A large-scale screening of the normalized mammalian mitochondrial gene expression profiles.

Anisimov, Sergey

Published in:
Genetical Research

DOI:
10.1017/S0016672305007718

Published: 2005-01-01

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal
A large-scale screening of the normalized mammalian mitochondrial gene expression profiles

SERGEY V. ANISIMOV*
Neuronal Survival Unit, Wallenberg Neuroscience Center, Lund University, 221 84 Lund, Sweden

(Received 24 January 2005 and in revised form 7 July 2005)

Summary
Mammalian mitochondrial genomes are organized in a conserved and extremely compact manner, encoding molecules that play a vital role in oxidative phosphorylation (OXPHOS) and carry out a number of other important biological functions. A large-scale screening of the normalized mitochondrial gene expression profiles generated from publicly available mammalian serial analysis of gene expression (SAGE) datasets (over 17.7 millions of tags) was performed in this study. Acquired SAGE libraries represent an extensive range of human, mouse, rat, bovine and swine cell and tissue samples (normal and pathological) in a variety of conditions. Using a straightforward in silico algorithm, variations in total mitochondrial gene expression, as well as in the expression of individual genes encoded by mitochondrial genomes are addressed, and common patterns in the species- and tissue-specific mitochondrial gene expression profiles are discussed.

1. Introduction
Mitochondrial function is essential to eukaryotic cells. Mitochondria carry out a variety of metabolic processes, their major functions being to supply the cell with energy generated by oxidative phosphorylation (OXPHOS) and to participate in protein synthesis (Fernandez-Silva et al., 2003). Both parts of mammalian genomes: nuclear DNA (nDNA) and maternally inherited mitochondrial DNA (mtDNA) are involved in OXPHOS (Enriquez et al., 1999). OXPHOS features, regulated in part by gene expression, vary significantly depending on cellular activities (Kagawa et al., 1999). Considering the direct involvement of mitochondria-encoded genes in many basic cellular processes (such as heat production, protein folding, calcium homeostasis, apoptosis, ageing and carcinogenesis) and human diseases, directed studies of mitochondrial gene expression in tissues and cell types might expand our understanding of the biological relevance of mitochondrial transcription.

Serial Analysis of Gene Expression (SAGE) is a sequence-based technique to study messenger RNA (mRNA) transcripts in cell populations (Velculescu et al., 1995). Two major principles underline SAGE: (i) short (10 bp) expressed sequenced tags (ESTs) are sufficient to identify individual gene products, and (ii) multiple tags (up to 100) can be concatenated (i.e. linked together as a chain or cascade) and identified by sequence analysis. SAGE results are reported in either absolute or relative numbers of tags, which permits direct comparisons between libraries. Advances in SAGE protocol and sequencing techniques, supplemented with the expansion of sequence information available in public databases, have further augmented the perception of SAGE as an important tool of modern molecular biology widely used in a number of applications.

In the current study, mitochondrial gene expression was evaluated in a wide array of normal and pathological samples using the complete set of currently available SAGE data for mammals, including Homo sapiens, Mus musculus, Rattus norvegicus, Bos taurus and Sus scrofa. The aim of this study was to perceive variations and commonalities in species-specific, tissue-specific and cell-state-specific profiles of mitochondrial gene expression in mammals. Taken together, over 17.7 million SAGE tags derived from SAGE libraries generated with commonly used NlaIII (Velculescu et al., 1995) and Sau3A I (Virlon et al.,

* Corresponding author: Sergey V. Anisimov, Neuronal Survival Unit, Department of Experimental Medical Science, Wallenberg Neuroscience Center, Lund University, BMC A10 221 84 Lund, Sweden. Tel: +46 46 222 4369. Fax: +46 46 222 0531. e-mail: Sergey.Anisimov@med.lu.se
1999) anchoring enzymes were subjected to the analysis. Clustering analysis has allowed identification of some common patterns in the profiles of mitochondrial gene expression and the potential biological significance of variations in the total mitochondrial gene expression is discussed.

2. Methods

(i) SAGE

SAGE libraries subjected to the analysis were generally constructed as previously described (Velculescu et al., 1995). Briefly, cDNA is prepared from poly (A)+ mRNA with biotinylated oligo (dT)$_n$ and restriction enzymes of type I (NlaIII/Sla3A I) and type II (BsmFI) are used to generate tags. After a concatemerization step, the DNA is cloned into pZeRO-1, PCR-amplified and sequenced. Sequencing results are typically analysed using the SAGE 2000 software (available at http://www.sagenet.org).

The complete set of tag abundance data for publicly available SAGE libraries was downloaded from the Gene Expression Omnibus (GEO) database (National Center of Biotechnology Information (NCBI); http://www.ncbi.nlm.nih.gov/geo/; Edgar et al., 2002). Duplicate dimers were extracted, and all linker-generated tags were excluded from the analysis. Furthermore, all publicly available SAGE libraries constructed by the author and co-workers (namely, human: GSM1515 and GSM1515, Potapova et al., 2002; mouse: GSM580, GSM1681, GSM1682, GSM1683, GSM1684, GSM7759, GSM11348 and GSM11349, Anisimov et al., 2002a,b,c and unpublished data) were subjected to a ‘clean-up’ procedure (Anisimov & Sharov, 2004), in which clones containing four or fewer tags were excluded. All SAGE libraries were annotated and sorted based on the number of tags sequenced. Non-informative (A)$_n$ sequences were extracted from SAGE libraries when detected, and tags per million (tpm) values were recalculated accordingly as the transcript’s raw tag count divided by the number of reliable tags in the library and multiplied by 10$^6$.

