A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K in human populations

Weiling Li[1], Lin Lin[2], Raunaq Malhotra[1,4], Lei Yang[3], Raj Acharya[1,5], and Mary Poss[3*]

[1] The School of Electrical Engineering and Computer Science,

[2] Department of Statistics, [3] Department of Biology and Veterinary and Biomedical Sciences

The Pennsylvania State University, University Park, PA, 16802, USA

[4] Commense Health, Cambridge, MA 02142, USA

[5] School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47405, USA

Running Title: Global genomic distribution of polymorphic HERV-K

* Corresponding author:

Mary Poss

814 404-7061

maryposs@gmail.com

Keywords HERV-K, human endogenous retrovirus, insertional polymorphism, 1000 Genomes Project, mixture model, visualization

Abstract

Human Endogenous Retrovirus type K (HERV-K) is the only HERV known to be insertionally polymorphic. It is possible that HERV-Ks contribute to human disease because people differ in both number and genomic location of these retroviruses. Indeed viral transcripts, proteins, and antibody against HERV-K are detected in cancers, auto-immune, and neurodegenerative diseases. However, attempts to link a polymorphic HERV-K with any disease have been frustrated in part because population frequency of HERV-K provirus at each site is lacking and it is challenging to identify closely related elements such as HERV-K from short read sequence data. We present an integrated and computationally robust approach that uses whole genome short read data to determine the occupation status at all sites reported to contain a HERV-K provirus. Our method estimates the proportion of fixed length genomic sequence (*k-mers*) from whole genome sequence data matching a reference set of *k-mers* unique to each HERV-K loci and applies mixture model-based clustering to account for low depth sequence data. Our analysis of 1000 Genomes Project Data (KGP) reveals numerous differences among the five KGP super-populations in the frequency of individual and co-occurring HERV-K proviruses; we provide a visualization tool to easily depict the prevalence of any combination of HERV-K among KGP populations. Further, the genome burden of polymorphic HERV-K is variable in humans, with East Asian (EAS) individuals having the fewest integration sites. Our study identifies population-specific sequence variation for several HERV-K proviruses. We expect these resources will advance research on HERV-K contributions to human diseases.

Author summary

Human Endogenous Retrovirus type K (HERV-K) is the youngest of retrovirus families in the human genome and is the only group that is polymorphic; a HERV-K can be present in one individual but absent from others. HERV-Ks could contribute to disease risk but establishing a link of a polymorphic HERV-K to a specific disease has been difficult. We develop an easy to use method that reveals the considerable variation existing among global populations in the frequency of individual and co-occurring polymorphic HERV-K, and in the total number of HERV-K that any individual has in their genome. Our study provides a global reference set of HERV-K genomic diversity and tools needed to determine the genomic landscape of HERV-K in any patient population.

Introduction

Endogenous retroviruses (ERVs) are derived from infectious retroviruses that integrated into a host germ cell at some time in the evolutionary history of a species [1–5]. ERVs in humans (HERVs) comprise up to 8% of the genome and have contributed important functions to their host [6–8]. The infection events that resulted in the contemporary profile of HERVs occurred prior to emergence of modern humans so most HERVs are fixed in human populations and those of closely related primates. However some HERVs retain the ability to replicate and reintegrate into germline so that individuals differ in the number and genomic location occupied by an ERV, a situation termed insertional polymorphism [9–11]. Among all families of HERVs, HERV-K is the only one known to be insertionally polymorphic in humans.

A full-length retroviral sequence is called a provirus and encodes several viral structural or regulatory proteins that are flanked by two long terminal repeats (5' or 3' LTR). While there are several HERV-K that are full length, none are infectious and most contain mutations or deletions that affect the open reading frames or truncate the virus. Further, the identical LTRs are substrates for homologous recombination, which deletes virus genes while retaining a single, or solo, LTR at the integration site [12–14]. Thus, in a population, a site could be unoccupied, occupied by a HERV-K provirus, or contain a solo LTR. Insertional polymorphism typically refers to the occupancy at a loci [15,16]. However the occupied site can contain a provirus or solo LTR and a provirus sequence can vary among individuals. Thus HERV-K and other HERVs can contribute to genomic diversity in the global human population in several ways [17].

HERVs from multiple families have been linked with both proliferative and degenerative diseases in humans [18–24] . Although there are known mechanisms by which a HERV can cause disease; for example, by inducing genome structural variation through recombination [25–29], affecting host gene expression [30], and inappropriate activation of an immune response by viral RNA or proteins [21], it has been difficult to establish an etiological role of a HERV in any disease. HERV-K specifically has been associated with breast and other cancers [3,31–35], and autoimmune diseases, such as rheumatoid arthritis [36,37], multiple sclerosis [20,38] and systemic lupus erythematosus [8,20,39] without definitive evidence of causality or of the specific loci involved. Recently, a HERV-K envelope protein was shown to recapitulate the

clinical and histological lesions characterizing Amyotropic Lateral Sclerosis [40,41], providing an important mechanistic advance of a role for a HERV-K protein in a disease.

In this paper, we focus on characterizing the genome landscape of known insertionally polymorphic HERV-K proviruses in the 1000 Genomes Project (KGP) data. We present a data-mining tool and a statistical framework that accommodates low depth data characteristic of the KGP - and often patient - data to estimate the presence or absence of a provirus at known HERV-K loci. Because combinations of HERV-K may act synergistically in the pathogenesis of a disease [42], we estimate the co-occurrence of polymorphic HERV-K proviruses in different populations and provide a tool to visualize HERV-K co-occurrence in global populations. Our results provide a reference of global population diversity in HERV-K proviruses at all currently known loci in the human genome and demonstrate that there are notable differences among population frequencies of HERV-Ks and the total number of HERV-Ks found in a person's genome.

Results

A model to estimate polymorphic HERV-K from whole genome sequence data.

