

largest clones of S-1 and S-2 ((shear sites vs. tags): (209 vs. 393) and (119 vs. 142)) (Figure 1D and Figure 2D). In all four samples, those variations were also similar in the minor clones of which the clone sizes did not exceed shear sites variations (approximately <200 variations) (See Additional file 1: Table S3 and Additional file 2: Table S1 for information on the ten largest clones). However, the variations covered by tags were significantly greater than those of shear sites, especially for large clones like those observed in the major clones of S-3 and S-4 ((shear sites vs. tags): (242 vs. 1751) and (222 vs. 2675)). The variations covered by tags and combinations were almost the same for all four samples ((tags vs. combinations): (393 vs. 296), (142 vs. 119), (1751 vs. 1192), and (2675 vs. 2038)).

Upon comparison of the tag system data with the shear site data, it was clear that both strategies yield essentially the same results when the size of clones is small enough to be covered by the number of shear site variations generated. However, the tag system provides a much better estimation of clonality when the number of sister cells in each clone exceeds shear site variations. Therefore, clone size was underestimated when considering only shear sites in expanded clones like samples S-3 and S-4. Given this, our tag system should be used for samples with different clonality status to avoid underestimation of the size of clones. See Additional file 2: Figure S3 for a simple comparison of shear site and tag variations.

Validation of the methodology

Our newly developed method - the tag system and the related data analysis - were successfully validated, internally. As mentioned above, the initial validation was done by analyzing samples from different HTLV-1-infected individuals (Figures 1 and 2). Finally, we conducted a comprehensive internal validation by using an appropriate control with known integration sites and clonality patterns to provide direct evidence for the effectiveness of our system in the clonality analysis. We designed a suitable control because there was not an appropriate control available. Using our system, we could evaluate the method and confirm its accuracy, sensitivity, and reproducibility. We selected two samples with the following special conditions as starting materials for preparing the control system.

Sample one (M): DNA from an acute ATL patient with 100% PVLs and a single integration site in the major clone (Figure 3A). The integration site of this sample was first checked with conventional splinkerette PCR, which detected a single major integration site. Subsequently, deep-sequencing data (tags only and combinations) showed that approximately 99% of the PVL accounted for the major clone with an integration site at

chromosome 12:94976747(-). A small numbers of clones occupied approximately 1% of the PVL of this sample. Those clones were only detected in the second trial samples for which the external PCR products were not diluted. Therefore, to simplify the overall analysis, we removed those low-abundance clones (data not shown).

Sample two (T): DNA was isolated from a fresh culture of TL-Om1, which is a registered monoclonal ATL cell line with 100% PVL and a single integration site at chromosome 1:121251270(-) in each cell (Figure 3A).

Having prepared these two samples, they were sonicated and mixed in proportions of 50:50 and 90:10 (Figure 3B). These known proportions were thus expected to generate specific patterns that could be verified with our subsequent analysis. We conducted two independent sets of trials.

In the first trial, samples were named as 'first trial control 1 ~ 4' and abbreviated as 1st T-cnt-1 ~ 4. Various amounts of DNA (μg) from samples M and T were mixed to prepare the final expected clone sizes as shown in Figure 3C. A 1- μL sample of a 10-fold dilution of external PCR product was used as the starting material for nested PCR for this trial. The samples were run in separate lanes of HiSeq 2000.

We named the samples of the second trial as second trial control-1 ~ 4 and abbreviated them as 2nd T-cnt-1 ~ 4. DNA samples were mixed similarly to that for the first trial except for sample four (Figure 3D). In contrast to the first trial, we used 1 μL of the external PCR product without any dilution as a starting material for the nested PCR. These samples were multiplexed and run in the same lane of HiSeq 2000. The purpose of the second trial was to test both method reproducibility and the effect that the dilutions had on the results.

The samples of both the first and second trials were analyzed under the same conditions, except where noted above. For each control sample, expected patterns and experimentally observed patterns were calculated for (a) raw sequence reads, (b) shear sites, (c) only tags, and (d) the combination of tags and shear sites (Figure 4). Figure 4 shows the data when the optimal conditions were considered. Additional file 1: Figure S3 includes most of the data accumulated during optimization of the method.

Evaluating the accuracy of the clonality analyzed based on shear sites vs. tags system

The 'absolute error', a technique used to evaluate system accuracy [61], was used to assess our method. The experimental values were subtracted from expected values (Figure 5A). Taking advantage of our control system (the first and second trial samples), the clone size was calculated by considering (a) sequencing reads without removing PCR duplicates, (b) only shear sites, (c) only tags, and (d) the combination of tags and shear sites (Figure 5B and C). The absolute errors of raw sequence reads for the first trial