(ii) Mitochondrial genomes

Complete mammalian mitochondrial genomes generally established as the ‘reference’ sequences were downloaded from the GenBank database (NCBI; http://www.ncbi.nlm.nih.gov/entrez/) as the following: Homo sapiens, accession number X93334 (a Caucasian from southern Sweden (Lund, Skåne); Arnason et al., 1996; Nilsson et al., 2003) and NC_001807 (an African (Yoruba ethnic group); Ingman et al., 2000); Mus musculus, AB042432 (Mizutani et al., 2001) and J01420 (Bibb et al., 1981); Rattus norvegicus, AY172581 (Rat Genome Sequencing Project Consortium, 2004) and NC_001665 (Gadaleta et al., 1989); Bos taurus, AY526085; and Sus scrofa, NC_000845 (Lin et al., 1999). Single nucleotide polymorphisms (SNPs) or genetic variations in mitochondrial sequences were addressed using the BLAST2 alignment tool (NCBI; http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html; Tatusova & Madden, 1999).

(iii) SAGE tag annotation

Complete mammalian mitochondrial genomes were searched for NlaIII and Sau3A I SAGE anchoring enzymes recognition sequences that were then annotated to the sequences for distinctive mitochondrial genes, ribosomal and transfer RNAs. These were then matched to individual SAGE libraries using the MS Access 2002 software package Query function. Individual queries were merged using MS Excel 2002 software, and calculations of relative expression of mitochondrial transcripts were performed using normalized tpm values.

(iv) Clustering analysis

Clustering analysis was performed using EPCLUST clustering, visualization and analysis software (http://ep.ebi.ac.uk/EP/EPCLUST/; credits to J. Vilo, P. Kemmeren & M. Kapushesky). Hierarchical clustering was performed via average linkage (average distance, UPGMA (unweighted pair group method with arithmetic mean)) clustering based on the linear correlation-based distance (Pearson, centred) method. K-mean clustering analysis was performed with initial cluster centres chosen by most distant (average) transcripts. The optimal number of clusters was determined empirically to produce the most balanced ratio of entries per cluster to the number of clusters.

To discover trends in the differences in mitochondrial gene expression in normal versus cancer and undifferentiated versus differentiated SAGE samples, average tag frequency values were calculated for the pools of SAGE libraries (e.g. prostatic cancer, 8 SAGE libraries). After ensuring that the average number of analysed tags per library was within the same order of magnitude in the groups being compared (e.g. normal prostatic tissue, 48,729 tags; prostatic cancer, 49,033 tags), average tag frequency values were compared using the following equation: 

$$\frac{[N_1 - k(N_1)^{1/2}] - [N_1 - k(N_1)^{1/2}] > 0, \text{ where } N_1 \text{ represents the larger tag frequency and } k = 3.29 \text{ for } P < 0.001, \text{ } k = 2.58 \text{ for } P < 0.01, \text{ } k = 1.96 \text{ for } P < 0.05 \text{ (Madden et al., 1997).}$$

3. Results

Tag frequency data for the complete set of publicly available mammalian SAGE libraries was downloaded.
Table 1. Summary of SAGE catalogs analyzed

<table>
<thead>
<tr>
<th>Species</th>
<th>GEO platform</th>
<th>Anchoring enzyme</th>
<th>Number of SAGE catalogs</th>
<th>Number of SAGE tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>GPL4</td>
<td>NlaIII</td>
<td>239</td>
<td>13 672 568</td>
</tr>
<tr>
<td></td>
<td>GPL6</td>
<td>Sau3A I</td>
<td>10</td>
<td>457 493</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>GPL11</td>
<td>NlaIII</td>
<td>43</td>
<td>2 029 235</td>
</tr>
<tr>
<td></td>
<td>GPL275</td>
<td>Sau3A I</td>
<td>9</td>
<td>218 412</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>GPL23</td>
<td>NlaIII</td>
<td>10</td>
<td>718 162</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>GPL223</td>
<td>NlaIII</td>
<td>10</td>
<td>586 001</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>GPL1270</td>
<td>NlaIII</td>
<td>2</td>
<td>84 616</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>323</td>
<td>17 766 487</td>
</tr>
</tbody>
</table>

GEO, Gene Expression Omnibus database (NCBI).