The goal of this research was to develop a computationally efficient and easy to use tool that could accurately report the status of all HERV-Ks with coding potential (provirus) from whole genome sequence (WGS) data. We use the KGP database to establish the global population diversity of each polymorphic HERV-K and the burden of HERV-K in individual

genomes to provide a foundation to study the role of HERV-K in human disease. Our method takes as input all reads that map to identified HERV-K elements in hg19. The rationale here is that polymorphic HERV-K are very similar to those in the reference genome and will map on existing elements. The recovered reads are reduced to *k-mers* and mapped to a reference set of *k-mers* representing all unique sites in every HERV-K in the database. The output is a ratio of subject *k-mers* (n) that are 100% match to the reference *k-mers* (T) (see methods for full details).

Our preliminary analysis of the KGP data demonstrated that our *k-mer*-based approach is sensitive to sequence depth; some HERV-K are represented by an almost continuous range of n/T from 0-1 (Fig 1A), making presence/absence classification difficult. A comparison of a subset of the 28 individuals in the KGP data that have both low and high sequence depth data shows how depth affects n/T (Fig 1B, see S2 Fig for data of all 28 persons). If read depth is greater than 20, there is less dispersion of n/T values, most likely because more reads are recovered from the mapped intervals. However, the majority of the KGP data is approximately 6x depth and thus to make use of this important resource, we developed a mixture model to cluster the n/T values from genomes sequenced at low depth. K was optimized to 50 because this value improved our model computational efficiency and output (Fig 1B, S1-S3 Methods, S1 Fig). The states, 'provirus', 'solo LTR', and 'absent' are preliminarily assigned to each cluster based on the high depth data (Fig 1B). Individuals with n/T=1 have the reference allele and n/T=0 indicates that the HERV-K is absent (no *k-mer*s to unique sites in the HERV-K were recovered from mapped sequence reads). The *k-mer*s derived from persons with low and

intermediate n/T values were mapped to each HERV-K to determine whether they localized only in the LTR (assign 'solo LTR') or in the coding region (assign 'provirus') (S3 Fig).

Fig 1. A mixture model to account for low depth WGS data

A) The plot displays the n/T value for 2535 individuals from KGP with low depth sequence data for chr12:55727215-55728183 when K=70. There is little resolution of values to enable assignment to provirus, solo LTR, or absent states.

B) The result of the mixture model on the same data. The individual clusters generated by the model are indicated by a unique color; in the example shown, there are four clusters. K has been optimized to 50 to enable clear clustering in the 'absent', 'solo LTR' and 'provirus' states. In this example, eight of the 28 individuals that have both low and high depth sequence data (see S1_Dataset:KGP) are shown to demonstrate the effect of sequence depth. The n/T ratio is 1 for persons with high depth data [red numbers, #6 and 12] who have the reference allele, while the corresponding low depth data [black numbers, yellow cluster] from the same individuals have n/T ranging from 0.7 to 0.9. There is less of an effect of depth for individuals who do not have the HERV-K (n/T=0). However, optimizing K facilitates separation of clusters for absent [red cluster, #23 and 28], and solo LTR [green cluster, #4 and 16]. States are confirmed by mapping the *k-mer*s from individuals in a cluster to the reference HERV-K (S3 Fig).

Prevalence of polymorphic HERV-K in each KGP super-population

The WGS data of each individual in the KGP dataset were evaluated using our analysis workflow. HERV-Ks on chrY were not considered. Twenty sites, omitting one at chr1:73594980 [see methods] were identified that were polymorphic for containing a HERV-K provirus. A

phylogenetic analysis of all HERV-Ks greater than 6 kbp shows that polymorphic HERV-Ks are closely related (S4 Fig). The prevalence of the 20 polymorphic HERV-Ks varied from 0.9% to 99.5% when averaged across the entire KGP dataset (Table 1). However, there were notable differences in prevalence at each site among the five super-populations (AFR, EAS, AMR, EUR, SAS). Of the 20, the prevalence of seven polymorphic HERV-Ks was greater than 90% and the difference between populations with the lowest and highest prevalence was less than 6.5% (Table 1). There was 100% occupancy for six of the seven high prevalence polymorphic HERV-Ks (98.8% for the seventh), indicating that the rate of conversion to solo LTR is low for viruses at these sites (S1 Table). Two polymorphic HERV-Ks had an overall prevalence of less than 10% in any population (Table 1) and we found no evidence of a solo LTR at these sites; both are found in individuals from AFR. Nine of the remaining 11 HERV-Ks are of interest because the difference between super-populations with the highest and lowest prevalence is between 28 and 80 percentage points (Table 1). Of note, the prevalence is lowest in EAS populations for the three HERV-Ks with the largest difference among super-populations.

Individuals from African populations differ significantly from the other four super-populations in the prevalence of ten of the polymorphic HERV-K, three of which occur in close proximity on chr19. (Table 1, S2_Dataset:compare_prevalence). EUR and AFR super-populations are significantly different at all but one of the 20 polymorphic HERV-K based on adjusted p-values (S2_Dataset:compare_prevalence).

Table 1. Provirus frequencies of polymorphic HERV-K.

173

	KGP pro-virus (%)	AFR	AMR	EAS	EUR	SAS	max-min
chr1:75842771[c]	42.88	26.76	56.53	6.02	68.91	66.80	62.89
chr3:112743479[a]	98.46	96.71	99.72	99.81	99.60	97.37	3.09
chr3:148281477	41.89	38.86	42.61	45.05	46.53	37.45	9.09
chr3:185280336[a]	99.49	98.06	100.00	100.00	100.00	100.00	1.94
chr4:69463709[c]	72.50	93.87	88.92	31.07	85.35	61.94	62.80
chr5:156084717[a]	99.41	98.36	99.72	100.00	99.80	99.60	1.64
chr6:57623896[a]	93.65	90.73	97.16	90.87	97.23	94.33	6.50
chr6:78427019[a]	97.71	95.52	97.16	99.61	97.23	99.60	4.10
chr7:4622057*[c]	47.50	61.14	30.11	58.25	36.44	41.50	31.02
chr8:12316492[c]	14.08	32.88	12.22	0	15.64	3.04	32.88
chr8:7355397[c]	18.66	39.16	12.50	6.02	11.29	15.99	33.14
chr10:27182399[a]	99.13	97.46	99.43	99.81	99.80	99.80	2.35
chr11:101565794[c]	63.04	80.87	77.27	6.99	86.53	63.16	79.54

chr12:55727215	72.19	72.80	80.40	63.30	80.99	65.79	17.69
chr12:58721242[c]	70.73	58.89	78.41	60.00	87.33	75.51	28.43
chr19:21841536[c]	26.98	39.16	11.93	32.23	10.69	32.39	28.47
chr19:22414379[c]	67.77	89.24	60.80	56.89	55.84	67.21	33.40
chr19:22457244[b]	0.87	3.29	0.00	0.00	0.00	0.00	3.29
chr22:18926187[a]	99.49	98.36	99.72	100.00	99.80	100.00	1.64
chrX:93606603[b]	2.25	7.32	2.27	0.00	0.00	0.00	7.32

For simplicity, only the starting coordinate is listed.

