PLOS ONE

RESEARCH ARTICLE

A computational method for prediction of matrix proteins in endogenous retroviruses

Yucheng Ma, Ruiling Liu*, Hongqiang Lv*, Jiuqiang Han, Dexing Zhong, Xinman Zhang

School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China

* meggie@mail.xjtu.edu.cn (RLL); hongqianglv@mail.xjtu.edu.cn (HQL)

Citation: Ma Y, Liu R, Lv H, Han J, Zhong D, Zhang X (2017) A computational method for prediction of matrix proteins in endogenous retroviruses. PLoS ONE 12(5): e0176909. https://doi.org/10.1371/journal.pone.0176909

Editor: Jean-Luc EPH Darlix, "INSERM", FRANCE

Received: September 2, 2016

Accepted: April 19, 2017

Published: May 4, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grant from National Natural Science Foundation of Shaanxi Province (No. 2012JQ8042) at http://www.sninfo.gov.cn and by grant from China Postdoctoral Science Foundation (No. 2015M580851) at http://jj.chinapostdoctor.org.cn. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Human endogenous retroviruses (HERVs) encode active retroviral proteins, which may be involved in the progression of cancer and other diseases. Matrix protein (MA), in group-specific antigen genes (*gag*) of retroviruses, is associated with the virus envelope glycoproteins in most mammalian retroviruses and may be involved in virus particle assembly, transport and budding. However, the amount of annotated MAs in ERVs is still at a low level so far. No computational method to predict the exact start and end coordinates of MAs in gags has been proposed yet. In this paper, a computational method to identify MAs in ERVs is proposed. A divide and conquer technique was designed and applied to the conventional prediction model to acquire better results when dealing with gene sequences with various lengths. Initiation sites and termination sites were predicted separately and then combined according to their intervals. Three different algorithms were applied and compared: weighted support vector machine (WSVM), weighted extreme learning machine (WELM) and random forest (RF). $G-mean$ (geometric mean of sensitivity and specificity) values of initiation sites and termination sites under 5-fold cross validation generated by random forest models are 0.9869 and 0.9755 respectively, highest among the algorithms applied. Our prediction models combine RF & WSVM algorithms to achieve the best prediction results. 98.4% of all the collected ERV sequences with complete MAs (125 in total) could be predicted exactly correct by the models. 94,671 HERV sequences from 118 families were scanned by the model, 104 new putative MAs were predicted in human chromosomes. Distributions of the putative MAs and optimizations of model parameters were also analyzed. The usage of our predicting method was also expanded to other retroviruses and satisfying results were acquired.

Introduction

Human endogenous retroviruses (HERVs) are remnants of ancient retroviral infections. HERVs and their related genetic elements make up 504 distinct families and compose ~8% of human genome [1]. Typical full-length HERVs are about 7-11kb in size and consist mainly of the coding regions for *gag*, *pro*, *pol*, and *env* genes, flanked on both 5'- and 3'- ends by long terminal repeats (LTR). Most HERVs in human genome have incomplete structures [2], which

Competing interests: The authors have declared that no competing interests exist.

contain multiple stop codons, insertions, deletions and frame shift mutations [3,4]. HERVs encode active retroviral proteins, which may exert important physiological functions in the body, but may also be involved in the progression of cancer and numerous human autoimmune, neurological and infectious diseases. In addition, HERVs regulate expression of the neighboring host genes and modify the genomic regulatory landscape [5].

The shortage of organs for transplantation is a major barrier to the treatment of organ failure. While porcine organs are considered promising, their use has been checked by concerns about transmission of porcine endogenous retroviruses (PERVs) to humans. The risk of infections of human recipients after xenotransplantations is now mainly represented by PERVs as these particles are part of the porcine genome. It was found that PERV infection of the HEK-293 cell line alters expression of HERV sequences [6]. However, this risk isn't impossible to overcome, considering that the possibility that PERVs can be inactivated for clinical application to porcine-to-human xenotransplantation has been demonstrated in a recent research [7]. The close relationship between HERVs and PERVs reminds us of the importance of bioinformatics research to be carried out combining different mammal ERVs.

Group-specific antigen (*gag*) is the genetic material that codes for the core structural proteins of a retrovirus. *Gag* is one of the three "main" genes found in all retroviruses (along with env and pol). *Gag*s have close relationship to many serious diseases such as AIDS and cancer. A previous research revealed that human endogenous retrovirus K *gag* coassembles with HIV-1 *gag* and reduces the release efficiency and infectivity of HIV-1 [8]. It was found 2 years ago in another research that prostate cancer progression correlates with increased humoral immune response to a human endogenous retrovirus *gag* protein [9]. However, the amount of *gag* in HERVs found by experimental methods is still at a low level. The lack of annotated *gag*s in HERVs is a barrier that has to be removed for the convenience of subsequent structure analysis and function study on *gag*s in HERVs. A computational method to predict *gag*s from HERVs has been brought up [10], but no computational method to predict exact start and end coordinates of interior genes of *gag*s which encode functional proteins has been proposed yet. RetroTector is a platform independent program package which could detect candidate long terminal repeats (LTR) in retroviruses as well as chains of conserved retroviral motifs (including motifs of MAs) fulfilling distances constraints [11]. However, RetroTector is based on sequence alignment, thus only motifs instead of the exact start and end coordinates of MAs could be predicted. And these conserved retroviral motifs are eventually combined and used as basis of the detection of retroviruses.

Gag contains around 1500 nucleotides, and encodes three separate proteins which form the building blocks of the viral core. The three proteins are:

1. Matrix protein, MA
2. Capsid protein, CA
3. Nucleocapsid protein, NC

Matrix protein (MA) is associated with the virus envelope glycoproteins in most mammalian retroviruses and may be involved in virus particle assembly, transport and budding. Membrane binding in HIV-1 replication process is mediated by the MA, a 132-residue polypeptide containing an N-terminal myristyl group that can adopt sequestered and exposed conformations[12]. Single amino acid changes in the HIV-1 matrix protein block virus particle production[13]. The length of a MA found in endogenous retroviruses varies from 88aa to 127aa according to records in National Center for Biotechnology Information (NCBI). Computational method to predict interior genes of *gag*s which encodes functional proteins would

benefit subsequent structure analysis and function study on *gag*s, but we have to overcome the difficulty of the shortage of annotated *gag* sequences in HERVs first.

Considering the importance of the relationship between *gag*s in HERVs and ERVs from other mammals, such as PERVs, *gag*s from different mammal ERVs could be combined to build up models for their interior gene prediction (i.e. MA).

In this paper, a computational model to identify MAs in ERVs was proposed. All ten parameters of divide physicochemical property scores (DPPS) [14] along with position weight matrix (PWM) were utilized to generate the feature space for MA prediction. An unconventional "divide and conquer" (D&C) technique was applied to acquire high prediction accuracy when dealing with sequences that are poorly conserved in their lengths (unlike the traditional D&C technique in computer science, D&C technique applied here is not intended to reduce the computational complexity of the algorithm, but to make conventional gene prediction algorithms designed for fix-length gene prediction also capable of predicting sequences with various lengths). Initiation sites and termination sites were predicted separately and then combined according to their intervals. Massive DNA sequences related with coding regions from 118 HERV families were scanned with the prediction model, which has high prediction accuracy under 5-fold cross validation test.