from the GEO database, for a total of 282 human, 96 mouse, 16 rat, 11 bovine and 2 swine SAGE libraries (see Supplemental Table 1.; all Supplemental Tables can be found at CUPwebsite). Among these, 2 human (Potapova et al., 2002) and 8 mouse (Anisimov et al., 2002a, b, c and unpublished data) SAGE libraries constructed and sequenced following the original protocol (Velculescu et al., 1995), with a number of minor modifications in the protocol (Kenzelmann & Mühlemann, 1999; Anisimov et al., 2002c), were analysed after being subjected to tag extraction procedures (Anisimov & Sharov, 2004). Taken together, acquired SAGE libraries have represented a wide spectrum of cell and tissue samples (normal and pathological) in a variety of conditions. The following criteria were applied to these when selecting libraries for the analysis of mitochondrial gene expression, and SAGE libraries were selected only if they represented (i) genetically unmodified species or cell cultures, (ii) untreated samples and (iii) SAGE libraries with a total of at least 10 000 reliable tags and a complete dataset available. For example, samples GSM14917 (human) and GSM3677 (mouse), representing telomerase overexpression-immortalized primary fibroblast culture and the Al×RAG-1/– TCR transgenic mouse, respectively, were excluded since they did not satisfy criterion (i). Similarly, samples GSM3832 (murine bone marrow-derived dendritic cells, LPS-treated) and GSM718 (normal human mammary gland epithelium; 1423 tags) were excluded as not satisfying criteria (ii) and (iii), respectively.

Three hundred and twenty-three SAGE libraries generated from human samples and mammalian species widely employed in general biology and molecular genetics were selected for further analysis. Together, these libraries consist of over 17.7 million SAGE tags. For Homo sapiens and Mus musculus, the GEO database provided SAGE libraries constructed using both NlaIII and Sau3A I anchoring enzymes common in SAGE applications, while rat, bovine and swine samples were all constructed using NlaIII anchoring enzyme (Table 1).

Complete ‘reference’ sequences of mammalian mitochondrial genomes were downloaded from the GenBank database. In cases where more than one complete mitochondrial genome sequence was established as ‘reference’, all these were analysed. Sequences were compared pair-wise using the BLAST2 sequence alignment tool, and a number of SNPs or genetic variations within anchoring enzyme (NlaIII and Sau3A I) recognition sites and corresponding SAGE tags were identified in human and rat complete mitochondrial genome sequences (Figs 1, 2). Sequences were formatted and individual ribosomal RNAs, mitochondrial genes and transfer RNAs were identified in these, followed by the identification of anchoring enzyme (NlaIII, all species; Sau3A I, Homo sapiens and Mus musculus) recognition sites. SAGE tags 10 bp long located immediately 3’ to the detected anchoring enzyme recognition sites were then retrieved from these sequences. Taken together, 47 and 25 human (for NlaIII and Sau3A I anchoring enzymes, respectively), 55 and 35 mouse (NlaIII and Sau3A I), 59 rat, 44 bovine and 53 swine (all – NlaIII) SAGE tags associated with mitochondrial H-strands were identified (see Supplemental Table 2). A number of these (1 and 1 for human, NlaIII and Sau3A I, respectively and 9 for rat, NlaIII) were originated from genetic variations and/or SNPs within the tags sequences. Two SAGE tags matching NADH dehydrogenase 1 (mt-Nd1) and cytochrome c oxidase subunit I (mt-CoI) rat mitochondrial genes were represented by 4 variant tags each, while tags matching rat mt-Co2, -Nd4 and -Nd5, human mt-Nd1 (NlaIII) and mt-Nd6 (Sau3A I) mitochondrial genes were each represented by 2 variant tags (Fig. 1, 2B). Supplemental Fig. 1 (all Supplemental Figures can be found at CUPwebsite) shows the relative dimensions and sequence breakdown, as well as the distribution and relative positions of SAGE NlaIII anchoring enzyme restriction sites, in human, mouse, rat, bovine and swine mitochondrial genomes, illustrating the abundance of mitochondrial gene expression data that could be extracted from SAGE samples.
Supplemental Table 3 lists the size of individual mammalian mitochondrial genome sequences. Owing to the existing overlap in the coding sequences of mt-Atp8/Atp6, mt-Atp6/Co3, mt-Nd4l/Nd4 (all species) and mt-Nd5/Nd6 (all but Homo sapiens) genes, sequence lengths of these genes were summed.

Primary analysis of mitochondrial gene expression profiles has demonstrated great variability in total mitochondrial gene expression. In the assessed datasets it ranged from 0.36% to 25.44% (calculated to the total (mitochondrial + nuclear) gene expression), with an average value of 4.59% for human, and from 0.46% to 22.64%, with an average value of 4.76%, for mouse. From 249 human SAGE libraries tested, 6 had a fraction of mitochondrial gene expression to the total number of analysed tags below 1%, and in 24 SAGE libraries it was above 10% (Table 2). Similarly, of 52 murine SAGE libraries tested, a fraction of mitochondrial gene expression was below 1% in 6 and above 10% in 13 SAGE libraries (Table 3). At the same time, initially smaller number of SAGE libraries generated from rat (10), bovine (10) and swine (2) samples did not lead to identification of many samples with extreme values of total mitochondrial gene expression. Average values of total mitochondrial gene expression were 7.19% for Rattus norvegicus, 2.14% for Bos taurus and 17.88% for Sus scrofa. Three of the rat samples (GSM581, Extraocular muscle; GSM1679 and GSM12532, Hippocampus) demonstrated a fraction of mitochondrial gene expression >10% (15.72%, 13.39% and 13.25% to the total number of analysed tags, respectively), while in 2 bovine samples (GSM11027, Madin-Darby bovine kidney (MDBK) cells and GSM24604, Circulating c/d T cells) this fraction was <1% (0.97% and 0.42%, respectively).