* The value given represents those with the tandem repeat

[a]: prevalence > 90%

[b]: low prevalence and no solo LTR

[c]: max-min difference is > 28%

underline: AFR significantly different from other 4 super populations.

See S2_Dataset:compare_prevalence for full data set.

The number of polymorphic HERV-Ks per individual

The HERV-K genome is close to 10 kbp. As there are 20 HERV-Ks that are polymorphic in human populations, we asked if some individuals carry a different burden of these repetitive, and potentially functional, viral elements than others. This was indeed the case. The number of

polymorphic HERV-K proviruses per person ranges from 7-18 (Fig 2, S2_Dataset:HERV-K per person). More than 63% of individuals from all super-populations except EAS carry 12 to 14 proviruses in their genome. Individuals from EAS have a lower burden with 69% of individuals carrying 9-11 HERV-K proviruses. 7% of AFR individuals have 16 or 17 proviruses compared to a maximum of 2% in other groups (S2_Dataset:HERV-K per person). These data highlight the importance of using a comprehensive approach to study the potential disease impact of polymorphic HERV-Ks, because total HERV-K burden in individuals might influence a disease phenotype without significant variation in prevalence of each provirus in a patient pool.

Fig 2. Histogram of the number of proviruses per individual among super-populations.

Co-occurrence of polymorphic HERV-Ks

Our data provide a comprehensive picture of sites occupied by HERV-K provirus in each genome, which enables analysis of polymorphic HERV-K co-occurrence in populations. We assessed combinations of three, four and five polymorphic HERV-Ks and found that there are many combinations of co-occurring viruses that are population-specific (S3_Dataset). To facilitate exploration of HERV-K combinations among KGP populations, we developed a D3.j visualization tool that allows a user to choose any combination of the 20 polymorphic HERV-K proviruses and display the co-occurrence prevalence among the 26 populations represented in the KGP data. As an example, we present a combination of four HERV-Ks to represent the variation that occurs in KGP individuals, which in this case ranges from 3% in EAS to 59% in EUR

(Fig 3A). We also determine that the three polymorphic HERV-Ks found on chr19 co-occur only from three AFR populations and in less than 2% of individuals (Fig 3B).

Fig 3. A visualization tool to examine co-occurrence of polymorphic HERV-Ks.

A) The co-occurrence of HERV-Ks at chr1:75842771-75849143, chr3:112743479-112752282, chr6:57623896-57628704, and chr12:58721242-58730698 in the 26 populations are represented based on their geographic location. The relative frequency for these four co-occurring HERV-Ks in each population bubble is displayed based on the color gradient shown in the scale at the top. The actual prevalence of the given combination of HERV-K provirus for each population and the cumulative prevalence for the super-population are shown in text on the right. Note that AFR and EAS have the lowest prevalence of these four polymorphic HERV-Ks.

B) As in (A) showing the co-occurrence of the three polymorphic HERV-Ks that are present on chr19 by population. This is a rare combination only found in two AFR populations and individuals in the Caribbean of African ancestry.

HERV-K status informs KGP super-populations

Because there are clearly population-specific differences in both individual HERV-K frequency and in the frequency of HERV-K co-occurrence, we explored whether the presence or absence of polymorphic HERV-Ks is sufficient to distinguish populations using Fisher's linear discriminant analysis (LDA) [43]. Based on the status 'provirus', 'solo LTR', or 'absence', there is little resolution of AFR, EUR, and EAS super-populations (Fig 4A). However, there is sufficient signature to separate AFR, EUR, and EAS if we utilize the n/T ratio on the 20 polymorphic HERV-

Ks (S5 Fig) and we further improve population separation if we use the n/T ratio for all 96 HERV-Ks (Fig 4B). This indicates that we are losing information by reducing the data to three states and that fixed HERV-K also contain signal for population of origin.

Fig 4. Linear discriminant analysis of HERV-K status among three super-populations.

A) LDA based on the states 'provirus', 'solo LTR' and 'absence' of the 20 polymorphic HERV-K for AFR, EAS, and EUR. AMR and SAS overlap these three populations and are removed for clarity B) LDA plot on n/T ratio of all 96 HERV-K discriminates AFR, EAS, and EUR super-populations. See S6 Fig for plots with all five super-populations.

An n/T = 1 indicates that we recovered all reads that map to the reference set T for a specific HERV-K. If there is a HERV-K allele that has not been reported in any database but that is common in a population, we expect n/T <1 because we require 100% match to reference set T and *k-mers* covering allelic sites will not be included. We assessed the density distributions of n/T plots for each of the 96 HERV-Ks for evidence of population-specific alleles (S4 Method, S7 Fig). Five HERV-Ks have some indication of population specific distributions (S1_Dataset:virus). The HERV-K at chr1:155596457-155605636, which we report as fixed, is notable because the reference allele (n/T=1) is only found in AFR (Fig 5A, S7 Fig). Individuals from most other populations have n/T near 0.5. We mapped *k-mer*s from individuals with n/T near 0.5 to the reference HERV-K sequence and confirmed that there is a loss of *k-mer*s at several sites covered by the unique reference *k-mer*s for this virus (S8 Fig). There are also cases where the reference allele is found in all populations except AFR (Fig 5B and see S7 Fig for additional examples).

