Doctoral thesis

Monitoring the clonal composition of HTLV-1-infected cells

HTLV-1 感染細胞のクローナリティ解析

Firouzi, Sanaz

フィルジ　サナーズ

(47-127338)

The Laboratory of Tumor Cell Biology

(Professor Watanabe, Toshiki)

Department of Medical Genome Science

Graduate School of Frontier Sciences

The University of Tokyo

2015

Table of Contents

Abstract and summary of results……Page-1

Introduction……Page-5

General background……..…Page-8
-Background: Clonality of HTLV-1 infected cells by previous studies…..…Page-12

Results -Section-1-A
General concepts……Page-19
Estimating the size of clones by shear sites……Page-19
Measuring the size of clones by the tag system……..Page-20
Estimating clone size by shear sites vs. tags……..Page-21
Oligoclonality index: shear sites vs. tag system……Page-21
Validation of the methodology…….....…..Page-31
Evaluating the accuracy of the clonality analyzed based on shear sites *vs.* tags system……..……Page-32
***In-silico* analysis…….……...Page-37**
Removing background noise……..…...Page-38
Mapping, reads coverage, and tag variations……..……...…Page-38
PCR-southern experiments…….……....…….Page-46

Results-section-1-B
Confirming reproducibility of results using clinical samples…….....……Page-49
Discussion-section-1-B
impact on Clinical Diagnosis: How does an accurate clonality assessment aid the clinician making therapeutic decisions?.....Page-49

Results-section-1-C
Examination of the accuracy of my new method by monitoring of clonality alterations over time…….…...…..Page-56
Discussion-section-1-C
Impact on Prognosis and Prevention: Demand for an effective prognostic indicator of ATL onset……Page-56

Results -Section-2……..……..……...Page-62

Final discussion……...Page-67

Experimental design, Material and Methods…….……...Page-71

References…….……Page-85

Acknowledgements……..…Page-95

論 文 の 内 容 の 要 旨

論文題目 **Monitoring the clonal composition of HTLV-1-infected cells（HTLV−1 感染細胞のクローナリティ解析）**

氏　　名　フィルジ　サナーズ

<Abstract>

Human T-cell leukemia virus type-I (HTLV-1) mainly survives *in vivo* by persistent proliferation of infected cells. HTLV-1 infection is the initial necessary event of multiple leukemogenic events that lead to adult T-cell leukemia (ATL) onset. As the first generation of studies on ATL risk factors, our Joint Study on Predisposing Factors of ATL Development (JSPFAD) group demonstrated that a proviral load (PVL) >4% is one of the risk factors for progression to ATL; however, PVL alone cannot predict development of the disease. Moreover, how the threshold of 4% and high levels of PVL are maintained and how they contribute to ATL onset remains to be elucidated.

Thus, with revolutionized insights, I have started a next generation of studies on prevention and revealing the molecular mechanisms of ATL development. For this purpose, I developed and validated an original methodology to detect clonality as accurate as qPCR while also taking advantage of the ability of deep sequencing to precisely characterize and distinguish large numbers of infected clones, based on provirus integration sites. This new methodology has been published, and is currently the most reliable method for accurate analysis of HTLV-1 clonality that can be used for clinical applications worldwide [Firouzi *et al. Genome medicine* 2014]. The realization of the potential clinical applications of this methodology will have far-reaching impacts on the diagnosis, prognosis, and treatment of infected individuals. Here I present my original methodology and part of pilot data on the clonal composition of HTLV-1 infected cells.

Summary of results

<Background and the necessity of this study>

ATL is a highly aggressive leukemia of T-cells, with an extremely poor prognosis and a short median survival time due to development of multidrug resistance. Prevention and treatment of ATL remain to be unresolved problems.

In 2002 JSPFAD was established as a nationwide collaborative study group to collect biomaterial samples from individuals infected with human T-cell leukemia virus type-I (HTLV-1) to facilitate research on the mechanisms and risk factors associated with ATL development. In the first generation of studies on ATL risk factors, JSPFAD assessed the correlation between disease outcome and proviral load (PVL). PVL represents the burden of HTLV-1 infection, defined as the percentage of infected cells among the total peripheral blood mononuclear cells (PBMCs), accurately measurable by qPCR. The JSPFAD initiative has currently collected > 9000 samples, all of which have had their PVL measured in our laboratory. PVL levels are different among infected individuals, with patients with malignant ATL having a significantly higher PVL than asymptomatic carriers (ACs). The initial JSPFAD study showed that a PVL >4% is one of the risk factors for progression to ATL [Iwanaga *et al. Blood* 2010]. However, some of ACs have abnormally high PVLs but do not develop ATL, and some infected individuals with low PVLs develop acute ATL. Thus, although an elevated PVL is currently the best-characterized risk factor associated with ATL development, a high PVL alone is not sufficient to predict disease progression and there is a need to discover additional predictive factors. Moreover, the mechanisms behind the maintenance of high levels of PVL and how these high PVLs lead to ATL onset need to be elucidated.

HTLV-1 infection is the initial necessary event among the multiple leukemogenic events that lead to ATL onset. HTLV-1 integrates into the human genome and maintains itself *in vivo* through persistent clonal growth of primarily infected cells. Following a long latency period of 40–60 years, about 5% of infected individuals convert from a polyclonal population of HTLV-1 infected cells into a monoclonal pattern that terminates in ATL onset. The monoclonal proliferation of HTLV-1-infected cells as a hallmark of ATL was

first detected by Southern blotting showing monoclonal bands [Yoshida *et al*. 1984]. Later, PCR-based analyses isolated HTLV-1 provirus integration sites and revealed that in addition to a monoclonal proliferation of infected cells, an oligoclonal or polyclonal proliferation occurs even in nonmalignant HTLV-1 carriers [Wattel *et al.* 1995, Etoh *et al.* 1997]. The overall proliferation levels of infected cells (PVL) are quantifiable by qPCR, and the general patterns of proliferation can be identified by the conventional techniques of Southern blot and inverse PCR. However, in-depth monitoring of the clonal composition of infected cells requires an advanced quantitative method that fulfills the three main criteria:

(1) High throughput isolation of a large numbers of integration sites (2) Detection of low abundance clones with high sensitivity even from the sample with low PVLs (3) Accurate measurement of the number of the infected cells in each clone (clone size).

Recently, a research group from the Imperial College of London devised a method that met only the first two criteria. Their method employed sonication to shear DNA to generate fragments of different lengths as a strategy for making unique fragments prior to PCR for the determination of clone size [Gillet *et al. Blood 2011]*. 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; thus, introducing high error with this method. Therefore, Gillet *et al.* used a calibration curve to statistically correct the shear site data. However, even with these statistical corrections, they had a bias of at least >20% in the prediction of large clones [Berry *et al. Bioinformatics* 2012].

I showed that a major problem with shear site strategy of Gillet *et al*. is that practically shear sites can provide <250 variations [Firouzi *et al. Genome medicine* 2014]. This number of variations is not enough to accurately estimate the size of clones because most of the time, the number of infected cells in each clone exceeds the number of variations of the shear sites. Because the incidence of large clones (clones with >250 infected cells) increases with disease progression from the healthy AC state to the malignant states of smoldering, chronic, or acute ATL, an accurate measurement of clone size, and particularly of large clones, is of great clinical significance. Because the method of Gillet *et al*. leads to an underestimation of the clone sizes, the development of an alternative methodology with a high accuracy is necessary for clinical applications.

<Results and Discussion>

<Approaches of our study to fulfill all three aforementioned criteria >

A novel methodology for accurate quantitative analysis of clonality with a potential far-reaching impact on worldwide clinical applications [Firouzi *et al. Genome medicine* 2014]

I conducted a comprehensive multidisciplinary study combining our expertise in the field of HTLV-1 with genomics and bioinformatics analysis. I took advantage of next-generation sequencing (NGS) technology, using a tag system and an i*n silico* analysis pipeline to develop and internally validate a new high-throughput methodology (Figure 1). Analyzing control samples with already known clone sizes ensured accurate measurement of the size of clones using this method. This high-throughput method enables specific isolation of HTLV-1 integration sites, and allows for accurate quantitative clonality analysis of not only the major clones and high-PVL samples but also low-abundance clones (minor clones) and samples with low PVLs (Figure 2, 3).. An original strategy to remove PCR bias and to measure clone size was developed using a tag system, in which 8-bp random nucleotides are incorporated at the end of DNA fragments. 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 PCR duplicates, and thereby more

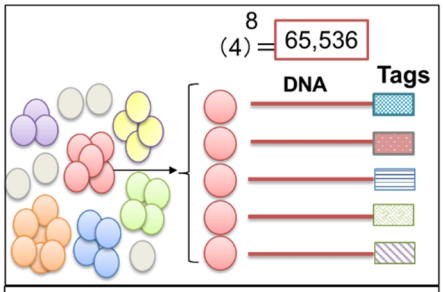


Figure 1. Prior to PCR, we incorporated 8-bp random tags into each DNA fragment to uniquely mark them. Random tags could theoretically provide approximately 65,536 variations. The number of potential variations is expected to amply cover large numbers of infected cells.

accurately estimate the original clonal abundance in the starting sample. Owing to their random design, the tags theoretically provide approximately 65,536 variations (4^8 = 65,536). This degree of potential variation is expected to provide a unique tag for a large number of sister cells in each clone (Figure 1). I proved my methodology to be reliable for isolating large numbers of integration sites and to be accurate for quantifying clone size. To the best of our knowledge, our methodology is the first in which the accurate size of clones is able to be experimentally measured without using any statistical corrections. This new methodology is currently the most reliable method for accurate analysis of HTLV-1 clonality available worldwide (Figure 2, 3) [Firouzi *et al. Genome medicine* 2014].

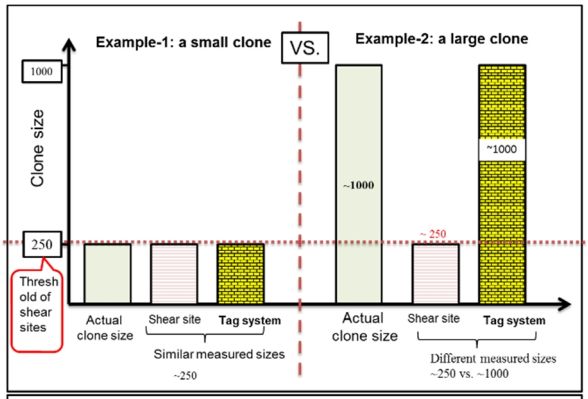


Figure 2. A simple comparison of shear site and tag variations
A simple image representing that the two approaches result in the similar clone sizes when the size of clones are small enough to be covered by shear site approach (Example-1). However, when the size of clones exceeds the shear site variations (~250 based on our experiments), clone sizes get underestimated by the shear site approach. Thus, the size of large clones are only measureable by the tag system (Example-2).

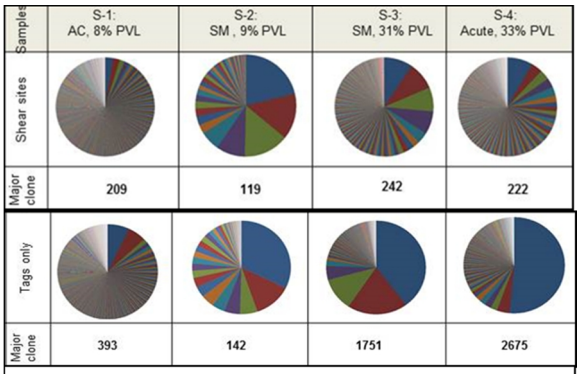


Figure 3. Clonality analysis of four clinical samples with different PVLs and disease states.The clone size of the first major clone (the largest clone) of each sample measured by shear site *vs.* tags system was [S-1: 209 *vs.* 393], [S-2:119 *vs.* 142], [S-3:242 *vs.* 1751],and [S-4:222 *vs.* 2675]. In the case of S-3 and S-4 which harboured large clones, the clone sizes measured by shear sites were significantly different from those of the tag system.

Results and discussion: "Analysis of Clinical samples"

A method to enable accurate quantification of clonality is in the center of this project

In this study, I aimed to connect the clonal composition of HTLV-1 infected cells with the diagnosis, prediction of prognosis, and elucidation of the mechanism underlying the multistep leukemogenesis of ATL. Similar to the project of JSPFAD on PVL, I have planned to study the clonality patterns of HTLV-1 infected individuals by taking advantage of my new methodology to accurately quantify the clonality of HTLV-1 infected cells. In the following sections, I present the results of my pilot data.

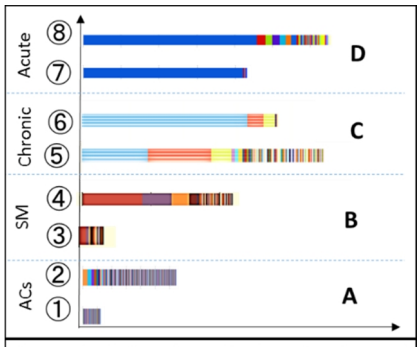


Figure 4. A) Acs showed Polyclonal patterns of infected cells. B)Smoldering patients showed polyclonal or Oligoclonal expansions C) Chronic samples showed Monoclonal patterns or largely expanded Oligoclonal patterns (major clones harbored large numbers of infected cells).D) Acute samples showed Monoclonal patterns.

Although the number of analyzed samples is limited, our initial data suggested different clonality patterns specific to individuals who were AC and those with the different subtypes of ATL (Figure 4). I analyzed samples from ACs, patients with the indolent types of ATL [smoldering (SM) and chronic] and patients with aggressive ATL (acute). Despite similar PVLs, AC vs. SM could be distinguished using clonality patterns (polyclonal vs. a shift towards oligoclonal). The clones of ACs showed a uniform distribution pattern with no large difference in clone size; however, clones of SM types had non-uniform sizes (Figure 4A-B). Chronic subtypes showed expanded oligoclonal patterns with a large shift to monoclonality (Figure 4C). All of the samples from patients with acute ATL harbored a largely expanded clone with a high absolute number of infected cells (Figure 4D). The clonality pattern of the chronic samples was more similar to the acute than the smoldering types (Figure 4C-D).

Due to diverse clinical manifestations and varying prognosis, ATL patients are categorized into distinct subtypes, based on standard clinical criteria: presence of organ involvement, leukemic manifestation, and levels of lactate dehydrogenase (LDH) and calcium. Currently, in clinical practice distinct treatment

strategies are used for the different subtypes of ATL [Tsukasaki *et al.* JCO 2009]. Therefore, classifying ATL patients into distinct subgroups is of high importance for selecting appropriate therapeutic interventions [Tsukasaki *et al.* Hematology 2013]. Considering the intimate link between ATL diagnosis and treatment, a more robust classification of ATL subtypes mediated by HTLV-1 clonal composition would be of fundamental clinical significance. Further examination of clonality patterns with a greater numbers of samples is necessary to validate the relationship between clonality patters and ATL subtypes, and to apply these patterns to diagnosis.

Impact on Prognosis and Prevention: An immediate demand for an effective prognostic indicator of ATL onset In a pilot study, I obtained data from four SM patients over 4 years. Two of the samples showed no progression in disease status (T1 = SM, T2 = SM); the other samples had progression into the chronic stage over the time course (T1 = SM, T2 = Chronic) (Figure 5). I detected a significant difference between the clonality patterns of the two sample sets independent of their PVL. Non-progressed samples manifested a polyclonal or oligoclonal expansion with a low number of infected cells (Figure 5B). However the progressed samples manifested monoclonal or largely expanded oligoclonal patterns (Figure 5A). Moreover, when I analyzed a particular individual over a time course of 6 years who progressed from AC to acute ATL (T1: AC, T2: AC, T3: Chronic, and T4: Acute), the major clone of the T4: acute state (showed with asterisk mark) was found to be dominant in earlier time points. This suggests the potential connection between clone size and the fate of the clone (Figure 6). In addition, I examined the effect of therapy on clonality patterns of patients before and after treatment. I could detect both stable and fluctuating clones from these samples. Most of the samples harbored a stable major clone before and after relapse (Figure 7A). However, in one patient I did find changes in size and order of the clones before and after treatment (Figure 7B).

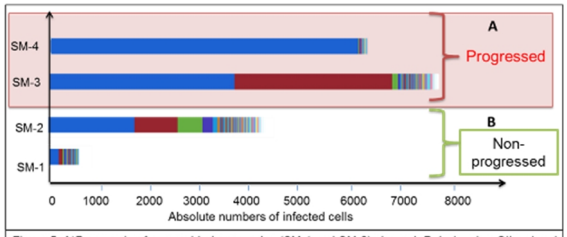


Figure 5. A)Progression free smoldering samples (SM-1 and SM-2) showed Polyclonal or Oligoclonal patterns (with small numbers of infected cells). B) Progressed samples (SM-3 and SM-4) showed largely expanded Oligoclonal and Monoclonal patterns, respectively.

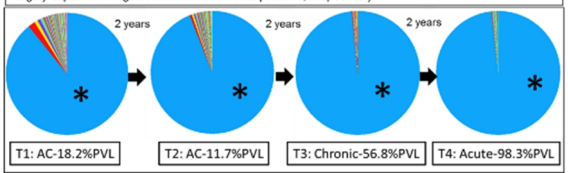


Figure 6. Tracking the clonality of a particular infected individual over a time course of 6 years. The clone marked by asterisks was integrated at chr2-74467952 position, and maintained over time.

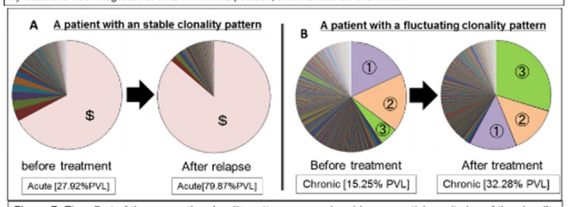


Figure 7. The effect of therapy on the clonality pattern was analyzed by sequential monitoring of the clonality patterns before and after treatment. A) The clone marked by $ was stable and dominant before and after therapy. B) Clones ①②③ were undergo change in size and order following treatment. Clone① and Clone ③ were dominant before and after therapy, respectively.

Although still preliminary, the data suggests that clonality patterns can be of prognostic use to patients. To pursue this, large-scale expansion of the project is recommended. This analysis should be helpful for decision making or developing timely and appropriate therapeutic intervention, based on clonality status of patients. 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 are a risk indicator of progression into ATL. The comparison of clonality patterns in individuals who progress from AC to development of ATL is expected to provide critical information on the clonality alterations that are associated with the transition from the AC to ATL state. Using this information as a prognostic indicator appears to be beneficial for the early detection of ATL onset, and eventually, ATL prevention. Accurate monitoring of the clonality patterns among infected individuals may help us to differentiate progressive and non-progressive patterns as well as assessments of the risk of disease development.

<Conclusion> In this next generation study on ATL risk factors, I have developed an original methodology to accurately monitor clonal composition of HTLV-1-infected cells. Our pilot data is promising and suggests possible applications of this methodology in enabling the molecular-based diagnosis of ATL, as well as predicting ATL development among HTLV-1-infected individuals. For this purpose, a cohort study to evaluate the clonal composition of infected cells is currently in progress. In summary overtime monitoring of clinical samples suggested the importance of our method for generating biologically meaningful information.

Introduction

Modern medicine has done much to eradicate and cure diseases. However, it has failed in certain areas, such as many types of cancer, and cancer remains one of the top 10 incurable diseases. A neoplasm, or tumor, is a cell population that has undergone unregulated cell growth. Most neoplasms are composed of clonally expanded cell populations [1-5]. 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. Using different approaches starting from early cytogenetic analysis of chromosomal abnormalities to later, more elaborate studies of mutation patterns with next generation sequencing (NGS) technologies, the clonal composition of different tumors has been analyzed [3, 6-9]. However, approaches like artificial labeling methods or the analysis of naturally occurring cytogenetic or genetic abnormalities are limited to analysis of progressing or already developed tumors. Therefore, comprehensive analysis of tumor development and its clonal evolution from early initiating events to the finally developed tumor is restricted

Among the different types of cancers, adult T-cell leukemia (ATL) is a remarkably unique neoplasm; because, Human T-cell Leukemia Virus type-1 (HTLV-1) is the direct cause of ATL, and HTLV-1 infection and integration of the provirus is intrinsic and unavoidable for ATL development [10-14]. HTLV-1 infection mainly occurs via breast- feeding among ATL patients [15]. Therefore, unlike most other malignancies, initiation of ATL can be traced back to the same initial time and event. Moreover, HTLV-1 is a retrovirus that integrates into the Human genome and persists in vivo during mitosis of host cells for a long latency period [12, 16]. Since infection is chiefly leads to a single integration per host cell; thus, each single infected cell can be uniquely characterized based on its integration site [17]. Using this feature as a clue provides us a strong advantage to track the changes and events from early infection to the final stage of malignancy (Figure 1). This makes ATL an appropriate and ideal model to study clonal composition and the dynamics of tumor development.

In this thesis, I used ATL as a model system to investigate the clonal composition of tumor development. The HTLV-1 integration site itself is a unique mark that clearly discriminates each single cell in the complex population of tumor cells. To elucidate clonal composition and to define clones based on integration sites, accurate isolation of as many integration sites and accurate measurement of infected cell numbers with the same integration site (clones) is necessary(Figure 1, 2).Previous studies have not been able to use a method to accurately analyze clonal composition[10, 12, 18-28].

Here I present my unique strategy that enables accurate monitoring of HTLV-1 infected cells[29]. In the first section, I will describe my original methodology to define and characterize clones based on provirus integration sites, and describe the biological data obtained by analyzing rare samples from HTLV-1 infected individuals and different subtypes of ATL. In the second section, I include the information on how to reveal intra-clonal composition of HTLV-1 infected clones by creating a link between integration sites and the genomic abnormality associated with each specific clone. The work for this section is still in progress; however, I have included my preliminary data (Figure 2).

I believe that robust monitoring and tracking of clonal dynamics using provirus integration sites and linking this information to the mutation profile of clones will allow for the assessment of clonal composition of HTLV-1-infected individuals from early infection to the final stage of ATL development. Moreover, such analysis not only enables clarification of the mechanisms underlying the multistep leukemogenic events of ATL development but also provides information that can be used as a model for studying the role of clonality in tumor development in general. (Figure 1-2).

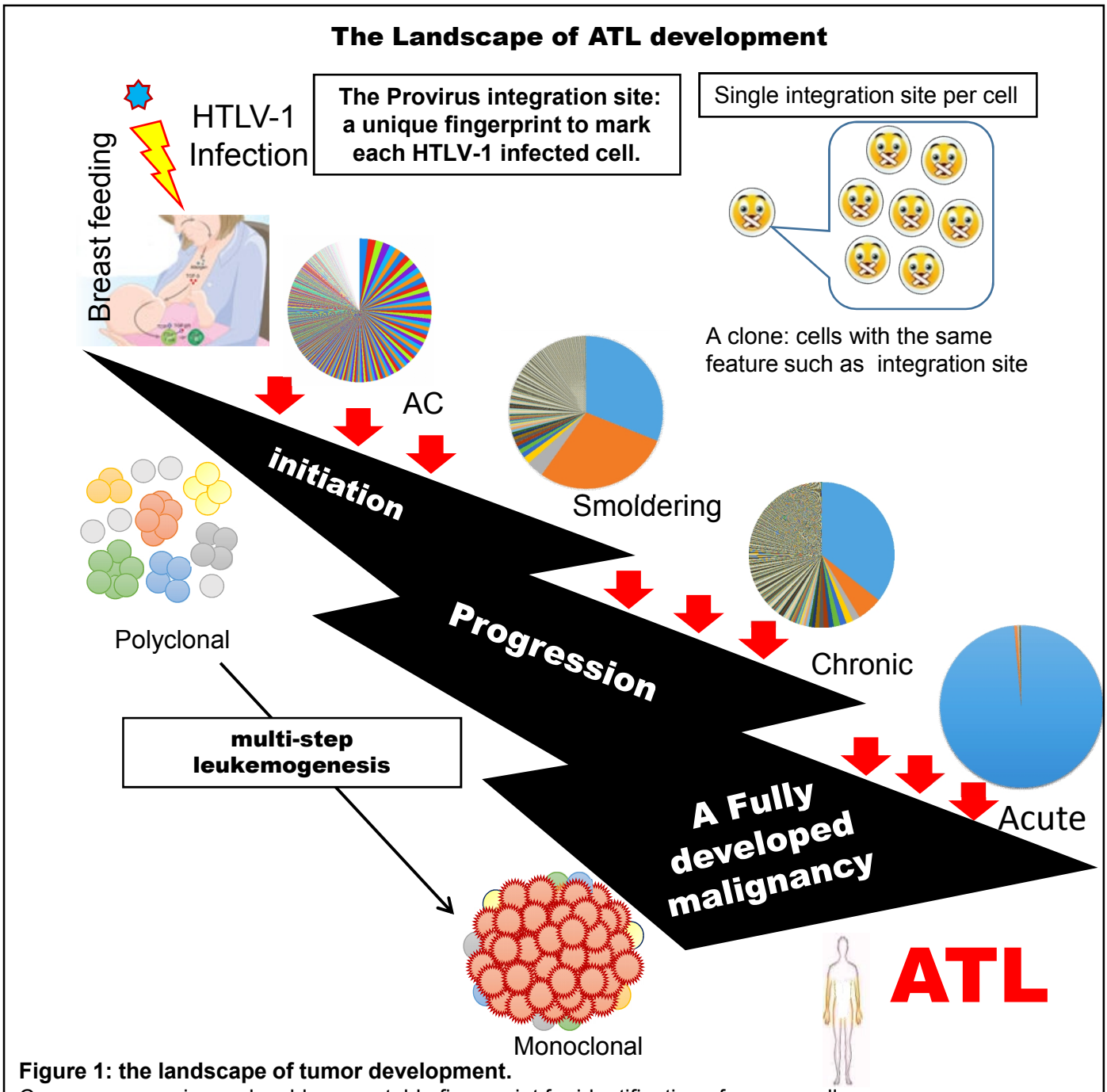
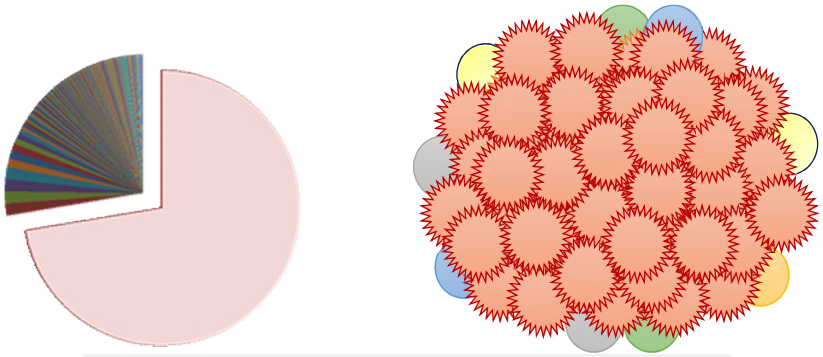


Figure 1: the landscape of tumor development.
Cancer genome is employable as a stable fingerprint for identification of cancer cells.
In most types of cancers, genomic marks such as chromosomal aberrations or single-nucleotide mutations can be used as genetic fingerprints of cancer cells. Human T-cell Leukemia Virus type-1 (HTLV-1) is the direct cause of Adult T-cell Leukemia. ATL is a malignancy that HTLV-1 infection and integration of provirus is intrinsic for its development. HTLV-1 infection mainly occurs via breast feeding among ATL patients. Therefore, unlike other malignancies initiation of ATL can be traced back to the same event. Moreover, HTLV-1 as a retrovirus integrates into the human genome and persist in vivo during mitosis of host cells for a long latency period. Since infection is chiefly a single integration per host cell, each single infected cell can be uniquely characterized based on its integration sites. Using this feature as a clue gives us a strong advantage to track the changes and events from early infection to the final stage of malignancy . Cancer development is known a Multistep evolutionary process during which a normal cell undergo change and abnormal growth and turn into cancer cells. In different kind of cancers genetic fingerprints have been used to monitor clonal expansion of cancer cells. Mainly this monitoring are limited to progression and fully developed malignancy. Because Initiation steps are difficult to be monitored. However, in the case of ATL. Because HTLV-1 infection through breast feeding is the early initiative event, the provirus integration sites can be used as a unique fingerprint to monitor each HTLV-1 infected cells. This makes ATL an appropriate model to monitor clonal expansion from early infection to the final stage of malignancy.

A Photo abstract of the present thesis

Current project

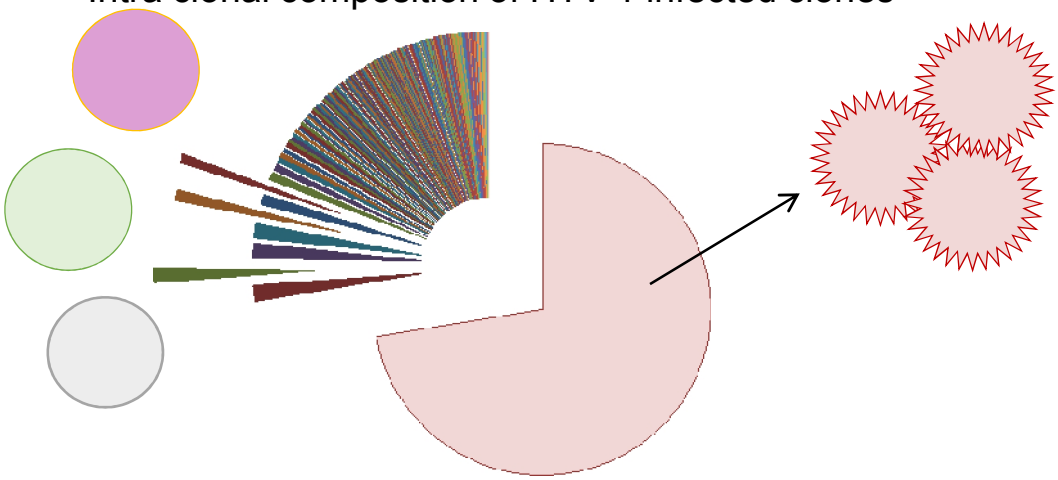
Clonality &
integration site analysis



Number of integration sites
Positions of integration sites
Accurate measurement of the clone size
(numbers of infected cells in each clone)

Future directions of this project (ongoing)

Intra clonal composition of HTV-1 infected clones



- A link between clonality data

based on HTLV-1 integration sites and mutation profiling of each clone.

- Genomic abnormalities associated with each specific clone.
- Clonal evolution through multistep leukemogenesis of ATL

Figure 2: Outlines of the present thesis.
(A)The main part of this thesis has been focused on defining HTLV-1 infected clones based on integration sites of the provirus. I have developed an original methodology which provides information on numbers of integration sites, positions of integration sites, and most importantly accurate numbers of infected cells in each clone (the clone size).
(B) My final ideal goal is to comprehensively demonstrate clonal composition of ATL from different perspectives. In the second part of this thesis I have discussed my plan to monitor clonality based on mutation profiling of each specific clone. For this purpose I will make a link between the data of integration sites-based clonality and mutation profiling- based clonality. I discussed my plan in the present thesis.

General Background

It has been >30 years since HTLV-1 was shown to be the causative agent of ATL [13, 30]. At the current time, although much research has been conducted on ATL and other HTLV-1-associated diseases like HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) and a vast amount of knowledge has been generated, we still cannot cure ATL [31, 32]. We do not even clearly know “what HTLV-1 is”, and “what ATL is”. Thus, it is important to review what we know about ATL and HTLV-1 to recognize the advantages and drawbacks of our previous strategies. We need to understand our advantages and try to increase them, as well as try to overcome our drawbacks. We also need to be open to new alternative strategies that may open new doors of knowledge about ATL. Using this philosophy, my study is empowered with the hope of making a difference and my final goal is to contribute to the field to be one step closer to the treatment of ATL and other HTLV-1-associated diseases.

Following HTLV-1 infection, 95% of HTLV-1-infected individuals show a polyclonal pattern of infected cells and remain as healthy asymptomatic carriers (ACs) throughout their life. However, approximately 5% of infected individuals change to a monoclonal pattern of infected cells and develop ATL after a long latency period [18, 19, 28, 29, 33, 34]. The factors that determine ATL development remain to be elucidated (Figure 3).

To discover the factors associated with ATL development, the Joint Study on Predisposing Factors of ATL Development (JSPFAD), a nationwide collaborative group, has been extensively investigating the proviral loads (PVLs) of infected individuals. In the first generation of studies on ATL risk factors, JSPFAD demonstrated that an elevated PVL is the best characterized risk factor associated with ATL development [35]. PVL represents the burden of HTLV-1 infection, i.e., the percentage of infected cells among the total peripheral blood mononuclear cells (PBMCs), which is accurately measurable by qPCR [36-38]. A PVL >4% is a strong indication of risk for progression to ATL. However, PVL alone cannot predict development of the disease [35], and it remains to be elucidated how the threshold of 4% and the overall high levels of PVL are maintained and how they contribute to ATL onset. A PVL <4% includes healthy infected individuals; thus, has been considered a “safety zone”; a PVL >4% includes both healthy individuals and ATL patients, and therefore has been considered as uncertainty zone (Figure4). Discovering further ATL risk factors and the underlying mechanisms of ATL development requires a comprehensive characterization of this PVL “uncertainty zone” (Figure 4).

Therefore, in the next generation study on ATL risk factors, I focused on the contents of PVLs that is the clonal composition of infected cells. Monitoring the clonal composition of infected cells is only achieved by determining the integration sites of the provirus and accurately measuring the size of every clone (Figure 4, 5).

I put my effort into developing a method to comprehensively monitor the clonal composition of HTLV-1 infected cells with a high level of accuracy. Investigating the clonality of HTLV-1 infected cells has previously been attempted using different conventional methods: Southern blot, inverse PCR and ligation-mediated PCR, and NGS technology [18, 26, 39-42]. However, the role of HTLV-1 integration site preference and the clonal composition of infected cells in disease outcome remain to be elucidated.

Asking the question, “What were the drawbacks of previous studies?” our experience led us to recognize that the level of accuracy and accessibility of information needed to be improved.

Subjective experience is true, but it may not be the totality of truth. With each of those methods we just observed a part of reality without being able to achieve a clear image of whole reality about clonal composition of HTLV-1 infected cells. In the following sections, I introduce information obtained from the previous studies, then describe the strategies I employed to analyze the clonal composition of HTLV-1 infected cells in the Results sections (Figure 2).

