Published OnlineFirst April 11, 2014; DOI: 10.1158/1078-0432.CCR-13-3169

disease progression in asymptomatic HTLV-1 carriers: a nationwide prospective study in Japan. Blood 2010;116:1211–9.

31. Yamagishi M, Nakano K, Miyake A, Yamochi T, Kagami Y, Tsutsumi A, et al. Polycomb-mediated loss of miR-31 activates NIK-dependent NF-kappaB pathway in adult T cell leukemia and other cancers. Cancer Cell 2012;21:121–35.

32. Asanuma S, Yamagishi M, Kawanami K, Nakano K, Sato-Otsubo A, Muto S, et al. Adult T-cell leukemia cells are characterized by abnormalities of Helios expression that promote T-cell growth. Cancer Sci 2013;104:1097–106.

33. Yamaguchi K, Kiyokawa T, Nakada K, Yul LS, Asou N, Ishii T, et al. Polyclonal integration of HTLV-I proviral DNA in lymphocytes from HTLV-I seropositive individuals: an intermediate state between the healthy carrier state and smouldering ATL. Br J Haematol 1988;68: 169–74.

34. Kamihira S, Iwanaga M, Doi Y, Sasaki D, Mori S, Tsurda K, et al. 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.

35. Masuda M, Maruyama T, Ohta T, Ito A, Hayashi T, Tsukasaki K, et al. CADM1 interacts with Tiam1 and promotes invasive phenotype of human T-cell leukemia virus type I-transformed cells and adult T-cell leukemia cells. J Biol Chem 2010;285: 15511–22.

Published OnlineFirst April 11, 2014; DOI: 10.1158/1078-0432.CCR-13-3169

AACR American Association for Cancer Research

Clinical Cancer Research

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 Kobayashi, Kazumi Nakano, Eri Watanabe, et al.

Clin Cancer Res 2014;20:2851-2861. Published OnlineFirst April 11, 2014.

Familial Clusters of HTLV-1-Associated Myelopathy/ Tropical Spastic Paraparesis

Satoshi Nozuma[1], Eiji Matsuura[1]*, Toshio Matsuzaki[2], Osamu Watanabe[1], Ryuji Kubota[2], Shuji Izumo[2], Hiroshi Takashima[1]

[1] Department of Neurology and Geriatrics, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima city, Japan, [2] Department of Molecular Pathology, Center for Chronic Viral Diseases, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima city, Japan

Abstract

Objective: HTLV-1 proviral loads (PVLs) and some genetic factors are reported to be associated with the development of HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP). However, there are very few reports on HAM/TSP having family history. We aimed to define the clinical features and laboratory indications associated with HAM/TSP having family history.

Methods: Records of 784 HAM/TSP patients who were hospitalized in Kagoshima University Hospital and related hospitals from 1987 to 2012 were reviewed. Using an unmatched case-control design, 40 patients of HAM/TSP having family history (f-HAM/TSP) were compared with 124 patients suffering from sporadic HAM/TSP, who were admitted in series over the last 10 years for associated clinical features.

Results: Of the 784 patients, 40 (5.1%) were f-HAM/TSP cases. Compared with sporadic cases, the age of onset was earlier (41.3 vs. 51.6 years, $p<0.001$), motor disability grades were lower (4.0 vs. 4.9, $p=0.043$) despite longer duration of illness (14.3 vs. 10.2 years, $p=0.026$), time elapsed between onset and wheelchair use in daily life was longer (18.3 vs. 10.0 years, $p=0.025$), cases with rapid disease progression were fewer (10.0% vs. 28.2%, $p=0.019$), and protein levels in cerebrospinal fluid (CSF) were significantly lower in f-HAM/TSP cases (29.9 vs. 42.5 mg, $p<0.001$). There was no difference in HTLV-1 PVLs, anti-HTLV-1 antibody titers in serum and CSF, or cell number and neopterin levels in CSF. Furthermore, HTLV-1 PVLs were lower in cases with rapid disease progression than in those with slow progression in both f-HAM/TSP and sporadic cases.

Conclusions: We demonstrated that HAM/TSP aggregates in the family, with a younger age of onset and a slow rate of progression in f-HAM/TSP cases compared with sporadic cases. These data also suggested that factors other than HTLV-1 PVLs contribute to the disease course of HAM/TSP.

Citation: Nozuma S, Matsuura E, Matsuzaki T, Watanabe O, Kubota R, et al. (2014) Familial Clusters of HTLV-1-Associated Myelopathy/Tropical Spastic Paraparesis. PLoS ONE 9(5): e86144. doi:10.1371/journal.pone.0086144

Editor: Steven Jacobson, National Institutes of Health, United States of America

Received June 6, 2013; **Accepted** December 5, 2013; **Published** May 6, 2014

Funding: This study was supported by Health and Labour Sciences Research Grants from the Ministry of Health Labour and Welfare and JSPS KAKENHI Grant Numbers 25293205 and 24133701. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pine@m.kufm.kagoshima-u.ac.jp

Introduction

HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) is characterized by slow progressive spastic paraparesis and positivity for anti-HTLV-1 antibodies in both serum and cerebrospinal fluid (CSF) [1,2]. Worldwide, at least 10–20 million people are infected with HTLV-1 [3]. However, although the majority of infected individuals remain lifelong asymptomatic carriers, approximately 2%–5% develop adult T-cell lymphomas [4,5] and another 0.25%–3.8% develop HAM/TSP [1,2]. Although the mechanisms underlying the development of HAM/TSP are not fully understood, several risk factors are closely associated with HAM/TSP. In particular, HTLV-1 proviral loads (PVLs) are significantly higher in HAM/TSP patients than in asymptomatic carriers and are also higher in genetic relatives of HAM/TSP patients than in non-HAM-related asymptomatic carriers [6]. Host genetic factors, including human leukocyte antigen (HLA) and non-HLA gene polymorphisms affect the occurrence of HAM/TSP [7], indicating that HTLV-1 PVLs and genetic backgrounds may influence individual susceptibility to HAM/TSP. Although several reports of familial adult T-cell lymphoma have been published [8,9], to our knowledge, there is only one case report of patient with HAM/TSP having family history (f-HAM/TSP) [10]. Hence, little is known about the prevalence and character of f-HAM/TSP cases. In this study, the characteristic clinical and laboratory features of f-HAM/TSP cases are defined and compared with those of sporadic cases.

Methods

Ethics Statement

This study was approved by the Institutional Review Boards of Kagoshima University. All participants provided written informed consent.

Design

We used an unmatched case-control design to identify the phenotypic features of f-HAM/TSP. f-HAM/TSP cases were identified as patients with multiple family members suffering from HAM/TSP. Controls were defined as HAM/TSP patients who were not genetically related to other HAM/TSP patients.

Subjects

f-HAM/TSP cases were extracted from our database of individuals diagnosed with HAM/TSP in Kagoshima University Hospital and related hospitals from 1987 to 2012. Controls included consecutive patients with sporadic HAM/TSP who were evaluated in our department between January 2002 and June 2012. HAM/TSP was diagnosed according to the World Health Organization diagnostic criteria, and the updated criteria of Castro-costa Belem [11]. Clinical information was obtained from the medical records of patient attendance at our hospital. In other cases, clinical data were obtained from the clinical records of patients or directly from the referring clinicians. Clinical variables included sex, age, age of onset, and initial symptoms. Neurological disabilities were assessed using Motor Disability Grading (MDG), modified from the Osame Motor Disability Scale of 0 to 10, as reported previously [12]. Motor disability grades were defined as follows: 5, needs one-hand support while walking; 6, needs two-hand support while walking; and 7, unable to walk but can crawl. We used a different assessment for the subgroup of more than grade 6 because their disease state significantly interfered with their lifestyle and necessitated the use of wheelchairs in daily life. The subgroup of patients with rapid progression was defined by deterioration of motor disability by more than three grades within two years. Anti-HTLV-1 antibody titers in serum and CSF were detected using enzyme-linked immunosorbent assays and particle agglutination methods (Fijirebio Inc, Tokyo, Japan). HTLV-1 PVLs in peripheral blood mononuclear cells (PBMCs) were assayed using quantitative PCR with the ABI PRISM 7700TM sequence detection system as reported previously [6].

Statistical Analysis

Data were analyzed using SPSS-20 (SPSS, Chicago, Illinois). Statistical analyses were performed using parametric (t-test) and non-parametric tests (Mann–Whitney test) for continuous variables and χ^2 (Pearsonχ^2 test/Fisher exact test) for categorical variables. Significant differences were then adjusted for potential confounders (age and sex) using multiple linear regression analysis. Survival was estimated according to the Kaplan–Meier method. The final endpoint was defined by a MDG score of 6. Patients with MDG scores of 6 almost wheelchair bound in daily life. The log rank test was used in Kaplan–Meier analyses. Differences were considered significant when $p<0.05$.

