Research Topics in Bioinformatics

Aarhus University

Bioinformatics Research center (BIRC)

Identification and characterisation of LTR-retrotransposons in the pig genome

Project report

Gernot Wolf

2010

Table of Contents

Summary

Endogenous retroviruses (ERVs) are prevalent in the genomes of all mammalian species examined so far. These elements represent germline integrations of exogenous retroviruses that persisted in their host organisms through the ages and became a perpetual part of the genome. Most ERVs found are inactive due to the accumulation of mutations and deletions that cause open reading frame interruptions and defect genes that code for non-functional proteins. However, in some animals, like mice and pigs, replication competent ERVs that are able to form infectious particles were identified. Some porcine endogenous retroviruses (PERVs) were found to be capable of infecting human cells in culture and immunodeficient mice, and therefore represent a major concern in Xenotransplantation of pig organs. Although these group of replication competent PERVs is well studied, little is known about other retroviral elements in the pig genome. The full sequencing of the pig genome is almost complete, so we aimed at analysing the retroviral load of the pig genome using conventional BLAST search approaches and two recently developed *de novo* LTR-retrotransposon/ERV predicting programs, *LTRharvest* and *LTRdigest*. We identified numerous full length retroviral genomes of which only partial sequences of the *pol* gene were known before and identified new primer binding sites and retroviral domains in these elements. In addition we describe two new groups of PERVs that contained numerous retroviral features but showed no homology to any published PERV sequence. Finally we estimated integration times of various PERV groups using an approach that is based on the molecular clock hypothesis. This study should give new insights into the diversity, structural organisation and evolutionary history of endogenous retroviruses in the pig genome.

1. Introduction

1.1 Life cycle and genome organisation of retroviruses

Retroviral particles consist of an encapsidated dimer of positive-sense single stranded RNA, enclosed in a capsid, which in turn is enclosed in a lipid bilayer envelope. After receptor- binding to a target cell, the virus enters the cytoplasm where it begins to uncoat and reverse transcribe its RNA to DNA. Reverse transcription is primed by a cellular transfer RNA (tRNA), binding to the viral primer binding site (PBS) which is complementary to the 3´end of the tRNA. The Polypurine tract (PPT) is a short stretch of purine-rich DNA that is not degraded by the Rnase H activity of the viral polymerase after synthesis of the first DNA strand. The second strand amplification is then primed by this strech of double stranded DNA.

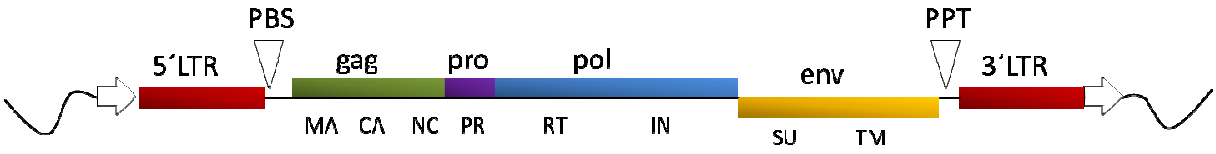


Figure 1: Genomic organisation of an integrated Provirus. Arrows indicate 4-6bp duplication sites that flank the provirus at both ends. The coding region normally consists of two open reading frames that are transcribed and translated to a gag-pro-pol polyprotein and the envelope (env) protein. The gag-pro-pol polyprotein is processed into structural (gag) and enzymatic active (pro, pol) proteins. LTR: long terminal repeat; PBS: primer binding site; MA: matrix; CA: capsid; NC: nucleocapsid; PR: protease; RT: reverse transcriptase; IN: integrase; SU: surface protein; TM: transmembrane protein.

Finally, the double-stranded DNA is integrated into the host genome. As an integrated provirus, the viral genome is transcribed by host cell encoded factors, starting from the 5´Long Terminal Repeat (LTR) that contains binding sites for transcription factors and enhancer elements (Figure 1). The 5´LTR and 3´LTR are identical after integration but have different functions. Whereas the 5`LTR serves as a strong promotor, the 3´LTR provides polyadenylation sites, necessary for mRNA stability. LTR-retrotransposons have a similar genomic organisation to endogenous retroviruses. The internal region is flanked by two LTRs and codes for *gag* and *pol* proteins. PBS and PPT sequences are found at the same position as on ERVs. Importantly, LTR-retrotransposons do not code for an *env* gene and therefore rely on a strictly intracellular replication cycle. Because of the high similarity of LTR-retrotransposons and ERVs these terms are often used synonymously.

1.2 The colonisation of the mammalian genome by ERVs

Because of the ability of retroviruses to stably integrate into the genome, it is possible that chromosomal integration also occurs in cells of the germline. Offspring that develope from these infected germ cells will carry the integrated provirus as part of their own genome and pass it on to the next generation in a mendellian manner. These Retroviruses are then referred to as endogenous retroviruses (ERVs). ERVs that have a fitness-reducing effect will be eliminated by selection over time whereas neutral ERVs may get fixated in a population and eventually in all members of a given species. When a species with a fixated ERV insertion splits up into new species, these will also harbor the ERV. There are also numerous examples of ERVs that have a positive effect on the host and that are therefore conserved in different species over time [1]. Furthermore ERVs, together with other transposable elements, have an important impact as drivers of genome evolution [2, 3]. It is believed that retroviruses started to populate mammalian genomes over a 100 million years ago, indicated by retroviral integrations that are shared by all mammalia lineages. Through evolution most of these integrated proviruses will eventually accumulate deletions and mutations since they are not subjected to purifying selection, therefore most ERVs will lose the ability to replicate relatively fast. However certain ERVs are still capable of at least some level of expression and replication within the host genome even after several tens of millions of years [4]. This is partially explained by the observation that ERVs can be complemented with retroviral proteins *in trans* by other copies of the same ERV species or by infecting exogenous retroviruses. Therefore ERVs that do not code for functional enzymes can still replicate providing that their essential regulatory sequences such as PBS, PPT, LTRs and the packaging signal (ψ) are intact.

In humans, about 8% of the genome is made up of the fossils of LTR-retrotransposons and endogenous retroviruses [5], most of them entered the germline after the separation of Old and New World monkeys (30-45 Myr ago) [6]. So far no replication competent ERVs have been found in humans or other primates but one subfamily (HML-2) of the HERV-K family is believed to have been still active in recent history [3]. In other mammalian species like the mouse, pig or koala replication competent ERVs that can form infectious particles were found [7-9].

1.3 Porcine endogenous retroviruses

Porcine endogenous retroviruses (PERVs) aroused special interest when it was shown that viral particles derived from PERVs are able to infect and replicate in human cells [9-11]. It has been also shown that PERVs can infect immunodeficient mice after transplantation of fetal pig pancreatic cells [12]. This is a major concern since Xenotransplantations from pigs to humans are widely used and alleviate the shortage of human donor organs. Unlike other infectious agents which can be eliminated by pathogen-free breeding, PERVs were found in all animals tested so far and cannot be simply removed. Although so far no trans-species infections of PERVs to humans in many *in vivo* porcine cell line or organ transplantation trials was shown [13] major concerns remain about the safety of Xenotransplantation of porcine organs. Three replication-competent gammaretrovirus subgroups of PERV (PERV-A, -B and C) have been identified in the pig genome [8, 10]. These subgroups share homologies in their *gag* and *pol* genes but differ in the receptor binding domain of the *env* gene. The variation in the *env* gene determines the different receptor usage and thus host restriction. Whereas PERV-A and PERV-B can infect human cells, PERV-C is mainly restricted to pig cells [14]. These three gammaretrovirus subgroups are relatively young ERV family, Screening of several species closely related to pigs revealed that PERV-A entered the pig lineage in the Miocene epoch (3,5 to 7,5 MYA) followed by PERV-B that was detected in a species (*P. larvatus*) that split from *S. scrofa* in the early Pleistocene epoch. PERV-C is the youngest member of this family and is believed to have entered the genome in the early Holocene epoch (0,1 to 1,5 MYA) [15]. Controversially, in another study it was reported that PERV-C is absent in wild boars and some domesticated pig strains indicating that PERV-C entered the pig genome only after the domestication of pig by humans occurred [16]. Interestingly, it seems that wild boars harbor less copies of all three PERV subclasses than domesticates pigs, indicating that PERV copy numbers are increased by inbreeding [16]. It has been also shown that PERV-A and PER-C can recombine and form infectious particles that can infect human cells [17].

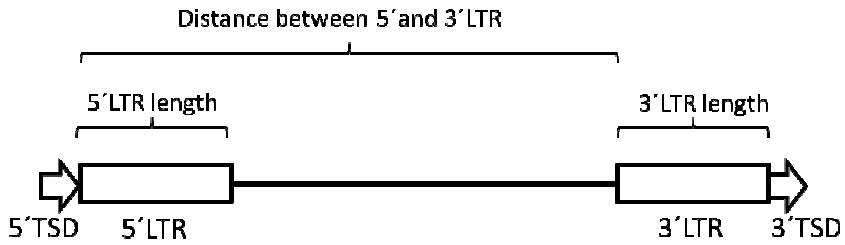
Due to the potential risk of trans-species infection of PERV-A, -B and C, this gammaretrovirus group was extensively studied but very little attention was paid to other groups of porcine endogenous retroviruses. The presence of other endogenous gamma- and betaretoviruses in the pig was shown by PCR amplification of partial retroviral

sequences with degenerated primers and low annealing temperatures but apart from these short sequences very little is known about these PERV groups [16, 18-20].