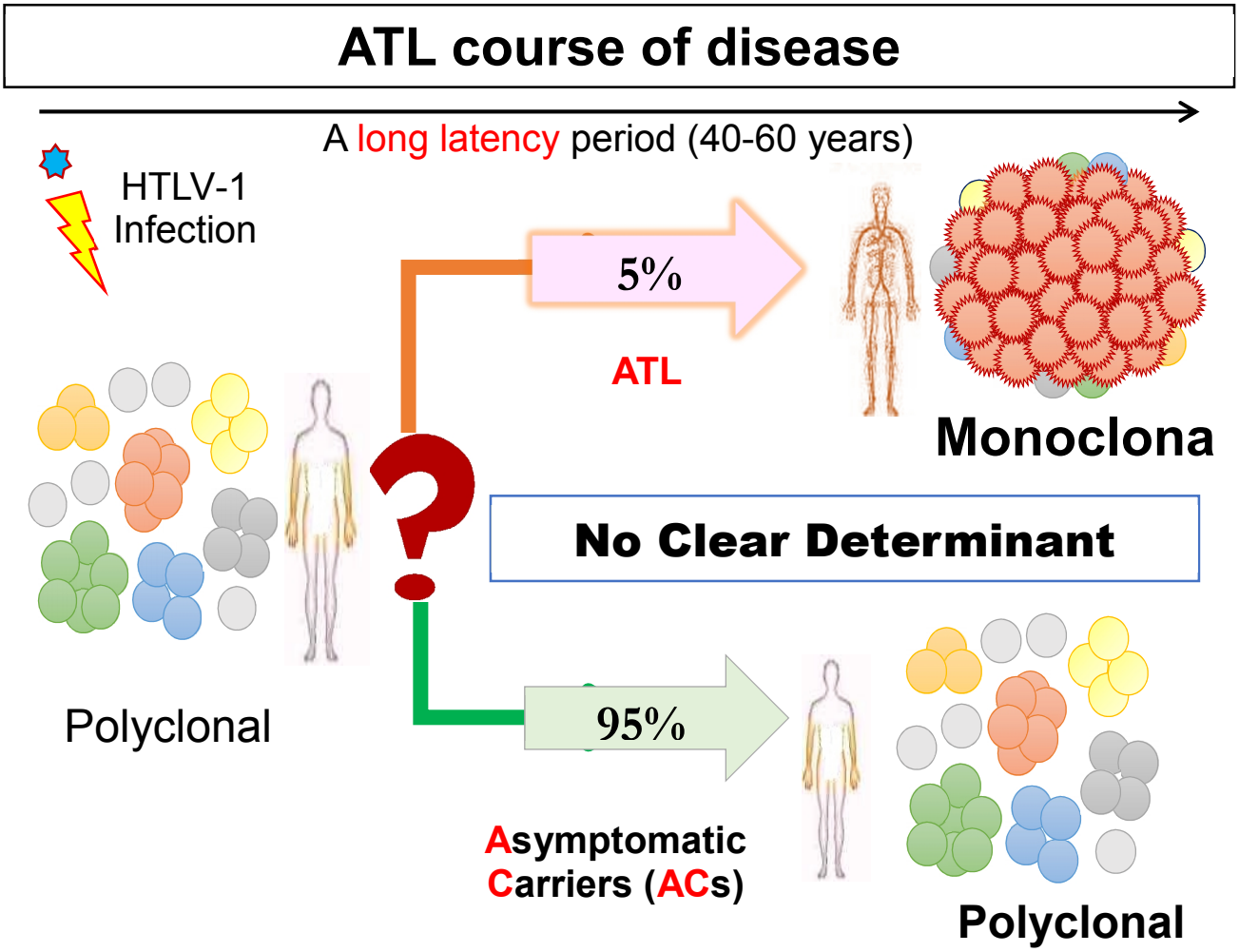


Figure 3: a background on the ATL course of disease.
Following HTLV-1 infection people show polyclonal population of infected cells.
After a long latency period about 95%of infected individuals, keep the initial polyclonal pattern and remain as ACs. But about 5% of them change into a monoclonal pattern and develop ATL. However there is no clear determinant to distinguish between the people who remain as AC and those who develop ATL.

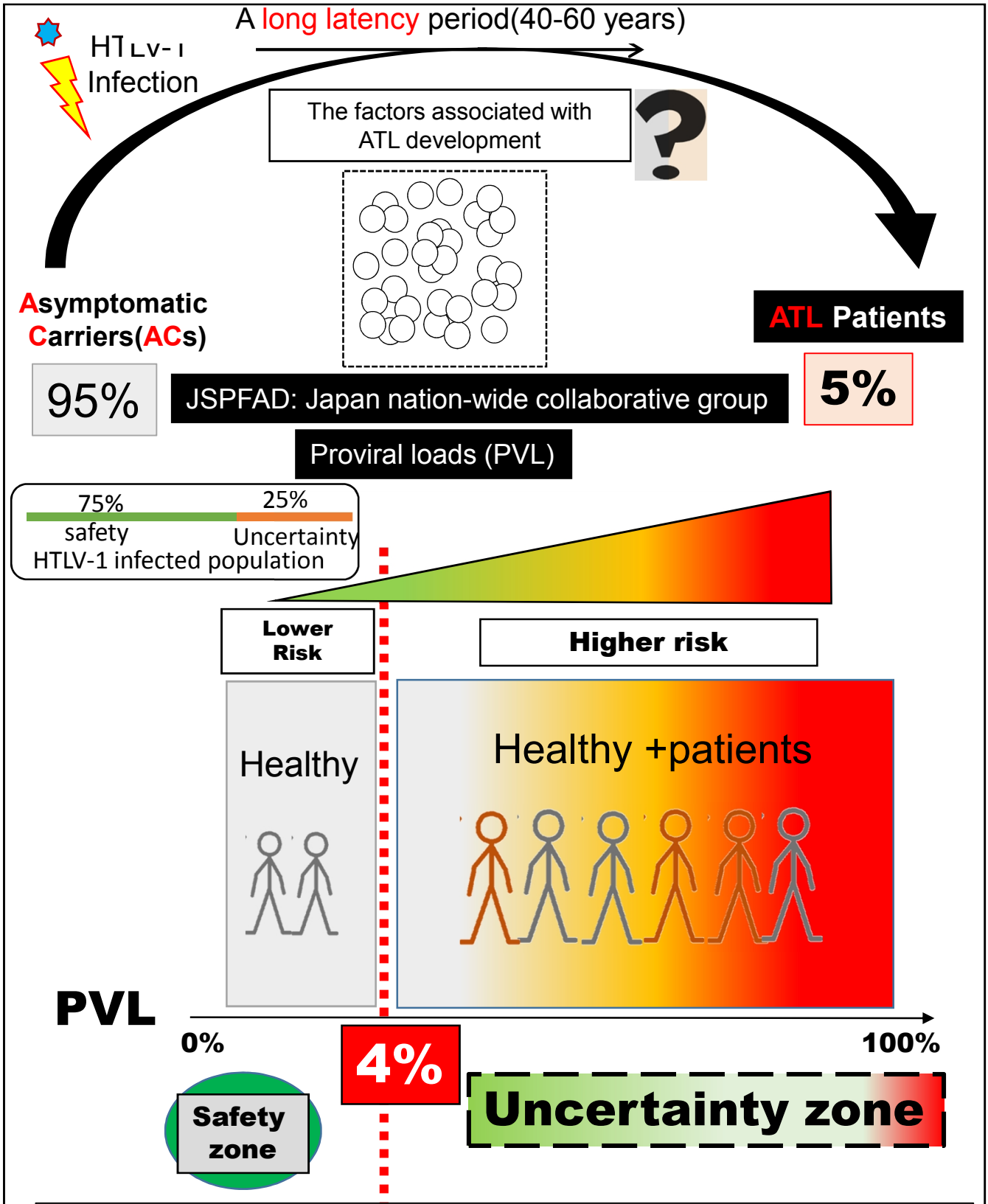


Figure 4: the necessity of determining the factors associated with ATL development
To discover the factors associated with ATL development, JSPFAD, which is a nationwide collaborative group, has been extensively investigating the levels of PVLs. PVL is the percentage of infected cells among total PBMCs, thus represents the total level of infection. Currently a PVL higher than 4% is the best factor associated with the risk of ATL development. Below 4% includes healthy infected individuals, thus considered as the safety zone. Above 4% includes both healthy individuals and ATL patients, so is considered as uncertainty zone. Discovering ATL risk factors requires comprehensive characterization of both zones.

A next generation study on ATL determinants

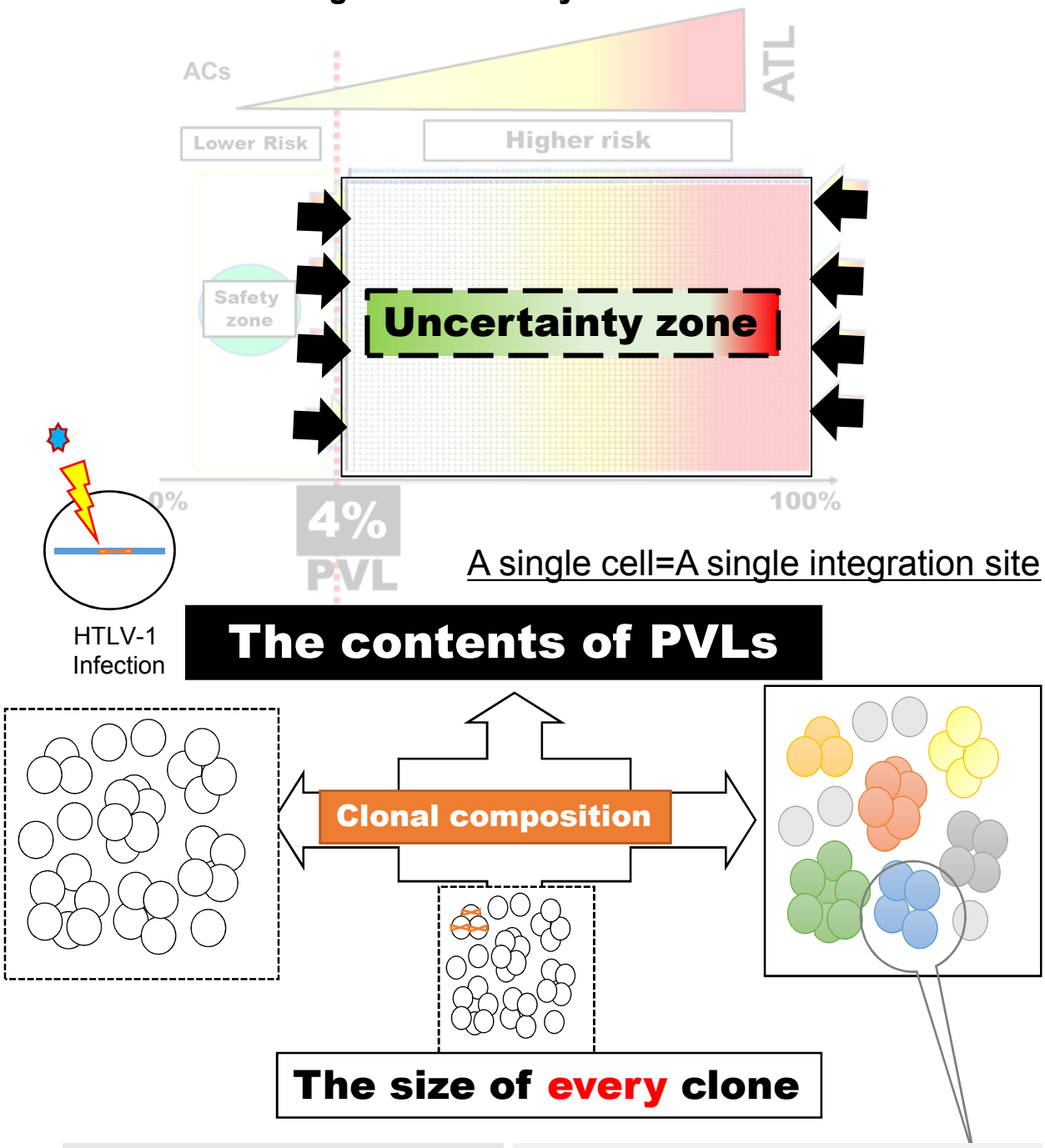


Figure 5: My approach to investigate factors determining ATL development.
As a next generation study on ATL risk factors, I have focused on the contents of PVLs that is the clonal composition of infected cells. Monitoring clonal composition of infected cells is only achievable by determining integration sites of provirus, and accurately measuring the size of every clone.

Background: Clonality of HTLV-1 infected cells by previous studies

Clonal proliferation of HTLV-1-infected cells was first detected as monoclonal-derived bands by southern blotting[39]. Early studies found that monoclonal integration of HTLV-1 is a hallmark of ATL cells [20]. Furthermore, it was suggested that detecting a monoclonal band is useful for diagnosis and is associated with a high risk of ATL development [12, 24]. 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[26, 40-42]. 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 [26, 43]. 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 [42]. This hypothesis was further supported by the detection of a particular HTLV-1 clone in the same carrier over the course of several years [21]. Two Miyazaki cohort studies focused on the maintenance and establishment of clonal expansion: Okayama et al. analyzed the maintenance of a pre-leukemic clone in an AC state several years prior to ATL onset [22], 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 [23].

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 [18, 41]. Furthermore, conventional techniques could not provide adequate information regarding the number of infected cells in each clone (clone size)[18]. To effectively track and monitor HTLV-1 clonal composition and dynamics, I 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 (Figure 6).

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 [44-48]. 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 [18]. 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 [18] (Figure 6, 7). 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 [18] (Figure 7). Therefore, Gillet et al. used a calibration curve to statistically correct the shear site data [18] (Figure 8). Later, Berry et al. introduced a statistical approach, and further addressed the difficulties of estimating clone size from shear site data [49]. Their approach estimates the size

of small clones with little error, but estimates for larger clones have greater error [49] (Figure 8). All estimations of this method mainly is based on relative size of clones. Gillet et al used a parameter which has been generally employed in economy to describe the distribution of the wealth in the society to convert the relative size of clones and general pattern of clonality of each sample into numbers. A parameter adopted from the Gini coefficient [49, 50] and termed the oligoclonality index was used to describe the size and distribution of HTLV-1 clones [18] (Figure 8). They showed that the oligoclonality index differs between ATL and non-malignant HTLV-1 infections (ACs). However they could not discriminate between different subtypes of ATL by relative measurement of the clone size and OCI. In figure 8, I simply depicted what the Gini coefficient is and how to measure it.

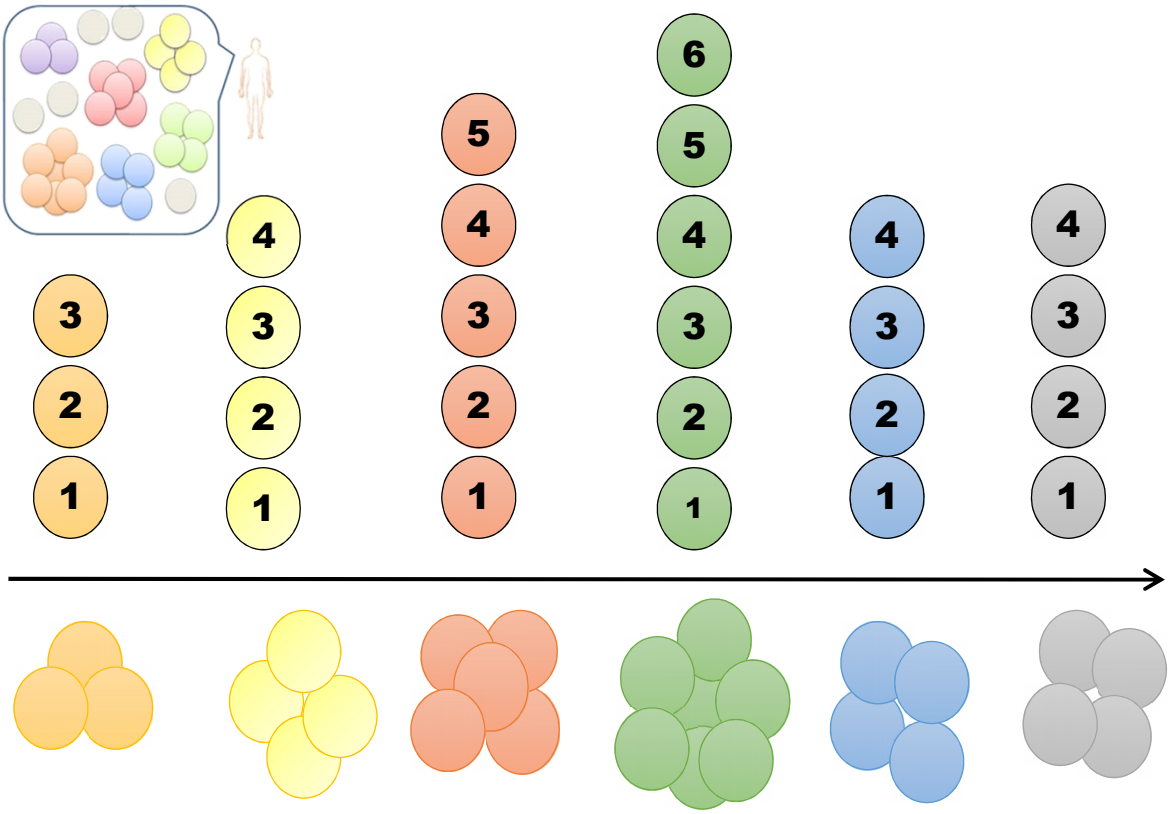
Followings I discuss about the method of Gillet et al. I tried to simplify the concepts to facilitate understanding of this method for the readers of present thesis. To understand the details please study all related sections here.

Here I 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, I 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, I conducted comprehensive internal validation experiments to assess the effectiveness and accuracy of my 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[29]. I present my methodology, which illustrates that employing the tag system is effective for improving quantification of clonal abundance.

How to measure the clone size ?

Accurate measurement of the size of every clone is essential.

Clone size: The number of infected cells in each clone



Two main strategies

Shear site system Gillet *et al.*

Tag system Firouzi *et al.*

My original methodology

Figure 6: what is the clone size and how to measure it.
The clone size is the number of infected cells in each clone.
Accurate measurement of the size of every clone is essential for monitoring clonality.
There are two main strategies for measuring the size of clones.
The shear site system and our original tag system.
In the present thesis I have explained and compared these two systems.

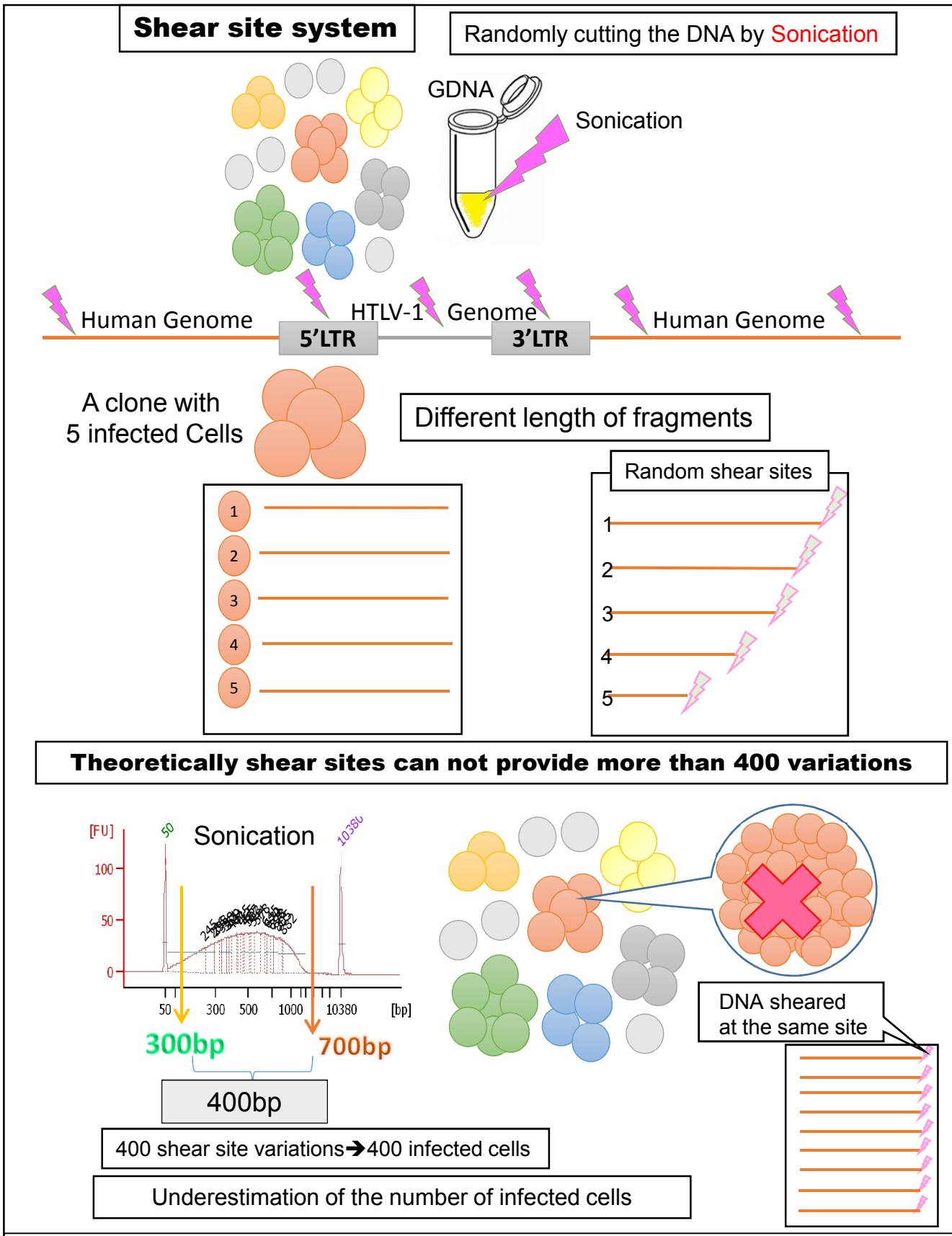


Figure 7: introducing shear site strategy of Gillet et al.
Shear site system is based on randomly cutting DNA by sonication. Having GDNA fragmented by sonication generates random shear sites. Therefore, Different unique lengths of fragments correspond to the original number of infected cells. However considering the lengths is not enough. Because sonication generates a size distribution of 300 to 700 bp, which can theoretically provide only 400 shear site variations. This amount of variation cannot cover more than 400 infected cells in each clone. Therefore as the size of clones increases the probability of DNA shearing at the same site will increase. This leads to underestimation of the number of infected cells.

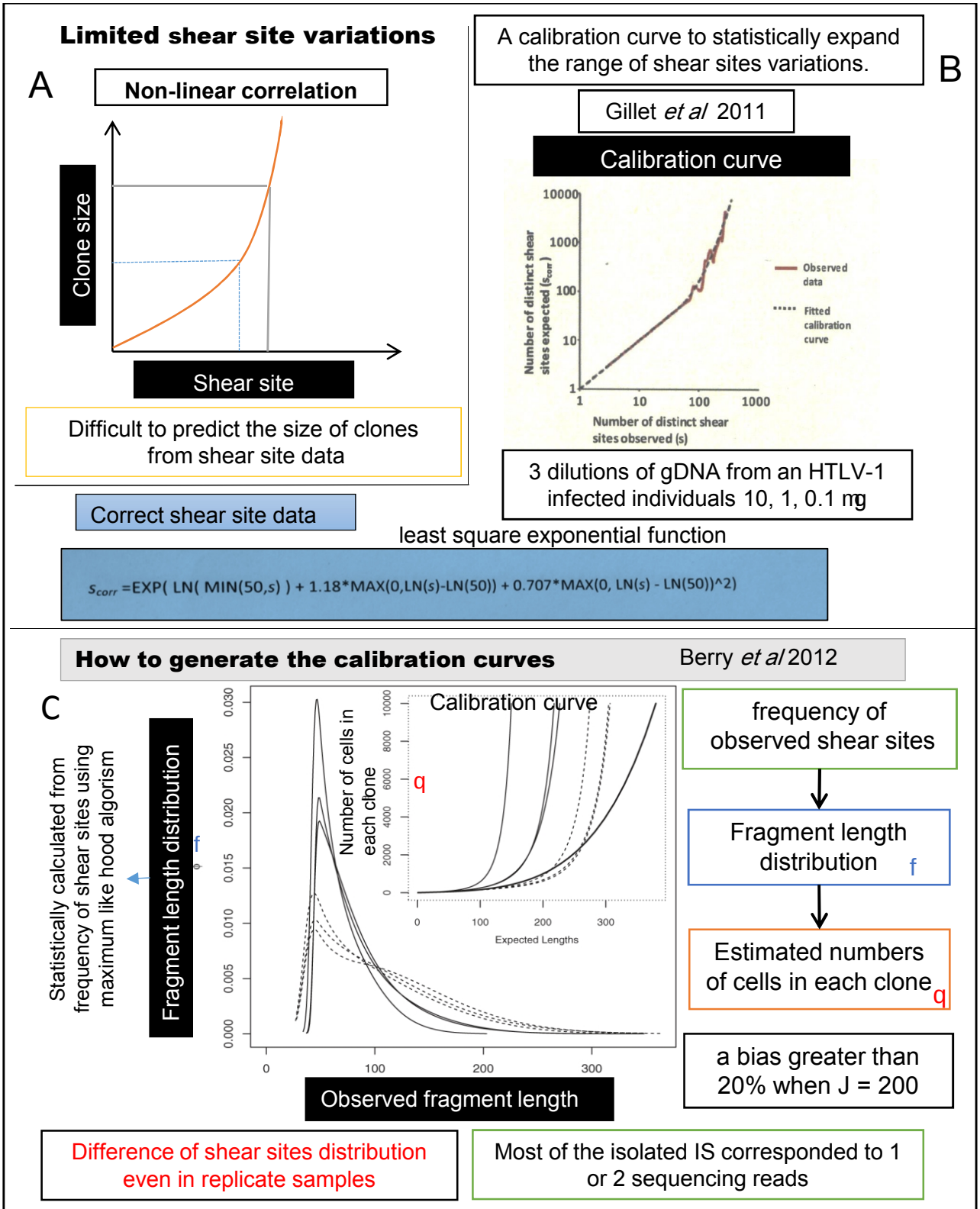


Figure 8: Statistical correction of shear sites data.
(A) Limited shear site variations leads to a non linear correlation between shear sites and clone size.Therefore it gets difficult to predict the size of clones from shear sites data, when the clone size exceeds shear sites variations. (B) Gillet et all introduced a calibration curve to statistically expand the range of shear site variations.They used 3 dilutions of gDNA to generate the graph. From the graph they retrieved this formulation to correct the shear site data. (C) Berry et al described how to generate the calibration curves .They converted Frequency of shear sites to fragment length distribution –the factor phi- by maximum like hood algorism.Then the phi were converted to (tetta) which is the estimated numbers of infected cells in each clone.They reported difference of shear sites distribution even in replicate samples because most of integration sites in their data corresponded to 1 or 2 sequencing reads. This reduce reliability of generated shear data. At least a bias greater that 20% will be introduced in the estimation of clone sizes when the factor J in other words the size of clones exceeds 200 infected cells.

Figure 8

D

The Gini coefficient

A relative measurement of inequality

In economics: the distribution of income in a society

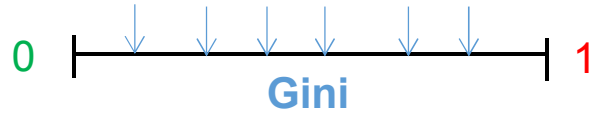
How evenly the income (or wealth) is distributed throughout a country.

How evenly the income is distributed ?

Society	Clonality
People	Integration sites
Relative Income	Relative clone size

Polyclonality **Perfect equality** 0-1 **Inequality** **Monoclonality**

Some of individuals have more income than others



Graphical representation of Gini coefficient

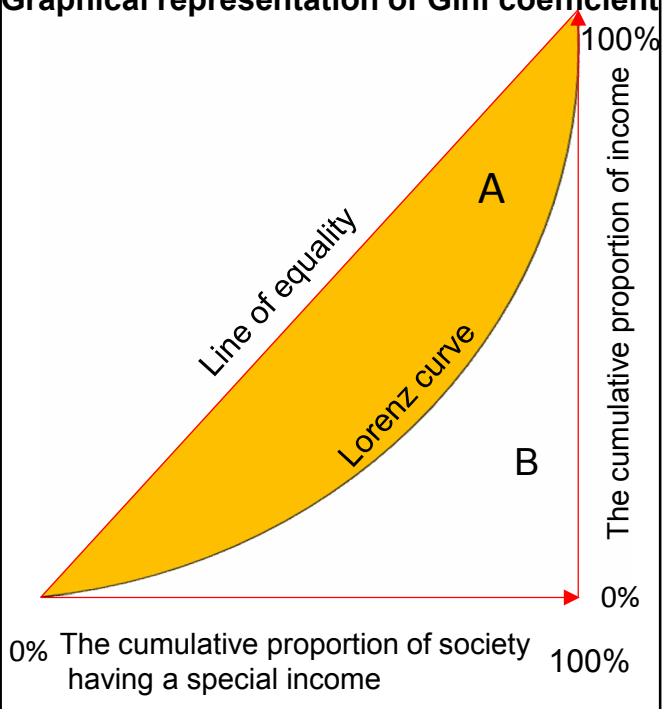


Figure 8 (D) Gini coefficient is a relative measurement of inequality. It has been mainly used in economics to show distribution of income in a society. In other words to show how evenly the income (or wealth) is distributed throughout a country. Gini coefficient ranges from zero to one. In the case of perfect equal distribution of income Gini is zero but when only one individual have all the income of society Gini value is one. If the society was clonality status of each sample, people correspond to integration sites and relative income corresponds to relative size of clones. Therefore it is expected that Perfect equality represents perfect polyclonality and inequality represents perfect monoclonality. This is the graphic representation of Gini coefficient. X axis is The cumulative proportion of society having a special income. Y axis is the cumulative proportion of income. The area between Lorenz curve and equality line is the value of Gini.

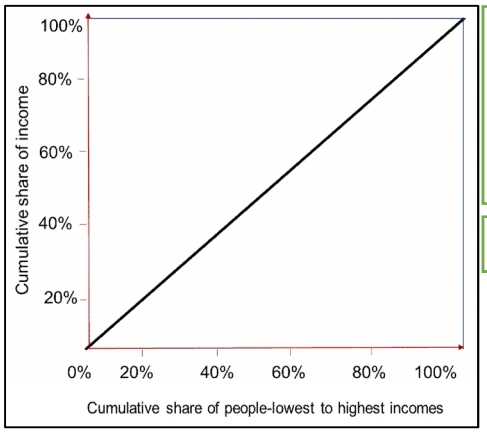
Cumulative proportion of Integration sites

Cumulative proportion of clone size

Person	Proportion of population (%)	Cumulative proportion of population (%)	Income (%)	Cumulative income (%)
A	10%	10%	10%	10%
B	10%	20%	10%	20%
C	10%	30%	10%	30%
D	10%	40%	10%	40%
E	10%	50%	10%	50%
F	10%	60%	10%	60%
G	10%	70%	10%	70%
H	10%	80%	10%	80%
I	10%	90%	10%	90%
J	10%	100%	10%	100%

Cumulative proportion of population (on the horizontal axis) against the **cumulative percentage of income** (on the vertical axis)

Graphic representation of a society with an equal distribution of income



An infected individual with an equal distribution of clones

Gini coefficient =0

Perfect equality

Figure 8 (E):
This is an example of a society in which. All people have the same income. Cumulative proportion of population corresponds to Cumulative proportion of Integration sites. Cumulative income corresponds to Cumulative proportion of clone size. To calculate Gini coefficient we should make a graph from cumulative proportion of population against cumulative percentage of income .

Results-section-1-A

General concepts

I originally designed our method to overcome the limitations of conventional techniques [26, 41] and the only existing high-throughput method [18]. In general, our method includes two main sets of wet experiments and an *in-silico* analysis. I used genomic DNA (gDNA) 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 (Figure 9).

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 [17]. 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*. [18], employing high-sensitivity deep sequencing allowed for the isolation of a large number of unique integration sites (UISs), including samples with low PVLs (Figure 10). I 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 10).

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 [44-47]. 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, I 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, I compare and discuss two main strategies, namely (1) shear sites and (2) a tag system, which enable DNA fragments to be uniquely marked (Figure 6).

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 [18]. 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 [18, 49] (Figure 7, 8).To directly assess the effectiveness of the shear site strategy, I 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 10A and 10B). Following library construction, however, the final product represented smaller size ranges, implying a relatively limited number of variations (Figure 10C). Finally, the number of PCR

amplicons with unique shear sites was retrieved from deep-sequencing data. See Figure 36 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 the next part) [18, 49]. For clarity, only the information relating to the major clone of each sample is provided in Figure 10D. 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 10). 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

I developed an alternate strategy to remove PCR bias and to estimate starting DNA.I 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 11).

I 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 11D). Similar variations of tags and shear sites were observed in the largest clones of S-1 and S-2 ((shear sites *vs.* tags): (209 *vs.* 393) and (119 *vs.* 142)) (Figure 10D and Figure 11D). 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 Table 1 and Table 2 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 Figure 12 for a simple comparison of shear site and tag variations.

Estimating clone size by shear sites vs. tags

Berry *et al.* estimated clone size using shear site data [49]. Considering that there is a nonlinear correlation between shear site variation (fragment lengths) and clone size, they proposed a statistical approach for correcting shear site data. They introduced J, ϕ, and θ as factors referring to possible fragment length, fragment length distribution, and expected number of parent fragments, respectively. J was determined empirically whereas ϕ and θ were statistically calculated. Because I did not use the same experimental setup or statistical estimations, I could not directly compare our data with those of Berry *et al*. . Alternatively, the generated variation in shear sites (factor J) might be indirectly compared. Berry *et al.* showed that the relative abundance of small clones was estimated with little bias, but estimating the size of large clones was problematic. For example, they observed a bias greater than 20% when J = 200 and θ was more than 1000. Consistent with their data, I observed underestimations in samples of which the clone sizes exceeded the generable shear site variations. I showed that these underestimations could be overcome by a large variety of tags. In the case of clinical samples (S-1, S-2, S-3, and S-4), maximum generated shear site variations were [S-3: 242] whereas maximum tag variations were [S-4: 2675].See figure 7 and 8.

Oligoclonality index: shear sites vs. tag system (Figure 13)

Gillet *et al*. introduced the oligoclonality index (OCI) as a parameter to describe HTLV-1 clonal distribution [51]. OCI was adapted from the Gini coefficient, which has been mainly used in economics to measure income inequality [52]. The ratio of the area between the line of equality (the 45-degree diagonal line) and the Lorenz curve graphically represents the Gini coefficient. Here, I calculated the Gini coefficient using Statsdirect software (http://www.statsdirect.com/), and similar to Gillet *et al*. termed it OCI. Our analysis is based on the data for shear sites, tags, and combinations without any statistical manipulation.

