The Endogenized K113 and K115 Retroviruses in the General Public

Thomas Reynolds, Dr. Keith Garrison

School of Science, Saint Mary's College of California, Moraga, California, United States of America

Abstract

The collection of retroviral insertions in the human genome, known as HERVs, are increasingly of interest to medical and genetic researchers. Their distribution in the overall population is not well-known, and some evidence exists to suggest that HERV-K113 and K115, two of the most intact proviral insertions, are not as new as current theory holds. This project was originally intended to be a general population study, focusing on the distribution of K113 and K115 in healthy individuals, but technical errors in DNA extraction and PCR optimization prevented us from gathering these data. However, a concurrent study of K113 in two different cancer cell lines determined that one of the cell lines, MCF-7, carried a K113 insertion. We identified novel SNPs in the K113 5' LTR of the MCF-7 cell line.

Table of Contents

Introduction

The Human Endogenous Retrovirus

The modern human genome is a evolutionary ensemble, composed from various sources over the course of history. Retroviruses are potential contributors, because their mode of replication involves integration of retroviral DNA into the host cell's genome.[1] In the normal course of retroviral infection the newly-inserted DNA rests in a dormant state before being activated, overriding the cell's natural processes to replicate new infectious retroviral particles. Retroviruses are either known or implicated as pathogens in multiple organisms, including human immunodeficiency virus (HIV).[1]

Approximately 8% of the human genome is derived from retroviruses that infected humanity's evolutionary ancestors; these sequences are collectively referred to as the Human Endogenous Retroviruses (HERVs). The term *endogenous* means that the retrovirus exists only as an insertion in another organism's genome. In contrast, the disease-causing virus mentioned above is exogenous. Insertions of retroviral genomes into the host genome are called *proviruses*. Endogenous retroviruses are believed to exist in their current state because of deactivating mutations that prevented further activation and replication. The only way an endogenized retrovirus can be transmitted from one individual to another is if the provirus exists in the genome of a germ cell (sperm or egg), which allows it to be inherited[2]. However, research is now looking into links, both positive and

---

[1] Nelson, David L., and Michael M. Cox. *Lehninger Principles of Biochemistry*. New York: W. H. Freeman, 2008. Print.

[2]Lee, Young Nam and Paul D. Beiniasz. "Reconstitution of an Infection Human Endogenous Retrovirus." PLoS Pathog. 3(1) 2007

pathological, between HERVs and a multitude of conditions including breast cancer and melanoma. [3,4]

At first, a HERV provirus consists of the following sections: a Long Terminal Repeat (LTR) which acts as a replication promoter, the *gag*, *pro*, *pol*, and *env* genes which all encode for either structural or transcriptional proteins, and finally another LTR.[4] Over time the genetic information of a HERV is degraded due to mutation. This can occur on a gradual scale ranging from a single nucleotide change to the entire provirus itself being deleted from the genome.[5]

HERV-K113 and -K115

This study was centered on two HERV sequences, designated K113 and K115. Both are part of the HERV-K family, named for a shared reverse transcription process that uses mRNA molecules associated with the amino acid lysine (Lys, K). These proviruses are notable because they are relatively intact: K113 has fully open reading frames (ORFs) for all of its proteins, which means that each of the four genes (*gag, pro, pol*, and *env*) are clearly distinct and translatable.[4] This strongly indicates that it is a relatively young provirus, although the actual method of dating HERV insertions is up for debate.[6]

---

[3] Burmeister, Thomas et al. "Insertional Polymorphisms of Endogenous HERV-K113 and HERV-K115 Retroviruses in Breast Cancer Patients and Age-Matched Controls." AIDS Research and Human Retroviruses. 20, 2004. pp. 1223-1229.

[4] Schiavetti, Francesca et al. "A Human Endogenous Retroviral Sequence Encoding an Antigen Recognized on Melanoma by Cytolytic T Lymphocytes." Cancer Research 62. 2002. pp. 5510-5516.

[5] Macfarlane, Catriona, and Peter Simmonds. "Allelic Variation of HERV-K(HML-2) Endogenous Retroviral Elements in Human Populations." J Mol Evol. 59, 2004. pp. 642-656

[6] Jha, A.R. *et al*. "Cross sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before *Homo sapiens*." Mol Biol Evol. 2009 Aug 10. [Epub ahead of print]

Experimental Phases

Over the course of the summer, the project underwent several changes in methodology, resulting in a significant departure from the original proposed study. It will be easier to describe this project in terms of several distinct phases, all of which centered around HERV-K113 and K115. Also, these phases were not discrete, but often occurred simultaneously with the changing demands of the project. Major results and discussion for the study will be presented as a whole.