1.4 Computational detection of transposable elements

A large number of computational tools for automated annotation of transposable elements like endogenous retroviruses and other repeat families has been developed. *Repeatmasker* screens DNA sequences for repeating elements using precompiled sequence libraries and special scoring matrices [21]. Although useful in many respects *Repeatmasker* and similar programs have some major disadvantages. First of all, a precompiled sequence library is needed to recognise transposable elements meaning that only elements with known sequence composition or elements closely related to known sequences can be detected. This may not be a problem in well studied model organisms like mouse or *Drosophila* since extensive repeat libraries are available for these animals [22]. However, most of the transposable element families are lineage specific and are not represented in species which are not closely related and even if, these elements might be highly divergent depending on the time of activity of the source repeat. The vast increase in fully sequenced vertebrate genomes available represents a new challenge for detection of transposable elements as only incomplete and inaccurate repeat libraries exist for newly sequenced genomes. Tools like *Repeatmasker* are therefore only of limited value in detecting new families of transposable elements in newly sequenced species. To overcome this problem a number of *de novo* methods for finding new repeat families were developed. These programs normally start with a self-comparison of the DNA query sequence to detect sequence similarities indicating the presence of a repeated sequence. In the next step, the identified repeated sequences are clustered to group related sequences into families. This method is applied by programs such as *REPuter* [23] , *RECON* [24], RAP [25] or PILER [26]. Although these programs work well to identify new highly redundant repeat families like short interspersed nuclear elements (SINEs) they might fail to detect more complex transposable elements at lower copy number as many LTR-retrotransposons and endogenous retroviruses. These elements have some additional

features that can be used for *de novo* detection. Most importantly, the long terminal repeats (LTRs) of LTR-retrotransposons flanking the coding region of the element are identical repeat of about 350 to 1000bp in newly integrated elements. The presence of two repeated elements of a certain size within a certain distance (the length of the coding



of LTR-retrotransposon features recognised by *de*
ninal repeat. TSD: Target site dublication.

region of a LTR-retrotransposon) is therefore a strong indicator for an integrated LTR-retrotransposon or endogenous retrovirus. Additionally LTRs are flanked by a short target site duplication (TSD) which is the result of the integration process, where a short 4 to 6 nucleotide sequence is duplicated (Figure 2). The internal region contains genes that code for the proteins necessary for retrotransposition or virus particle formation and often have conserved domains that can be recognised. Programs specifically designed for *de novo* detection of LTR-retrotransposons and endogenous retroviruses that use these features in post processing steps to detect LTR-retrotransposons include *LTR-STRUC* [27], *LTR_par* [28] and LTR_FINDER [29].

A recently developed program, *LTRharvest* [30] implements already existing models of LTR recognition but also introduces some new algorithms and features. The major advantages of this program are the high speed of the computation that allows the screen of a mammalian chromosome within a few minutes on a standard computer and a high flexibility in the parameter settings. The latter enables the user to optimise the settings for a more precise recognition of a certain group of LTR-retrotransposons.

LTRdigest [31] was developed as an extension to *LTRharvest* that allows a further automated analysis of LTR elements identified by *LTRharvest*. The program uses three additional features for LTR-retrotransposon recognition: First, the elements are screened for Primer binding sites (PBS), that are essential for reverse transcription of LTR-retrotransposons and endogenous retroviruses. For this a region directly downstream of

the 5´LTR is analysed for homologies to the 3´end of a library of tRNAs. The length of the region that is scanned for a PBS as well as the region of the tRNA that is considered are variable and can be set by the user. Second, candidate elements are screened for a Polypurine tract (PPT) which is normally located near the 3´LTR. This short sequence is characterised by a purine-rich base composition and identified by a hidden Markov model (HMM) based algorithms [31]. Last, protein domains within the two LTRs are detected by profile hidden Markov models (pHMMs) [31, 32] which are publicly available for a large number of protein families [33]. A collection of protein domain models of interest in HMMER format can be created by the user. *LTRdigest* translates the internal region of candidate elements in all six possible reading frames and screens the amino acid sequence for homologies to the protein domains in the local HMMER formatted library.

For each analysed sequence several output files are generated, including an overview for all hits in a tabular format. All information about position and length of the candidate sequences, position and motif of PBS and PPT and protein domain hits are found in this files. Files in multiple FASTA format include the full length sequences of all candidates and, in separated files, the 5´and 3´LTR sequences. Elements inserted on the minus-strand are automatically reverse complemented if identification of LTR-retrotransposon features like PBS, PPT and protein domains allowed strand determination of the candidate element by *LTRdigest*. In this study we will employ *LTRharvest* and *LTRdigest* to screen the nearly fully sequenced pig genome for endogenous retroviruses. Known and possibly new ERV genomes will be grouped in families and further characterised by local BLAST and BLASTX search.

2. Results

2.1 *LTRharvest/LTRdigest* prediction in the pig genome

All sequenced chromosomes of *Sus scrofa* (Nov.2009 SGSC Sscrofa9.2/susScr2) were downloaded from the UCSC Genome Browser homepage and analysed by *LTRharvest* which was installed as part of *GenomeTools*, version 1.3.4 (http://genometools.org). First, enhanced suffix arrays for each chromosome were created by the program *Suffixerator* according to the *LTRharvest* manual [30]. This step is required to create all the files necessary to run *LTRharvest*. To enable a more accurate search for endogenous

retroviruses and LTR-retrotransposons, some of the parameters were changed as follows: The minimum length of each LTR was set from 100bp (default settings) to 350bp. By adjusting this value we hoped to decrease wrong hits caused by porcine SINE elements (PRE-1). These elements were estimated to populate the pig genome at a frequency of $1x10^6$ [34, 35] and could therefore be misleadingly identified as LTRs whenever two elements are located within a certain distance. The average size of PRE-1 elements does not exceed 230bp [34, 35] and most retroviral LTRs are at least 400bp long so we chose 350bp as minimum LTR length. Also the minimum distance of the two LTR starting positions were changed from 1000bp (default) to 4000bp in order to exclude shorter sequences that are very unlikely to represent LTR-retrotransposons or endogenous retroviruses which are normally between 7000bp and 12000bp long. The remaining settings were left as default [30]. With these parameters all pig chromosomes were analysed by LTRharvest.

In the next step we screened the candidate elements, extracted by *LTRharvest* with *LTRdigest* for the presence of PBS and PPT. Since no complete tRNA library for *Sus scrofa* is available we downloaded all known tRNAs of Bos taurus from the *Genomic tRNA database* (http://lowelab.ucsc.edu/GtRNAdb) in a multiple FASTA format that can be read by *LTRdigest*. Most PBS of endogenous retroviruses in vertebrates start with a 5´-TGG base triplet which is complementary to the CCA-3´end of most mature tRNAs. These three bases at the 3´end of tRNAs is normally not coded by the tRNA gene and added in a post-transcriptional step. The standard setting for the minimum and maximum PBS offset (distance from 5´LTR end to PBS start) is 0 and 5, respectively. Since the first three bases of the PBS will in most cases not match a tRNA 3´end of the library which is composed of the genomic tRNA sequences and not of the mature tRNAs, only PBS that are located not further than 2bp downstream of the 5´LTR would be detected by *LTRdigest*. We therefore increased the maximum PBS offset to 10. The tabular result files for each chromosome were analysed and depicted in Tab. 1.

Chr.	length	hits	hits with PPT	hits w. PBS	hits w. PPT and PBS	total length of hits	% of chromosome
1	295529705	315	139	80	28	3119774	1,06
2	140133492	223	89	61	21	2259563	1,61
3	123599780	98	34	22	5	927963	0,75
4	136254946	117	38	26	7	1104158	0,81
5	100516970	86	34	18	5	835747	0,83
6	123305171	112	39	30	3	1138789	0,92
7	136409062	126	51	32	11	1252180	0,92
8	119985671	92	44	25	10	885175	0,74
9	132468591	141	60	40	18	1382240	1,04
10	66736929	68	28	20	6	677645	1,02
11	79814395	70	34	15	6	674826	0,85
12	57431344	83	23	24	0	769899	1,34
13	145235301	122	40	27	5	1230431	0,85
14	148510138	120	42	32	11	1158159	0,78
15	134541103	114	50	24	5	1163066	0,86
16	77435658	60	29	10	7	608369	0,79
17	64395339	53	29	8	1	564708	0,88
18	54309914	23	13	4	1	224526	0,41
X	125871292	335	157	85	45	3371039	2,68
total	2262484801	2358	973	583	195	23348257	1,03

Tab. 1: Statistical analysis of *LTRharvest/LTRdigest* output. Predicted LTR-retrotransposons for each chromosome, predicted number of elements with PBS and PPT as well as the total length of the predicted elements are depicted. No hits were found on chromosome M.

A total of 2358 candidate elements were detected by LTRharvest, taken together the length of these elements accounts for about 1% of the analysed genome. Of these elements, 973 were screened positively for a PPT, 583 for a PBS. 195 candidates had both a PPT and a PBS that was recognised by *LTRdigest*. Primer binding sites were highly diverse in sequence but were matching preferably to a tRNAs that transfer a glycine amino acid (Figure 3). Most of the known PERVs use a PBS complementary to a glycine tRNA.