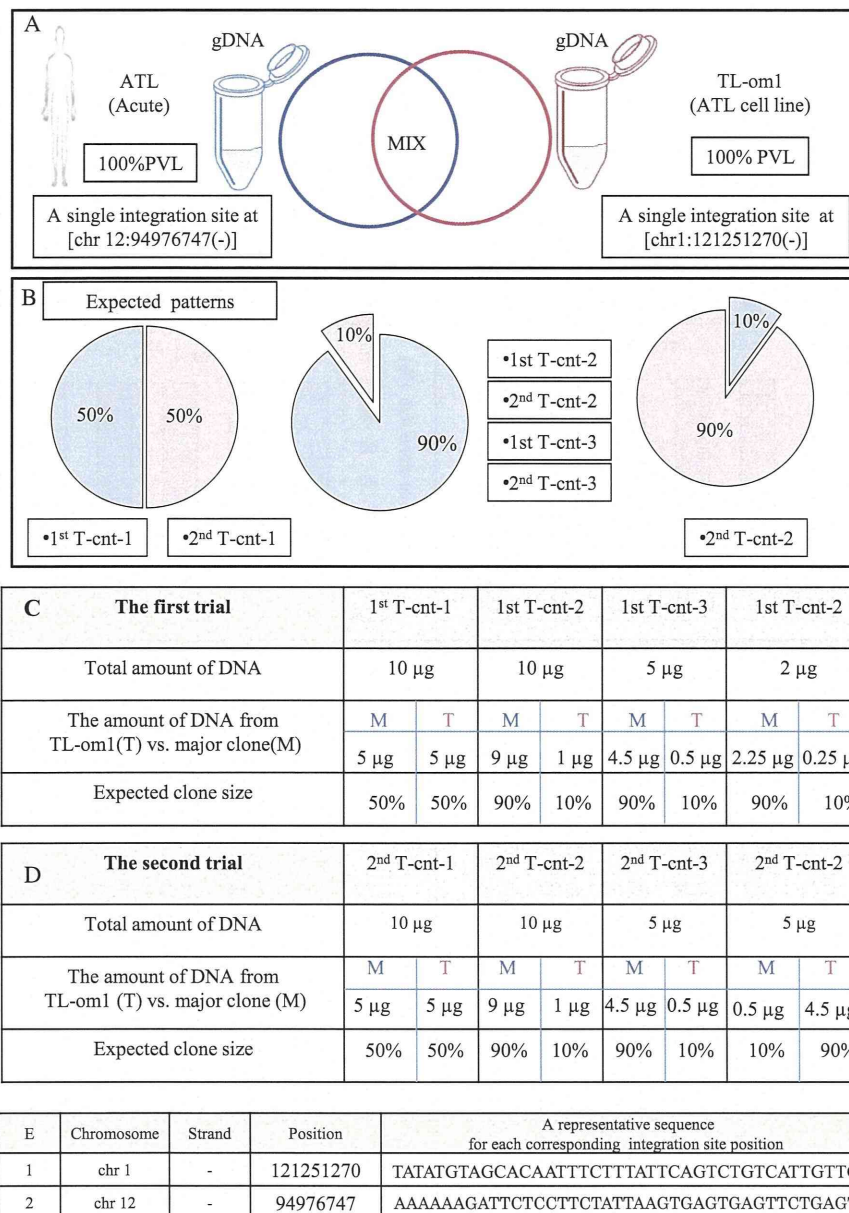


Figure 3 Preparing the control system. (A) The control system was designed by mixing sonicated genomic DNA (gDNA) of TL-Om1 with that of an ATL patient in proportions of 50:50 and 90:10. TL-Om1 is a standard ATL cell line with 100% PVL and a known single integration site at (chr1:121251270(-)). The patient sample was from an acute type of ATL with 100% PVL and a single integration site at (chr 12:94976747(-)). (B) The expected clonality patterns: (50% vs. 50%), (90% vs. 10%), and (10% vs. 90%) were generated by mixing gDNA from an ATL sample with that from TL-Om1. (C, D) Full details of the first trial's and the second trial's samples including: name of samples, total amount of DNA (µg), the amount of DNA (µg) from TL-Om1 (T) vs. major clone (M), and expected clone size are provided. (E) Integration site position of TL-Om1 and the major clone of ATL sample.

samples were 23.58, 6.26, 4.57, and 5.72, whereas those of the second trial samples were 44.66, 9.50, 6.88, and 60.24. The magnitude of errors in the first trial was lower than that of the second trial probably due to the dilution of the external PCR products in the first trial. However because dilution reduced the number of covered integration sites, it

should be done sparingly and with the purpose of the experiments in mind. The errors when considering only shear sites were 1.72, 34.33, 21.76, and 18.73 for the first trial and 0.47, 38.29, 36.72, and 40.47 for the second trial. Underestimations caused by low shear site variation did not affect the relative size of clones when the expected size of the

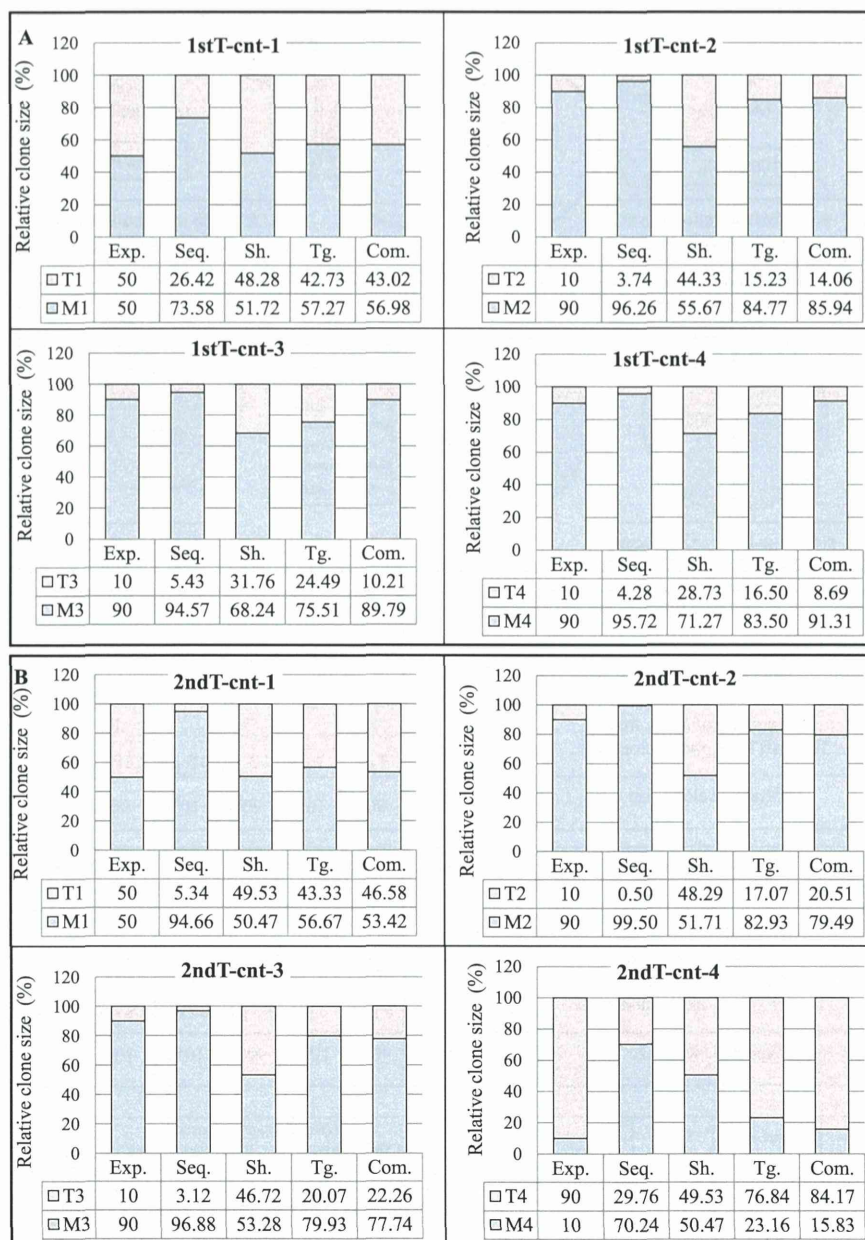


Figure 4 Validation of the tag system. For each control sample, both the expected and the experimentally observed patterns of raw sequence reads, shear sites, and the combination of tags and shear sites are represented in the bar graphs. Abbreviations: Com.: Combinations, Exp.: expected pattern, Seq.: raw sequencing data without removing PCR duplicates, Sh.: Shear sites, Tg.: Tags. **(A)** Clone size data of the first trial samples: Data were obtained considering the final optimal conditions: (Bowtie parameters: -v 3 - - best, and filtering condition: (merging approach) JT-10). **(B)** Clone size data of the second trial samples: Data were obtained considering the final optimal conditions: (Bowtie parameters: -v 3 - - best, and filtering condition: (merging approach) JT-10-1%). See Additional file 1: Figure S4 for information on merging approach.

clones was 50% vs. 50%. In this situation, shear sites had the smallest error: 1.72 for 1st T-cnt-1 and 0.47 for 2nd T-cnt-1.