Generally, OCI values range from zero to one. A low OCI indicates a more equal distribution of clones; for example, an OCI of zero represents clones of equal sizes of uniform distribution (perfect equality). A higher OCI indicates a more unequal clone size distribution. An OCI of one represents perfect monoclonality (Figure 13).

The OCI of the samples S-1, S-2, S-3, and S-4 were 0.54, 0.67, 0.68, 0.63, respectively, for shear sites and 0.60, 0.67, 0.84, and 0.80 for combinations. Referring to section of Results and discussion, the clone sizes of S-1 and S-2 measured based on shear site data were similar to those of tags and combinations. Consistent with those data, the OCIs of S-1 and S-2 were similar (shear sites vs. tags vs. combinations: [S-1:0.54 vs. 0.62 vs. 0.60], [S-2: 0.67 vs. 0.68 vs. 0.67]). In the case of S-3 and S-4, however, because the clone size was underestimated by shear sites, the OCI calculated based on shear site data differed from that of tags and combinations (shear sites vs. tags vs. combinations: [S-3:0.67 vs. 0.87 vs. 0.84], [S-4: 0.63 vs. 0.88 vs. 0.80]). Although these samples were categorized based on accurately measured clone sizes, they could not be clearly discriminated based on their OCI. In reference to the limitations of the Gini coefficient addressed in economics, because the Gini coefficient is a relative measure, countries may have identical Gini coefficients even with different income distributions [53]. This problem makes interpretation of the Gini coefficient (and thus OCI) controversial. Therefore, S-3 and S-4, even with different sizes and distributions of clones, had a similar OCI (0.84 vs. 0.80). These data suggest that accurately measured clone sizes are more desirable than OCI for discriminating ATL subtypes.

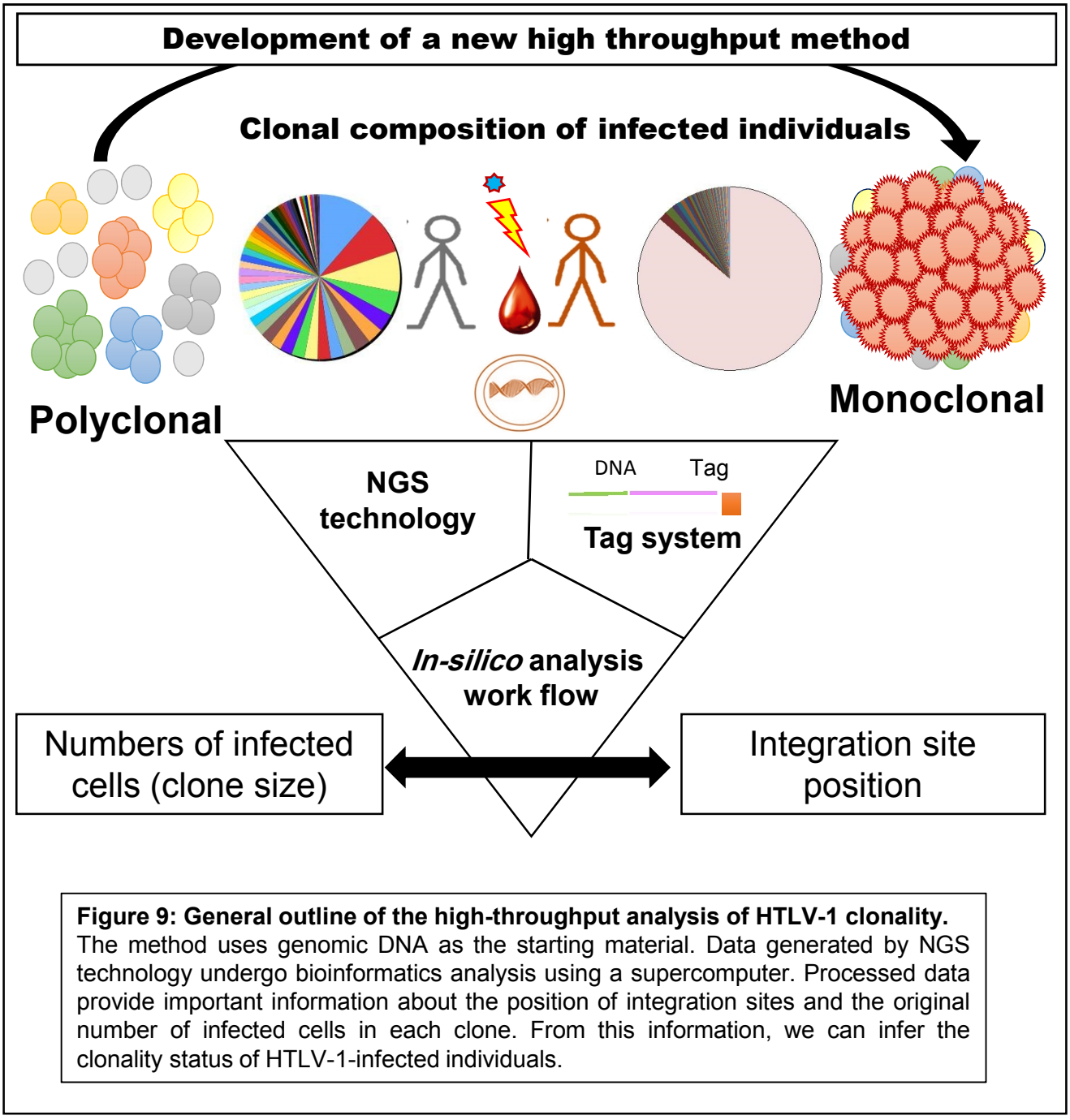


Figure 9: General outline of the high-throughput analysis of HTLV-1 clonality. The method uses genomic DNA as the starting material. Data generated by NGS technology undergo bioinformatics analysis using a supercomputer. Processed data provide important information about the position of integration sites and the original number of infected cells in each clone. From this information, we can infer the clonality status of HTLV-1-infected individuals.

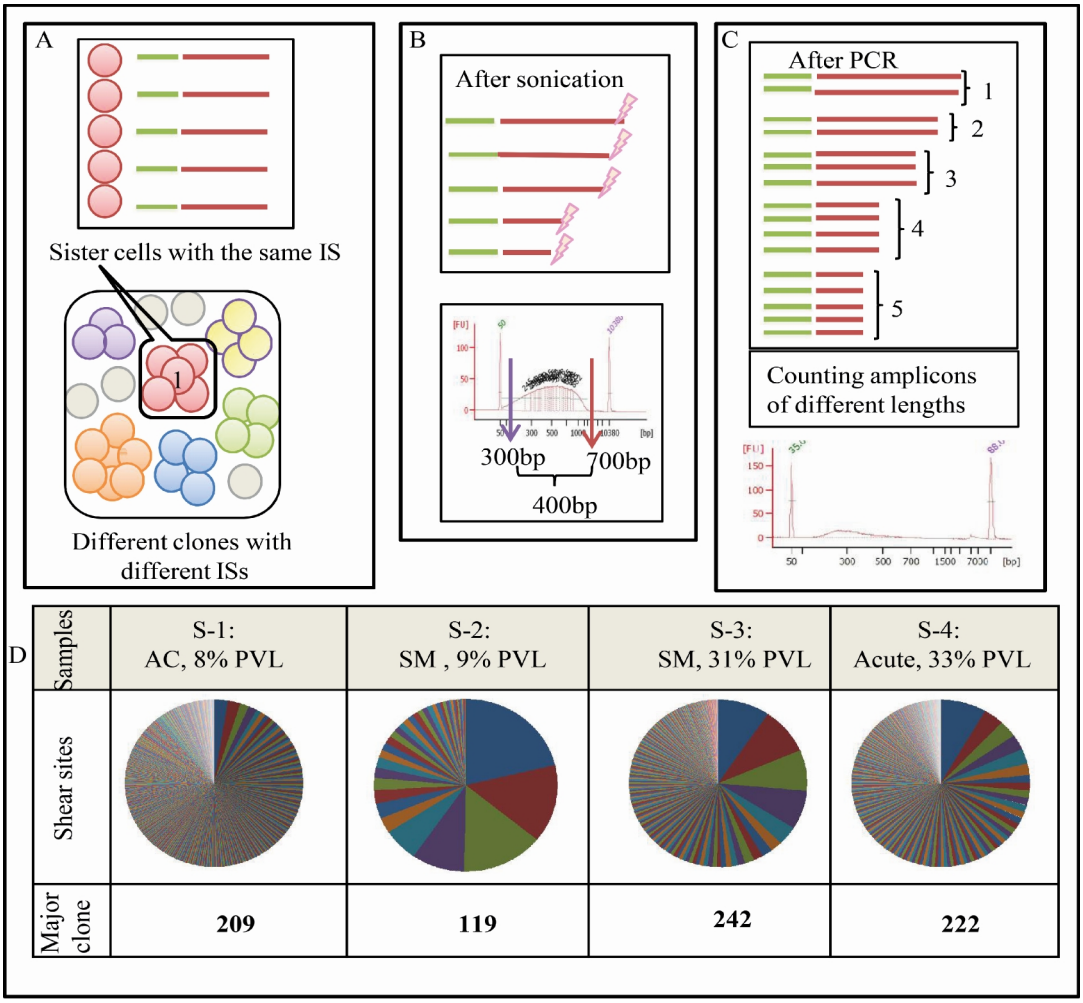
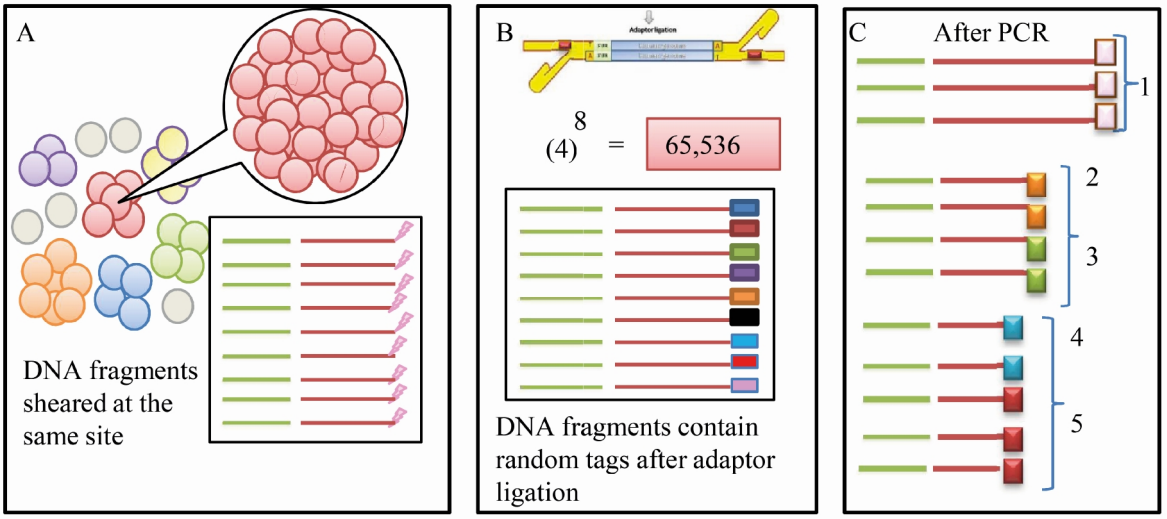


Figure 10: Estimating clone size by 'shear sites'.
(A) Depicted is the complex population of uninfected cells (grey circles) together with infected clones (circles of different colors). A clone is shown as a group of sister cells (circles of the same color) having the same integration site (IS). Different 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 my 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.



D

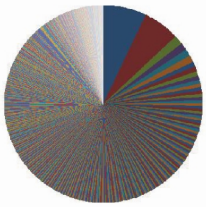
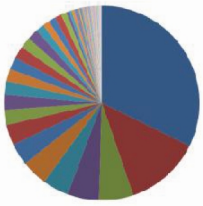
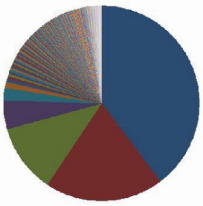
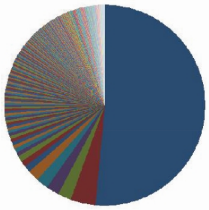
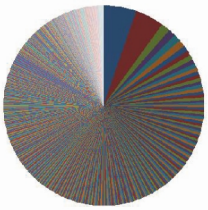
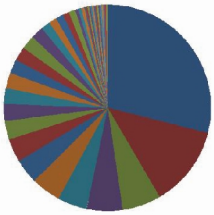
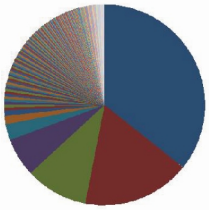
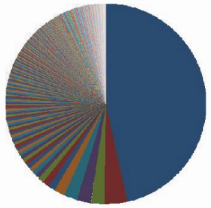
Samples	S-1: AC, 8% PVL	S-2: SM , 9% PVL	S-3: SM, 31% PVL	S-4: Acute, 33% PVL
Tags only				
Major clone	**393**	**142**	**1751**	**2675**
Combinations				
Major clone	**269**	**119**	**1192**	**2038**

Figure 11. 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.

Table 1. The top 10 clones isolated from sample S-1, S-2, S-3, and S-4.

		Chromosome	Strand	Position	Tags	Relative size (%)	Combinations	Relative size (%)	Shear sites*	Relative size (%)
Top 10 clones of S-1	1	chr11	+	41829319	393	7.08	269	5.30	209	2.60
	2	chr11	-	37042565	329	5.93	235	4.63	130	1.62
	3	chr7	-	121751243	83	1.50	74	1.46	43	0.54
	4	chr13	+	69268469	67	1.21	58	1.14	124	1.54
	5	chr18	-	46701081	65	1.17	58	1.14	30	0.37
	6	chr17	+	18847529	60	1.08	58	1.14	180	2.24
	7	chr15	-	37836845	46	0.83	44	0.87	29	0.36
	8	chr2	-	100184973	44	0.79	42	0.83	22	0.27
	9	chr6	+	10852456	42	0.76	40	0.79	72	0.90
	10	chr8	-	35831701	39	0.70	37	0.73	36	0.45
Top 10 clones of S-2	1	chr15	+	59364370	142	32.27	119	28.95	119	21.25
	2	chr13	-	74706141	55	12.50	52	12.65	11	1.96
	3	chr3	+	28073332	25	5.68	25	6.08	11	1.96
	4	chr21	-	44242161	23	5.23	23	5.60	6	1.07
	5	chr18	-	38428907	19	4.32	19	4.62	1	0.18
	6	chrX	-	107427783	19	4.32	19	4.62	2	0.36
	7	chr13	-	84177236	16	3.64	16	3.89	5	0.89
	8	chr21	+	25834766	15	3.41	15	3.65	7	1.25
	9	chr2	+	234346116	11	2.50	10	2.43	6	1.07
	10	chr7	-	99740574	11	2.50	11	2.68	2	0.36
Top 10 clones of S-3	1	chr4	-	563543	1751	39.76	1192	35.72	242	9.48
	2	chr20	+	58007381	863	19.60	579	17.35	232	9.08
	3	chr5	+	62579369	502	11.40	336	10.07	202	7.91
	4	chr6	+	133958124	210	4.77	207	6.20	191	7.48
	5	chr3	-	126392282	91	2.07	86	2.58	94	3.68
	6	chr3	+	178928610	43	0.98	43	1.29	54	2.11
	7	chr8	+	119096533	27	0.61	27	0.81	43	1.68
	8	chr10	-	111698526	18	0.41	18	0.54	9	0.35
	9	chr13	+	21355493	17	0.39	17	0.51	24	0.94
	10	chr18	+	62126326	16	0.36	14	0.42	10	0.39
Top 10 clones of S-4	1	chrX	-	83705328	2675	51.50	2038	46.54	222	8.35
	2	chr14	+	30655896	209	4.02	160	3.65	87	3.27
	3	chr14	+	49676335	112	2.16	97	2.22	77	2.90
	4	chr6	-	85461536	108	2.08	95	2.17	80	3.01
	5	chr16	-	17339636	102	1.96	97	2.22	98	3.69
	6	chr8	+	96129917	93	1.79	75	1.71	59	2.22
	7	chr1	+	4032445	55	1.06	48	1.10	22	0.83
	8	chr7	+	140001929	50	0.96	49	1.12	40	1.50
	9	chr21	+	35571080	49	0.94	41	0.94	38	1.43
	10	chr1	-	56007274	35	0.67	32	0.73	38	1.43

*Orders and numbers of integration sites in shear sites has been matched to those of tags and combinations.

The reported positions of integration sites can be readily searched by common genome browsers such as Blast of NCBI or Blat of UCSC.

Table 2 Information on integration site positions of top-10 clones for each sample

		Chromosome	Strand	Position	A representative sequence for each corresponding integration site position
Top 10 clones of S-1	1	chr11	+	41829319	AATATCTAGTTAAAGGAGGCTGTGAGAATTAGAAAATATA
	2	chr11	-	37042565	GAAATGCATAGCACTAAATGCTCACAAGAGAAAGCAGGAA
	3	chr7	-	121751243	CATAGTTATAAAAACCACTTTACAATGTTCATCTCATACT
	4	chr13	+	69268469	GGAGACCTTTTATCTTTTCTTTTATAATCACTTAATGGTA
	5	chr18	-	46701081	CTTTCTCACCCTTTCTAGTTAGTAAAATCCTAGGGAATAT
	6	chr17	+	18847529	AATTCATAATTACCAAAACTTTGGTAGATGTCCTTAAGTA
	7	chr15	-	37836845	TTTAGCTTCTACTTATAAGTGAAAACATGCAGTATTTGAT
	8	chr2	-	100184973	ACCCTTGGCTTGGTCCCAGGATCAAATCTCTTTTCAAAAA
	9	chr6	+	10852456	GAATAATATGTTGAAGAGTTTTGGATTTTACCTTTCTTAA
	10	chr8	-	35831701	CTGAGCAACGTATCTTTCTCTTCTTTAAACTTTGAGTTTT
Top 10 clones of S-2	1	chr15	+	59364370	GCTATGACTGATGAAAGTGGTGGTACCTAAAGATTGGGGT
	2	chr13	-	74706141	TGTAGAAGCATAGTGGAAATAGGATGTTGAAGGACAGACG
	3	chr3	+	28073332	AAATCAATGTTCTAAGATATGTACTAAGTGCTAGGGAACA
	4	chr21	-	44242161	CAAGCTCACTGATGTTTTCTTCTGTTTGATCAATTCTGCT
	5	chr18	-	38428907	GATTATTAGTATGATAATTACTTTACCAATCTGGTTGCAG
	6	chrX	-	107427783	CAAAAGAGTCCAAACACTAAAGCAACCTTGCTTTATAACA
	7	chr13	-	84177236	CACAGGTTCTAGGAATTAGTGTGTGGATATCTTTGTGGGG
	8	chr21	+	25834766	ATACTCCTGTTCAGGGAAAAATTTGAGCCGGTTTTCAGCA
	9	chr2	+	234346116	GACTCCGATGGTGGGTACCACACATGCTTATCCTTCTCAT
	10	chr7	-	99740574	AGTTTTCAACCCAGTACAGGGACTGTTACATAGCATCTTC
Top 10 clones of S-3	1	chr4	-	563543	CATTGTTTGT GTACCTATGT ACCAGCCTTTTCAAATGAGG
	2	chr20	+	58007381	TATGTTTCCT TACATTACTT ACTAATAGTA ATAAATAGCA
	3	chr5	+	62579369	GTCTCAGATTCCATGCTCTGAGAAAAGTGTGTATGAATTT
	4	chr6	+	133958124	GGTTTTTTTTTTTTTTTTTTTTTTTTGCCATTGTAGGAT
	5	chr3	-	126392282	GTTAACCTCTTGAATTTTGAGAACTAAGGTTAGATGCCTG
	6	chr3	+	178928610	ACTAGGTAATAAAGTCATGAACTAAACAGACCTTCATTGA
	7	chr8	+	119096533	ATGTCCCAGATGCTACCTCCTGGGATCAACTGCAAGTCGT
	8	chr10	-	111698526	CTCCATGGTCTTCCCTGAACACCTCCTACTGCCCTGCAAC
	9	chr13	+	21355493	AAAAGGTCTAGCCGTTGCAACTCAGTGGCATCCCCATCAC
	10	chr18	+	62126326	CATAATCACTTTAAATGTGATTGGAATAAATTCTCCAGTT
Top 10 clones of S-4	1	chrX	-	83705328	CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAG
	2	chr14	+	30655896	AAAACTAAGTTTCAGCTCACAGTATTAGAGTGGGTTACAT
	3	chr14	+	49676335	GTGACTCAAAACAAAAAACAACACACTTACAGTCTTTTTA
	4	chr6	-	85461536	GAAGTTAACACTGATCTCTAATTAGTAAAGCTGTAGACTC
	5	chr16	-	17339636	CATTGTATCCTTCAGTCACCCATGAGAGATTGGATTTAGG
	6	chr8	+	96129917	ACTAGGCTGTGGACAAAAATGACATATGTCTCTTCCGGGC
	7	chr1	+	4032445	GGTTTCTAAAAGAATAGGTGCAAGTCTGTCATTGTGCTAA
	8	chr7	+	140001929	CACTTTCCCCATTGATGGTTGTGACACTTAAGCCCTCTTG
	9	chr21	+	35571080	CGGTGAGACCCCTGAAATACGAGTCATCCCCACTCCTGAC
	10	chr1	-	56007274	GGACACTTACTGTGAATTAGCTTGCAGGACTGGAAGTTGC

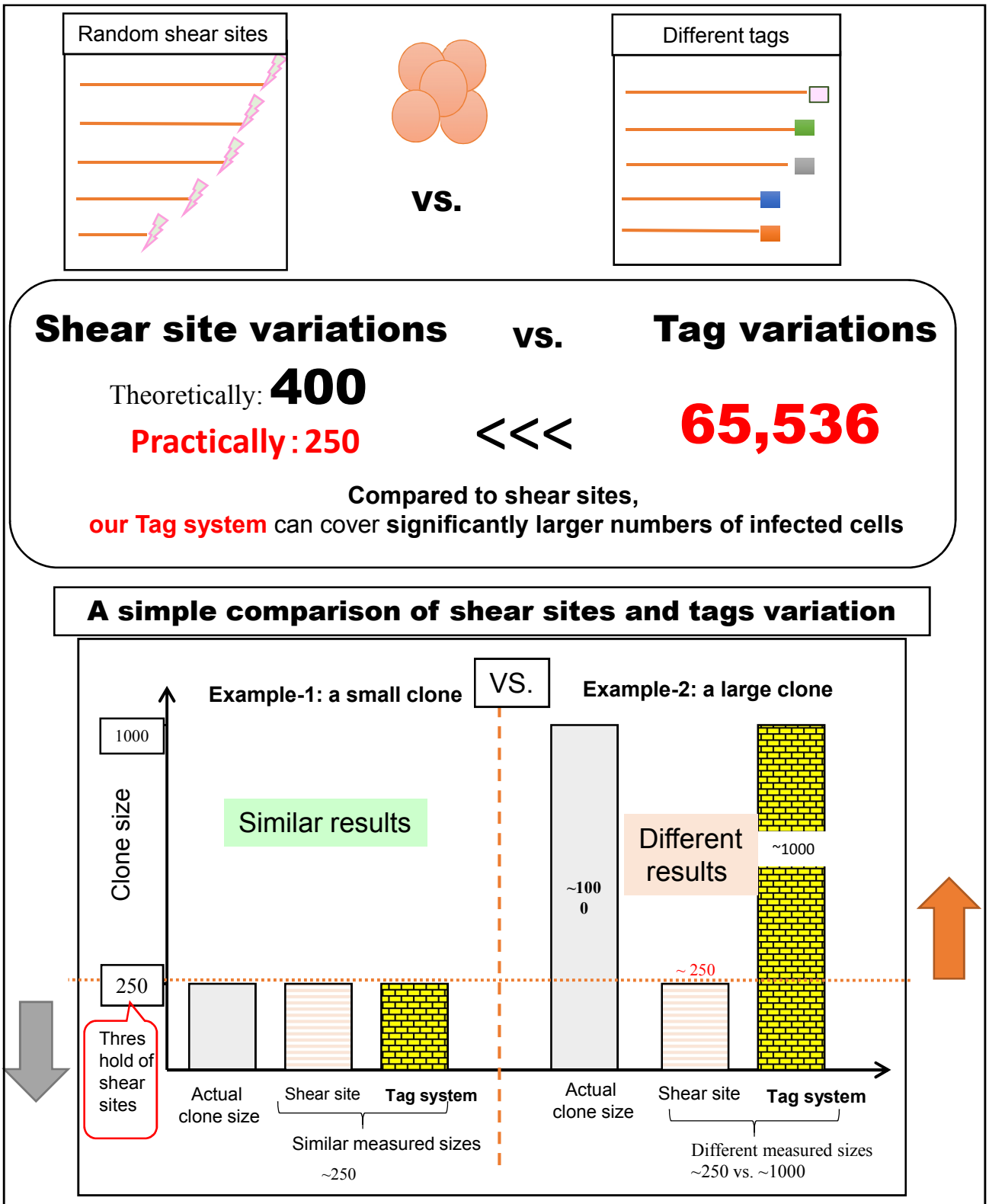
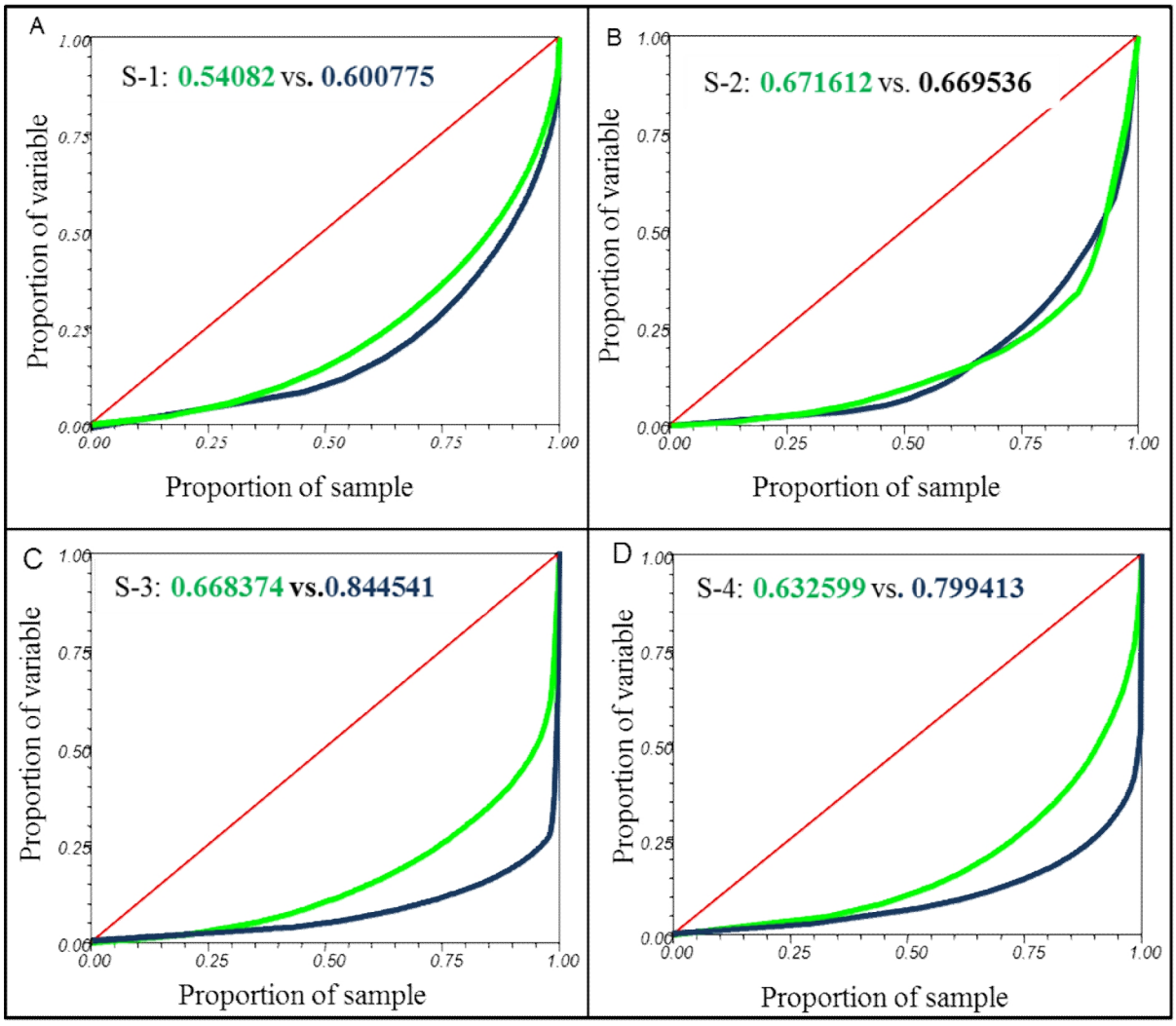


Figure 12: A simple comparison of shear site and tag variations
A simple image representing that the two approaches result in the similar clone sizes when the size of clones are small enough to be covered by shear site approach (Example-1). However, when the size of clones exceeds the shear site variations (242 based on our experiments), clone sizes get underestimated by the shear site approach. Thus, the size of large clones are only measureable by the tag system (Example-2). 400 is the theoretical upper-limit of shear site variations.



S-1: AC, 8% PVL — 0.60

S-2: SM , 9% PVL — 0.67

S-3: SM, 31% PVL — 0.84

S-4: Acute, 33% PVL — 0.8

Different clonality patterns

Similar Gini coefficients

Figure 13: Oligoclonality index for shear sites vs. combinations
The Gini coefficient and Lorenz curve were analyzed by StatsDirect software and are represented as an oligoclonality index (OCI). The red, 45-degree diagonal lines are the lines of equality. The green and blue curves are Lorenz curves of shear sites and combination data, respectively. (A) Lorenz curves and the values of OCI for S-1 (shear sites vs. combinations: 0.54082 vs. 0.600775). (B) Lorenz curves and the values of OCI for S-2 (shear sites vs. combinations: S-2:0.671612 vs. 0.669536). (C) Lorenz curves and the values of OCI for S-2 (shear sites vs. combinations: 0.668374 vs. 0.844541). (D) Lorenz curves and the values of OCI for S-2 (shear sites vs. combinations: 0.632599 vs. 0.799413)

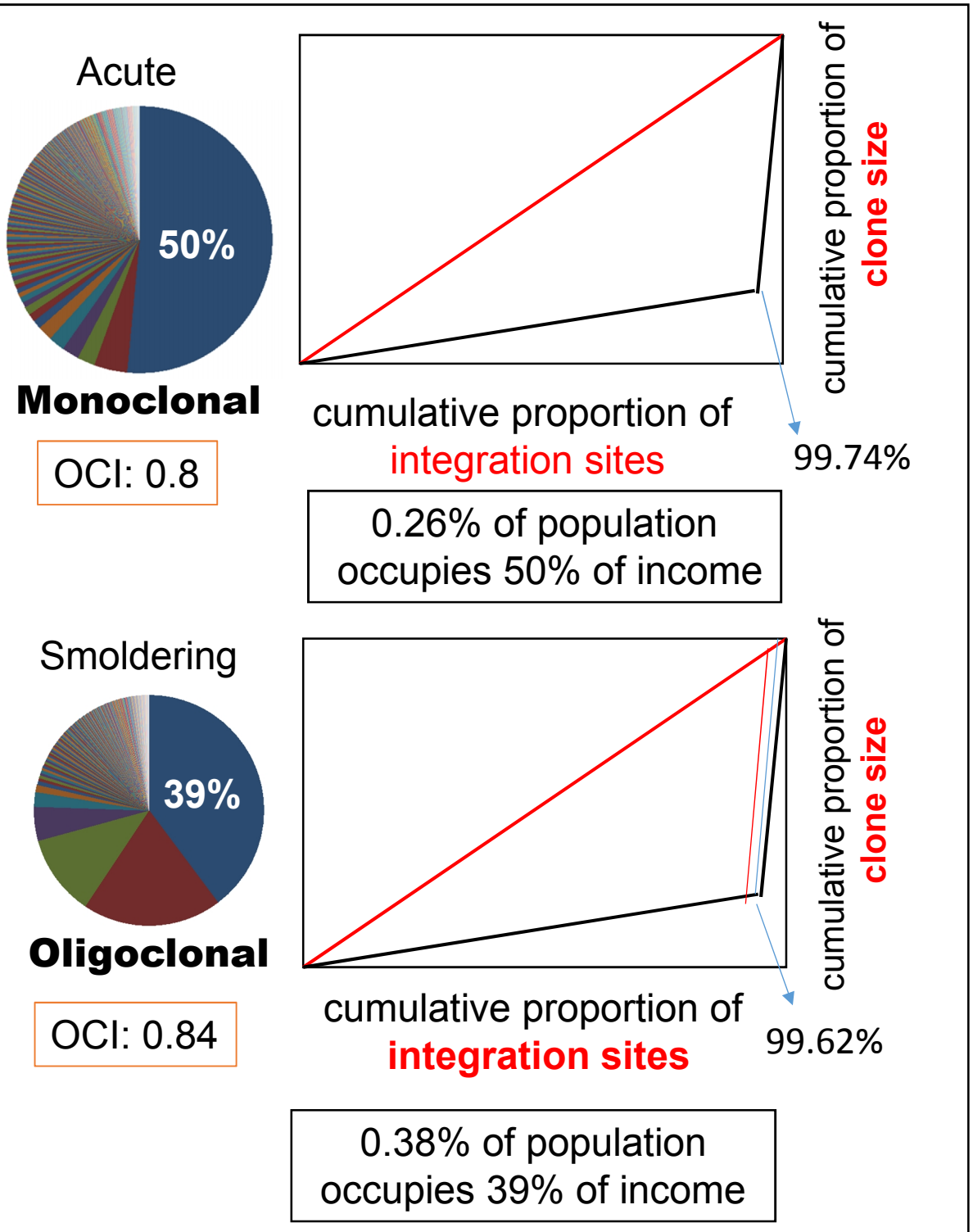
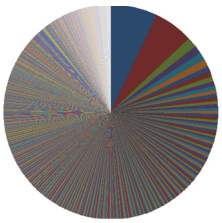


Figure 13 (E) :

Oligoclonality index/ Gini index is mainly determined by **majority of population**

Difference in limited numbers of clones (oligo or monoclones)(top-10 clones) does not affect the area between Lorenz curve and equality line, the value of Gini

Figure 13

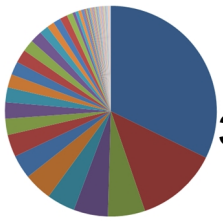


393

7.08%

0.60

AC

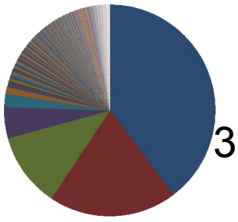


142

32.27%

0.67

SM

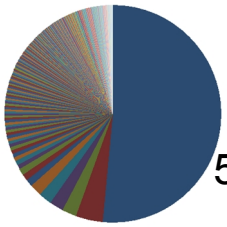


1751

39.76%

0.84

SM



2675

51.50%

0.8

Acute

Tags

		Chromosome	Strand	Position	Tags	Relative size (%)	Combinations	Relative size (%)	Shear sites*	Relative size (%)
Top 10 clones of S-1	1	chr11	+	41829319	393	7.08	269	5.30	209	2.60
	2	chr11	-	37042565	329	5.93	235	4.63	130	1.62
	3	chr7	-	121751243	83	1.50	74	1.46	43	0.54
	4	chr13	+	69268469	67	1.21	58	1.14	124	1.54
	5	chr18	-	46701081	65	1.17	58	1.14	30	0.37
	6	chr17	+	18847529	60	1.08	58	1.14	180	2.24
	7	chr15	-	37836845	46	0.83	44	0.87	29	0.36
	8	chr2	-	100184973	44	0.79	42	0.83	22	0.27
	9	chr6	+	10852456	42	0.76	40	0.79	72	0.90
	10	chr8	-	35831701	39	0.70	37	0.73	36	0.45
Top 10 clones of S-2	1	chr15	+	593[illegible]	142	32.27	119	28.95	119	21.25
	2	chr13	-	74706141	55	12.50	52	12.65	11	1.96
	3	chr3	+	28073332	25	5.68	25	6.08	11	1.96
	4	chr21	-	44242161	23	5.23	23	5.60	6	1.07
	5	chr18	-	38428907	19	4.32	19	4.62	1	0.18
	6	chrX	-	107427783	19	4.32	19	4.62	2	0.36
	7	chr13	-	84177236	16	3.64	16	3.89	5	0.89
	8	chr21	+	25834766	15	3.41	15	3.65	7	1.25
	9	chr2	+	234346116	11	2.50	10	2.43	6	1.07
	10	chr7	-	99740574	11	2.50	11	2.68	2	0.36
Top 10 clones of S-3	1	chr4	-	563[illegible]	1751	39.76	1192	35.72	242	9.48
	2	chr20	+	5800738[illegible]	863	19.60	579	17.35	232	9.08
	3	chr5	+	62579369	502	11.40	336	10.07	202	7.91
	4	chr6	+	133958124	210	4.77	207	6.20	191	7.48
	5	chr3	-	126392282	91	2.07	86	2.58	94	3.68
	6	chr3	+	178928610	43	0.98	43	1.29	54	2.11
	7	chr8	+	119096533	27	0.61	27	0.81	43	1.68
	8	chr10	-	111698526	18	0.41	18	0.54	9	0.35
	9	chr13	+	21355493	17	0.39	17	0.51	24	0.94
	10	chr18	+	62126326	16	0.36	14	0.42	10	0.39
Top 10 clones of S-4	1	chrX	-	8370[illegible]	2675	51.50	2038	46.54	222	8.35
	2	chr14	+	3065589[illegible]	209	4.02	160	3.65	87	3.27
	3	chr14	+	49676335	112	2.16	97	2.22	77	2.90
	4	chr6	-	85461536	108	2.08	95	2.17	80	3.01
	5	chr16	-	17339636	102	1.96	97	2.22	98	3.69
	6	chr8	+	96129917	93	1.79	75	1.71	59	2.22
	7	chr1	+	4032445	55	1.06	48	1.10	22	0.83
	8	chr7	+	140001929	50	0.96	49	1.12	40	1.50
	9	chr21	+	35571080	49	0.94	41	0.94	38	1.43
	10	chr1	-	56007274	35	0.67	32	0.73	38	1.43

Tag system for measuring size of clones

Absolute numbers of infected cells

Especially top-10 clones

Figure 13 (F) : Gini coefficient of AC was different from Acute sample.
SM and Acute samples with different clonality patterns showed similar Gini coefficients.
This is the disadvantage of GINI which can not differentiate polyclonality from oligoclonality and oligoclonality from monoclonality.
Later I will show that accurate measurement of the number of infected cells particularly the size of top-10 clones are essential to differentiate clonality status of different ATL subtypes. Therefore I put my major effort into accurate measurement of the absolute size of clones by the tag system.