Fig 5. Population specificity of HERV-K alleles. A) n/T plot for chr1:155596457-155605636 colored by each of the 5 super-populations. Only individuals from AFR and a few from AMR have the HERV-K reference sequence. B) Plot of chr5:156084717-156093896 colored by each of the 5 super-populations. In this case, all populations except AFR have the reference allele and all super-populations have an alternative allele that is not present in the databases.

Discussion

Our research provides a tool to mine whole genome sequence data to collectively evaluate the status of HERV-K provirus at known polymorphic and fixed sites in the human genome. The tool incorporates a statistical clustering algorithm to accommodate low depth sequence data and a visualization tool to explore the co-occurrence of polymorphic HERV-K in the global populations represented in the KGP data. There are numerous significant differences in the prevalence of individual and co-occurring polymorphic HERV-K among the five KGP super-populations. It is notable that individuals from EAS carry a lower total burden of HERV-K than other represented populations. These data provide a comprehensive framework of HERV-K genomic diversity to advance studies on potential roles for HERV-K in human disease, which have been alluring yet difficult to establish [19,20,22].

Tools developed to interrogate ERV insertional polymorphism typically exploit the unique signature created by the host-virus junction [11,44,45]. These approaches indicate that a site is occupied by an ERV but not whether there is a provirus associated with the site, which is more difficult to accomplish with short read sequence data. Our analysis tool provides an

277 efficient means to detect occupancy and provirus status in one step. We decrease
278 computational time by analyzing only the set of reads that map to existing HERV-K in the
279 reference genome. This approach is justified because the polymorphic HERV-K that are missing
280 from the human reference are closely related to those in the reference genome assembly and
281 hence reads derived from them map to a related HERV-K in the reference. We employ *k-mer*
282 counting methods, which also increase computational efficiency. A reference set of *k-mer*s that
283 is unique to a HERV-K is generated for each location in the genome and the proportion of reads
284 from the query set that maps to the *k-mer* reference set is reported as a continuous variable;
285 there is no threshold of read count or depth imposed for classification. Instead we utilize a
286 mixture model to cluster values and assign the same HERV-K status to the entire cluster.
287 Clusters with n/T of 1 have all the unique *k-mer*s identified in the HERV-K reference set. We
288 classify other clusters by determining if *k-mer*s mapped on the reference allele are distributed
289 at sites in the coding portion of the genome or only in the LTR. This approach led to the
290 interesting finding that several HERV-K could have population specific alleles.
291
292 Wildschutte *et al* [11] have conducted the most comprehensive study of HERV-K
293 prevalence in the KGP data to date. While the goal of that paper was to identify new
294 polymorphic insertions of either provirus or solo LTR in the KGP data, their analyses provide the
295 prevalence of some polymorphic HERV-K provirus for comparison with our results. There are
296 five HERV-K previously reported in Subramanian *et al* 2011 [10] that were not included in the
297 Wildschutte paper [11]; all are polymorphic in our analysis (range 43-99%, see Table 1 and
298 S1_Dataset:virus-column N). Seven polymorphic HERV-K, which Wildschutte *et al* [11] indicate

299 occur in greater than 98% of KGP individuals, are fixed in our study. Our estimated prevalence
300 for 14 HERV-K differs from that reported in Wildschutte et al [11] by 5% or more. Of these 14,
301 the prevalence estimates at chr1:155596457-155605636 are most divergent. Our data show
302 this site is fixed for provirus and Wildschutte *et al* [11] report that only 14% of the KGP data, all
303 from AFR, have a HERV-K provirus integration. Our plots for chr1:155596457-155605636 show
304 that AFR individuals carry the reference allele at this site (n/T near 1, Fig 5A) and all other
305 individuals have n/T near 0.5. The *k-mer*s from individuals with low n/T values for
306 chr1:155596457-155605636 map to only a subset of sites marked by unique *k-mer*s in the
307 coding region (S8 Fig), which is consistent with sequence polymorphism or a deletion at these
308 positions. The reference set T is small for this HERV-K and therefore overall coverage of the
309 genome is low. Because Wildschutte *et al* [11] used a minimum coverage threshold for their *k-*
310 *mer* mapping method, it is possible that alleles present in non-AFR populations would be
311 outside their inclusion criteria. There is a similar signal for alleles, represented by lower n/T
312 values, at the other 13 HERV-K sites although the differences between our prevalence
313 estimates and those of Wildschutte et al [11] are small (S1_Dataset:virus). In most cases these
314 putative alleles are found in all populations at different frequencies but in five there is some
315 degree of population specificity (Fig 5, S7 Fig, S1_Dataset:virus). Our results indicate that there
316 could be considerably more sequence variation in HERV-K among human populations than
317 previously appreciated. These data also suggest that using HERV-K consensus sequences to
318 study pathogenic potential could miss important features of HERV-K polymorphism, which can
319 be characterized by both the site occupancy status (presence/absence) and, when present, by
320 sequence differences in among individuals.

HERV-Ks are the youngest family of endogenous retroviruses in humans and consequently they share considerable sequence identity. This has the effect of limiting the number of unique sites associated with some HERV-K, which decreases the size of the reference set T (S1_Dataset:virus). Another example of a polymorphic HERV-K with a small set T is chr8:12316492, reported to be human specific (9), which shares a recent common ancestor with two older HERV-K (chr8:12073970 and chr8:8054700) all located at 8p23.1. Our data indicate that 14% of KGP individuals have the reference allele and most n/T values are less than 0.4 and fall into two non-zero clusters (S9 Fig). These appear to represent various structural variations (truncation or deletion) because there are several peaks in both the LTR and in the 5' coding region. Thus although an n/T ratio of 0 or 1 reliably indicates absence and presence of the reference HERV-K, respectively, when T is small, sequence polymorphism and a deletion event can be difficult to distinguish from a solo LTR. However, because our mixture model statistically clusters similar n/T values, all individuals in a cluster have the same status (e.g allele or solo LTR) even if we do not know what that state is.