Materials and methods

Datasets

All available amino acid sequences of ERVs from various organisms were collected from NCBI at http://www.ncbi.nlm.nih.gov [15–68]. Among them, all 129 sequences of ERVs with MAs annotated in their *gag*s were used to build up the prediction model (please refer to S1 File for details). One hundred and twenty five of them are with both initiation sites and termination sites and the other four are with initiation sites only.

"Divide and conquer"

In computer science, divide and conquer is an algorithm design paradigm based on multi-branched recursion. A divide and conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same or related type, until these become simple enough to be solved directly. The solutions to the sub-problems are then combined to give a solution to the original problem.

Traditional gene prediction methods would lose their accuracy or even feasibility when dealing with gene sequences with large length variations (because the dimension of the feature space couldn't be constant), even though they could bring out ideal prediction results when dealing with fixed length gene sequences.

The idea of D&C inspired us to solve such problem. Unlike conventional D&C, the recursion level in our problem is only 2 because our main purpose is dividing the original problem into two simpler sub-problems instead of reducing the time complexity of it. Our prediction method focuses more on the boundaries of the genes instead of the interior areas, because the former has more to do with gene prediction. We broke down the problem into two simpler sub-problems (fixed length gene prediction) and then combined the solutions of them to generate prediction results of the original gene prediction problem. The two sub-problems are initiation site and termination site prediction, which could be done well with traditional gene prediction method because we could consider them as fixed length gene prediction problems. We just need to predict the fixed length flanking residues of the initiation site and termination site to deduce the precise locations of them. Then we just need to find a reasonable combination of the predicted initiation sites and termination sites to generate gene prediction results

with high accuracy. We predicted the initiation site and termination site separately, and then regarded the sequence between them as a candidate MA sequence only when it has an appropriate length. The advantage of this divide and conquer technique is that feasibility and high accuracy could be obtained at the same time.

Sample preparation

Training samples are prepared from the amino acid sequences with MA annotations. Positive training samples for initiation sites consisted of *s*-aa-long subsequences starting from the initiation sites. The best prediction result was obtained when *s* was set to be 15 (please refer to Discussion part for more details). Likewise, positive training samples for termination sites consisted of 15-aa-long subsequences ending at the termination sites. Negative training samples consisted of 15-aa-long subsequences from regions either without MAs or overlapping with MAs but not sharing the same initiation or termination sites with them. To overcome the difficulty of the lack of positive training samples, we generated negative training samples with a size 5 times as large as the positive sample size and took the imbalanced data problem into our consideration in the modelling process. Thus the training sets for initiation site prediction model and termination site prediction model were built separately.

Feature selection

Combining position characteristics of sequences and physicochemical properties, a hybrid feature space construction approach was proposed.

Position weight matrix (PWM) [69] was applied to extract the position characteristic of sequences. A PWM has one row for each symbol of the alphabet: 20 rows for 20 kinds of amino acids in this case. It also has one column for each position of the 15-aa-long pattern. So a 20×15 matrix was built to represent the different frequencies of 20 kinds of amino acid appearing on various positions of the 15-aa-long motifs in this case. To construct a PWM, a basic position frequency matrix (PFM) is created by counting the occurrences of each nucleotide at each position at first. From the PFM, a position probability matrix (PPM) can be created by dividing that former nucleotide count at each position by the number of sequences, thereby normalising the values. Formally, given a set X of N aligned sequences of length l, the elements of the PPM M^{PPM} are calculated:

$$M_{k,j}^{PPM} = \frac{1}{N}\sum_{i=1}^{N} I(X_{i,j} = k) \quad (1)$$

Where $i \in (1,\ldots,N)$, $j \in (1,\ldots,l)$, k is the set of symbols in the alphabet and $I(a = k)$ is an indicator function where $I(a = k)$ is 1 if $a = k$ and 0 otherwise.

Most often the elements in PWMs are calculated as log likelihoods. That is, the elements of the PWM are transformed using a background model b so that:

$$M_{k,j}^{PWM} = \log(M_{k,j}^{PPM}/b_k) \quad (2)$$

The above equation describes how an element in the PWM M^{PWM} is calculated. The simplest background model assumes that each letter appears equally frequently in the dataset. That is, the value of $b_k = 1 / |k|$ for all symbols in the alphabet ($|k| = 20$ for amino acids, so $b_k = 0.05$).

After generating the PWM matrix with positive sequences (please refer to S5 File for details of PWM matrix of MA initiation sites and S6 File for details of PWM matrix of MA termination sites), a mapping method is used to extract the position characteristic of any 15-aa-long

sequence V. Assign each amino acid of V with its corresponding value in the matrix according to its position. Then a 15-dimension-vector V^{Pos} to represent the position characteristic of the original 15-aa-long sequence could be generated.

$$V_j^{Pos} = M_{k,j}^{PWM} \quad (3)$$

Where $j \in (1,\ldots,l)$, $k = V_j$, $l = 15$.

All ten parameters of the divided physicochemical property scores (DPPS) [14] were selected to extract the physicochemical properties of sequences. The parameters consist of 4 electronic properties, 2 steric properties, 2 hydrophobic properties and 2 hydrogen bond properties. Similarly, when dealing with a 15-aa-long sequence, the sequence was mapped into a 10×15 matrix to represent its physicochemical properties.

Combining the above two kinds of information, (1+10)×15 = 165 features in total were extracted from each 15-aa-long sequence for prediction. To get a persuasive performance comparison of different prediction models, we ran the following binary classifiers on the same 165-dimensional feature space.

Binary classifiers

Three binary classifiers based on different principles were applied to predict the initiation sites and termination sites of the MA sequences:

WSVM classifier. The support vector machine (SVM) is a supervised machine learning algorithm based on the statistical learning theory [70]. The basic thought of SVM is to map the original data into a high dimensional feature space through a nonlinear mapping function and then construct a hyper plane as a discriminative surface between the positive and negative data [71]. Weighted SVM (WSVM) is able to deal with data with imbalanced class distribution while maintaining a good performance. In this paper, WSVM was employed to solve both the initiation site prediction and the termination site prediction, which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

WELM classifier. Extreme learning machine (ELM) is a kind of artificial neural network and works for the "generalized" single-hidden-layer feed forward networks (SLFNs), the hidden layer (or called feature mapping) in ELM need not to be tuned. Compared with traditional computational intelligence techniques, ELM provides better generalization performance at a much faster learning speed. It has milder optimization constraints and with least human intervention [72]. Weighted ELM (WELM) also works well with data with imbalanced class distribution, and it is available at http://www.ntu.edu.sg/home/egbhuang/. In the paper, WELM was also used to solve the classification problem of unbalanced training samples of MA.

RF classifier. Random forest (RF) is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode) of the classes (classification) or mean prediction (regression) of the individual trees[73]. Random forest correct for decision trees' habit of overfitting to their training set. Random forest algorithm is also employed to solve the MA prediction problem, and it is available at https://cran.r-project.org/web/packages/randomForest/.

