


METHODOLOGY

Open Access



An advanced enrichment method for rare somatic retroelement insertions sequencing

Alexander Y. Komkov^{1,2}, Anastasia A. Minervina¹, Gaiaz A. Nugmanov¹, Mariia V. Saliutina¹, Konstantin V. Khodosevich³, Yuri B. Lebedev¹ and Ilgar Z. Mamedov^{1,4*} 

Abstract

Background: There is increasing evidence that the transpositional activity of retroelements (REs) is not limited to germ line cells, but often occurs in tumor and normal somatic cells. Somatic transpositions were found in several human tissues and are especially typical for the brain. Several computational and experimental approaches for detection of somatic retroelement insertions was developed in the past few years. These approaches were successfully applied to detect somatic insertions in clonally expanded tumor cells. At the same time, identification of somatic insertions presented in small proportion of cells, such as neurons, remains a considerable challenge.

Results: In this study, we developed a normalization procedure for library enrichment by DNA sequences corresponding to rare somatic RE insertions. Two rounds of normalization increased the number of fragments adjacent to somatic REs in the sequenced sample by more than 26-fold, and the number of identified somatic REs was increased by 8-fold.

Conclusions: The developed technique can be used in combination with vast majority of modern RE identification approaches and can dramatically increase their capacity to detect rare somatic RE insertions in different types of cells.

Keywords: Somatic retroelement insertions, Genomic normalization, Kamchatka Crab duplex-specific nuclease

Background

In the past decade the rapidly growing number of whole genome sequencing studies proved the somatic variability to be the common property of genomes of both malignant and normal human cells [1–3]. This somatic variability includes single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and somatic insertions of active retroelements (REs) of L1, Alu and SVA subfamilies. Somatic RE insertions were found in several types of malignancies including lung, colorectal and prostate cancers [4–6]. Studies of somatic RE insertions in normal cells were mainly focused on human brain since RE transpositions were shown to be associated with human adult neurogenesis [7–9]. In other normal human tissues somatic RE variations are still poorly studied [10].

The modern experimental approaches for detection of somatic RE insertions is based on targeted high-throughput sequencing of genome fragments adjacent to RE insertions (TIP-Seq [11], RC-Seq [12], L1-Seq [13], TE-NGS [14]). However, even though the sequencing capacity of HTS technologies is growing rapidly somatic REs studies are still limited to few tissue samples, especially in case of low somatic insertions rate. At the moment, it is almost impossible to proceed the routine screening for somatic retroposition events in a sufficient number of individual cell genomes even using the most robust Illumina NovaSeq platform. Existing hybridization [12] and amplification-based enrichment techniques [11, 15] partially solve this problem allowing to increase the concentration of active RE subfamilies in sequencing libraries. Enrichment capacity achievable in these methods is sufficient to detect somatic RE insertions in most rapidly dividing cell samples such as tumor or embryonic cells where the proportion of somatic RE carrying cells is high. However, somatic RE insertions (especially from large subgroups) presented in one or few cells of

*Correspondence: lmamedov@mx.ibch.ru

¹Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Miiklukho-Maklaya str. 16/10, 117997 Moscow, Russia

⁴Pirogov Russian National Research Medical University, Ostrovitianov str. 1, 117997 Moscow, Russia

Full list of author information is available at the end of the article



entire tissue sample remain almost undetectable among overwhelming majority of molecules corresponding to fixed and polymorphic ones. For instance approximately 4,000 AluYa5 insertions are present in genomic DNA of each cell. Consequently, up to 800,000,000 molecules in AluYa5-enriched library represent fixed and polymorphic insertions in a 100,000 diploid cells sample whereas each somatic insertion can be presented in this sample by just several molecules. Thus, identification of rare somatic insertions without their specific enrichment is cost ineffective and looks like finding a needle in a haystack.

Another challenging point in somatic RE studies is the estimation of the number of cells in which a particular insertion is present. Most high-throughput sequencing library preparation techniques employ PCR amplification which inevitably introduce significant quantitative bias. As a result, the number of sequencing reads corresponding to each particular somatic insertion provides no assessment of the number of cells bearing this insertion even with usage of random fragmentation points for removing PCR duplicates.

Here we present the first approach for specific enrichment for rare somatic RE insertions in sequencing libraries. The method based on normalization procedure with utilization of Kamchatka Crab duplex-specific nuclease which allows to eliminate abundant DNA sequences and thus to increase the concentration of rare DNA sequences in the library. “Unique molecular identifiers” (UMIs) [16, 17] are used to remove PCR duplicates and estimate the true number of cells bearing a particular insertion. The method was employed for identification of AluYa5 somatic insertions in a sample of 50,000 nuclei from the adult human brain.

Results

The rationale of the method

The proposed method allows to identify rare somatic RE insertions (present in a single or few cells) using less sequencing reads. Furthermore, the method allows to quantify the number of cells that bear a particular insertion. There are three principal steps in the procedure:

1) Obtaining the genome fragments adjacent to RE insertions. In this study we performed selective amplification of the regions flanking retroelements of an evolutionary young AluYa5 subfamily using previously described technique [15, 18–20] with several modifications (see Fig. 1 and selective amplification section below). Obtained amplicon contained sequences flanking AluYa5 insertion (about 90%) present in each cell, somatic AluYa5 insertion and sequences flanking insertions belonging to other Alu subgroups depleted during AluYa5-specific amplification. Sequences of non-Ya5 and somatic AluYa5 insertions were presented at a low level in the amplicon and were

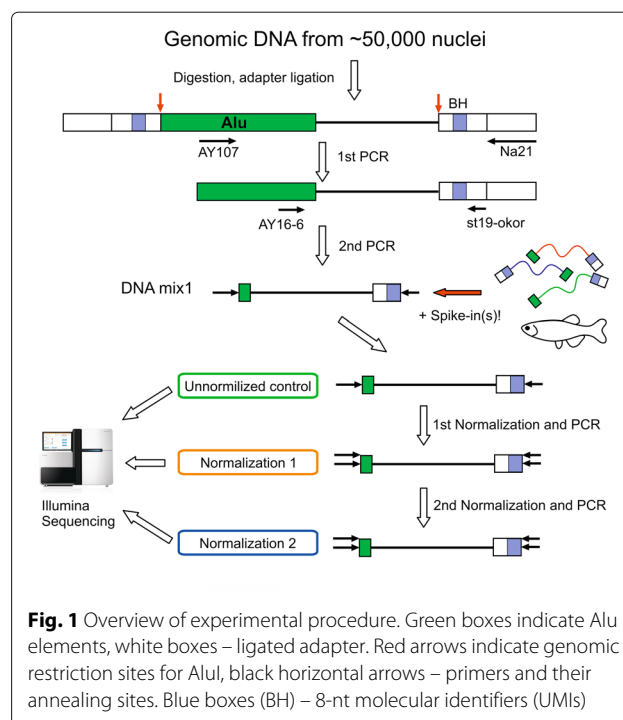


Fig. 1 Overview of experimental procedure. Green boxes indicate Alu elements, white boxes – ligated adapter. Red arrows indicate genomic restriction sites for AluI, black horizontal arrows – primers and their annealing sites. Blue boxes (BH) – 8-nt molecular identifiers (UMIs)

used for tracing changes of amplicon composition during subsequent normalization stages.

