UPPSALA
UNIVERSITET

Identification of Long Interspersed Nuclear Elements in the chimpanzee genome

Ludwig Hedberg

Degree project in biology, Master of science (1 year), 2011
Examensarbete i biologi 30 hp till magisterexamen, 2011
Biology Education Centre and Department of Immunology, Genetics and Pathology, Uppsala University
Supervisor: Lars Feuk

Table of contents

Summary

Long Interspersed Nuclear Elements or LINEs are a class of repetitive DNA that is common in all mammalian genomes and comprise approximately 20% of the human genome sequence. The aim of this study was to identify and validate novel LINE sequences in the chimpanzee genome. Different experimental validation strategies were evaluated and applied to candidate LINEs that were based on computational predictions. Our experiments with PCR and Sanger sequencing shows that in 8 of 10 regions tested, we were able to validate the presence of a LINE element in the region predicted from the computational analysis. For 7 of the 8 validated regions, our analysis verified that it belonged to the same LINE family as predicted.

Introduction

Genomes

In 2001 the first draft sequences of the human genome were completed with articles in Nature (1) and Science (2) from the Human Genome Project consortium and Celera, respectively, making headlines around the world. The first sequence draft had covered more than 95% of the genome but lacked completion of the regions with repetitive DNA and the sequences close to the telomeres. It took until 2004 (3) before a majority of the assembly gaps had been filled in. Both the original draft sequence and the finishing sequence added in 2004 were carried out with cloned DNA from bacterial artificial chromosomes that were then sequenced using primarily Sanger sequencing. A lot of knowledge about the genome was gained from filling in the gaps and improving the assembly from 2001 and 2004. As an example, the estimation of the number of active genes was changed from 30-40,000 in the draft sequence to 20-25,000 in the finished assembly.

Once the sequencing of the human genome was completed, several other mammalian species were sequenced. An important continuation of the human genome project was the sequencing of the chimpanzee (*Pan troglodytes*) genome which was completed in 2005 (4). The main focus of the analysis performed in the initial publication of the sequenced chimpanzee genome was to compare it to the human genome. The chimpanzee constituted a closer relative to humans compared to other species sequenced at that time. The similarity between humans and chimpanzees can differ from 95% to 99% depending on how it is calculated. It is 95% if insertions and deletions are included but 99% if it is calculated purely based on single nucleotide differences (4). Reference assemblies have been updated as more information has become available, and the current well annotated assemblies for human and chimpanzee are called HG18 (build 36) and PanTro2, respectively (5).

Retrotransposons

The human genome consists of about 3 billion base pairs (haploid size) set in 23 chromosomes pairs, approximately half of this is made up of repetitive DNA sequence (1). Repetitive DNA can be separated into different classes depending on their sequence and distribution. One important class of repeats is called transposons, and includes two subclasses called DNA transposons and retrotransposons. DNA transposons are a class of elements defined by their ability to move around the genome as DNA and insert themselves into new sites in the genome (6). These transposons are not active in the human genome but have been during primate evolution (7). Retrotransposons represent a different subgroup, and have the ability to relocate in the genome in a different way than DNA transposons. Retrotransposon contain genes that allow them to be transcribed into RNA and then reinserted into the genome at a new location by use of reverse transcriptase (6). The retrotransposons can then be further divided into groups that contain or lack long terminal repeats (LTRs). In humans the LTRs originate from retroviruses, in the form of human endogenous retroviruses (HERVs). LTRs make up about 8% of the human genome but have a very limited activity if they are still active (8). The other group is the non-LTR retrotransposons that can be divided into SINEs (Short interspersed nuclear element) and LINE (Long interspersed nuclear element).

The most common type of SINE is called an Alu repeat. Full length Alu elements are approximately 300 base pairs (bp) long and are built up by two sequences that flank a region rich in adenines. There are around 1 million copies of Alu elements in the genome, which equates to ~10.5% of the genome (table 1). The name of the Alu element comes from the fact that it is recognized by the Alu restriction endonuclease. Alu elements are still mobile in the human genome; however they do not encode for any proteins and are thus dependent upon the machinery from other retrotransposons to retranspose (9).