Results

Clinical characteristics of f-HAM/TSP

Of the 784 patients diagnosed with HAM/TSP between January 1987 and June 2012, 40 (5.1%) were f-HAM/TSP. The sex ratio was 33 males : 7 females. Of these 40 cases, 10 had parents or children (25.0%), 27 had siblings (67.5%), and three had other relatives (7.5%) diagnosed with HAM/TSP. Three individuals from one family were diagnosed with HAM/TSP, whereas only two individuals were diagnosed with HAM/TSP in all other families. In f-HAM/TSP cases, the age of onset was earlier (41.3 vs. 51.6 years, $p<0.001$), cases with rapid progression were fewer (10.0% vs. 28.2%, $p = 0.019$), motor disability grades were lower (4.0 vs. 4.9, $p = 0.043$) despite longer duration of illness (14.3 vs. 10.2 years, $p = 0.026$), and time elapsed between onset and wheelchair use in daily life was longer (18.3 vs. 10.0 years, $p = 0.025$) compared with sporadic cases. Sex and initial symptoms did not differ significantly between f-HAM/TSP and sporadic cases (Table 1). Twelve patients of f-HAM/TSP, and 38 of the 128 sporadic cases reached endpoint MDG scores of 6. Significant differences were then adjusted for potential confounders (age and sex) using multivariate analysis. Age of onset, duration of illness, MDG scores, and time elapsed between onset and wheelchair use in daily life remained significantly different after multivariate analysis (Table 1). The proportion of patients with rapid progression did not differ significantly between the groups, although there was a trend toward a higher proportion in sporadic cases. Kaplan–Meier analyses revealed that approximately 30% of both f-HAM/TSP and sporadic cases needed a wheelchair in daily life in 15 years after onset, and approximately 50% of patients from both groups needed it in 20 years after onset (Figure 1). Although sporadic patients needed wheelchairs earlier in most cases, the difference in the ratio of the patients with MDG score above six was not statistically significant between the groups. Finally, we compared differences in the age of onset between parent–child and sibling cases in f-HAM/TSP cases. Age of onset in parent–child f-HAM/TSP cases was significantly younger than that in sibling f-HAM/TSP cases (29.9±10.0 vs. 45.1±13.0 years, $p = 0.002$).

Laboratory parameters and PVLs in f-HAM/TSP cases

Protein levels in CSF were significantly lower in f-HAM/TSP cases than in sporadic cases (29.9 vs. 42.5 mg/dl, $p<0.001$). This difference in CSF protein level remained significant after multivariate analysis. Anti-HTLV-1 antibody titers in serum and CSF, and cell numbers and neopterin levels in CSF were not significantly different between two groups. Moreover, HTLV-1 PVLs did not differ significantly. (Table 2).

Clinical and laboratory findings in patients with rapid disease progression

Previous studies suggest that an older age of onset is associated with rapid disease progression. Similar findings are found in the present study. The percentage of rapid progression tended to increase with older age of onset in both f-HAM/TSP and sporadic groups (Figure 2). We compared the characteristics of 124 sporadic HAM/TSP patients with rapid and slow progression who were admitted to Kagoshima University Hospital in series during the last 10 years (Table 3). Patients with rapid progression were significantly older at onset than those with slow progression (62.3 vs. 47.4 years, $p<0.001$), although sex and initial symptoms did not differ significantly between rapid and slow progression groups. However, the time elapsed between onset and wheelchair use in daily life was markedly shorter among patients with rapid progression (1.5 vs. 14.4 years, $p<0.001$). Cell numbers, protein levels, and anti-HTLV-1 antibody titers in CSF were significantly higher in patients with rapid progression than in those with slow progression (11.6 vs. 3.2, $p<0.001$; 55.3 vs. 36.7 mg/dl, $p<0.001$; 1,251 vs. 416, $p<0.014$, respectively). Interestingly, HTLV-1 PVLs were significantly lower in patients with rapid progression than in those with slow progression (370 vs. 1,245 copies, $p<0.001$). Furthermore, we compared the differences between women and men in patients with rapid progression because the reason remains unknown why HAM/TSP is common in female

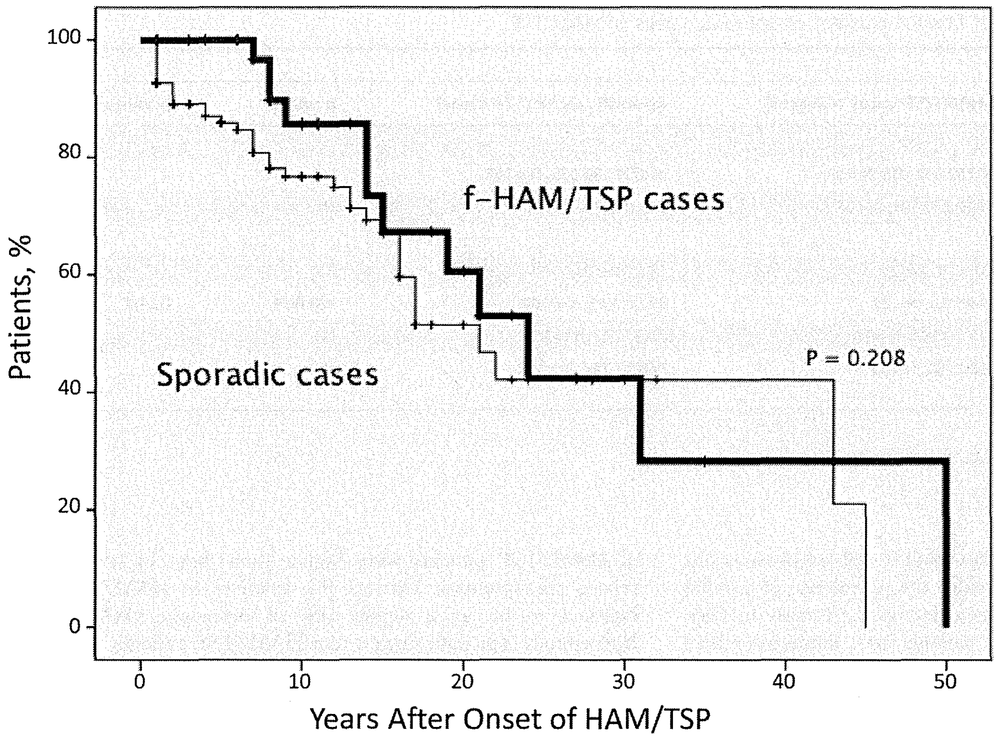


Figure 1. Kaplan-Meier estimates of the time from disease onset to assignment of motor disability scores of 6. In sporadic cases, more patients reached the score of six at an early stage; however, the difference was not significant. Approximately 30% of both f-HAM/TSP cases and sporadic cases needed a wheelchair in daily life in 15 years after onset and approximately 50% of patients from both groups needed a wheelchair in 20 years after onset.
doi:10.1371/journal.pone.0086144.g001

than in male. There was no significant difference between women and men in the age of onset (61.5 y.o.±12.6 vs. 62.7 y.o.±12.5), in the incidence of rapid progression (26.3% vs. 32.3%) and in MDG score (5.4 vs. 5.0; mean).

Discussion

We demonstrated that among 784 HAM/TSP patients, 40 (5.1%) had family members with the disease. The lifetime risk of developing HAM/TSP is 0.25% of HTLV-1 carriers in Japan

Table 1. Clinical features of f-HAM/TSP cases or sporadic cases of HAM/TSP.

	f-HAM/TSP cases (40 cases)	Sporadic cases (124 cases)	p value	p value†
Female ratio (%)	78.8% (7 males : 33 females)	66.4% (31 males : 93 females)	NS	
Age	55.6±13.0 (23–79)	61.8±12.5 (15–83)	**0.008**	
Age of onset	41.3±13.9 (14–65)	51.6±15.9 (13–78)	**<0.001**	**0.017**
Duration of illness (years)	14.3±11.4 (1–49)	10.2±9.6 (0–45)	**0.026**	**0.017**
Initial symptoms				
Gait disturbance	50.0%	52.4%	NS	
Urinary disturbance	32.5%	26.6%	NS	
Sensory disturbance	12.5%	14.5%	NS	
Others	5%	6.5%	NS	
Rapid disease progression	4 cases (10.0%)	35 cases (28.2%)	**0.019**	0.069
Motor disability score	4.0±2.0 (0–7)	4.9±1.5 (0–8)	**0.043**	**0.036**
Score more than 6	12 cases (30.0%)	38 cases (30.7%)	NS	
Time elapsed between onset and wheelchair use in daily life (years)	18.3±12.4 (7–50)	10.0±10.4 (1–45)	**0.025**	**0.020**

Data are presented as mean values ± s.d., (range),
†Adjusted for age and sex.
doi:10.1371/journal.pone.0086144.t001

Table 2. Laboratory findings of familial clusters or sporadic cases of HAM/TSP.

	f-HAM/TSP cases (40cases)	Sporadic cases (124 cases)	p value	p value†
Anti-HTLV-1 antibodies*				
Titer in Serum	20,787±31,004, N = 37	31,009±36,075, N = 109	NS	
Titer in CSF	2,310±11,741, N = 31	672±1,274, N = 111	NS	
Cerebrospinal fluid				
Cell number (/mm^3)	3.0±2.5, N = 25	5.7±10.0, N = 109	NS	
Protein (mg/dl)	29.9±9.4, N = 22	42.5±19.3, N = 109	**<0.001**	**0.007**
Neopterin (pmol/ml)	83.2±118.1, N = 18	38.3±56.8, N = 35	NS	
HTLV-1 proviral loads (Copies/10^4 PBMCs)	930±781, N = 32	968±1,746, N = 101	NS	

* Particle Aggregation Method.
Data are presented as mean values ± s.d., N = sample number,
†Adjusted for age and sex.
doi:10.1371/journal.pone.0086144.t002

[13]. Although clustering of familial adult T-cell lymphomas has been reported [8,9], to our knowledge the prevalence of familial clusters of HAM/TSP has not been described. A study in Peru showed that 30% of HAM/TSP patients have family members with paralytic neurological disorders, but the cause of paralysis was not evaluated [14]. In the present study, we included f-HAM/TSP diagnosed in medical institutions and excluded cases with a family history of neurological disorders. Thus, the actual incidence rates of f-HAM/TSP may be higher than those reported here. Interestingly, although HTLV-1 PVL has been associated with the development and clinical progression of HAM/TSP [15–17], there was no significant difference between f-HAM/TSP and sporadic cases in the present study. Because previous studies reported that HTLV-1 PVLs of asymptomatic carriers in relatives of HAM/TSP patients were higher than those in non-HAM-related asymptomatic carriers [6], relatives of HAM/TSP are believed to be at a higher risk of developing HAM/TSP. Interestingly, our data suggest that HAM/TSP patients aggregate in families and factors other than HTLV-1 PVLs may contribute to HAM/TSP.