Supplemental Table 2 shows a distribution of SAGE tag frequencies in mammalian mitochondrial genomes. Clearly, this distribution is reflective of the relative expression of individual mitochondrial transcripts in the sample. In many samples, a significant proportion of total mitochondrial expression was granted by only a few transcripts. In the most abundant array of human NlaIII-derived SAGE libraries, the mt-Co2 most 3′ tag alone (CCCAT-CGTCC) accounted for 12.5% of the total mitochondrial gene expression (by the averaged data). Relative frequency values for individual tags have reached extreme ranges in certain libraries: in a human PC3 cell SAGE library (Potapova et al., 2002), for example, the same tag has accounted for 26.7% of total mitochondrial gene expression (and 0.37%...
Mitochondrial gene expression profile screening

Fig. 2. Distribution of SAGE anchoring enzyme restriction sites in mouse (A) and rat (B) mitochondrial genome. Circular mitochondrial DNA is plotted in linear form. Numbers at the end of each segment correspond to the base number of the RNA encoded by the heavy strand (H strand); oblique strokes indicate overlaps in the coding sequences; the vertical arrow (↑) indicate the starting position; scale bars correspond to 1 kb. Two complete mouse (GenBank accession nos. AB042432 and J01420) and rat (AY172581 and NC_001665) mitochondrial DNA samples representing Mus musculus and Rattus norvegicus strains, respectively, were used for each analysis. Shaded areas represent tRNAs (black) and control region (D-loop; grey). Left-headed arrows (↓) indicate areas encoded by the light strand (L strand). Filled triangles under the bars (▲) and empty triangles above the bars (▼) indicate NlaIII and Sau3A I restriction sites, respectively. Filled asterisks (☆) indicate potential single nucleotide polymorphisms (SNPs) or genetic variations in SAGE tags for NlaIII (mt-Nd1, G/C/GT/C/ctcacc; mt-Co1, CTGG/G/GGATGTATC; mt-Co2, AGCC/G/TCCCT; mt-Nd4, GGAGGC/G/ACAA; mt-Nd5, ATCAATGT/ATA) and Sau3A I (mt-Rnr2, C/G/TAG/A/AATA and T/G/GTGT/G/ATA; mt-Nd1, A/GGGATGAGCA/C; mt-Cytb, AT/T/CCTACTAT) anchoring enzymes in rat mitochondrial sequences. Empty asterisks (☆) indicate potential SNPs in the recognition sequences of NlaIII (mt-Nd1, CAT/G/A) and Sau3A I (mt-Nd5, GAT/CC, C/GATC and GA C/T A/C/T) anchoring enzymes in rat mitochondrial sequences.

of total gene expression), while tags matching tRNA-Ser and mt-Co1 have accounted for 22.8% and 14.3% of total mitochondrial gene expression, respectively. At the same time, many other SAGE libraries have different mitochondrial expression profiles. Predictably, similar diversity was observed in other mammalian species and in datasets generated using an alternative SAGE anchoring enzyme (Sau3A I).
To address the cellular basis for the evident dissimilarity in mitochondrial gene expression profiles in an integrated way, clustering analysis was performed using EPCLUST clustering, visualization and analysis software. Hierarchical clustering was performed using a correlation-measure-based distance (Pearson, centred)/average linkage (average distance) clustering method and hierarchical trees built for individual datasets (Fig. 3; Supplemental Figs 2, 3). The optimal number of K-mean clusters (\textit{Homo sapiens}: \textit{NlaIII}, 10, \textit{Sau3A I}, 4; \textit{Mus musculus}: \textit{NlaIII}, 5, \textit{Sau3A I}, 3; \textit{Rattus norvegicus}: \textit{NlaIII}, 3; \textit{Bos taurus}: \textit{NlaIII}, 3) was determined empirically and these have been built for individual datasets resulting in pattern recognition (Figs 3, 4 and Supplemental Fig. 3). Notable diversity was observed among possible transcription profiles (Fig. 4, Supplemental Fig. 3), with either relatively high expression characteristics for just a few individual transcripts (e.g. Supplemental Fig. 3B, Cluster 4; Supplemental Fig. 3J), or with expression profiles being more proportional (e.g. Fig. 4, Cluster 5). It was anticipated that samples generated from the tissues/cell types with similar characteristics would fall into the same cluster, as demonstrated in Fig. 3 where a complete dataset for human \textit{Sau3A I} SAGE libraries is comprised of only two major transcription profiles, clearly associated with either kidney tissue structures or U937 monoblast/early monocyte cell line (Fig. 3). However, the analysis of more complex datasets failed to provide solid support for that hypothesis, and the analysis of abundant SAGE datasets based on sample origin (tissue or sample type, etc.) has identified very few groups of these falling into the same cluster. For example, while in a human \textit{NlaIII} SAGE dataset broken into 10 K-mean clusters 5 of 12 samples falling in Cluster 7 were generated from normal mammary gland tissue (Fig. 4), one more sample of the same origin was detected in Cluster 5 and 2 in Cluster 10. Similarly, 12 of 20 samples falling in Cluster 2 represented either carcinomas (e.g. GSM14753, GSM1730) or tumours generating metastatic carcinomas (GSM696) or