2) Normalization using duplex-specific DNase. At this stage, the amplicon is denatured and then slow renatured so that the abundant DNA molecules find their complementary pairs and return to the double-stranded (ds) state, while the rare molecules lag behind and remain single-stranded (ss). Subsequent treatment by duplex-specific DNase from Kamchatka crab [21] eliminates dsDNA leaving ssDNA intact. After the amplification the relative abundance of molecules with low concentration in the original mix (including the flanks of somatic REs) is increased. This procedure is repeated twice to increase the enrichment efficiency.

3) Sequencing of the normalized amplicons by Illumina and data analysis.

Obtaining the genome fragments adjacent to RE insertions

Fifty thousand nuclei were extracted from the frozen human brain sample (frontal cortex). Genomic DNA was extracted and used for selective amplification using suppression PCR. This procedure included DNA digestion by AluI endonuclease followed by ligation of suppressive adapters (see Fig. 1). Each molecule of the ligated adapter contains a “unique molecular identifier” (UMI) – a random sequence of 8 partly degenerated nucleotides (see Additional file 1 for oligonucleotide sequences). As a result, each of the ligated DNA molecules is marked by

one of 6561 different 8-nt oligomers prior to the amplification. UMIs allow to estimate the number of cells bearing a particular somatic insertion in case of sufficient sequencing depth. Sequences with identical UMI indicate a single ligation event and the number of different UMI corresponds to the number of cells containing each RE insertion. Following the adapter ligation two rounds of selective PCR were performed. In the first round, primer AY107 [20] was used for the selective amplification of insertions belonging to AluYa5 and AluYa8 subfamilies. The second primer (Na21) anneals to the 5' part of the ligated adapter. In the second round of amplification, a nested pair of primers was used: AY16-6 anneals to the 5' end of an Alu element and St19okor primer to the middle part of the ligated adapter. As a result, each molecule in the amplicon contains two common parts at the ends (a 16 bp part of an Alu and a 27 bp adapter which includes the UMI) and a unique genomic flanking sequence for each insertion between (see Fig. 1) them.

Spike-in controls

To monitor subsequent normalization, four artificial DNA fragments were added to the amplicon. These fragments ranging from 240 to 418 bp contain four different sequences from the genome of zebrafish (*Danio rerio*) which have the ends identical to those presented in all other fragments in the amplicon (a 16 bp part of an Alu and a 27 bp adapter introduced by step-out PCR). Two of these fragments (240 bp and 389 bp in length) were added in a concentration corresponding to a somatic insertion that is presented in five out of 50,000 cells whereas two other (259 bp and 418 bp in length) in the concentration corresponding to an insertion that is presented in one out of 50,000 cells (see [Methods](#)). Following the addition of spike-in controls, the mixture was divided into two equal aliquots. One aliquot was sequenced and used as unnormalized control whereas the other one was subjected to normalization using duplex-specific endonuclease.

Normalization using the duplex-specific endonuclease

The amplicon was denatured, renatured and treated by the thermostable duplex-specific endonuclease. During renaturation DNA fragments with high concentration find their complementary chains and anneal to form dsDNA

whereas fragments with low concentration remain single-stranded in the mix. As a result of subsequent digestion by duplex-specific DNase, the majority of highly abundant fragments (corresponding to fixed AluYa5 insertions) were digested whereas rare fragments (including somatic AluYa5 insertions, spike-in controls and previously depleted other Alus such as AluYb8) remained intact. The normalized amplicon was reamplified with the primers used for the second round of selective amplification (AY16-6/St19okor) and again split to two equal portions. The first portion ("normalization 1") was ligated to the Illumina adapters and sequenced. The second portion was subjected to second round of normalization, reamplified ("normalization 2"), ligated to the Illumina adapters and sequenced.

Sequencing and data analysis

Three libraries ("unnormalized", "normalization 1" and "normalization 2") were sequenced using Illumina HiSeq. More than 47 millions of sequencing reads were obtained (see [Table 1](#) for details). The vast majority of reads in the "unnormalized" library represented the sequences flanking AluYa5 insertions. About 80% of reads represented known AluYa5 insertions (annotated in Human Genome Browser, in databases of polymorphic REs and previous studies [22–24]), while 11% of sequences corresponded to the flanks of polymorphic or germline AluYa5 insertions found in the genome of the same donor in our previous study [15]. About 9% of sequencing reads originated from the Alu insertions of other subfamilies. The Alu subfamily composition of normalized libraries significantly changed as a result of the normalization process ([Table 1](#)). As expected the number of sequencing reads comprising highly abundant flanks of known AluYa5 and AluYa8 insertions is decreased while the number or reads corresponding to flanking regions of non-Ya5 Alu copies with low concentration before normalization is increased. The depletion of Ya5 flanks does not affect somatic Alu Ya5 insertions which concentration is also increase in the course of normalization. The identification of potentially somatic insertions was performed as previously described [15, 18]. Briefly, all sequencing reads were mapped to the reference human genome (hg38) and the obtained coordinates were compared to the coordinates of fixed

Table 1 Distribution of sequencing reads

Sequencing reads	Unnormalized	Normalization 1	Normalization 2
Total	13,736,244	16,991,713	16,533,472
Unambiguously mapped to the human genome (hg38)	8,376,753	7,406,323	5,765,382
Corresponding to known AluYa5 and AluYa8	6,484,320	1,180,602	198,024
Number of known AluYa5 and AluYa8	2134	2392	2248
Corresponding to other known Alu	743,310	5,617,692	4,496,177
Corresponding to somatic Alu	56	609	1525

and polymorphic Alu insertions. To filter out the insertions present in all tissues of the donor, the remaining coordinates were compared to the previously identified Alu coordinates from four other tissues (cerebellum, sub-ventricular zone, dentate gyrus and myocardium) of the same individual [18]. Only the insertions that did not match any RE insertion in the human genome and were absent from the other four tissues of the same individual were considered potentially somatic. Additionally, all artificial sequences (e.g. chimeric reads, PCR fragments resulting from mispriming, etc) were filtered out using previously described stringent algorithms [18]. Genomic coordinates, sequencing reads and the distribution of UMIs is shown in Additional file 2.