Table 1: Comparison of the fraction of the genome made up of different classes of transposons

Type of Transposon	Procentage	Full length size
LINE-1	17%	6kb
LTRs	8%	Mixed
Alu	10,5%	300bp
Other transposons	17%	Mixed

LINE-1 (L1) is the biggest group of transposons and makes up about 17% of the human genome with its 500000 copies (1). The number of LINEs has thus grown significantly since its original insertion into the eukaryote ancestor genome approximately 150 million years ago. The full length LINE is 6 kilobases (kb) and has two open reading frames (ORFs), including a 5’ UTR, a 3’ UTR downstream and a poly-a tail signal. The ORFs encode a reverse transcriptase with endonuclease ability and a RNA binding protein (10). These proteins make it possible for the LINE to retranspose in the genome without exogenous factors. However, most LINEs do not make it unharmed when inserting in a new position in the genome and are truncated, mutated or rearranged to render them inactive. There are about 100 LINEs in the human genome that still have the ability to actively transpose (11).

Retrotransposons have been shown to have an impact on primate’s genomes by mediating inversions and deletions (12). Based on their proven role in genome evolution and rearrangements, it is important to characterize the location and type of transposable elements in human and primate genomes. Better maps of these elements will help elucidate the differences between primates and how and when these changes occurred. As an example SINEs have been much more active in humans than Chimpanzees since our last common ancestor (4).

Next generation sequencing

Next generation sequencing is a term used to describe the novel sequencing techniques developed after the completion of the human genome project. There are several technologies that differ in the chemistry of the sequencing reaction, but they are all characterized by a significantly higher throughput and lower cost per base sequenced, as compared to traditional Sanger sequencing. In 2004 Roche/454 released the first modern high-throughput sequencing machine. Soon after, Applied Biosystems SOLiD (Sequencing by Oligonucleotide Ligation and Detection) and Illumina Solexa platforms came onto the market each with different approaches for making sequencing faster and more affordable. The lower cost of sequencing and higher yields have made it possible not only to sequence the entire genomes of many organisms, but also to sequence individuals of a species and thus gaining additional insight into genome variation.

The basis for the SOLiD system are emulsion PCR to amplify the chosen fragments to be sequenced. The sequencing works by incorporating a pentamer on each cycle where the two first bases are read by fluorescence. The universal primer that starts each sequencing run are of precise length that is shortened with 1 base between different primer rounds thus creating an overlap and reading every base more than once by different pentamers. Overlapping the pentamers creates a proofreading element to the system and thus a higher degree of certainty. The use of ligase and the pentamers instead of a conventional polymerase is unique to the SOLiD system. SOLiD is under rapid development and from 2008 to 2011 the length of reads has increased from 25 to 75bp. SOLiD sequencing can be based on different types of libraries (e.g. single fragment, mate-pair and paired-end) and depending upon which of these are used the maximum length will differ.

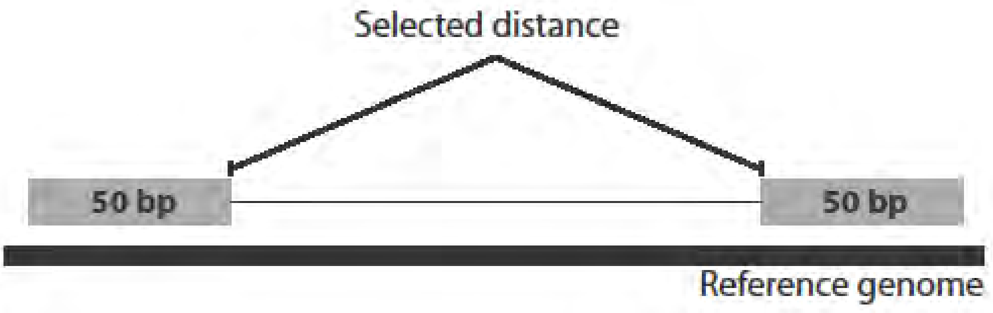


Figure 1; Mate-pair with a length of 50bp each, located at a known distance from each other and with both mates mapping to the reference genome.

Mate-pairs are sequences that are the outmost bases on the fragment being sequenced (see figure 1). The approximate length of the fragments is known, and when mapping the two mate-pair reads to the reference genome they would be expected to map at the expected fragment distance. Based on this expected mapping, strategies have been developed to identify deletions and insertions in the donor DNA, by searching for clusters of mapped mates where the observed distance differs significantly from the expected distance. The mapping information can also be used to find inversions if one mate in the pair maps to the reference in an inverted orientation. Finally, smaller insertions, deletions and single nucleotide variants can be detected by identification of mismatches or small gaps within mapped reads.

Project overview

My project was a part of a larger study of how LINEs have transposed in the chimpanzee genome when compared to the human genome. In this project a method to use mate-pair reads from SOLiD sequencing in which only one mate have mapped uniquely to the reference genome was implemented. The read in the mate-pair that did not map to the genome (called orphan) would be re-aligned against a computationally constructed reference sequence made out of all human specific LINEs. This reference sequence was based on all full length LINEs that are unique for the human genome when compared to the chimpanzee genome. The orphans that indicated the existence of a new LINE in the chimpanzee were further filtered to create a candidate list of putative novel LINEs. This list was the starting point for a curation and validation effort to identify novel LINEs in the chimpanzee.