Compared with sporadic HAM/TSP, the clinical characteristics of f-HAM/TSP have a younger age of onset and longer time elapsed between onset and wheelchair use in daily life. Although we were unable to identify the reason for earlier onset among f-HAM/TSP cases, one can speculate that mild symptoms, such as urinary and sensory disturbances, may be identified earlier by family members who are familiar with HAM/TSP symptoms. However, the present data show no difference in initial symptoms

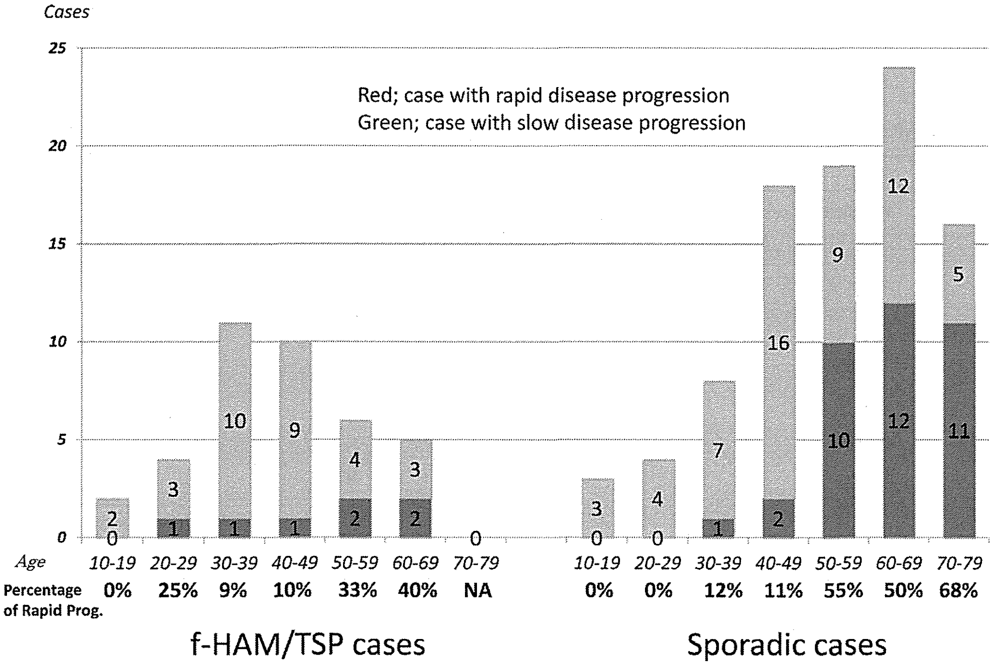


Figure 2. Age-specific proportions of rapid disease progression. The proportion of cases with rapid disease progression tended to increase with the older age of onset.
doi:10.1371/journal.pone.0086144.g002

Table 3. Clinical and laboratory findings of sporadic HAM/TSP with rapid/slow disease progression.

Type of disease progression	Rapid progression	Slow progression	p value
Female ratio (%)	71.4% (10 males : 25 females)	76.4% (21 males : 68 females)	NS
Age of onset	62.3±9.6, N = 35	47.4±15.9, N = 89	**<0.001**
Age of onset of f-HAM/TSP cases	60.5±3.7, N = 4	39.2±12.9, N = 36	**0.002**
Duration between onset and inability to walk alone (years)	1.5±0.9, N = 13	14.4±10.4, N = 25	**<0.001**
Anti-HTLV-1 antibodies*			
Titer in Serum	31,894±36,845, N = 34	30,608±35,965, N = 75	NS
Titer in CSF	1,251±1,800, N = 34	416±852, N = 77	**0.014**
Cerebrospinal fluid			
Cell number (/mm^3)	11.6±16.6, N = 34	3.2±3.5, N = 75	**<0.001**
Protein (mg/dl)	55.3±24.3, N = 34	36.7±13.0, N = 75	**<0.001**
Neopterin (pmol/ml)	74.9±107.9, N = 8	27.4±23.4, N = 27	0.255
HTLV-1 proviral loads (Copies/10^4 PBMCs)	370±327, N = 32	1,245±2,046, N = 69	**<0.001**

* Particle Aggregation Method.
Data are presented as mean values ± s.d., N = sample number.
doi:10.1371/journal.pone.0086144.t003

between f-HAM/TSP and sporadic cases. In all cases, the age of onset and initial symptoms of HAM/TSP were evaluated by the neurologists during hospitalization. Because inflammatory processes are less marked in f-HAM/TSP cases, as indicated by significantly lower protein levels in CSF, f-HAM/TSP cases may show slow progression of disease.

We need to discuss the possibility that the two groups compared represent different mode of HTLV transmission, i.e. vertical vs. sexual transmission. To clarify genetic backgrounds, sporadic HAM/TSP with seropositive carrier family members may be a more appropriate control, but are not available at present. The incidence of female cases showing no significant differences between f-HAM/TSP and sporadic cases, and between rapid and slow disease progression, might suggest less possibility of sporadic cases due to sexual transmission.

Although the subgroup of patients with rapid progression has not been clearly defined, previous studies suggest that rapid progression occurs in 10%–30% of all patients with HAM/TSP [12,14,16], and is associated with an older age of onset [14–16]. In the present study, the age of onset in patients with rapid progression was significantly older than that in patients with slow progression between f-HAM/TSP and sporadic cases, and the proportion of patients with rapid progression increased with the older age of onset (Figure 2). Among sporadic cases, cell numbers and protein levels in CSF were significantly higher in patients with rapid progression, suggesting that inflammation is more active in the spinal cords of patients with rapid progression and that cytotoxic T-lymphocyte (CTL) immune responses may be more intensive. Therefore, lower PVLs in PBMCs of patients with rapid disease progression may be attributed to the strong killing ability of the CTL. However, PVLs were higher in PBMCs of patients with HAM/TSP than in asymptomatic carriers [6]. In addition, the killing ability of CTLs in patients with HAM/TSP does not differ from that in asymptomatic carriers [18]. Hence, strong immune responses may be associated with the disease course. The onset of disease may require other factors that lead to strong immune responses. A late onset may also be associated with alterations of the immune function in HTLV-1-infected patients. Indeed, an increased age has been associated with autoimmune disorders, such as myasthenia gravis and rheumatoid arthritis, and may be partly explained by immune intolerance and accumulation of autoantibodies in older individuals [19,20].

In conclusion, we demonstrated that patients with HAM/TSP aggregate in some families. Compared with sporadic cases, the age of onset was younger and rates of disease progression were slower among familial cases, whereas HTLV-1 PVLs did not differ between f-HAM/TSP and sporadic groups. The present data suggest that factors other than HTLV-1 PVLs contribute to the disease course of HAM/TSP. Our data also suggested strong immune responses in the spinal cord of HAM/TSP patients with rapid progression. Further studies on HTLV-1, immune response to HTLV-1 and genetic factor in patients with rapid progression might provide new insights into HAM/TSP pathogenesis.

Acknowledgments

We would like to express the deepest appreciation to Yuka Komai for collecting information from patients and Dr. Moe Moe Aye for reading our manuscript.

Author Contributions

Conceived and designed the experiments: HT SI OW. Performed the experiments: SN EM. Analyzed the data: SN EM. Contributed reagents/materials/analysis tools: SN EM TM RK. Wrote the paper: SN EM.

References

1. Gessain A, Barin F, Vernant JC, Gout O, Maurs L, et al. (1985) Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet 2: 407–410.
2. Osame M, Usuku K, Izumo S, Ijichi N, Amitani H, et al. (1986) HTLV-I associated myelopathy, a new clinical entity. Lancet 1: 1031–1032.
3. Proietti FA, Carneiro-Proietti AB, Catalan-Soares BC, Murphy EL (2005) Global epidemiology of HTLV-I infection and associated diseases. Oncogene 24: 6058–6068.
4. Hinuma Y, Nagata K, Hanaoka M, Nakai M, Matsumoto T, et al. (1981) Adult T-cell leukemia: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. Proc Natl Acad Sci U S A 78: 6476–6480.