Our approach provides human disease researchers with a rapid means to determine if the frequency, and overall burden of the 96 HERV-K proviruses evaluated differ between a patient data set and populations represented in KGP. The visualization tool allows investigators to determine if HERV-Ks co-occur in certain clinical settings. The potential that HERV-K has multiple allelic forms in different populations is worthy of further analysis because a sequence allele could also contribute to a disease condition.

Materials and methods

HERV-K proviruses

The 96 HERV-K proviruses previously reported [10,11,32,46] were supplemented with HERV-K alleles present in the NCBI nt database (November 2016 release). We required that any allele of a HERV-K from the nt database have at least 2kb of reference-matching host flanking sequence to confirm genome location. In total, 234 alleles were collected at the 96 known HERV-K loci (92 in hg19, and 4 from the nt database). The location information and virus features are summarized in S1_Dataset: virus.

Developing a *k-mer* based detection model

We identified the *k-mer*s that correspond to unique sequence characterizing each HERV-K. *K-mer*s are substrings (subsequences) of length *k* that exist in a string (DNA sequence). The length *k* is determined empirically (S1 Method). Each *k-mer* is labeled with the corresponding viruses in which it is observed.

Only those *k-mer*s referring to a single virus, unique *k-mer*s, are selected for the set T. Where multiple alleles of a HERV-K are available, *k-mer*s unique to all alleles at that location comprise T. Multiple 2bps different *k-mer*s (such as SNPs) corresponding to the same location on the virus, are merged into a single entry for the purposes of computing T. We map unique *k-mer*s back to the corresponding alleles to determine depth of the HERV-K (S3 Fig) and whether *k-mer*s are located in LTRs. (S1_Dataset: virus)

365

366

367 **Analysis of 1000 Genome Project (KGP) Data**

368 To develop a method to recover sequences containing information on HERV-K we
369 leverage the fact that HERV-Ks are closely related. Thus, most sequence reads obtained from an
370 individual with a polymorphic HERV-K that is absent in the human reference, hg19, will map to
371 the location of one of the closely related HERV-K that is present the human genome reference.
372 A file with the coordinates for all reported HERV-K insertions is used to extract mapped reads
373 from a genome sequence file (S1_Dataset:bed, which provides the coordinates for both hg19
374 and hg38). Note that the KGP data were mapped to GRCh37, which includes the decoy
375 sequence hs37d5. This decoy contains the HERV-K at chr1:73594980_73595948 and is not
376 present in hg19. Thus, we did not recover any reads for this HERV-K, which is polymorphic but
377 reportedly at high prevalence in most populations [11].

378

379 The KGP data were downloaded in aligned Binary Alignment/Map (BAM) format
380 (ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/data/). It contains data for 2,535 individuals
381 (S1_Dataset:KGP) sequenced via low-depth whole-genome sequencing (mean depth = 6.98X).
382 The individuals represent 26 populations, derived from 5 super-populations, including African
383 (AFR), Admixed America (AMR), East Asian (EAS), European (EUR), and South Asian (SAS)
384 [47,48]. Of 2,535 individuals, 28 also have high-depth DNA sequences (mean depth = 48.06X),
385 which we use as a pilot dataset for our clustering methods, described below and in
386 Supplementary Methods.

Our computational framework to indicate the status of each known HERV-K provirus is based on the *n/T* ratio, which is the proportion of *k-mer*s from each individual that are identical to the reference set T for each HERV-K. Sequence reads are extracted from a mapped file of whole human genome sequence data based on coordinates corresponding to each annotated HERV-K. The reads are *k-meri*zed and mapped to the set T, which represents all unique *k-mer*s assigned to each HERV-K in the reference set. We use exact match to map *k-mer*s data set to the unique *k-mer*s references. The *n/T* ratio is used as an indicator of the presence of each HERV-K. Using a hash table (S5 Method), it takes 15 minutes to generate the n/T matrix for 100 files. The source code for the entire process is at https://github.com/lwl1112/polymorphicHERV

Dirichlet process Gaussian mixture model (DPGMM)

Because n (the number of *k-mer*s obtained from a persons' sequence data, that map to a specific HERV-K) is affected by sequencing depth, we utilized a statistical model to cluster n/T for each HERV-K for each individual and assigned HERV-K status to a cluster. Mixture modeling is arguably the most widely used statistical method for clustering. In this analysis, we follow the work proposed by Lin et al. 2015 [49], which employs a Gaussian Mixture Model (GMM) with density function given by

$$f(x \mid \theta) = \sum_{j=1}^{M} \pi_j \, N(\mu_j, \Sigma_j) \,, \qquad (1)$$

and with a mixture of relatively large number M Gaussian components (denoted by $N(\mu_j, \Sigma_j)$, for j = 1:M) to represent the data density. To allow a flexible modeling approach, we employ the standard Bayesian (truncated) Dirichlet Process prior for the parameters $\theta = (\pi_j, \mu_j, \Sigma_j, j = 1{:}M$ [50,51]. The idea is that some of the mixture probabilities (π_j) can be zero, hence the actual number of mixture components needed may be smaller than the upper bound M. This mechanism allows automatic determination of the number of mixture components needed by the data set at hand. Given a fitted model via the Bayesian expectation–maximization algorithm, in terms of estimates of all parameters θ, we identify clusters by aggregating Gaussian components. Merging components into clusters can be done by associating each of the Gaussian components to the closest mode of $f(x|\theta)$. Hence, the number of modes identified is the realized number of clusters. [See S2 Method for full detail]

Co-occurrence of polymorphic HERV-K

We consider that both the individual frequency of a HERV-K and the co-occurrence of multiple HERV-K could differ among populations.

The time of a brute-force approach for finding all combinations C_m of size m from p polymorphic HERV-K is $(\sum_{m=1}^{p} \binom{p}{m} = 2^p - 1)$, which is not efficient and redundant. We employed the Apriori algorithm [52], which is commonly used for finding frequent pattern sets; in our case indicating which polymorphic HERV-K frequently appear together. It first generates combinations C_m (initialized to 1). In the optimization, frequent combinations F_m are returned from candidates C_m when frequency exceeds the minimum threshold of co-occurrence. F_m are

then self-joined to generate combinations C_{m+1} of size *m* +1 and out of which F_{m+1} satisfy the minimum co-occurrence. In each pass, candidate combinations are pruned so as to avoid generating all combinations, which reduces running time significantly.