<table>
<thead>
<tr>
<th>N</th>
<th>SAGE library</th>
<th>AE</th>
<th>Sample</th>
<th>Mt&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GSM14751</td>
<td>\textit{NlaIII}</td>
<td>Skin, melanoma</td>
<td>0.36%</td>
</tr>
<tr>
<td>2</td>
<td>GSM754</td>
<td>\textit{NlaIII}</td>
<td>Prostate, cell line</td>
<td>0.42%</td>
</tr>
<tr>
<td>3</td>
<td>GSM41382</td>
<td>\textit{NlaIII}</td>
<td>Cartilage chondrosarcoma grade 3</td>
<td>0.80%</td>
</tr>
<tr>
<td>4</td>
<td>GSM668</td>
<td>\textit{NlaIII}</td>
<td>Kidney, embryonic cell line 293, uninduced cells</td>
<td>0.87%</td>
</tr>
<tr>
<td>5</td>
<td>GSM675</td>
<td>\textit{NlaIII}</td>
<td>Ovary</td>
<td>0.97%</td>
</tr>
<tr>
<td>6</td>
<td>GSM32266</td>
<td>\textit{NlaIII}</td>
<td>Microvascular endothelial cells exposed to sustained high shear stress</td>
<td>0.99%</td>
</tr>
<tr>
<td>1</td>
<td>GSM786</td>
<td>\textit{NlaIII}</td>
<td>Frontal cortex</td>
<td>10.14%</td>
</tr>
<tr>
<td>2</td>
<td>GSM7800</td>
<td>\textit{NlaIII}</td>
<td>Primary gastric cancer tissue, poorly differentiated (scirrhous type)</td>
<td>10.24%</td>
</tr>
<tr>
<td>3</td>
<td>GSM1499</td>
<td>\textit{NlaIII}</td>
<td>Normal heart tissue</td>
<td>10.30%</td>
</tr>
<tr>
<td>4</td>
<td>GSM14780</td>
<td>\textit{NlaIII}</td>
<td>Normal gastric epithelial tissue from the antrum</td>
<td>10.37%</td>
</tr>
<tr>
<td>5</td>
<td>GSM725</td>
<td>\textit{NlaIII}</td>
<td>Mammary gland, adenocarcinoma</td>
<td>10.75%</td>
</tr>
<tr>
<td>6</td>
<td>GSM14742</td>
<td>\textit{NlaIII}</td>
<td>Brain cortex, oligodendroglialoma</td>
<td>10.76%</td>
</tr>
<tr>
<td>7</td>
<td>GSM2451</td>
<td>\textit{NlaIII}</td>
<td>Brain, astrocytoma</td>
<td>11.08%</td>
</tr>
<tr>
<td>8</td>
<td>GSM14783</td>
<td>\textit{NlaIII}</td>
<td>Brain anaplastic medulloblastoma</td>
<td>11.15%</td>
</tr>
<tr>
<td>9</td>
<td>GSM14796</td>
<td>\textit{NlaIII}</td>
<td>Brain, substantia nigra</td>
<td>11.49%</td>
</tr>
<tr>
<td>10</td>
<td>GSM728</td>
<td>\textit{NlaIII}</td>
<td>Normal colonic epithelium</td>
<td>11.75%</td>
</tr>
<tr>
<td>11</td>
<td>GSM6956</td>
<td>\textit{NlaIII}</td>
<td>Brain, normal, greater than 95% white matter</td>
<td>11.77%</td>
</tr>
<tr>
<td>12</td>
<td>GSM763</td>
<td>\textit{NlaIII}</td>
<td>Brain, normal</td>
<td>12.11%</td>
</tr>
<tr>
<td>13</td>
<td>GSM755</td>
<td>\textit{NlaIII}</td>
<td>Adenocarcinoma, colon</td>
<td>12.23%</td>
</tr>
<tr>
<td>14</td>
<td>GSM729</td>
<td>\textit{NlaIII}</td>
<td>Normal colonic epithelium</td>
<td>13.53%</td>
</tr>
<tr>
<td>15</td>
<td>GSM10427</td>
<td>\textit{Sau3A I}</td>
<td>Kidney distal convoluted tubule</td>
<td>13.59%</td>
</tr>
<tr>
<td>16</td>
<td>GSM676</td>
<td>\textit{NlaIII}</td>
<td>Brain, normal, greater than 95% white matter</td>
<td>13.64%</td>
</tr>
<tr>
<td>17</td>
<td>GSM10425</td>
<td>\textit{Sau3A I}</td>
<td>Kidney medullary thick ascending limb</td>
<td>13.73%</td>
</tr>
<tr>
<td>18</td>
<td>GSM784</td>
<td>\textit{NlaIII}</td>
<td>Normal gastric epithelial tissues</td>
<td>13.79%</td>
</tr>
<tr>
<td>19</td>
<td>GSM819</td>
<td>\textit{NlaIII}</td>
<td>Normal muscle (old)</td>
<td>16.01%</td>
</tr>
<tr>
<td>20</td>
<td>GSM824</td>
<td>\textit{NlaIII}</td>
<td>Normal muscle (young)</td>
<td>16.66%</td>
</tr>
<tr>
<td>21</td>
<td>GSM14771</td>
<td>\textit{NlaIII}</td>
<td>Placenta</td>
<td>17.95%</td>
</tr>
<tr>
<td>22</td>
<td>GSM10424</td>
<td>\textit{Sau3A I}</td>
<td>Kidney proximal straight tubule</td>
<td>18.79%</td>
</tr>
<tr>
<td>23</td>
<td>GSM708</td>
<td>\textit{NlaIII}</td>
<td>Normal kidney tissue</td>
<td>19.31%</td>
</tr>
<tr>
<td>24</td>
<td>GSM713</td>
<td>\textit{NlaIII}</td>
<td>Brain, normal thalamus</td>
<td>25.44%</td>
</tr>
</tbody>
</table>

Indexes (GSM#) represent GEO database accession numbers for SAGE libraries. AE, anchoring enzyme.