Computational analysis

The data that was the starting point for this project was handled by Johan Stenninger as part of his master thesis in bioinformatics (13). I will not go into details on all the work he did but a brief outline and a summary of our joint work will be given to provide an understanding of the data that was the basis for the current project.

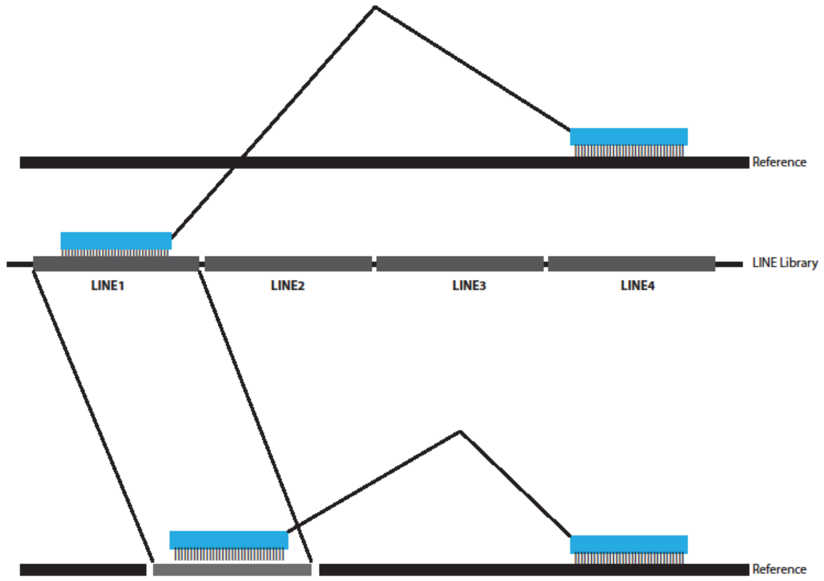


Figure 2: Mapping an orphan to the constructed reference of LINES and moving it to a position in the Chimpanzee reference.

The starting point for all the data analysis was three mate pair libraries from two different chimpanzees sequenced by SOLiD. The resulting data was mapped to the reference genomes for chimpanzee and human. Since the focus was to find new full length LINEs, an artificial reference sequence with human specific full length LINEs was constructed. The idea was to map mates from orphan-pairs to this library to see if the orphan reads mapped to LINEs not present in the chimpanzee reference (see figure 2). An orphan-pair is a pair in which one mate have not mapped, but the other mate has mapped correctly somewhere in the reference. The mate that was not orphan, i.e. already mapped to the reference sequence, could then be used to identify a putative insertion point for the LINE element into the reference assembly. All mates mapping to LINEs from the library were later grouped together into small clusters based on their location in the chimpanzee reference.

When a cluster was found, both the type of LINE and the insert position that this cluster predicted had to be calculated. The position at which a mate mapped at in the LINE was used to determine an offset position in the LINE so that a putative starting position for the LINE could be determined (Figure 3). The clustering in the previous step was of great assistance since it allowed us to get rid of artifacts and to achieve a better overall prediction of the insert position. Since the type of LINE was available from the artificial reference of LINEs this was used as well to reach an estimate of the length of the element. The measure of both start and length meant that a fictional LINE could be presented in the genome browser at UCSC (14).