Statistical analysis of HERV-K frequencies across populations

We make statistical comparisons across 5 super-populations for the following three problems. For each problem, there are $\binom{5}{2}$ = 10 families of 1-to-1 comparisons conducted. The 'prop-test' function in R is used to test whether the proportions for two super-populations are the same.

1) individual prevalence of polymorphic HERV-K. (20 comparisons for each polymorphic HERV-K in a family)

2) the number of polymorphic HERV-K present per individual. (21 comparisons as the number of co-occurring polymorphic HERV-K is from 0 to 20)

3) the co-occurrence for combinations of polymorphic HERV-K.

Therefore, multiple hypotheses would be conducted on frequencies *F* across super-populations $P_{1...5}$ as follows:

Null hypothesis, H_0: $F_{P_i} = F_{P_j}$, where i ≠ j;

Alternative hypothesis, H_A: $F_{P_i} \neq F_{P_j}$, where i ≠ j.

A separate P-value is computed for each test and the Benjamini-Hochberg procedure [53] is used to account for multiple comparisons.

Visualization in D3.js

We utilized D3.js (Data Driven Documents) [54], an open-source java script library to create an interactive visualization to display co-occurrence of polymorphic HERV-Ks in human populations. Our visualization system includes two modules, a welcome page and a result page. Input JSON data include locations of polymorphic HERV-K, population information, and the 0/1 (absence / presence) matrix. (See S3 Method). Source code is available at:

https://github.com/lwl1112/polymorphicHERV/tree/master/visualization

Acknowledgements

WL was supported in part by the Louis S. and Sara S. Michael Endowed Graduate Fellowship in Engineering and the Fred A. and Susan Breidenbach Graduate Fellowship in Engineering.

References

1. Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc Natl Acad Sci. 2013;110: 20146–20151. doi:10.1073/pnas.1315419110

2. Feschotte C, Gilbert C. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet. 2012;13: 283–296. doi:10.1038/nrg3199

3. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat Rev

Microbiol. 2012;10: 395–406. doi:10.1038/nrmicro2783

4. Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. Virus Genes. 2003;26: 291–315. Available: http://www.ncbi.nlm.nih.gov/pubmed/12876457

5. Weiss RA. The discovery of endogenous retroviruses. Retrovirology. 2006;3: 67. doi:10.1186/1742-4690-3-67

6. Jern P, Coffin JM. Effects of Retroviruses on Host Genome Function. Annu Rev Genet. 2008;42: 709–732. doi:10.1146/annurev.genet.42.110807.091501

7. Löwer R, Löwer J, Kurth R. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A. 1996;93: 5177–84. Available: http://www.ncbi.nlm.nih.gov/pubmed/8643549

8. Bannert N, Kurth R. Retroelements and the human genome: New perspectives on an old relation. Proc Natl Acad Sci. 2004;101: 14572–14579. doi:10.1073/pnas.0404838101

9. Moyes D, Griffiths DJ, Venables PJ. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet. 2007;23: 326–333. doi:10.1016/j.tig.2007.05.004

10. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011;8: 90. doi:10.1186/1742-4690-8-90

11. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci. 2016;113: E2326–E2334. doi:10.1073/pnas.1602336113

12. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, et al. Rate of Recombinational Deletion among Human Endogenous Retroviruses. J Virol. 2007;81: 9437–9442. doi:10.1128/JVI.02216-06

13. Medstrand P, Mager DL. Human-specific integrations of the HERV-K endogenous retrovirus

family. J Virol. 1998;72: 9782–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/9811713

14. Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: Implications for human and viral evolution. Proc Natl Acad Sci. 2004;101: 1668–1672. doi:10.1073/pnas.0307885100

15. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. J Virol. 2005;79: 12507–12514. doi:10.1128/JVI.79.19.12507-12514.2005

16. Marchi E, Kanapin A, Magiorkinis G, Belshaw R. Unfixed Endogenous Retroviral Insertions in the Human Population. J Virol. 2014;88: 9529–9537. doi:10.1128/JVI.00919-14

17. Shin W, Lee J, Son S-Y, Ahn K, Kim H-S, Han K. Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome. Cordaux R, editor. PLoS One. 2013;8: e60605. doi:10.1371/journal.pone.0060605

18. Gröger V, Cynis H. Human Endogenous Retroviruses and Their Putative Role in the Development of Autoimmune Disorders Such as Multiple Sclerosis. Front Microbiol. 2018;9. doi:10.3389/fmicb.2018.00265

19. Young GR, Stoye JP, Kassiotis G. Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. BioEssays. 2013;35: 794–803. doi:10.1002/bies.201300049

20. Ryan FP. Human endogenous retroviruses in health and disease: a symbiotic perspective. JRSM. 2004;97: 560–565. doi:10.1258/jrsm.97.12.560

21. Volkman HE, Stetson DB. The enemy within: endogenous retroelements and autoimmune disease. Nat Immunol. 2014;15: 415–422. doi:10.1038/ni.2872

22. Magiorkinis G, Belshaw R, Katzourakis A. “There and back again”: revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. Philos Trans

R Soc B Biol Sci. 2013;368: 20120504–20120504. doi:10.1098/rstb.2012.0504

23. Löwer R. The pathogenic potential of endogenous retroviruses: facts and fantasies. Trends Microbiol. 1999;7: 350–356. doi:10.1016/S0966-842X(99)01565-6

24. Hohn O, Hanke K, Bannert N. HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. Front Oncol. 2013;3. doi:10.3389/fonc.2013.00246

25. Hughes JF. Human Endogenous Retroviral Elements as Indicators of Ectopic Recombination Events in the Primate Genome. Genetics. 2005;171: 1183–1194. doi:10.1534/genetics.105.043976

26. Hughes JF, Coffin JM. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat Genet. 2001;29: 487–489. doi:10.1038/ng775