The errors were reduced in the data using the tag system: 7.27, 5.23, 14.49, and 6.50 for the first trial, and 6.67, 7.07, 10.07, and 13.16 for the second trial. In the

case of the combination of tags and shear sites, errors were: 6.98, 4.06, 0.21, and 1.31 for the first trial and 3.42, 10.51, 12.26, and 5.83 for the second trial. Interestingly, the samples 'tags only' and 'combinations' showed similar error levels. Based on these data, our system showed lower absolute errors than when considering only shear

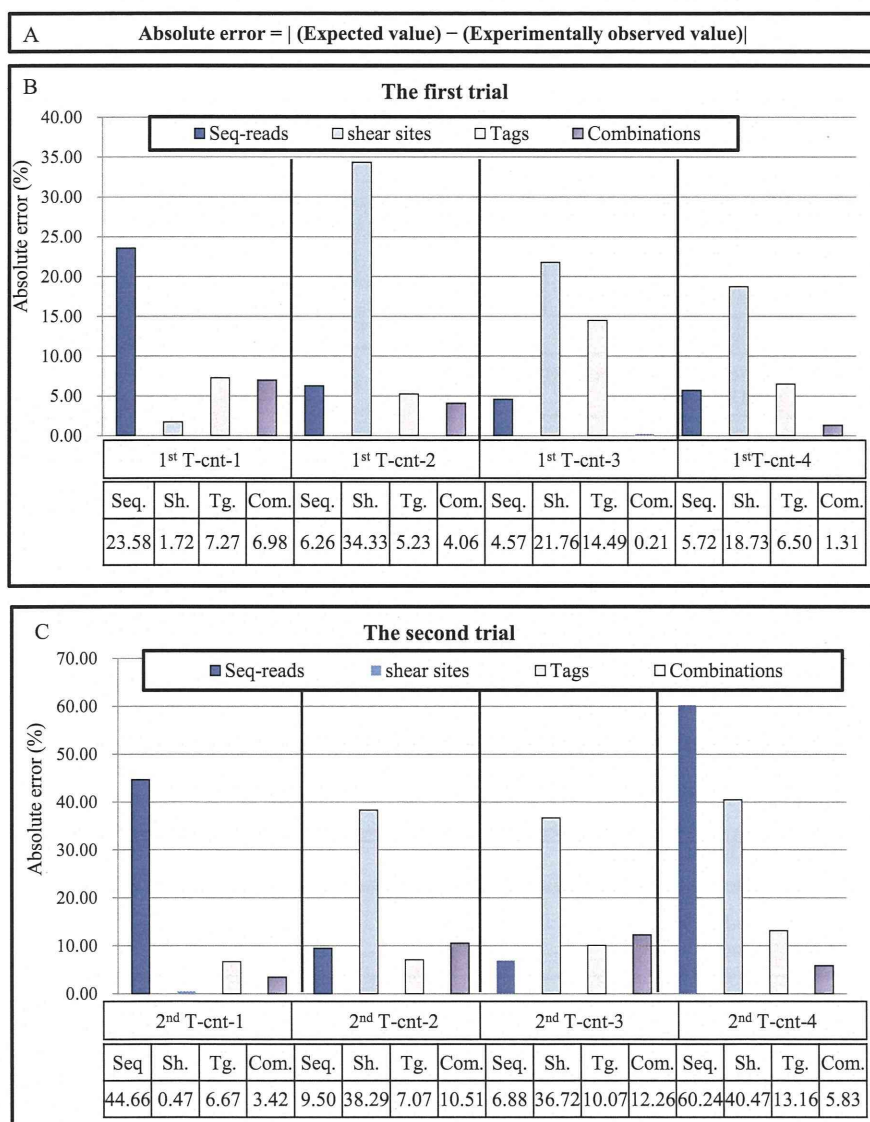


Figure 5 Evaluating the accuracy of the clonality analysis. (A) Absolute error is calculated by subtracting the expected values from the experimentally observed values. (B, C) The accuracy of the method is evaluated by calculating the absolute error of the clone size estimation of the control samples (see Figure 3). The y axis represents the percentage of absolute errors in different conditions including: (1) raw sequencing reads without removing duplicated PCR, (2) only shear sites, (3) only tags, and (4) the combination of tags and shear sites. The absolute errors of the final optimal condition: the first trial: (Bowtie parameters: -v 3 - - best, and filtering condition: (merging approach) JT-10), and the second trial: (Bowtie parameters: -v 3 - - best, and filtering condition: (merging approach) JT-10-1%) are presented in this figure. Please refer to Additional file 1: Figure S6 for the absolute errors in all examined conditions. (B) The absolute errors of the first trial. (C) The absolute errors of the second trial. See Additional file 1: Figure S4 for information on merging approach.

sites (Figure 5) (Additional file 1: Figure S4). Owing to differences in analyzed samples and system setups, we could not directly compare our data with published data [22,46]. Indirect evidence, however, provided by shear site analysis of our own data illustrated that our system has lower absolute errors than using the shear site-based methodology.