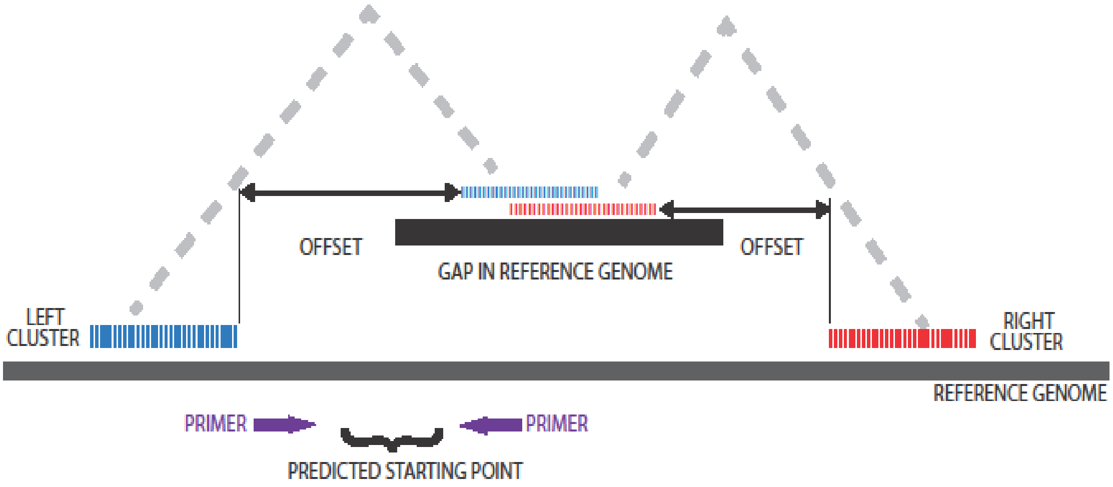


Figure 3; Clusters on both sides of a gap predicts that it covers a LINE and provides a predicted start and end of the LINE. The offset is the difference between where in the LINE the orphan maps and where the other mate in the mate-pair is located. Primers were designed to create one PCR spanning across the predicted starting point and another across end point.

Aims

Sequencing of a complete genome yields large amounts of sequence data that does not map to the reference assembly. Some of this data comes from mate-pairs where only one of the mates has mapped to the reference. The overall aim of the research project was to use this data to identify and validate LINEs in the chimpanzee genome that is not present in the current reference assembly. Another aim of the project was to fill existing gaps in the reference assembly in the cases that the gap corresponds to a LINE element. The aim for the experimental part of the project described in this thesis was to filter computationally predicted LINEs and experimentally validate novel LINEs to achieve a better annotation of these elements in the chimpanzee genome.

Material and Method

Bioinformatical analysis

During the initial survey of the predicted LINEs obtained from the orphan read mapping it was clear that additional information was required to be able to adequately interpret the data. The number of orphan reads mapping to each LINE was used to establish the sequence coverage in each LINE and a higher ratio of coverage in the LINE was used as a means to prioritize targets. Using coverage as a criterion for filtering significantly shortened the list of validation targets, but still the specific site where the putative novel LINE would be inserted into the reference genome was ambiguous. Every mate pair in a cluster was used to predict where in the reference sequence a novel LINE would be inserted. This location was predicted using the information about where the mate had mapped within the human LINE. As outlined in Figure 2, the mapping within the LINE in combination with knowledge about the approximate distance between the mate pairs makes it possible to predict an insertion interval. The insertion point can never be exactly determined as there is some uncertainty about the insert size in the sequenced library. However, by using all of the different the predictions the uncertainty could be substantially reduced in most cases. A relatively small predicted insertion interval is required to be able to design validation experiments.

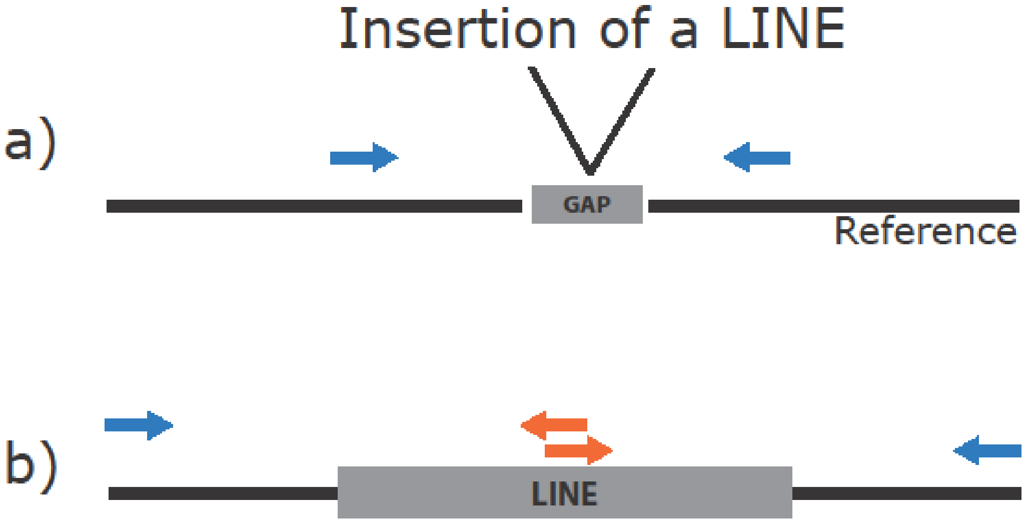


Figure 4 a) Often, a gap or other poor sequence is located where the LINE has been predicted to be making short PCR difficult. The problem is that a PCR spanning across the entire region would be 7-10kb long. b) By using the human reference and information on the type of LINE that is there in humans, primers could be designed within the LINE and generate much smaller PCR products.