5. Uchiyama T, Yodoi J, Sagawa K, Takatsuki K, Uchino H (1977) Adult T-cell leukemia: clinical and hematologic features of 16 cases. Blood 50: 481–492.
6. Nagai M, Usuku K, Matsumoto W, Kodama D, Takenouchi N, et al. (1998) Analysis of HTLV-I proviral load in 202 HAM/TSP patients and 243 asymptomatic HTLV-I carriers: high proviral load strongly predisposes to HAM/TSP. J Neurovirol 4: 586–593.
7. Saito M (2010) Immunogenetics and the Pathological Mechanisms of Human T-Cell Leukemia VirusType 1- (HTLV-1-)Associated Myelopathy/Tropical Spastic Paraparesis (HAM/TSP). Interdiscip Perspect Infect Dis 2010: 478461.
8. Pombo-de-Oliveira MS, Carvalho SM, Borducchi D, Dobbin J, Salvador J, et al. (2001) Adult T-cell leukemia/lymphoma and cluster of HTLV-I associated diseases in Brazilian settings. Leuk Lymphoma 42: 135–144.
9. Miyamoto Y, Yamaguchi K, Nishimura H, Takatsuki K, Motoori T, et al. (1985) Familial adult T-cell leukemia. Cancer 55: 181–185.
10. Mori M, Ban N, Kinoshita K (1988) Familial occurrence of HTLV-I–associated myelopathy. Ann Neurol 23: 100.
11. De Castro-Costa CM, Araujo AQ, Barreto MM, Takayanagui OM, Sohler MP, et al. (2006) Proposal for diagnostic criteria of tropical spastic paraparesis/HTLV-I-associated myelopathy (TSP/HAM). AIDS Res Hum Retroviruses 22: 931–935.
12. Nakagawa M, Izumo S, Ijichi S, Kubota H, Arimura K, et al. (1995) HTLV-I-associated myelopathy: analysis of 213 patients based on clinical features and laboratory findings. J Neurovirol 1: 50–61.
13. Kaplan JE, Osame M, Kubota H, Igata A, Nishitani H, et al. (1990) The risk of development of HTLV-I-associated myelopathy/tropical spastic paraparesis among persons infected with HTLV-I. J Acquir Immune Defic Syndr 3: 1096–1101.
14. Gotuzzo E, Cabrera J, Deza L, Verdonck K, Vandamme AM, et al. (2004) Clinical characteristics of patients in Peru with human T cell lymphotropic virus type 1-associated tropical spastic paraparesis. Clin Infect Dis 39: 939–944.
15. Matsuzaki T, Nakagawa M, Nagai M, Usuku K, Higuchi I, et al. (2001) HTLV-I proviral load correlates with progression of motor disability in HAM/TSP: analysis of 239 HAM/TSP patients including 64 patients followed up for 10 years. J Neurovirol 7: 228–234.
16. Olindo S, Cabre P, Lezin A, Merle H, Saint-Vil M, et al. (2006) Natural history of human T-lymphotropic virus 1-associated myelopathy: a 14-year follow-up study. Arch Neurol 63: 1560–1566.
17. Takenouchi N, Yamano Y, Usuku K, Osame M, Izumo S (2003) Usefulness of proviral load measurement for monitoring of disease activity in individual patients with human T-lymphotropic virus type I-associated myelopathy/tropical spastic paraparesis. J Neurovirol 9: 29–35.
18. Asquith B, Mosley AJ, Barfield A, Marshall SE, Heaps A, et al. (2005) A functional CD8+ cell assay reveals individual variation in CD8+ cell antiviral efficacy and explains differences in human T-lymphotropic virus type 1 proviral load. J Gen Virol 86: 1515–1523.
19. Manoussakis MN, Tzioufas AG, Silis MP, Pange PJ, Goudevenos J, et al. (1987) High prevalence of anti-cardiolipin and other autoantibodies in a healthy elderly population. Clin Exp Immunol 69: 557–565.
20. Aprahamian T, Takemura Y, Goukassian D, Walsh K (2008) Ageing is associated with diminished apoptotic cell clearance in vivo. Clin Exp Immunol 152: 448–455.

Genome Medicine

METHOD Open Access

Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites

Sanaz Firouzi[1], Yosvany López[2], Yutaka Suzuki[2], Kenta Nakai[3], Sumio Sugano[1], Tadanori Yamochi[1*] and Toshiki Watanabe[1*]

Abstract

Transformation and clonal proliferation of T-cells infected with human T-cell leukemia virus type-I (HTLV-1) cause adult T-cell leukemia. We took advantage of next-generation sequencing technology to develop and internally validate a new methodology for isolating integration sites and estimating the number of cells in each HTLV-1-infected clone (clone size). Initial analysis was performed with DNA samples from infected individuals. We then used appropriate controls with known integration sites and clonality status to confirm the accuracy of our system, which indeed had the least errors among the currently available techniques. Results suggest potential clinical and biological applications of the new method.

Background

It has been more than 30 years since human T-cell leukemia virus type-I (HTLV-1) was shown to be the causative agent of adult T-cell leukemia (ATL) [1,2]. However, understanding the true nature of the multiple leukemogenic events [3] that are essential for this aggressive transformation remains elusive [4-9]. Although approximately 5% of HTLV-1-infected individuals develop ATL after a long latency period, the majority remain asymptomatic carriers (ACs) throughout their lifetimes. However, there are not enough clear determinants to distinguish between individuals who eventually develop ATL and those who remain as ACs [10,11]. To discover the factors associated with disease development, long-term prospective studies have assessed the correlation between disease outcome and proviral load (PVL), that is, the percentage of infected cells among the total peripheral blood mononuclear cells (PBMCs) [10-12]. The 'Joint Study on Predisposing Factors of ATL Development' (JSPFAD) [13] showed that a PVL higher than 4% is one of the indications of risk for progression to ATL [10]. Although an elevated PVL is currently the best characterized factor associated with a high risk of ATL development, a high PVL alone is not sufficient for disease prediction, suggesting the need to discover additional predictive factors [10,11].

Because ATL is a malignancy caused by HTLV-1 infection, both the integration of provirus into the host genome and the clonal expansion of infected cells are highly critical leukemogenic events [6,7,14,15]. Although many studies have addressed these aspects, the mechanism of HTLV-1 clonal expansion has not been elucidated [15-35]. Accurate monitoring for changes in clonality occurring before, during, and after ATL development is of great interest and of major clinical significance not only to clarify the underlying mechanisms but also to discover reliable predictive biomarkers for disease progression.

A broad range of evidence strongly supports that most neoplasms are composed of clonally expanded cell populations [36-38]. Owing to its biological significance, the concept of clonal expansion in cancer biology has been investigated using a variety of approaches in many tumor types [36-39], including ATL [6,15,16,18-20,22,24,29-32].

* Correspondence: yamochi@mgs.k.u-tokyo.ac.jp; tnabe@k.u-tokyo.ac.jp
[1]Department of Medical Genome Science, Graduate School of Frontier Sciences, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
Full list of author information is available at the end of the article

Clonal proliferation of HTLV-1-infected cells was first detected as monoclonal-derived bands by southern blotting [33]. Early studies found that monoclonal integration of HTLV-1 is a hallmark of ATL cells [16]. Furthermore, it was suggested that detecting a monoclonal band is useful for diagnosis and is associated with a high risk of ATL development [29,30]. Subsequent PCR-based methods included inverse PCR, linker-mediated PCR, and inverse long PCR, which enabled analysis of samples with clonality below the detection threshold of southern blotting [17,25,31,34]. Based on the observed banding patterns, the clonality of the samples was described as having undergone monoclonal, oligoclonal, or polyclonal expansion. Such PCR-based analyses revealed that, in addition to a monoclonal proliferation of infected cells, a monoclonal or polyclonal proliferation occurs even in non-malignant HTLV-1 carriers [31,35]. Moreover, considering the stability of the HTLV-1 proviral sequence, it was hypothesized that maintaining a high PVL is achieved by persistent clonal proliferation of infected cells *in vivo* [25]. This hypothesis was further supported by the detection of a particular HTLV-1 clone in the same carrier over the course of several years [18]. Two Miyazaki cohort studies focused on the maintenance and establishment of clonal expansion: Okayama *et al.* analyzed the maintenance of a preleukemic clone in an AC state several years prior to ATL onset [19], and Tanaka *et al.* assessed the establishment of clonal expansion by comparing the clonality status of long-term carriers with that of seroconverters. They showed that some of the clones from long-term carriers were stable and large enough to be consistently detectable by inverse long PCR; however, those from seroconverters were unstable and rarely detectable over time [20].

Knowledge provided by conventional studies has shed light on the next challenges worthy of further investigation. Owing to technical hurdles, however, previous studies isolated small numbers of integration sites from highly abundant clones and detected low abundant clones in a non-reproducible manner [22,34]. Furthermore, conventional techniques could not provide adequate information regarding the number of infected cells in each clone (clone size) [22]. To effectively track and monitor HTLV-1 clonal composition and dynamics, we considered devising a new method that would not only enable the high-throughput isolation of integration sites but also provide an accurate measurement of clone size.

PCR is a necessary step for the integration site isolation and clonality analysis. However, bias in the amplification of DNA fragments (owing to issues such as extreme fragment length and high GC content) is intrinsic to any PCR-based method [40-45]. Different fragment amplification efficiencies make it difficult to calculate the amount of starting DNA (the original distribution of template DNA) from PCR products. Hence, estimating HTLV-1 clonal abundance, which requires calculating the number of starting DNA fragments, is only achievable by avoiding the PCR bias.

Recently, Bangham's research group analyzed HTLV-1 clonality and integration site preference by a high-throughput method [22]. In the method developed by Gillet *et al.*, clone sizes were estimated using length of DNA fragments (shear sites generated by sonication) as a strategy for removing PCR bias [22]. Owing to the limited variation in DNA fragment size observed with shearing, the probability of generating starting fragments of the same lengths is high, leading to a nonlinear relationship between fragment length and clone size [22,46]. Therefore, Gillet *et al.* used a calibration curve to statistically correct the shear site data [22]. Later, Berry *et al.* introduced a statistical approach, and further addressed the difficulties of estimating clone size from shear site data [46]. Their approach estimates the size of small clones with little error, but estimates for larger clones have greater error [46]. A parameter adopted from the Gini coefficient [47,48] and termed the oligoclonality index was used to describe the size and distribution of HTLV-1 clones [22]. It has been demonstrated that the oligoclonality index differs between malignant and non-malignant HTLV-1 infections, and also a high PVL of HTLV-1-associated myelopathy is due to cells harboring large numbers of unique integration sites [22]. Furthermore, genome-wide integration site profiling of clinical samples revealed that the abundance of a given clone *in vivo* correlates with the features of the flanking host genome [22,24]; although there was not a specific hotspot, HTLV-1 more frequently integrated in transcriptionally active regions of the host genome [22,24]. These findings further clarified the characteristics of HTLV-1 integration sites, and strongly suggested the importance of HTLV-1 clonal expansion *in vivo*.