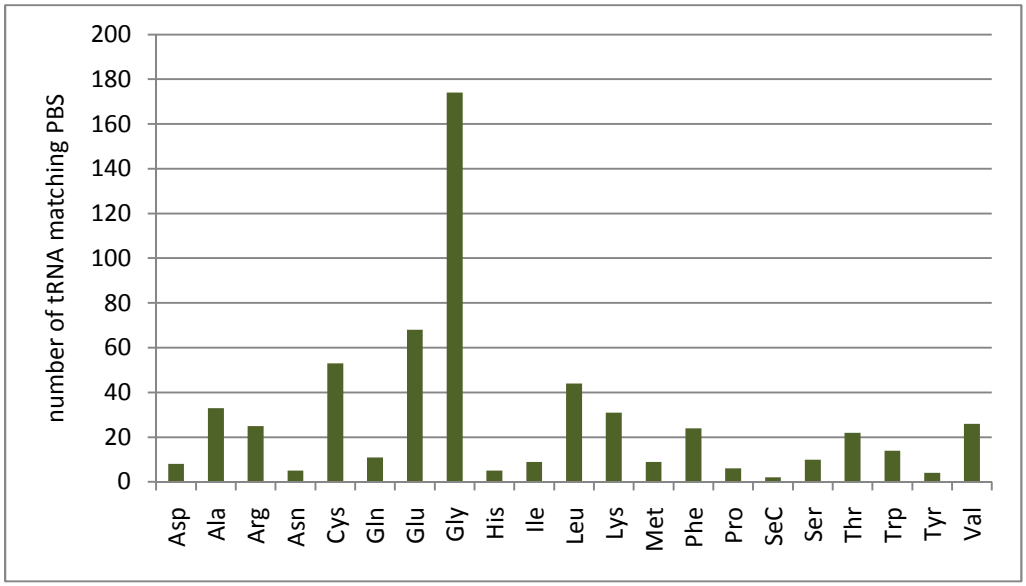


Figure 3: Matching transfer RNAs (tRNAs) to 583 identified PBS sequences

2.2 Local BLAST search of candidate elements

To identify known PERV groups among the detected candidates, local BLAST search of a local database with published partial PERV sequences [18, 20] was performed. In parallel, a genome-wide BLAST search (*UCSC Genome Browser*) of the published PERV sequences was done (Tab. 2). The partial PERV-γ3 sequence is highly homologue to PERV-γ5 and matched to the same regions of the pig genome. Therefore we did not consider them as two separated groups. The same applies to PERV-β1 and PERV-β2.

PERV	Accession number	UCSC BLAST	candidate elements
γ1	AF274705	25	19
γ 2	AF274706	63	35
γ 3/5	AF274707 / AF274709	10	1
γ 4	AF274708	1	1
γ 7	AF511111	1	1
γ 8	AF511112	0	0
γ 9	AF511113	0	0
γ 10	AF511114	12	1
β1/2	AF274710 / AF274711	16	3
β3	AF274712	21	5
β4	AF274713	1	0
β5	AF511115	1	0
total		151	66

Tab. 2: Hit numbers of known PERV sequences in candidate elements and in the pig genome. Hits were counted when at least 50% of the query sequence was at least 85% identical to the target sequence.

Hit numbers of most PERV families was considerably higher in the genome-wide search, indicating that many retroviral sequences were not recognised by *LTRharvest*. These elements may have partially deleted LTRs or no intact target site duplications.

2.3 Screening of candidate elements for retroviral domain families

Our installation of *LTRdigest* did not support the detection of protein domain families by pHMMs, therefore all 2358 candidate elements collected by LTRharvest were screened by BLASTX search (translated DNA sequences in all six reading frames) against a local database of domain families that were downloaded as FASTA files from the Pfam homepage. 21 different protein domains families that are associated with proteins of LTR-retrotransposons and retroviruses were included in the local library (**Error! Reference source not found.**). The maximum e-value for the BLASTX search was set to 0,01 to exclude random hits.

Pfam accession	Pfam ID	Pfam description	hit number
PF01141	Gag_p12	Gag polyprotein inner coat protein12	228
PF01140	Gag_Ma	Matrix protein (MA), p15	196
PF00429	TLV-coat	ENV polyprotein (coat)	168
PF00077	RVP	Retroviral aspartyl protease	115
PF02093	Gag_p30	Gag P30 core shell protein	115
PF04160	Borrelia_orf X	Orf-X protein	97
PF00098	zf-CCHC	Zinc knuckle domain	96
PF00665	rve	Integrase core domain	88
PF00075	Rnase_H	Ribonuclease H domain	75
PF00607	Gag-p24	Gag gene protein 24	42
PF02022	Integrase_Zn	Integrase Zinc binding domain	34
PF00517	GP41	Envelope polyprotein GP41	32
PF02337	Gag_p10	Retroviral GAG p10 protein	31
PF00552	Integrase	Integrase DNA binding domain	29
PF00692	dUTPase	dUTPase domain	25
PF03708	Avian gp85	Avian retrovirus env. protein, gp85	14
PF06817	RVT-thumb	Reverse transcriptase thumb domain	9
PF00516	GP120	Envelope glycoprotein gp120	8
PF02813	RetroM	Retroviral M domain	8
PF03408	Foamy_virus-ENV	Foamy virus envelope protein	5
PF09590	Env-gp36	Env-gp36 protein (HERV/MMTV type)	3

Tab. 3: Pfam domain families used for local BLASTX search of candidate elements. Hit number is the number of elements within the candidates that had homologies to at least one of the seed sequences of the family

588 elements out of 2358 matched to at least one retroviral domain family. We were interested in whether the presence of a PBS or PPT detected by *LTRdigest* correlated with a higher number of pfam hits. For this we divided our candidate elements into different groups according to how many pfam domains the elements matched and counted PBS and PPT hits for these groups.

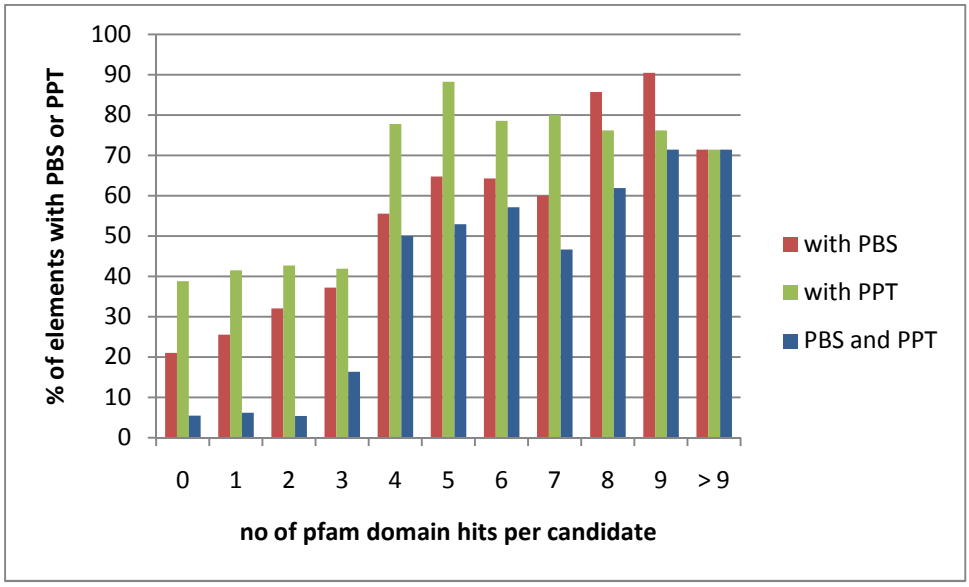


Figure 4: Presence of PBS and PPT in candidate elements, grouped according to the number of pfam domain matches

Candidates that matched to pfam domains were more likely to also contain PBS or PPT sequences but the correlation was less than expected. However, the presence of both PBS and PPT seemed to correlate with retroviral domain quite well. Only about 5% of the elements that did not match to any of the retroviral protein domains were positive for both a PBS and a PPT, but more than 50% of the elements with more than 3 retroviral domain matches. To identify candidates that may represent formerly uncharacterised retroviral genomes, we cross-checked elements for pfam domain hits, matches to known PERV sequences and PBS and PPT (Tab. 4).

no of pfam hits per element	all candidates	matches to known PERVs	with PBS and PPT
0	1770	0	96
1	357	0	22
2	75	0	4
3	43	0	7
4	18	0	9
5	17	3	9
6	14	5	8
7	15	12	7
8	21	19	13
9	21	20	15
10	2	2	1
11	1	1	1
12	4	4	3
total	2358	66	195

Tab. 4: Cross-check of candidates for matches to pfam domains, known PERVs and PBS and PPT

All candidates that were identified as known PERVs matched to at least 5 retroviral protein domains but also several element were identified that matched to several pfam domains but had no homology to any known PERV sequence. One possibility is that some elements may have large deletion in the region complementary to the published partial PERV sequences. Therefore we screened all candidates by BLAST search against available full length PERV-γ1 and –γ2 sequences. For the remaining PERV groups, no full length sequences were published before, so we screened the candidates for homologies to our own full length sequences that showed homologies to these PERV groups. Furthermore we tried to cluster elements that did not match to any known PERV to identify new groups.

2.4 Family classification of putative LTR retrotransposons

In addition to the 66 elements that were classified by local BLAST search against published PERV sequences we identified several candidates that belong to known PERV groups but had deletions in the critical regions. Furthermore we found two new PERV groups that showed no homology to any known PERV sequence, designated as cluster1 and cluster2 (Figure 5). All together we classified 97 porcine endogenous retroviruses,

including 21 elements that were clustered in two groups that have not been described before.