Validation of the methodology

My 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 10 and 12). Finally, I 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 my system in the clonality analysis. I designed a suitable control because there was not an appropriate control available. Using my system, I could evaluate the method and confirm its accuracy, sensitivity, and reproducibility. I selected two samples with the following special conditions as starting materials for preparing the control system (Figure 14).

Sample one (M): DNA from an acute ATL patient with 100% PVLs and a single integration site in the major clone (Figure 14). 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, I 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 14A).

Having prepared these two samples, they were sonicated and mixed in proportions of 50:50 and 90:10 (Figure 14B). These known proportions were thus expected to generate specific patterns that could be verified with my 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 14C. 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.

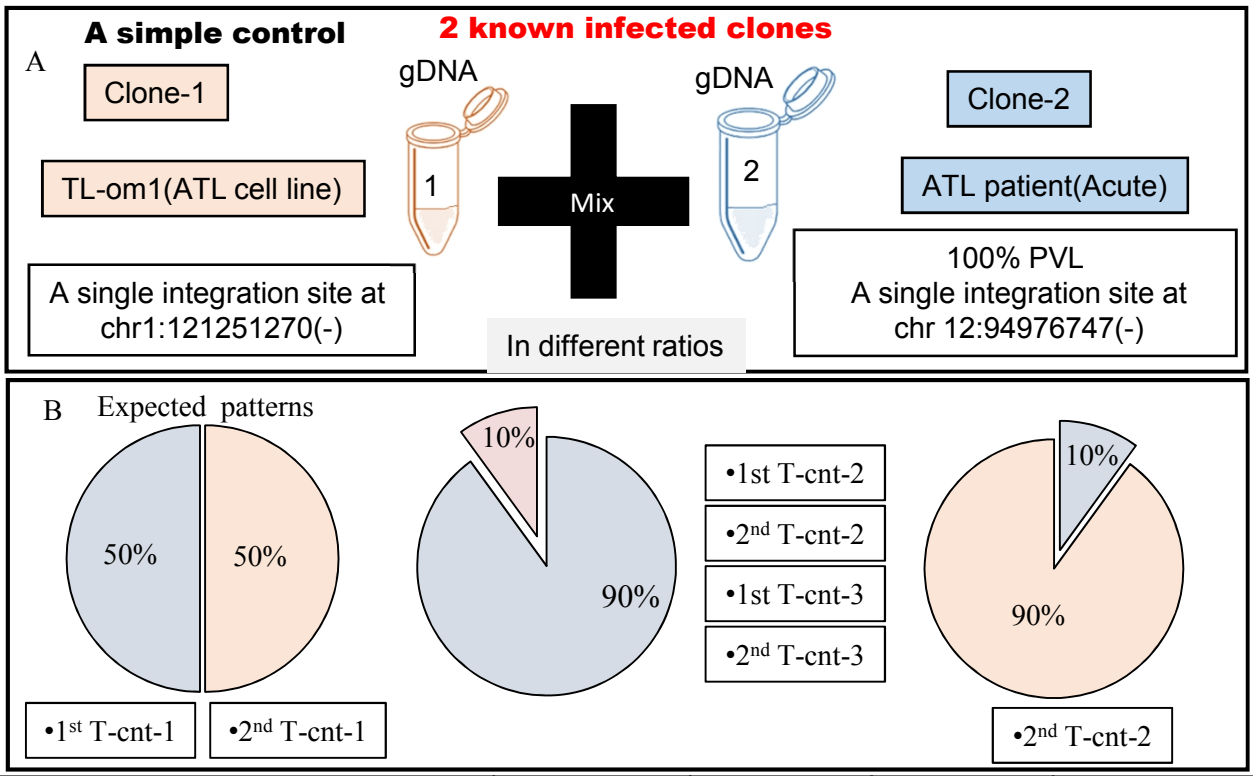
I named the samples of the second trial as second trial control-1 ~ 4 and abbreviated them as 2nd T-cnt-1 ~ 4. DNA samples were mixed similarly to that for the first trial except for sample four (Figure 14D). In contrast to the first trial, I used 1 μL of the external PCR product without any dilution 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 15). Figure 15 shows the data when the optimal conditions were considered.

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

The 'absolute error', a technique used to evaluate system accuracy [54], was used to assess my method. The experimental values were subtracted from expected values (Figure 16A). Taking advantage of my 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 16B and C). The absolute errors of raw sequence reads for the first trial 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 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, my system showed lower absolute errors than when considering only shear sites (Figure 16). Owing to differences in analyzed samples and system setups, I could not directly compare my data with published data [18, 49]. Indirect evidence, however, provided by shear site analysis of my own data illustrated that my system has lower absolute errors than using the shear site-based methodology. See summary of absolute errors data in Figure 17.



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 14: Preparing the control system.
(A) I prepared a simple control system by mixing two known infected clones in different ratios.Clone-1 was from Tlom-1 which is an ATL cell line with a single IS. and Clone-2 was from an acute patient with 100% PVL and a single integration site. 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.

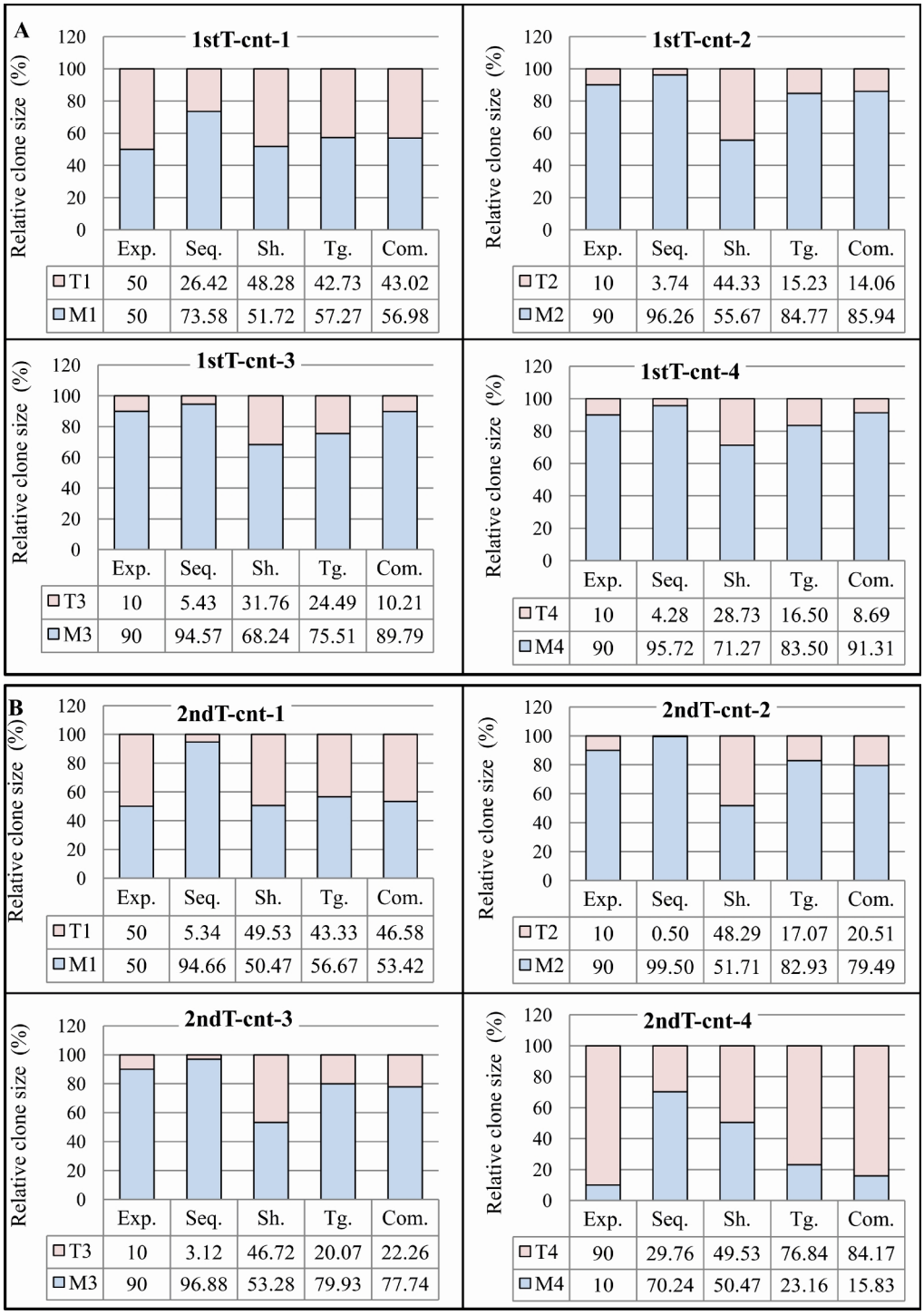
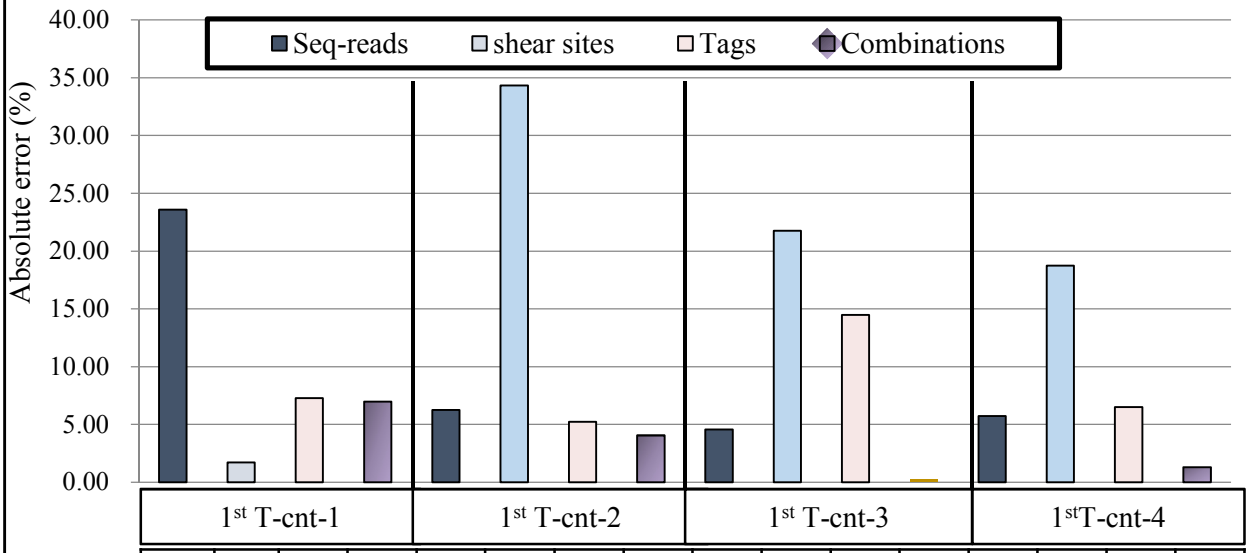


Figure 15: Validation of the tag system. For each control sample, both the expected and the experimentally observed patterns of raw sequence reads, shear sites, and the combination of tags and shear sites are represented in the bar graphs. Abbreviations: Com.: Combinations, Exp.: expected pattern, Seq.: 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%).

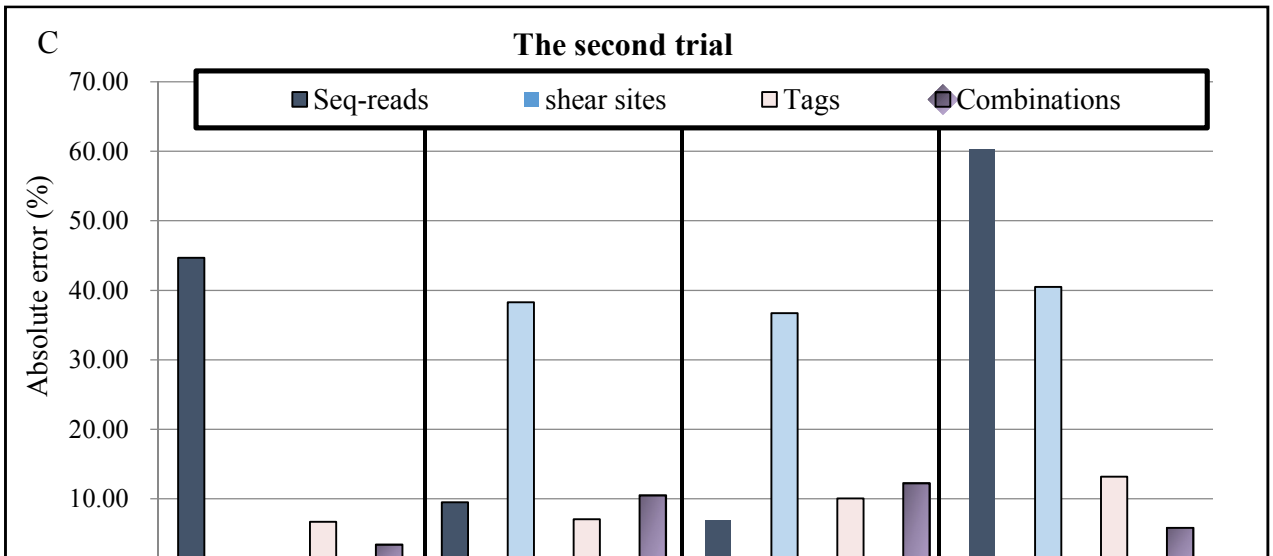
A **Absolute error = | (Expected value) − (Experimentally observed value)|**

B



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

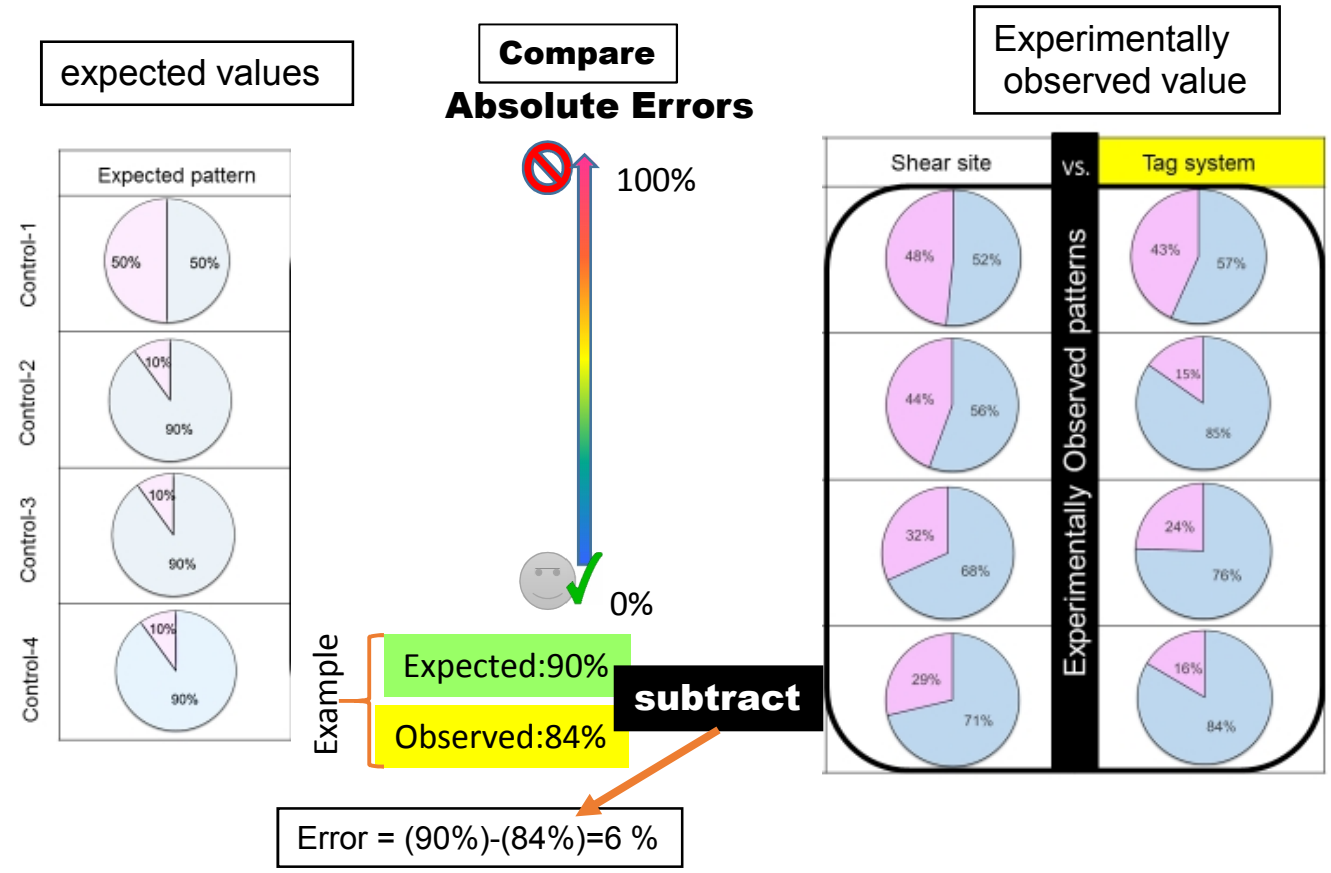
C



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 16: 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. 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.

Absolute error=I Experimentally observed value- Expected value I



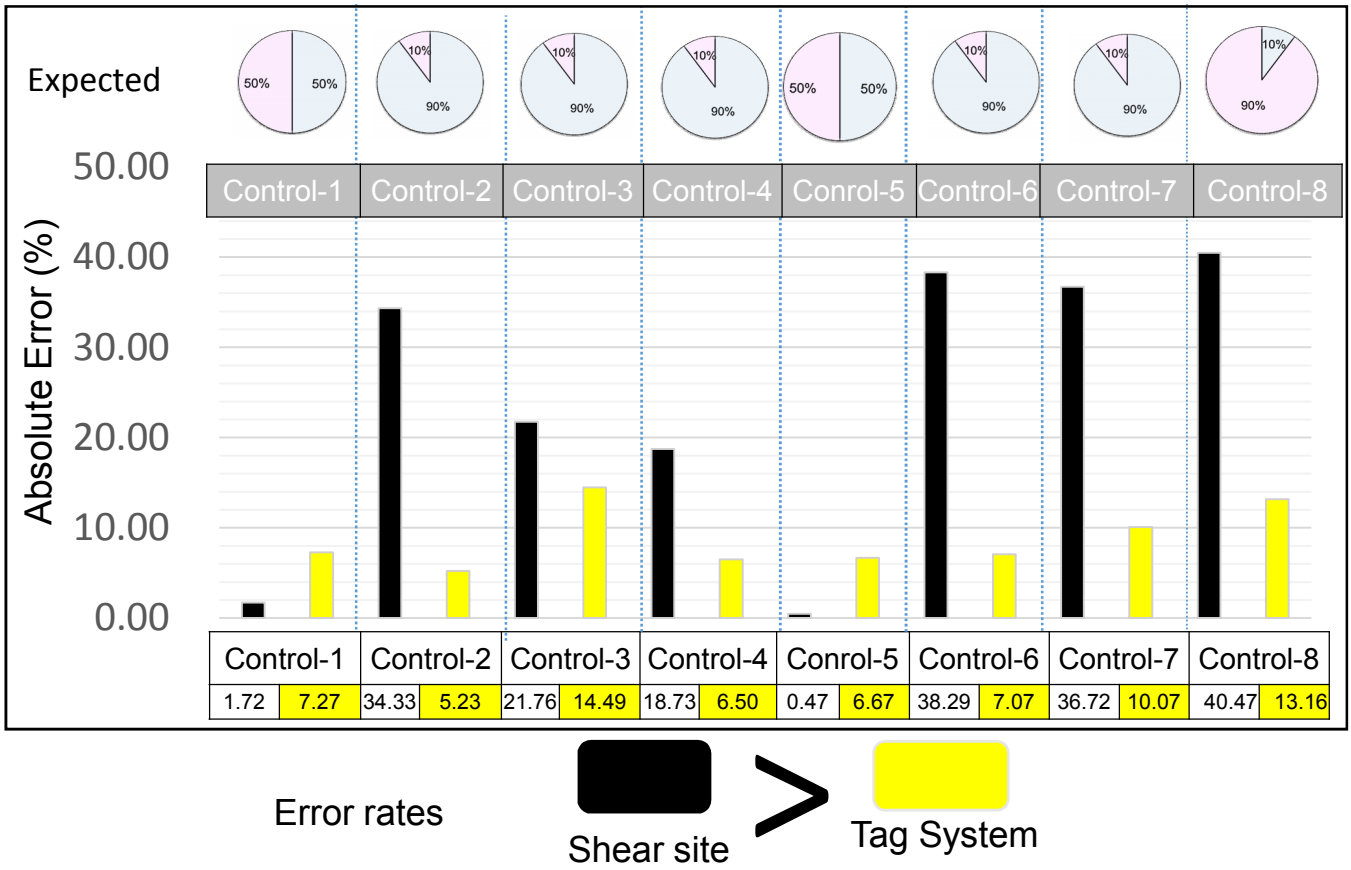


Figure 17: Summary of Absolute errors to evaluate the measurement,
I calculated Absolute Errors by subtracting the expected values from experimentally observed values. My tag system showed significantly lower absolute errors than shear site strategy.

In-silico analysis

Raw sequencing data were processed according to the workflow described in Figure 18. 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 later. 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].

Processing, management, and analysis of the large amount of data generated by deep sequencing require special infrastructures and bioinformatics skills. I designed a data analysis and interpretation pipeline specific for HTLV-1 integration sites and clonality studies. The workflow is provided in Figure 18. First, the raw data for high-throughput sequencing were checked for quality by the FastQC tool. I 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 18). The downstream 23 nucleotides, which represented LTR-specific primers, were also trimmed before further analysis. I 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]. I employed various parameters of Bowtie and different lengths of read three to obtain the optimal mapping yield. 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.

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 18).

Removing background noise

Data obtained from next-generation sequencers are not error free[45, 61-64]. There are many reports on the error rate of Illumina sequencers [65, 66]. 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 [67]. In our study, consistent with the data of Kivioja *et al*. [67], 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, I tested different filtering conditions to remove background noise. Here, I report my proven filtering approach (Figure 19).

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, 67]. Therefore, I 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 (Figure 20).

I used a filtering system, which I 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. I referred to this filtering approach as 'Join Tag- remove10' (JT-10) in the Figure legends. To facilitate understanding, these filtering conditions are illustrated in (Figure 19). I provided the absolute error data for different filtering conditions in Figure 21.

Mapping, reads coverage, and tag variations

As later described in Materials and methods, incorporating 5-bp random nucleotides downstream of the region specific for read-1 sequencing primer was necessary to generate high-quality sequencing reads. These 5-bp random nucleotides were not used for S-1, S-2, S-3, and S-4 samples, and thus resulted in a low sequence quality. I handled low quality reads by keeping only the generated sequencing reads that were uniquely mapped, similar to the strategy of Heng Li *et al.* [63]. I utilized different mapping software including Bowtie [58] and Burrows-Wheeler Aligner (BWA). The number of uniquely mapped reads was not significantly different. Bowtie was used for further analysis owing to convenience and higher speed. Final mapped reads were: S-1: 2,758,423; S-2: 281,941; S-3: 4,315,531; and S-4: 11,870,957. Owing to high sequencing quality, the number of uniquely mapped reads was greater for control samples (Table 3).

Different numbers of generated sequencing reads were analyzed and evaluated. Maximum, minimum, and average mapped reads of our analyzed samples were [1st T-cnt-1: 27,962,532], [S-2: 281,941], and 10,485,747, respectively. These numbers are comparable to those of published methods in which a maximum, minimum, and average mapped reads of 107509, 4659, and 31961 were used [51].

Depending on the purpose of any particular experiment, a high or a low coverage may be selected. Our data suggest that, although a low coverage (for example S-2: 281,941) can

provide some estimation of clonality, a higher coverage may ensure a more reliable and representative picture of clonal composition and isolate a larger number of integration sites. Based on these data, I recommend use of about 2–3 million mapped reads for each analysis. Sequencing and mapping errors are intrinsic to NGS data [63, 68-71]. Therefore, occasional generation of false positive integration sites is unavoidable in this kind of analysis. Considering all the characteristics in NGS data analysis, I designed the study and analysis steps to avoid errors as much as possible, and accurately generate and interpret data.

Tags are randomly generated nucleotides incorporated into splinkerette adaptors. Since 7-bp is the default length of read-2 in Illumina, I had initially used 7-bp tags for optimization (S-1, S-2, and S-3). Later it became possible to increase the length of read-2 to 8-bp. Therefore, I analyzed samples with 8-bp tags (S-4, first trial samples, and second trial samples). I also analyzed 2nd trial samples with both 7-bp and 8-bp tags, and compared the results. The measured clone sizes were not significantly different, but barely better in the case of 8-bp tags (data are not shown). Therefore, I used 8 bp as the optimal length for tags in my analysis.

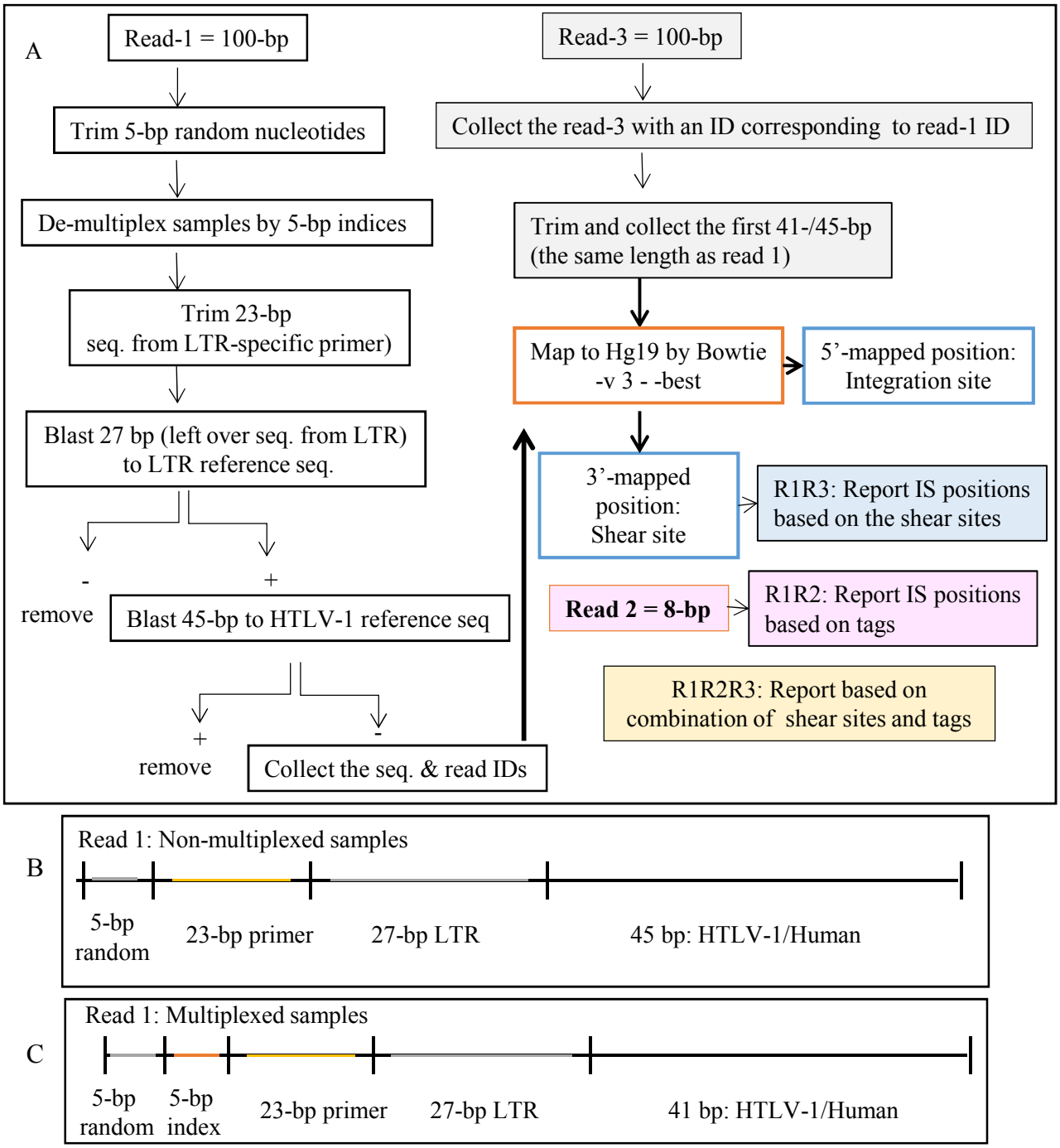
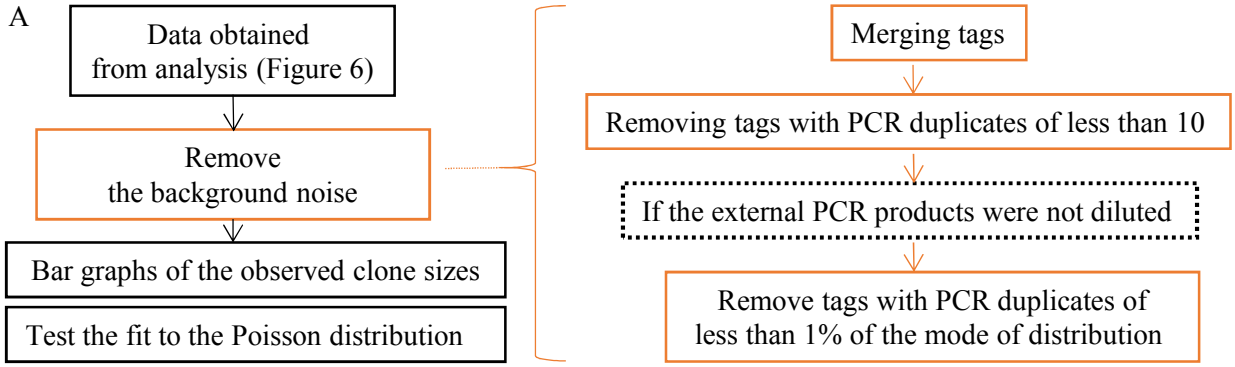


Figure 18: *In-silico* analysis work flow. (A) Illumina HiSeq 2000 platform outputs raw data of (Read-1 = 100 bp), (Read-3 = 100 bp), and (Read-2 = 8 bp). Data were analyzed according to this work flow after checking quality with the FastQC tool. In the case of Read-1, the first 5 bp were trimmed, and 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.