Boundary combination

When the putative MA initiation sites and termination sites were predicted, a proper combination method should be proposed to get the prediction of the entire MA sequences. As we may acquire more than one putative initiation sites and more than one putative termination sites

(making up even more putative boundary pairs) in one unannotated *gag* sequence, a method that could abandon the redundant false putative results and leave only one putative boundary pair as the final prediction results is required. Such requirement could be accomplished through the following 2 steps:

1. Choose the putative initiation sites and termination sites predicted by the RF models (please refer to S7 File for details of RF model for MA initiation sites and S8 File for details of RF model for MA termination sites) that possess distances within the range of the lengths of MA sequences (88 to 127 aa) as candidate boundary pairs.
2. Leave the candidate boundary pair that is predicted to be possible boundary pairs by WSVM models (please refer to S9 File for details of WSVM model for MA initiation sites and S10 File for details of WSVM model for MA termination sites) as well and also has the highest production value of its initiation site decision value and termination site decision value generated from the WSVM models as the final MA prediction result of the unannotated gag sequence. (A decision value is an important basis for the prediction result generated by a WSVM model. It is generated according to the degree of similarity between the predicted sample and training samples. It ranges from 0 to 1. The larger the decision value, the more likely the prediction result is positive, vise versa.)

Advantage of this technique:

1. Provides a way to rule out redundant MA boundary predictions and leave the most probable boundary pair as the final prediction result.
2. The final results have the advantages of both the RF models and WSVM models. They possess high sensitivity value provided by the RF models and high specificity value brought by the WSVM models (please refer to Results part for more details). By using the prediction results of RF models as candidate boundary pairs, we could reduce the omission rate of positive boundary pair. And by applying the WSVM model, false positive boundary pairs were ruled out as much as we could.

Performance assessment

N-fold cross-validation and Jack-knife test are usually used to illuminate the performance of a prediction model. Since 5-fold and 10-fold cross-validation were found to work better than Jack- knife test [74], 5-fold and 10-fold cross-validation were employed to assess the performance of the models in this paper.

True positive (TP) and false negative (FN) are the number of positive samples that are predicted to be positive and negative respectively. Analogously, true negative (TN) and false positive (FP) are used to denote the number of negative samples that are predicted to be negative and positive respectively.

Sensitivity S_n (also called the true positive rate) measures the proportion of positive samples that are correctly identified as such. Specificity S_p (also called the true negative rate) measures the proportion of negative samples that are correctly identified as such.

Overall accuracy *ACC* denotes the proportion of the testing samples correctly predicted. Usually *ACC* is used to measure the effectiveness of a classifier. Unfortunately, in presence of imbalanced data, this metric may fail to provide adequate information about the performance of the classifier. For instance, when given a binary classification problem consisting of 1 percent positive sample and 99 percent negative class, any dumb classifier would easily achieve 99 percent accuracy by classifying all the samples as negative.

Matthew's correlation coefficient *MCC* is also used in machine learning as a measure of the quality of binary classifications and it's generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

G – mean is an evaluation metric adopted in this paper to give more insight into the accuracy obtained within each class. As the geometric mean of the prediction accuracy of the positive samples and the prediction accuracy of the negative samples, *G – mean* could provide reasonable evaluation for the performance of the prediction model when dealing with imbalanced data. As with the 1:99 example, *G – mean* could be as low as 0 when the classifier is dump and could only classify all the samples as negative.

In this paper, *G – mean* under 5-fold cross-validation was selected as the major performance evaluation measure of the models to provide basis for parameter selection of models. S_n, S_p, *ACC* and *MCC* were also calculated as a supplemental reference.

$$\begin{cases} S_n = \dfrac{TP}{TP + FN} \\ S_p = \dfrac{TN}{TN + FP} \\ ACC = (TP + TN)/(TP + TN + FN + FP) \\ MCC = \dfrac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ G - mean = \sqrt{S_n S_p} = \sqrt{\dfrac{TP}{TP + FN} \times \dfrac{TN}{TN + FP}} \end{cases} \quad (4)$$

Putative MA detection

Once the RF models with high accuracy for initiation and termination site prediction were trained, putative MA sequences could be obtained by applying sliding window technique. When an unannotated sequence was analyzed, a 15-aa-long sliding window was used to "observe" the sequence. As with the initiation site prediction, the prediction result of the 15-aa-long subsequence in the window could be acquired when the subsequence was put into the RF model which was previously trained to distinguish MA initiation sites. Then we can obtain the putative MA initiation site after sliding the window on the entire sequence. Similarly, we can obtain the putative MA terminal site on the same sequence as well. With the reasonable combination narrated in the 'Boundary combination' part, we could then finally decide whether the sequence between the putative MA initiation site and terminal site is a putative MA sequence or not.

Results

Performance of the method

Accuracy of the prediction of MA boundaries. The performance of the prediction models was tested by 5-fold cross-validation and shown in Table 1. From Table 1, we can find that RF models have the best prediction results. They have the highest S_n, *ACC*, *MCC* and *G – mean* values for MA prediction of both initiation sites and termination sites (*G – mean* values of initiation sites and termination sites are 0.9869 and 0.9755 respectively), while WSVM models have the highest S_p values.

Accuracy of the prediction of MA. All of the 125 ERV sequences collected with complete MAs were used to test the prediction performance of our prediction model (please refer to S2

Table 1. Prediction performance of models applying different algorithms.

MA boundary type	Algorithm	Sn	Sp	ACC	MCC	G-mean
MA initiation sites	WSVM	0.9023	**1**	0.9837	0.9406	0.9497
	WELM	0.9605	0.9992	0.9928	0.9738	0.9795
	RF	**0.9767**	0.9974	**0.9939**	**0.9783**	**0.9869**
MA termination sites	WSVM	0.9152	**1**	0.9859	0.9484	0.9561
	WELM	0.9456	0.9922	0.9844	0.9439	0.9683
	RF	**0.9536**	0.9982	**0.9908**	**0.9667**	**0.9755**

https://doi.org/10.1371/journal.pone.0176909.t001

File for more details). 123 of them were predicted completely correct. This means that 98.4% of the sequences could be predicted completely correct. The other 2 were predicted with only 2aa position deviations in their terminal sites. It is worth mentioning that all the initiation sites were predicted completely correct.

Putative MA detection results

The proposed model was used to search for new putative MAs of 118 HERV families from sequences without MA annotations. A total of 94,671 DNA sequences (please refer to S3 File for details) corresponding to coding regions of HERVs from RepeatMasker have been scanned. 104 new putative MAs (please refer to S4 File for details) were predicted in coding region sequences. The exact locations of these new putative MAs of HERVs in the human chromosomes have been described with CIRCOS [75] software and shown in Fig 1.

The angles of the dark red lines represent the exact positions of new putative MAs of HERVs in the human chromosomes. The length of the line is proportional to the product of the decision values of the initiation site and termination site of the corresponding MA.

Discussion

Conservative property of MA boundaries

Motifs of sequences adjacent to origins and terminals of MAs in ERVs were generated based on WebLogo version 2.8.2 (http://weblogo.berkeley.edu/logo.cgi) and shown in Fig 2. From the figure, we can defer that sequences on both ends of MAs in ERVs are fairly conservative. This explains why satisfying results could be obtained from models created to predict the origins and terminals of MAs in ERVs separately.

Distribution of the putative MAs

The number of MAs in HERVs of the 24 human chromosomes and the number of MAs per bp in HERVs of the 24 human chromosomes were shown in Fig 3.