Here we introduce a method that overcomes many of the limitations of currently available methods. Taking advantage of next-generation sequencing (NGS) technology, nested-splinkerette PCR, and a tag system, we designed a new high-throughput method that enables specific isolation of HTLV-1 integration sites and, most importantly, allows for the quantification of clonality not only from the major clones and high-PVL samples but also from low-abundance clones (minor clones) and samples with low PVLs. Moreover, we conducted comprehensive internal validation experiments to assess the effectiveness and accuracy of our new methodology. A preliminary validation was conducted by analyzing DNA samples from HTLV-1-infected individuals with different PVLs and disease status. Subsequently, an internal validation was performed that included an appropriate control with known integration sites and clonality patterns. We present our methodology, which illustrates that employing the tag system is effective for improving quantification of clonal abundance.

Methods

Our clonality analysis method included two main aspects: (1) wet experiments, and (2) *in silico* analysis (Additional file 1: Figure S1). A general explanation of materials and methods is provided here, and detailed protocols of the wet experiments are included in Additional file 1: Notes. The *in silico* analysis is further described in Results and discussion.

NGS data have been deposited in the Sequence Read Archive of NCBI with access number of (SRP038906).

Wet experiments

Biological samples: specimens and cell lines

Specimens: In total five clinical samples were provided by a biomaterial bank of HTLV-1 carriers, JSPFAD [13,49]. The clinical samples were a part of those collected with an informed consent as a collaborative project of JSPFAD. The project was approved by the Institute of Medical Sciences, the University of Tokyo (IMSUT) Human Genome Research Ethics Committee. Information about the disease status of samples was obtained from JSPFAD database in which HTLV-1-infected individuals were diagnosed based on the Shimoyama criteria [50]. In brief, genomic DNA from PBMCs was isolated using a QIAGEN Blood kit. PVLs were measured by real-time PCR using the ABI PRISM 7000 Sequence Detection System as described in [10].

Cell lines: An IL2-dependent TL-Om1 cell line [51] was maintained in RPMI 1640 medium supplemented with 10% heat-inactivated fetal calf serum (GIBCO), 1% penicillin-streptomycin (GIBCO), and 10 ng/mL IL2 (R&D systems). The same conditions as those of patient samples were used to extract DNA and measure PVL.

Illumina-specific library construction

We employed a library preparation protocol specifically designed to isolate HTLV-1 integration sites. The final products in the library that we generated contained all the specific sequences necessary for the Illumina HiSeq 2000 platform (Additional file 1: Figure S2). These products included a 5′-flow cell binding sequence, a region compatible with read-1 sequencing primer, 5-bp random nucleotides, 5-bp known barcodes for multiplexing samples, HTLV-1 long terminal repeat (LTR), human or HTLV-1 genomic DNA, a region compatible with read-2 and read-3 sequencing primers, 8-bp random tags, and a 3′-flow cell binding sequence from 5′ to 3′, respectively (Additional file 1: Figure S2B).

Incorporating the 5-bp random nucleotides downstream of the region compatible with the read-1 sequencing primer was critical and resulted in high-quality sequence data. We used a library designed without the first 5-bp of random nucleotides as input for the HiSeq 2000 sequencer in our first samples (S-1, S-2, S-3, and S-4). Because all fragments began with the same LTR sequence, clusters generated in the flow cells could not be differentiated appropriately. These samples resulted in low-quality sequence data (see Additional file 1: Notes). Designing the first 5-bp randomly resulted in high-quality sequence data for the remaining samples because clusters were differentiated with no problem during the first five cycles of sequencing (data not shown).

Our library construction pipeline comprised the following four steps (Additional file 1: Figure S2) (Additional file 1: Notes):

(1) DNA isolation: DNA was extracted as described above, and the concentration of extracted DNA was measured with a NanoDrop 2000 spectrophotometer (Thermo Scientific). We recommend using 10 μg of DNA as the starting material. However, in practice there are some rare clinical samples with limited DNA available. In order to be able to handle those samples, the method was also optimized for 5 μg and 2 μg of starting DNA.

(2) Fragmentation: According to the protocol provided in Supplementary Notes, the starting template DNA was sheared by sonication. The resulting fragments represented a size range of 300 to 700 bp as checked by an Agilent 2100 Bioanalyzer and DNA 7500 kit (Figure 1B).

(3) Pre-PCR manipulations: Four steps of end repair, A-tailing, adaptor ligation, and size selection were performed as described in Additional file 1: Notes.

(4) PCR: To amplify the junction between the genome and the viral insert, we used nested-splinkerette PCR (a variant of ligation-mediated PCR [52,53]) (Additional file 1: Figure S2). We confirmed that the technique specifically amplifies HTLV-1 integration sites; since there was no non-specific amplification neither from human endogenous retroviruses nor from an exogenous retrovirus such as HIV (see Additional file 1: Table S1 and Additional file 2: Figure S1).

Information on oligonucleotides, including adaptors and primers, and the LTR and HTLV-1 reference sequences [54] are provided in Additional file 1: Table S1. The final PCR products were sequenced using the HiSeq 2000 platform.

In silico analysis

Raw sequencing data were processed according to the workflow described in the Results and discussion section.

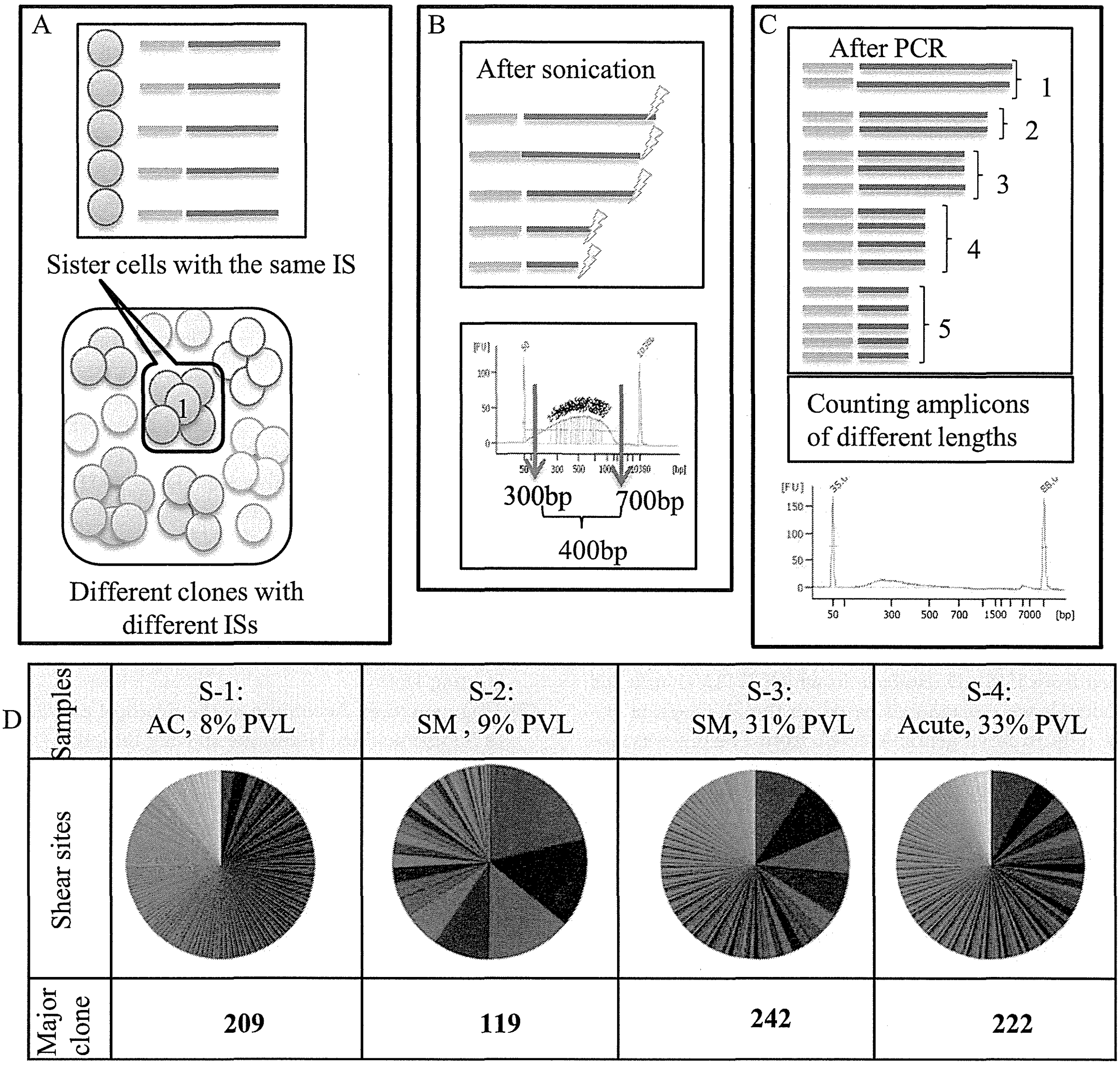


Figure 1 Estimating clone size by 'shear sites'. Also see Additional file 2: Figure S2 for a simple image from an integration site and its shear sites. **(A)** Depicted is the complex population of uninfected cells (grey circles) together with infected clones (circles of different colors). A clone is shown as a group of sister cells (circles of the same color) having the same integration site (IS). Different clones are distinguishable based on differing integration sites, and thus the number of integration sites represents the number of infected clones. For example, the six different unique integration sites refer to six unique clones. **(B)** Genomic DNA fragmented by sonication generates random shear sites (fragments of different length). Fragment size, measured by an Agilent Bioanalyzer, ranged from 300 to 700 bp. This size range can theoretically provide approximately 400 variations. **(C)** The size distribution of fragments decreased following amplification by integration-site-specific PCR. From the deep sequencing data, the original number of starting fragments could be estimated by removing PCR duplicates and counting fragments with different lengths. For example, five different lengths of PCR amplicons represent five infected sister cells. **(D)** We analyzed four samples, including (S-1: asymptomatic carrier (AC), (8% PVL)), (S-2: smoldering (SM), (9% PVL)), (S-3: smoldering, (31% PVL)), and (S-4: acute, (33% PVL)). Using our method, the clone sizes were quantified by considering only shear sites. The first major clone (the largest clone) of each sample was mapped to (chr 11-41829319 (+)), (chr 15: 59364370 (+)), (chr 4-563543 (-)), and (chr X - 83705328 (-)), respectively. The shear site variations of each major clone were 209, 119, 242, and 222, respectively. Different colors on the pie graphs indicate different integration sites, and the size of each piece represents the clone size.