Figure 19.

A



B

	Chromosome	Strand	Position	Tags	PCR duplicates
1	chr 1	-	121251270	ACATGCTA	24
2	chr 1	-	121251270	ACAATCTA	32
3	chr 1	-	121251270	ACATTCTG	22
4	chr 1	-	121251270	ACATACTA	11
5	chr 1	-	121251270	ACATTCTA	28
6	chr 1	-	121251270	CTAGAGGT	39
7	chr 1	-	121251270	CTATAGGT	27
8	chr 1	-	121251270	CAGAAGGC	12
9	chr 1	-	121251270	AAGAACGC	12
10	chr 1	-	121251270	AAGAAGGC	56

Merge the tags with one mismatch

	Chromosome	Strand	Position	Tags	PCR duplicates
1	chr 1	-	121251270	ACATGCTA-ACAATCTA-ACATTCTG-ACATACTA-ACATTCTA	117
2	chr 1	-	121251270	CAGAAGGC-AAGAACGC-AAGAAGGC	80
3	chr 1	-	121251270	CTAGAGGT-CTATAGGT	66

The frequency of unique tags before filtering: **10**

The frequency of unique tags after filtering: **3**

Figure 19

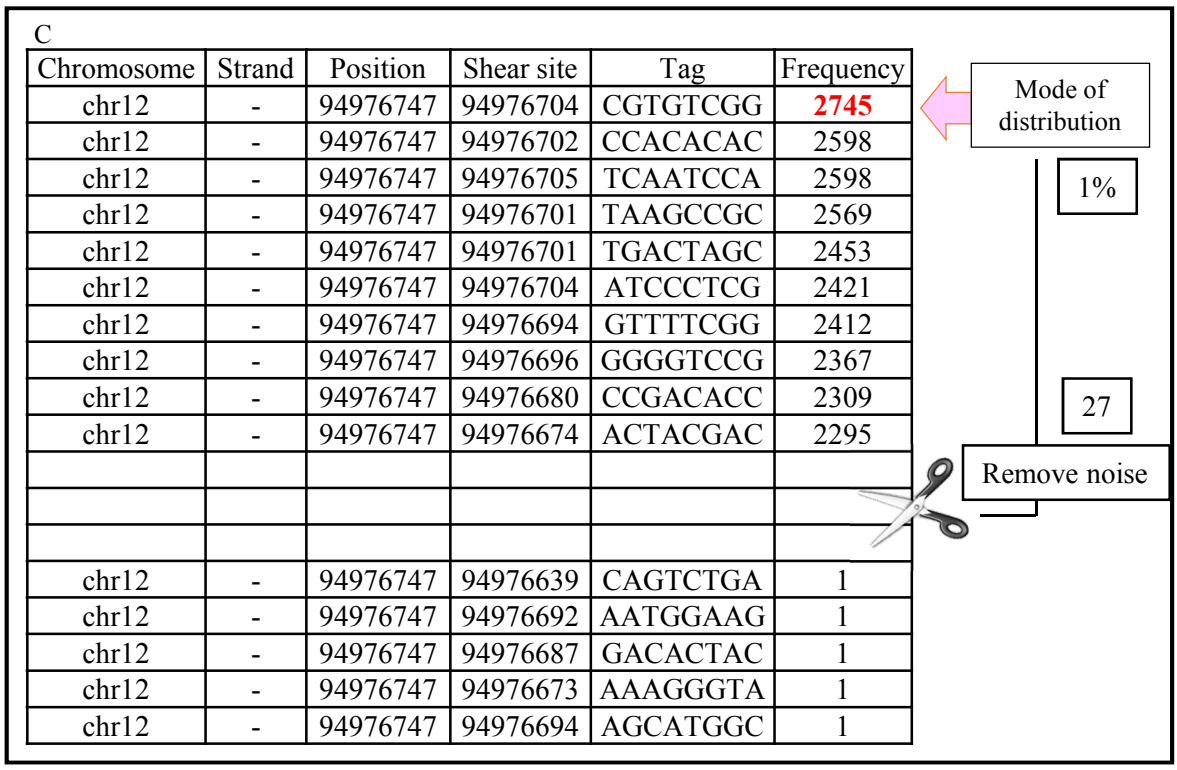


Figure 19: The filtering system for removing background noise
(A) After completing the data analysis as described in Figure 18 we performed a workflow, which included: (1) removing the background noise, (2) fitting data to the Poisson distribution, and (3) preparing graphs of the observed clone sizes. Filtering was done separately for each clone. The background noise was removed by merging tags that differed by one nucleotide (one mismatch permission). Tags with less than ten PCR duplicates were then removed. In the case of the second trial's control samples 1-4 for which the external PCR products were not diluted, tags with PCR duplicates less than 1% of the mode of distribution, were removed (See Supplementary Figure S6, Additional file 1). (B) A simple diagram of merging tags is presented. (C) The external PCR products were not diluted in the second trial (control samples 1-4). For these samples, in addition to the merging approach, tags with PCR duplicates less than 1% of the mode of distribution were removed. Mode of the above depicted distribution is 2745 (indicated in red typeface). In such distribution, tags with a frequency less than 27 (1% of 2745) were removed.

Figure 20.

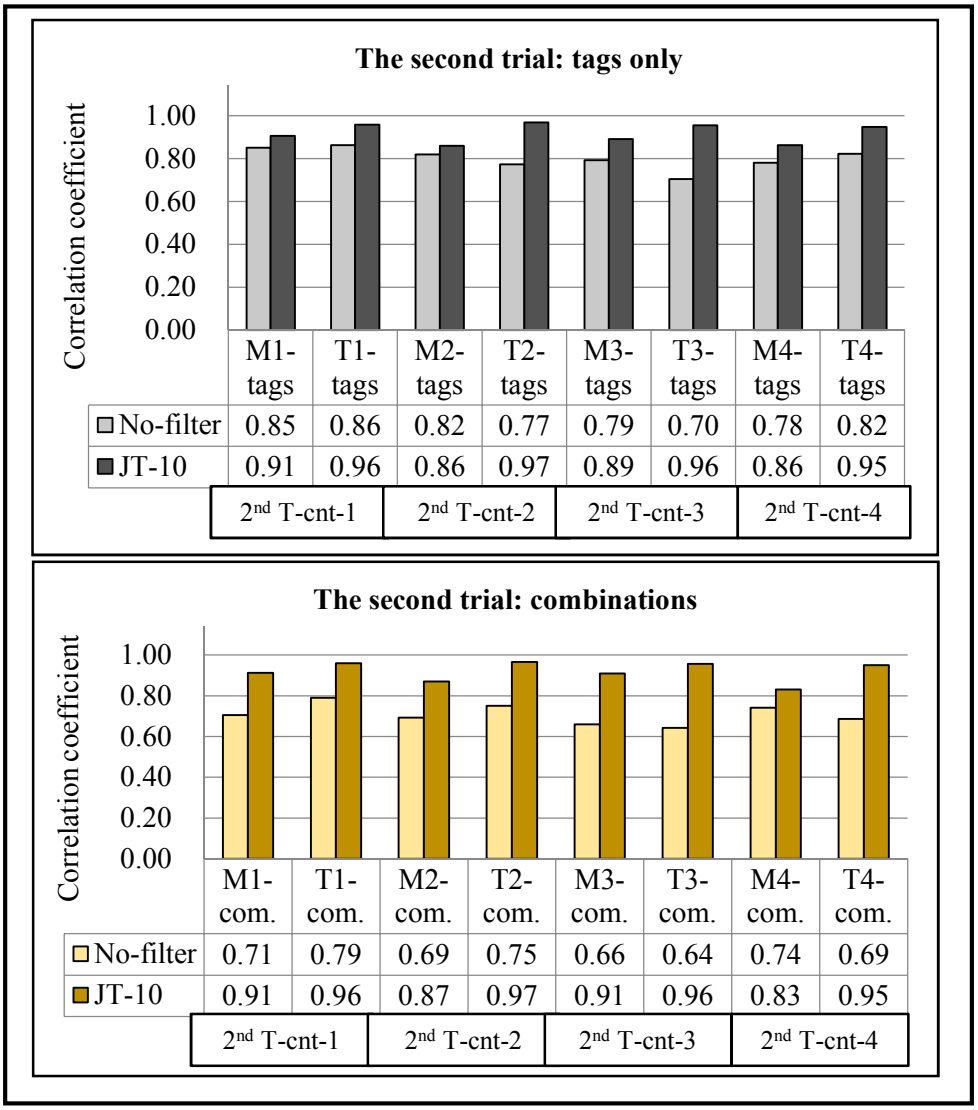
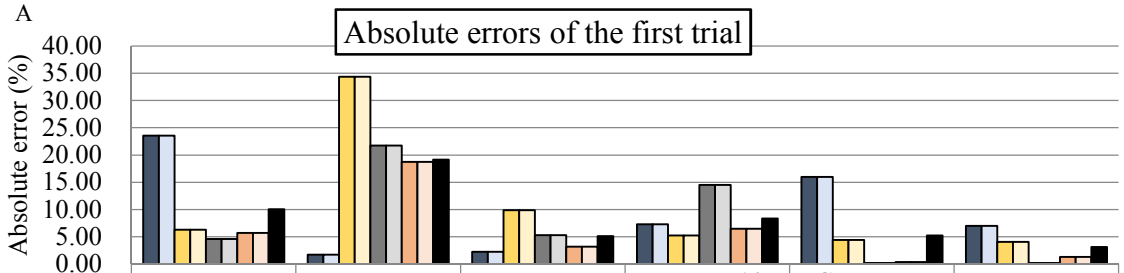


Figure 20: Testing of the data fit to the Poisson distribution
The efficacy of filtering was determined by checking the fit to the Poisson distribution. Distribution of tags for each sample was analyzed by R-package "gamlss.tr", and the correlation coefficient before filtering has been compared to that after filtering.
(A) The second trial: tags only.
(B) The second trial: the combination of tags and shear sites.

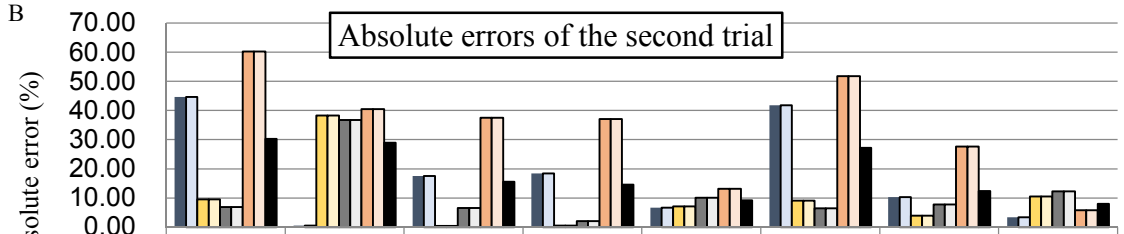
Figure 21

A



	Seq-reads	shear sites	Tags-Non filt.	Tags-JT-10-filt.	Com.-Non filt.	JT-10-com
M1	23.58	1.72	2.23	7.27	16.00	6.98
T1	23.58	1.72	2.23	7.27	16.00	6.98
M2	6.26	34.33	9.90	5.23	4.44	4.06
T2	6.26	34.33	9.90	5.23	4.44	4.06
M3	4.57	21.76	5.27	14.49	0.21	0.21
T3	4.57	21.76	5.27	14.49	0.21	0.21
M4	5.72	18.73	3.19	6.50	0.35	1.31
T4	5.72	18.73	3.19	6.50	0.35	1.31
average	10.03	19.14	5.15	8.37	5.25	3.14

B



	Seq-reads	shear sites	Tags-Non filt.	Tags-JT-10-filt.	Tags-JT-10-1%-filt.	Com.-Non fil.	Com-JT-10	Com-JT-10-1%
M1	44.66	0.47	17.57	18.43	6.67	41.77	10.32	3.42
T1	44.66	0.47	17.57	18.43	6.67	41.77	10.32	3.42
M2	9.50	38.29	0.45	0.48	7.07	9.03	3.88	10.51
T2	9.50	38.29	0.45	0.48	7.07	9.03	3.88	10.51
M3	6.88	36.72	6.51	2.08	10.07	6.41	7.75	12.26
T3	6.88	36.72	6.51	2.08	10.07	6.41	7.75	12.26
M4	60.24	40.47	37.54	37.12	13.16	51.76	27.59	5.83
T4	60.24	40.47	37.54	37.12	13.16	51.76	27.59	5.83
average	30.32	28.99	15.52	14.53	9.24	27.24	12.39	8.01

Figure 21: Evaluating the accuracy of clonality analysis for differing conditions
The method was evaluated by calculating “absolute error”, which is further described in Results and discussion, Fig. 5, and Fig. 3. Absolute error was calculated for different conditions including: (a) raw sequence reads, (b) shear sites, (c) only tags, and (d) combination of tags and shear sites. In the case of tags only and combinations, both the non-filtered data and the data filtered with the “merging” approach (JT-10 and JT-10-1%) are provided. (A) Absolute errors of the first trial. (B) Absolute errors of the second trial.

Table 3.

Sample	Status	PVL (%)	Barcode	Total reads	Reads from 5'LTR	Reads from 3'LTR	Reads uniquely mapped to Human genome
S-1	AC	7.56	No	101,697,565	37,429,937	64,267,628	2,758,423
S-2	SM	9.01	No	102,690,388	38,344,138	64,346,250	281,941
S-3	SM	31.15	No	96,569,010	31,068,714	65,500,296	4,315,531
S-4	Acute	32.56	No	111,838,665	34,456,523	77,382,142	11,870,957
1st T-cnt-1	TLom1/Acute	100	No	135,665,814	58,159,788	77,506,026	27,962,532
1st T-cnt-2	TLom1/Acute	100	No	108,939,606	46,902,164	62,037,442	22,456,195
1st T-cnt-3	TLom1/Acute	100	No	105,244,134	44,280,981	60,963,153	20,294,502
1st T-cnt-4	TLom1/Acute	100	No	92,804,419	38,245,287	54,559,132	19,736,034
2nd T-cnt-1	TLom1/Acute	100	ACAGT	20,653,487	8,877,796	11,775,691	3,580,966
2nd T-cnt-2	TLom1/Acute	100	GGCTA	31,909,311	13,607,338	18,301,973	5,937,997
2nd T-cnt-3	TLom1/Acute	100	TTACG	15,686,210	6,774,110	8,912,100	2,683,504
2nd T-cnt-4	TLom1/Acute	100	GCTAC	22,110,335	9,443,089	12,667,246	3,950,379

Table 3. Sample information and mapping results

Sample information including the disease status, PVL, barcodes, total sequencing reads, and numbers of reads from 5'-LTR, reads from 3'-LTR, and reads uniquely mapped to human genome are presented in this table. Sequencing errors have not included here. The first eight samples were sequenced in separate lanes, and the remaining four samples were barcoded and sequenced in one lane of HiSeq 2000. *In silico* analysis was done by our own Perl scripts, and the sequencing reads were mapped to Hg19 by Bowtie-1 with -v 3 - -best parameters and the same length of read-1 and read-3 were used for mapping. See (Figure 3. Preparing the control system) for more information regarding the last eight samples. PVL of TL-om1 and the control Acute sample were 100% and 100.42%, respectively. In the main manuscript we referred to values of PVLs rounded to zero decimal places: 8%, 9%, 31%, 33%, and 100%. 8-bp random tags were used for all samples except S-1, S-2 and S-3 for which 7-bp tags were used (see Supplementary Notes). Raw sequencing data have been deposited in the Sequence Read Archive with access number of [SRP038906].

PCR-southern experiments

I used PCR-southern to provide reference data about the relative size of the major clone in sample S-4 because I believed this classical approach had less bias than the other conventional methods. Because HTLV-1-infected clones contain the same LTR sequence, they are detectable using a common LTR probe by PCR-southern. Size of clones was estimated by the strength of the bands detected from each specific clone. Furthermore, I compared the relative size of clones as measured by PCR-southern with that obtained using the Shear site or tag system. Relative size of clones based on PCR-southern data was similar to that of my tag system, which validated the accuracy of my tag system. The following is the detailed information obtained from the PCR-southern experiments (Figure 22).

Two bands detected from S-4 were sequenced: the upper band from chromosome X [83705328 (-), the Blue clone] and the lower band the amplification of the 5-LTR-containing part of the HTLV-1 genome. No bands were detected from the Red clone [the second clone of S-4; chromosome 14: 30655896(+)].

I compared the quantitative size of the clones measured by shear site and tag system with the qualitative estimation of clone size based on PCR-southern data.

The shear sites data showed that the Blue clone was 2.6 times larger than the Red clone. If the size estimation by shear sites (222 vs. 87 for Blue vs. Red, respectively) was accurate, a weak band should at least be detected from the red clone using PCR-southern methods. However, this method detected a strong monoclonal band from the Blue clone, and no band from the Red clone. This was consistent with the tag data, which estimated the size of the Blue clone to be 12.8 times larger than the Red clone (2675 vs. 209 for Blue vs. Red, respectively).

I compared the clone size data obtained from S-4 with those of S-1. I first considered the size of the first clone, and then the number of integration sites. S-1 was used as a clone size control. Consistent with shear site and tag data, PCR-southern did not detect any major band from S-1. The number of integration sites isolated from each sample (S-1: and S-4) were 1030 and 384, respectively.

Based on the shear site data, the size of the first clone in the sample S-1 was 209 and that of S-4 was 222. However, based on the tag data, the size of first clone in the sample S-1 was 393 and that of S-4 was 2675. If the size estimation as modeled from the shear site data were accurate, I would expect that the sample S-4 should show a smear-like pattern, similar to that of S-1 on the PCR-southern data. However, the data from the PCR-southern method detected a monoclonal band from S-4, which was consistent with the data from tag system (i.e., that the size of the clones from S-1 and S-4 was significantly different) (Figure 22). Taken together, I believe these data further support the accuracy of clone size as measured by the tag system.

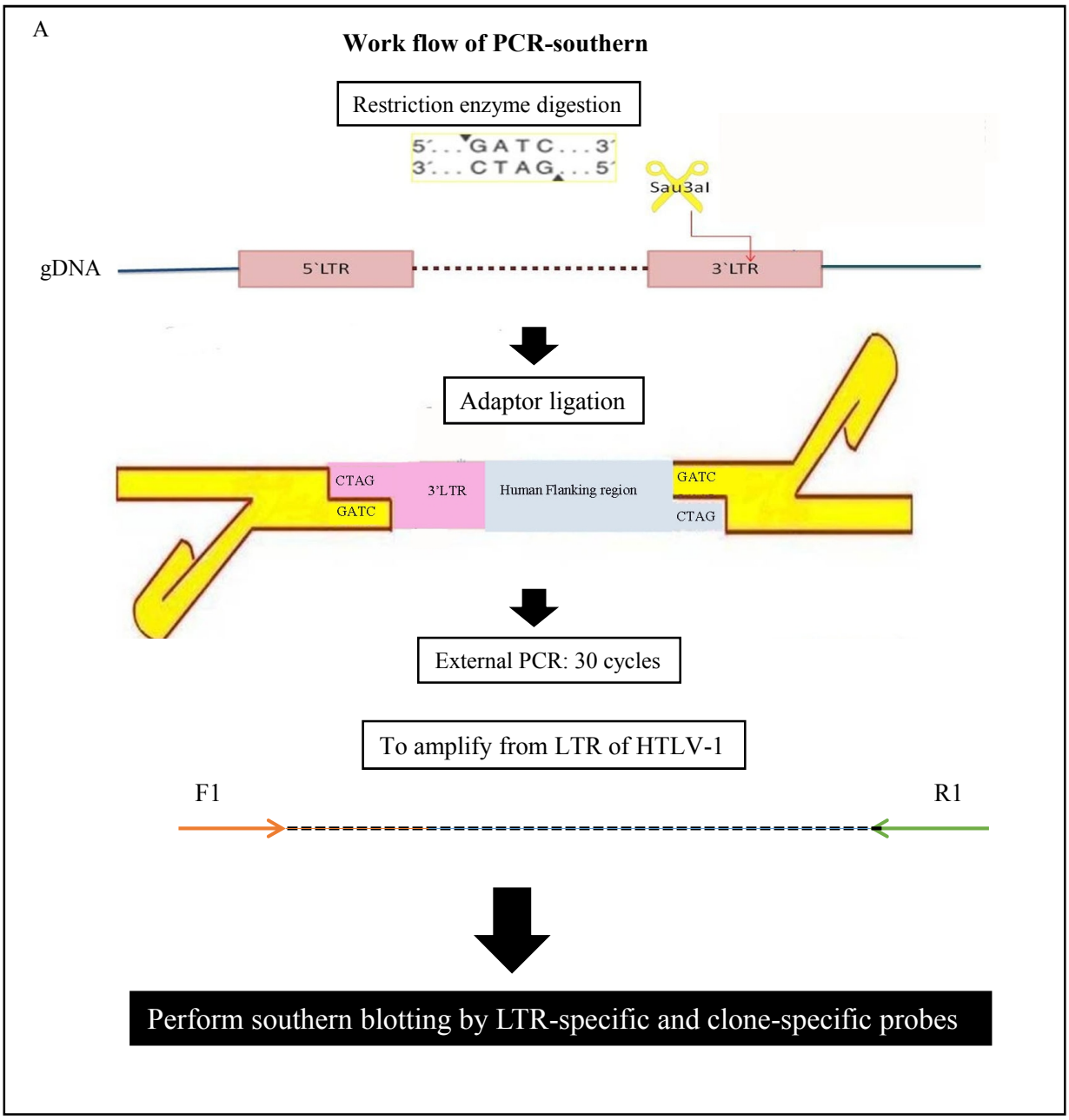


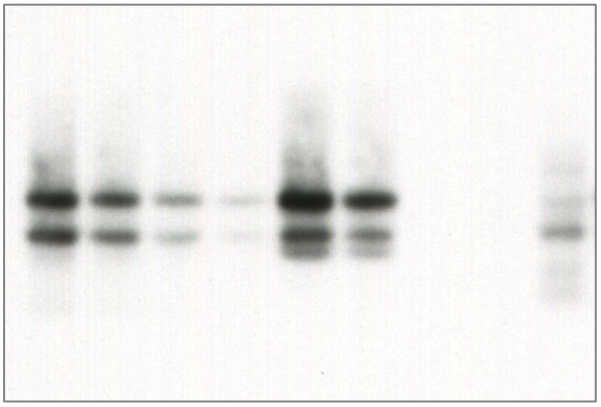
Figure 22: Detecting the major clone of S-4 by PCR-southern
Work flow of PCR-southern: restriction enzyme digestion, adaptor ligation, external PCR amplification and southern blotting. Perform experiments as described in additional protocols. (B) clones detected by a LTR-specific probe (C) chrx:83705328(-) clone detected by the clone-specific probe (D) Comparison of the clone size estimated by shear sites and the tag system with that of PCR-southern.

Figure 22: Detecting the major clone of S-4 by PCR-southern

B

Detected by LTR-specific probe

1 2 3 4 5 6 7 8 9



Order of samples

① S-4 (10 ㎕)
② S-4 (5 ㎕
③ S-4 (2.5 ㎕)
④ S-4 (1 ㎕)
⑤ Tl-om1 (10 ㎕)
⑥ TL-om1 (5 ㎕)
⑦ HIV cell line (U1) (10 ㎕)
⑧ Normal PBMC (10 ㎕)
⑨ S-1 (10 ㎕)

LTR-specific probe (7 5-bp)
TGTGTACTAAATTTCTCTCCTGAGAGTGCT
ATAGGATGGGCTGTCGCTGGCTCCGAGCCA
ACGGAGTCGCCGGTA

C

Detected by clone-specific probe

1 2 3 4 5 6 7 8

Order of samples

① S-4 (10 ㎕)
② S-4 (5 ㎕
③ S-4 (2.5 ㎕)
④ S-4 (1 ㎕)
⑤ Tl-om1 (10 ㎕)
⑥ TL-om1 (5 ㎕)
⑦ HIV cell line (U1) (10 ㎕)
⑧ Normal PBMC (10 ㎕)

Blue clone-specific probe (171-bp)
(chrx:83705328 (-))
GGTGAGATTGCTTTCTTGTAGGCAGTATA
TAGTGGAGTGATGGTTTTTTTTGTTGTTGT
CCATTTAGCCAGTCTATATATTTTAAGTG
GAAAGTTTAATTCATTTATATTCAAAATC
ATAATTGATATGTGAATATTTATTCCTGTC
ATTTTACTAGTTGATTTCTGGTGG

D

Sample	Shear site	Tag system	PCR- southern	Comment
S-1	**209**	393	A smear-like pattern	All three approaches showed similar results
S-4	**222**	2675	A strong monoclonal band	Tag system and PCR-southern : consistent results Shear sites vs. PCR-southern: controversial results

Approach	Blue vs. Red	Proportion of Blue vs. Red
Shear sites	222 vs. 87	2.6 vs.1
Tag system	2675 vs. 209	12.8 vs. 1

Also see Figure 10 and 11
for data of shear sites and tags

Results-section-1-B

Confirming reproducibility of results using clinical samples

We selected samples from ACs and different subtypes of ATL that had differing PVLs [AC (n = 12), SM (n = 9), Chronic (n = 12), and Acute (n = 10)] (Figure 23). Sample information has been provided in Table 4. All ACs and some SM patients showed large numbers of small clones with a uniform distribution pattern. In addition, ACs showed large numbers of integration sites ranging from 100 to 5000, independent of their PVLs. The number of isolated integration sites depended on each sample (see Table 4). Based on our analysis, size of clones in most ACs was >200 infected cells (Figures 10, 11; Table 1). When the clones of ACs were arranged in descending order, the size of the first and the second largest clone was not significantly different. Moreover, as expected, both tag and shear site analysis showed similar results because the clone sizes were below the practical threshold of shear site variations (<250) (Figure 12).

In samples from patients with smoldering ATL, polyclonal and oligoclonal patterns were observed (Figure 24). In some of these samples the size of clones exceeded the number of shear site variations, which caused a difference in the clonality patterns observed by the shear site method and our tag system. The biological significance of the accurate measurement of clone sizes was further highlighted when we analyzed samples from these patients over time (results have been described later).

For all of the analyzed chronic samples, the size of the clones, and particularly the major clone, exceeded the number of shear site variations. Therefore, the shear site method underestimated the size of clones, and the clonality patterns observed were similar to those of ACs (Figures 24–26). However, our tag system more accurately measured the size of clones, and monoclonal patterns or largely expanded oligoclonal patterns were detected from these chronic samples (Figure 25).

Similarly, in the case of the acute samples, the size of clones was underestimated by the shear site method; thus, giving rise clonality patterns that were comparable with those of ACs. However, our tag system accurately detected monoclonal expansion of large clones in these acute samples (Figure 26).

Taken together, these results show that our method enabled accurate analysis of clinical samples with complex clonality patterns. Only our Tag system could detect differences in the clone size of ACs and the different ATL subtypes (Figure 27), which were missed using the shear site method.

Discussion-section-1-B

Impact on Clinical Diagnosis: How does an accurate clonality assessment aid the clinician making therapeutic decisions?

Although the preliminary results have limited numbers of analyzed samples, this initial data suggests different clonality patterns that are specific to AC and to the different subtypes of ATL. Despite similar PVLs, ACs could be distinguished from patients with SM ATL clonality patterns (polyclonal vs. a shift towards oligoclonal for AC vs. SM, respectively). The AC clones showed a uniform distribution pattern with no large differences in clone size; however, clones of the SM types had non-uniform sizes. Chronic subtypes showed expanded oligoclonal patterns, with a large shift to monoclonality, and all acute samples harbored a large expanded clone with a high

absolute number of infected cells. The clonality pattern of the chronic samples was more similar to the acute than the smoldering types.

Because of their diverse prognosis and clinical manifestations, ATL patients are categorized into distinct subtypes based on standard clinical criteria: presence of organ involvement, leukemic manifestation, and levels of lactate dehydrogenase (LDH) and calcium [72]. Currently, distinct treatment strategies are used for the different subtypes of ATL. Therefore, classifying ATL patients into distinct subgroups is of a high importance for selecting appropriate therapeutic interventions [16, 72, 73]. Considering such an intimate link between ATL diagnosis and treatment, a more robust classification of ATL subtypes mediated by the HTLV-1 clonal composition is of fundamental clinical significance. Further examination of the clonality patterns using large numbers of samples is necessary to confirm the relationship between clonality patterns and ATL subtypes, and to apply these clonality patterns to diagnosis and treatment.

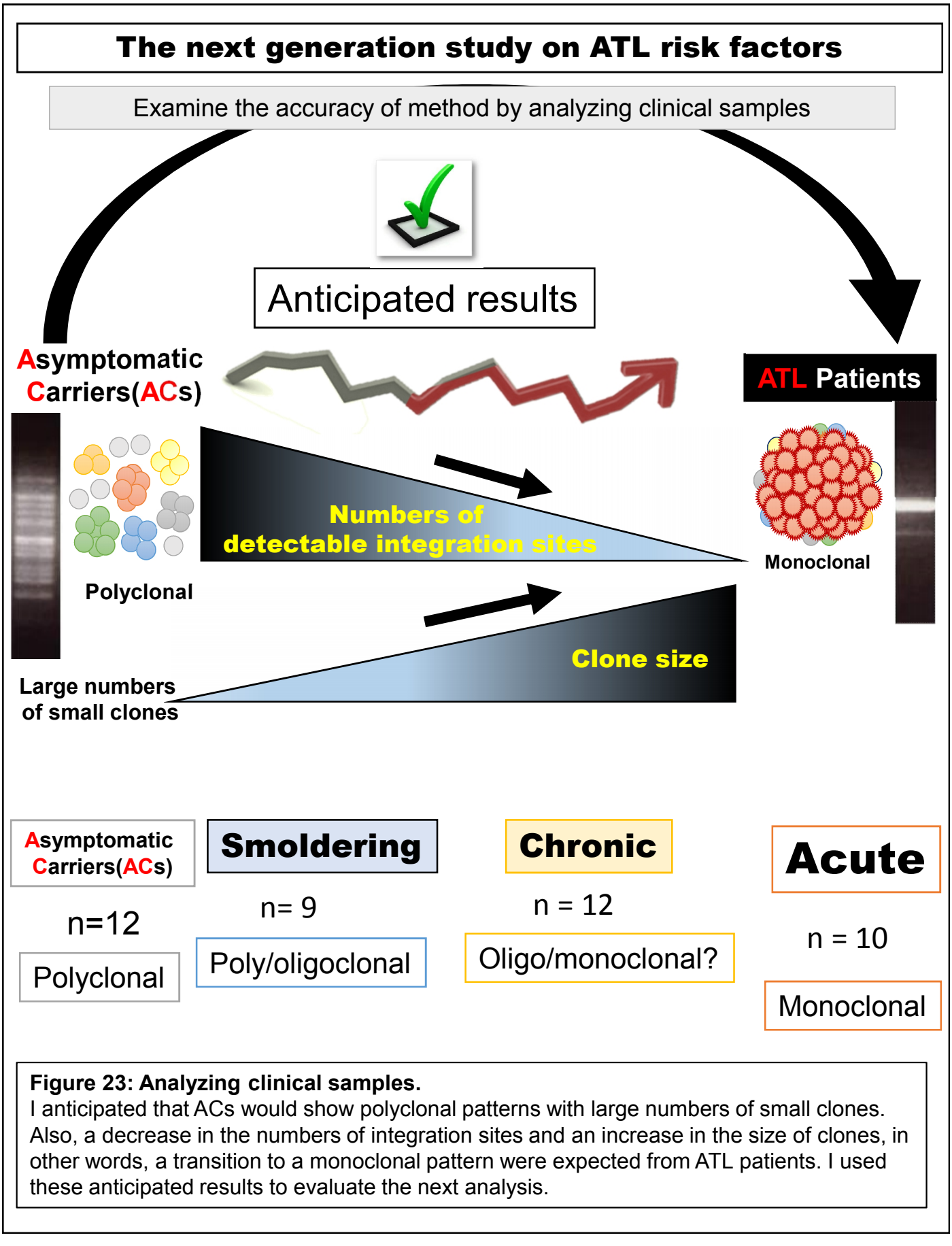


Figure 23: Analyzing clinical samples.
I anticipated that ACs would show polyclonal patterns with large numbers of small clones. Also, a decrease in the numbers of integration sites and an increase in the size of clones, in other words, a transition to a monoclonal pattern were expected from ATL patients. I used these anticipated results to evaluate the next analysis.

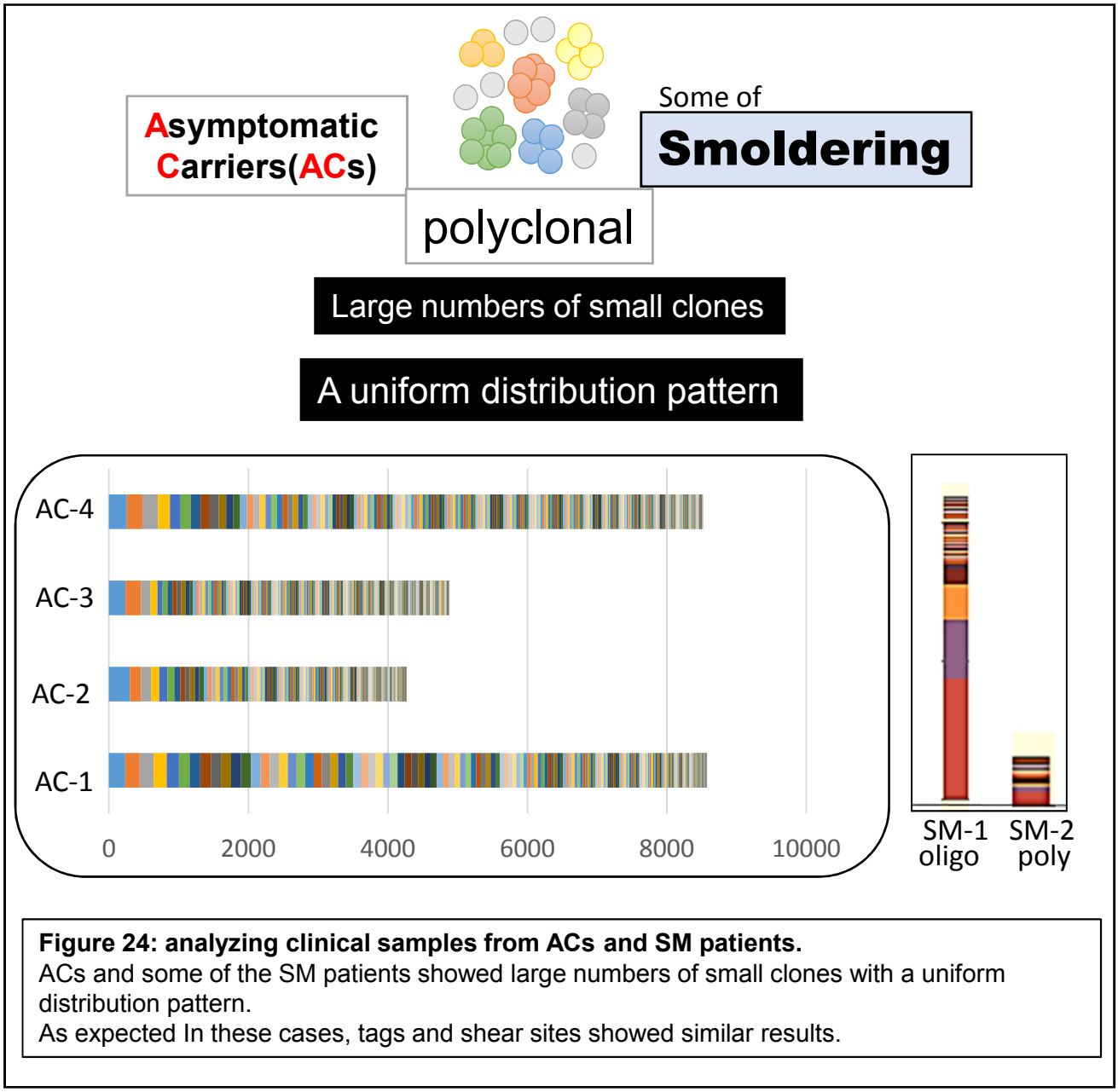


Figure 24: analyzing clinical samples from ACs and SM patients.
ACs and some of the SM patients showed large numbers of small clones with a uniform distribution pattern.
As expected In these cases, tags and shear sites showed similar results.