Evaluation of the method efficiency for library enrichment for somatic RE insertions

The efficiency of normalization was evaluated by direct counting of the number of somatic insertions, sequencing reads and UMIs corresponding to somatic insertions and spike-in controls (see Table 2). The number of identified putative somatic insertions increased more than 3.5-fold (from 47 to 171) after the first round of normalization and 8-fold (from 47 to 378) after the second round compared to the “unnormalized” library. Pearson’s Chi-squared test indicated a significant increase in the proportion of somatic insertions relative to fixed ones ($p = 9.7 * 10^{-5}$ for “unnormalized” versus “normalization 1”; $p = 4.5 * 10^{-13}$ for “normalization 1” versus “normalization 2”; $p < 2.2 * 10^{-16}$ for “unnormalized” versus “normalization 2”). The number of sequencing reads representing somatic insertions increased from 56 in “unnormalized” library to 609 and 1525 after the first and the second rounds of normalization respectively. 38 out of 378 insertions identified in the “normalization 2”

library had more than one UMI indicating that these insertions were initially present in more than one cell. Only one out of four spike-in controls was detected in the “unnormalized” library. Two spike-in controls were identified in the “normalization 1” library whereas three out of four spike-in controls were detected in the “normalization 2” (see Table 2). The number of sequencing reads corresponding to spike-in controls also increased from one in the “unnormalized” to nine in the “normalization 2” library.

We additionally employed quantitative PCR (qPCR) as another method to estimate efficiency of normalization. To this end, we used primer pairs that corresponded to sequences flanking three fixed AluYa5 insertions, four randomly selected somatic insertions having more than one UMI and four spike-in controls (Fig. 2 and Additional file 3). The qPCR data indicated that the concentration of fixed AluYa5 insertions decreased by approximately 4-30 fold after the first round of normalization and by 8-30 fold after the second round (Fig. 2, orange dots). Oppositely, the concentration of spike-in controls increased by 8-30 fold for the ones added in concentration of five cells and by 130-250 fold for the sequences added at concentration corresponding to one cell per 50,000. Thus, the increase in the concentration of spike-in controls depended on the initial abundance in the amplicon before normalization. After the second round of normalization, the concentration of spike-in controls additionally increased by 2-8 fold. (Fig. 2, green dots). Furthermore, the selected somatic insertions initially presented at higher concentrations compared to the spike-in controls were also significantly enriched in the course of normalization (Fig. 2 blue dots). Thus, the ratio between highly abundant and rare sequences of the initial amplicon was greatly decreased by normalization leading to more universal distribution

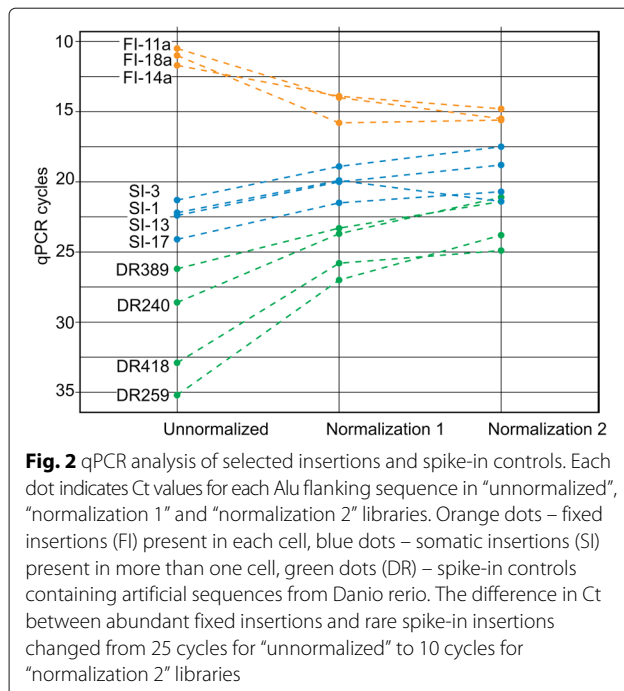
Table 2 Number of sequencing reads and UMIs corresponding to putative somatic insertions and spike-in controls

Putative somatic Alu insertions	Unnormalized	Normalization 1	Normalization 2
Sequencing reads	56(47) ^a	609(218) ^a	1,525(461) ^a
Total number	47	171	378
Number with UMI count > 1	0	24	38
Spike-in controls			
DR240 (in 5 cells ^b)	0	4(3) ^a	7(4) ^a
DR389 (in 5 cells ^b)	1(1) ^a	1(1) ^a	1(1) ^a
DR259 (in 1 cell ^c)	0	0	1(1) ^a
DR418 (in 1 cell ^c)	0	0	0

^aNumber of UMIs is given in parentheses

^bCorresponds to an insertion present in 5 out of 50,000 cells

^cCorresponds to an insertion present in 1 out of 50,000 cells



of RE frequencies in the amplicon. Strikingly, as shown in Fig. 2, the difference between the most abundant and the rarest sequence in our experiment changed from nearly 25 qPCR cycles (that is roughly 33,000,000-fold difference in concentration) to only 10 cycles (corresponding to 1000-fold concentration difference).

Parameters of amplicon library normalization

More generally, the effect of normalization is described by the normalized entropy measure that evaluates distribution uniformity of sequencing reads per insertion (The normalized entropy equals one if each insertion is covered by an equal number of sequencing reads, and asymptotically approaches zero as the reads per insertion count becomes more biased). For the “unnormalized” library, the normalized entropy was estimated at 0.62 (See [Methods](#) section for details). After the first and second rounds of normalization the entropy was increased up to 0.85 and 0.92 respectively. Thus we conclude that normalization makes the distribution of reads per insertions more even and increase the total number of different insertions detected, hence leading to the more efficient discovery of low represented insertions.

Renaturation of an amplicon during normalization is a complex process where many different types of molecules are hybridized to each other. For each group of molecules with the identical nucleotide sequence the speed of renaturation is mainly proportional to concentration although other factors including molecules length and GC content are also important. To evaluate the impact of these

two factors on the normalization efficiency we plotted the number of sequencing reads corresponding to each Alu insertion from Ya5 (highly abundant before normalization) and Yb8 (rare before normalization) subfamilies versus the length of each fragment (Fig. 3a). No relation between fragments length and normalization efficiency was observed. The impact of GC content on the normalization efficiency was more complex (Fig. 3b). We observed a lower normalization rate for AT rich fragments during the first round of normalization. However, during the second round, the normalization rate for AT rich fragments was similar to their counterparts with higher GC content.