In-silico analysis

Processing, management, and analysis of the large amount of data generated by deep sequencing require special infrastructures and bioinformatics skills. We designed a data analysis and interpretation pipeline specific for HTLV-1 integration sites and clonality studies. The workflow is provided in Figure 6. First, the raw data for

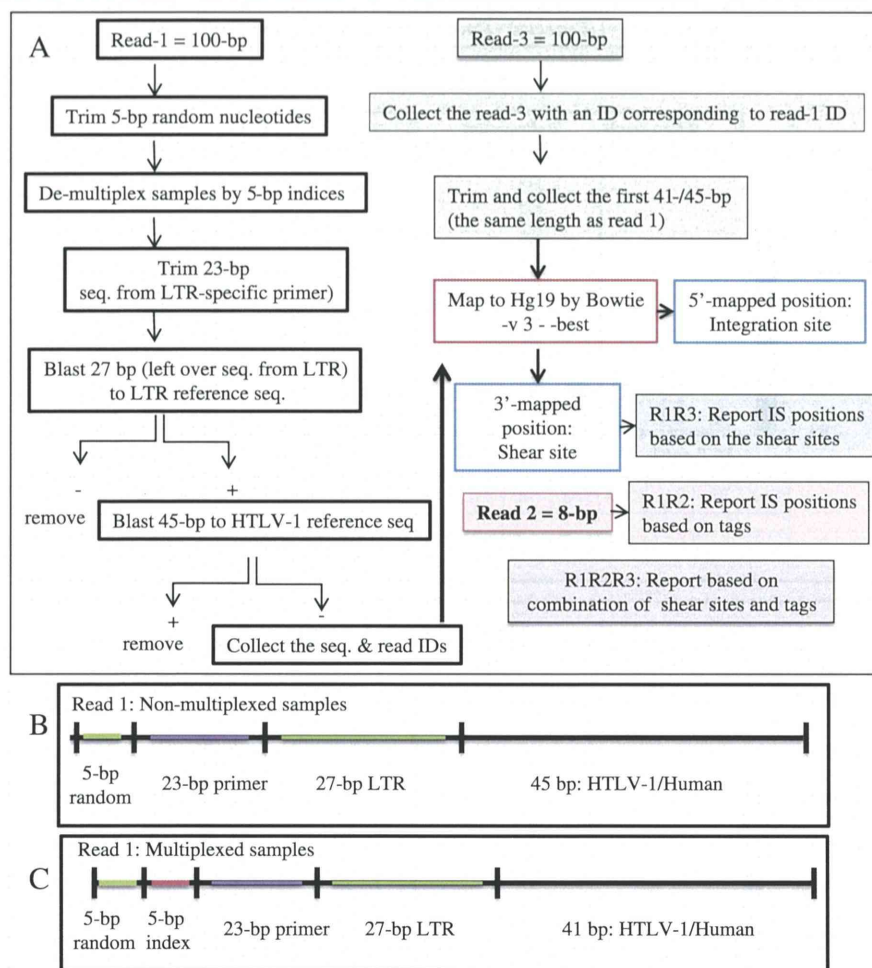


Figure 6 *In-silico* analysis work flow. **(A)** Illumina HiSeq 2000 platform outputs raw data of (Read-1 = 100 bp), (Read-3 = 100 bp), and (Read-2 = 8 bp). Data were analyzed according to this work flow after checking quality with the FastQC tool. In the case of Read-1, the first 5 bp were trimmed, and the next 5 bp were used to de-multiplex indexed samples. The downstream 23 bp, which correspond to the LTR primer (F2), were then removed. The next 27 bp were subjected to a blast search against the LTR reference sequence. For the blast search reads, the remaining 41/45 bp were subjected to a blast search against an HTLV-1 reference sequence. Reads were confirmed to be from HTLV-1 was removed, and the sequences and IDs from the remaining reads which considered as human, were collected. Subsequently, Read-3 with IDs corresponding to Read-1's IDs were collected. The first 41/45 bp of Read-3 were trimmed and collected to have the same length as Read-1. The paired sequences of Read-1 and Read-3 (same lengths) were mapped against hg19 by Bowtie with -v 3 - -best parameters. The 5'-mapped positions were considered to be integration sites and the 3'-mapped positions as shear sites. Read-2 information was used to retrieve the clone size based on tags. Finally, the clone size was computed by combining tag and shear site information. All the analyses were done by our own Perl scripts, which resulted in the following reports. Report R1R3: the distribution of unique shear sites per integration site. Report R1R2: the distribution of unique tags per integration site. Report R1R2R3: the distribution of unique tags and shear sites per integration site. **(B, C)** The structure of Read-1 for the non-multiplexed and multiplexed samples.

high-throughput sequencing were checked for quality by the FastQC tool. We then removed the first 5-bp random nucleotides from read-1 and de-multiplexed those samples that were run in the same lane of the HiSeq 2000 based on 5-bp of the known sequence (Figure 6 and Additional file 1: Figure S2). The downstream 23 nucleotides, which represented LTR-specific primers, were also trimmed before further analysis. We then separated the remaining sequence of read one into two different datasets: (1) LTR sequence and (2) HTLV-1 or

human sequence. The former comprises the 27-bp sequence remaining from the LTR, whereas the latter is composed of the 41-bp or 45-bp HTLV-1 or human sequence. In the case of multiplexed and non-multiplexed samples, different lengths (that is, 41-bp and 45-bp) were available for analysis. Both sets were subjected to blast analysis against LTR and HTLV-1 reference sequences with one or two mismatches permitted, respectively. Reads for which the sequence did not match HTLV-1 were presumed to be human as long as their

27-bp LTR sequences matched the LTR reference sequence. The resulting human reads were mapped to the human genome (hg19) using Bowtie 1.0.0 [58]. We employed various parameters of Bowtie and different lengths of read three to obtain the optimal mapping yield (Additional file 1: Table S2). These conditions were achieved when a maximum of three mismatches were permitted (-v parameter) and when the best alignment regarding the number of mismatches was reported (-best parameter). In addition, use of the same length of read-1 as in read-3 allowed for better mapping results. Mapping results are further discussed in Additional file 1: Notes.

The 5'-mapped regions were considered to be the positions of integration sites and reported as (chromosome: position: (strand)) for example, (chr1:121251270: (-)). In addition, 3'-mapped regions from read-3 were reported as shear sites for each corresponding position. Information on the tags, obtained from read-2, was used to determine the size of clones as described in subsection: Measuring the size of clones by the tag system. Final outputs of our analysis - the three main reports: R1R3, R1R2, and R1R2R3 - include information on shear sites, tags, and a combination of tags and shear sites, respectively (Figure 6).

Removing background noise

Data obtained from next-generation sequencers are not error free [40,62-65]. There are many reports on the error rate of Illumina sequencers [66,67]. Teemu Kivioja *et al.* recently developed a system named unique molecular identifiers (UMIs) for quantifying mRNAs and employed filtering criteria to remove false UMIs generated by sequencing errors [68]. In our study, consistent with the data of Kivioja *et al.* [68], the sequencing errors produced false tags with low frequencies. A filtering system was required to remove those tags, which could affect interpretation of our clonality data and reduce the accuracy of the clone size measurement. To minimize the effect of sequencing errors on data interpretation, we tested different filtering conditions to remove background noise. Here, we report our proven filtering approach (Additional file 1: Figure S4).