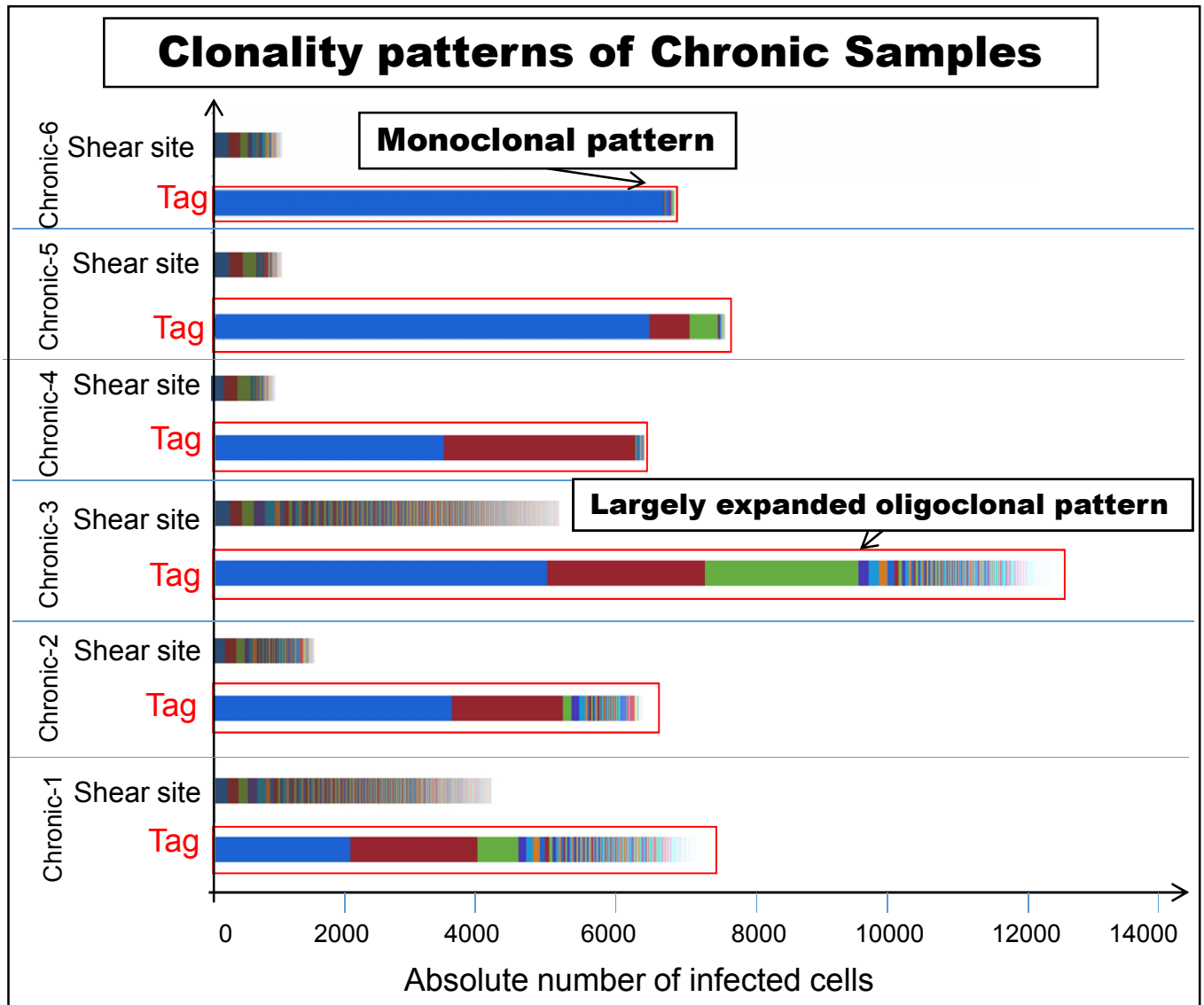


Figure 25: analyzing clinical samples from chronic patients.
In the case of chronic samples, because shear sites underestimated the size of clones, The clonality patterns were so similar to those of ACs. However, by our tag system, monoclonal patterns or largely expanded oligoclonal patterns were detected from chronic samples.

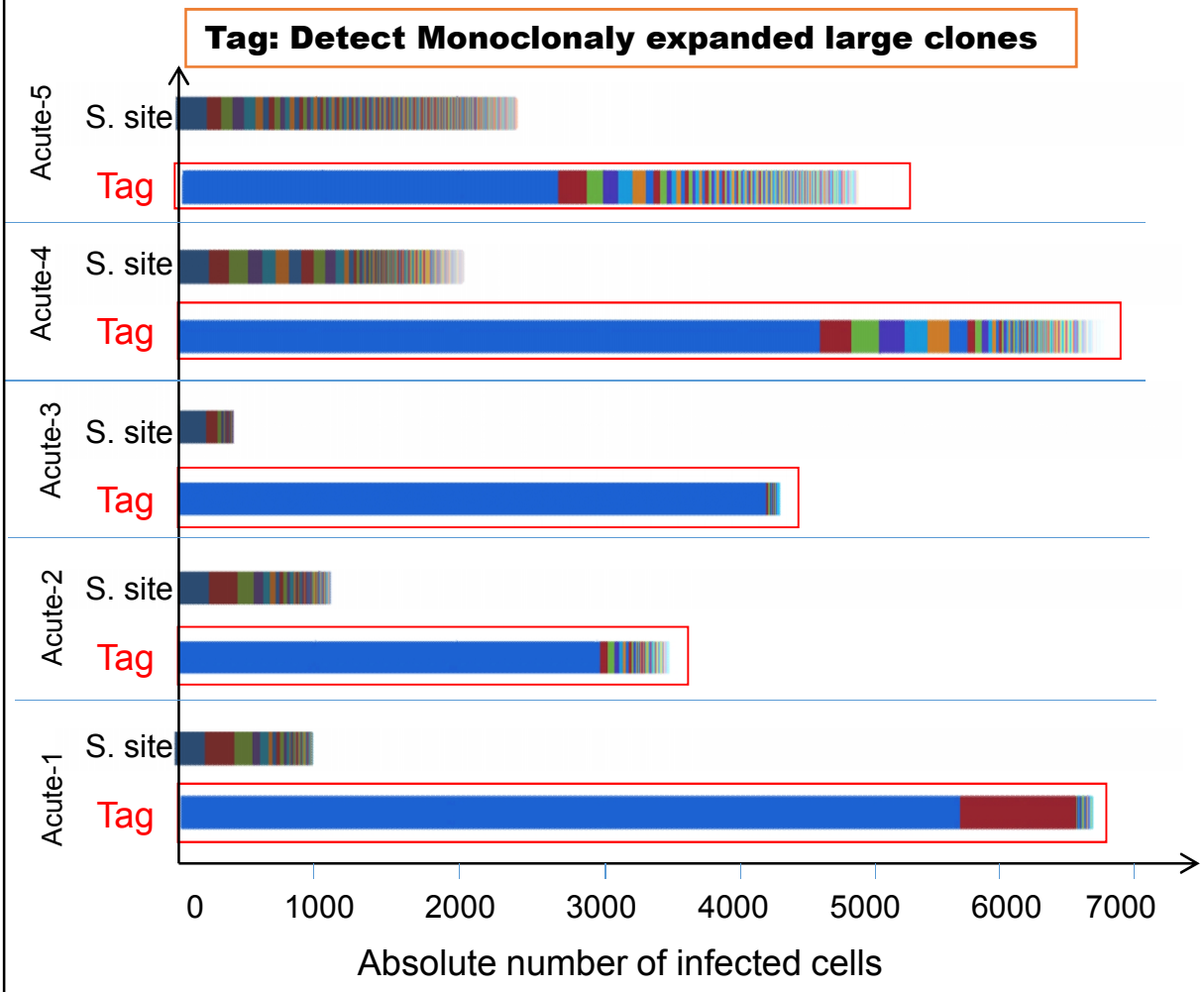
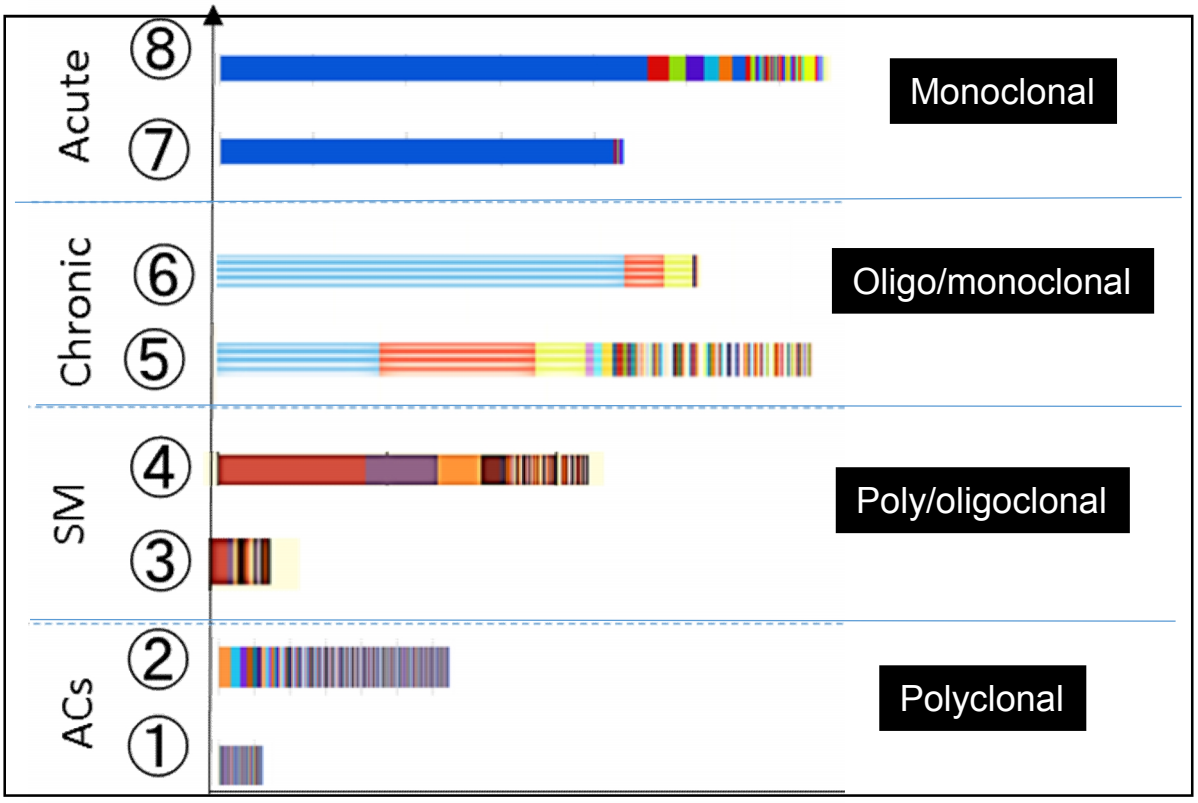


Figure 26: analyzing clinical samples from acute patients.
Also In the case of acute samples, the size of clones was underestimated by shear sites, and clonality patterns were comparable to those of ACs. However the tag system accurately detected monoclonaly expanded large clones in Acute samples.

Clonality patterns demonstrated by my tag system

By accurate measurement of clonality
Only Tag system could detect differences in the clonality patterns of ACs and different ATL subtypes



My method enables accurate analysis of clinical samples with complex clonality patterns

Figure 27: A summary of clonality of ACs and different subtypes of ATL.
only the Tag system could detect differences in clone size of ACs and different ATL subtypes.

Results-section-1-C

Examination of the accuracy of my new method by monitoring of clonality alterations over time

To examine the accuracy of my new method for monitoring clonality, I analyzed four sets of rare sequential samples over time. Here I present these initial results that include the following sets of patients:

- Set 1: ACs with no change in clinical status.
- Set 2: SM patients who remained SM over time compared to those who progressed to a chronic state
- Set 3: One AC who changed to acute ATL
- Set 4: Patients before and after medical therapy

In ACs with no change of clinical status (set 1), the clonality pattern remained polyclonal over the time course of analysis. Thousands of small clones were detected at each time point, and hundreds of observed clones were identically detected over both time points (Figure 28).

In set 2, I analyzed two groups of patients with SM ATL, some who progressed and some who remained progression free. Similar clonality patterns were detected by the shear site method from both groups. However, my tag system found largely expanded clones that were specific to the progressed smoldering samples (Figure 29).

In set 3, I sequentially monitored the clonality of an AC that changed over to chronic and eventually acute ATL over a time course of 6 years. The major clones of the leukemic state were already dominant in the early state of being AC. The clonality pattern of this patient in the AC state was significantly different from that of a typical AC, suggesting that the early detection of an expanded clone may help to predict the prognosis of infected individuals (Figure 32).

In addition, I monitored the effect of therapy on the dynamics of clonality (set 4). I could detect both stable clones and fluctuating clones after chemotherapy (Figure 31).

In summary, monitoring of clinical samples over time suggested the importance of my method in generating biologically meaningful information.

Set 1 detected large numbers of small clones that survived over time in patients that remained ACs, while Set 2 confirmed the presence of large clones only in samples that showed progression. Set 3 allowed for the earlier detection of a leukemic clone and its effect on the prognosis of infected individuals, while Set 4 revealed the presence of stable and fluctuating clones after therapy.

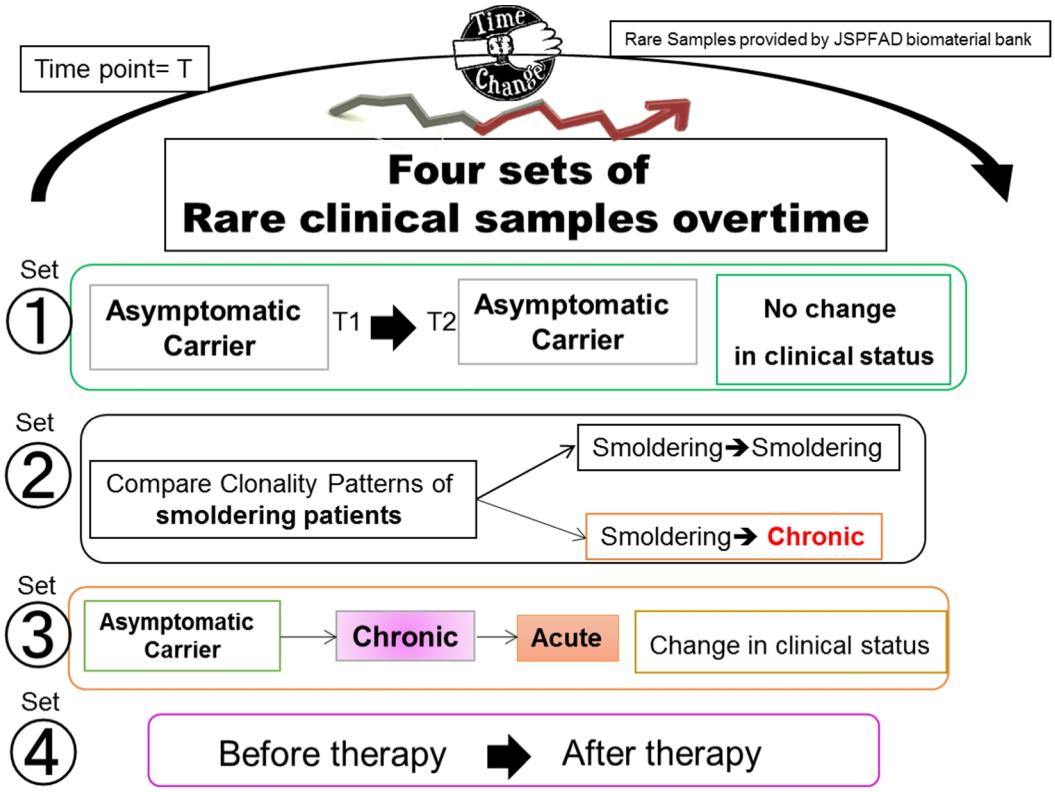
Discussion-section-1-C

Impact on Prognosis and Prevention: Demand for an effective prognostic indicator of ATL onset

In a pilot study, I obtained data from four SM patients over a time course of 4 years. Two of the samples were without progression in disease status (t1 = SM, t2 = SM), while the other two samples progressed into the chronic stage over time (t1 = SM, t2 = Chronic). I detected a significant difference between the clonality patterns of the two groups independent of their PVL. Progression-free samples manifested polyclonal or oligoclonally expanded clones with low numbers of infected cells, while the progressed samples manifested a monoclonal or largely expanded oligoclonal pattern. I also examined the effect of therapy on clonality patterns and the prognosis of different patients. For this purpose, I monitored the clonality patterns of patients

before and after treatment. Most of the samples harbored a stable major clone before and after relapse. However, I could find fluctuating large clones (change in order and size of some large clones) in some chronic patients before and after treatment.

Although still preliminary, this data suggests that the clonality patterns can potentially be used to evaluate the prognosis of patients. Thus, I recommend expanding this into a large-scale genomic project. This analysis may be helpful in clinical decision-making, such as determining the most appropriate and timely therapeutic interventions, based on the clonality status of patients. The information from previous studies on HTLV-1 clonality and my own data suggest 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 may a risk indicator of progression into ATL. Comparing the clonality patterns of the infected cells of patients who progress from AC to ATL is expected to provide highly critical information on the clonality alterations, associated with the transition from the AC to ATL state. Using the changes in the clonal composition of infected individuals as a prognostic indicator appears to be beneficial for early detection of ATL onset, and, in turn, ATL prevention. Accurate monitoring of clonality patterns among infected individuals may help us to differentiate progressive and non-progressive clonality patterns and assessments of the risk of disease development. For this purpose, I propose undertaking a cohort study, similar to that of JSPFAD on PVLs, to be conducted on the clonality patterns of sequential samples from individuals over time.



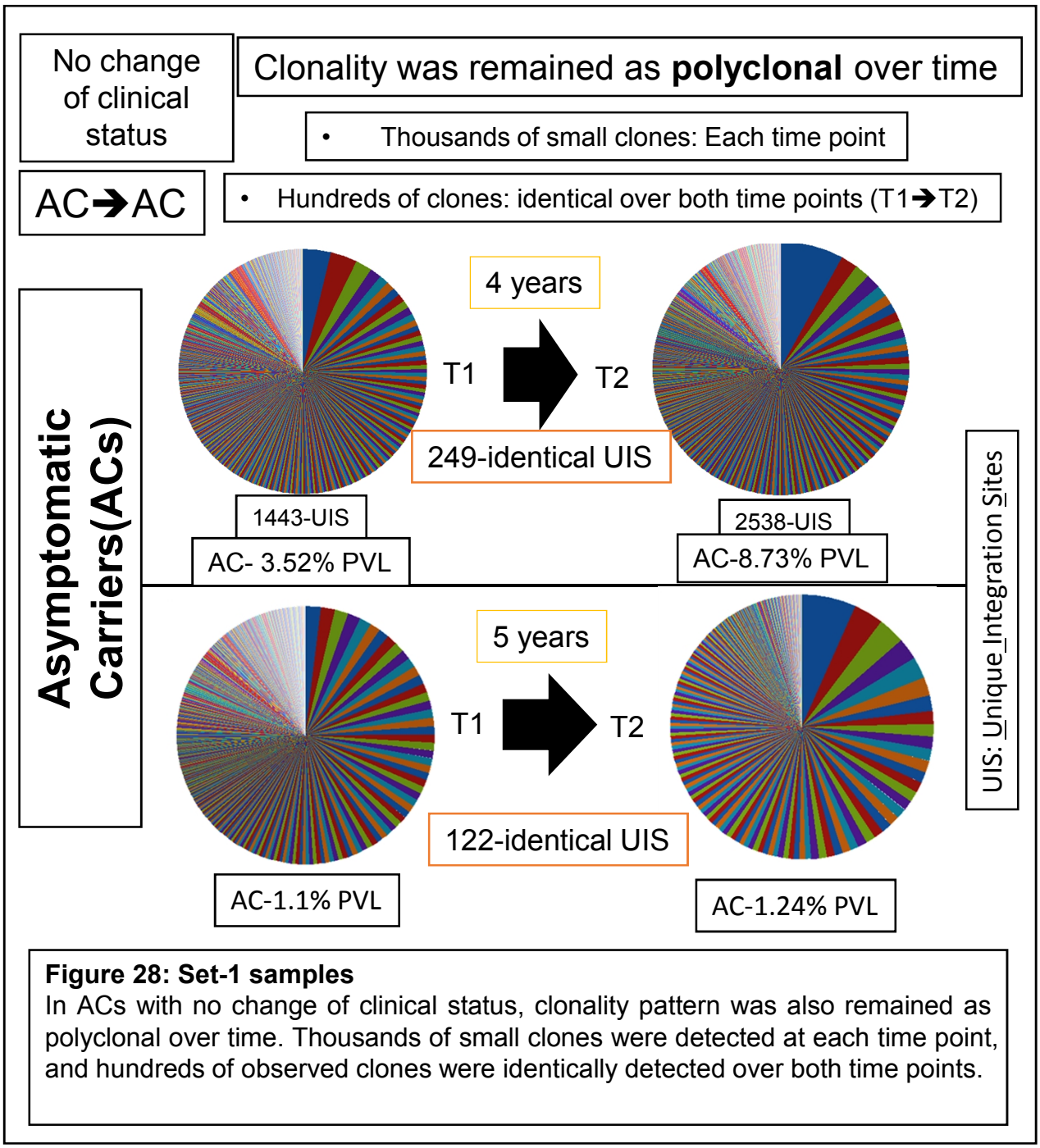


Figure 28: Set-1 samples
In ACs with no change of clinical status, clonality pattern was also remained as polyclonal over time. Thousands of small clones were detected at each time point, and hundreds of observed clones were identically detected over both time points.

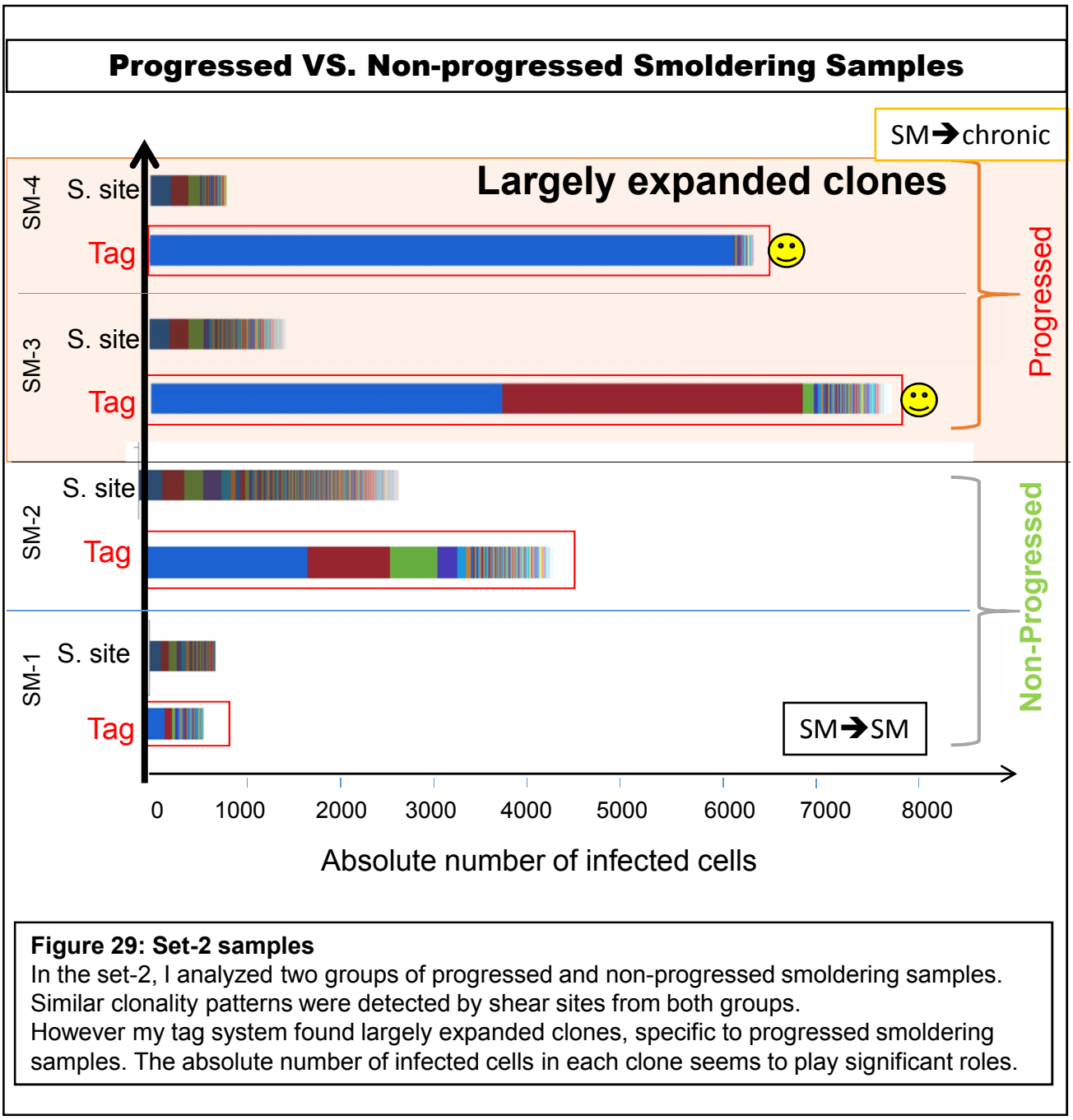


Figure 29: Set-2 samples
In the set-2, I analyzed two groups of progressed and non-progressed smoldering samples. Similar clonality patterns were detected by shear sites from both groups.
However my tag system found largely expanded clones, specific to progressed smoldering samples. The absolute number of infected cells in each clone seems to play significant roles.

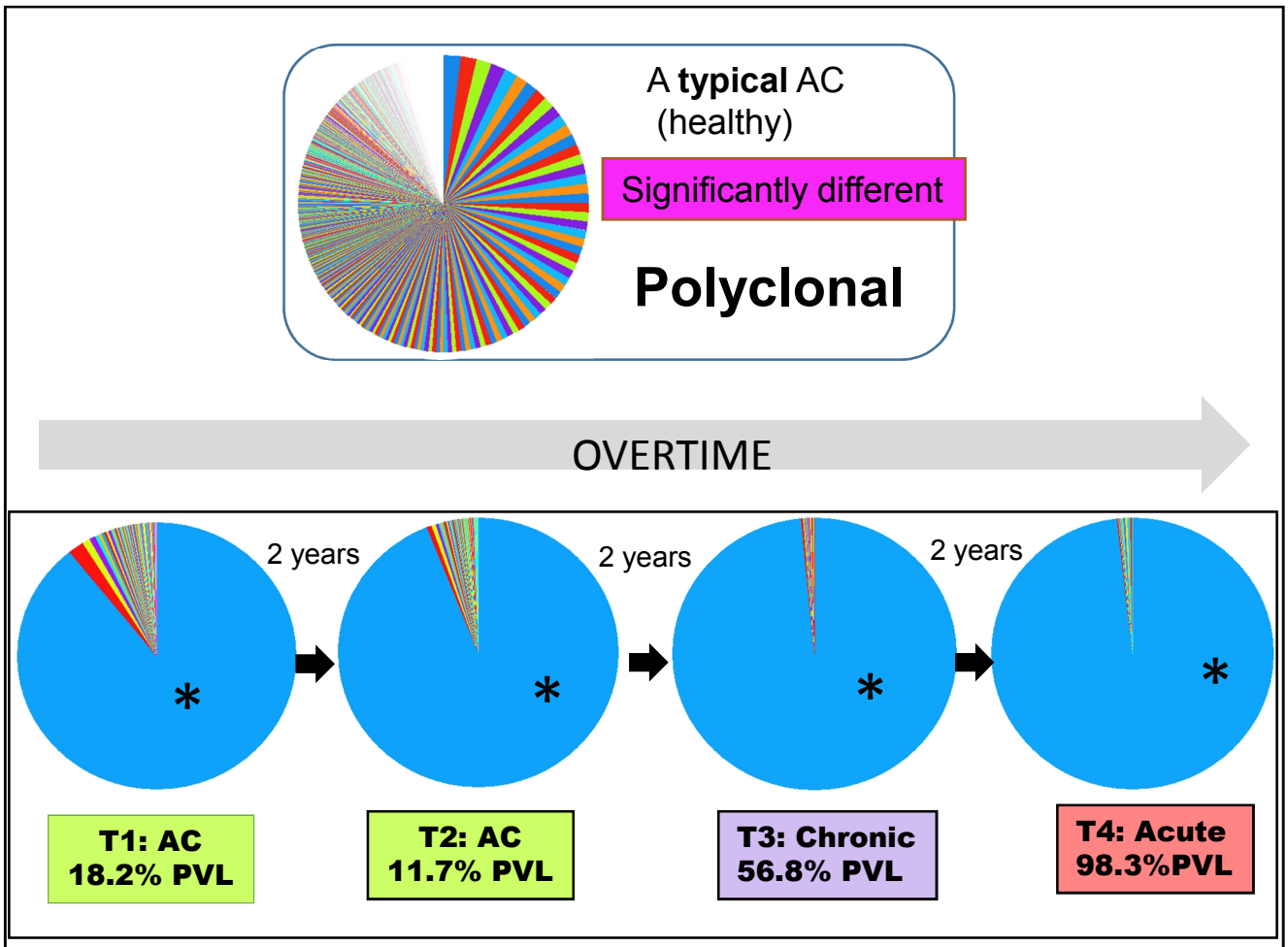


Figure 30: Set-3 samples
In the set-3, I sequentially monitored clonality of an AC who changed to chronic and eventually Acute ATL over a time course of 6 years.
The major clones of leukemic state were already dominant in early state of being AC.
The clonality pattern of this patient in state of AC was significantly different from that of a typical AC. Such kind of earlier detection of expanded clone and may help to predict prognosis of infected individuals.

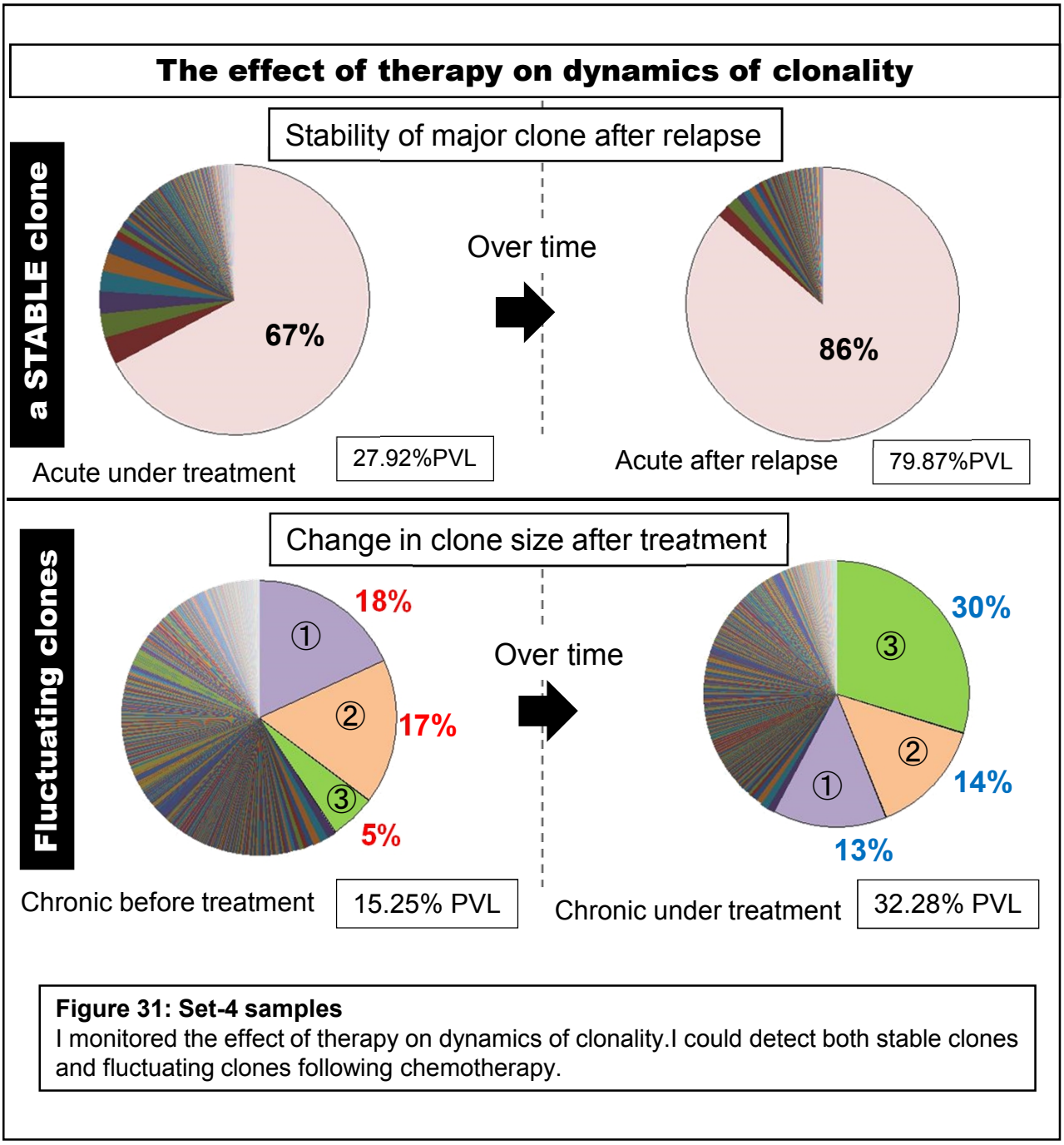


Figure 31: Set-4 samples
I monitored the effect of therapy on dynamics of clonality.I could detect both stable clones and fluctuating clones following chemotherapy.