Phase I: Preparing for Sample Collection

At first, the study was intended to be a public survey of K113 and K115 incidence in the general population of Moraga, California including students at Saint Mary's College. While the proviruses had been studied in various disease states, no study had examined it in a healthy population with any great detail. The goal was to collect confidentially genomic DNA from 100 volunteers, which would greatly increase the amount of data on HERVs in healthy individuals overall. Genetic material would be donated in buccal cell swabs, and extracted by a Chelex-based protocol as used in the Biology 01-Lab class. PCR would then amplify the flanking region and/or the 5' LTR of each proviral insertion, generating a complete genotype for each individual. 5' LTR amplicons would be sent to the UC Berkeley DNA Sequencing Facility, both as a confirmation of the PCR-based genotyping assay and to identify novel LTR polymorphisms in the general population. More detailed procedures for this part of the study will follow in **Phases III-VII.**

As part of the informed consent process (see below), donors could also divulge information about their gender and ethnicity. Insertion frequencies would be correlated with this donor information, to determine novel trends in the population. Also, contact information was collected that could be used in a future, follow-up study involving HERVs and immune responses.

Donors would be approached and DNA collected in Dryden or Oliver Halls. Potential donors would be asked to not donate if they had a history of cancer or HIV infection, but health histories would never be requested. The Institutional Review Board (IRB) requested certain changes in the collection protocol to help protect donor confidentiality. For the sake of donor privacy the collection procedure was relocated to the Molecular Biology room at Brousseau Hall (#209), and only the primary researcher would be present. Recruitment activities in Dryden/Oliver Hall would only assign appointment times for sample collection at Brousseau 209. Recruitment materials (Appendix I), an oral script for donor-researcher interaction (Appendix II), an Informed Consent Document (Appendix III), and a suitable method of correlating donors to their samples (Appendix IV) were written and submitted for updated approval. The revisions were accepted by the IRB on July 7th and the public survey began soon afterward.

Phase II: DNA Extraction

Original Public Survey Procedures

8 individuals agreed under the terms of the Informed Consent Document to donate buccal cells for the study. Their samples are identified by pseudo-random code numbers generated in Microsoft Excel 2008 using the following formula: =DEC2HEX(RANDBETWEEN(1,1000). By using these pseudo-random numbers, no sample could be linked to a donor by knowing when they donated. In contrast, genotypes of sequentially-numbered samples could be correlated to donors if one knew when the individual donated. Samples were tracked in an Excel workbook, with conditional formatting that would highlight a sample code that had been used twice (Appendix V). This made sure that each sample, and donor, was assigned a unique identifier. Number assignment occurred after the donor had left Brousseau 209, assuring total confidentiality.

A single paper database correlated each sample number with its donor (Appendix IV). It was kept in Dr. Garrison's locking file cabinet in his locked office, and it could only be removed and accessed by

either Dr. Garrison or the primary researcher. The only times it left the office were during sample collections, when the database was stored in the locked drawer BROU211-67 in the Advanced Classes and Research room, Brousseau 211. Again, only Dr. Garrison and the primary researcher had access to this drawer. BROU211-67 was also the storage site for the experimental notebook. Informed Consent Documents were stored in Dr. Garrison's file cabinet.

After a donor signed the Informed Consent Document, a Puritan Cap-Shure swab was removed from its sterile packaging and handed to the donor. They were asked to open the cap, extend the swab, and swab both cheeks for 15 seconds. After that, the swab was sealed and handed back to the primary researcher, the code number was assigned, and the swab left in a sterile plastic box to air-dry overnight (collections took place in the afternoon). The donor was not swabbed if they had consumed caffeine up to 30 minutes prior.

The DNA was extracted using a Chelex-based procedure validated in the Bio-001 lab course. The air-dried swab tip was inserted into a 1.5 mL microcentrifuge tube with 1.0 mL of 1X phosphate-buffered saline (PBS), and vigorously spun to dislodge the collected buccal cells. The buccal cell suspension was then centrifuged for one minute at 6,200*g* in an Eppendorf 5415C microcentrifuge. The supernatant from the centrifugation was drawn off, and the pellet resuspended in a 5% (w/v) Chelex-100 bead suspension. The sample was boiled for 10 minutes, placed in an ice-bath for 2 minutes, and then recentrifuged at 6,200*g* again for 30 seconds. 200 µL of the supernatant was collected and stored as the genetic template extract. All waste was treated as biohazardous and disposed of accordingly.