Optimization of model parameters

Model parameters were optimized according to the prediction performance they eventually bring about. A parameter was settled when it could bring about the best prediction performance. To rule out random factors as much as possible, the whole prediction process was rerun for 10 times and the average of the model performance measurement values was calculated whenever a parameter value changes during the parameter optimization process. To

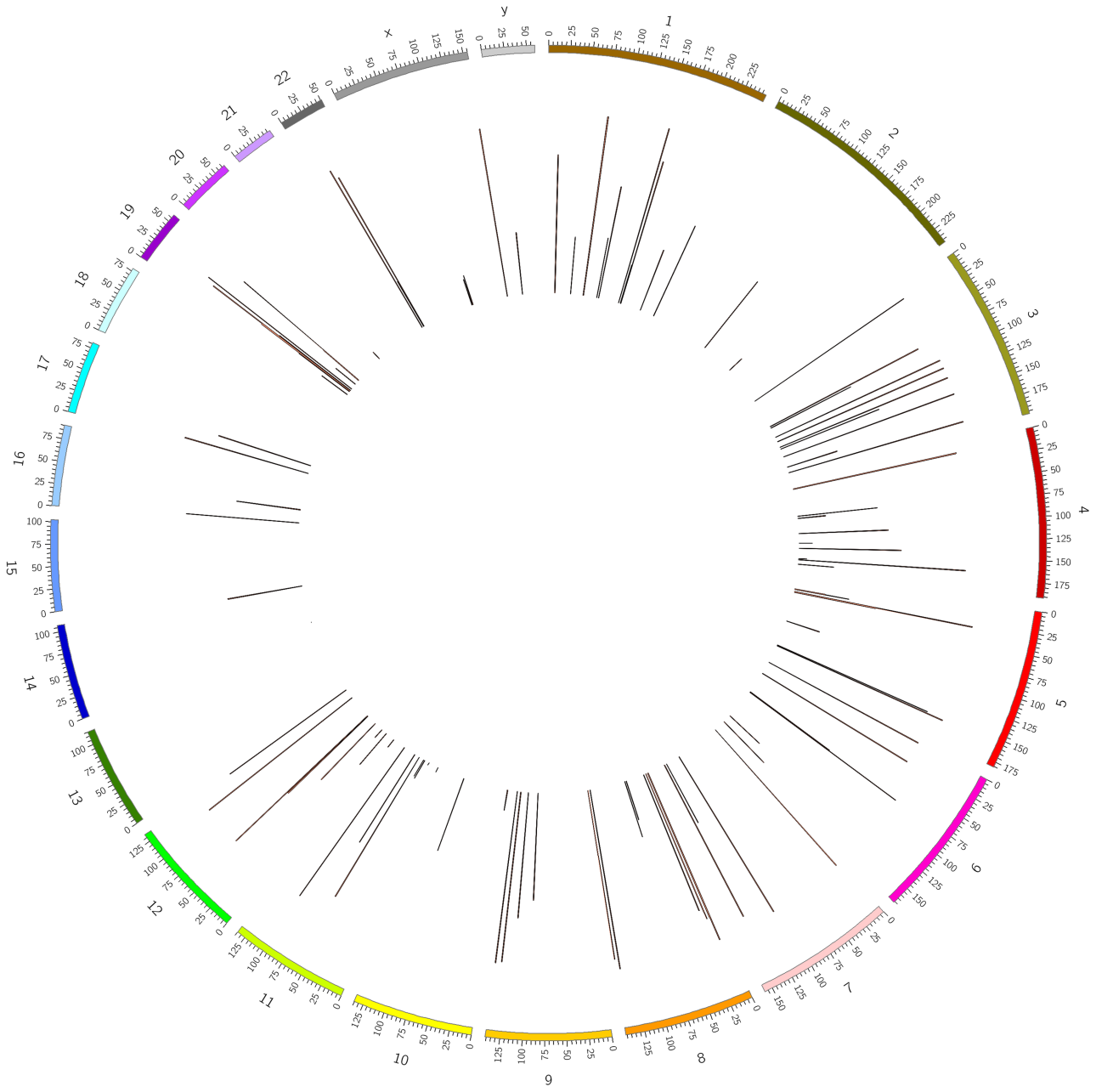

Fig 1. Exact locations of 104 new putative MAs of HERVs in the human chromosomes.

https://doi.org/10.1371/journal.pone.0176909.g001

choose the best values of parameters in the models, we adopted the method of cross validation based on grid search, avoiding the arbitrary and capricious behaviour.

The window length was selected according to its prediction performance. It was found that 15 is the best window length when the performance of the initiation site prediction model and the termination site prediction model are comprehensively considered.

The optimization details of model parameters in WSVM, WELM, RF models were shown in Table 2.

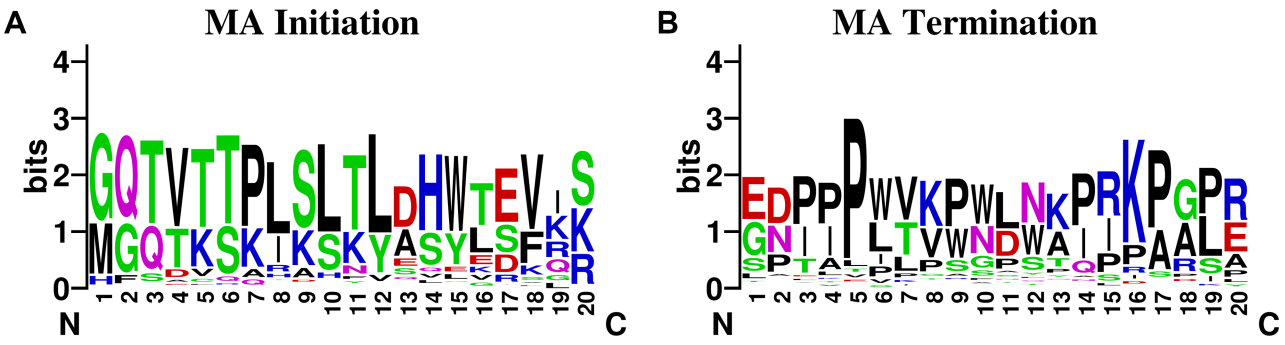

Fig 2. Motifs of residues adjacent to boundaries of MAs in ERV sequences. It shows motifs of surrounding residues of ERVs' (A) MA initiation sites, (B) MA Termination sites.

https://doi.org/10.1371/journal.pone.0176909.g002

Predicting effectiveness on other retroviruses

Retroviruses from different genuses might have different gag structures [76]. To test the predicting effectiveness on other retroviruses, all available gag sequences with MA annotated in Retroviridae, including Alpharetrovirus, Betaretrovirus, Gammaretrovirus, Deltaretrovirus, Epsilonretrovirus, Lentivirus and Spumavirus were collected from NCBI and used as source of training and testing sets to examine the effectiveness of our method. (No MAs were found annotated in Epsilonretrovirus and Spumavirus. Please refer to S11 File for more details about annotated MAs in Retroviridaes.)

Prediction models based on similar strategy were built for MAs in retroviruses from various genuses. Their effectiveness was also tested (Prediction source code is available at SourceForge, with the download URL: https://sourceforge.net/projects/ma-detection/files/MA%20prediction.zip/download). The dataset summary, model parameters and prediction results were shown in Tables 3 & 4.