Considering that tags are designed randomly, each tag has an equal probability of being observed. Hence, the distribution of tags should be fitted to the zero truncated Poisson distribution [59,68]. Therefore, we test data fit to the Poisson distribution to determine the efficacy of each filtering condition. The distribution of tags for each sample was measured by the R-package 'gamlss.tr' [59], and the correlation coefficient was compared before and after filtering (Additional file 1: Figure S6).

We used a filtering system, which we named the merging approach. The merging approach was conducted by

clustering the tags and allowing only one mismatch so that unique tags, differing only in one nucleotide (one-mismatch permission), were merged. Subsequently, if the frequency of observed tag reads (PCR duplicates) was greater than 10, those unique tags were employed in further analysis. Otherwise, they were considered as artifacts. We referred to this filtering approach as 'Join Tag- remove10' (JT-10) in the Figure legends. To facilitate understanding, these filtering conditions are illustrated in Additional file 1: Figure S4.

Final discussion

The advent of NGS technologies holds promise to reveal the complex nature of neoplasms and to move past the limitations of previous methods. Using different approaches starting from early cytogenetic analysis to later, more elaborate studies with NGS technologies, the clonal composition of different tumors has been analyzed [36-39]. Robust monitoring and tracking of clonal dynamics using provirus integration sites allow for the assessment of clonal composition of HTLV-1-infected individuals from early infection to the final stage of ATL development. To meet the technical requirements for such type of analysis, we combined our expertise in the field of HTLV-1 research and NGS analysis and developed the high-throughput methodology described herein.

Gillet *et al.* also recently introduced a high-throughput method to extensively characterize HTLV-1 integration site preferences and quantify clonality (further discussed in Additional file 1: Notes) [22]. They statistically analyzed shear site data to estimate clone size. According to their published data [22,46] and as well as our current data, the limited variation in shear sites leads to an underestimation of the size of large clones. Considering that the incidence of large clones increases with disease progression from the healthy AC state to the malignant states of smoldering, chronic, or acute [22,46], an accurate measurement of clone size - particularly large clones - is of great clinical significance.

Our study is the first in which the size of large clones was experimentally measured without using statistical estimation. We have provided details of the method design, optimized experiment protocols, and *in-silico* data processing workflow. To validate our methodology and assess its accuracy, we analyzed eight control samples with known integration sites and clone sizes, and four clinical samples. We subjected the samples to deep sequencing so that they had enough read coverage for each integration site and to ensure accurate measurement of clone size (See Additional file 1: Notes). We proved our methodology to be reliable for isolating large numbers of integration sites and to be accurate for quantifying clone size. Because the tag system could

provide a sufficient number of variations regardless of clone size, we were able to demonstrate that the measurements are accurate.

Preliminary experiments on the clinical samples with differing PVLs and disease status showed different clonality patterns specific to AC and different ATL disease subtypes. S-1 was selected to represent still-healthy but infected individuals (ACs), S-2 and S-3 to represent indolent types of ATL, and S-4 to represent a typical aggressive type of ATL. Despite similar PVLs, S-1 and S-2 could be distinguished based on clonality patterns (polyclonal vs. a shift towards oligoclonal): S-1: AC, 8% PVL, and S-2: SM, 9% PVL. The clones of AC showed a uniform distribution pattern with no large difference in clone size; clones of S-2, however, had non-uniform size. S-2 and S-3 (S-3: SM, 31% PVL) are both smoldering subtypes of ATL progression with differing PVLs (9% vs. 31%) and showed similar clonality patterns but a different number of infected cells in each clone. S-3 and S-4 had similar PVL (S-4: acute, 33% PVL) but exhibited different clonality patterns: oligoclonal for S-3 (three or four relatively large clones at the top surrounded with other clones) vs. monoclonal for S-4 (a large major clone surrounded with some small clones in the background). After ranking the clones in order of descending size, we noted that the size of the largest clone in the acute sample was 10 times that of the next clone (tags: (chr X: 83705328 (-)) = 2675 vs. (chr 14: 30655896 (+)) = 209). Relative size of the major clone (chr X: 83705328 (-)) was also estimated by another method (PCR-southern) (detailed information is provided in Additional file 2: Figure S3 and Additional file 2: Supporting experiments). Samples with distinct disease status (AC, SM, and acute) manifested different clone sizes (Additional file 1: Table S3 and Additional file 2: Table S1 include the number of infected cells in the top 10 clones), but S-1 vs. S-2 (0.60 vs. 0.67) and S-3 vs. S-4 (0.84 vs. 0.80) could not be discriminated based on their oligoclonality index (Additional file 1: Figure S7) (See Additional file 1: Notes for further discussion). Therefore, it can be inferred that, with an accurate measurement of clone size, the application of this method will aid in the discrimination of ATL subtypes. These results suggest a possible association between disease status, PVLs, and clonality patterns. Hence, HTLV-1-infected individuals could be classified in different groups based on their clonality patterns, which could ultimately affect their choice of therapy and estimation of prognosis.

Moreover, by interpreting information from previous studies on HTLV-1 clonality [15,18-20,22,27,31,32,35] and considering the data provided in our present paper, it appears that ACs harbor a polyclonal population of HTLV-1-infected cells, whereas ATL patients show monoclonal patterns. Thus, changes in the clonality pattern and onset

of a clonal expansion of HTLV-1-infected cells seem to be potentially applicable as a prognostic indicator of ATL onset. For these purposes, it is necessary to analyze appropriate pools of samples from ACs and different subtypes of ATL and to conduct a cohort study on the clonality patterns of the sequential samples available over time.

Conclusions

We took advantage of next-generation sequencing technology, a tag system, and an *in-silico* analysis pipeline to develop and internally validate a new high-throughput methodology. The method was proved to accurately measure the size of clones by analyzing control samples with already known clone sizes and clinical samples. We also discussed the novelty, significance, and applications of our method, and compared it with the only existing high-throughput method devised by Gillet *et al.* [22]. Employing our new methodology and the analysis of an appropriate pool of samples provided by JSPFAD [13] will be helpful not only for diagnosis and prediction but also for elaborated understanding of the underlying mechanism of ATL development. The methodology described here could be adapted to investigate and quantify other genome-integrating elements (such as proviruses, transposons, and vectors in gene therapy). In addition, the tag system can be used for quantifying DNA/RNA fragments in RNA expression [68] or in metagenomics for determining the size of bacterial populations.

Additional files

Additional file 1: Supplementary data include (1) Supplementary Notes: 'Supplementary materials and method' and 'Supplementary results and discussion' (2) Supplementary figures and tables: seven figures, and three tables provided in a PDF file.