The extraction procedure had been repeatedly validated in a classroom setting. However, subsequent PCR failures necessitated a number of optimization procedures.

Cell Lines With QIAGEN Extraction

Dr. Garrison procured suspensions of cells from the MCF-7 and T47D cell lines. These came from the UCSF Division of Experimental Medicine. DNA was extracted from the cell lines using the

QIAGEN DNEasy Blood and Tissue Kit with the following procedure.

Approximately 0.5 mL of each cell line suspension (total count≈1x10^7 cells) was centrifuged at 15,800*g* in the Eppendorf microcentrifuge for 5 minutes. One pellet of each sample was set aside for storage at -80°C, and the other pellet resuspended in 200 µL of PBS. This was done in the Cell Culture Room (Brousseau 200). 20 µL of proteinase K solution and 200 µL lysis buffer (AL) were added, and the samples incubated in a 56°C water bath for 10 minutes. After the lysis buffer was added, work proceeded in the Molecular Biology lab room (Brousseau 209).

200 µL of 98% EtOH was added post-incubation to stop the lysis reaction, and the entire mixture was pipetted onto a DNEasy Spin Column and centrifuged for 1 minute at 6,200*g.* This was followed by another 1 minute centrifugation at 6,200*g* with 500 µL of Wash Buffer 1 (AW1), and a 3 minute centrifugation at 15,800*g* with Wash Buffer 2 (AW2). Flow-through was discarded as biohazardous waste after each spin. Finally the spin column was inserted into a 1.5 mL microcentrifuge tube, and the DNA eluted with 200 µL elution buffer (AE) in a 1 minute, 6,200*g* spin.

0.8% LE agarose electrophoresis gels were ran with 5 µL cell line DNA Extract and 1 µL Ambresco EZ-View Three Loading Buffer. 5 µL of λ HindIII DNA Ladder was used for reference. Gel #7 (Fig. 1f) showed that the QIAGEN extraction worked splendidly with MCF-7 and T47D, and the procedure was successfully repeated and confirmed with Gel #13 (Fig. 1l).

Chelex-QIAGEN Comparison with Buccal Cells

Because the QIAGEN extraction of MCF-7 and T47D had worked, two donors were asked to come back and re-donate cells for optimization and comparison. These were samples 15F and 295, one Caucasian male and female. They each swabbed with two different swabs: the Cap-Shure swabs and the Fisherbrand polyester swabs used in Bio-001L, and for the two extraction procedures–4 swabs per donor overall. These were identified as PC (Polyester-Chelex), CC (Cotton-Chelex), PQ (Polyester-QIAGEN), and CQ (Cotton-QIAGEN). The donors were swabbed for 1 minute per swab, to

increase yield, and the DNA was extracted immediately post-donation. The QIAGEN extractions were done with a User-Developed Protocol for saliva samples (Appendix VIII).

As can be seen in Gel #16 (Fig. 1m), yield for all four collection-extraction procedures was disappointingly poor. Only the QIAGEN extractions showed any visible product, with a slight yield increase if the polyester swab had been used. This result was one of the reasons for the termination of the swab-based Moraga study.

STAN Samples

At the beginning of August, Aashish Jha at UCSF donated 28 samples from his own project to the study. These all came from blood donors at the Stanford Blood Center: 16 were in the form of cell suspensions in freezing medium (serum +DMSO), and the other 12 as cell pellet residuals. Jha also contributed ethnicity and gender information for the Stanford samples.

The IRB was notified of their use by a submitted, written update. Because the Chelex-based protocol was already considered ineffective, all 28 were extracted by the QIAGEN DNEasy protocol.

Figures 1o, 1r, and 1t show that the QIAGEN extractions frequently worked well, with visible yields in 13 of the 15 extracts assayed. Because these had succeeded, it was hoped that a study of K113/K115 insertion frequency could still take place, even if it was not in the original context.

Phase III: PCR

Original Survey Procedures

PCR was originally carried out in 20 µL reaction volume, with 10 µL of Applied Biosystems PCR Fast Master Mix, 5 µL of genomic extract, 1 µL of each primer (Appendix IX) at 20 µM concentration, and 3 µL of sterile water. This conformed with the procedures recommended by Applied Biosystems. For the initial survey K113 was the only insertion examined; K115 had been disregarded due to time and budget constraints.