From Tables 3 & 4, we can find that our prediction models could bring out favourable prediction results on sequences from various genuses of Retroviridae. Thus our predicting

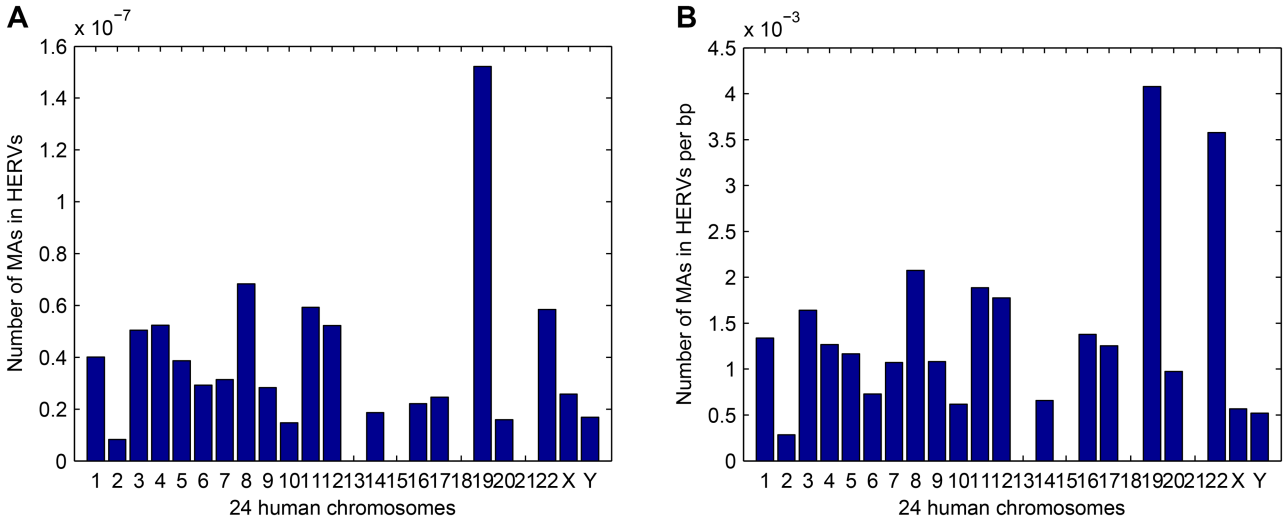

Fig 3. Distribution of MAs. (A) The number of MAs in HERVs of the 24 human chromosomes. (B)The number of MAs per bp in HERVs of the 24 human chromosomes.

https://doi.org/10.1371/journal.pone.0176909.g003

Table 2. Optimization details of parameters in WSVM, WELM, RF models.

MA boundary type	Algorithm	Model parameters	Step size in search	Value of optimized parameters
MA initiation sites	WSVM	c	0.0001	0.1895
		g	0.0001	0.0625
	WELM	Number of hidden neurons	50	2000
		C	50	9300
	RF	Number of trees	5	160
		mtry	5	80
MA termination sites	WSVM	c	0.0001	0.5743
		g	0.0001	0.1895
	WELM	Number of hidden neurons	50	1600
		C	50	5100
	RF	Number of trees	5	140
		mtry	5	50

https://doi.org/10.1371/journal.pone.0176909.t002

strategy (focusing on predicting MA start and end coordinates by combining RF and WSVM) is extensible to various genuses of Retroviridae.

Gammaretroviruses, such as murine leukemia viruses (MLVs), encode, in addition to the canonical gag, pol, and env proteins that will form progeny virus particles, a protein called "glycogag" (glycosylated Gag) [77]. All available glycosylated Gag sequences with MA annotated were downloaded from NCBI and scanned by our prediction model for Gammaretrovirus. All of their annotated boundaries were predicted totally correct. (Please refer to S12 File for more details about MA prediction in glycosylated Gags). It seems that the prediction of MAs in glycogag is not a special issue distinguished from normal Gags. This consist with the

Table 3. Prediction performance of models applied to MA boundaries from different retrovirus genuses.

MA Boundary Type	Organism	Number of sequences	Algorithm	Sn	Sp	Acc	MCC	G-mean
MA Initiation Sites	Alpharetrovirus	141	WSVM	0.9931	1	0.9988	0.9958	0.9965
			RF	0.9931	1	0.9988	0.9958	0.9965
	Betaretrovirus	95	WSVM	0.9895	1	0.9928	0.9936	0.9947
			RF	0.9979	0.9994	0.9991	0.9969	0.9986
	Gammaretrovirus	482	WSVM	0.9726	0.998	0.9938	0.9775	0.9852
			RF	0.9807	0.9965	0.9938	0.9779	0.9885
	Deltaretrovirus	234	WSVM	0.9812	1	0.9969	0.9887	0.9905
			RF	0.9872	1	0.9979	0.9923	0.9936
	Lentivirus	17272	WSVM	0.9619	0.9987	0.9926	0.9732	0.9801
			RF	0.9746	0.9982	0.9942	0.9792	0.9863
MA Initiation Sites	Alpharetrovirus	140	WSVM	0.9857	1	0.9976	0.9914	0.9928
			RF	0.9907	1	0.9985	0.9944	0.9953
	Betaretrovirus	98	WSVM	0.99	1	0.9983	0.9939	0.9949
			RF	1	1	1	1	1
	Gammaretrovirus	347	WSVM	0.9447	0.9959	0.9873	0.954	0.9699
			RF	0.9516	0.9978	0.9901	0.9639	0.9743
	Deltaretrovirus	181	WSVM	0.9892	0.9945	0.9936	0.9773	0.9918
			RF	0.9891	0.9989	0.9972	0.9901	0.9939
	Lentivirus	18234	WSVM	0.9074	0.9998	0.9844	0.9433	0.9523
			RF	0.9625	0.9983	0.9923	0.9723	0.9802

https://doi.org/10.1371/journal.pone.0176909.t003

Table 4. Prediction performance on sequences with intact MAs from different retrovirus genuses.

Organism	Intact Seq Amount	Init Acc Amount	Init Acc Rate	Term Acc Amount	Term Acc Rate	Boundaries Acc Amount	Boundaries Acc Rate
Alpharetrovirus	139	139	1	138	0.9928	138	0.9928
Betaretrovirus	95	95	1	95	1	95	1
Gammaretrovirus	341	336	0.9853	336	0.9853	332	0.9736
Deltaretrovirus	179	178	0.9944	171	0.9553	170	0.9497
Lentivirus	16292	15057	0.9242	15190	0.9324	14196	0.8713

https://doi.org/10.1371/journal.pone.0176909.t004

previous research that glycogag protein is identical in primary sequence to Gag except that it contains 88 additional residues at its N terminus [78].

Limits of the model

MA prediction based on identifying boundaries of MAs has the advantage of high efficiency and accuracy. However, some ERVs may not have a typical MA, like HERVL. Our prediction focuses on prediction of MAs with typical structures, thus it is not suitable for predicting non-canonical MAs.