Before designing assays for validation of predicted LINEs, a strategy for how to perform the PCR for validation had to be chosen. The proposed methods were a long-range PCR spanning the entire predicted LINE or, alternatively, to design a shorter PCR where one primer was situated in the LINE and the other just outside the LINE. As a long range PCR would yield the full LINE element, which could then be further validated by sequencing, this approach was initially chosen. Significant effort was put into optimizing reactions to amplify the full length LINEs in several regions. However, this turned out to be very challenging. Even after extensive optimization, it was difficult to successfully amplify targets over 5kb in length within regions of highly repetitive sequence. Based on these results, it was decided to change to a different strategy based on shorter PCRs. New primers were designed for most regions. Not all regions that were targeted with long-range PCR were amenable to short PCR designs, and therefore not all previous targets could be included. With the shorter PCR method more than one assay had to be designed for each region in order to increase the rate of validation and in order to validate that the LINE element was indeed a full length LINE.

For the short PCR product approach, the primers were designed with one primer in the reference sequence before the predicted LINE insertion site and one primer within the LINE. In the instances where it was considered likely that no novel LINE insertion had occurred and the novel LINE had been predicted due to poor annotation or gap in the chimpanzee reference sequence, primers were made from both sides of the LINE in order to determine where the LINE started and ended. When possible, the primers situated in the middle of the LINE were overlapping so that a full sequence could be obtained in a future Sanger sequencing. When this was not possible an effort was made to try and keep the distance between them as short as practically possible. All primers were designed by initially using the human reference sequence to minimize the effect of gaps in the chimpanzee genome. The primers were designed with the aid of Primer3 (15). Primer3 is a free software program that facilitates the choice of possible primer pairs from a given sequence. Primer3 takes into account everything from melting temperature, GC content to possible secondary structure of the primers and how complementary the primers are to each other.

The primers that had been designed with the aid of Primer3 were then tested with *in-silico* PCR. *In-silico* PCR is a theoretically made PCR where the primers are used on a reference genome to see if a product could be obtained. The *in-silico* PCR at UCSC (University of California, Santa Cruz) will give no result both when too many products or no products would be the outcome. This is important to note since there is a large number of highly identical LINEs in the primate genome, making both outcomes are highly likely. The primers that passed testing by *in-silico* PCR were then further investigated by using BLAT (BLAST Like Alignment Tool) (16) and to align them against both reference genomes. BLAT is a program that aligns the tested sequence against a selected reference to see if the sequence is contained within the reference. BLAT was used for chimpanzee as a complement to *in-silico* PCR, because some primers were located in gaps in the sequence and no results would were then obtained from *in-silico* PCR. By using BLAT to ensure that the primer did not align to multiple other places in the chimpanzee genome, an assumption could be made that any PCR product would be the one we wanted if at least one primer was unique (even if the other was located in a LINE)

These repetitive sequences can create problems for BLAT as BLAT can give the result that the primer exists all over the genome. BLAT tries to match the entered sequence to the chosen reference genome. A limitation that is noteworthy especially when trying to align primers is that it will not align sequences shorter than 20bp. This means that a 20bp primer with one mismatch will yield no matches. For shorter sequences, BLAST can be used instead of BLAT, but this algorithm is significantly slower than BLAT.