The Flanking reaction used the K113 5' Flanking primer and the 3' Flanking primer, and would amplify the area around the insertion locus *if the provirus was not*

present. Therefore, a Flanking product indicated the presence of one K113(-) chromosome. The Flanking reaction was also called the Insertion reaction.

The LTR Reaction used the K113 5' Flanking primer and the LTR-Reverse (LTR-R) primer, replicating the K113 5' LTR and part of the *gag* gene *if the provirus was present*.

For each individual there had to be product from at least one reaction. However this protocol only worked once, in the Flanking/Insertion reaction for Sample 054 (Figures 1a, 1c).

Public Survey Optimization

The first attempt to optimize the PCR was by varying the quantity of template DNA present. As can be seen in Figure 1b, varying the template volume did somewhat affect the yield for 054-Flanking, but did not produce a result in 054-LTR. From then on, the maximum possible template volume of 8 µL was used in all reactions with AB's Fast PCR Master Mix.

Figure 1d shows that the increased template DNA volume did not generate any product with 3 other samples in Flanking or LTR reactions. At this point, the efficacy of the Master Mix was in doubt.

Acting on the advice of Dr. Hansell, the next test used the primary researcher's own DNA to examine the TPA25 *Alu* repetitive sequence. Since the primary researcher is a known heterozygote for this *Alu* insertion, the test would determine the viability of AB Fast PCR. Figure 1d shows that neither TR-TPA25 reaction worked. As a control, the TR-TPA25 reaction was repeated using the GE Dry PCR Bead protocol used in Bio-001L that had generated the TPA-25 heterozygous genotype. The results from this test, as seen in Figure 1e, did not agree with the prior genotype, but the presence of product at all strongly implicated the AB Fast PCR Master Mix as a likely culprit in the K113 failures. For the sake of comparison, three Flanking reactions were performed with swab samples and Dry PCR Beads: 1E6, 316, and 1C8. The results are also unclear, as none of the products correspond to the known K113 amplicon sizes.

A call to Applied Biosystems confirmed that Fast PCR Master Mix was formulated for fast-style thermal cyclers, and was not

compatible with the Perkin-Elmer 2400 or Eppendorf thermal cyclers available. Another PCR mix would have to be used.

Curiously, K113 (and later K115) PCR using AB Fast PCR Master Mix worked well with the MCF-7 and T47D extracts when large amounts of template DNA were added to the reaction (Figures 1g, 1h, 1i, 1j). These results allowed MCF-7 to be genotyped as K113-heterozygous and T47D as a null homozygote.

PCR Mix Comparison and STAN Sample Amplifications

Because AB Fast PCR Master Mix was shown to not work with the public samples, the assay switched over to using Invitrogen Platinum PCR SuperMix. Samples were prepared with 45 µL of SuperMix, 3 µL of DNA template, and 1 µL of each primer at 20 µM concentration.

Again, the Invitrogen SuperMix failed to produce results with the public swab samples, regardless of the extraction procedure or type of swab used. At that time no other PCR mix was available, and the Moraga public survey was terminated due to technical shortcomings and time constraints (Figures 1h, 1i, 1j, 1m, 1n).

The STAN samples tended to work much better with the Invitrogen PCR mix, but success on a per-reaction basis was mixed at best. Of the 146 STAN-Invitrogen amplifications attempted, only 42 were observed to give product–a 29% success rate. Flanking products were eventually observed for all the suspended STAN samples except for STAN65 and STAN79 (Figures 1p, 1s, 1t, 1u,). However, as can be seen with STAN80 in Figures 1y and 1z, a successful reaction was not necessarily reproducible. Indeed, MCF-7 and T47D frequently failed to give product with either PCR mix, even after their genotypes were confirmed (Figures 1p, 1s, 1t, 1u).

With the Invitrogen Mix, the LTR reaction frequently gave Flanking product (Figures 1u, 1w, 1y), even though the primers had been properly paired. STAN71B was seen to produce LTR, indicating it to be heterozygous (Figure 1v), but it also tended to produce Flanking product instead. The 100 µM stock primer purity was assessed with single-primer PCR using MCF-7: if any reaction, with only primer, managed to give product, then that primer was contaminated. Figure 1x shows no product from any of the 3 single-

primer tests, but repeated failures with MCF-7 and Invitrogen PCR Mix makes this result somewhat ambiguous. The STAN PCRs' repeated failures, with no apparent cause, led to the termination of the HERV K113 frequency study of STAN samples.