<sup>a</sup> Total mitochondrial gene expression (as percentage of the total number of analyzed tags).
preceding carcinoma development (GSM780). At the same time, many other carcinoma samples were detected in further clusters. Only one group of samples that fell entirely into a single cluster was identified within the course of analysis. Mitochondrial profiles of all 4 SAGE samples generated from human retina (\textit{NlaIII} dataset; GSM571-574), including human retinal pigment epithelium (RPE), peripheral and central retina (Sharon et al., 2002), were found in Cluster 2 (Fig. 4). Notably, all these samples also fell into a single cluster when the dataset (which contains 239 entries) was analysed with the number of K-mean clusters increased up to 51.

### 4. Discussion

The mammalian mitochondrial genome is a circular molecule organized in a conserved and extremely compact manner. The mitochondrial DNA (mtDNA) encodes 2 ribosomal RNAs and 22 transfer RNAs that play an essential role in mitochondrial translation. A non-coding mitochondrial control region contains the main regulatory sequences for transcription and replication initiation, including a triple-stranded D-loop. Though the great majority of mitochondrial proteins are encoded by the nuclear genome, mtDNA encodes 13 proteins involved in oxidative phosphorylation (for review see Hiendleder & Wolf, 2003 and Fernandez-Silva et al., 2003). Individual mammalian mitochondrial-genome-encoded genes have demonstrated a number of other important biological functions, such as serving as histocompatibility antigens for the NADH dehydrogenase 1 (\textit{mt-Nd1}) and cytochrome c oxidase subunit I (\textit{mt-Co1})-derived peptides (Loveland et al., 1990; Morse et al., 1996) or playing a role in protein folding for 12S and 16S ribosomal RNA (\textit{mt-Rnr1} and -Rnr2; Suljoadiokusumo et al., 2001).

Mammalian mitochondrial genomes share a significant level of sequence homology and genes in these are organized in a virtually identical manner (Supplemental Fig. 1). Though there have been reports on the variations in mtDNA structure (triplex/duplex forms) in large versus small mammals (Annex & Williams, 1990), it is expected that basic mitochondrial gene expression patterns should be similar within mammalian species. Numerous SNPs or genetic variations were detected in human and rat reference mitochondrial sequences, and average allele distribution could be estimated from the pooled datasets. For example, the G-allele of human \textit{mt-Nd6} gene CTCCCGAAT \textit{C/G} tag (\textit{Sau3A I} anchoring enzyme) was not detected in either 11 SAGE libraries constructed using the pooled RNA from 9 individuals (Chabardes-Garonne et al., 2003) or human U937 monoblast/early monocyte cell line, while the A-allele of human \textit{mt-Nd1} gene \textit{G}/\textit{A}CCAACCTCC tag (\textit{NlaIII}) was detected in 7 SAGE libraries of 239 tested, with the average allele frequency being only 0.004 (Fig. 3A; Supplemental Table 2). Similarly, of
4 potential allelic variants of rat mt-Nd1 gene GC/CCTCACCC tag (NlaIII), only 3 were detected in 10 SAGE libraries, with the average allele frequency being 0.991, 0.005 and 0.004 for CT, GT and CC allelic variants, respectively (Supplemental Fig. 3E). In contrast, mitochondrial genome sequence identity was observed for mouse reference sequences, consistent with earlier observations and based on a common descent of all mouse inbred strains along the maternal line (Silver, 1995).

In the current study, a large-scale screening (323 SAGE libraries, over 17.7 million SAGE tags) of normalized mitochondrial gene expression profiles was performed. Normalization via recalculating tag frequency in individual SAGE libraries to tag per million (tpm) values was employed as a generally accepted normalization method widely used in the field. A great variability in the total mitochondrial gene expression was observed (ranging from 0.36% to 25.44% of the total gene expression in human and from 0.46% to 22.64% in mouse), consistent with the wide range of cell and tissue types screened. The results of this screening confirm that total mitochondrial gene expression varies considerably depending on a cell/tissue type. Since the majority of genomic-scale studies employ tissues rather than individual cells or pure cellular populations, the main type of cell present in the tissue determines the number of mitochondria per cell and tissue/organism metabolic activity will result in tissue-specific variations in the total mitochondrial gene expression and that of individual mitochondrial-genome-encoded genes. There are therefore considerations that discrete cell types are able to partially ‘mask’ each other’s gene expression profiles within the tissue, due to the potential disparity of their own overlapped distinct gene expression profiles. It is thus important to evaluate these complex tissues and cell populations as an integral function of the total gene expression of the cells constituting the sample. Moreover, some SAGE tags could be associated with more than one transcript. As a search with the SAGEmap (http://www.ncbi.nlm.nih.gov/projects/SAGE) annotation tool demonstrates, this is true for many 10 bp SAGE
Mitochondrial gene expression profile screening