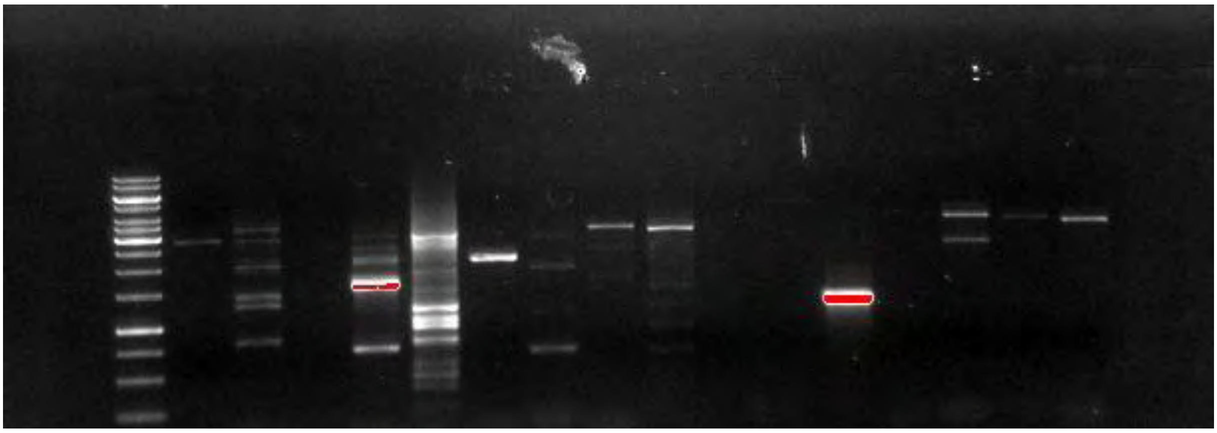
Experimental validation parameters

The long range PCRs were preformed with AccuTaq (Sigma-Aldrich) (0.05 units/µL) run with AccuTac LA buffer (1x), dNTP (500 µM for each), 5% dimethyl sulfoxide, 400nM of each primer and 40ng of template DNA. The reaction was run with 10mins at 50C° followed by 30s at 95C° and then cycled 35 times with 15s denaturation at 94C°, annealing at 60C° for 20s and extension at 68C° for 15min. To end the PCR a final extension at 68C° for 10 min was applied. The short PCRs were preformed with PCR buffer (1x) from Qiagen, 0.5mM MgCl, dNTP (200 µM for each), Qiagen Hot Star+ Taq (0.005U/ µL) and run with a hot start at 95C° for 5 mins and then cycled 35 times with a denaturation at 94C° for 30s, annealing at 60C° for 30s and extension for 2mins at 72C°. A final extension for 7mins at 72C° ended the program.

Extraction of DNA that had been gel purified was performed by using Qiagens Qiaquick gel extraction kit for DNA extraction according to the manufacturer's protocol. The DNA was then further purified by ethanol precipitation and the pellet was dissolved in water.

Results

From the computer analysis 151 candidates were extracted and out of these 64 were analyzed further with regard to existing annotations in the region, sequence coverage of the LINE and size of the insertion point interval for a potential novel LINE. These 64 were the regions that had the most supporting reads from the clustering analysis. From the evaluation of the 64 regions the top 20 were selected for further study to select targets that covered all possibilities of results and to ensure that the sequence context around the insertion point made a PCR design possible. The challenges of designing PCR reactions come from the fact that in these regions unique sequences suitable for primer placement are scarce. These regions contain many gaps and repetitive sequences and in some cases it was not possible to find unique sequences for primers.



. Figure 5; Testing the primers in human showed that 12 out of 16 PCRs worked.

The primers were first tested on human DNA to see if they would give a single band or if multiple products would be obtained. Different PCR protocols were tested to minimize background amplification and obtain a clear band of the expected size. After the primers had been tested on human DNA they were run with chimpanzee DNA. All primers were tested with chimpanzee DNA even if they had not worked with human DNA, as there might be individual and species differences leading to PCR failure. Primers that did not work on human DNA also did not work with chimpanzee, as can be seen in figure 5. The results from the first test confirmed the length that had been anticipated in 8 out of 12 cases. All the PCRs were run in bigger volume so that enough DNA was obtained for subsequent sequencing experiments. The DNA was gel purified unless only a single band could be observed in the gel. This also allowed us in one case to try and sequence more than one band (which had two bands in the right size range). When more than one band was present in the gel the strongest band closest to the expected size was excised from the gel.

Table 1: Regions that were experimentally tested for LINE insertions and summary of results. *For this two bands were excised from the gel and only one worked and is presented.

Chromosome	Start	Stop	Side of LINE	Primer direction	Number of bp	Validating	Correctly Predicted LINE
chr11	21511262	21517569	Left	Forward	-	No	-
				Reverse	480	Yes	Yes
			Right	Forward	-	No	-
				Reverse	-	No	-
chr13	65578438	65584151	Right	Forward	210	Yes	-Yes
				Reverse	200	Yes	-Yes
chr19	20423257	20429362	Right	Forward	-	No	-
				Reverse	-	No	-
chr3	22982134	22988331	Left	Forward	180	Yes	Yes
				Reverse	760	Yes	Yes
			Right*	Forward	550	No	Yes
				Reverse	550	No	Yes
chr3	121208854	121215306	Left	Forward	760	Yes	Yes
				Reverse	680	Yes	Yes
chr3	142372394	142378430	Left	Forward	250	Yes	-
				Reverse	400	No	-
chr5	79163950	79170099	Left	Forward	725	Yes	Yes
				Reverse	710	Yes	Yes
chr6	111998069	112004205	Left	Forward	610	Yes	Yes
				Reverse	240	Yes	Yes
chr7	151119811	151125652	Right	Forward	635	Yes	Yes
				Reverse	780	Yes	Yes
			Left	Forward	280	Yes	Yes
				Reverse	125	Yes	Yes

The sequencing experiments worked well enough for further analysis for 20 out of 26 reactions (table 1). These were aligned with BLAT against both the human and chimpanzee genomes. Alignment against only the chimpanzee does not always work as the sequenced DNA might correspond to a gap in the reference assembly. Out of 10 regions tested, 7 were confirmed to contain the LINE we had predicted from the computational analyses. One additional LINE could be verified but the type could not be confirmed to be the same as the one that was predicted. The sequences that could not be successfully aligned against chimpanzee due to gaps were tested against the constructed reference of human LINEs that had been used to find targets for validation. Of the two that were not validated one gave no results in the sequencing and for the other only one primer worked, but the resulting sequence was too short to reach into the putative LINE.