Phase IV: Sequencing

PCR Cleanup

LTR product was prepared using the QIAGEN QIAQuick PCR Purification Kit protocol; like the DNEasy protocol it is spin column-based. 5 volumes of purification buffer (PB) are added to a single volume of PCR product. If the solution has been treated with pH indicator and is orange or purple, 10 µL of 3.0 M sodium acetate is added to turn the product-buffer solution yellow.

The entire solution was loaded into a collection spin-column and centrifuged for 1 minute, binding the DNA. Biohazardous flow-through was discarded and the column spun again with 0.75 mL of PE solution for 1 minute. Once again, flow-through is discarded and the "empty" column spun for another 1 minute. This cleared residual ethanol from the PE solution out of the DNA. Finally, the column is placed in a 1.5 mL microcentrifuge tube and spun with 50 µL Elution Buffer (EB). All centrifugation is done at maximum possible speed (15,800*g*). Survival of the cleaned LTR amplicon was confirmed by a 1.2% agarose gel (Figures 1k, 1cc). Both K113 and K115 LTR amplicons from MCF-7 visibly survived the clean-up procedure, but the K115 LTRs from STAN71B and T47D did not. There was no other way to determine amplicon viability. The LTR from STAN71B was not prepared for sequencing, because the Stanford-based public study had been terminated by that point. The T47D K115 LTR was prepared for sequencing; it was hoped that enough template amplicon had survived to make sequencing viable.

Sequencing Preparation

The UC Berkeley DNA Sequencing Facility required that 100 ng of PCR product-type DNA be added to 0.8 pmol of a single primer, and diluted to 13 µL total volume. HERV-K LTR samples for sequencing were prepared with 3 µL of LTR amplicon, 1.6 µL of primer (either 5' Flanking or LTR-R) at .5 µM, and 8.4 µL of sterile water. The amount of template

amplicon in a cleaned LTR reaction was estimated by visual comparison with pBR322 DNA in standardized concentrations, and determined to be <50 ng/μL. This proportion was used for both MCF-7 LTR samples; because the T47D K115 LTR was ambiguously weak and did not visibly survive the cleaning process, 8.4 μL of T47D K115 LTR was used with 1.6 μL of primer and 3 μL sterile water.

The prepared samples were then driven to the UCB DNA Sequencing deposit point, in the Life Sciences Addition building on the UC Berkeley campus. Within 36 hours, an AB1 and a SEQ file for each sample was returned by email. Both were often necessary for the final phase of the project. Of the seven samples sent to Berkeley, only three were successfully sequenced: MCF-7/K113 (Flanking Primer), MCF-7/K113 (LTR-R), and MCF-7/K115 (LTR-R).

Phase V: In Silico Analysis

Sequences were inspected and trimmed in 4Peaks (mekentosj.com); the primary tasks were to correct easily-identifiable nucleotides marked Unknown ("N"), and start the sequence as close to the beginning of the LTR reference as possible.

Alignments, translations, and BLASTs were performed in Geneious Basic 4.7.4 (geneious.com). Reverse complements were generated for sequences made with the LTR-R primer. The primary references used for alignments were the complete K113 and K115 genomes (AY037928 and AY037929, respectively). Other references included AC016582 (*Homo sapiens* chromosome 19 complete sequence) and the HERV-K10 reference sequence (HUMERVKA).