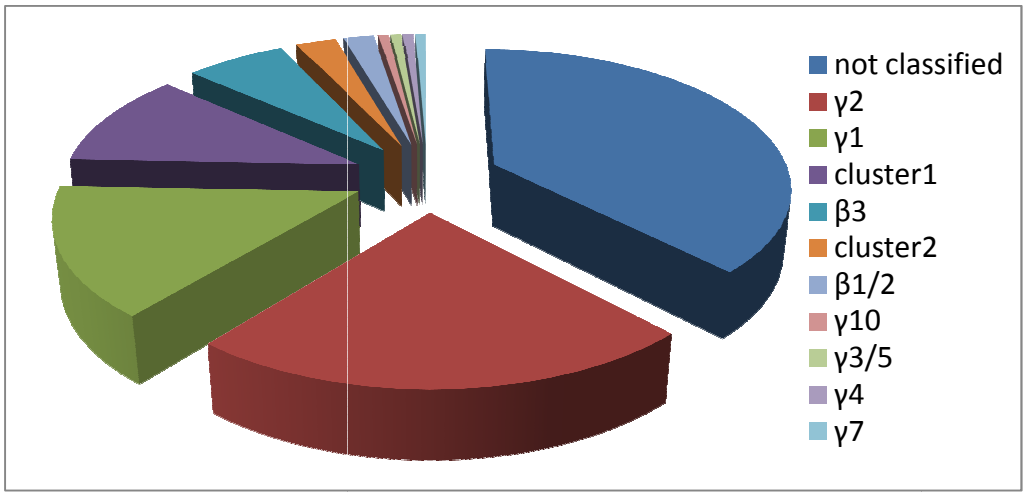


Figure 5: Classification of 156 candidate elements that matched to three or more pfam domains

In the following section all identified elements belonging to known PERV groups and the two newly described PERV clusters are further characterised. Elements of the same group were aligned and regions coding for retroviral domains were annotated. Local BLASTX search was performed and the maximal regions that matched to pfam domains (Tab. 3) were marked, regardless of stop codons or frame shifts in the coding regions. For this analysis we added one more protein domain to our Pfam library, RVT_1 (Pfam accession: PF00078). This domain is found in retroviral reverse transcriptase (RT) protein but also in some Long interspersed nuclear element (LINE) RTs. Candidates were also screened for intact open reading frames (>300bp) and for homologies to known transposable element in the pig genome. For this, a local database containing all known transposable elements that were found in pigs was downloaded from the *Repbase Update* homepage [22]. Local BLAST and BLASTX searches as well as alignments and graphical depictions of annotated sequences were performed using tools and programs that are part of *CLC bio* (Main Workbench 5.7.1).

2.5 Characterisation of identified LTR retrotransposons families

2.5.1 PERV-γ1 group

As mentioned above, the PERV-γ1 group is very well described because of their potential to infect human cells. All together we could identify 23 PERV-γ1 elements among our candidates, a closer analysis by screening these elements for published *envA*, *envB* and *envC* sequences revealed that 13 elements belong to the PERV-A class, 8 to the PERV-B class and one element to the PERV-C class. One element had a large deletion in the entire *env* coding region and thus could not be classified by this approach. Besides the *env* gene, PERV-A,-B and –C also differ in the PBS they use for reverse transcription. Whereas all PERV-A and –B sequences examined so far use a PBS complementary to a Glycine tRNA, PERV-C have a Proline-PBS [8, 36]. Since an intact PBS, complementary to a Proline tRNA was found in the element with a missing *env* gene, we conclude that ERV belongs to the PERV-C class. Local BLAST search of a representative PERV-γ1 against our retroviral domain family database (Pfam) identified 4 *gag* domains (Gag-Ma, Gag_p12, Gag_p30 and Gag_p24), followed by a Zinc knuckle domain (zf-CCHC) that mediates RNA binding of the gag-protein to the retroviral RNA (Figure 6, top). Further downstream we found a Protease domain (RVP) which is part of the expressed retroviral Protease protein that catalyses processing of the retroviral proteins out of the precursor polyprotein. The Protease itself is expressed as part of the gag-pol polyprotein and processed by Protease enzymes that were incorporated in the viral particle upon viral assembly. Domains of a reverse transcriptase (RVT-1) and integrase (rve) that were found are both part of the retroviral reverse transcriptase that mediates reverse transcription and integration into the genome. Finally, a large region was matched to a domain that is part of the TLV-coat family and is located at the 3´end of the sequence. This domain represents the env gene which allows a retrovirus to infect new cells and is the major difference between LTR-retrotransposons and endogenous retroviruses.

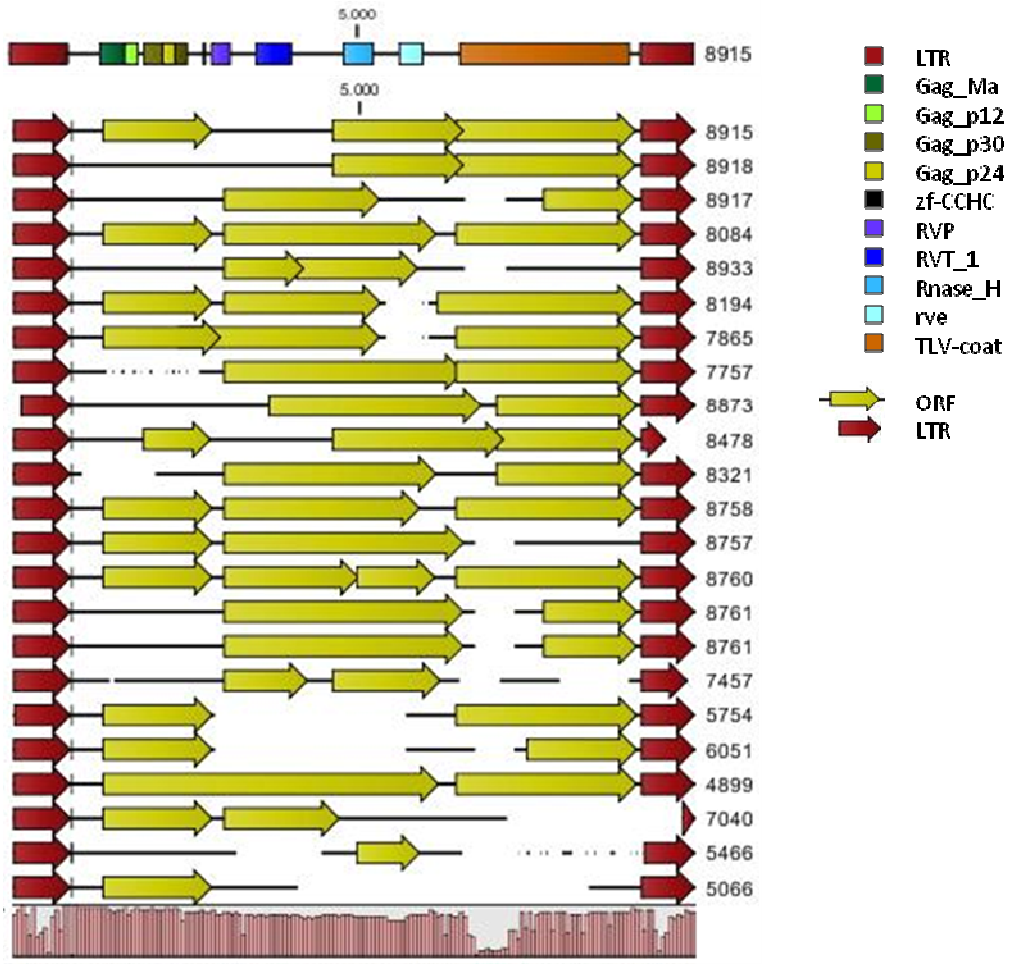


Figure 6: Alignment of 23 PERV-γ1 endogenous retroviruses. Top: Annotation of DNA regions, coding for retroviral proteins for a representative sequence of the PERV group. Bottom: Conservation score of the alignment.

2.5.2 PERV-γ2 group

Shortly after the sequences of the three PERV-γ1 classes PERV-A, PERV-B and PERV-C were found in the pig genome, two groups independently reported the existence of a new type of PERV, designated as PERV-E [16] and PERV-γ2 [18]. The *gag*, *pol* and *env* genes of PERV-γ2 are closely related to those of human endogenous retrovirus (HERV) but these porcine and human endogenous retroviruses have no common endogenous ancestor. Although PERV-γ2 endogenous retroviruses were found in all *Suidea* (Old World pigs) species examined, they were absent in the genome of two genera of peccary (*Tayassuidae*), which are believed to have separated from the *Suidea* lineage 20 million years ago [18]. These findings indicate that the PERV-γ2 group has entered the pig lineage after the separation of *Suidea* and *Tayassuidae* but before speciation within the *Suidea* lineage. No intact full length open reading frames were found in the PERV-γ2 sequences

examined, but a copy number estimation by diluting nested PCRs revealed that domesticate pig breeds harbor more PERV-γ2 integrations then wild boars. This supports the hypothesis that some members of this PERV group were still active during the domestication process of pigs by humans which is estimated to have occurred less than 5000 years ago [16]. 36 elements of the 2358 candidates were identified as PERV-γ2 and further characterised. Full length elements were about 8900bp in length and had LTRs of about 430bp. Local BLASTX search against our local database of retroviral domain families (Pfam) identified the regions coding for *gag*, *pol* and *env* proteins (Figure 7).