Additional file 2: Additional supporting data include (1) Additional supporting protocols and (2) Additional supporting experiments: four figures and one table provided in a PDF file.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TW, TY, YS, SS, and SF conceived the project. SF designed and carried out the experiments and wrote the manuscript. YL prepared the Perl scripts. YL and SF performed *in-silico* data analysis. SF and TY analyzed and interpret the data. YS, SS and SF contributed in sequencing the samples. YS and KN contributed to *in-silico* data analysis. TY, YL, TW and YS assisted in drafting the manuscript. TY and YS advised the direction of study. TW supervised the study. All authors read and approved the final manuscript.

Acknowledgements

We gratefully appreciate: JSPFAD for providing clinical samples; M. Nakashima and T. Akashi for maintenance of JSPFAD; Sung-Joon Park, Riu Yamashita, and Kuo-ching Liang for their invaluable advice on *in-silico* analysis; K. Abe, K. Imamura, T. Horiuchi, and M. Tosaka for sequencing technical support; Sara Firouzi and Unes Firouzi for comments on the design of figures. SF expresses deep respect and gratitude to the NITORI scholarship foundation for supporting her during undergraduate studies.

Computational analyses were provided by the Super Computer System, Human Genome Center, Institute of Medical Science, at The University of Tokyo.

Funding

This work was supported by the Japanese Society for the Promotion of Science (JSPS) - DC1 (24.6916 to SF); Third Term Comprehensive Control Research for Cancer, Ministry of Health, Labour and Welfare (H24-G-004 to TW); JSPS KAKENHI (23390250 to TW, 24591383 to TY); and MEXT KAKENHI (221S0001 to TW, 221S0002 to YS).

Author details

¹Department of Medical Genome Science, Graduate School of Frontier Sciences, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. ²Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 277-8561, Japan. ³Human Genome Center, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

Received: 11 March 2014 Accepted: 20 June 2014

Published: 27 June 2014

References

1. Takatsuki K: Discovery of adult T-cell leukemia. *Retrovirology* 2005, **2**:16.
2. Gallo RC: The discovery of the first human retrovirus: HTLV-1 and HTLV-2. *Retrovirology* 2005, **2**:17.
3. Okamoto T, Ohno Y, Tsugane S, Watanabe S, Shimoyama M, Tajima K, Miwa M, Shimotohno K: Multi-step carcinogenesis model for adult T-cell leukemia. *Jpn J Cancer Res* 1989, **80**:191-195.
4. Yamagishi M, Watanabe T: Molecular hallmarks of adult T cell leukemia. *Front Microbiol* 2012, **3**:334.
5. Matsuoka M, Jeang K-T: Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. *Nat Rev Cancer* 2007, **7**:270-280.
6. Yamaguchi K, Watanabe T: Human T lymphotropic virus type-I and adult T-cell leukemia in Japan. *Int J Hematol* 2002, **76**:240-245.
7. Bangham CRM: HTLV-1. In *eLS*. Edited by John W, Sons L. Chichester, UK: John Wiley & Sons, Ltd; 2011.
8. Tsukasaki K, Tobinai K: Biology and treatment of HTLV-1 associated T-cell lymphomas. *Best Pract Res Clin Haematol* 2013, **26**:3-14.
9. Nakano K, Watanabe T: HTLV-1 Rex: the courier of viral messages making use of the host vehicle. *Front Microbiol* 2012, **3**:330.
10. Iwanaga M, Watanabe T, Utsunomiya A, Okayama A, Uchimarui K, Koh KR, Ogata M, Kikuchi H, Sagara Y, Uozumi K, Mochizuki M, Tsukasaki K, Saburi Y, Yamamura M, Tanaka J, Moriuchi Y, Hino S, Kamihira S, Yamaguchi K: Human T-cell leukemia virus type I (HTLV-1) proviral load and disease progression in asymptomatic HTLV-1 carriers: a nationwide prospective study in Japan. *Blood* 2010, **116**:1211-1219.
11. Iwanaga M, Watanabe T, Yamaguchi K: Adult T-Cell Leukemia: A Review of Epidemiological Evidence. *Front Microbiol* 2012, **3**:322.
12. Taylor GP, Tosswill JH, Matutes E, Daenke S, Hall S, Bain BJ, Davis R, Thomas D, Rossor M, Bangham CR, Weber JN: Prospective study of HTLV-I infection in an initially asymptomatic cohort. *J Acquir Immune Defic Syndr* 1999, **22**:92-100.
13. Yamaguchi K, Uozumi K, Taguchi H, Kikuchi H, Okayama A, Kamihira S, Hino S, Nosaka K, Watanabe T: Nationwide Cohort Study of HTLV-1 Carriers in Japan: Joint Study on Predisposing Factors of ATL Development (JSPFAD). *AIDS Res Hum Retroviruses* 2007, **23**:581-600.
14. Taylor GP, Matsuoka M: Natural history of adult T-cell leukemia/lymphoma and approaches to therapy. *Oncogene* 2005, **24**:6047-6057.
15. Tsukasaki K, Tsushima H, Yamamura M, Hata T, Murata K, Maeda T, Atogami S, Sohda H, Momita S, Ideda S, Katamine S, Yamada Y, Kamihira S, Tomonaga M: Integration Patterns of HTLV-I Provirus in Relation to the Clinical Course of ATL: Frequent Clonal Change at Crisis From Indolent Disease. *Blood* 1997, **89**:948-956.
16. Yoshida M, Seiki M, Yamaguchi K, Takatsuki K: Monoclonal integration of human T-cell leukemia provirus in all primary tumors of adult T-cell leukemia suggests causative role of human T-cell leukemia virus in the disease. *Proc Natl Acad Sci U S A* 1984, **81**:2534-2537.
17. Cavrois M, Gessain A, Wain-Hobson S, Wattel E: Proliferation of HTLV-1 infected circulating cells in vivo in all asymptomatic carriers and patients with TSP/HAM. *Oncogene* 1996, **12**:2419-2423.
18. Etoh K, Tamiya S, Yamaguchi K, Okayama A, Tsubouchi H, Ideta T, Mueller N, Takatsuki K, Matsuoka M: Persistent clonal proliferation of human T-lymphotropic virus type I-infected cells in vivo. *Cancer Res* 1997, **57**:4862-4867.
19. Okayama A, Stuver S, Matsuoka M, Ishizaki J, Tanaka G-i, Kubuki Y, Mueller N, Hsieh C-c, Tachibana N, Tsubouchi H: Role of HTLV-1 proviral DNA load and clonality in the development of adult T-cell leukemia/lymphoma in asymptomatic carriers. *Int J Cancer* 2004, **110**:621-625.
20. G-i T, Okayama A, Watanabe T, Aizawa S, Stuver S, Mueller N, Hsieh C-C, Tsubouchi H: The clonal expansion of human T lymphotropic virus type 1-infected T cells: a comparison between seroconverters and long-term carriers. *J Infect Dis* 2005, **191**:1140-1147.
21. Derse D, Crise B, Li Y, Princlair G, Lum N, Stewart C, McGrath CF, Hughes SH, Munroe DJ, Wu X: Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. *J Virol* 2007, **81**:6731-6741.
22. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, Berry C, Bangham CRM, Taylor GP, Bangham CRM: The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* 2011, **117**:3113-3122.
23. Cook LB, Rowan AG, Melamed A, Taylor GP, Bangham CRM: HTLV-1-infected T cells contain a single integrated provirus in natural infection. *Blood* 2012, **120**:3488-3490.
24. Melamed A, Laydon DJ, Gillet NA, Tanaka Y, Taylor GP, Bangham CRM: Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection. *PLoS Pathog* 2013, **9**:e1003271.
25. Wattel E, Vartanian JP, Pannetier C, Wain-Hobson S: Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. *J Virol* 1995, **69**:2863-2868.
26. Gabet AS, Mortreux F, Talarmin A, Plumelle Y, Leclercq I, Leroy A, Gessain A, Clity E, Joubert M, Wattel E: High circulating proviral load with oligoclonal expansion of HTLV-1 bearing T cells in HTLV-1 carriers with strongyloidiasis. *Oncogene* 2000, **19**:4954-4960.
27. Kamihira S, Iwanaga M, Doi Y, Sasaki D, Mori S, Tsurda K, Nagai K, Uno N, Hasegawa H, Yanagihara K, Morinaga Y, Tsukasaki K, Taniguchi H: Heterogeneity in clonal nature in the smoldering subtype of adult T-cell leukemia: continuity from carrier status to smoldering ATL. *Int J Hematol* 2012, **95**:399-408.
28. Satoh M, Toma H, Sugahara K, Etoh K, Shiroma Y, Kiyuna S, Takara M, Matsuoka M, Yamaguchi K, Nakada K, Fujita K, Kojima S, Hori E, Tanaka Y, Kamihira S, Sato Y, Watanabe T: Involvement of IL-2/IL-2R system activation by parasite antigen in polyclonal expansion of CD4(+)25(+) HTLV-1-infected T-cells in human carriers of both HTLV-1 and S-stercoalis. *Oncogene* 2002, **21**:2466-2475.
29. Ikeda S, Momita S, Kinoshita K-i, Kamihira S, Moriuchi Y, Tsukasaki K, Ito M, Kanda T, Moriuchi R, Nakamura T: Clinical course of human T-lymphotropic virus type I carriers with molecularly detectable monoclonal proliferation of T lymphocytes: defining a low-and high-risk population. *Blood* 1993, **82**:2017-2024.
30. Yamaguchi K, Seiki M, Yoshida M, Nishimura H, Kawano F, Takatsuki K: The detection of human T cell leukemia virus proviral DNA and its application for classification and diagnosis of T cell malignancy. *Blood* 1984, **63**:1235-1240.
31. Takemoto S, Matsuoka M, Yamaguchi K, Takatsuki K: A novel diagnostic method of adult T-cell leukemia: monoclonal integration of human T-cell lymphotropic virus type I provirus DNA detected by inverse polymerase chain reaction. *Blood* 1994, **84**:3080-3085.
32. Imaizumi Y, Iwanaga M, Tsukasaki K, Hata T, Tomonaga M, Ikeda S: Natural course of HTLV-1 carriers with monoclonal proliferation of T lymphocytes ("pre-ATL") in a 20-year follow-up study. *Blood* 2005, **105**:903-904.
33. Wongstaal F, Hahn B, Manzari V, Colombini S, Franchini G, Gelmann EP, Gallo RC: A survey of human leukemias for sequences of a human retrovirus. *Nature* 1983, **302**:626-628.
34. Cavrois M, Wain-Hobson S, Wattel E: Stochastic events in the amplification of HTLV-I integration sites by linker-mediated PCR. *Res Virol* 1995, **146**:179-184.
35. Ohshima K, Mukai Y, Shiraki H, Suzumiya J, Tashiro K, Kikuchi M: Clonal integration and expression of human T-cell lymphotropic virus type I in