Discussion

Main findings

In starting this study the hypothesis was postulated that the similarity between human and chimpanzee could be used to identify new LINEs in the less explored chimpanzee genome. What was found instead was that most LINEs identified were identical in the human and chimpanzee, and that they were lacking from the chimpanzee reference sequence due to poor assembly and annotation of the chimpanzee genome. The results show that the strategy of using orphan reads to try and extract more knowledge from new sequence information is a good proposition and additional new methods using this general strategy should be set up.

Annotation and comparison

The consensus right now is that the sequence similarity between human and chimpanzee is around 98%. However, the similarity depends on how the comparison is made, as the human reference genome has been worked on extensively to fill gaps and complete the sequence. A comparison of the current assemblies would yield a much lower similarity between human and chimpanzee, as o the chimpanzee reference genome has approximately 250,000 gaps as compared to 387 for the human reference. There are also considerable differences in where the gaps are located. In the human genome there are about half as many bases predicted in gaps, but as seen in Table 2 these are of a considerably larger size.

Table 2: Number of gaps and total gap length in human and chimpanzee.

	Human		Chimpanzee	
Length cutoff	# Gaps	Total length	# Gaps	Total length
<100000	293	12 338 192	246 269	127 541 627
None	387	226 162 028	246 448	440 962 440

The misalignment that occurs because of gaps makes direct comparisons in gap regions unreliable at best, and this is definitely an issue in regions around LINEs and other repetitive elements. The unreliability also goes over onto all annotations that have been made to the reference. In many places a program called RepeatMasker (17) has used consensus sequences for transposons and other repetitive DNA sequences to annotate them in the reference genome. A problem is that this program can often produce a very fragmented picture of how a particular stretch of the genome looks as evident in Figure 6. Examination of both the human and chimpanzee sequences simultaneously leads to a very different picture. The sequences in figure 4 are the same if manually annotated using BLAT and not using the Liftover (a tool that translates sequences between different assemblies) and Repeat masker. After having completed this project I propose that an addition should be made to the RepeatMasker program. When the first run and annotation has been made a second run should be made where the program will try to merge LINE fragments into larger elements. In this second run a comparison with similar species should be done as well.

Assay design

To be able to design primers for the assay in this project the human reference genome had to be used, this was because our targets were not present in the chimpanzee reference. One of the problems with this is that many assumptions have to be made. Even if human and chimpanzee genomes are close they are not the same. After the initial screening several targets were removed as a primer pair that was specific enough could not be designed for them. One of the problematic areas is that this study is focused on regions that have many repetitive sequences and many of these are known to often be truncated, mutated or inversed. Because of all these challenges it was not surprising to see more than one band occur in several of the lanes after the PCR.

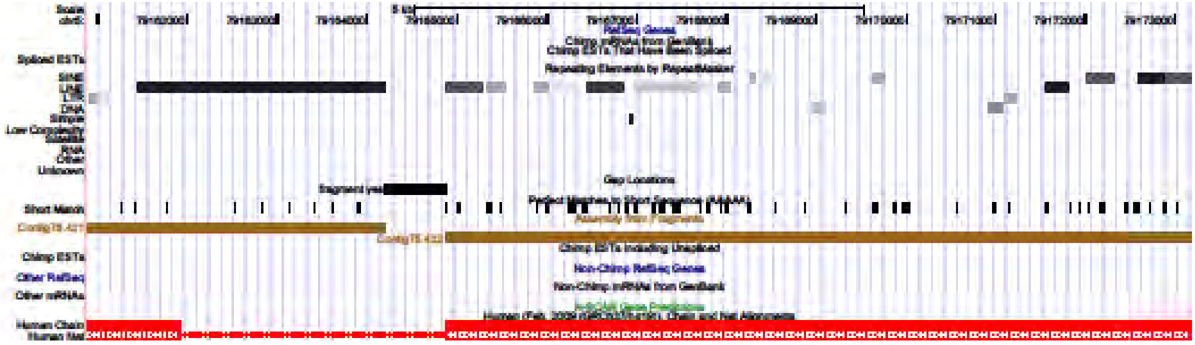


Figure 6; Repeat Masker has identified fragments of several LINEs instead of one of full length LINE due to a gap located inside it. The same gap has resulted in Liftover not being able to accurately translate the sequence coordinates from the human to the chimpanzee assembly even though BLAT can do this. *http://genome.ucsc.edu*