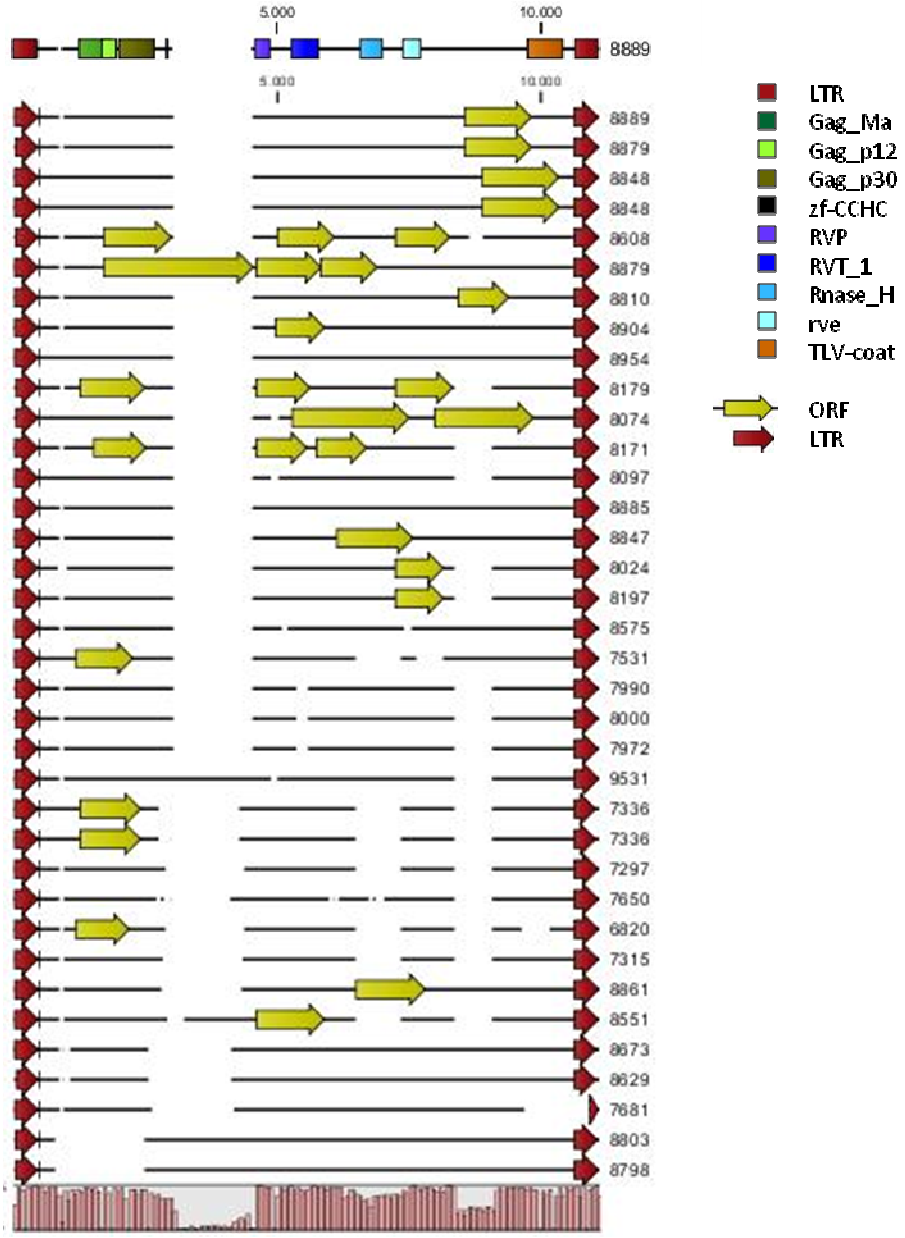


Figure 7: Alignment of 36 PERV-γ2 endogenous retroviruses. Top: Annotation of DNA regions, coding for retroviral proteins for a representative sequence of the PERV group. Bottom: Conservation score of the alignment.

All 36 elements were screened for ORFs (longer than 300bp) and aligned. Several truncated ORFs were identified but none that codes for a full length retroviral protein. The largest part of the aligned sequences was highly conserved, with some exceptions. Some elements had a large insertion (about 1400bp) downstream of the region coding for *gag* domains. The inserted sequence was conserved between these elements, indicating that they were still able to replicate, possibly by *in trans* complementation with the disrupted factors. Additionally, 16 elements sequences were missing an about 600bp long sequence, possibly within the coding region for the *env* gene. We cannot say whether this gap should be regarded as a deletion or an insertion since about half of the elements are missing this sequence and none of our sequences had intact open reading frames. One possibility would be that some elements lost their ability to express an intact *env* protein either by deletion or insertion of these 600bp but still replicated after endogenisation in an intracellular replication manner (like LTR-retrotransposons).

Figure 8: Sequence Logo of the Primer binding site region of all 36 PERV-γ2 sequences.

All sequences had a conserved primer binding site, complementary to a Glycine tRNA, similar to PERV-γ1. The 7th nucleotide of the PBS showed a dimorphism, about half of the elements had an Adenine, the other half a Guanine base at this position (Figure 8). However both PBS versions are matching to different Glycine-tRNAs and are therefore likely to allow priming of the reverse transcription reaction.

2.5.3 PERV-γ3 to -γ10 groups

Apart from PERV-γ1 and PERV-γ2, more gammaretroviral groups were reported in the pig genome [18]. However, we could only identify a few of these elements among our candidates. As discussed above, our settings of LTRharvest obviously did not work well for the detection of these elements (Tab. 2). One element identified matched to PERV-γ3/5. A local BLAST search of these element against a local database with all known Transposable elements (TEs), downloaded from the *Repbase Update* homepage revealed

that the LTRs (identified by *LTRharvest*) of PERV-γ3/5 are highly similar to a porcine LINE. We conclude that these repetitive sequence was recognised as LTRs by mistake. Nevertheless, this element had homologies to the published PERV-γ3 and -γ5 sequences and we also identified *gag*, *pol* and env domains within the sequence (Figure 9).

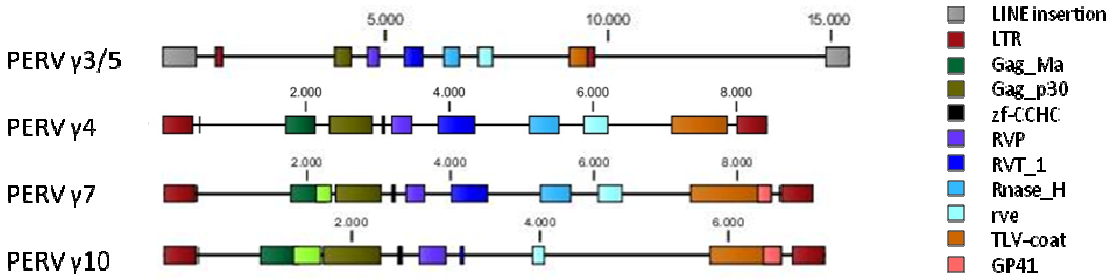


Figure 9: PERV-γ3/5 to PERV-γ10 elements with annotated retroviral domains

We therefore manually searched for alternative LTRs and found a repetitive region of 190bp that may represent the actual, truncated LTRs of this endogenous retrovirus. A primer binding site near the possible 5´LTR could not be identified. The sequence, identified as PERV-γ4 contained *gag*, *pol* and *env* domains and a PBS complementary to a Threonine tRNA. As for PERV-γ3/5 no intact ORFs were found. Each one PERV-γ7 and one PERV-γ8 was identified. Both matched to *gag*, *pol* and *env* domains and had one short open reading frames in the *pol* and *env* region, respectively. Notably, the LTRs of the PERV-γ10 element were only 352bp long. If that is the normal LTR length for this PERV group we may have missed most of the integrated copies because we set the minimum LTR length to low (350bp). PERV-γ7 had a PBS complementary to a Threonine tRNA but no PBS could be identified in PERV-γ10.

2.5.4 PERV-β1/2 group

Although the copy number of the PERV-β1 and PERV-β2 groups were estimated to be 8 to 16 and 4 to 8, respectively [18], and we found 16 matching regions in the pig genome (Tab. 2) only three of our candidate elements showed a significant homology to the published partial PERV–β1/2 *pol* sequences. Interestingly two of the elements were almost identical (99,7%) which might indicate a very recent activity of these elements. But

since we did not find any intact ORF in these elements and the distance between the two integrations was relatively short (220kb), we assume that the high similarity of these elements might be the result of a duplication of a chromosomal region that included one of the elements.

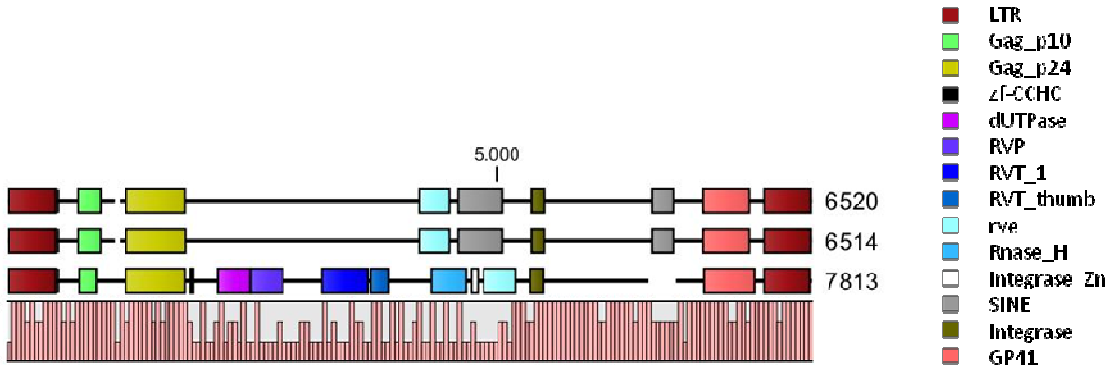


Figure 10: PERV-β1/2 elements with annotated retroviral domains

Another indication that these two highly similar elements lost their ability to replicate long time ago is that we found two SINE insertion in the coding region and in contrast to the third PERV-β1/2 elements only one retroviral *pol* domain (rve) was found (Figure 10).