Fig. 4. Mitochondrial gene expression patterns as identified by K-mean clusters in human NlaIII-derived SAGE dataset (239 libraries). Individual SAGE tags (including variant tags) are spotted at abscissa in the order following mitochondrial H-strand (see Supplemental Table 2). Most 3’-located tags for mt-Rnr2, -Co1, -Co2, -Nd4 and -Cytb genes are denoted with arrows in the representative cluster. Normalized tpm values are spotted on the ordinate. Groups of numerous individual samples fall into the limited number of clusters with analogous expression profiles of individual mitochondrial transcripts. Note the diversity among possible transcription profiles with either relatively high expression characteristics for a few individual transcripts (e.g. Clusters 2 and 3), or with more proportional expression profiles (e.g. Clusters 5 and 9).

tags. For example, tag ATACTGACAT (most 3’ tag for murine mt-Co3 gene (NlaIII dataset), highly abundant in many SAGE libraries) matches 543 different UniGene clusters, 15 of which are considered ‘reliable’. Though this is a common concern in all studies utilizing conventional SAGE technology (i.e. not ‘LongSAGE’ or its analogues), it has been taken into account that since there is currently no way to distinguish between such multiple matches, a certain degree of bias could therefore be introduced into the analysis. Moreover, some mitochondrial-encoded genes do not bear restriction sites for anchoring enzymes routinely employed in SAGE (such as NlaIII). Examples include mt-Atp8 and mt-Nd4f genes for Homo sapiens, and mt-Nd1, mt-Nd4l and mt-Nd6 for Mus musculus. Supplementing SAGE data with an alternative anchoring enzyme (e.g. Sau3A I or MmeI) or alternative technique (such as hybridization-based microarray technology) should address this issue, generating the most comprehensive and accurate expression profiles.

Tables 2 and 3 list human and murine SAGE libraries with extreme values of total mitochondrial gene expression, the highest levels belonging mostly to brain, muscle (cardiac and skeletal) and kidney tissue, as well as to placenta and several tumour subtypes. High requirements for energy production and/or protein synthesis are known for all of these. Justifiably, the large amount of energy produced by OXPHOS in mitochondria is directly associated with continuous contraction of a muscle fibre, or transmembrane ion pumping involved in the transmission of a nerve signal or the modification of primary urine. Within all these tissues, regional variations in total mitochondrial gene expression could be associated with different spectra of cell subtypes or activities. For example, most of the central nervous system regions (including those studied by SAGE, such as brain cortex, thalamus and substantia nigra) differ in the relative presence of neuronal cell subtype fractions, and their firing rate.

Expression of individual SAGE tags varied to an even greater extent: while some tags (including variant ones) were either not detected in all SAGE libraries studied or had low frequency (see Supplemental Table 2), others demonstrated generally high frequency levels, with extreme abundance in certain cell types. Maximum normalized values detected were 103 686 tpm for mouse NlaIII (GCTGCCCTCC, mt-Co1, GSM15116, old heart; Anisimov et al., 2005)
and 59,623 for Sl3A1 (TGCTTCAATA, mt-Co2, GSM4874, medial striatum) datasets, 38,589 for rat (GATGCCTCCCC, mt-Co1, GSM581, extracerebral muscle; Cheng & Porter, 2002) and 32,868 for human (CCCATGGTCC, mt-Co2, GSM708, normal kidney tissue) NlaIII datasets. Notably, all these represented the most 3’ tags of the corresponding genes; this is further illustrated by the examples presented in Fig. 4 and Supplemental Fig. 3. While 4 bp SAGE anchoring enzyme recognition sites are abundant in mammalian mitochondrial genomes (see Figs 1, 2 and Supplemental Fig. 1), and many mitochondrial transcripts bear more than one SAGE tag, most SAGE tags originate from the most 3’ anchoring enzyme cleavage sites. Still, a number of SAGE tags match more 5’ anchoring enzyme cleavage sites in mitochondrial genes (Supplemental Table 2). It is therefore appropriate to evaluate the expression of individual mitochondrial-encoded genes in a cumulative way, i.e. as a sum of all corresponding transcripts. Furthermore, a few mitochondrial genes, such as mt-Atp6/8 and mt-Nd4/4l (Supplemental Fig. 1, Supplemental Table 3) have 3’-to-5’-end overlaps and are transcribed together, with separate proteins being produced from these by frameshifting during the subsequent translation. At the same time, shifts in the distribution of frequency values for multiple tags related to the individual genes encoded by the mitochondrial genome might be indicative of defects in mitochondrial transcription potentially associated with ageing and disease (Kowald & Kirkwood, 2000).