Supporting information

S1 File. ERV sequences with MA annotations collected for research in this paper.
(XLS)

S2 File. MA prediction results of all 125 ERV sequences collected with complete MAs in this paper.
(XLS)

S3 File. Details of 94,671 DNA sequences corresponding to coding regions of HERVs from RepeatMasker.
(FSA)

S4 File. Details of 104 putative MAs in HERVs.
(XLS)

S5 File. Details of PWM matrix of MA initiation sites.
(MAT)

S6 File. Details of PWM matrix of MA termination sites.
(MAT)

S7 File. Details of RF model for MA initiation sites.
(MAT)

S8 File. Details of RF model for MA termination sites.
(MAT)

S9 File. Details of WSVM model for MA initiation sites.
(MAT)

S10 File. Details of WSVM model for MA termination sites.
(MAT)

S11 File. Details of all collected Retroviridae sequences with MAs annotated.
(XLSX)

S12 File. Details about MA prediction in glycosylated Gags.
(XLSX)

Acknowledgments

We are grateful to our colleagues in the School of Electronic and Information Engineering, Xi'an Jiaotong University for their help during the course of this work, in particular Dr. Shanxin Zhang and Dr. Ze Liu for critical reading and helpful discussions on the manuscript.

Author Contributions

Conceptualization: YCM RLL.

Data curation: YCM DXZ XMZ.

Formal analysis: YCM HQL.

Funding acquisition: RLL HQL.

Investigation: YCM RLL.

Methodology: YCM HQL.

Project administration: JQH.

Resources: DXZ XMZ.

Software: YCM JQH.

Supervision: RLL JQH.

Validation: YCM RLL.

Visualization: YCM RLL.

Writing – original draft: YCM RLL.

Writing – review & editing: RLL HQL.

References

1. Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26: 291–315. PMID: 12876457
2. Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. Annu Rev Genet 42: 709–732. https://doi.org/10.1146/annurev.genet.42.110807.091501 PMID: 18694346
3. de Parseval N, Casella JF, Gressin L, Heidmann T (2001) Characterization of the three HERV-H proviruses with an open envelope reading frame encompassing the immunosuppressive domain and evolutionary history in primates. Virology 279: 558–569. https://doi.org/10.1006/viro.2000.0737 PMID: 11162811
4. Kim HS (2001) Sequence and phylogeny of HERV-W pol fragments. Aids Research and Human Retroviruses 17: 1665–1671. https://doi.org/10.1089/088922201753342086 PMID: 11779355
5. Suntsova M, Garazha A, Ivanova A, Kaminsky D, Zhavoronkov A, Buzdin A (2015) Molecular functions of human endogenous retroviruses in health and disease. Cellular and Molecular Life Sciences 72: 3653–3675. https://doi.org/10.1007/s00018-015-1947-6 PMID: 26082181
6. Machnik G, Klimacka-Nawrot E, Sypniewski D, Matczynska D, Galka S, Bednarek I, et al. (2014) Porcine Endogenous Retrovirus (PERV) Infection of HEK-293 Cell Line Alters Expression of Human Endogenous Retrovirus (HERV-W) Sequences. Folia Biologica 60: 35–46. PMID: 24594055

7. Yang LH, Guell M, Niu D, George H, Lesha E, Grishin D, et al. (2015) Genome-wide inactivation of porcine endogenous retroviruses (PERVs). Science 350: 1101–1104. https://doi.org/10.1126/science.aad1191 PMID: 26456528
8. Monde K, Contreras-Galindo R, Kaplan MH, Markovitz DM, Ono A (2012) Human Endogenous Retrovirus K Gag Coassembles with HIV-1 Gag and Reduces the Release Efficiency and Infectivity of HIV-1. Journal of Virology 86: 11194–11208. https://doi.org/10.1128/JVI.00301-12 PMID: 22855497
9. Reis BS, Jungbluth AA, Frosina D, Holz M, Ritter E, Nakayama E, et al. (2013) Prostate Cancer Progression Correlates with Increased Humoral Immune Response to a Human Endogenous Retrovirus GAG Protein. Clinical Cancer Research 19: 6112–6125. https://doi.org/10.1158/1078-0432.CCR-12-3580 PMID: 24081977
10. Villesen P, Aagaard L, Wiuf C, Pedersen FS (2004) Identification of endogenous retroviral reading frames in the human genome. Retrovirology 1: 32. https://doi.org/10.1186/1742-4690-1-32 PMID: 15476554
11. Sperber GO, Airola T, Jern P, Blomberg J (2007) Automated recognition of retroviral sequences in genomic data—RetroTector (c). Nucleic Acids Research 35: 4964–4976. https://doi.org/10.1093/nar/gkm515 PMID: 17636050
12. Summers MF, Saad JS (2009) STRUCTURAL BASIS FOR TARGETING HIV-1 GAG PROTEINS TO THE PLASMA MEMBRANE FOR VIRUS ASSEMBLY. Proceedings of the National Academy of Science. pp. 11364–11369.
13. Freed EO, Orenstein JM, Buckler-White AJ, Martin MA (1994) Single amino acid changes in the human immunodeficiency virus type 1 matrix protein block virus particle production. Journal of Virology 68: 5311–5320. PMID: 8035531
14. Tian F, Yang L, Lv F, Yang Q, Zhou P (2009) In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach. Amino Acids 36: 535–554. https://doi.org/10.1007/s00726-008-0116-8 PMID: 18575802
15. Ono M, Yasunaga T, Miyata T, Ushikubo H (1986) Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J Virol 60: 589–598. PMID: 3021993
16. Boller K, Janssen O, Schuldes H, Tonjes RR, Kurth R (1997) Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. J Virol 71: 4581–4588. PMID: 9151852
17. Tonjes RR, Boller K, Limbach C, Lugert R, Kurth R (1997) Characterization of human endogenous retrovirus type K virus-like particles generated from recombinant baculoviruses. Virology 233: 280–291. https://doi.org/10.1006/viro.1997.8614 PMID: 9217052
18. Akiyoshi DE, Denaro M, Zhu H, Greenstein JL, Banerjee P, Fishman JA (1998) Identification of a full-length cDNA for an endogenous retrovirus of miniature swine. J Virol 72: 4503–4507. PMID: 9557749
19. Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J (1999) Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr Biol 9: 861–868. PMID: 10469592
20. Tonjes RR, Czauderna F, Kurth R (1999) Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K. J Virol 73: 9187–9195. PMID: 10516026
21. Mayer J, Sauter M, Racz A, Scherer D, Mueller-Lantzsch N, Meese E (1999) An almost-intact human endogenous retrovirus K on human chromosome 7. Nat Genet 21: 257–258. https://doi.org/10.1038/6766 PMID: 10080172
22. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, et al. (1999) The DNA sequence of human chromosome 22. Nature 402: 489–495. https://doi.org/10.1038/990031 PMID: 10591208
23. Voisset C, Bouton O, Bedin F, Duret L, Mandrand B, Mallet F, et al. (2000) Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family. AIDS Res Hum Retroviruses 16: 731–740. https://doi.org/10.1089/088922200308738 PMID: 10826480
24. Deng YM, Tuch BE, Rawlinson WD (2000) Transmission of porcine endogenous retroviruses in severe combined immunodeficient mice xenotransplanted with fetal porcine pancreatic cells. Transplantation 70: 1010–1016. PMID: 11045635
25. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11: 1531–1535. PMID: 11591322
26. Sugimoto J, Matsuura N, Kinjo Y, Takasu N, Oda T, Jinno Y (2001) Transcriptionally active HERV-K genes: identification, isolation, and chromosomal mapping. Genomics 72: 137–144. https://doi.org/10.1006/geno.2001.6473 PMID: 11401426
27. Krach U, Fischer N, Czauderna F, Tonjes RR (2001) Comparison of replication-competent molecular clones of porcine endogenous retrovirus class A and class B derived from pig and human cells. J Virol 75: 5465–5472. https://doi.org/10.1128/JVI.75.12.5465-5472.2001 PMID: 11356953