27. Romanish MT, Cohen CJ, Mager DL. Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer. Semin Cancer Biol. 2010;20: 246–253. doi:10.1016/j.semcancer.2010.05.005

28. Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum Mol Genet. 2000;9: 2563–72. Available: http://www.ncbi.nlm.nih.gov/pubmed/11030762

29. Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. Cell. 2010;143: 837–847. doi:10.1016/j.cell.2010.10.027

30. Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. Gene. 2009;448: 105–114. doi:10.1016/j.gene.2009.06.020

31. Simmons W. The Role of Human Endogenous Retroviruses (HERV-K) in the Pathogenesis of

544 Human Cancers. Mol Biol. 2016;05. doi:10.4172/2168-9547.1000169

545 32. Wildschutte JH, Ram D, Subramanian R, Stevens VL, Coffin JM. The distribution of insertionally
546 polymorphic endogenous retroviruses in breast cancer patients and cancer-free controls.
547 Retrovirology. 2014;11: 62. doi:10.1186/s12977-014-0062-3

548 33. Kassiotis G, Stoye JP. Making a virtue of necessity: the pleiotropic role of human endogenous
549 retroviruses in cancer. Philos Trans R Soc B Biol Sci. 2017;372: 20160277.
550 doi:10.1098/rstb.2016.0277

551 34. Johanning GL, Malouf GG, Zheng X, Esteva FJ, Weinstein JN, Wang-Johanning F, et al.
552 Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast
553 cancer phenotype. Sci Rep. 2017;7: 41960. doi:10.1038/srep41960

554 35. Bhardwaj N, Coffin JM. Endogenous Retroviruses and Human Cancer: Is There Anything to the
555 Rumors? Cell Host Microbe. 2014;15: 255–259. doi:10.1016/j.chom.2014.02.013

556 36. Hanke K, Hohn O, Bannert N. HERV-K(HML-2), a seemingly silent subtenant - but still waters run
557 deep. APMIS. 2016;124: 67–87. doi:10.1111/apm.12475

558 37. Trela M, Nelson PN, Rylance PB. The role of molecular mimicry and other factors in the
559 association of Human Endogenous Retroviruses and autoimmunity. APMIS. 2016;124: 88–104.
560 doi:10.1111/apm.12487

561 38. Antony JM, DesLauriers AM, Bhat RK, Ellestad KK, Power C. Human endogenous retroviruses
562 and multiple sclerosis: Innocent bystanders or disease determinants? Biochim Biophys Acta - Mol
563 Basis Dis. 2011;1812: 162–176. doi:10.1016/j.bbadis.2010.07.016

564 39. Tugnet N. Human Endogenous Retroviruses (HERVs) and Autoimmune Rheumatic Disease: Is
565 There a Link? Open For Sci J. 2013;7: 13–21. doi:10.2174/1874312901307010013

566 40. Li W, Lee M-H, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human endogenous retrovirus-
567 K contributes to motor neuron disease. Sci Transl Med. 2015;7: 307ra153-307ra153.

568 doi:10.1126/scitranslmed.aac8201

569 41. Douville RN, Nath A. Human Endogenous Retrovirus-K and TDP-43 Expression Bridges ALS and
570 HIV Neuropathology. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.01986

571 42. Nexø BA, Villesen P, Nissen KK, Lindegaard HM, Rossing P, Petersen T, et al. Are human
572 endogenous retroviruses triggers of autoimmune diseases? Unveiling associations of three
573 diseases and viral loci. Immunol Res. 2016;64: 55–63. doi:10.1007/s12026-015-8671-z

574 43. Fukunaga K. Introduction to statistical pattern recognition. Academic press; 1990.

575 44. Ciuffi A, Ronen K, Brady T, Malani N, Wang G, Berry CC, et al. Methods for integration site
576 distribution analyses in animal cell genomes. Methods. 2009;47: 261–268.
577 doi:10.1016/j.ymeth.2008.10.028

578 45. Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. Mobile element scanning
579 (ME-Scan) by targeted high-throughput sequencing. BMC Genomics. 2010;11: 410.
580 doi:10.1186/1471-2164-11-410

581 46. Bhardwaj N, Montesion M, Roy F, Coffin J. Differential Expression of HERV-K (HML-2) Proviruses
582 in Cells and Virions of the Teratocarcinoma Cell Line Tera-1. Viruses. 2015;7: 939–968.
583 doi:10.3390/v7030939

584 47. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated
585 map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81.
586 doi:10.1038/nature15394

587 48. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference
588 for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

589 49. Lin L, Chan C, West M. Discriminative variable subsets in Bayesian classification with mixture
590 models, with application in flow cytometry studies. Biostatistics. 2015; kxv021.
591 doi:10.1093/biostatistics/kxv021

592 50. Escobar MD, West M. Bayesian Density Estimation and Inference Using Mixtures. J Am Stat
593 Assoc. 1995;90: 577. doi:10.2307/2291069

594 51. Ishwaran H, James LF. Gibbs Sampling Methods for Stick-Breaking Priors. J Am Stat Assoc.
595 2001;96: 161–173. doi:10.1198/016214501750332758