2.5.5 PERV-β3 group

Initially, we could only identify four PERV-β3 elements among all candidate elements. Six more were found by BLAST search against the identified full length PERV-β3 sequences. These elements were missing a large part of the *pol* coding region and were thus not detected by BLAST search against the published partial *pol* sequence (Figure 11).

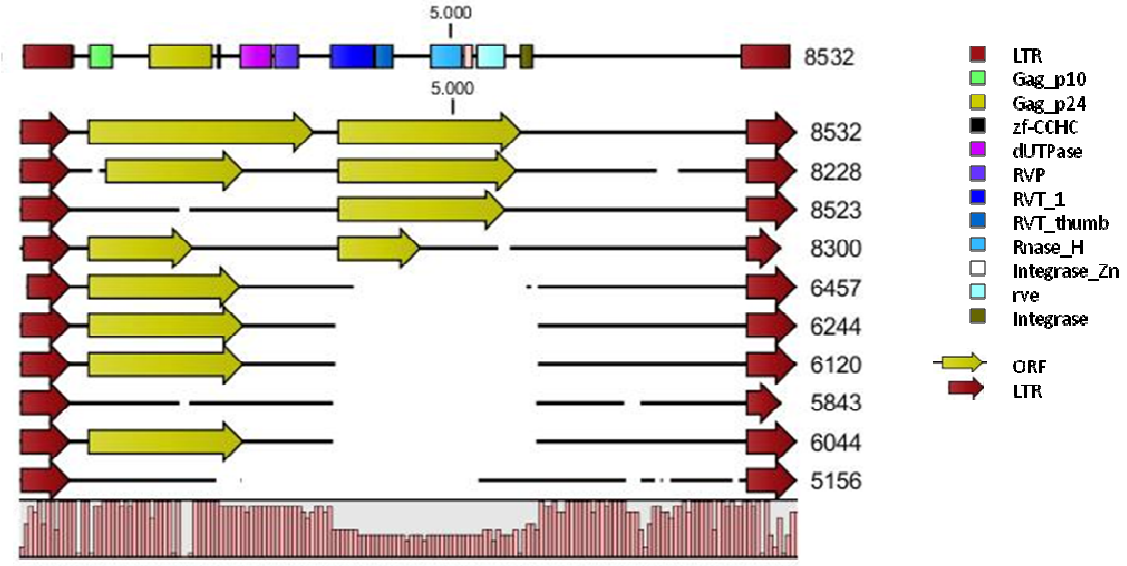


Figure 11: PERV-β3 elements. Top: Annotation of DNA regions, coding for retroviral proteins for a representative sequence of the PERV group. Bottom: Conservation score of the alignment.

Interestingly, five elements have almost identical deletions which may indicate that these sequences are the result of *in trans* complementation of a internally deleted PERV-β3 element with intact signal sequences like a PBS. Indeed, the PBS of these elements was highly conserved and matched to a Lysine tRNA (Figure 12). Strikingly, the remaining part of these elements was also highly conserved.

Figure 12: Sequence logo of PERV-β3 PBS

Large intact open reading frames were found in these elements, indicating a relatively recent integration time. Although retroviral *gag* and *pol* domains were present, no *env* domains could be found. Possibly these PERVs had an *env* gene that has no homology to the ones in our retroviral protein domain library or this gene was lost after integration and the elements replicated intracellularly in a LTR-retrotransposon like manner.

2.5.6 New PERV cluster 1

17 elements, that showed no homologies to published PERV sequences were clustered into a new PERV group. There is strong evidence that these sequences are of retroviral origin. Domains of retroviral *gag*, *pol* and *env* proteins were found in the right order (Figure 13) and the elements share a common motif for a PBS that is complementary to a Threonine tRNA. The PBS was highly conserved among all elements, only the 13th nucleotide showed some dimorphism (Figure 14).

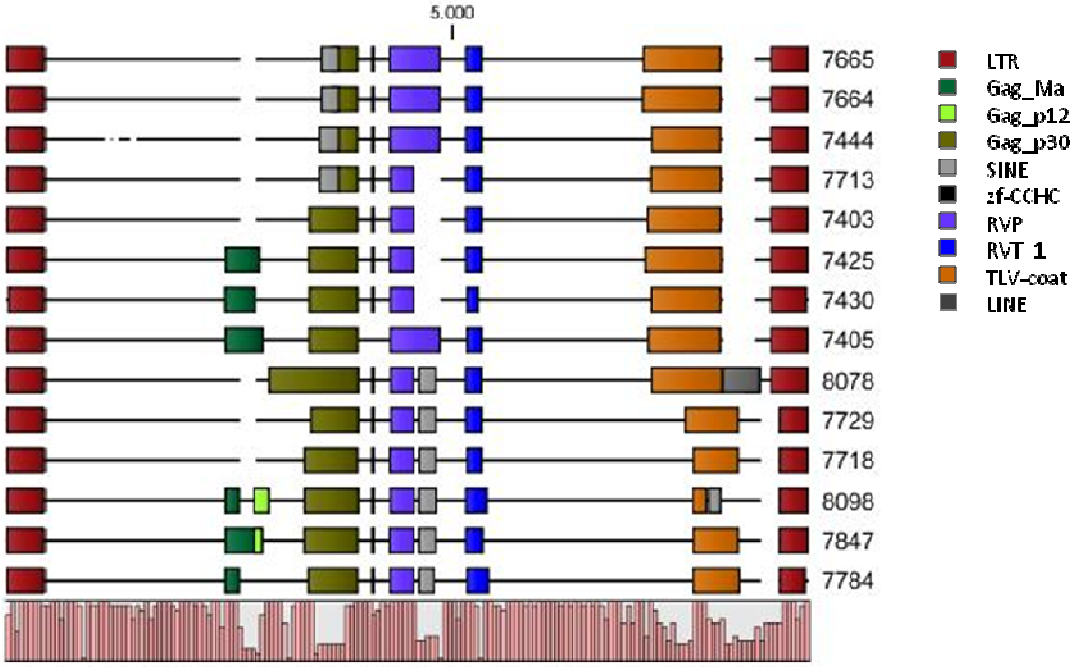


Figure 13: Unknown cluster of 14 retroviral sequences. Retroviral domains of all sequences were annotated. Only 14 of the 17 identified members of this group are shown in the alignment.

This might by a natural variation rather than a mutation since both PBS variations match perfectly to different Threonine tRNAs. We conducted BLAST searches of these elements with against the elements we could identify as known PERVs but no homologies were found, confirming again that this elements represent a new PERV group. According to the similarity of the domain organisation to gammaretroviruses we assume that this group is also of the γ-type.

Figure 14: Sequence logo of the PBS of a new PERV group (Cluster1)

No open reading frames were found in this group but we detected several SINE insertions. Since we found these insertions at the same position in different elements we assume that some elements were still replicating after the SINEs integrated into the retroviral genome even though this obviously resulted in defective ORFs or non-functional proteins (Figure 13).

2.5.7 New PERV cluster2

A second cluster of formerly unknown PERVs was identified. The four elements that form this cluster showed some homologies at the 3´end of the coding region to cluster1 PERVs but were clearly distinct in their LTR and *gag-pol* coding sequences. All members of this family matched to *gag* and *pol* domain, three of them additionally to an *env* domain (Figure 15). Several SINE insertions were found and interestingly two elements showed a very similar SINE integration pattern. These elements were almost identical in their sequence and only 100kb separated from each other, therefore we suggest that a duplication of a chromosomal region including the retroviral insertion occurred. Three of these elements had a PBS directly adjacent to the 5´LTR that matched to a Proline tRNA.

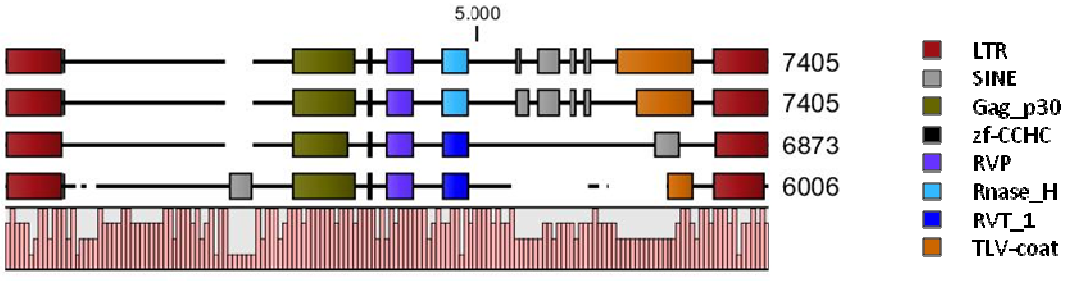


Figure 15: Unknown cluster of 4 retroviral sequences. Retroviral domains of all sequences were annotated.

2.6 Age estimation by LTR divergence and molecular clock hypothesis

As mentioned above, the two LTRs flanking the retroviral genome are identical after chromosomal integration but will then accumulate mutations over time. This gives us the opportunity to estimate the integration time by comparing the 5´- and 3´LTRs and counting the mutations in these sequences. Analysis of primate ERVs has shown that these elements accumulate mutations with a rate of approximately 2,3 x 10^{-9} to 5,0 x 10^{-9} substitutions per nucleotide per year [37]. This mutation rate was also used in a study to estimate the age of PERV-γ1 [38]. For our calculations we assumed an average mutation rate of 3,65 x 10^{-9}. The LTRs of all classified elements were aligned and nucleotide substitutions were counted. Insertions and deletions were not considered as mutations since we cannot estimate the frequency of these events. We also considered the different LTR lengths in our calculation.