- carriers detected by polymerase chain reaction and inverse PCR. *Am J Hematol* 1997, **54**:306–312.
36. Nowell PC: The clonal evolution of tumor cell populations. *Science* 1976, **194**:23–28.
37. Greaves M, Maley CC: Clonal evolution in cancer. *Nature* 2012, **481**:306–313.
38. Sprouffske K, Merlo Lauren MF, Gerrish Philip J, Maley Carlo C, Sniegowski Paul D: Cancer in Light of Experimental Evolution. *Curr Biol* 2012, **22**:R762–R771.
39. Melo FDSE, Vermeulen L, Fessler E, Medema JP: Cancer heterogeneity—a multifaceted view. *EMBO Rep* 2013, **14**:686–695.
40. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 2011, **12**:R18.
41. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF: PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* 2005, **71**:8966–8969.
42. Polz MF, Cavanaugh CM: Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* 1998, **64**:3724–3730.
43. Suzuki MT, Giovannoni SJ: Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* 1996, **62**:625–630.
44. Dabney J, Meyer M: Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 2012, **52**:87.
45. Raabe CA, Tang TH, Brosius J, Rozhdestvensky TS: Biases in small RNA deep sequencing data. *Nucleic Acids Res* 2014, **42**:1414–1426.
46. Berry CC, Gillet NA, Melamed A, Gormley N, Bangham CRM, Bushman FD: Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* 2012, **28**:755–762.
47. Gini C: Sulla misura della concentrazione e della variabilità dei caratteri. *Transactions of the Real Istituto Veneto di Scienze* 1914, **LIII**:1203.
48. Gini C: Concentration and dependency ratios (in Italian). *English Translation in Rivista di Politica Economica* 1997, **87**:769–789.
49. Biomaterial resource bank of HTLV-1 carriers, Joint Study on Predisposing Factors of ATL Development (JSPFAD). [<http://htlv1.org/old/bank-en.html>]
50. Shimoyama M: Diagnostic criteria and classification of clinical subtypes of adult T-cell leukaemia-lymphoma. A report from the Lymphoma Study Group (1984–87). *Br J Haematol* 1991, **79**:428–437.
51. Sugamura K, Fujii M, Kannagi M, Sakitani M, Takeuchi M, Hinuma Y: Cell surface phenotypes and expression of viral antigens of various human cell lines carrying human T-cell leukemia virus. *Int J Cancer* 1984, **34**:221–228.
52. Devon RS, Porteous DJ, Brookes AJ: Splinkerettes—improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res* 1995, **23**:1644–1645.
53. Uren AG, Mikkers H, Kool J, van der Weyden L, Lund AH, Wilson CH, Rance R, Jonkers J, van Lohuizen M, Berns A, Adams DJ: A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat Protoc* 2009, **4**:789–798.
54. Seiki M, Hattori S, Hirayama Y, Yoshida M: Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc Natl Acad Sci* 1983, **80**:3618–3622.
55. Human Genome Center (HGC), the institute of medical Science, the University of Tokyo. [<http://www.hgc.jp/>]
56. FastQC: A quality control tool for high throughput sequence data. [<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>]
57. UCSC Genome Browser. [<http://genome.ucsc.edu/>]
58. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009, **10**:R25.
59. Stasinopoulos DM, Rigby RA: Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw* 2007, **23**:1–46.
60. StatsDirect Medical Statistics Software. [<http://www.statsdirect.com/>]
61. Chatburn RL: Evaluation of instrument error and method agreement. *AANA J* 1996, **64**:261–268.
62. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 2008, **36**:e105.
63. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER: Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods* 2008, **5**:183–188.
64. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: A large genome centre's improvements to the Illumina sequencing system. *Nat Methods* 2008, **5**:1005–1010.
65. Li H, Ruan J, Durbin R: Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008, **18**:1851–1858.
66. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S: Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 2011, **39**:e90–e90.
67. Minoche AE, Dohm JC, Himmelbauer H: Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* 2011, **12**:R112.
68. 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**:72–74.

