of pitfalls in the identification of SAGE tags were presented by Frank Baas, who showed that polymorphisms in or near the anchoring enzyme recognition site occur and will

therefore result in different tags for one single gene. It was suggested that a database of confirmed sequence tags with all the possible polymorphic variants should be generated.

One of the main disadvantages of SAGE was that large amounts of tissue were needed for the construction of SAGE libraries. This has hampered research by many groups who are interested in the analysis of a specific cell type and its specific physiological conditions within a mammalian tissue. Two approaches for the construction of SAGE libraries from small amounts of tissues were presented. Jean-Marc Elalouf (Cea Saclay, DBCM, Gif/Yvette, France) presented SADE, a micro assay for SAGE. Combination of single-step mRNA purification and modified reverse transcriptase and cloning procedures resulted in a protocol in which only 100 000 cells are needed. Preliminary experiments suggest that with even smaller amounts of cells SAGE libraries can be constructed. Nicole Datson (Leiden University, The Netherlands) presented micro SAGE, a procedure which also reduces the amount of starting material compared with conventional SAGE protocols. Both presentations showed that SAGE on small samples is achievable. Another, limiting step in SAGE is the sequence analysis. Wilhelm Ansorge (EMBL, Heidelberg, FRG, Germany) presented data on a novel automated DNA sequencing system which would allow high throughput sequencing on gels. The potential throughput of the system is estimated to be up to 1 megabase per day.

SAGE has grown to maturity and is currently being used for both the identification of biochemical pathways and gene discovery. With the technical developments, generation of SAGE libraries is no longer a major problem. It is therefore extremely important to pay attention to all the variables that can influence the gene expression pattern of the samples studied. For instance, a tumor sample might be obtained after chemotherapy. In such a case it is possible that the treatment greatly affected biological parameters and thus also the composition of the transcriptoom. In an ideal case one might be able to analyze cells obtained from control and knock-out mice (or yeast) isolated under similar conditions. This will guarantee minimal experimental variation and genetic homogeneity. Analysis of the enormous amount of data generated is currently the bottleneck. This calls for the development of better and more user-friendly interfaces of SAGE software and the identification of tags. Since the data generated from SAGE can be stored indefinitely in an easily interchangeable format, exchange of SAGE data will advance transcriptome research. The CGAP approach is a good example of data sharing which will be of great benefit to the scientific community. In order to make optimal use of the data generated, the starting material must be carefully defined and great care taken to avoid additional variation in conditions which might affect gene expression. The participants of this meeting agreed that sharing of data would benefit all and the development of a SAGE consortium might be useful. The participants also expressed that a follow-up of the meeting is essential and we are all looking forward to the second international SAGE meeting.