Conclusions

The results show that the similarity between humans and chimpanzees might be slightly larger than what has been suggested, as many of the LINEs we identified turned out to be assembly errors or poor annotation in the chimpanzee reference sequence. The study also shows that there are possible strategies to use left-over data from next generation sequencing to identify novel insertion variants. The experimental work also highlight that methods that are often considered routine might be very challenging when taken towards the edge of their capability.

Acknowledgements

I would like to express my gratitude to Lars Feuk for the opportunity to do this study in his group and to the other members of his group, Anna Johansson, Ammar Zaghlool, Jonatan Halvardsson and Johan Stenninger for all the support and help. Also great thanks are also due to Uppsala Genome Center for help with the sequencing and assistance in the laboratory.

References

1. **Lander, E. S.** Initial sequencing and analysis of the human genome. *Nature.* February 15, 2001, Vol. 409, p. 860-921.

2. **Venter, J. C.** The Sequence of the Human Genome. *Science.* 2001, Vol. 291, p. 1304-1351.

3. **Consortium, International Human Genome Sequencing.** Finishing the euchromatic sequence of the human genome. *Nature.* 2004, Vol. 431, p. 931-945.

4. **Consortium, The Chimpanzee Sequencing and Analysis.** Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature.* 2005, Vol. 437, p. 69-87

5. **Kent W. J., Sugnet C. W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D.** The human genome browser at UCSC. *Genome Research.* 2002, p 996-1006.

6. **Craig, N.L., Craigie R., Gellert, M.,Lambowitz, A. M.** *Mobile DNA II.* Washington : American Society for Microbiology, 2002.

7. **Pace J. K., Feschotte, C.** The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Research.* 2007, Vol. 17, p. 422-432.

8. **Mills, R. E., Bennet, E. A., Iskow, R. C., Devine, S. E.** Which transposable elements are active in the human genome? *Trends in genetics.* 2007, Vol. 23, p. 183-191.

9. **Dewannieux, M., Esnault, C. Heidmann, T.** LINE mediated retrotransposition of marked Alu sequences. *Nature Genetics.* 2003, Vol. 35, p. 41-48.

10. **Swergold, G.D.** Identification, characterization and cell specificity of a human LINE-1 promoter. *Molecular Cell Biology.* 1990, Vol. 10, p. 6718-6729.

11. **Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., Kazazian, H. H.,.** Hot L1 account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Science.* 2003, Vol. 100, p. 5280-5285.

12. **Lee, J., Han, K., Meyer, T. J., Kim, H-S., Batzer, M. A.,.** Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLoS ONE.* 3, 2008, Vol. 12, doi:10.1371.

13. **Stenninger, J.** Uppsala Universitet. [Online] [Cited: 03 10, 2011.] http://www.ibg.uu.se/upload/2011-03-09_095100_093/final%20complete%20thesis_johan_stenninger_110309.pdf.

14. **Fujita P. A., Rhead B., Zweig A. S., Hinrichs A. S., Karolchik D., Cline M. S., Goldman M., Barber G. P., Clawson H., Coelho A., Diekhans M., Dreszer T. R., Giardine B. M., Harte R. A., Hillman-Jackson J., Hsu F., Kirkup V., Kuhn R. M., Learned K., Li C. H., Meyer L. R., Pohl A., Raney B. J., Rosenblo.** The UCSC Genome Browser database: update 2011. *Nucleic Acids Research.* 2010, D876-882.

15. Primer3 Input. [Online] 03 10, 2011. http://frodo.wi.mit.edu/primer3/input.htm.

16. **Kent, W. J.** BLAT - the BLAST-like alignment tool. *Genome Research.* 2002, p. 656-664.

17. **Smit, A. F. A., Hubley, R. & Green, P.** RepeatMasker Open-3.0. [Online] 03 10, 2011. http://www.repeatmasker.org.

18. **Nuo Yang, Haig H. Kazazian, Jr,.** L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nature structural and molecular biology.* 2006, Vol. 13, p. 763-771.

19. Sequencing Instruments Products at Applied Biosystems. *Applied Biosystems.* [Online] [Cited: 03 10, 2011.] https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=catNavigate2&catID=607060.