Genome-wide mapping of proviral integration sites in aggresive virus-induced leukemia

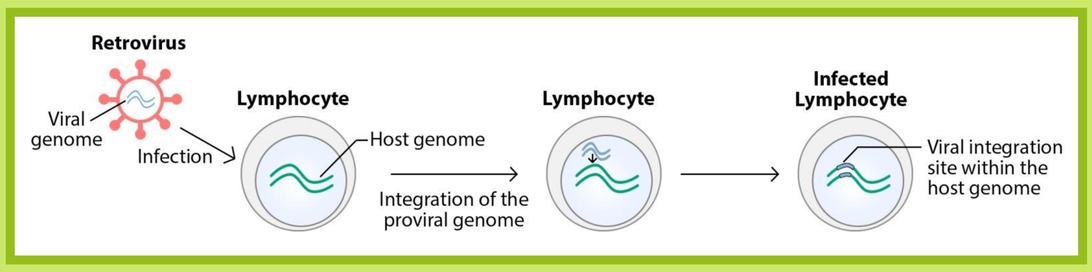
Nicolas Rosewick[1,2], Vincent Hahaut[1], Maria Artesi[1], Keith Durkin[1], Ambroise Marçais[3], Philip Griebel[4], Natasa Arsic[4], Véronique Avettand-Fenoel[5], Arsène Burny[2],Carole Charlier[1], Olivier Hermine[3], Michel Georges[1] & Anne Van den Broeke[1,2]

[1] Unit of Animal Genomics, GIGA-R, Université de Liège (ULg), 4000 Liège, Belgium, [2] Laboratory of Experimental Hematology, Institut Jules Bordet, Université Libre de Bruxelles (ULB), 1000 Brussels, Belgium, [3] Service d'hématologie, Hôpital Universitaire Necker, Université René Descartes, Assistance publique hôpitaux de Paris, Institut Imagine, Paris, France. U1163-ERL8254, [4] Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon S7N 5E3, Canada, [5] Service de Virologie, Hôpital Necker-Enfants Malades, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

BACKGROUND

A hallmark of retroviral infection is the integration of retro-transcribed viral DNA (provirus) within the host genome. Identifying proviral integration sites within the genome is critical for a better understanding of viral pathogenesis. We developed an improved next generation sequencing (NGS) based method to profile thousands of viral integration site in a single experiment, providing an unprecedented way to characterize the clonal distribution of infected cells.

Human T-Cell Leukemia virus type-1 (HTLV-1) and Bovine Leukemia Virus (BLV) are two functionally and structurally related delta-retroviruses. These viruses infect T (HTLV-1) and B (BLV) lymphocytes, provoking a polyclonal expansion that will evolve into an aggressive lethal monoclonal leukemia in ~5% of individuals following decades of latency. It is generally assumed that oncogenic changes are largely dependent on virus-encoded products and especially the trans-activating effects of the Tax (paradoxically silent at the tumor stage) and HTLV-1 HBZ oncoproteins, while progression to acute leukemia is independent of viral integration site that has been found to be extremely variable between tumors.