Results

K113 Genotypes

Sample Name	Source	K113	Sex	Ethnicity
1C8	Moraga Study	(?/?)	M	Hispanic
054	Moraga Study	(-/?)	F	Caucasian/ European
15F	Moraga Study	(-/?)	F	Caucasian/ European
2F4	Moraga Study	(?/?)	M	Caucasian/ European
1E6	Moraga Study	(?/?)	F	Caucasian/ European
316	Moraga Study	(?/?)	F	Caucasian/ European
24B	Moraga Study	(?/?)	F	Caucasian/ European
295	Moraga Study	(-/?)	M	Caucasian/ European
MCF-7	ATCC Cell Line	(+/-)	F	Caucasian/ European
T47D	ATCC Cell Line	(-/-)	F	Caucasian/ European
STAN58	Stanford Blood	(-/?)	F	Asian/Oriental/ Hawaiian/Eskimo
STAN62	Stanford Blood	(-/?)	F	Caucasian/ European
STAN63	Stanford Blood	(-/?)	M	Caucasian/ European
STAN64	Stanford Blood	(-/?)	F	Caucasian/ European
STAN65	Stanford Blood	(?/?)	M	Caucasian/ European
STAN66	Stanford Blood	(-/?)	F	Caucasian/ European
STAN67	Stanford Blood	(-/?)	N/A	N/A
STAN68	Stanford Blood	(-/?)	F	Caucasian/ European
STAN70	Stanford Blood	(-/?)	N/A	Caucasian/ European
STAN71 A	Stanford Blood	(-/?)	N/A	N/A
STAN71 B	Stanford Blood	(+/-)	N/A	N/A
STAN77	Stanford Blood	(-/?)	F	Cauc./Euro/ Cent./S. American
STAN79	Stanford Blood	(?/?)	F	Native American
STAN80	Stanford Blood	(-/?)	M	African American
STAN81	Stanford Blood	(-/?)	M	Asian/Oriental/ Hawaiian/Eskimo
STAN82	Stanford Blood	(-/?)	M	Caucasian/ European
16	Stanford Blood	(?/?)	M	Caucasian/ European
31	Stanford Blood	(?/?)	F	Asian/Oriental/ Hawaiian/Eskimo

Sample Name	Source	K113	Sex	Ethnicity
34	Stanford Blood	(?/?)	M	Caucasian/ European
43	Stanford Blood	(?/?)	M	Caucasian/ European
46	Stanford Blood	(?/?)	F	Caucasian/ European
51	Stanford Blood	(?/?)	F	Caucasian/ European
53	Stanford Blood	(?/?)	M	Caucasian/ European
55	Stanford Blood	(?/?)	M	Caucasian/ European
55A	Stanford Blood	(?/?)	M	Caucasian/ European
59	Stanford Blood	(?/?)	N/A	N/A
60	Stanford Blood	(?/?)	M	Caucasian/ European
61	Stanford Blood	(?/?)	N/A	N/A

All samples denoted as (?/?) did not produce Flanking or LTR product, and a (-/?) genotype indicates that only Flanking product was ever isolated. There is a very good chance that a null homozygote would be labelled as (-/?), and that the vast majority of these samples are in fact null homozygotes. The high rate of PCR failure that occurred during the study induces a degree of uncertainty with negative LTR reactions.

K115 Genotypes

Sample Name	Source	K113	Sex	Ethnicity
MCF-7	ATCC Cell Line	(-/?)	F	Caucasian/ European
T47D	ATCC Cell Line	(-/?)	F	Caucasian/ European
STAN79	Stanford Blood	(?/?)	F	Native American
STAN80	Stanford Blood	(?/?)	M	African American
STAN81	Stanford Blood	(?/?)	M	Asian/Oriental/ Hawaiian/Eskimo
STAN82	Stanford Blood	(?/?)	M	Caucasian/ European

Both cell lines, STAN80, and STAN81 gave visible LTR product, but sequencing of the MCF-7 product indicated that it was actually the 5' LTR for K113 (see below). The genotype for the cell lines, partial negative, can be proven by comparing Flanking product (Figure 1aa) with cell line K113 Flanking (Figure 1h)–the K115 Flanking product is approximately 550 bp long, while K113 Flanking product is only 350 bp.

Sequencing

Sample: 86-LTRR (MCF-7, K113 5' LTR)

Because this was a reverse-primed sample, Geneious was used to generate a

reverse-complement sequence (Fig. 2). Since read quality of a sequencing reaction degenerates over time, the last nucleotides recorded were the first of the LTR sequence and the first 5 bases (TGTGG) were lost.

Sample: 72-MCF7-LTR2 (MCF-7, K113 5' LTR)

This was the third sample sent to Berkeley, as the first attempt at a forward-primed MCF-7 sample had failed. The end-of-sequence degradation consumed the last 170 bases of the LTR (Fig. 2), but fortunately this section was well-read by 86-LTRR.