The initial forward read (100-bp) was termed Read-1 and the reverse read (100-bp) was termed Read-3 and an index read (8-bp) was termed Read-2. In brief, analysis programs were written in Perl language and run on a supercomputer system provided by The University of Tokyo's Human Genome Center at The Institute of Medical Science [55]. The sequencing output was check for quality using the FastQC tool [56]. The regions corresponding to the LTR and HTLV-1 genome were subjected to a blast search against the reference sequences described in Additional file 1: Table S1. Following isolation of the integration sites, the flanking human sequences were mapped to the human genome (hg19) (the UCSC genome browser [57]) by Bowtie 1.0.0 [58]. The final processed data included information about shear sites (R1R3), tags (R1R2), and a combination of tags and shear sites (R1R2R3). Fitting the data to the zero truncated Poisson distribution for retrieving correlation coefficients were done by the R-package 'gamlss.tr' [59]. The Gini coefficient was calculated by StatsDirect medical statistics software [60].

Results and discussion

General concepts

We originally designed our method to overcome the limitations of conventional techniques [31,34] and to make improvements in the only existing high-throughput method [22]. In general, our method includes two main sets of wet experiments and an *in-silico* analysis. We used genomic DNA as the starting material to prepare an appropriate library for Illumina sequencing. Subsequently, deep-sequencing data were analyzed by a supercomputer. The resulting information represents the clonality status of each sample (Additional file 1: Figure S1).

There are complex populations of infected clones and uninfected cells in a given HTLV-1 infected individual. High-throughput clonality analysis requires monitoring two main characteristics of clones: HTLV-1 integration sites and the number of infected cells in each clone (clone size). Each HTLV-1-infected cell naturally harbors only a single integration site [23]. Therefore, the number of detected unique integration sites corresponds to the number of infected clones. Based on our analysis, which is consistent with the data of Gillet *et al.* [22], employing high-sensitivity deep sequencing allowed for the isolation of a large number of unique integration sites (UISs), including samples with low PVLs (Figure 1). We analyzed four samples from HTLV-1-infected individuals with different PVLs, disease status, and expected clonality patterns. The samples include S-1: AC (8% PVL); S-2: smoldering ATL (SM) (9% PVL); S-3: SM (31% PVL); and S-4: acute ATL (33% PVL). Based on the final optimized conditions, 1030, 39, 265, and 384 UISs were isolated from each sample, respectively (Figure 1).

The most challenging aspect of our clonality analysis was estimating the number of infected cells in each clone. Although a necessary step in the analysis, PCR introduces a bias in the frequency of starting DNA material [40-45]. Because amplification causes significant changes in the initial frequency of starting materials, PCR products cannot be used directly to estimate the amount of the starting DNA material. To overcome this problem, we needed to manipulate DNA fragments to make them unique prior to PCR amplification. Thus, if each DNA fragment could be marked with a unique feature, it would then be possible to calculate its frequency based on the frequency of that unique feature. When a single unique stretch of DNA is amplified by PCR, the resulting product is a cluster of identical fragments termed PCR duplicates. Therefore, to estimate the frequency of starting DNA fragments, one should count the number of clusters with unique features. The remaining technical question then becomes how to mark the starting DNA prior to PCR amplification. In the following section, we compare and discuss two main strategies, namely (1) shear sites and (2) a tag system, which enable DNA fragments to be uniquely marked.

Estimating the size of clones by shear sites

The first strategy, described by Gillet *et al.*, relies on shearing DNA by sonication, resulting in fragments of random length [22]. Sonication-derived shear sites were thus used as a distinguishing feature to make fragments unique prior to PCR. Clone sizes were then estimated by statistical approaches [22,46].

To directly assess the effectiveness of the shear site strategy, we analyzed the clonality of the aforementioned clinical samples (S-1, S-2, S-3, and S-4). Genomic DNA was cleaved by sonication with fragments in the 300- to 700-bp range, theoretically providing approximately 400 possible variations in fragment size (Figure 1A and B). Following library construction, however, the final product represented smaller size ranges, implying a relatively limited number of variations (Figure 1C). Finally, the number of PCR amplicons with unique shear sites was retrieved from deep-sequencing data. See Additional file 2: Figure S2 for a simple image from an integration site and its shear sites. The data obtained from the shear site experiments were not fitted to calibration curves or statistical treatments, which were used by Gillet *et al.* and Berry *et al.*, respectively (See Additional file 1: Notes) [22,46]. For clarity, only the information relating to the major clone of each sample is provided in Figure 1D. The shear-site variations of the major clone were 209, 119, 242, and 222 for samples S-1 through S-4, respectively. Even in the case of control samples with 100% PVLs, the shear sites did not provide more than 225 variations (see Validation of the methodology). However, it was expected that samples with differing PVLs and disease status would harbor varying numbers of sister cells, at least in their major clones. Similar variations of shear sites were observed in major clones of AC, SM, and acute samples. These data suggest that, because the number of sister cells in each clone exceeded the shear site variations, the size of the clones was underestimated (Figure 1). This is most problematic in the case of large clones and leads to an underestimation of the clone size.

Measuring the size of clones by the tag system

We developed an alternate strategy to remove PCR bias and to estimate starting DNA. We designed a tag system in which 8-bp random nucleotides are incorporated at the end of DNA fragments during adaptor ligation step. Each tag acts as a molecular barcode, which gives each DNA fragment a unique signature prior to PCR. Information on the frequency of observed tags from the deep-sequencing data can be used to remove the PCR duplicates and thereby estimate the original clonal abundance in the starting sample. Owing to their random design, the tags could theoretically provide approximately 65,536 variations. This degree of potential variation is expected to provide a unique tag for a large number of sister cells in each clone (Figure 2).

We analyzed samples S-1, S-2, S-3, and S-4 to assess the effectiveness of our tag system for estimating clone size. The major clone of each sample showed tag variations of 393, 142, 1751, and 2675, respectively (Figure 2D). Similar variations of tags and shear sites were observed in the