Validation of putative somatic insertions

To test the validity of the method for identification of real somatic insertions we randomly selected 12 out of 38 putative insertions (see [Table 2](#)) with UMI number >1. We designed a pair of primers (For1 and For2, see [Additional file 1](#), PCR validation) corresponding to genomic 5' flanking region of each insertion and used them in combination with Alu 5' end specific primer (AY16-6) in two-step semi-nested PCR amplification reaction starting from “normalization 2” library. We also performed the same PCR amplification with the unnormalized library obtained from control non-brain tissue (myocardium) from the same donor. As a result, we obtained PCR products of expected length for 10 out of 12 tested insertions in “normalization 2” but not in control (myocardium) library. One insertion failed to generate expected PCR product and another one was observed in both tissues. The resulting PCR products (see [Fig. 4](#) for electrophorogram) were sequenced with the corresponding genomic primer by Sanger method (See [Additional file 4](#) for the results of amplification and sequencing). All the sequencing reactions confirmed presence of putative Alu insertion with attached 5' unique genomic sequence in the “normalization 2” library of the frontal cortex. To further validate somatic Alu insertions we isolated genomic DNA from another piece of the frontal cortex from the same individual. We used the same two-step nested PCR principle (see [Methods](#) for details) to amplify the 3' adjacent genomic flank for all 12 putative somatic Alu insertions. We failed to detect corresponding 3' adjacent genomic flanks for all 12 insertions using this approach. Based on these results we are unable to confidently prove and claim the somatic Alu insertions in the human brain.

Discussion

Somatic mosaicism resulting from new RE insertions was proposed to play a significant role in adult organism in particular contributing to individual neurons plasticity [8, 25]. RE activity might also be involved in brain

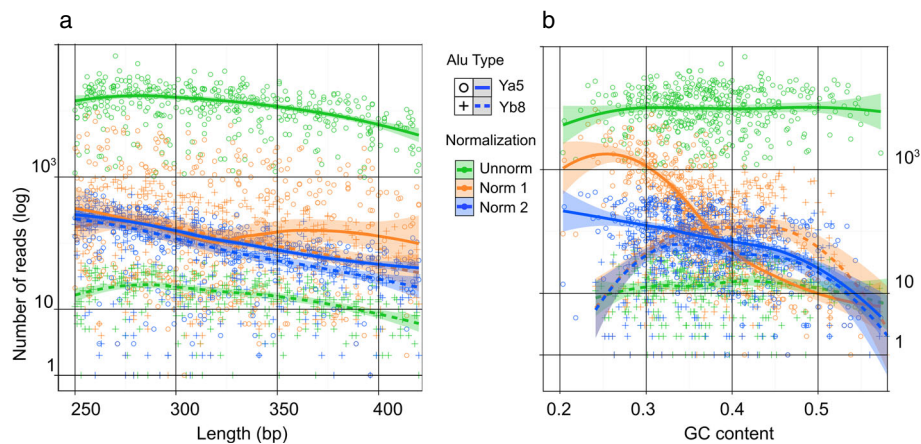


Fig. 3 Effect of fragment length and GC content on normalization efficiency. The number of sequencing reads of rare (corresponding to AluYb8) and abundant (corresponding to AluYa5) flanks is plotted against fragment length **(a)** and GC content **(b)**, respectively, in “unnormalized”, “normalization 1” and “normalization 2” datasets. Y axis – number of reads (logarithmic scale). X axis is length of fragments **(a)** or their GC content **(b)**. Green circles and green crosses indicate Ya5 and Yb8 insertions in “unnormalized” library; orange circles and orange crosses indicate Ya5 and Yb8 insertions in “normalization 1” library; blue circles and blue crosses indicate Ya5 and Yb8 insertions in “normalization 2” library. Trendlines were fit to data using generalized additive models, shaded area indicate confidence interval (CI=0.95) for trendlines

disorders including Rett syndrome [9] and schizophrenia [26]. The most valid method to find new somatic RE insertions is their direct detection by high-throughput sequencing of genomic DNA. Although the capacity of modern sequencing platforms is rapidly increasing it is still expensive to study the distribution of somatic RE insertions (especially rare) in thousands of individual cells or many tissue samples. Even with the use of current protocols for enrichment in RE sequences only a minor fraction (up to 0.01% [15, 27]) of HTS reads is comprised by the somatic elements. In this study, we propose a tool that can significantly improve the capacity of most methods to identify rare somatic RE insertions. The entire process suppose two types of enrichment procedures: (1) selection of sequences flanking RE insertions of a particular subfamily by one of existing methods and (2) enrichment for sequences representing rare (including potentially somatic) insertions (normalization). The better results at the first enrichment stage are achievable using vectorette PCR [11] or suppressive PCR techniques [15]. As a result more than 90% of the final amplicon is comprised by DNA fragments that flank RE insertions of the selected subgroup. During the second enrichment stage (employed in this study) highly abundant fragments are diminished in the amplicon, while rare sequences (including the fragments corresponding to potentially somatic RE insertions) are enriched. Thus, two successive rounds of normalization led to more than 26-fold increase in the number of potentially somatic REs flanks in a sequenced sample. The efficiency of this strategy is confirmed by both direct sequencing and qPCR of rare insertions and spike-in controls.

Along with a more than 26-fold increase in the number of sequencing reads representing potentially somatic REs, the number of identified insertions increased by 7.9-fold (from 48 to 379) and the UMI number increased by 9.2-fold (from 51 to 468). The difference between the increments of the sequencing reads and potentially somatic insertions might be explained by the limited number of somatic REs present in 50,000 cells. Therefore, the enrichment by normalization increases the number of reads, while the number of identified insertions starts reaching a plateau.

In this study we employed two successive rounds of normalization. The first normalization round resulted in a 10-fold increase in the number of sequencing reads corresponding to potentially somatic insertions and 3.5-fold increase in the number of identified insertions. After the second round of normalization there was an additional 2.5-fold increase in both the number of reads and the number of potentially somatic insertions. The difference in the efficiency of the first and second rounds of normalization probably reflects the principal limitation of the method of enrichment for low abundant fragments under selected conditions (renaturation time and DNA concentration).

UMIs are increasingly applied in the HTS-based methods to reduce the biasing effect of PCR and sequencing on quantitative information about particular sequences in the initial sample and to correct for PCR and sequencing errors [16, 17, 28]. For instance, UMIs were used recently for the quantitative assessment of T cell repertoire diversity in course of aging [29]. Although deep oversequencing is usually required for the accurate estimation of UMI