Sample: 56-M115R (MCF-7, K113 5' LTR)

This reverse-primed sample was supposed to be the K115 5' LTR, but aligning its reverse complement with the K115 Reference Sequence (AY037929) showed that it lacked the characteristic 8-bp deletion from 115:T to 122:A (Fig. 2c). A blastn search for part of 56-M115R's flanking region product returned *Homo sapiens* chromosome 19 with an E-value of 4.02×10^{2}, which is the chromosomal locus of HERV-K113. Numerous other species were represented in the blastn results (Fig. 2), suggesting that the flanking area for 56-M115R is a common repetitive sequence–the character of the K113 insertion point.

The evidence indicates that sample 56-M115R actually contains amplified K113 5' LTR, instead of any part of the K115 provirus. If this is the case, then necessarily the band seen in Fig. 1bb were not K115 at all, although Fig. 1aa might contain K115-Flanking amplicon (the corresponding sequencing reaction did not work). Contaminated or mislabeled primer solutions could have lead to the incorrect amplicon. Since K113 and K115 both use the same LTR-R primer, amplifying the incorrect provirus was an unfortunate but entirely possible occurrence. 56-M115R still gave usable data for the K113 alignment, so it can still be considered a success for the study.

Alignments

Alignment 1: K113:AY037928, 56-M115R, 72-MCF7-LTR2, and 86-LTRR

This alignment (Fig. 3) represents all three K113-5' LTR samples collected from MCF-7, compared to the standard reference sequence AY037928 used by

Boller[7], Beimforde[8], Macfarlane[9], Ruprecht[10] *et al*. The end-sequence degradation has been trimmed from 56-M115R and 86-LTRR, but left intact in 72-MCF7-LTR2.

Position	In K113-AY037928:	In M-115R:	In 72-MCF7-LTR2:	In 86-LTRR:
174	G	A	A	A
581	C	T	T	T
812	C	G	G	G

Alignment 2: 56-M115R, K113 5' LTR, and K115 5' LTR

This alignment was used to confirm the identity of 56-M115R as K113, not K115 (Fig. 3a). The 8-bp deletion can clearly be seen in the K115-AY037929 sequence, whereas both 56-M115R and K113-AY037928 have TTAATCTA. Overall, more clear differences can be seen in K115 in comparison:

Nucleotide Position	In 56-M115R:	In K113-AY037928:	In K115-AY037929
155	C	C	T
156	A	A	G
174	A	G	G
179	G	G	A
393	C	C	A
502	C	C	G
581	T	C	T
695	C	C	A
802	T	T	G
812	G	C	C
857	T	T	G
928	G	G	A
	Differences: 3	Differences: 1	Differences: 9

This chart does not list disagreements due to sequence ambiguities (marked as N in each sequence; Fig. 2).

---

[7] Boller, Klaus, Kurt Schönfeld, Stefanie Lischer, Nicole Fischer, Andreas Hoffman, Reinhard Kurth, and Ralf R. Tönjes. "Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles." Journal of General Virology 89 (2008): 567-72. PubMed. Web. 29 July 2009.

[8] Beimforde, Nadine, Kirsten Hanke, Ismahen Ammar, Reinhard Kurth, and Norbert Bannert. "Molecular cloning and functional characterization of the human endogenous retrovirus K113." Virology 371 (2008): 216-25. PubMed. Web.

[9] Macfarlane, Catriona, and Peter Simmonds. "Allelic Variation of HERV-K(HML-2) Endogenous Retroviral Elements in Human Populations." Journal of Molecular Evolution 59 (2004): 642-56. PubMed. Web.

[10] Ruprecht, Klemens, Humberto Ferreira, Aline Flockerzi, Silke Wahl, Marlies Sauter, Jens Mayer, and Nikolaus Meuller-Lantzsch. "Human Endogenous Retrovirus Family HERV-K(HML-2) RNA Transcripts Are Selectively Packaged into Retroviral Particles Produced by the Human Germ Cell Tumor Line Tera-1 and Originate Mainly from a Provirus on Chromosome 22q11.21." Journal of Virology 92.20 (2008). PubMed. Web.

Discussion

Genotyping Study Analysis

The public study's failure cannot be isolated to a known cause at this time. The PCR optimization examined almost every possible element involved: PCR mix, primer concentration, thermal cycler, DNA concentration, &c. The question of DNA was ruled out by using different swabs, sources, and extraction procedures, and three different PCR mix formulations were tested with a small degree of success. With the multiple variables and underlying uncertainty of many results, this study exemplifies the weaknesses of PCR.