doi:10.1186/gm568

Cite this article as: Firouzi et al.: Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites. *Genome Medicine* 2014 **6**:46.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



CADM1 Expression and Stepwise Downregulation of CD7 Are Closely Associated with Clonal Expansion of HTLV-I-Infected Cells in Adult T-cell Leukemia/Lymphoma

Seiichiro Kobayashi¹, Kazumi Nakano⁵, Eri Watanabe², Tomohiro Ishigaki², Nobuhiro Ohno³, Koichiro Yuji³, Naoki Oyaizu⁴, Satomi Asanuma⁵, Makoto Yamagishi⁵, Tadanori Yamochi⁵, Nobukazu Watanabe², Arinobu Tojo^{1,3}, Toshiki Watanabe⁵, and Kaoru Uchimaru³

Abstract

Purpose: Cell adhesion molecule 1 (CADM1), initially identified as a tumor suppressor gene, has recently been reported to be ectopically expressed in primary adult T-cell leukemia-lymphoma (ATL) cells. We incorporated CADM1 into flow-cytometric analysis to reveal oncogenic mechanisms in human T-cell lymphotropic virus type I (HTLV-I) infection by purifying cells from the intermediate stages of ATL development.

Experimental Design: We isolated CADM1- and CD7-expressing peripheral blood mononuclear cells of asymptomatic carriers and ATLs using multicolor flow cytometry. Fluorescence-activated cell sorted (FACS) subpopulations were subjected to clonal expansion and gene expression analysis.

Results: HTLV-I-infected cells were efficiently enriched in CADM1⁺ subpopulations (D, CADM1^{pos} CD7^{dim} and N, CADM1^{pos} CD7^{neg}). Clonally expanding cells were detected exclusively in these subpopulations in asymptomatic carriers with high proviral load, suggesting that the appearance of D and N could be a surrogate marker of progression from asymptomatic carrier to early ATL. Further disease progression was accompanied by an increase in N with a reciprocal decrease in D, indicating clonal evolution from D to N. The gene expression profiles of D and N in asymptomatic carriers showed similarities to those of indolent ATLs, suggesting that these subpopulations represent premalignant cells. This is further supported by the molecular hallmarks of ATL, that is, drastic downregulation of miR-31 and upregulation of abnormal *Helios* transcripts.

Conclusion: The CADM1 versus CD7 plot accurately reflects disease progression in HTLV-I infection, and CADM1⁺ cells with downregulated CD7 in asymptomatic carriers have common properties with those in indolent ATLs. *Clin Cancer Res*; 20(11); 2851–61. ©2014 AACR.

Introduction

Human T-cell lymphotropic virus type I (HTLV-I) is a human retrovirus that causes HTLV-I-associated diseases, such as adult T-cell leukemia-lymphoma (ATL), HTLV-I-associated myelopathy/tropical spastic paraparesis, and HTLV-I uveitis (1–3). In Japan, the estimated lifetime risk of developing ATL in HTLV-I carriers is 6% to 7% for males

and 2% to 3% for females (4–6). It takes several decades for HTLV-I-infected cells to reach the final stage of multistep oncogenesis, which is clinically recognized as aggressive ATL (acute-type and lymphoma-type; ref. 7). Molecular interaction of viral genes [e.g., Tax and the HTLV-I basic leucine zipper (HBZ) gene] with the cellular machinery causes various genetic and epigenetic alterations (7–11). However, difficulties in purifying HTLV-I-infected cells *in vivo* seem to have hindered understanding of the genetic events that are directly involved in the multistep oncogenesis of ATL.

Upregulation or aberrant expression of cell surface markers, such as CCR4 and CD25, is useful for diagnosis of ATL and has been utilized for molecular-targeted therapy (12, 13). However, the expression levels of these markers vary among patients, which often make it difficult to identify ATL cells specifically based on the immunophenotype. Previously, we focused on downregulated markers in acute-type ATL cells, such as CD3 and CD7, and successfully purified ATL cells using the CD3 versus CD7 plot of CD4⁺ cells (14). Analysis of other clinical subtypes

Authors' affiliations: ¹Division of Molecular Therapy; ²Laboratory of Diagnostic Medicine, Division of Stem Cell Therapy; ³Department of Hematology/Oncology, Research Hospital; ⁴Clinical Laboratory, Research Hospital, Institute of Medical Science; and ⁵Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

Note: Supplementary data for this article are available at Clinical Cancer Research Online (<http://clincancerres.aacrjournals.org/>).

Corresponding Author: Kaoru Uchimaru, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan. Phone: 81-3-5449-5542; Fax: 81-3-5449-5429; E-mail: uchimaru@ims.u-tokyo.ac.jp

doi: 10.1158/1078-0432.CCR-13-3169

©2014 American Association for Cancer Research.