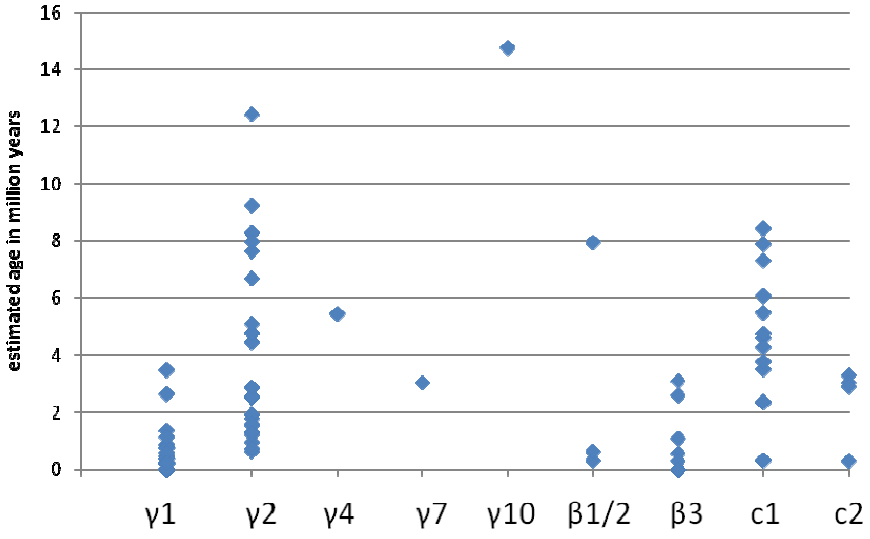


Figure 16: Age estimation of classified PERVs by the molecular clock hypothesis. c1,c2: Newly described Cluster 1 and 2

According to our estimation all PERV-γ1 elements integrated into the pig genome less than 4 million years ago and only three elements more than 2 million years ago (all three were of the PERV-A type) (Figure 16). This finding fits with the results of another study, where the age of the oldest PERV-A integration was estimated to have occurred 3,5 to 7,5 million years ago [15]. Most of the elements investigated had perfect or nearly perfect matching repeats, indicating a very recent activity which also concurs with the observation that domesticated pigs (like *DUROC*) carry more PERV-γ1 copies in the genome than wild boars [16]. PERV-γ2 elements showed a higher variability in their LTR differences. Some elements of this family

seem to have integrated more than 8 million years ago whereas other PERV-γ2 members had almost identical LTRs which points towards a much more recent integration time. Information about the age of PERV-γ2 in the literature are rare. Patience *et al.* [18] did not find PERV-γ integrations in the genome of New World peccaries which have separated from the Old World pigs about 20 million years ago but in all Old World pig species examined. However, it seems that some members of this family has been still active within the last few thousand years given that domesticated pigs obviously harbor more PERV-γ2 copies than wild boars [16]. It is possible that some members of this family remained replication competent over millions of years or alternatively that they were reactivates at several time points through history by *in trans* complementing elements. It is difficult to tell anything about the age of PERV-γ4, –γ7, -γ10, and -β1/2 groups since we could only identify a few members of this family. Interestingly, all PERV-β3 sequences appeared to be relatively young, with 5 elements that had perfect matching LTRs. This also fits with the observation that these elements had long intact open reading frames. The PERV families that have not been described before (Cluster1, and Cluster2) had both one member with nearly intact LTRs but also several members that appear to be older. Since we could not find any open reading frames in these ERVs we assume it is unlikely that they have been active in recent history. Only a very rough estimation is possible with the method we employed here. Mutation rates are only estimation themselves and will in reality depend strongly on factors like selective pressure and integration site.

3. Discussion

Endogenous retroviruses play an important role in genome evolution and have marked effects on cell function and metabolism. A recent paper proposes a role of an endogenous long terminal repeats in the development of human lymphoma [39]. Apart from the risk of reactivates human endogenous retroviruses in the onset of disease, replication competent porcine endogenous retroviruses (PERV) raised concerns in the widespread use of pig organs for Xenotransplantation [9-12]. In this study we screen the nearly completely sequenced pig genome for retroviral sequences using recently developed computational tools [30, 31]. We identified 23 members of the well described PERV-γ1 group and found several intact open reading frames (ORFs) in several elements (Figure 6). In contrast, only truncated ORFs were found in retroviral genomes that belong to the PERV-γ2 group (Figure 7). As for the

remaining PERV-γ groups that have been described before, no intact ORFs but several retroviral domains and intact primer binding sites (PBS) could be identified. A discussed above, our settings for the LTR-retrotransposon prediction by *LTRharvest* were not optimal for detecting elements of these groups. Possibly, some of these retroviral sequences have only very short or deleted LTRs that did not fit to our applied settings. For known endogenous retroviruses it may therefore be more efficient to use BLAST search of published sequences against the full genome, followed by manually extracting these elements. However, the aim of *LTRharvest* is to identify new retroviral genomes of which no published sequences are available. Since mammalian genome are full of repeated elements that are not of retroviral origin, it is important to screen the detected candidate elements for additional retroviral features to exclude wrong hits and delimit possible candidates. The presence of retroviral domains in a candidate is probably the strongest indicator for an endogenous retrovirus and can be normally automatically detected by *LTRdigest*. Since our installation did not support this function, we screened all candidates for retroviral domains by local BLAST search. After this step we could easily identify the elements with the highest probability of being a LTR-retrotransposon or endogenous retrovirus. We further excluded known PERV elements by local BLAST search against published PERV sequences from our selection and received a relatively small number of candidate elements that are likely to represent new PERV genomes. By local BLAST search of these elements against each other we identified two cluster of PERVs that showed no homology to known PERV groups but contained several retroviral features like PBS and *gag*, *pol* and *env* domains. A considerable number of elements that are likely to be of retroviral origin remained uncharacterised (Figure 5). Additionally we cannot exclude that some of the candidates that did not match to any of the retroviral domains in our pfam library are elements coding for retroviral proteins that lack any known retroviral domain and represent completely new groups of endogenous retroviruses. We should also mention that many retroviral elements will not be detected by *LTRharvest* when they lack intact LTR regions or target site duplications (Tab. 2). In general, the stricter the settings for *LTRharvest* are chosen, the more retroviral elements will not be detected. On the other hand, less stringent settings will lead to more background noise by mistakenly identified candidates. Setting should therefore be optimized and adapted, depending on which genome is analysed and which elements should be identified. Our approach to estimate the integration time of the analyses elements is somewhat

problematic since it presumes that retroviral genomes accumulate mutations at a more or less constant rate. In reality, this rate might be biased by several factors like evolutionary selection or the influence of flanking sequences and the surrounding chromatin structure. Despite these limitations our estimated age of the oldest PERV-γ1 integrations fit well with the results from another study, in which species closely related to pigs were screened for PERV-γ1 insertions. The PERVγ-2 elements we identified seem to bee of an older origin than PERV-γ1 since we found several members that dated back later than the oldest PERV-γ1 insertion (Figure 16). Some of these elements might have retained their ability to replicate over a few million years. Alternatively, the younger elements could represent *de novo* germline insertions of a circulating exogenous PERV-γ2 at a later time point in history. But since all elements showed high similarity to each other and it seems to be unlikely that a circulating exogenous retrovirus did not change for millions of years we believe all these elements had one or a few common ancestors that inserted into the germline within a relatively short period. The age distribution of PERV-β3 elements closely resembles the one of PERV-γ1, indicating that this group has entered the pig lineage about the same time but in contrast to PERV-γ1 already lost their ability to replicate. Endogenous retroviruses of the newly described Cluster2 might have entered the lineage about the same time as PERV-γ1 and -β3 but we would need to analyse more elements to strengthen this hypothesis. In contrast, elements of the PERV cluster1 might have entered the genome earlier, approximately at the same time as PERV-γ2. However, all these assumptions are highly speculative and would need to be reassessed by PCR screening of closely related species for the presence of these PERV groups.

In conclusion, we here characterise a number of already known but not well described PERV genomes and identify two new groups of porcine endogenous retroviruses that were predicted by computational programs. Additionally, many promising candidates that were predicted but not further classified proof this approach as a powerful method for the *de novo* LTR-retrotransposon prediction.