28. Bartosch B, Weiss RA, Takeuchi Y (2002) PCR-based cloning and immunocytological titration of infectious porcine endogenous retrovirus subgroup A and B. J Gen Virol 83: 2231–2240. https://doi.org/10.1099/0022-1317-83-9-2231 PMID: 12185278

29. Griffiths DJ, Voisset C, Venables PJ, Weiss RA (2002) Novel endogenous retrovirus in rabbits previously reported as human retrovirus 5. J Virol 76: 7094–7102. https://doi.org/10.1128/JVI.76.14.7094-7102.2002 PMID: 12072509

30. Scobie L, Taylor S, Wood JC, Suling KM, Quinn G, Meikle S, et al. (2004) Absence of replication-competent human-tropic porcine endogenous retroviruses in the germ line DNA of inbred miniature Swine. J Virol 78: 2502–2509. https://doi.org/10.1128/JVI.78.5.2502-2509.2004 PMID: 14963152

31. Bartosch B, Stefanidis D, Myers R, Weiss R, Patience C, Takeuchi Y (2004) Evidence and consequence of porcine endogenous retrovirus recombination. J Virol 78: 13880–13890. https://doi.org/10.1128/JVI.78.24.13880-13890.2004 PMID: 15564496

32. Martin J, Han C, Gordon LA, Terry A, Prabhakar S, She X, et al. (2004) The sequence and analysis of duplication-rich human chromosome 16. Nature 432: 988–994. https://doi.org/10.1038/nature03187 PMID: 15616553

33. Preuss T, Fischer N, Boller K, Tonjes RR (2006) Isolation and characterization of an infectious replication-competent molecular clone of ecotropic porcine endogenous retrovirus class C. J Virol 80: 10258–10261. https://doi.org/10.1128/JVI.01140-06 PMID: 17005704

34. Scherer SE, Muzny DM, Buhay CJ, Chen R, Cree A, Ding Y, et al. (2006) The finished DNA sequence of human chromosome 12. Nature 440: 346–351. https://doi.org/10.1038/nature04569 PMID: 16541075

35. Hirschl S, Schanab O, Seppele H, Waltenberger A, Humer J, Wolff K, et al. (2007) Sequence variability of retroviral particles derived from human melanoma cells melanoma-associated retrovirus. Virus Res 123: 211–215. https://doi.org/10.1016/j.virusres.2006.08.010 PMID: 17005285

36. Kim NY, Lee D, Lee J, Park EW, Jung WW, Yang JM, et al. (2009) Characterization of the replication-competent porcine endogenous retrovirus class B molecular clone originated from Korean domestic pig. Virus Genes 39: 210–216. https://doi.org/10.1007/s11262-009-0377-7 PMID: 19543822

37. Jung WY, Kim JE, Jung KC, Jin DI, Moran C, Park EW, et al. (2010) Comparison of PERV genomic locations between Asian and European pigs. Anim Genet 41: 89–92. https://doi.org/10.1111/j.1365-2052.2009.01953.x PMID: 19781037

38. Ma Y, Lv M, Xu S, Wu J, Tian K, Zhang J (2010) Identification of full-length proviral DNA of porcine endogenous retrovirus from Chinese Wuzhishan miniature pigs inbred. Comp Immunol Microbiol Infect Dis 33: 323–331. https://doi.org/10.1016/j.cimid.2008.10.007 PMID: 19070900

39. Yu SL, Jung WY, Jung KC, Cho IC, Lim HT, Jin DI, et al. (2012) Characterization of porcine endogenous retrovirus clones from the NIH miniature pig BAC library. J Biomed Biotechnol 2012: 482568. https://doi.org/10.1155/2012/482568 PMID: 21912484

40. Xiang S, Ma Y, Yan Q, Lv M, Zhao X, Yin H, et al. (2013) Construction and characterization of an infectious replication competent clone of porcine endogenous retrovirus from Chinese miniature pigs. Virol J 10: 228. https://doi.org/10.1186/1743-422X-10-228 PMID: 23837947

41. Escalera-Zamudio M, Mendoza ML, Heeger F, Loza-Rubio E, Rojas-Anaya E, Mendez-Ojeda ML, et al. (2015) A novel endogenous betaretrovirus in the common vampire bat (Desmodus rotundus) suggests multiple independent infection and cross-species transmission events. J Virol 89: 5180–5184. https://doi.org/10.1128/JVI.03452-14 PMID: 25717107

42. Pothlichet J, Heidmann T, Mangeney M (2006) A recombinant endogenous retrovirus amplified in a mouse neuroblastoma is involved in tumor growth in vivo. Int J Cancer 119: 815–822. https://doi.org/10.1002/ijc.21935 PMID: 16550601

43. Bartman T, Murasko DM, Blank KJ (1995) A replication-competent, endogenous retrovirus from an aged DBA/2 mouse contains the complete env from Emv-3 and a novel gag partially related to AKT-8. J Virol 69: 3224–3228. PMID: 7707556

44. Niebert M, Rogel-Gaillard C, Chardon P, Tonjes RR (2002) Characterization of chromosomally assigned replication-competent gamma porcine endogenous retroviruses derived from a large white pig and expression in human cells. J Virol 76: 2714–2720. https://doi.org/10.1128/JVI.76.6.2714-2720.2002 PMID: 11861838

45. Cingoz O, Paprotka T, Delviks-Frankenberry KA, Wildt S, Hu WS, Pathak VK, et al. (2012) Characterization, mapping, and distribution of the two XMRV parental proviruses. J Virol 86: 328–338. https://doi.org/10.1128/JVI.06022-11 PMID: 22031947

46. Tang HB, Ouyang K, Ma L, Bai A, Qin S, Chen F, et al. (2015) Complete Genome Sequence of a Porcine Endogenous Retrovirus Isolated from a Bama Minipig in Guangxi, Southern China. Genome Announc 3.