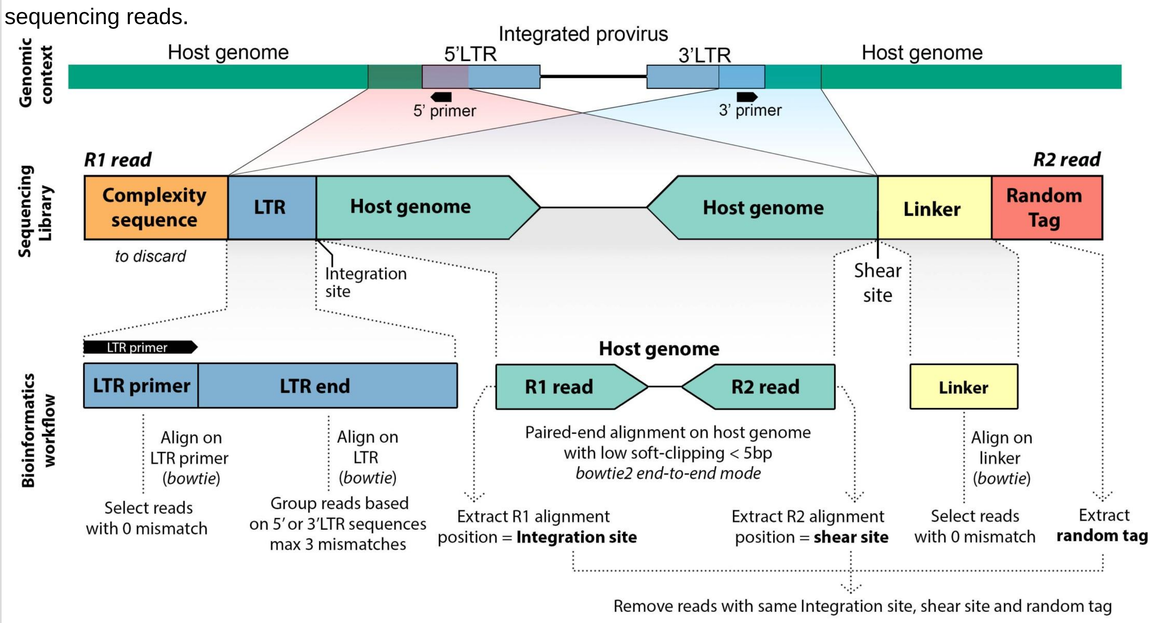


AIM

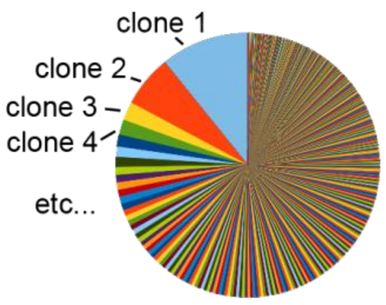
Use NGS to profile the integration sites genome-wide and quantify the abundance of the corresponding clones in HTLV-1 and BLV induced leukemia.

DEEP DNA SEQUENCING TO MAP PROVIRAL INTEGRATION SITES AND QUANTIFY INFECTED CLONES

Viral integration sites were identified by sequencing DNA fragments spanning the virus-host junction. Sequencing was performed on an Illumina MiSeq plateform. After alignment to the host genome the sequencing reads (the sequenced DNA fragments) were analyzed in *R* to extract proviral integration sites. First, to avoid signal from inter-sample contamination (samples from the same sequencing run) we verified that all sequencing reads from a particular sample harbor the correct/same linker sequence. In order to remove PCR duplicates we introduced a random tag (a random sequence of 8 nucletotides) in the R2 read unique to each DNA fragment sequenced. Reads with the same shear site (alignment position) and same random tag sequence were flagged as PCR duplicates and merged. The abundance of a specific clone (infected cells with the same integration site) was then computed relative to the total number of sequencing reads.



Pie-chart representing the clone abundance distribution in a leukemia sample



R packages used for the analysis :

- GenomicRanges
- BioStrings
- ShortRead
- Stringdist

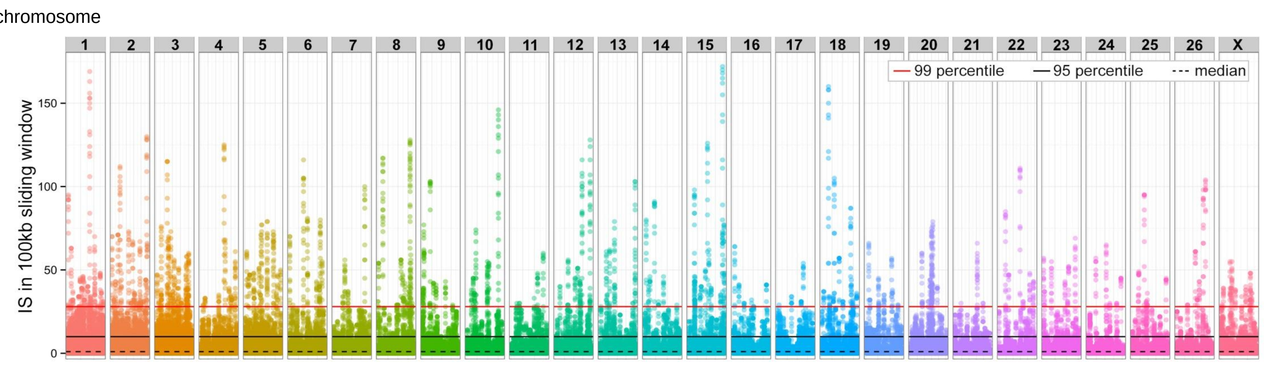
Plot:

- ggplot2
- Gviz

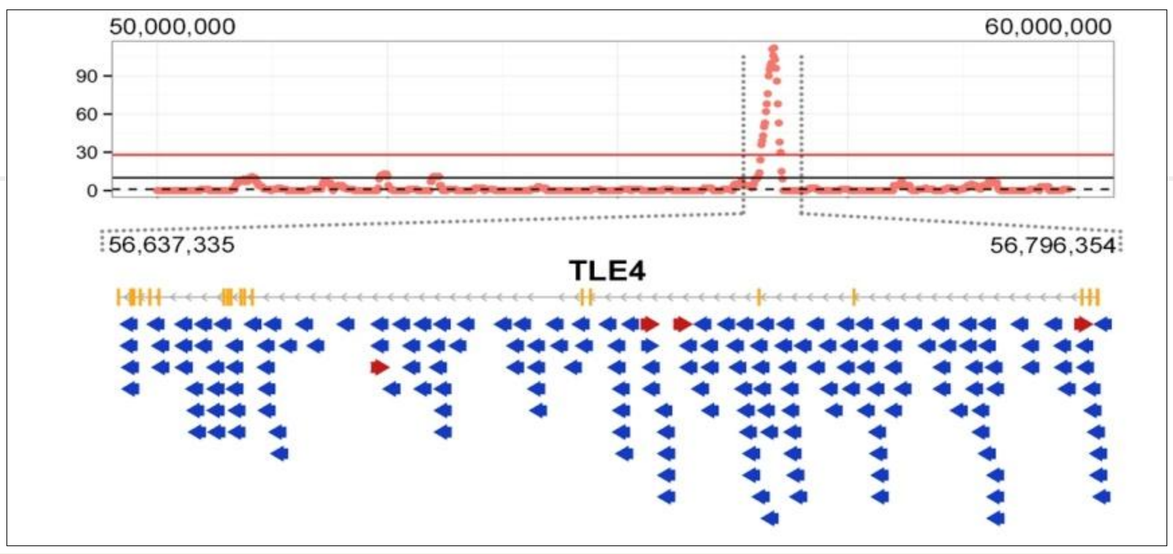
Application 1 : Identification of proviral integration hotspots in Bovine Leukemia Virus-induced leukemia