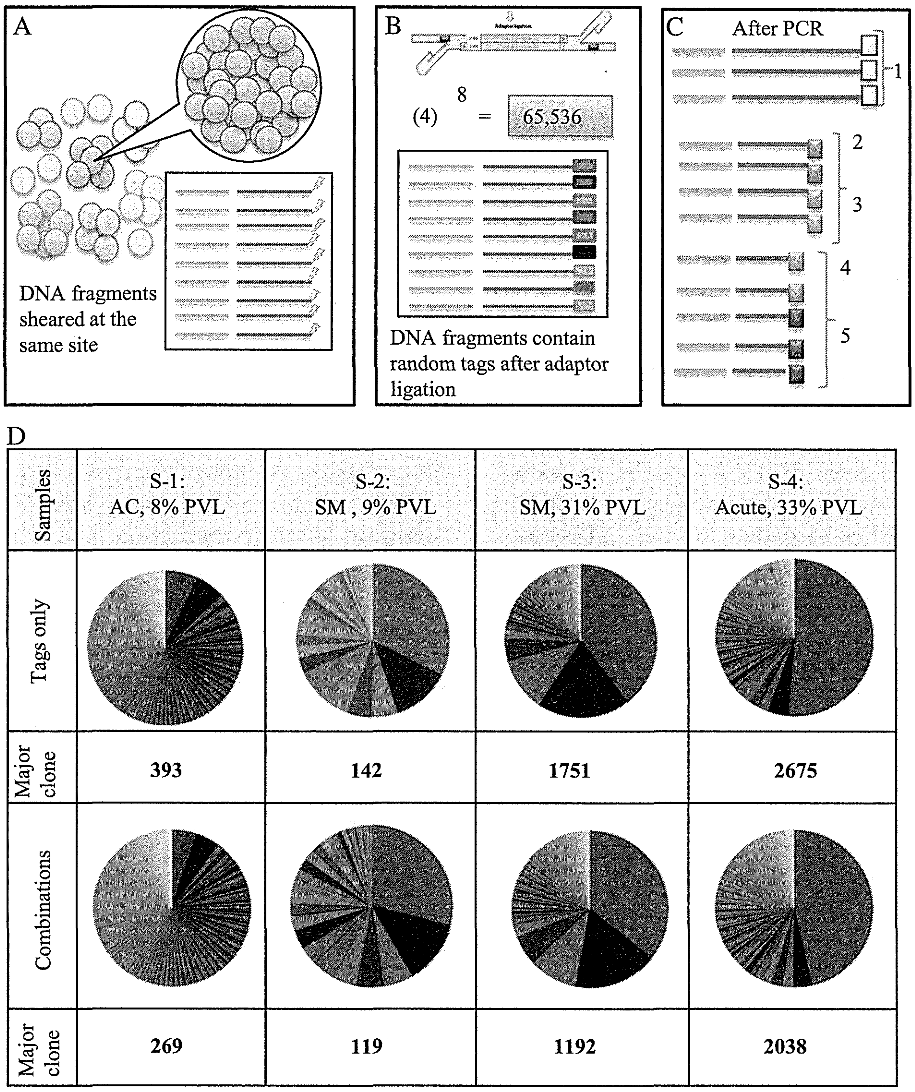


Figure 2 Measuring clone size using the tag system. (A) The depiction above shows that shear site variations are not able to cover all sister cells in large clones. As the number of the sister cells in a given clone increases, the probability of DNA shearing at the same site increases. **(B)** Prior to PCR, we incorporated 8-bp random tags into each DNA fragment to uniquely mark them. Random tags could theoretically provide approximately 65,536 variations. The number of potential variations is expected to amply cover large numbers of the sister cells. **(C)** The tag information was used to remove PCR duplicates and to estimate the original number of starting fragments. If the fragments had the same shear sites but different tags, they were counted separately. For example, here five different combinations of tags and shear sites represent five infected cells. **(D)** Samples: S-1, S-2, S-3, and S-4 were analyzed by the final optimal condition (Bowtie parameters: -v 3 - - best, and filtering condition: (merging approach) JT-10). Clone size was measured by tags only or by the combination of shear sites and tags. The covered variations were (393,142, 1751, and 2675) and (269, 119, 1192, and 2038), respectively.

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.

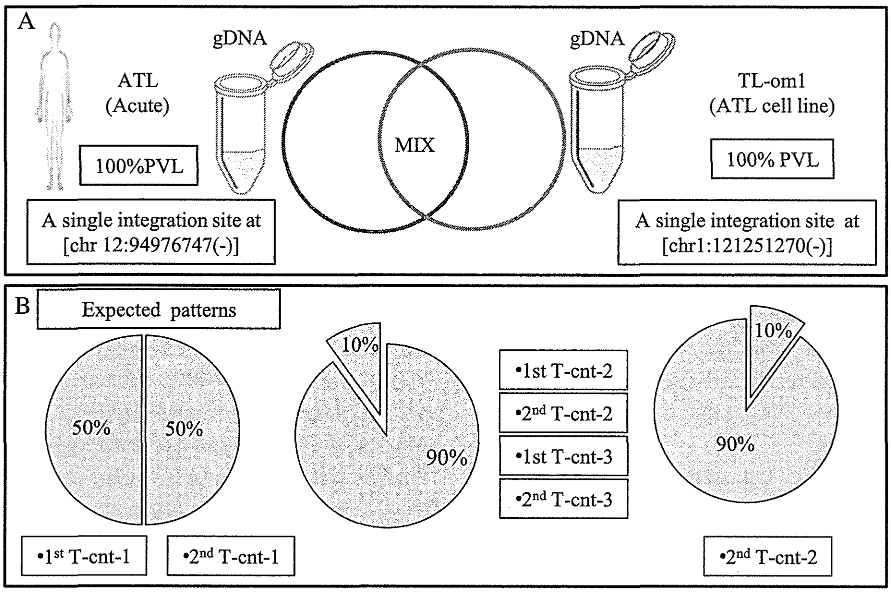
In the first trial, samples were named as 'first trial control 1 ~ 4' and abbreviated as 1st T-cnt-1 ~ 4. Various amounts of DNA (μg) from samples M and T were mixed to prepare the final expected clone sizes as shown in Figure 3C. A 1-μL sample of a 10-fold dilution of external PCR product was used as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



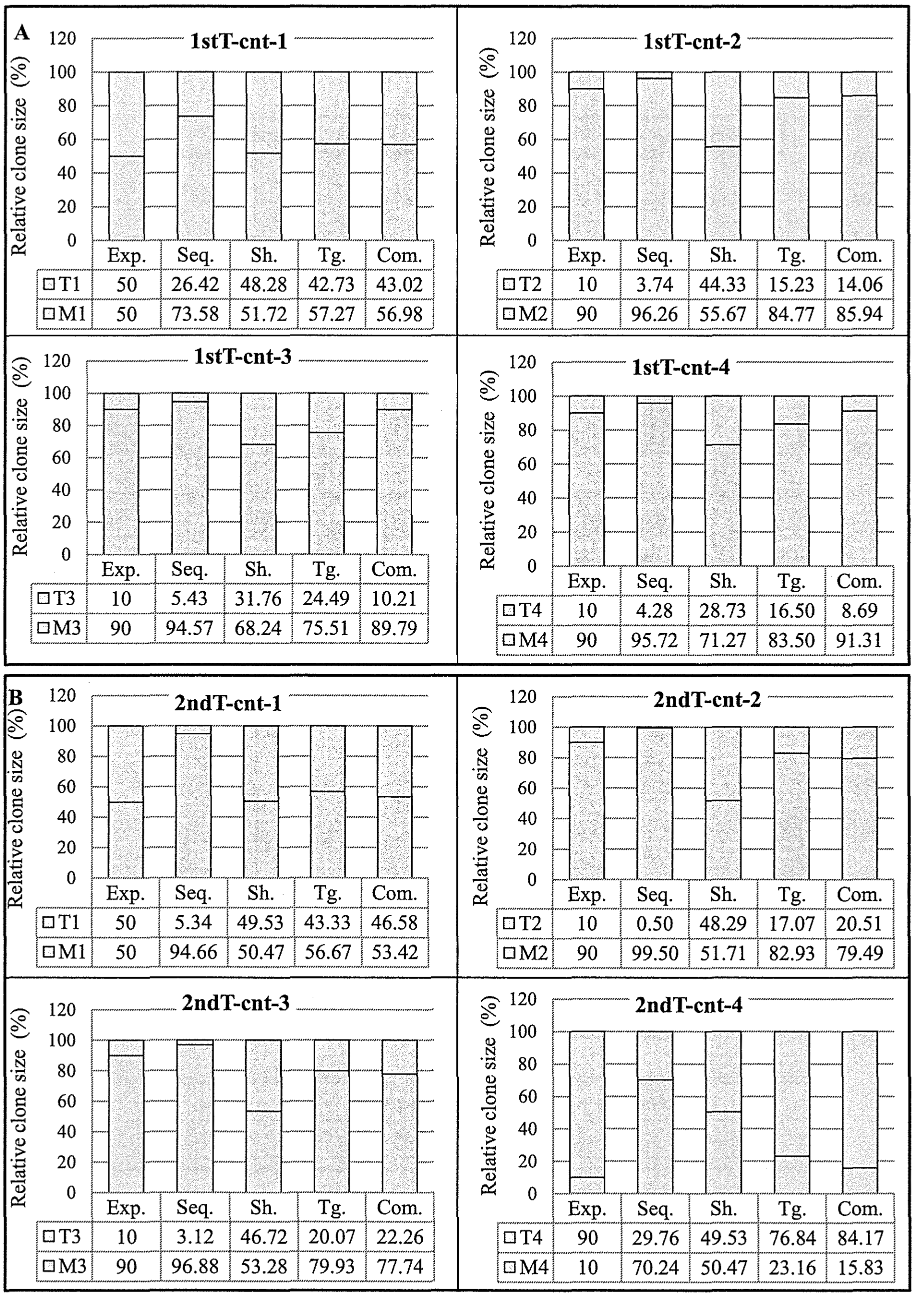
C The first trial	1st T-cnt-1		1st T-cnt-2		1st T-cnt-3		1st T-cnt-2	
Total amount of DNA	10 µg		10 µg		5 µg		2 µg	
The amount of DNA from TL-om1(T) vs. major clone(M)	M	T	M	T	M	T	M	T
	5 µg	5 µg	9 µg	1 µg	4.5 µg	0.5 µg	2.25 µg	0.25 µg
Expected clone size	50%	50%	90%	10%	90%	10%	90%	10%

D The second trial	2nd T-cnt-1		2nd T-cnt-2		2nd T-cnt-3		2nd T-cnt-2	
Total amount of DNA	10 µg		10 µg		5 µg		5 µg	
The amount of DNA from TL-om1 (T) vs. major clone (M)	M	T	M	T	M	T	M	T
	5 µg	5 µg	9 µg	1 µg	4.5 µg	0.5 µg	0.5 µg	4.5 µg
Expected clone size	50%	50%	90%	10%	90%	10%	10%	90%

E	Chromosome	Strand	Position	A representative sequence for each corresponding integration site position
1	chr 1	-	121251270	TATATGTAGCACAATTTCTTTATTCAGTCTGTCATTGTTG
2	chr 12	-	94976747	AAAAAAGATTCTCCTTCTATTAAGTGAGTGAGTTCTGAGT

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



1stT-cnt-1

	Exp.	Seq.	Sh.	Tg.	Com.
□T1	50	26.42	48.28	42.73	43.02
□M1	50	73.58	51.72	57.27	56.98

1stT-cnt-2

	Exp.	Seq.	Sh.	Tg.	Com.
□T2	10	3.74	44.33	15.23	14.06
□M2	90	96.26	55.67	84.77	85.94