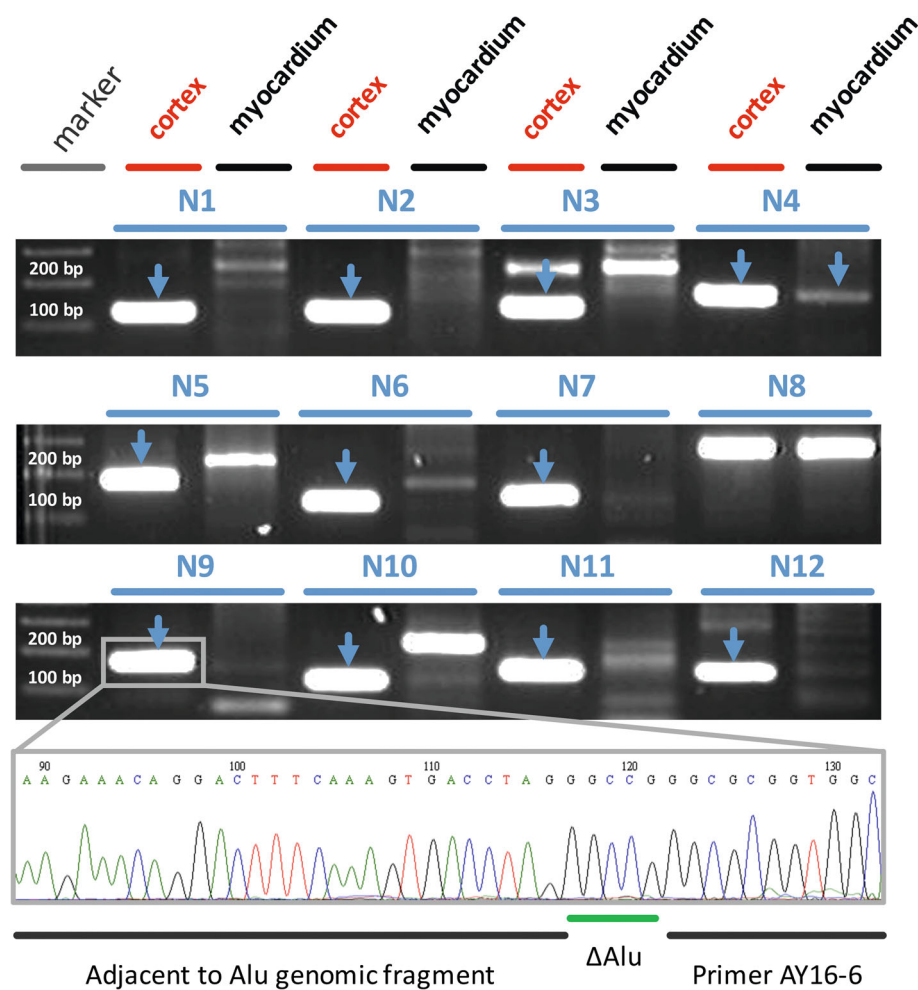


Fig. 4 Results of putative somatic Alu locus-specific PCR. Marker – 100 bp DNA ladder (Evrogen); cortex – results of semi-nested PCR with “normalization 2” library; myocardium – results of semi-nested PCR with control library (myocardium from the same donor); N1-N12 somatic insertions with UMI count more than 1. Blue arrows indicate PCR products of expected size. For N4 insertion amplification product of expected size was observed in both “normalization 2” and control myocardium PCR reactions. For N8 insertion amplification product of expected size was not detected neither in “normalization 2” nor in control myocardium PCR reaction. An example (N9 insertion) of capillary sequencing result is shown at the bottom

based events [28], some unique quantitative traits could be obtained even with smaller sequencing depth. Here we ligate UMIs before introducing any quantitative bias by selective PCR or bridge amplification on the solid phase of the Illumina sequencing machine. Thus, the number of UMIs ligated to the fragments with identical sequences corresponds to the number of cells bearing this particular insertion.

In this study we found 38 potentially somatic AluYa5 insertions (Table 2) which were characterized by more than one UMI per insertion. Therefore, these ones represent the most promising pool of putative somatic insertions detected in this study. However, we were unable to validate these insertions by direct amplification and Sanger sequencing of both 5' and 3' insertion flanks. The final validation of these insertions can be done by

identification of target site duplication (TSD) - the main characteristic signature of retroposition event [30]. Thus, the phenomenon of retrotranspositional Alu activity in the human brain remains to be confirmed in the future studies. Simultaneous sequencing of both 5' and 3' RE flanks for TSD identification as well as the developed normalization based enrichment technique could significantly improve existing methods for the rare somatic RE insertions profiling.

Conclusions

Somatic RE activity in humans and other mammals has been intensively studied over the last several years. Several studies reported a significant rate of insertional mutagenesis mediated by de novo integrations of REs not only in cancer, but also in normal human tissues including

the brain. However, current enrichment protocols do not provide enough power for the detection of novel RE integrations and thus the sensitivity for somatic RE detection is usually enhanced by increasing the number of sequencing reads, which is cost consuming. The described approach can increase the efficiency of existing RE identification methods decreasing the number of sequencing reads required for the confident estimation of somatic REs abundance. Furthermore, the method allows to analyze much larger samples (tens of thousands cells) than usually studied nowadays (from 1 to hundreds of cells) with an almost comprehensive identification of very rare somatic RE insertions. The use of UMIs provides quantitative information on the distribution of REs. The direct estimation of the number of cells bearing each particular insertion can provide information on the period of RE retroposition activity in studied tissues, which could be linked to the stage of the disease progress or normal tissue development.

Methods

Nuclei isolation and DNA extraction

100 mg frozen tissue from postmortal human cortex (72 year old male individual) was used for nuclei isolation. All following manipulations were performed on ice. Tissue sample was homogenized in Dounce tissue grinder in 10 ml of nuclei extraction buffer (10 mM Hepes, 3 mM $MgCl_2$, 5 mM $CaCl_2$, 0.32 M sucrose, 0.2% Triton X-100). Homogenate was layered over equal volume of sucrose solution (0.64 M sucrose, 1× PBS, 0.2% Triton X-100) and centrifuged for 15 min at 1600 g, +4 °C. The sediment was resuspended in 1 ml 1× PBS and centrifuged for 10 min at 450 g, +4 °C. The obtained nuclei fraction was resuspended in 200 μ l 1× PBS, stained by trypan blue and counted in hemocytometer. A portion of the suspension containing approximately 50,000 nuclei was taken for downstream analysis starting from genomic DNA extraction by standard phenol/chloroform method.

AluYa5 flanking fragments library preparation

Genomic DNA was digested by incubation with AluI (Fermentas) endonuclease (10 U) for 12 hours. Fragmented DNA was purified by AmPure XP beads (Beckman Coulter) and ligated to suppressive adapters. The 10 μ l ligation mixture contained 50 pmoles of each st19BH and st20BH adapters, 10 U of T4 DNA ligase in a T4 reaction buffer (both Promega) and digested genomic DNA. The reaction was carried out overnight at +4 °C. Ligated fragments were incubated for 2 hours with 3 U of restriction enzyme AluI in 1× Y Tango buffer to decrease the number of chimeric molecules. Restriction products were purified using QIAquick PCR Purification Kit (Qiagen).

DNA amplification for library preparation was performed in two subsequent suppression PCR steps.