Results -Section-2

Following analysis of samples for their clonality, based on the integration sites of the provirus, I proceed to more in-depth characterization of clones based on their mutation patterns.

My initial data on integration of site-based clonality suggested that a number of mapped reads of the human flanking regions in ATL patients was lower than that in ACs, particularly in the acute ATL samples, where a large clone typically occupied >97% of the PVL. The number of reads with at least one reported alignment was 142618 (3.57%), and that of reads that failed to align was 3857382 (96.43%) (Figure 32). To find the reason for the non-mapped reads, I used this sample for further analysis. I took the human flanking regions before mapping and clustered by cd-hit, then performed multiple alignments and constructed a phylogenetic tree, based on the mutation patterns of the flanking human region of integration site data (approximately 50–100 bp in length) (Figure 32). Although the data are still preliminary and need further confirmation, I think that there are intraclonal mutations associated with each clone. ATL is known for harboring chromosomal abnormalities; however, comprehensive characterization of these mutation patterns has not been conducted [74, 75]. With these new insights, I decided to clarify the mutation patterns associated with the cells in each specific HTLV-1 clone.

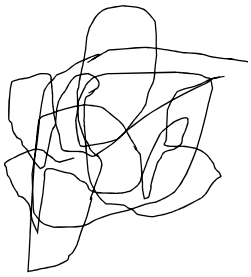
For this purpose I am going to take advantage of my method for clonality analysis, based on integration site data, and link that data with the mutation profiling of each clone. The process I will use is outlined in Figures 33 and 34.

To retrieve the genomic abnormalities, associated with each clone, I will use a combination of long and short sequencing reads. Long sequencing reads allow for coverage of larger distances from the site of integration, and are necessary for my analysis. For this purpose, I will use Nanopore technology, a third generation sequencing technology that generates long reads up to 50–100 kbp. To overcome sequencing errors, I will combine the long reads with short reads of high sequences as shown in Figure 33. From this, I will reconstruct the genome at a single cell level by overlapping sequencing reads. Before conducting any amplification, I will incorporate barcoding tags to each DNA fragment to remove amplification bias and retrieve the original frequency of starting fragments.

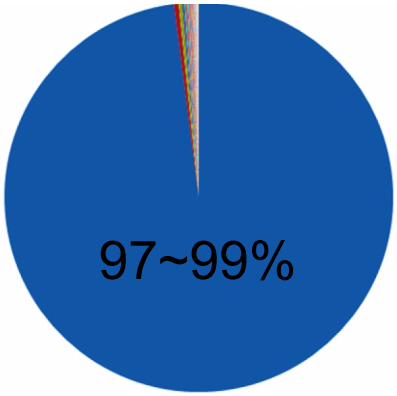
To capture the integration site of interest, I need to design a custom sequencing system. Because a long captured target DNA is also necessary, commercially available kits were not suitable. Companies providing capturing services can only capture fragments of 500 bp; therefore, I have designed my own capture system as described in Figures 33 and 34. I am now ready to proceed with the actual analysis of the samples.

A

Integration site analysis



Bulk gDNA



Acute (51.8% PVL)

Integration site at
chr 7 (-) 9408532

Independent experiments: gDNA and MDA product

reads with at least one reported alignment: 142618 (3.57%)
reads that failed to align: 3857382 (96.43%)

B

Mutation profiling of Human Flanking Region(50-100bp)

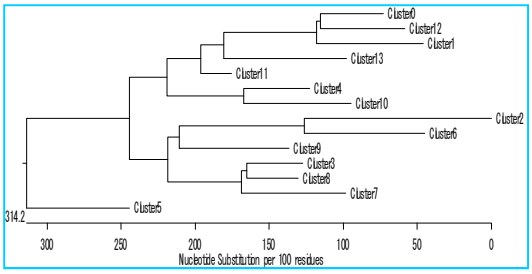
Construct phylogenetic tree based on the mutation patterns of flanking Human region of integration site data (~up to 100-bp length).

clustering the target sequences by cd-hit software → multiple alignment of clusters by ClustalW

```
Majority ---------XXXXXXACXATXTGAGAACXXXTXTGAAXCXAAAXXTXXXXXX
         ---------+---------+---------+---------+---------+--
                  10        20        30        40        50
         ---------+---------+---------+---------+---------+--
Cluster0 -------ACTATGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Cluster1 --ACTATGCTATTGGTCTATGTGTCTATAGATCGGAAGAGCA
Cluster2 -------AATCAAGGGGGAAGCCAGGGTGCCTCCTAAGCTGCTTACC
Cluster3 ACATGTCCACCCCACACCCTGCGTTAGCTACTATAAATAA
Cluster4 ------------AGTAATGTGCTTGAATCATCCCGACACCATCACTACCCTC
Cluster5 -AGTTGGGGCGTCGTACCGCCTGTGAGCGCGCCTTTAGTCG
Cluster6 -------ACACGGCTGAACTCCAGGCACATCTGGGGACCTCGTATGC
Cluster7 -----------AATTATCCTCTTGTAGATATCAGGTATGTCTGATTGTTTT
Cluster8 ----------CAGATACCCAGTGAAAGTAAATATGCATTTACCCTTTTGC
Cluster9 ----ACCGCGCCATCACGGGATGGGAATTACTGTGGGGCTAGAT
Cluster10 ------CGGGAGAAGAAAACACAAAAACTCACAGGCTACAATGACA
Cluster11 ----------AAGGGACTACATTTAAATTCATCTGAAACATAAGGTTTCG
Cluster12 -------AATAGGACACTACTGAAGTACAGTAAAGGACACAAATCTA
Cluster13 ---------ATTTTCCCTTTATCAAAACGTTTATACAAACAAACATATC
```

→ cunstruct phylogenetic tree

Signify the **relationship** and **distance** between the target sequences



Adapt the workflow for analysis in a large genomic scale

Infer evolutionary relationship among ATL clones

Figure 32: integration site analysis of a patient with a very low mapped reads
(A) Patient information, integration site and clone size data, and mapping yield.
(B) Mutation profile of this sample and phylogenetic tree of mutation variations.

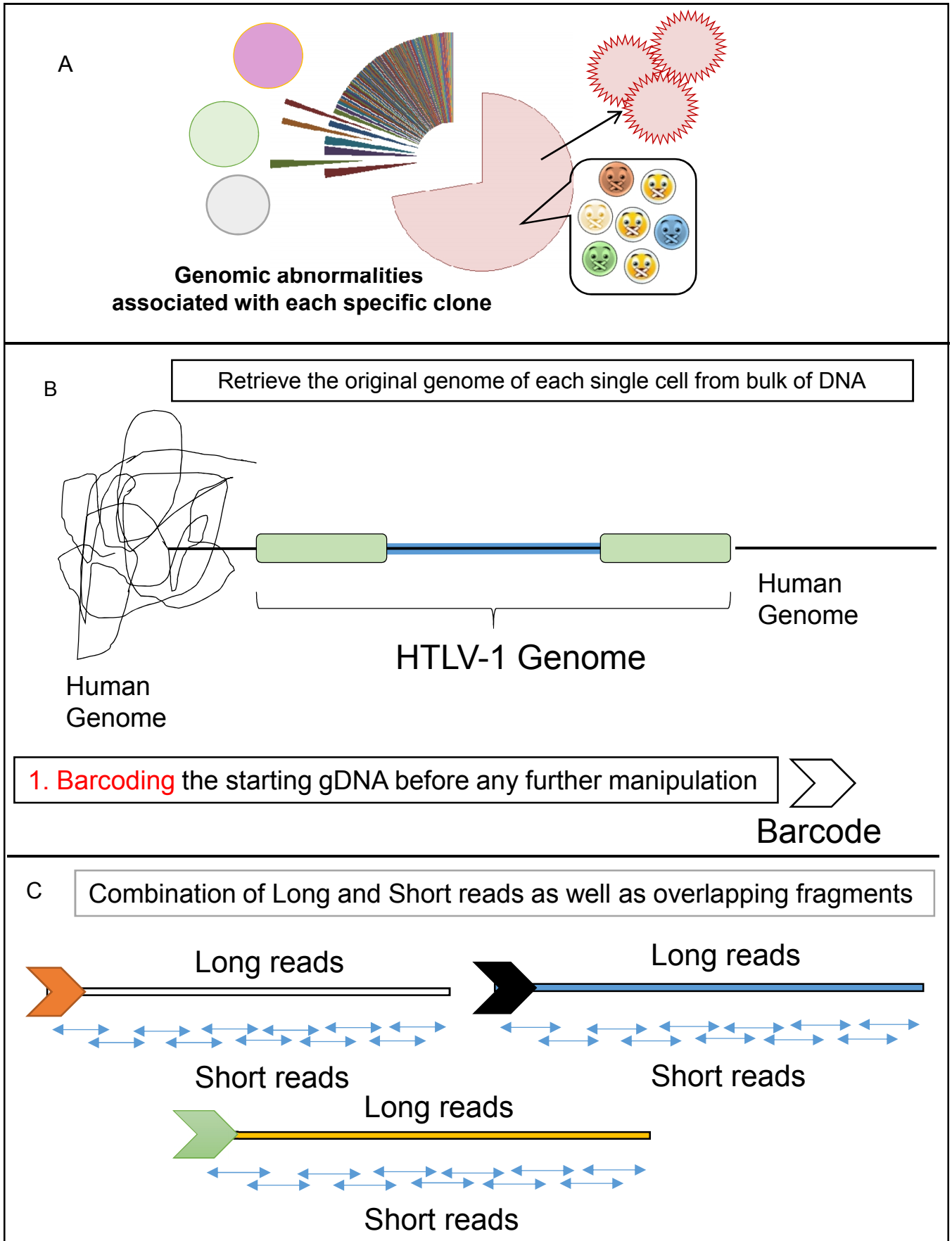


Figure 33: How to investigate genomic abnormalities associated with each specific clone

D

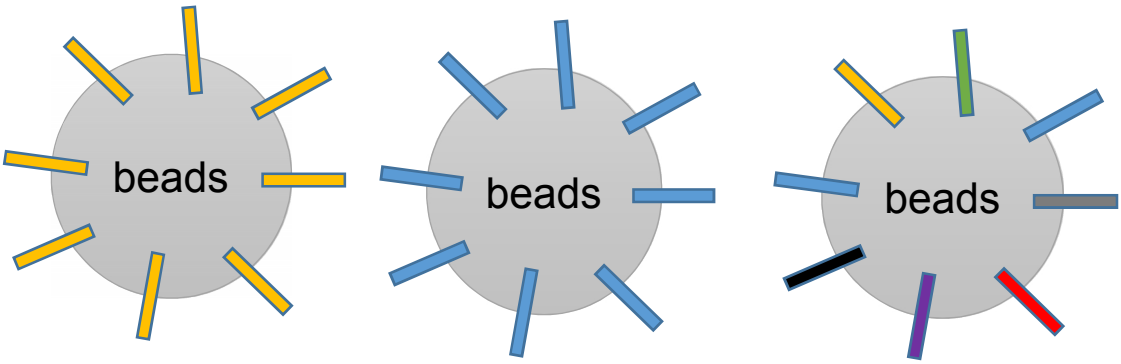
Retrieve the original genome of each single cell

Link to data of integration sites

Capture the integration site of interest

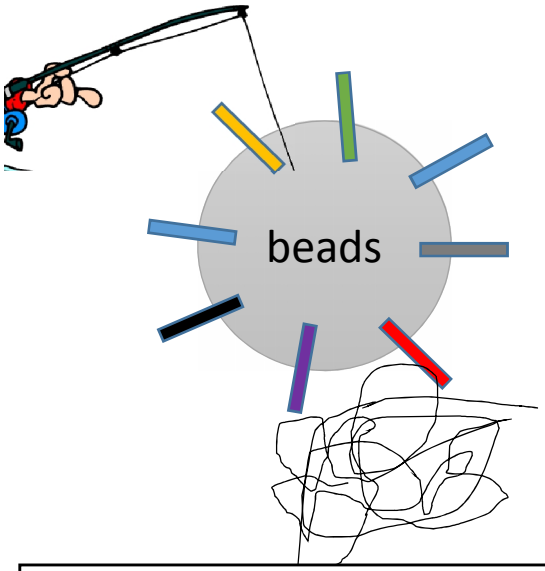
E

Custom capture system with specific beads



Homo-probe beads

Hetero-probe beads



- The size of Beads
- The lengths and design of baits
- The length of target DNA
- Immobilization strategy

Capture target fragments as long as possible

Capture Genome of each specific integration site

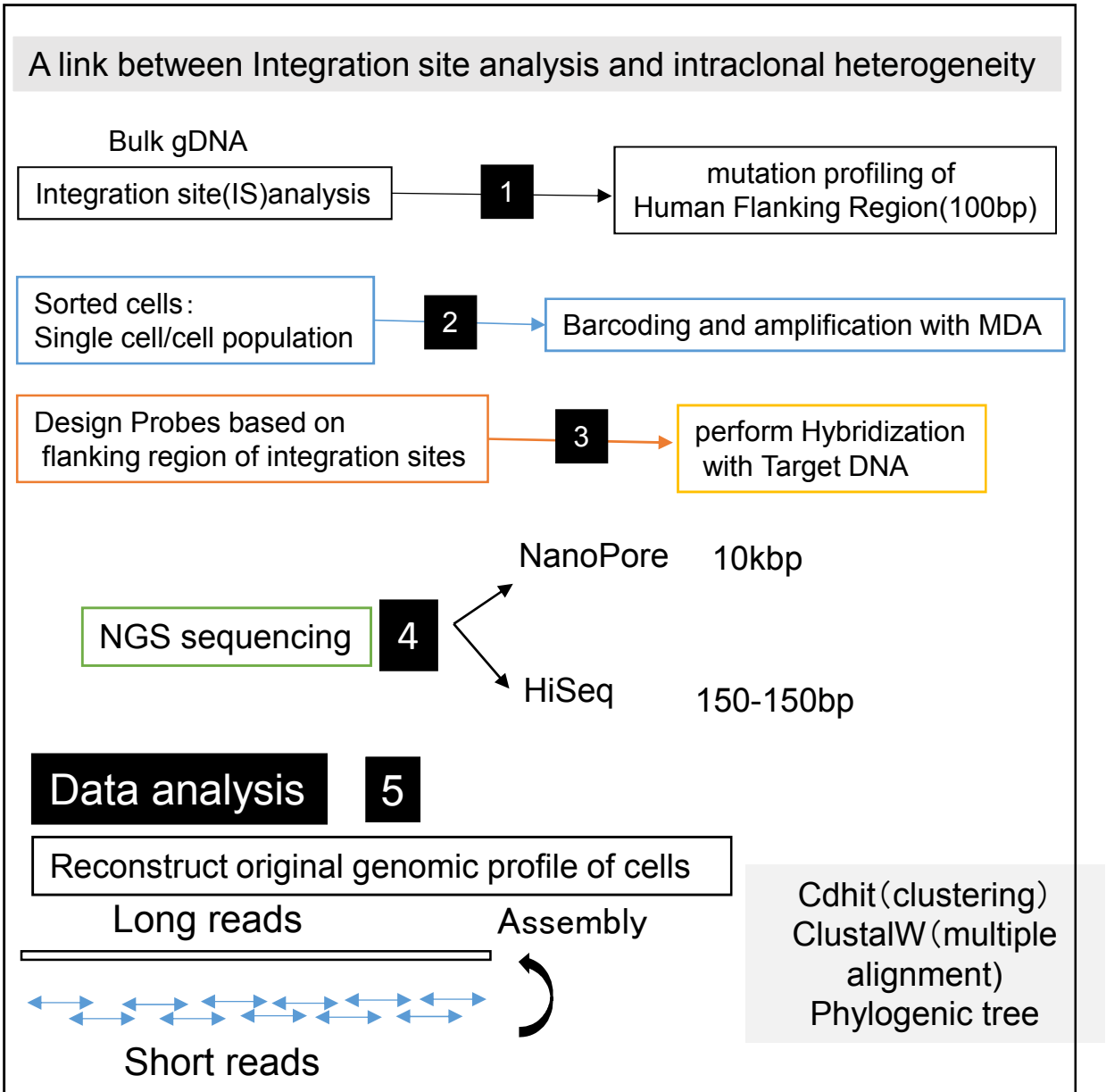


Figure 34:
workflow of making a link between integration sites data and mutation profiling of each clone

Final Discussion

I approached to the concept of clonal expansion in cancer while using ATL as an appropriate model. I employed HTLV-1 integration sites as markers to uniquely characterize every single infected cell in each clone. My novel method of assessing clonality (the tag system) has enabled accurate measurement of the absolute numbers of infected cells in each clone (clone size) for the first time.

The two main characteristics of ATL cells include the site of integration and mutation profiling of the cells, which allows for monitoring the clonal composition of ATL from two different perspectives. In the present thesis I mainly investigated integration site mediated clonality analysis which provided enough and accurate information to further elucidate a clear image from clonal composition of ATL. Using this system of accurate monitoring, I was able to detect clonal dynamics and alterations of clonality sequentially over time as well as at independent time points. This accurate and comprehensive information is expected to lead to a more accurate molecular diagnosis or aiding in the prediction of prognosis for HTLV-1 infected individuals, as well as clarifying the mechanisms underlying multistep leukemogenesis of ATL.

I have provided details of the method design, optimized experiment protocols, and *in-silico* data processing workflow. I published the first part of data including development of methodology and its validation by analyzing eight control samples with known integration sites and clone sizes, and four clinical samples. (Firouzi et al)[29]. I 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. I proved my 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, it enabled accurate measurements of the clone size. 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 (polyclonalvs. 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, I 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 confirmed by another method (PCR-southern) (detailed information is provided in results-section 1). Samples with distinct disease status (AC, SM, and acute) manifested different clone sizes, 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 (the index that introduced by Gilet et al). Therefore, it can be inferred that, with an accurate measurement of clone size (particularly absolute number of infected cells), the application of this method will aid in the discrimination of ATL subtypes.

Moreover, taking advantage of my new method to detect clonality (Firouzi et al. Genome Medicine)[29], I further analyzed a greater number of samples from asymptomatic carriers and different the subtypes of ATL. ACs and some SM patients showed large numbers of small clones with uniform distribution patterns, while the absolute number of infected cells in analyzed ACs was significantly lower than other subtypes. In most cases, the number of infected cells in the largest clones of ACs was >300 infected cells. Samples from ACs showed large numbers of integration sites, ranging from 100 to 5000. Independent of PVL, the numbers of isolated integration sites were intrinsic to each sample. In general, the number of detected integration sites was higher in ACs than in ATL patients.

In the case of SM patients, polyclonal or oligoclonal patterns were detected, with absolute numbers of infected cells more than ACs but less than chronic samples. Chronic samples showed expanded oligoclonal patterns or monoclonal patterns, while the absolute number of infected cells was typically less than acute samples. Acute patients showed monoclonal patterns, with most harboring a single large clone among a background of very small clones. The numbers of detected background small clones differed depending on each sample.

In samples whose absolute numbers of infected cells exceeded 250, which is the theoretical limit of the number of different shear sizes able to be produced, the shear site method (Gillet et al.)[18] underestimated the clone size, while my new tag system more accurately measured the size of clones. Consequently, the relative measurement and statistical estimation of the clone size by Melamed et al. and Gillet at al. was unable to find differences in clonality among the different subtypes of ATL, while my method allowed for discrimination between ACs and the different subtypes of ATL. Therefore, with an accurate measurement of clone size, my method is able to aid in the discrimination of ATL subtypes. The results presented in this thesis suggest a possible association between disease status and clonality patterns. Hence HTLV-1-infected individuals may be able to be classified into different groups, based on their clonality patterns, ultimately affecting their diagnosis, choice of therapy, and prognosis.

To further validate my methodology and its accuracy, I analyzed rare sequential samples that were collected over time. The four sets of sequential samples, which were analyzed, included patients with AC and no change in disease status (set 1), SM samples with progression and without progression in disease status (set 2), samples from a patient initially with AC who progressed to ATL over time (set 3), and samples from patients before and after therapy (set 4).

I detected a large number of integration sites from ACs over time. Even with low PVLs averaging 1%, my method detected hundreds to thousands of integration sites. Each infected individual displayed a proportion of clones that were constantly detected over time, while new integration sites were also detected over time. These new integration sites may be because of newly emerging clones generated from new infections; however, this still needs further validation.

I detected dynamic clonal alterations in samples of patients with SM, chronic, and acute ATL. While some clones kept growing over time, some were undetectable or lost, and some clones shrunk, while others remained stable. If the newly detected integration sites were isolated accurately by a reproducible manner, it can be suggested that they are newly emerging clones generated due to new infections. Sequential detection of common clones or newly isolated clones in ACs and the different subtypes of ATL only suggest the possibility for the presence of

both persistent and newly emerging clones. However, to have a correct conclusion, this issue should be further examined and validated.

I analyzed samples from patients with SM ATL; one subset of these patients progressed over the time course, while the other subset remained stable. Comparison of the clonality measures of the shear site strategy (Gillet et al.) with my new tag method (Firouzi et al.) showed that the shear sites strategy could not detect differences in clone sizes between these two groups, while I was able to observe different clonality patterns from progressed and non-progressed SM patients. Non-progressed SM patients who remained SM over time showed polyclonal or oligoclonal patterns with low numbers of infected cells in each clone. Progressed SM patients who changed over to chronic ATL showed largely expanded oligoclonal or monoclonal patterns with high numbers of infected cells in each clone. These data suggest that measurement of clonality may be useful in predicting the prognosis of HTLV-1 infected individuals.

In this study, I was also able to analyze the rare samples from an AC who finally progressed to acute ATL over the time span of 6 years. The major clone detected during the acute state was already dominant when the patient was AC. The clonality pattern of this patient at AC was abnormally monoclonal, differing from typical ACs who have polyclonal patterns. From these data, I inferred that accurate measurement of the clone size may help to predict the probability of progression years before developing ATL. In addition, although the size of clones was abnormally large when the patient was in AC stage, it took approximately 6 years to develop into the acute state. This suggests the possibility of additional hits necessary to develop into ATL. I hypothesize that expansion of a clone increases the probability of accumulating genetic abnormalities within the expanded clone.

Sets of samples before and after therapy were also analyzed in this thesis. In the first case, the most abundant clone did not change in size following therapy, but was stable over time. In case 2, the abundance and order of clones changed after treatment, suggesting the presence of both stable and fluctuating clones.

Taken together, my data shows that an accurate measurement of clone size is essential to obtain biologically significant information from the clonal composition of infected individuals.

Provirus integration site-mediated clonality analysis holds the promise to allow us to reach the final goal of demonstrating the genetic abnormalities, associated with individual clones. In analyzing information from ACs and ATL samples, I found that many acute samples had low mapping quality, despite identical sequencing read and analysis conditions with other samples. Using multiple alignments of these reads, I found similarities that led us to hypothesize that the low uniquely mapping reads in the acute samples is due to a high frequency of mutation around the site of integration.

ATL is an aggressive malignancy with many chromosomal abnormalities reported [74, 75]. Because I found mutations even in the flanking regions of integration sites, I hypothesize that mutation rates must be high in the genome of ATL cells. To further investigate this, I propose to combine my current method for clonality analysis using on provirus integration sites with a mutation profile of each specific clone obtained by target sequencing, exome sequencing, and whole genome sequencing. This analysis will enable us to further characterize ATL clone, and holds promise of clarifying the multiple steps leading to development of ATL; and answer the question that “When, where and at which step clonal expansion and transformation occurs?”

In conclusion, I 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. I also discussed the novelty, significance, and applications of my method, and compared it with the only existing high-throughput method devised by Gillet *et al.* [18]. Employing our new methodology and the analysis of an appropriate pool of samples provided by JSPFAD[76] 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 [67] or in metagenomics for determining the size of bacterial populations.

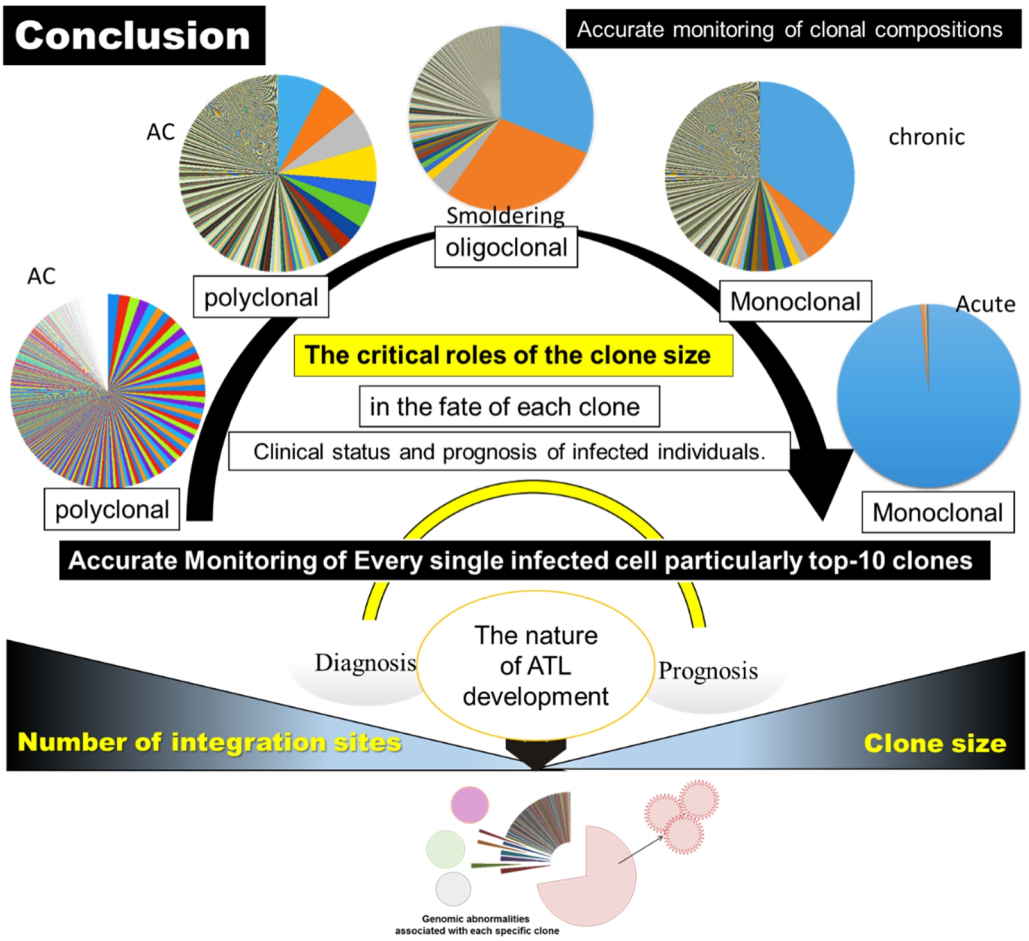


Figure 35. Accurate monitoring of clonal compositions suggested the critical roles of the clone size in the fate of each clone, clinical status and prognosis of infected individuals. Monitoring of clinical samples suggested the importance of our method for generating biologically meaningful information. My data suggested different clonality patterns specific to AC and different subtypes of ATL. In this thesis I proposed clinical application of my methodology for diagnosis and predicting prognosis of infected individuals. I will further validate the observed results and I will also try to reveal intraclonal composition of clones by making a link between IS-based clonality data and clone specific mutation profiles. I hope my data shed light on the mechanism of ATL development and be beneficial for discovering treatment of ATL patients.

Experimental design, Material and Methods

Our clonality analysis method included two main aspects: (1) wet experiments, and (2) in silico analysis (Figure 18-19). Detailed protocols of the wet experiments are included in this section. 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 [76, 77]. 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 [78]. 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 [35].

Cell lines: An IL2-dependent TL-Om1 cell line [38, 79] 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

I employed a library preparation protocol specifically designed to isolate HTLV-1 integration sites. The final products in the library that I generated contained all the specific sequences necessary for the Illumina HiSeq 2000 platform (Figure 36). 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 (Figure 36).

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. I 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).

My library construction pipeline comprised the following four steps:

(1) DNA isolation: DNA was extracted as described above, and the concentration of extracted DNA was measured with a NanoDrop 2000 spectrophotometer (Thermo Scientific). I

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 later in section of DNA fragmentation, 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). (Figure 10B).

(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, I used nested-splinkerette PCR (a variant of ligation-mediated PCR[80, 81] (Figure 36).

I 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 (Figure 38).

Information on oligonucleotides, including adaptors and primers, and the LTR and HTLV-1 reference sequences [82] are provided in Table 4. The final PCR products were sequenced using the HiSeq 2000 platform (Results-section-1-A). Data from Samples of results-section-1-B and -C were obtained by HiSeq 2500 platform.

Wet experiments were performed according the following protocols, and compatible with Illumina. Following Abbreviations were used:

Catalog Number: **#**; Reaction: **rxn**; and Milli-Q water: **MQ**

DNA fragmentation

Shear starting DNA fragments using following equipments and operation settings.

Equipments: Covaris™ S220 System (Applied Biosystems ®) #4465653.

Micro Tube (6×16mm) Round bottom glass tube, AFA fiber with Snap-Cap, Covaris #520045.

Sample: 10 µg DNA in 100 µl MQ. Sonication conditions are similar to Gillet *et al.[18]* [18].

Set the following operation settings for Covaris:

	Cycle 1	Cycle 2
Duty cycle	20%	5%
Intensity	5	3
Cycles per burst	200	200
Time	5 sec	90 sec
Temperature	6-8°C	6-8°C

Check the size of fragments by Agilent 2100 Bioanalyzer Instruments- Agilent DNA 7500 Kit according to the instructions of manufacturer. This product should represent a size range of 300-700 bp (Figure 10B).

End repair and 5'-phosphorylation (for 1 rxn)

End repair converts 3'- and 5'- protruding ends of DNA fragments to blunt ends.

T4 DNA polymerase fills in the 5'overhangs to form blunt ends .It has 5'➔3'polymerase activity and 3'➔5'exonuclease activity, and does not have any 5'➔3'exonuclease function.

Klenow enzyme removes 3'overhangs with a 3'➔5'exonuclease activity. It does not have any 5'➔3'exonuclease activity.

T4 polynucleotide kinase catalyses the transfer of gamma-phosphate from ATP to 5'-OH group of single/double strand DNAs/RNAs. This enzyme phosphorylates the fragments at 5'-end, and makes them ready for ligation reaction.

Set up the following end repair reaction.

Component	μl per tube	Information
Sample	100	The product of Sonication
MQ	43	
T4 DNA polymerase (5U/ μl)	3	Takara (#2040A)
10x DNA polymerase buffer	20	Takara (#2040A)
ATP (10 mM)	20	Takara (#4041)
dNTP mix (25 mM)	8	Invitrogen (#10297-018)
Klenow enzyme (5U/ μl)	1	Takara (#2140A)
T4 polynucleotide kinase (10U/ μl)	5	NEB (#M0201L)
Total	200	Incubate at 20°C for 30 min.

Clean up with PCR purification kit (Qiagen #28104) and elute in 67 μl MQ.

A-tailing (for 1X rxn)

Klenow fragment exo- is the large fragment of DNA polymerase-I with a 5'➔3' polymerase activity without any 3'➔5'exonuclease activity. This enzyme leaves a single base 3'-overhang. Set up the following reaction:

Component	μl per tube	Information
Sample	67	The product of end repair
NEB buffer 2	10	
dATP (1 mM)	20	Takara (#4026)
klenow fragment exo- (15U)	3	NEB (#M0212S)
Total	100	Incubate at 37°C for 30 min

Clean up with Qiaquick PCR purification kit (Qiagen #28104), and elute in 60 μl MQ.

Adaptor ligation

Adaptors were designed compatible with Illumina (Figure 36). Adaptor mixture was prepared as described previously [81]. The sequences of adaptors are provided in Table 4. Split the product of A-tailing into 2 tubes of 30 μl then set up the following reaction in a 500 μl thin-layer PCR tube.

Component	μl per tube	Information
Sample	30	The product of A-tailing
Adaptor mix (25 μM)	4	HPLC purified oligonucleotides
T4 DNA ligase buffer (10x)	5	NEB (#M1801)
T4 DNA ligase enzyme (3U/ μl)	5	NEB (#M1801)
MQ	6	
Total	50	Incubate at 20°C for 2 hours

Clean up with MiniElute PCR purification kit (Qiagen #28004), and elute in 20 μl MQ.

Size selection

Size selection is a necessary sample preparation step to remove adaptors carried over from the ligation reaction. Carried over adaptors work as PCR primers, thus interferes the tag data, and lead to a final tag distribution without any sign of PCR amplification (data not shown).

Perform following size selection steps to remove carried over adaptors:

- Prepare a polyacrylamide gel (10% TBE PAGE: 2 mm thick). Pre run 100 volt for 15 min
- Load 100 bp DNA ladder and 20 μl of samples in the wells.
- Run 160 volt for 60 min
- Stain with Ethidium bromide (EtBr) (100 ml 0.5x TBE + 5 μl EtBr) for 10 min.
- Cut the gel from 200-bp to 1000-bp.
- Slice the gel fragment and inset into a 1.5 ml tube which has a pore at the bottom.
- Centrifuge at 12000 rpm for 8 min.
- Add 700 μl of 1x LoTE buffer and centrifuge at 12000 rpm for 2 min.
- Add 1x LoTE buffer up to 2 ml, and Incubate at 65°C for 15 min in a water bath.
- Apply on the Colum (Costar Spin-X centrifuge TUBE Filters 0.22 μm pore CA membrane, #8160).
- Use Millipore Amicon Ultra Centrifugal Filters, 0.5 ml, 100K # UFC 510024 to purify and concentrate (repeat 4 times, each time 500 μl supernatant).
- Wash the filter twice with 500 μl MQ**.**
- End up with purified DNA in 20 μl MQ.