47. van der Kuyl AC, Mang R, Dekker JT, Goudsmit J (1997) Complete nucleotide sequence of simian endogenous type D retrovirus with intact genome organization: evidence for ancestry to simian retrovirus and baboon endogenous virus. J Virol 71: 3666–3676. PMID: 9094640

48. Triviai I, Ziegler M, Bergholz U, Oler AJ, Stubig T, Prassolov V, et al. (2014) Endogenous retrovirus induces leukemia in a xenograft mouse model for primary myelofibrosis. Proc Natl Acad Sci U S A 111: 8595–8600. https://doi.org/10.1073/pnas.1401215111 PMID: 24912157

49. Onions D, Côté C, Love B, Toms B, Koduri S, Armstrong A, et al. (2011) Ensuring the safety of vaccine cell substrates by massively parallel sequencing of the transcriptome. Vaccine 29: 7117–7121. https://doi.org/10.1016/j.vaccine.2011.05.071 PMID: 21651935

50. Czauderna F, Fischer N, Boller K, Kurth R, Tonjes RR (2000) Establishment and characterization of molecular clones of porcine endogenous retroviruses replicating on human cells. J Virol 74: 4028–4038. PMID: 10756014

51. Marsh AK, Willer DO, Skokovets O, Iwajomo OH, Chan JK, MacDonald KS (2012) Evaluation of cynomolgus macaque (Macaca fascicularis) endogenous retrovirus expression following simian immunodeficiency virus infection. PLoS One 7: e40158. https://doi.org/10.1371/journal.pone.0040158 PMID: 22768246

52. Huder JB, Boni J, Hatt JM, Soldati G, Lutz H, Schupbach J (2002) Identification and characterization of two closely related unclassifiable endogenous retroviruses in pythons (Python molurus and Python curtus). J Virol 76: 7607–7615. https://doi.org/10.1128/JVI.76.15.7607-7615.2002 PMID: 12097574

53. Wu HL, Leon EJ, Wallace LT, Nimiyongskul FA, Buechler MB, Newman LP, et al. (2016) Identification and spontaneous immune targeting of an endogenous retrovirus K envelope protein in the Indian rhesus macaque model of human disease. Retrovirology 13: 6. https://doi.org/10.1186/s12977-016-0238-0 PMID: 26767784

54. Lower R, Boller K, Hasenmaier B, Korbmacher C, Muller-Lantzsch N, Lower J, et al. (1993) Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci U S A 90: 4480–4484. PMID: 8506289

55. Miller AD, Bergholz U, Ziegler M, Stocking C (2008) Identification of the myelin protein plasmolipin as the cell entry receptor for Mus caroli endogenous retrovirus. J Virol 82: 6862–6868. https://doi.org/10.1128/JVI.00397-08 PMID: 18463156

56. Anai Y, Ochi H, Watanabe S, Nakagawa S, Kawamura M, Gojobori T, et al. (2012) Infectious endogenous retroviruses in cats and emergence of recombinant viruses. J Virol 86: 8634–8644. https://doi.org/10.1128/JVI.00280-12 PMID: 22674983

57. Beck-Engeser GB, Ahrends T, Knittel G, Wabl R, Metzner M, Eilat D, et al. (2015) Infectivity and insertional mutagenesis of endogenous retrovirus in autoimmune NZB and B/W mice. J Gen Virol 96: 3396–3410. https://doi.org/10.1099/jgv.0.000271 PMID: 26315139

58. Herring C, Quinn G, Bower R, Parsons N, Logan NA, Brawley A, et al. (2001) Mapping full-length porcine endogenous retroviruses in a large white pig. J Virol 75: 12252–12265. https://doi.org/10.1128/JVI.75.24.12252-12265.2001 PMID: 11711616

59. Pothlichet J, Mangeney M, Heidmann T (2006) Mobility and integration sites of a murine C57BL/6 melanoma endogenous retrovirus involved in tumor progression in vivo. Int J Cancer 119: 1869–1877. https://doi.org/10.1002/ijc.22066 PMID: 16708391

60. Shimode S, Nakagawa S, Miyazawa T (2015) Multiple invasions of an infectious retrovirus in cat genomes. Sci Rep 5: 8164. https://doi.org/10.1038/srep08164 PMID: 25641657

61. Herr W (1984) Nucleotide sequence of AKV murine leukemia virus. J Virol 49: 471–478. PMID: 6319746

62. Copeland NG, Jenkins NA, Nexo B, Schultz AM, Rein A, Mikkelsen T, et al. (1988) Poorly expressed endogenous ecotropic provirus of DBA/2 mice encodes a mutant Pr65gag protein that is not myristylated. J Virol 62: 479–487. PMID: 2826810

63. Cornelis G, Vernochet C, Carradec Q, Souquere S, Mulot B, Catzeflis F, et al. (2015) Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. Proc Natl Acad Sci U S A 112: E487–496. https://doi.org/10.1073/pnas.1417000112 PMID: 25605903

64. Wu T, Yan Y, Kozak CA (2005) Rmcf2, a xenotropic provirus in the Asian mouse species Mus castaneus, blocks infection by polytropic mouse gammaretroviruses. J Virol 79: 9677–9684. https://doi.org/10.1128/JVI.79.15.9677-9684.2005 PMID: 16014929

65. Yoshikawa R, Miyaho RN, Hashimoto A, Abe M, Yasuda J, Miyazawa T (2015) Suppression of production of baboon endogenous virus by dominant negative mutants of cellular factors involved in multivesicular body sorting pathway. Virus Res 196: 128–134. https://doi.org/10.1016/j.virusres.2014.11.020 PMID: 25463055

66. Mendoza R, Vaughan AE, Miller AD (2011) The left half of the XMRV retrovirus is present in an endogenous retrovirus of NIH/3T3 Swiss mouse cells. J Virol 85: 9247–9248. https://doi.org/10.1128/JVI.05137-11 PMID: 21697491

67. Hanger JJ, Bromham LD, McKee JJ, O'Brien TM, Robinson WF (2000) The nucleotide sequence of koala (Phascolarctos cinereus) retrovirus: a novel type C endogenous virus related to Gibbon ape leukemia virus. J Virol 74: 4264–4272. PMID: 10756041

68. Policastro PF, Fredholm M, Wilson MC (1989) Truncated gag products encoded by Gv-1-responsive endogenous retrovirus loci. J Virol 63: 4136–4147. PMID: 2789292

69. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the Perceptron Algorithm to Distinguish Translational Initiation Sites in Escherichia-Coli. Nucleic Acids Research 10: 2997–3011. PMID: 7048259

70. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10: 988–999. https://doi.org/10.1109/72.788640 PMID: 18252602

71. Cui Y, Han J, Zhong D, Liu R (2013) A novel computational method for the identification of plant alternative splice sites. Biochem Biophys Res Commun 431: 221–224. https://doi.org/10.1016/j.bbrc.2012.12.131 PMID: 23313482

72. Huang GB, Wang DH, Lan Y (2011) Extreme learning machines: a survey. International Journal of Machine Learning & Cybernetics 2: 107–122.

73. Ho TK. Random Decision Forests; 1995. pp. 278–282 vol.271.

74. Breiman L, Spector P. Submodel selection and evaluation in regression—the X-random case; 1992. pp. 291–319.

75. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639–1645. https://doi.org/10.1101/gr.092759.109 PMID: 19541911

76. Wills JW, Craven RC (1991) Form, function, and use of retroviral gag proteins. Aids 5: 639–654. PMID: 1883539

77. Ahi YS, Zhang S, Thappeta Y, Denman A, Feizpour A, Gummuluru S, et al. (2016) Functional Interplay Between Murine Leukemia Virus Glycogag, Serinc5, and Surface Glycoprotein Governs Virus Entry, with Opposite Effects on Gammaretroviral and Ebolavirus Glycoproteins. Mbio 7.

78. Prats AC, De BG, Wang P, Darlix JL (1989) CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus. Journal of Molecular Biology 205: 363–372. PMID: 2538626