Each of 20 first step PCR reaction (25 μ l) contained 1/20 of the total amount of ligation reaction, 0.4 μ M AluYa5 specific primer (AY107), 0.16 μ M Na21 primer, dNTPs (0.125 μ M each), 1 U of Tersus polymerase in 1× Tersus Plus buffer (both Evrogen). The amplification profile was as follows: 72 °C for 4 min, followed by 12 cycles of 20 s at 94 °C, 15 s at 65 °C and 1 min at 72 °C. PCR products were combined, purified with the QIAquick PCR Purification Kit (Qiagen). Each of two second step PCR reaction (25 μ l) contained 1/160 of the first PCR products, 0.4 μ M of each AY16-6 and st19okor primers, dNTPs (0.5 μ M each), 1 U of Tersus polymerase in 1× Tersus Plus buffer. The amplification profile was as follows: 20 s at 94 °C, 15 s at 60 °C, 1 min at 72 °C, 9 cycles. PCR product was purified and loaded on agarose gel. Fragments ranging from 250 to 450 bp were cut and purified using QIAquick Gel Extraction kit (Qiagen).

Spike-in controls preparation

Four different loci of zebrafish genome were selected for the preparation of artificial spike-in controls. Four different PCR reactions (25 μ l) containing 20 ng of zebrafish genomic DNA, dNTPs (0.125 μ M each), 1 U of Tersus polymerase and 0.4 μ M of each DR primers (see Additional file 1, primers for spike-in preparation) in 1× Tersus Plus buffer were performed. Forward primer contained the 16 nucleotides of AluYa5 at the 5' end. The amplification profile was as follows: 20 s at 94 °C, 15 s at 60 °C, 1 min at 72 °C, 9 cycles. Obtained PCR products were phosphorylated using T4 polynucleotide kinase (Promega) in the appropriate buffer. Phosphorylated PCR products were ligated to St19BH/St20BH adapter as described above. On the last step PCR reaction with ligated fragments and 0.4 μ M of each AY16-6/St19okor primers was performed. PCR products were purified by Cleanup mini PCR Purification Kit (Evrogen) and their concentration was measured by Qubit. As a result four DNA fragments with the ends identical to those of the constructed AluYa5 flanking fragments library and having four different flanking sequences 240, 259, 389 and 418 bp long inside were obtained. 0.6×10^{-9} ng of DR259, 1×10^{-9} ng of DR418, 2.2×10^{-9} ng of DR240 and 3.6×10^{-9} ng of DR389 were added to 4.2 ng of AluYa5 flanking fragments library that corresponds to the insertions present in one (DR259 and DR418) or 5 (DR240 and DR389) out of 50,000 cells. AluYa5 flanking fragments library with added spike-in controls hereafter is called DNA mix 1.

Normalization with Kamchatka Crab duplex-specific nuclease (DSN)

An aliquot (1/6 part) of the obtained DNA mix 1 were used for "unnormalized" control library preparation. Each of 5 PCR reaction tubes (25 μ l) contained 1/30 of the DNA mix 1, 0.8 μ M of each AY16-ind301 (contains sample

barcode 301) and st19okor primers, 0.25 μM each of dNTPs, 1 U of Encyclo polymerase in the 1 \times Encyclo reaction buffer (both Evrogen). The amplification profile was as follows: 9 cycles of 20 s at 94 $^{\circ}\text{C}$, 15 s at 60 $^{\circ}\text{C}$, 1 min at 72 $^{\circ}\text{C}$. PCR products were combined and purified using QIAquick PCR Purification Kit (Qiagen).

Same volume aliquot of DNA mix 1 was subjected to PCR as described above except for primers used for amplification (AY16-6 without sample barcode and st19okor, 13 cycles). 480 ng (3 μl) of the purified PCR product was mixed with 1 μl of 4 \times Hybridisation Buffer (200 mM HEPES pH 7.5, 2M NaCl). Reaction mixture was overlaid by mineral oil drop, denatured at 97 $^{\circ}\text{C}$ for 3 min, chilled to 76 $^{\circ}\text{C}$ with ramp 0.1 $^{\circ}\text{C}/\text{s}$ and renatured at 76 $^{\circ}\text{C}$ for 4 hours. After renaturation 5 μl of 2 \times DSN Master Buffer and 1 μl (1 U/ μl) of DSN solution (both Evrogen), preheated at 76 $^{\circ}\text{C}$, were added to the reaction consecutively. Incubation was continued at 76 $^{\circ}\text{C}$ for 15 min. 10 μl of 2 \times Stop Solution (Evrogen) was added to the reaction to inactivate DSN. The resulted normalization product was immediately purified using AMPure XP beads (Beckman Coulter, USA) and redissolved in 30 μl of water.

First aliquot (15 μl) was reamplified with AY16-ind302/st19okor primers and Encyclo polymerase for 9 cycles as described above resulting in “normalization 1” library. Second aliquot (15 μl) was reamplified with AY16-6/st19okor primers and used for second normalization as described above except of higher DNA concentration (1800ng in 3 μl). After the second normalization DNA was purified using AMPure XP beads and reamplified with AY16-ind304/st19okor primers and Encyclo polymerase for 9 cycles as described above resulting in “normalization 2” library.

Sequencing and data analysis

Three libraries (“unnormalized”, “normalization 1” and “normalization 2”) each containing sample barcode were ligated to Illumina Truseq adapters using standard protocol and sequenced on HiSeq 2000 platform (paired end 2 \times 100). Data analysis includes four main stages: 1) initial fastq files processing; 2) mapping to the reference human genome; 3) fixed, polymorphic and germline insertions filtration and 4) artifacts removal. During initial fastq processing we remove sequences introduced in course of library preparation (primers, adapters and UMIs) and also the 5 bp 5' Alu fragment from the reads. Sequences of UMIs and 5 bp 5' Alu fragments were kept for subsequent analysis. Processed reads were mapped to the reference human genome (hg38 assembly) using Bowtie2 software with the default parameters. For the downstream analysis we extracted only coordinates of concordantly and uniquely mapped reads. Identical coordinates were merged and then intersected with the coordinates of all known Alu insertions from

hg38, dbRIP and our own sequencing datasets including samples obtained from lab members. The coordinates that did not match to any of known Alu were intersected with control tissue libraries (cerebellum, subventricular zone, dentate gyrus, and myocardium) from the same individual. Next we removed various artifacts of sample preparation including: a) sequencing reads containing restriction sites AluI (chimeras formed during ligation); b) sequencing reads mapping to the regions containing restriction sites (chimeras formed during ligation with PCR or sequencing errors in AluI restriction site); c) putative insertions located in immediate proximity to genomic non-Alu annealing site for AY16-6 primer (result of mispriming). We used threshold of 4 mismatches since more than 95% of randomly selected genomic 11mers (the length of AY16-6 primer) has more than 4 mismatches; d) putative insertions having the first 12 nucleotides of the flank identical to the sequences of known Alu insertions flanks (results of template switch during PCR); e) putative insertions with more than one mismatch from Ya5 consensus (GGCCG) in the 5 bp 5' Alu fragment. The remaining coordinates were considered as sites of putative somatic insertions.