PCR

Although inverse-PCR has been a widely used approach for isolating integration sites, it is relatively inefficient when compared with other PCR-based approaches. It has been reported that splinkerette PCR has become the most widely accepted technique for the amplification of viral and transposon insertion sites [83, 84]. I used a nested-splinkerette PCR to amplify the junction between the human genome and the HTLV-1 insertion (Figure 36). Information on the sequence of primers is available in Table 4. Perform external and nested PCR as described below

External PCR

Component	μl per tube	Information
Template DNA	20	The product of size selection
Primer **F1** (10 μM)	2	
Primer **R1** (10 μM)	2	
10x buffer I	5	
Accuprime taq high fidelity (5U/ μl)	0.2	Invitrogen (#12346-094)
MQ	20.8	
Total	50	

Set the ramping of thermo cycler to 1.9°C /sec.

PCR conditions	Temperature	Time	Cycling
Initial denaturation	94°C	5 min	1 cycle
Denaturation	94°C	50 sec	25 cycles
Combined annealing & extension	68°C	3 min	25 cycles
Final extension	68°C	10 min	1 cycle
Hold @ 4°C			

Mix PCR products of the same sample 50+50+50 μl total: 150μl.

Clean up with Qiaquick PCR purification kit (Qiagen #28104), and elute in 150 μl MQ.

*(optional) Use 1 μl of 10-fold diluted the external PCR for nested PCR.

Nested PCR

Component	μl per tube	Information
DNA	1	*Input from product of external PCR
Primer **P1F2** (10 μM)	2	
Primer **NFCB** (10 μM)	2	
10x buffer I	5	
Accuprime taq high fidelity(5U/ μl)	0.2	Invitrogen (#12346-094)
MQ	39.8	
Total	50	

Set the ramping of thermo cycler to 1.9°C /sec.

PCR conditions	Temperature	Time	Cycling
Initial denaturation	94°C	5 min	1 cycle
Denaturation	94°C	50 sec	30 cycles
Combined annealing & extension	68°C	3 min	30 cycles
Final extension	68°C	10 min	1 cycle
Hold @ 4°C			

Clean up with Qiaquick PCR purification kit (Qiagen #28104), and elute in 50 μl MQ.

Check the size distribution and concentration of PCR products by Agilent 2100 Bioanalyzer Instruments using Agilent DNA 7500 Kit according to the instructions of manufacturer, and conduct sequencing by Illumina HiSeq platform.

Additional supporting protocols:

Restriction enzyme digestion, adaptor ligation, external PCR, nested PCR, and southern blotting

Restriction enzyme digestion

Set up following reaction:

Incubate at 37℃ overnight (12-16h). Heat inactivate the digested DNA at 65℃ for 20 minutes.

Component	μl per tube	
Genomic DNA	10	2 μg
Sau3A1 enzyme (5U/μl)	4	10 units
10XNEB buffer1.1	4	
MQ	22	
Final volume	40	

Adaptor Ligation

The sequence of adaptors:

Long-strand adaptor (61nt)
CGAAGAGTAACCGTTGCTAGGAGAGACCGTGGCTGAATGAGACTGGTGTCGACACTAGTG
G

Short-strand adaptor (48nt)
GATCCCACTAGTGTCGACACCAGTCTCTAATTTTTTTTTTCAAAAAAA

Contains hairpin and GATC 5' overhang

'Long-strand adaptor 'and' Short-strand adaptor' need to be purified by HPLC.

Adaptor ligation reaction

Component	µl per tube	
DNA	6	Sau3AI-digested product (300ng)
Adaptor mix (25 µM)	1	HPLC purified oligonucleotides
T4 DNA-ligase (20 U µl)	0.5	NEB (#M1801)
10× T4 DNA-ligase buffer	4	NEB (#M1801)
MQ	28	
Final volume	40	

Incubate the ligation reaction at 20 °C for 2 hours.

Clean up with PCR purification kit (Qiagen #28104) and elute in 50 µl MQ.

External PCR

Component	µl per tube	Information
Template DNA	20	The product of ligation
Primer **F1** (10 µM)	2	
Primer **R1** (10 µM)	2	
10x buffer I	5	
Accuprime taq high fidelity (5U/ µl)	0.2	Invitrogen (#12346-094)
MQ	20.8	
Total	50	

Set the ramping of thermo cycler to 1.9°C /sec.

PCR conditions	Temperature	Time	Cycling
Initial denaturation	94°C	5 min	1 cycle
Denaturation	94°C	50 sec	30 cycles
Combined annealing & extension	68°C	3 min	30 cycles
Final extension	68°C	10 min	1 cycle
Hold @ 4°C			

Clean up with Qiaquick PCR purification kit (Qiagen #28104), and elute in 50 µl MQ.

*(optional) Use 1 µl of 10-fold diluted the external PCR for nested PCR.

Nested PCR

Component	μl per tube	Information
DNA	1	*Input from product of external PCR
Primer **F2** (10 μM)	2	
Primer **R2** (10 μM)	2	
10x buffer I	5	
Accuprime taq high fidelity(5U/ μl)	0.2	Invitrogen (#12346-094)
MQ	39.8	
Total	50	

Set the ramping of thermo cycler to 1.9°C /sec.

PCR conditions	Temperature	Time	Cycling
Initial denaturation	94°C	5 min	1 cycle
Denaturation	94°C	50 sec	30 cycles
Combined annealing & extension	68°C	3 min	30 cycles
Final extension	68°C	10 min	1 cycle
Hold @ 4°C			

PCR-southern

I conducted a PCR-southern as described followings.

Sample preparation:

Digest 2μg of gDNA by Sau3AI restriction enzyme according to aforementioned protocol.

Perform adaptor ligation as described above. Amplify the ligation product by 30 cycles of an external PCR using LTR-specific (F1) and adaptor-specific (R1) primers.

F1: TACCGGCGACTCCGTTGGCT

R1: CGAAGAGTAACCGTTGCTAGGAGAGACC

Electrophorese the PCR products on a 3% TAE agarose gel, and then transfer on a nylon membrane (Biodyne® Nylon Transfer Membranes (B) of Pall cooperation) for 6 hours. Wash the membrane by 2x Saline Sodium Citrate (SSC) buffer on a shaker at room temperature for 10 min.

Prepare following probes, and then label them using TaKaRa *Bca*BEST Labeling Kit (cat No. 6046) according to the manufacturer's instructions.

[α-32P] dCTP, 0.250 mCi (NEG-513H) was purchased from Perkin Elmer.

LTR-specific (75-bp)

TGTGTACTAAATTTCTCTCCTGAGAGTGCTATAGGATGGGCTGTCGCTGGCTCCGAGCCAA
CGGAGTCGCCGGTA

Blue clone-specific (chr-x) probe (171-bp)

GGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCA
TTTAGCCAGTCTATATATTTTAAGTGGAAAGTTTAATTCATTTATATTCAAAATCATAATTGAT
ATGTGAATATTTATTCCTGTCATTTTACTAGTTGATTTCTGGTGG

Red clone-specific (chr-14) probe (196-bp)

GCTCACAGTATTAGAGTGGGTTACATTTTAAGTAGAAAAACATTTGGTTATATCATTGTCCTT
ATAGCATGATTCTGACTTATTTGCATAAACAAATATTTATGTTCTTGTTTATGTATTTTTGTAA
AACAATATCTATAGGAAAAGTAGGCCTATCCTATAAACCCCCGGAAGGGAAGGTTGATTCA
GACACAGT

Pre-hybridize the membrane at 65°C for 12 hours and hybridize it at 65°C for 12 hours (on a rotator).

Use 2x10^6 cpm of the labeled probes for hybridization.

Wash the hybridized membrane as followings.

All buffers must contain 0.1% SDS.

2x SSC: at 65°C, 10 min, repeat 3 times

0.5x SSC: at 65°C, 10 min, repeat 3 times

0.2x SSC: at 65 °C, on a shaker water bath for 30 min.

0.1x SSC: at 65°C, on a shaker water bath for 30 min.

Expose the membrane on Carestream Health X-OMAT AR (XAR) Autoradiography Film (KODAK 1651454) for about 3-4 hours. Process the exposed film using FPM 800A, Fuji Film instrument.

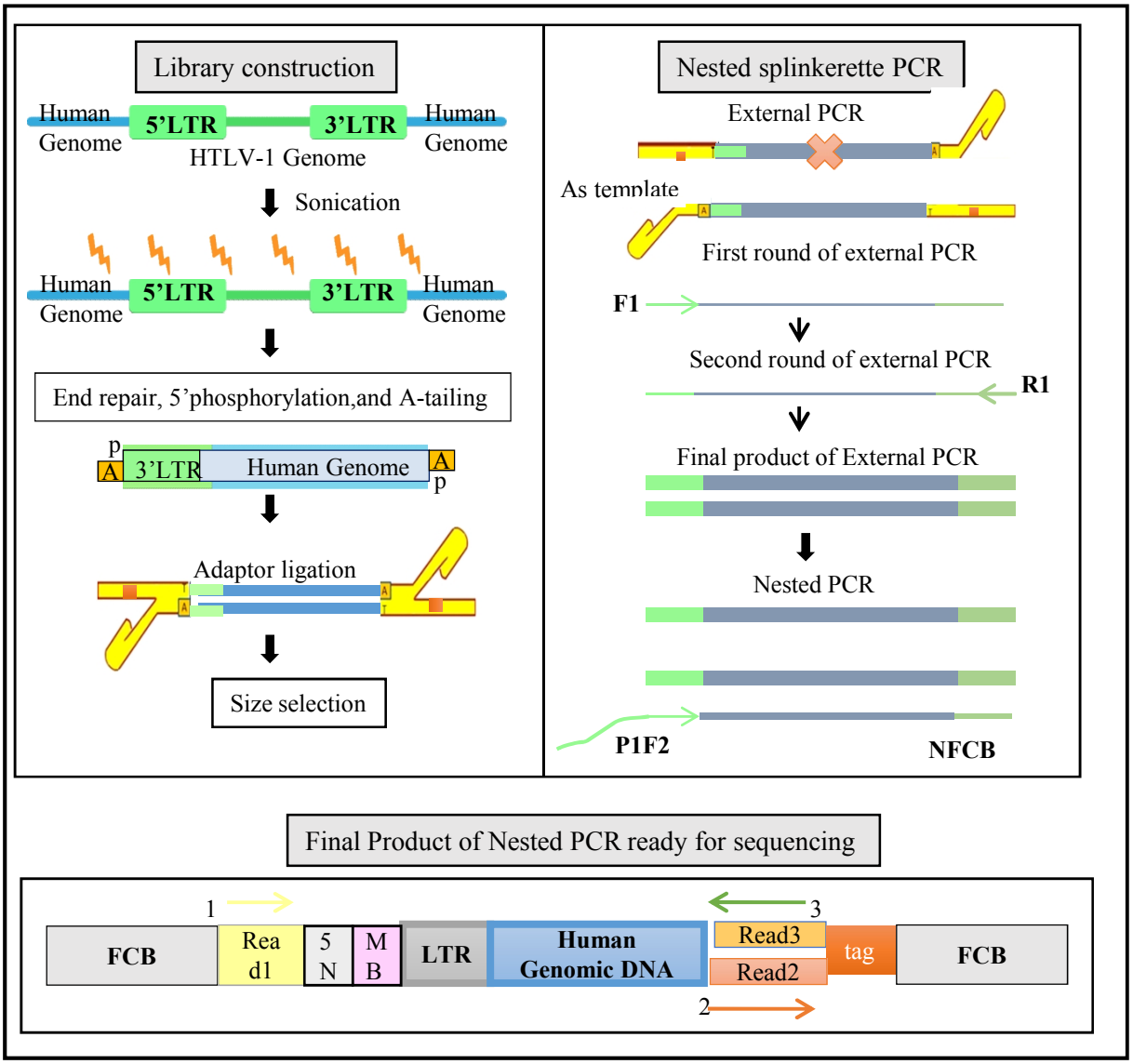


Figure 36 . Outline of the library preparation for sequencing

(A) Design of primers and the hairpin adaptors leads to specific amplification from integration sites. F1 is the LTR-specific primer with a sequence complementary to the bottom strand of the target DNA. R1 is the adaptor-specific primer with a sequence identical to that of the adaptor. This primer can only undergo amplification until the second cycle of PCR, when the complementary strand is produced by amplification from the F1 primer. After amplification of the target region in external PCR, 1 µl of this product is used as the starting material for nested PCR. Alternatively, the external PCR product is diluted 10-fold, and 1 µl is used for nested PCR.

(B) The final product, ready for sequencing, includes the following regions:

FCB = flow cell binding sequence: 3'-/5'-

Read 1: compatible with the read-1 sequencing primer (5'-read)

Read 2: compatible with the read-2 sequencing primer (8-bp tag read)

Read 3: compatible with the read-3 sequencing primer (3'-read)

5N: 5-bp random nucleotides

MB: 5-bp known multiplexing barcodes including: [barcode 1: ACAGT], [barcode 2: GGCTA], [barcode 3: TTACG], and [barcode 4: GCTAC]

Tag: 8-bp randomly generated nucleotides

Amplified target region: Fragments, amplified from 5'-LTR or 3'-LTR, harbor a portion of HTLV-1 genome or the flanking human genome, respectively. Subsequent *in silico* analysis of sequencing data discriminates the flanking human genome from HTLV-1 genome (see Figure 6).

Figure 37

A simple image from an integration site and its shear sites

Read-1 starts with LTR sequence of HTLV-1: "CCAGCGACAGCCCATCCTATAGCACTCTCAGGAGAGAAATTTAGTACACA".

This region follows with a sequence which its position can be determined by mapping to the human genome. The starting position of human flanking region is defined as integration site (marked by a red asterisk*) .

Read-3 contains fragments with different lengths which were generated by sonication. Mapping results of the fragments with different lengths provides positions of shear sites (marked by a dolor sign $).

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCATTTAGCCAGTCTATATATTTTAAGTGGAAAGTTTAATTCATTTATATTCAA$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCATTTAGCCAGTCTATATATTTTAAGTGGAAAGTTTAATTC$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCATTTAGCCAGTCTATATATTTTAAGTGGAAA$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCATTTAGCCAGTCTATATATTTTAAGTGG$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCATTTAGCCAGTCTATATATTTTA$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCATTTAGCCAGTCTATATATT$

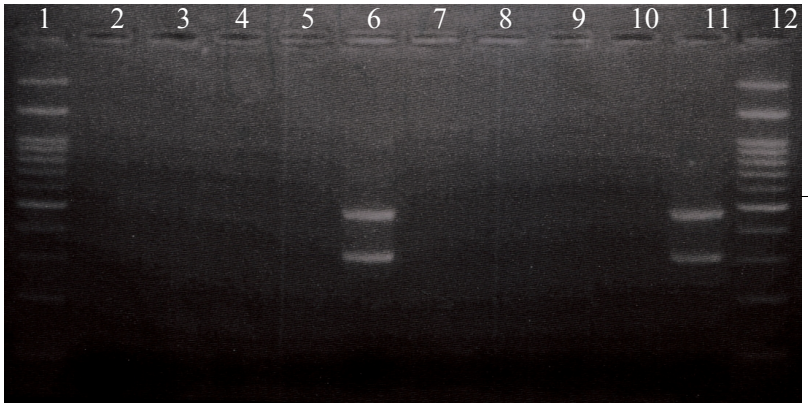
*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTGATGGTTTTTTTTGTTGTTGTCCA$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGTGGAGTG$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATATAGT$

*CCTTTATAGGTGAGATTGCTTTCTTGTAGGCAGTATAT$

*Integration site
$ Shear site



1.	100-bp DNA ladder
2.	U1
3.	OM-10
4.	ACH2
5.	Normal PBMC
6.	TL-Om1
7.	U1
8.	OM-10
9.	ACH2
10.	Normal PBMC
11.	TL-Om1
12.	100-bp DNA ladder

Figure 38.
Checking the specificity of technique for isolation of HTLV-1 integration sites by conventional nested- splinkerette PCR using positive and negative controls
The specificity of method was examined by conventional nested-splinkerette PCR. I analyzed 3 cell lines from HIV (U1, OM10.1 and ACH-2) as a control for exogenous retroviruses. Neither bands nor any smear-like pattern were detected in any of them. gDNA of normal PBMC and TL-Om1 were used as negative and positive controls respectively.

Table-4: required oligo nucleotides and sequencing information

ID of oligonucleotides	Description	Length	Sequence
Illu-long 8N	Long adaptor strand, includes 8-bp random nucleotides, HPLC purification	86	AACCGTTGCTAGGAGAGACCCAAGCAGA AGACGGCATACGAGATNNNNNNNNGTGA CTGGAGTTCAGACGTGTGCTCTTCCGATCT
Illu-short-p	Short adaptor stand, 5'- phosphorylated, HPLC purification	27	P-GATCGGAAGAGCGTTTTTTTTCAAAAA
R1	Adaptor primer for External PCR (first PCR)	28	CGAAGAGTAACCGTTGCTAGGAGAGACC
NFCB	Adaptor primer for Nested PCR	21	CAAGCAGAAGACGGCATACGA
F1	LTR-specific primer for External PCR	20	TACCGGCGACTCCGTTGGCT
F2	LTR-specific primer for Nested PCR (second PCR)	23	CCAGCGACAGCCCATCCTATAGC
P1F2 (No-Index)	Includes necessary sequences for Illumina and LTR-specific primer for Nested PCR	81	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCTC CAGCGACAGCCCATCCTATAGC
P1F2 (Index1-ACAGT)	Include necessary sequences for Illumina, 5-bp-N, 5-bp- barcode, LTR-specific primer for Nested PCR (F2)	91	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCTN NNNNACAGTCCAGCGACAGCCCATCCTAT AGC
P1F2 (Index2-GGCTA)		91	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCTN NNNNGGCTACCAGCGACAGCCCATCCTAT AGC
P1F2 (Index3-TTACG)		91	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCTN NNNNTTACGCCAGCGACAGCCCATCCTAT AGC
P1F2 (Index4-GCTAC)		91	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCTN NNNNGCTACCCAGCGACAGCCCATCCTAT AGC
LTR	A portion of HTLV-1 long terminal repeat (GenBank: J02029.1)	27	ACTCTCAGGAGAGAAATTTAGTACACA
HTLV-1	A portion of HTLV-1 genome (GenBank: J02029.1)	50	GTTGGGGGCTCGTCCGGGATACGAGCGCC CCTTTATTCCCTAGGCAATGG

References

1. Nowell PC: **Citation Classic - the Clonal Evolution of Tumor-Cell Populations.** *Current Contents/Life Sciences* 1988**:**19-19.

2. Greaves M, Maley CC: **Clonal evolution in cancer.** *Nature* 2012, **481:**306-313.

3. Melo FDE, Vermeulen L, Fessler E, Medema JP: **Cancer heterogeneity-a multifaceted view.** *Embo Reports* 2013, **14:**686-695.

4. Sprouffske K, Merlo LM, Gerrish PJ, Maley CC, Sniegowski PD: **Cancer in light of experimental evolution.** *Curr Biol* 2012, **22:**R762-771.

5. Merlo LMF, Pepper JW, Reid BJ, Maley CC: **Cancer as an evolutionary and ecological process.** *Nat Rev Cancer* 2006, **6:**924-935.

6. Nowell PC: **Citation Classic - the Clonal Evolution of Tumor-Cell Populations.** *Current Contents/Clinical Medicine* 1988**:**18-18.

7. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, Fisher R, McGranahan N, Matthews N, Santos CR, et al: **Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing.** *Nat Genet* 2014, **46:**225-233.

8. Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, Dawson KJ, Iorio F, Nik-Zainal S, Bignell GR, et al: **Heterogeneity of genomic evolution and mutational profiles in multiple myeloma.** *Nat Commun* 2014, **5:**2997.

9. Anderson K, Lutz C, van Delft FW, Bateman CM, Guo Y, Colman SM, Kempski H, Moorman AV, Titley I, Swansbury J, et al: **Genetic variegation of clonal architecture and propagating cells in leukaemia.** *Nature* 2011, **469:**356-361.

10. Yamaguchi K, Watanabe T: **Human T lymphotropic virus type-I and adult T-cell leukemia in Japan.** *International Journal of Hematology* 2002, **76:**240-245.

11. 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.** *Japanese Journal of Cancer Research* 1989, **80:**191-195.

12. 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.

13. Gallo RC: **The discovery of the first human retrovirus: HTLV-1 and HTLV-2.** *Retrovirology* 2005, **2**.

14. Tsukasaki K, Tsushima H, Yamamura M, Hata T, Murata K, Maeda T, Atogami S, Sohda H, Momita S, Ideda S, et al: **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.

15. Fujino T, Nagata Y: **HTLV-I transmission from mother to child.** *J Reprod Immunol* 2000, **47:**197-206.

16. Tsukasaki K, Tobinai K: **Biology and treatment of HTLV-1 associated T-cell lymphomas.** *Best Practice & Research Clinical Haematology* 2013, **26:**3-14.

17. Cook LB, Rowan AG, Melamed A, Taylor GP, Bangham CR: **HTLV-1-infected T cells contain a single integrated provirus in natural infection.** *Blood* 2012, **120:**3488-3490.

18. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, Berry C, Bushman FD, Taylor GP, Bangham CR: **The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones.** *Blood* 2011, **117:**3113-3122.

19. Melamed A, Laydon DJ, Gillet NA, Tanaka Y, Taylor GP, Bangham CR: **Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection.** *PLoS Pathog* 2013, **9:**e1003271.

20. 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.** *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* 1984, **81:**2534-2537.

21. 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 Research* 1997, **57:**4862-4867.

22. Okayama A, Stuver S, Matsuoka M, Ishizaki J, Tanaka G, Kubuki Y, Mueller N, Hsieh 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.** *International Journal of Cancer* 2004, **110:**621-625.

23. Tanaka G, Okayama A, Watanabe T, Aizawa S, Stuver S, Mueller N, Hsieh CC, Tsubouchi H: **The clonal expansion of human T lymphotropic virus type 1-infected T cells: A comparison between seroconverters and long-term carriers.** *Journal of Infectious Diseases* 2005, **191:**1140-1147.

24. Ikeda S, Momita S, Kinoshita K, Kamihira S, Moriuchi Y, Tsukasaki K, Ito M, Kanda T, Moriuchi R, Nakamura T, Tomonaga M: **Clinical Course of Human T-Lymphotropic Virus Type-I Carriers with Molecularly Detectable Monoclonal Proliferation of T-Lymphocytes - Defining a Low-Risk and High-Risk Population.** *Blood* 1993, **82:**2017-2024.

25. 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.

26. 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.

27. Cook LB, Melamed A, Niederer H, Valganon M, Laydon D, Foroni L, Taylor GP, Matsuoka M, Bangham CR: **The role of HTLV-1 clonality, proviral structure, and genomic integration site in adult T-cell leukemia/lymphoma.** *Blood* 2014, **123:**3925-3931.

28. Niederer HA, Bangham CR: **Integration site and clonal expansion in human chronic retroviral infection and gene therapy.** *Viruses* 2014, **6:**4140-4164.

29. Firouzi S, Lopez Y, Suzuki Y, Nakai K, Sugano S, Yamochi T, Watanabe T: **Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites.** *Genome Med* 2014, **6:**46.

30. Takatsuki K: **Discovery of adult T-cell leukemia.** *Retrovirology* 2005, **2:**16.

31. Iwanaga M, Watanabe T, Yamaguchi K: **Adult T-cell leukemia: a review of epidemiological evidence.** *Frontiers in Microbiology* 2012, **3**.

32. Matsuoka M, Jeang KT: **Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation.** *Nature Reviews Cancer* 2007, **7:**270-280.

33. Taylor GP, Matsuoka M: **Natural history of adult T-cell leukemia/lymphoma and approaches to therapy.** *Oncogene* 2005, **24:**6047-6057.

34. Taylor GP, Tosswill JHC, Matutes E, Daenke S, Hall S, Bain BJ, Davis R, Thomas D, Rossor M, Bangham CRM, Weber JN: **Prospective study of HTLV-I infection in an initially asymptomatic cohort.** *Journal of Acquired Immune Deficiency Syndromes* 1999, **22:**92-100.

35. Iwanaga M, Watanabe T, Utsunomiya A, Okayama A, Uchimaru K, Koh KR, Ogata M, Kikuchi H, Sagara Y, Uozumi K, et al: **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.

36. Kamihira S, Dateki N, Sugahara K, Hayashi T, Harasawa H, Minami S, Hirakata Y, Yamada Y: **Significance of HTLV-1 proviral load quantification by real-time PCR as a surrogate marker for HTLV-1-infected cell count.** *Clin Lab Haematol* 2003, **25:**111-117.

37. Kamihira S, Yamano Y, Iwanaga M, Sasaki D, Satake M, Okayama A, Umeki K, Kubota R, Izumo S, Yamaguchi K, Watanabe T: **Intra- and inter-laboratory variability in human T-cell leukemia virus type-1 proviral load quantification using real-time polymerase chain reaction assays: a multi-center study.** *Cancer Sci* 2010, **101:**2361-2367.

38. Kuramitsu M, Okuma K, Yamagishi M, Yamochi T, Firouzi S, Momose H, Mizukami T, Takizawa K, Araki K, Sugamura K, et al: **Identification of TL-Om1, an Adult T-Cell Leukemia (ATL) Cell Line, as Reference Material for Quantitative PCR for Human T-Lymphotropic Virus 1.** *J Clin Microbiol* 2015, **53:**587-596.

39. 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.

40. Cavrois M, Gessain A, WainHobson 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.

41. Cavrois M, Wainhobson S, Wattel E: **Stochastic Events in the Amplification of Htlv-I Integration Sites by Linker-Mediated Pcr.** *Research in Virology* 1995, **146:**179-184.

42. Wattel E, Vartanian JP, Pannetier C, Wainhobson S: **Clonal Expansion of Human T-Cell Leukemia-Virus Type I-Infected Cells in Asymptomatic and Symptomatic Carriers without Malignancy.** *Journal of Virology* 1995, **69:**2863-2868.

43. 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. *American Journal of Hematology* 1997, **54:**306-312.

44. 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.

45. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.** *Genome Biology* 2011, **12**.

46. 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-+.

47. Polz MF, Cavanaugh CM: **Bias in template-to-product ratios in multitemplate PCR.** *Appl Environ Microbiol* 1998, **64:**3724-3730.

48. 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.

49. 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.

50. Gini C: **Concentration and dependency ratios (in Italian). EnglishTranslation in Rivista di Politica Economica** 1997, **87:**769-789.

51. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, Berry C, Bushman FD, 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.

52. Gini C: **Sulla misura della concentrazione e della variabilita dei caratteri.** *Transactions of the Real Istituto Veneto di Scienze* 1914, **LIII**.

53. Maio FGD: **Income inequality measures.** *Journal of Epidemiology and Community Health* 2007, **61:**849-852.

54. Chatburn RL: **Evaluation of instrument error and method agreement.** *AANA J* 1996, **64:**261-268.

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.** *Journal of Statistical Software* 2007, **23**.

60. **StatsDirect Medical Statistics Software.**
[http://www.statsdirect.com/].

61. 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.

62. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, et al: **Whole-genome sequencing and variant discovery in C. elegans.** *Nat Methods* 2008, **5:**183-188.

63. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Research* 2008, **18:**1851-1858.

64. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nat Methods* 2008, **5:**1005-1010.

65. Minoche AE, Dohm JC, Himmelbauer H: **Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems.** *Genome Biology* 2011, **12**.

66. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al: **Sequence-specific error profile of Illumina sequencers.** *Nucleic Acids Research* 2011, **39**.

67. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J: **Counting absolute numbers of molecules using unique molecular identifiers.** *Nature Methods* 2012, **9:**72-U183.

68. Bravo HC, Irizarry RA: **Model-based quality assessment and base-calling for second-generation sequencing data.** *Biometrics* 2010, **66:**665-674.

69. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al: **Sequence-specific error profile of Illumina sequencers.** *Nucleic Acids Res* 2011, **39:**e90.

70. 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

71. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Research* 2008, **36:**e105.

72. Tsukasaki K, Hermine O, Bazarbachi A, Ratner L, Ramos JC, Harrington W, Jr., O'Mahony D, Janik JE, Bittencourt AL, Taylor GP, et al: **Definition, prognostic factors, treatment, and response criteria of adult T-cell leukemia-lymphoma: a proposal from an international consensus meeting.** *J Clin Oncol* 2009, **27:**453-459.

73. Tsukasaki K: **[Treatment for adult T-cell leukemia-lymphoma].** *Gan To Kagaku Ryoho* 2009, **36:**756-761.

74. Kamada N, Sakurai M, Miyamoto K, Sanada I, Sadamori N, Fukuhara S, Abe S, Shiraishi Y, Abe T, Kaneko Y, et al.: **Chromosome abnormalities in adult T-cell leukemia/lymphoma: a karyotype review committee report.** *Cancer Res* 1992, **52:**1481-1493.

75. Miyamoto K, Tomita N, Ishii A, Nonaka H, Kondo T, Tanaka T, Kitajima K: **Chromosome abnormalities of leukemia cells in adult patients with T-cell leukemia.** *J Natl Cancer Inst* 1984, **73:**353-362.

76. **Biomaterial resource bank of HTLV-1 carriers, Joint Study on Predisposing Factors of ATL Development (JSPFAD).**
[http://htlv1.org/old/bank-en.html].

77. 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 Research and Human Retroviruses* 2007, **23:**582-582.

78. Shimoyama M: **Diagnostic-Criteria and Classification of Clinical Subtypes of Adult T-Cell Leukemia-Lymphoma - a Report from the Lymphoma-Study-Group (1984-87).** *British Journal of Haematology* 1991, **79:**428-437.

79. 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.** *International Journal of Cancer* 1984, **34:**221-228.

80. Devon RS, Porteous DJ, Brookes AJ: **Splinkerettes--improved vectorettes for greater efficiency in PCR walking.** *Nucleic Acids Res* 1995, **23:**1644-1645.

81. 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.

82. 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 U S A* 1983, **80:**3618-3622.

83. 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.** *Nature protocols* 2009, **4:**789-798.

84. Devon RS, Porteous DJ, Brookes AJ: **Splinkerettes--improved vectorettes for greater efficiency in PCR walking.** *Nucleic acids research* 1995, **23:**1644-1645.

Acknowledgements

After receiving my Bachelor's degree of cellular and molecular biology-Biotechnology in Iran, I came to Japan for my higher education in pursuit of my dreams to become a medical researcher. As a graduate student at the University of Tokyo, I joined to the laboratory of Prof. Toshiki Watanabe which is prominent for research on HTLV-1. About 30 years ago, Japanese researchers contributed to discovery of HTLV-1 as the causative agent of ATL (Takatsuki et al 1977). Both Iran and Japan are endemic areas of HTLV-1 infection. In Iran, a high HTLV-1 sero-prevalence rate of 3 to 0.77% was found in blood donors of Mash-had area, however there is no established research group studying HTLV-1 and ATL in Iran (Safai et al 1996 and Abbaszadegan et al 2003). Also in Japan, a high prevalence rate of 0.66% in men and 1.02% in women were reported. Estimated 1.08 million Japanese are infected with HTLV-1, and approximately 1,000 people die annually from ATL. [M. Satake et al 2012]. ATL is a highly aggressive leukemia of T-cells, with an extremely poor prognosis and short median survival time, due to development of multi-drug resistance among patients with aggressive ATL. However, understanding the true nature of the multiple leukemogenic events that are essential for this aggressive transformation remains elusive.

Through years of having the opportunity to pursue my study as a graduate student in Japan; I learnt to keep high my enthusiasm and preserve to fulfill my dreams. I found that great ideas are good but one needs to really painstakingly develop and follow them with curiosity, drive, and especially persistence. I learnt to think different and make difference. And not to give up when my idea is rejected. Being rejected is not the end but a new start to understand my drawbacks and further strengthen my knowledge and skills. I learnt an invaluable lesson to persist, persist, persist, and never give up until my idea is accepted. Keeping these lessons in my mind, I am always grateful to those who provided me with confidence to persist and realize my dreams. Being acknowledged and trusted to conduct research in the field of HTLV-1 (in Prof. Watanabe's Lab) and collaborate with highly prominent laboratories in the field of Genomics (Prof Sugano_Suzuki's lab) and in silico analysis (Prof. Nakai's Lab) is one of my lasting achievements which open the doors for my future success.

Life for me is a great opportunity to make a big difference in other people's life to the best of my ability. I firmly believe that health is the most valuable blessing one can receive, enabling us to enjoy our surroundings and find satisfaction in our accomplishments. The privilege of returning sick people to health and restoring their happiness is a great joy for me. This is my main motivation which drives me to become a medical researcher and contribute to the happiness of society by my research results. My desire to work hard and not to give up in being beneficial for others and contributing to the society in every aspect of my life (particularly through my research) until the end of my life is what I would call the greatest achievement. I am highly motivated to help HTLV-1 infected individuals and ATL patients all over the world through my research.

This research project would have not been possible without the supports of many people.

I wish to express my gratitude to my supervisor, Professor Toshiki Watanabe who was abundantly helpful and offered invaluable assistance, support and guidance. Thank you Prof. Watanabe for having me in your laboratory.

It is a pleasure to thank, those who made this thesis possible, Professor Sumio Sugano and especially to Professor Yutaka Suzuki for their invaluable comments and sequencing the samples.

I would like to thank to Professor Kenta Nakai for giving me the wonderful opportunity to collaborate with his laboratory.

I owe my sincere and deepest gratitude is to Yamochi-sensei for all his supports, guidance and patience in my training. I will never forget the lessons which he taught me.

Special thanks to Professor Sato (associated professor) for proof reading of this thesis, and to Dr.Nakano (assistant professor), Dr.Yamagishi (PD) and all my lab members for their invaluable assistance.

I would like thank 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, all lab members of the laboratory of Prof. Nakai, Prof. Sugano, Prof. Suzuki and S. Aoki for her help during my master course.

I gratefully appreciate: JSPFAD for providing clinical samples; M. Nakashima and T. Akashi for maintenance of JSPFAD.

I am also highly grateful to Prof. Uchimaru for providing clinical samples, and his helpful comments on interpretation of clinical results.

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

This work was supported by the Japanese Society for the Promotion of Science (JSPS) - DC1 (24.6916 to SF). Hereby I express my gratitude for their support during 3 years of my PhD.

I express my deep respect and gratitude to the NITORI scholarship foundation for supporting me during 2 years of my undergraduate studies.

I sincerely appreciate patience and hard work and friendship of Yosvany Lopez in bioinformatics analysis of this thesis. I never forget your helps Lopez. Always thank you.

I wish to express my love and gratitude to my beloved families in Iran (my mother (Maryam Firouzi) and my father (Reza Firouzi), my sister (Sara) and my brother (Unes) & in Japan (especial thanks to my uncle (Abbas Firouzi) who was my father in japan, and his family: Miwa-san, Miran and Reon); for their understanding & endless love, through the duration of my life and education.

Finally, many uncountable thanks to Amir Farmanbar for his useful advices on statistical analysis, and especially for his supports and positive energies during writing the present thesis.

Thanks God for everything…