Literature

1. Goodier, J.L. and H.H. Kazazian, Jr., *Retrotransposons revisited: the restraint and rehabilitation of parasites.* Cell, 2008. **135**(1): p. 23-35.
2. Kazazian, H.H., Jr., *Mobile elements: drivers of genome evolution.* Science, 2004. **303**(5664): p. 1626-32.
3. Khodosevich, K., Y. Lebedev, and E. Sverdlov, *Endogenous retroviruses and human evolution.* Comp Funct Genomics, 2002. **3**(6): p. 494-8.
4. Stoye, J.P., *Endogenous retroviruses: still active after all these years?* Curr Biol, 2001. **11**(22): p. R914-6.
5. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
6. Sverdlov, E.D., *Retroviruses and primate evolution.* Bioessays, 2000. **22**(2): p. 161-71.
7. Stocking, C. and C.A. Kozak, *Murine endogenous retroviruses.* Cell Mol Life Sci, 2008. **65**(21): p. 3383-98.
8. Le Tissier, P., et al., *Two sets of human-tropic pig retrovirus.* Nature, 1997. **389**(6652): p. 681-2.
9. Fiebig, U., et al., *Transspecies transmission of the endogenous koala retrovirus.* J Virol, 2006. **80**(11): p. 5651-4.
10. Patience, C., Y. Takeuchi, and R.A. Weiss, *Infection of human cells by an endogenous retrovirus of pigs.* Nat Med, 1997. **3**(3): p. 282-6.
11. Czauderna, F., et al., *Establishment and characterization of molecular clones of porcine endogenous retroviruses replicating on human cells.* J Virol, 2000. **74**(9): p. 4028-38.
12. Deng, Y.M., B.E. Tuch, and W.D. Rawlinson, *Transmission of porcine endogenous retroviruses in severe combined immunodeficient mice xenotransplanted with fetal porcine pancreatic cells.* Transplantation, 2000. **70**(7): p. 1010-6.
13. Paradis, K., et al., *Search for cross-species transmission of porcine endogenous retrovirus in patients treated with living pig tissue. The XEN 111 Study Group.* Science, 1999. **285**(5431): p. 1236-41.
14. Takeuchi, Y., et al., *Host range and interference studies of three classes of pig endogenous retrovirus.* J Virol, 1998. **72**(12): p. 9986-91.
15. Niebert, M. and R.R. Tonjes, *Evolutionary spread and recombination of porcine endogenous retroviruses in the suiformes.* J Virol, 2005. **79**(1): p. 649-54.
16. Mang, R., et al., *Identification of a novel type C porcine endogenous retrovirus: evidence that copy number of endogenous retroviruses increases during host inbreeding.* J Gen Virol, 2001. **82**(Pt 8): p. 1829-34.
17. Bartosch, B., et al., *Evidence and consequence of porcine endogenous retrovirus recombination.* J Virol, 2004. **78**(24): p. 13880-90.
18. Patience, C., et al., *Multiple groups of novel retroviral genomes in pigs and related species.* J Virol, 2001. **75**(6): p. 2771-5.
19. Ericsson, T., et al., *Identification of novel porcine endogenous betaretrovirus sequences in miniature swine.* J Virol, 2001. **75**(6): p. 2765-70.
20. Klymiuk, N., et al., *Characterization of porcine endogenous retrovirus gamma pro-pol nucleotide sequences.* J Virol, 2002. **76**(22): p. 11738-43.
21. Tarailo-Graovac, M. and N. Chen, *Using RepeatMasker to identify repetitive elements in genomic sequences.* Curr Protoc Bioinformatics, 2009. **Chapter 4**: p. Unit 4 10.

22. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.
23. Kurtz, S., et al., *REPuter: the manifold applications of repeat analysis on a genomic scale.* Nucleic Acids Res, 2001. **29**(22): p. 4633-42.
24. Bao, Z. and S.R. Eddy, *Automated de novo identification of repeat sequence families in sequenced genomes.* Genome Res, 2002. **12**(8): p. 1269-76.
25. Campagna, D., et al., *RAP: a new computer program for de novo identification of repeated sequences in whole genomes.* Bioinformatics, 2005. **21**(5): p. 582-8.
26. Edgar, R.C. and E.W. Myers, *PILER: identification and classification of genomic repeats.* Bioinformatics, 2005. **21 Suppl 1**: p. i152-8.
27. McCarthy, E.M. and J.F. McDonald, *LTR_STRUC: a novel search and identification program for LTR retrotransposons.* Bioinformatics, 2003. **19**(3): p. 362-7.
28. Kalyanaraman, A. and S. Aluru, *Efficient algorithms and software for detection of full-length LTR retrotransposons.* J Bioinform Comput Biol, 2006. **4**(2): p. 197-216.
29. Xu, Z. and H. Wang, *LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W265-8.
30. Ellinghaus, D., S. Kurtz, and U. Willhoeft, *LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons.* BMC Bioinformatics, 2008. **9**: p. 18.
31. Steinbiss, S., et al., *Fine-grained annotation and classification of de novo predicted LTR retrotransposons.* Nucleic Acids Res, 2009. **37**(21): p. 7002-13.
32. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.
33. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2008. **36**(Database issue): p. D281-8.
34. Takahashi, H., T. Awata, and H. Yasue, *Characterization of swine short interspersed repetitive sequences.* Anim Genet, 1992. **23**(5): p. 443-8.
35. Yasue, H. and Y. Wada, *A swine SINE (PRE-1 sequence) distribution in swine-related animal species and its phylogenetic analysis in swine genome.* Anim Genet, 1996. **27**(2): p. 95-8.
36. Akiyoshi, D.E., et al., *Identification of a full-length cDNA for an endogenous retrovirus of miniature swine.* J Virol, 1998. **72**(5): p. 4503-7.
37. Johnson, W.E. and J.M. Coffin, *Constructing primate phylogenies from ancient retrovirus sequences.* Proc Natl Acad Sci U S A, 1999. **96**(18): p. 10254-60.
38. Tonjes, R.R. and M. Niebert, *Relative age of proviral porcine endogenous retrovirus sequences in Sus scrofa based on the molecular clock hypothesis.* J Virol, 2003. **77**(22): p. 12363-8.
39. Lamprecht, B., et al., *Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma.* Nat Med. **16**(5): p. 571-9, 1p following 579.

Supplementary data

Tab. 1: Chromosomal position of classified LTR-retrotransposons in the pig genome

PERV group	chromosome	start	end	length
β1/2	chr8	38594232	38600751	6519
β1/2	chr8	38823069	38829582	6513
β1/2	chrX	65342601	65350413	7812
β3	chr1	244167773	244174016	6243
β3	chr11	30445594	30451713	6119
β3	chr11	31932512	31937667	5155
β3	chr15	9782530	9788986	6456
β3	chr2	132383120	132391347	8227
β3	chrX	124410708	124419239	8531
β3	chrX	124433234	124439277	6043
β3	chrX	47401440	47409739	8299
β3	chrX	65642061	65647903	5842
β3	chrX	69521389	69529911	8522
cluster1	chr1	205992832	206000496	7664
cluster1	chr1	206116860	206124523	7663
cluster1	chr1	72409300	72416707	7407
cluster1	chr11	13190065	13197793	7728
cluster1	chr18	36891323	36899068	7745
cluster1	chr2	12794390	12802426	8036
cluster1	chr2	38025001	38032735	7734
cluster1	chr2	70945046	70954552	9506
cluster1	chr3	26741751	26750735	8984
cluster1	chr4	27430622	27438468	7846
cluster1	chr8	44433418	44440820	7402
cluster1	chr9	33252501	33265000	12499
cluster1	chrX	124579266	124586709	7443
cluster1	chrX	124659058	124666770	7712
cluster1	chrX	31219118	31226542	7424
cluster1	chrX	62435684	62446258	10574
cluster1	chrX	99094177	99101894	7717
cluster2	chr14	63096027	63102032	6005
cluster2	chrX	70600253	70607125	6872
cluster2	chrX	95338269	95345673	7404
cluster2	chrX	95440427	95447831	7404
γ1	chr1	138418098	138426181	8083
γ1	chr1	275966368	275974124	7756
γ1	chr1	40135413	40141166	5753
γ1	chr1	40216893	40222943	6050
γ1	chr10	65647784	65656730	8946

γ1	chr12	25923871	25932064	8193
γ1	chr13	106418102	106427019	8917
γ1	chr13	109324472	109333417	8945
γ1	chr13	109333422	109338898	5476
γ1	chr14	64965230	64972892	7662
γ1	chr17	3541380	3552021	10641
γ1	chr17	3607474	3615959	8485
γ1	chr2	46839767	46847223	7456
γ1	chr3	46346530	46355287	8757
γ1	chr4	46187514	46196274	8760
γ1	chr4	46259261	46268021	8760
γ1	chr7	113235746	113243613	7867
γ1	chr8	12214588	12223347	8759
γ1	chr8	44474694	44483610	8916
γ1	chr9	130569858	130578614	8756
γ1	chrX	110721675	110726573	4898
γ1	chrX	70857742	70862894	5152
γ1	chrX	77092996	77101919	8923
γ10	chr15	9796901	9803931	7030
γ2	chr1	149918569	149927244	8675
γ2	chr1	150142662	150151291	8629
γ2	chr1	293839378	293846908	7530
γ2	chr1	84812586	84821386	8800
γ2	chr10	65738454	65747034	8580
γ2	chr13	52134489	52142138	7649
γ2	chr14	112456880	112465833	8953
γ2	chr15	23153633	23160452	6819
γ2	chr16	11514740	11523906	9166
γ2	chr2	138941045	138948359	7314
γ2	chr2	139989337	139998868	9531
γ2	chr2	55274800	55283610	8810
γ2	chr2	56683211	56692030	8819
γ2	chr2	62001994	62010067	8073
γ2	chr2	66890134	66898053	7919
γ2	chr4	47774484	47782580	8096
γ2	chr6	10014100	10022707	8607
γ2	chr6	83229913	83238984	9071
γ2	chr7	105029243	105037429	8186
γ2	chr7	25535353	25550270	14917
γ2	chr7	61122604	61131482	8878
γ2	chr7	99847272	99856748	9476
γ2	chr9	1069580	1077579	7999
γ2	chr9	1206293	1214264	7971
γ2	chr9	1307496	1315485	7989
γ2	chr9	40929818	40938696	8878

γ2	chrX	46511409	46519587	8178
γ2	chrX	47092852	47101699	8847
γ2	chrX	47201251	47210098	8847
γ2	chrX	48798499	48806522	8023
γ2	chrX	49069448	49077644	8196
γ2	chrX	50926776	50935855	9079
γ2	chrX	77138337	77145676	7339
γ2	chrX	77249888	77257227	7339
γ2	chrX	78187139	78196018	8879
γ2	chrX	78296906	78305795	8889
γ3/5	chr11	35526267	35541700	15433
γ4	chr1	53594652	53603082	8430
γ7	chrX	60953412	60962472	9060