Statistical analysis

To evaluate the statistical significance of sequencing library enrichment by putative somatic insertions we applied Pearson's Chi-squared test. The P values were calculated using the `chisq.test` function from R [31]. The normalized entropy measure on a distribution of reads per insertion for a sample was calculated using the following formula:

$$H_n = \sum_{i=0}^n p_i * \log_2(p_i) \div \log_2(|D|),$$

where H_n is normalized entropy, p_i is a proportion of reads in the i -th insertion to the overall number of reads, $|D|$ is a size of the distribution (total number of identified insertions).

To correct sequencing errors in UMIs corresponding to each putative somatic Alu insertion we built a graph where UMI sequences were vertices and hamming distances between them were edges. Each strongly connected component in the graph with one “parental” UMI was deleted. Number of remaining vertices was considered as a corrected number of UMIs in the input set for each particular somatic RE insertion.

Quantitative PCR analysis of selected AluYa5 insertions and spike-in controls

qPCR was performed to measure relative quantities of three fixed, four selected somatic and four artificial

spike-in AluYa5 insertions. Each pair of primers was designed to align to unique genomic region between 5' end of the Alu element and nearest AluI restriction site. Each of 15 μ l PCR reactions contained 2.5 ng of template DNA ("unnormalized", "normalization 1" or "normalization 2" libraries), 0.17 μ M of each direct and reverse primers (see Additional file 1, primers for qPCR) in 1 \times qPCR-HS SYBR mix (Evrogen). Three technical replicates for each PCR reaction were performed. The changes in relative quantities were evaluated using delta-delta Ct method.

Amplification of putative somatic Alu insertions and Sanger sequencing

For 5' flank: First multiplex PCR reactions (25 μ l) contained 6 ng of the template DNA ("normalization 2" library), 0.2 μ M of each of 12 genomic For1 primers (see Additional file 1) and 0.2 μ M of Alu specific primer (AY16-6), 0.25 μ M each of dNTPs, 1 U of Encyclo polymerase in the 1 \times Encyclo reaction buffer (both Evrogen). The amplification profile was as follows: 20 cycles of 20 s at 94 $^{\circ}$ C, 15 s at 60 $^{\circ}$ C, 1 min at 72 $^{\circ}$ C. 2 μ l of 25-fold diluted PCR product was used as a template in each of 12 second (semi-nested) PCR reactions. Each of 12 separate reactions contained all the same components except that corresponding genomic For1 was replaced with nester For2 primer for each tube/locus. The amplification profile was as follows: 30 cycles of 20 s at 94 $^{\circ}$ C, 15 s at 60 $^{\circ}$ C, 1 min at 72 $^{\circ}$ C. Control library (unnormalized myocardium from the same donor) was amplified in exactly the same way. PCR products were loaded on 1,5% agarose gel and purified using QIAquick Gel Extraction kit (Qiagen). For 3' flank: genomic DNA was isolated from 50,000 nuclei obtained from another piece of frontal cortex as described above. First multiplex PCR reactions (50 μ l, same amplification profile as for 5' flank, 35 cycles) contained 300 ng of the template genomic DNA, 0.2 μ M of each of 12 genomic Rev1 primers (see Additional file 1) and 0.2 μ M of Alu specific primer (AY102), 0.25 μ M each of dNTP, 2 U of Encyclo polymerase in the 1 \times Encyclo reaction buffer (both Evrogen). 2 μ l of 25-fold diluted PCR product was used as a template in each of 12 second (nested) PCR reactions (25 μ l, same amplification profile as for 5', 30 cycles). Each of 12 separate reactions contained all the same components except that AY237 primer was added instead of AY102 and corresponding genomic Rev1 was replaced with nester Rev2 primer for each tube/locus. Each of the purified PCR products (10 for 5' flank and 4 for 3' flank) was sequenced with the corresponding For2 or Rev2 genomic primer on ABI PRISM 3500 (Applied Biosystems).

Additional files

Additional file 1: This file contains a table with sequences of oligonucleotide primers used in this study. (PDF 70 kb)

Additional file 2: This file contains a table with coordinates of identified putative somatic RE insertions in "unnormalized", "normalization 1" and "normalization 2" libraries. (XLS 84 kb)

Additional file 3: This file contains a table with result of qPCR analysis of selected insertions and spike-in controls. (PDF 37 kb)

Additional file 4: This file contains a table with results of PCR validation and capillary sequencing of 12 randomly selected putative somatic Alu insertions. (XLS 70 kb)

Abbreviations

HTS: High-throughput sequencing; qPCR: Quantitative polymerase chain reaction; RE: Retroelement; UMI: Unique molecular identifier

Acknowledgements

We thank reviewers for their comments on the manuscript.

Funding

This work was supported by Russian Science Foundation grant RSF 18-14-00244 to IZM and partially supported by Russian Foundation for Basic Research grants [RFBR 16-04-00779 to IZM (sequencing data analysis), RFBR 17-04-01280 and RFBR 16-34-01100 to AYK (spike-in preparation and UMI technology testing)]. AYK is personally supported by Russian President's Fellowship [SP-4059.2016.4].

Availability of data and materials

The raw Illumina sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) with accession number SRX1113412.

Authors' contributions

IZM conceived the idea for the study. IZM and AYK developed experiment scheme and designed oligonucleotides. AYK, AAM and MVS performed experiments. GAN, IZM and AYK analysed sequencing data. IZM, AYK, KVK and YBL prepared the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study was approved by the local ethics committee and conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Miklukho-Maklaya str. 16/10, 117997 Moscow, Russia. ²Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Samory Mashela str. 1, 117997 Moscow, Russia. ³Biotech Research and Innovation Centre, Copenhagen University, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. ⁴Pirogov Russian National Research Medical University, Ostrovitianov str. 1, 117997 Moscow, Russia.

Received: 9 April 2018 Accepted: 15 October 2018

Published online: 31 October 2018

References

1. Iourov I, Vorsanova S, Yurov Y. Somatic Genome Variations in Health and Disease. *Curr Genomics*. 2010;11(6):387–96. <https://doi.org/10.2174/138920210793176065>.

2. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP. Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci*. 2012;109(44):18018–23. <https://doi.org/10.1073/pnas.1213736109>.
3. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, Walsh CA. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Rep*. 2014;8(5):1280–9. <https://doi.org/10.1016/j.celrep.2014.07.043>.
4. Iskov RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. *Cell*. 2010;141(7):1253–61. <https://doi.org/10.1016/j.cell.2010.05.020>.
5. Lee E, Iskov R, Yang L, Gokcumen O, Haseley P, Luquette LJ, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs R, Kucherlapati RA, Lee C, Kharchenko PV, Park PJ. Landscape of Somatic Retrotransposition in Human Cancers. *Science*. 2012;337(6097):967–71. <https://doi.org/10.1126/science.1222077>.
6. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, Wheeler S, Upton KR, Shukla R, Faulkner GJ, Largaespada DA, Kazazian HH. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res*. 2012;22(12):2328–38. <https://doi.org/10.1101/gr.145235.112>.
7. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH. L1 retrotransposition in human neural progenitor cells. *Nature*. 2009;460(7259):1127–31. <https://doi.org/10.1038/nature08248>.
8. Kuwabara T, Hsieh J, Muotri A, Yeo G, Warashina M, Lie DC, Moore L, Nakashima K, Asashima M, Gage FH. Wnt-mediated activation of NeuroD1 and retroelements during adult neurogenesis. *Nat Neurosci*. 2009;12(9):1097–105. <https://doi.org/10.1038/nn.2360>.
9. Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, Gage FH. L1 retrotransposition in neurons is modulated by MeCP2. *Nature*. 2010;468(7322):443–6. <https://doi.org/10.1038/nature09544>.
10. Xing J, Witherspoon DJ, Jorde LB. Mobile element biology: new possibilities with high-throughput sequencing. *Trends Genet*. 2013;29(5):280–9. <https://doi.org/10.1016/j.tig.2012.12.002>.
11. Tang Z, Sterankac JP, Maa S, Grivainisa M, Rodic NR, Huang CRL, Shihc IM, Wangc TL, Boeke JD, Fenyöa D, Burns KH. Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. *Proc Natl Acad Sci*. 2017;114(5):733–40. <https://doi.org/10.1073/pnas.1619797114>.
12. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, Ewing AD, Salvador-Palomeque C, van der Knaap MS, Brennan PM, Vanderver A, Faulkner GJ. Ubiquitous L1 Mosaicism in Hippocampal Neurons. *Cell*. 2015;161(2):228–39. <https://doi.org/10.1016/j.cell.2015.03.026>.
13. Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, Manda SS, Abril G, Pereira G, Makohon-Moore A, Looijenga LHJ, Gillis AJM, Hruban RH, Anders RA, Romans KE, Pandey A, Iacobuzio-Donahue CA, Vogelstein B, Kinzler KW, Kazazian HH, Solyom S. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res*. 2015;25(10):1536–45. <https://doi.org/10.1101/gr.196238.115>.
14. Kvikstad EM, Piazza P, Taylor JC, Lunter G. A high throughput screen for active human transposable elements. *BMC Genomics*. 2018;19(1):115. <https://doi.org/10.1186/s12864-018-4485-4>.
15. Kurnosov AA, Ustyugova SV, Nazarov VI, Minervina AA, Komkov AY, Shugay M, Pogorelyy MV, Khodosevich KV, Mamedov IZ, Lebedev YB. The Evidence for Increased L1 Activity in the Site of Human Adult Brain Neurogenesis. *PLoS ONE*. 2015;10(2):e0117854. <https://doi.org/10.1371/journal.pone.0117854>.
16. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci*. 2011;108(23):9530–5. <https://doi.org/10.1073/pnas.1105422108>.
17. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*. 2012;9(1):72–4. <https://doi.org/10.1038/nmeth.1778>.
18. Kurnosov AA, Ustyugova SV, Pogorelyy MV, Komkov AY, Bolotin DA, Khodosevich KV, Mamedov IZ, Lebedev YB. A novel approach to identification of somatic retroelements' insertions in human genome. *Russ J Bioorg Chem*. 2013;39(4):417–25. <https://doi.org/10.1134/S1068162013040110>.
19. Mamedov I, Batrak A, Buzdin A, Arzumanyan E, Lebedev Y, Sverdlov ED. Genome-wide comparison of differences in the integration sites of interspersed repeats between closely related genomes. *Nucleic Acids Res*. 2002;30(14):71.
20. Mamedov IZ. Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach. *Nucleic Acids Res*. 2005;33(2):16. <https://doi.org/10.1093/nar/gni018>.
21. Shagin DA, Rebrikov DV, Kozhemyako VB, Altschuler IM, Shcheglov AS, Zhulidov PA, Bogdanova EA, Sraroverov VA, Lukyanov S. A novel method for SNP detection using a new Duplex-Specific Nuclease from Crab hepatopancreas. *Genome Res*. 2002;12(12):1935–42. <https://doi.org/10.1101/gr.547002.1>.
22. Witherspoon DJ, Zhang Y, Xing J, Watkins WS, Ha H, Batzer MA, Jorde LB. Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome Res*. 2013;23(7):1170–81. <https://doi.org/10.1101/gr.148973.112>.
23. Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, Urban AE, Grubert F, Lam HYK, Lee W-P, Busby M, Indap AR, Garrison E, Huff C, Xing J, Snyder MP, Jorde LB, Batzer MA, Korbel JO, Marth GT. A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genet*. 2011;7(8):e1002236. <https://doi.org/10.1371/journal.pgen.1002236>.
24. Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakhshi M, Sahinalp SC, Eichler EE. Alu repeat discovery and characterization within human genomes. *Genome Res*. 2011;21(6):840–9. <https://doi.org/10.1101/gr.115956.110>.
25. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci*. 2014;15(8):497–506. <https://doi.org/10.1038/nrn3730>.
26. Bundo M, Toyoshima M, Okada Y, Akamatsu W, Ueda J, Nemoto-Miyauchi T, Sunaga F, Toritsuka M, Ikawa D, Kakita A, Kato M, Kasai K, Kishimoto T, Nawa H, Okano H, Yoshikawa T, Kato T, Iwamoto K. Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. *Neuron*. 2014;81(2):306–13. <https://doi.org/10.1016/j.neuron.2013.10.053>.
27. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustinchin S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, c Faulkner GJ. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*. 2011;479(7374):534–7. <https://doi.org/10.1038/nature10531>.
28. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, Bolotin DA, Staroverov DB, Putintseva EV, Plevova K, Linnemann C, Shagin D, Pospisilova S, Lukyanov S, Schumacher TN, Chudakov DM. Towards error-free profiling of immune repertoires. *Nat Methods*. 2014;11(6):653–5. <https://doi.org/10.1038/nmeth.2960>.
29. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, Bolotin DA, Lukyanov S, Bogdanova EA, Mamedov IZ, Lebedev YB, Chudakov DM. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *J Immunol*. 2014;192(6):2689–98. <https://doi.org/10.4049/jimmunol.1302064>.
30. Evrony GD, Lee E, Park PJ, Walsh CA. Resolving rates of mutation in the brain using single-neuron genomics. *eLife*. 2016;5:e12966. <https://doi.org/10.7554/eLife.12966>.
31. Team RC. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2013. <http://www.R-project.org/>.