In order to identify mitochondrial gene expression profiles with patterns characteristic for individual tissues or cell types, an extensive clustering analysis (by K-means) was performed. Figs 3 and 4 and Supplemental Fig. 3B, D, F, H and J clearly demonstrate that groups of individual samples fall into the limited number of clusters with analogous expression profiles of individual mitochondrial transcripts, and it was anticipated that samples generated from the tissues/cell types with similar characteristics would fall into the same cluster. However, the analysis of complex datasets (such as human and murine NlaIII datasets) failed to provide solid support for that hypothesis. Although a few groups of samples of the same sample origin (tissue or sample type, etc.) demonstrated a trend towards similar mitochondrial expression profiles, SAGE libraries generated from human retina was the only identified group falling completely into a single cluster (human NlaIII dataset, Cluster 2; Fig. 4). It could therefore be suggested that the origin of the tissue and sample could determine mitochondrial gene expression profile on rare occasions.

Numerous reports validate the assumption that mitochondrial gene expression (i.e. the abundance of mtRNA transcripts) is a function of cell type and state, as well as of the metabolic/respiratory activity of the proband (Annex & Williams, 1990; Kagawa et al., 1999; Unami et al., 2004). Importantly, the replication and transcription of the mitochondrial genome depends on nuclear gene products, and post-transcriptional mechanisms play an important role in regulating mitochondrial expression (Ostronoff et al., 1996; Escriva et al., 1999; Wu et al., 2002). While the mtDNA/mitochondrion ratio is essentially constant in all cell types in mammalian species, the number of mitochondria per cell and the amount of mtDNA per cell seem to be closely regulated within a given cell type but differ widely between cell types (Robin & Wong, 1988). Total mitochondrial gene expression seems to be a direct function of ageing, and generally declines with age, as has been demonstrated in a number of studies (Bodyak et al., 2002; Lu et al., 2004), while the activity of mitochondrial enzymes is suggested as a biochemical marker of ageing (Navarro et al., 2004). Importantly, ageing might affect the expression of individual mitochondrial genes to various extents (Tollet-Egnell et al., 2001). Quantitatively accurate data from both publicly available ageing-related SAGE studies (Welle et al., 2000; Anisimov et al., 2005) provide a degree of support for this hypothesis. It has been demonstrated that ageing in human skeletal muscle is associated with lower expression levels for individual mitochondrial transcripts, including those for mt-Nd1, mt-Co2, mt-Atp8, mt-Atp6, mt-Nd3 and mt-Nd4 (Welle et al., 2000); moreover, preliminary SAGE data for mouse cardiac muscle suggest that ageing in this tissue might be associated with lower expression levels for a few mitochondrial transcripts, such as mt-Atp8 and mt-Nd5 (Anisimov et al., 2005).

There are indications that other important physiological and pathological processes, such as cell differentiation (Angenieux et al., 2001), immortalization (Kim et al., 2001) or tumorigenesis (Ojala et al., 2003; Baris et al., 2004), are also associated with alterations in total mitochondrial gene expression. However, the applied algorithm failed to associate these processes with characteristic alterations in expression of mitochondrial transcripts. Instead, when analysing SAGE libraries representing undifferentiated versus differentiated or normal versus cancer cells diverse changes in the expression of individual mitochondrial genes were detected. Though total mitochondrial gene expression values have demonstrated a trend to increase in association with tumorigenesis, no regular patterns in these or in the expression profiles of individual mitochondrial genes have been revealed. Supplemental Table 4 shows an example of such an analysis performed for normal versus cancer cells generated from mammary gland (average data from 8 vs 21 SAGE libraries, respectively), pancreas (7 vs 5) and prostate (6 vs 8).
Statistical analysis of averaged mitochondrial gene expression data demonstrates a notable diversity in the observed alterations caused by cellular transformation. Observed alterations in the expression of a few mitochondrial transcripts (mt-Rrn1, mt-Rrn2, mt-Co2, mt-Co3) are in opposite directions in different tissues, illustrating the complexity of this issue. Similar analysis performed for undifferentiated versus differentiated cells also failed to reveal regular patterns in expression profiles of individual mitochondrial genes (data not shown).

In this study a large-scale screening of the normalized mitochondrial gene expression profiles generated from publicly available mammalian SAGE datasets was performed, and variations in total mitochondrial gene expression, as well as in the expression of individual genes encoded by mitochondrial genomes, were analysed. Though the analysis demonstrates that total mitochondrial gene expression is a function of the cell type/cell state of the sample, the approach failed to reveal a direct link between the former and expression profiles of individual mitochondrial genes. At the same time, the trends in the diversity of mitochondrial gene expression profiles are remarkable, and it could be suggested that on some occasions the origin of tissue and sample could determine the mitochondrial gene expression profile.

Accumulated data illustrate the utility of SAGE technology that, unlike microarray technology, can generate quantitatively accurate and normalized gene expression information. The Gene Expression Omnibus database (GEO, NCBI) has become a convenient and broad public tool for comparative gene expression studies. A moderate volume of important biological information can be extracted from this database using straightforward in silico algorithms and subjected to further, more directional studies. Being validated with conservative molecular biology methods, such information can serve as an important tool in discovering delicate mechanisms of vital cellular functions.

This study was supported by a grant from The United States Army Medical Research Acquisition Activity (USAAMRAA) Award No. W81XWH-04-1-0366, National Institute of Health Grant Number 1 R21 NS043717-01A1 and by The Swedish Parkinson Foundation. The author is grateful to all GEO database contributors and to Professor Patrik Brundin, Assistant Professor Jiu-Yi Li, Dr Emma Lane and Ana Sofia Correia for their critical review of this manuscript.

References