1stT-cnt-3

	Exp.	Seq.	Sh.	Tg.	Com.
□T3	10	5.43	31.76	24.49	10.21
□M3	90	94.57	68.24	75.51	89.79

1stT-cnt-4

	Exp.	Seq.	Sh.	Tg.	Com.
□T4	10	4.28	28.73	16.50	8.69
□M4	90	95.72	71.27	83.50	91.31

2ndT-cnt-1

	Exp.	Seq.	Sh.	Tg.	Com.
□T1	50	5.34	49.53	43.33	46.58
□M1	50	94.66	50.47	56.67	53.42

2ndT-cnt-2

	Exp.	Seq.	Sh.	Tg.	Com.
□T2	10	0.50	48.29	17.07	20.51
□M2	90	99.50	51.71	82.93	79.49

2ndT-cnt-3

	Exp.	Seq.	Sh.	Tg.	Com.
□T3	10	3.12	46.72	20.07	22.26
□M3	90	96.88	53.28	79.93	77.74

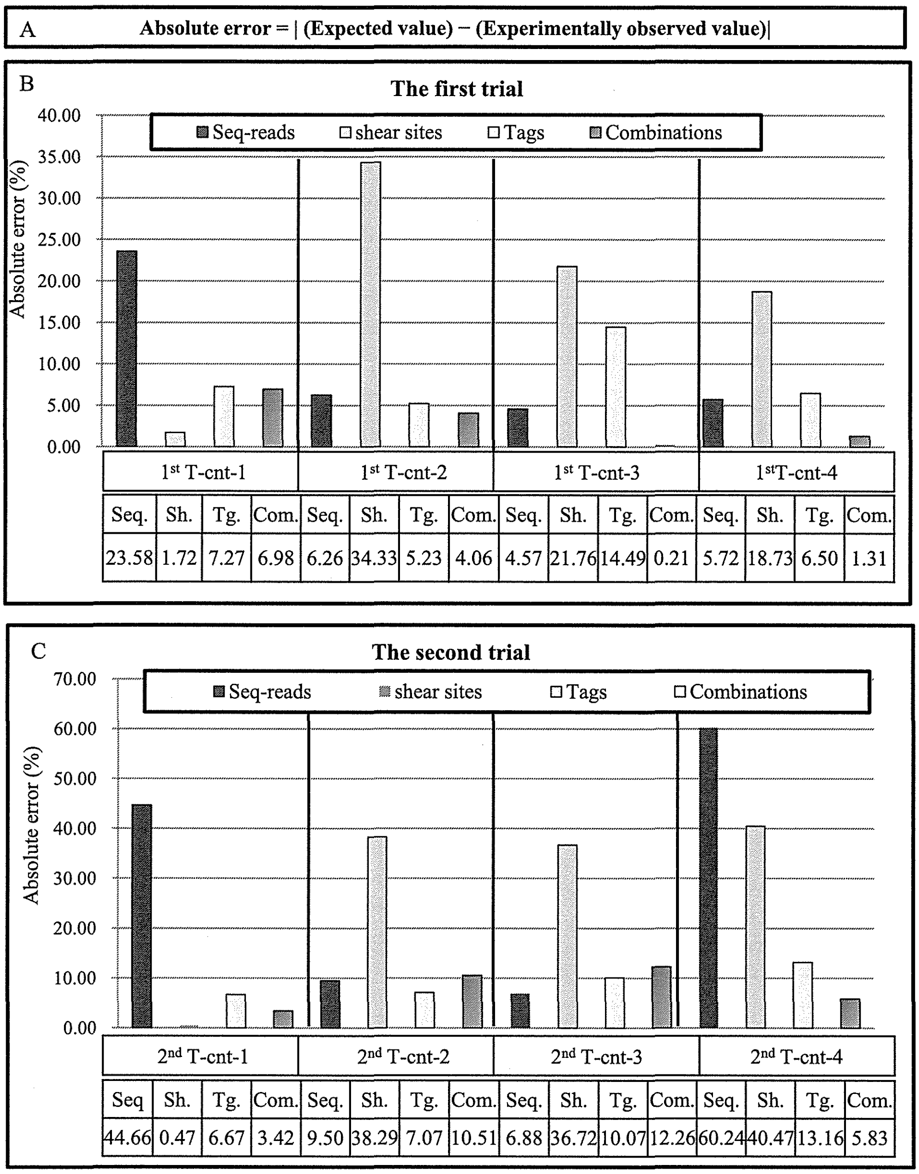
2ndT-cnt-4

	Exp.	Seq.	Sh.	Tg.	Com.
□T4	90	29.76	49.53	76.84	84.17
□M4	10	70.24	50.47	23.16	15.83

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



1st T-cnt-1				1st T-cnt-2				1st T-cnt-3				1stT-cnt-4			
Seq.	Sh.	Tg.	Com.	Seq.	Sh.	Tg.	Com.	Seq.	Sh.	Tg.	Com.	Seq.	Sh.	Tg.	Com.
23.58	1.72	7.27	6.98	6.26	34.33	5.23	4.06	4.57	21.76	14.49	0.21	5.72	18.73	6.50	1.31

2nd T-cnt-1				2nd T-cnt-2				2nd T-cnt-3				2nd T-cnt-4			
Seq	Sh.	Tg.	Com.	Seq.	Sh.	Tg.	Com.	Seq.	Sh.	Tg.	Com.	Seq.	Sh.	Tg.	Com.
44.66	0.47	6.67	3.42	9.50	38.29	7.07	10.51	6.88	36.72	10.07	12.26	60.24	40.47	13.16	5.83

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

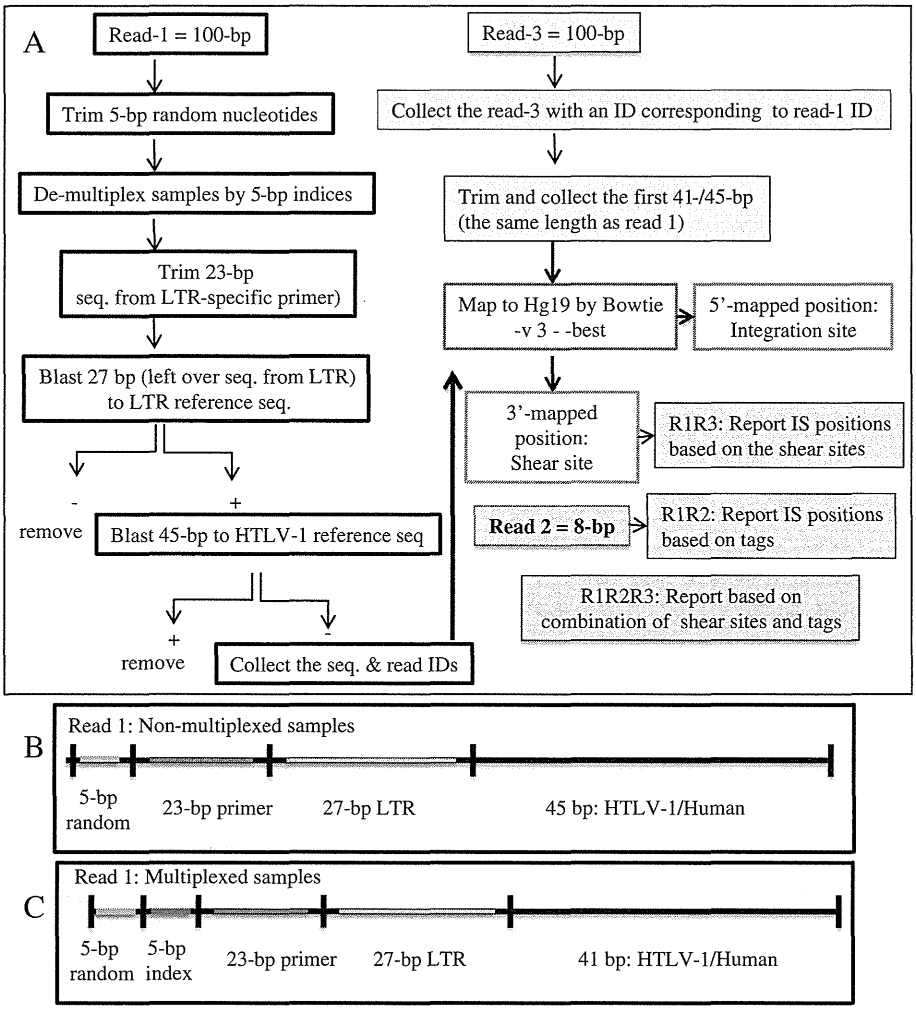

A
Read-1 = 100-bp
Trim 5-bp random nucleotides
De-multiplex samples by 5-bp indices
Trim 23-bp seq. from LTR-specific primer)
Blast 27 bp (left over seq. from LTR) to LTR reference seq.
-
remove
+
Blast 45-bp to HTLV-1 reference seq
+
remove
-
Collect the seq. & read IDs
Read-3 = 100-bp
Collect the read-3 with an ID corresponding to read-1 ID
Trim and collect the first 41-/45-bp (the same length as read 1)
Map to Hg19 by Bowtie -v 3 - -best
5'-mapped position: Integration site
3'-mapped position: Shear site
R1R3: Report IS positions based on the shear sites
Read 2 = 8-bp
R1R2: Report IS positions based on tags
R1R2R3: Report based on combination of shear sites and tags
B
Read 1: Non-multiplexed samples
5-bp random
23-bp primer
27-bp LTR
45 bp: HTLV-1/Human
C
Read 1: Multiplexed samples
5-bp random
5-bp index
23-bp primer
27-bp LTR
41 bp: HTLV-1/Human

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