Genome-wide profiling of 10 infected sheep at different time points post-infecton (from 1 month to 3 years) identifies ~130.000 unique integration sites (IS) and reveals hotspots of proviral integration in the vicinity of host cancer-driver genes.

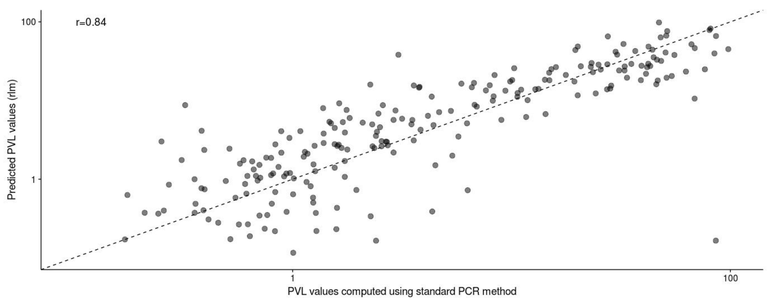
Rosewick et al. 2017, Nature Communications doi:10,1038/ncomms/15264



Proviral integration hotspot localized within the cancer-related gene TLE4. Each arrow represents a single integration site



Application 2 : Predicting proviral load (percentage of infected cells) from the number of virus-host sequencing reads



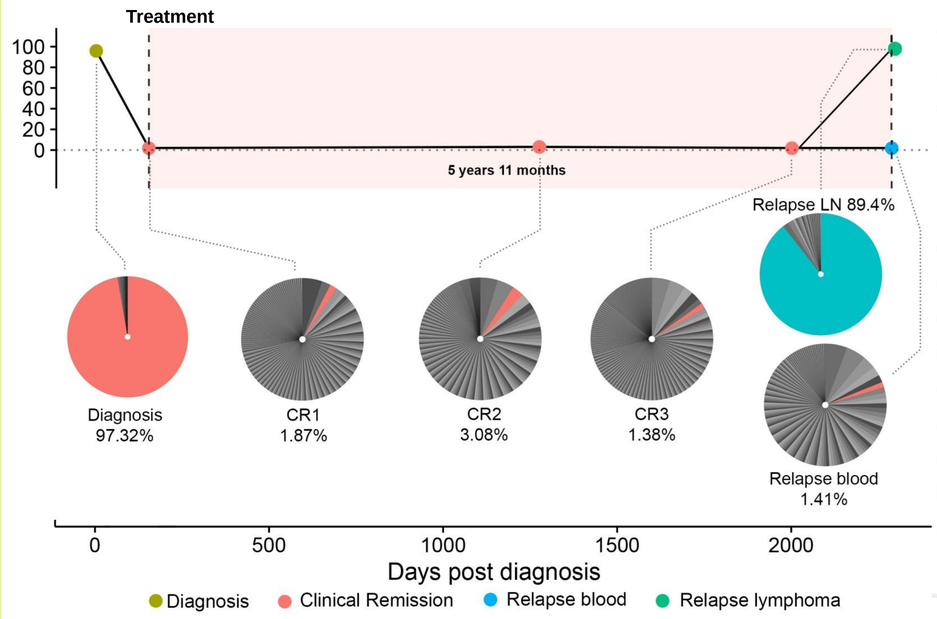
Proviral load (PVL) is an important prognostic marker

Robust linear regression

rlm(PVL ~ reads)

Corr = 0.84

TRANSLATION INTO THE CLINICS : MONITORING ATL PATIENTS



We used our method to follow-up the progression of the leukemic clone in patients diagnosed with Adult T-cell Leukemia (ATL).

The method enables the follow-up of the dominant tumor clone detected at diagnosis during the course of the disease. This improves prediction of relapse (increase of the tumor clone abundance). The tumor clone during clinical remission (pink background) is not systematically the most abundant one.

SUMMARY

We used a deep DNA sequencing approach to profile proviral integrations in delta-retrovirus induced leukemia. Our method identifies thousands of proviral integration sites in a single experiment, revealing interactions between the provirus and functionally important host genes. Our approach can be used in clinical settings, enabling the follow-up of leukemia patients and better predict progression. The core analysis was performed in R using a mix of CRAN and Bioconductor packages (GenomicRanges, ShortRead, stringdist) reinforcing *R* as one of the main actors in bioinformatics related analysis.