The production of Flanking and LTR amplicons with cell line DNA (MCF-7, T47D) and the Applied Biosystems Fast PCR Master Mix is easily the strangest result of the study. Using the undiluted genetic extract for each cell line should not have given a product, as the template DNA should have been far too concentrated, and the Fast PCR Master Mix was not formulated to work with the thermal cyclers at Saint Mary's. The success of these reactions* are unexplainable but very useful to us.

Under ideal circumstances a pilot study would have been performed beforehand, to successfully optimize the extraction and genotyping procedures and scale up the process to accommodate the number of samples we intended to process. However, this would likely have taken the entire time allotted to the research project.

MCF-7 Genotype

One of the reasons MCF-7 was selected for this study was that it is known to produce HERV-K transcripts in culture, primarily for the *env* gene[11]. MCF-7 is derived from adenocarcinomic breast tissue and is a frequent subject of study.

---

* The only known problem that occurred using cell-line DNA and AB Fast PCR Master Mix was the incorrect amplification of K113 5' LTR in Sample M-115R. It can still be considered a useful and productive reaction.

[11] Wang-Johanning, Feng *et al*. "Expression of Human Endogenous Retrovirus K Envelope Transcripts in Human Breast Cancer." Clinical Cancer Research. 2001. <http://clincancerres.aacrjournals.org/cgi/content/abstract/7/6/1553>

Ejthadi *et al*. have confirmed that HERV-K exists in MCF-7; especially of interest is the fact that HERV-K10 expression is elevated in MCF-7 cells stimulated with estrogen.[12] A heterozygous genotype for the cell line means that it can be used as a positive control for future genotyping studies. It can also repeatedly generate both Flanking and LTR product for confirmatory sequencing.

SNPs in K113

Haplotypic Confirmation of Findings by Jha *et al*.

The SNPs 174:G>A and 581:C>T in the MCF-7 K113 5' LTR confirm an earlier genotyping result by Jha *et al*., showing that K113 exhibits both insertional and genetic polymorphic character in different ethnicities[13]. While it was not noted as a polymorphism, 629:C was common to all 4 aligned sequences and is the third of Jha's haplotype-defining SNPs (Fig. 3, this report). The haplotype of 174:A, 581:T, and 629:C is, to date, the only haplotype observed in Caucasian individuals like MCF-7's progenitor.

If anything, this study further proves the point that future K113 pathological studies must take into account the possibility of haplotypic influence. The SNPs observed do not occur in any known functional region of the LTR[14], but their presence indicates the possibility of haplotypic SNPs in functional regions elsewhere in the K113 proviral sequence. If haplotypes of K113 functional regions were found to exist, it could have profound implications for HERV pathogenicity.

Novel SNP Reported in K113

A new SNP was also discovered in the MCF-7 K113 5' LTR: 812:C>G. It occurred in multiple PCR reactions and in at least 2 of the aligned sequences. 812:C>G can be clearly discerned in Alignment 2 (Fig. 3b). It was also bidirectionally confirmed with forward (72-MCF7-LTR2) and reverse-

---

[12] Ejthadi, H. Davari *et al*. "A novel multiplex RT-PCR system detects human endogenous retrovirus-K in breast cancer." Arch Viriol. 2008. 150:177-184. PubMed.

[13] Jha, A.R. *et al*. "Cross sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before *Homo sapiens*." Mol Biol Evol. 2009 Aug 10. [Epub ahead of print]

complement (56-M115R and 86-LTRR) sequences. The 72-MCF7-LTR2 sequence had already registered a few ambiguous bases, but it had not yet entered end-of-sequence degradation.

This SNP does not occur within Jha's functional regions, but 812:C>G is only 10 bp downstream from the Poly-A signal[14] It will be interesting to explore further the 812:C>G SNP. It could have functional significance for the level of HERV transcription, and possibly relate to the transformation process of the original tumor. However, the SNP may have originated during the in vitro culturing of the cell line over time.

Conclusion

As the most intact examples of the HERV-K family, K113 and K115 merit much further attention. As shown in this study, the nature of these proviruses is much more complex than previously thought. Since the occurrence of haplotypes has been confirmed, future genotyping studies must also screen for and consider rates of haplotype distribution in the population. Sequential polymorphisms can be assayed for functional impact, both in the production of retroviral particles and in the immunological implications of carcinogenic HERV pathology.

[14] Jha, A.R. *et al*. "Cross sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before *Homo sapiens*." Mol Biol Evol. 2009 Aug 10. [Epub ahead of print]