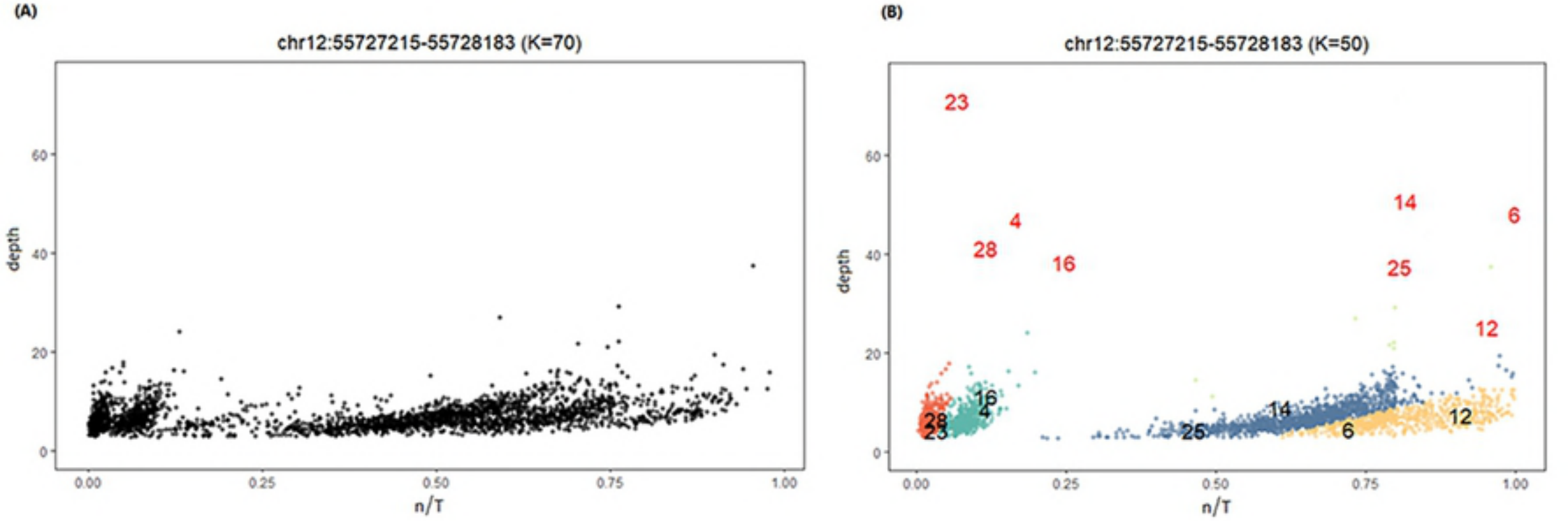
596 52. Huang L, Chen H, Wang X, Chen G. A fast algorithm for mining association rules. J Comput Sci
597 Technol. 2000;15: 619–624. doi:10.1007/BF02948845

598 53. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach
599 to multiple testing. J R Stat Soc Ser B. 1995; 289–300. doi:10.2307/2346101

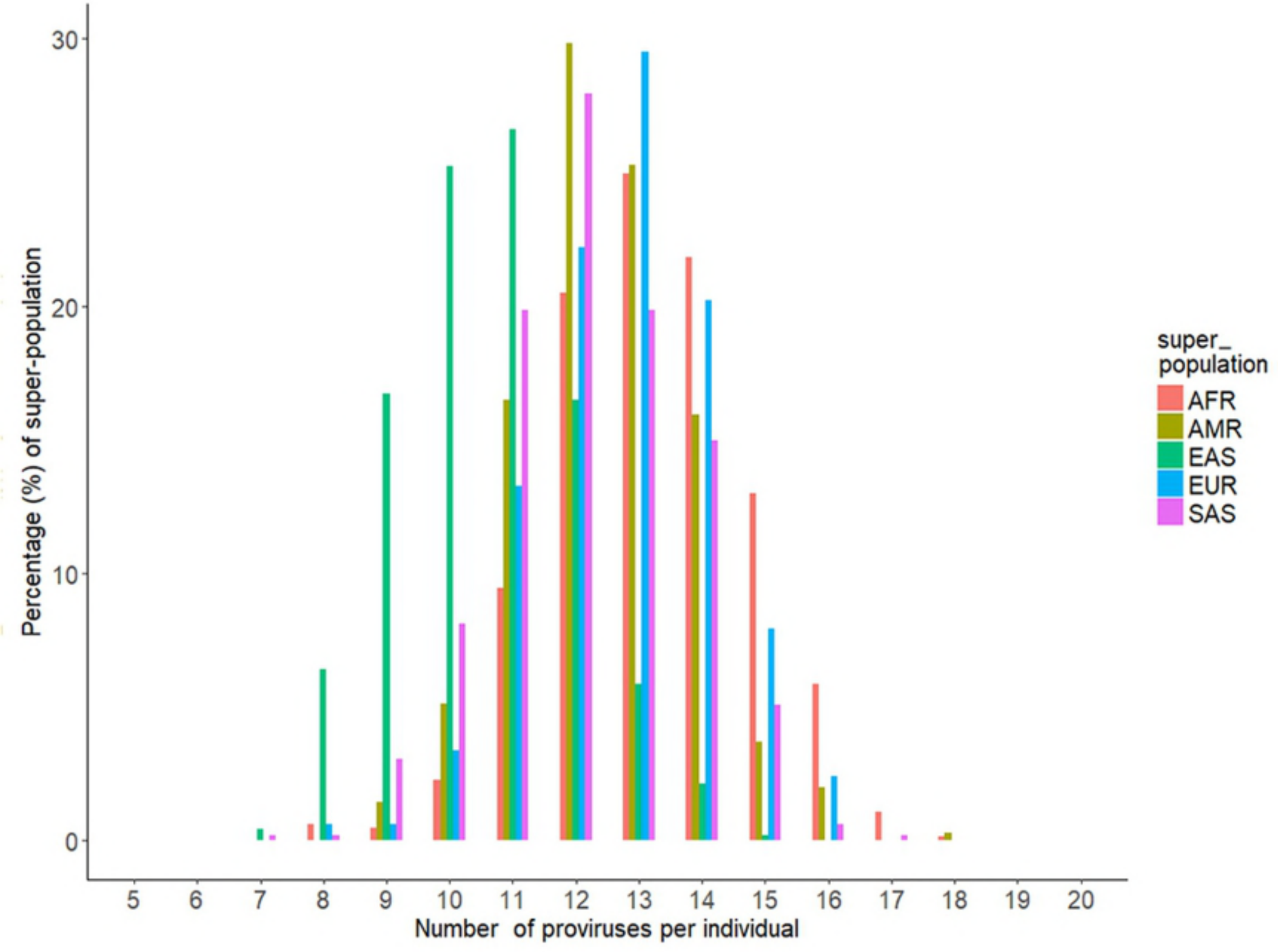
600 54. Bostock M, Ogievetsky V, Heer J. D^3 Data-Driven Documents. IEEE Trans Vis Comput Graph.
601 2011;17: 2301–2309. doi:10.1109/TVCG.2011.185

602

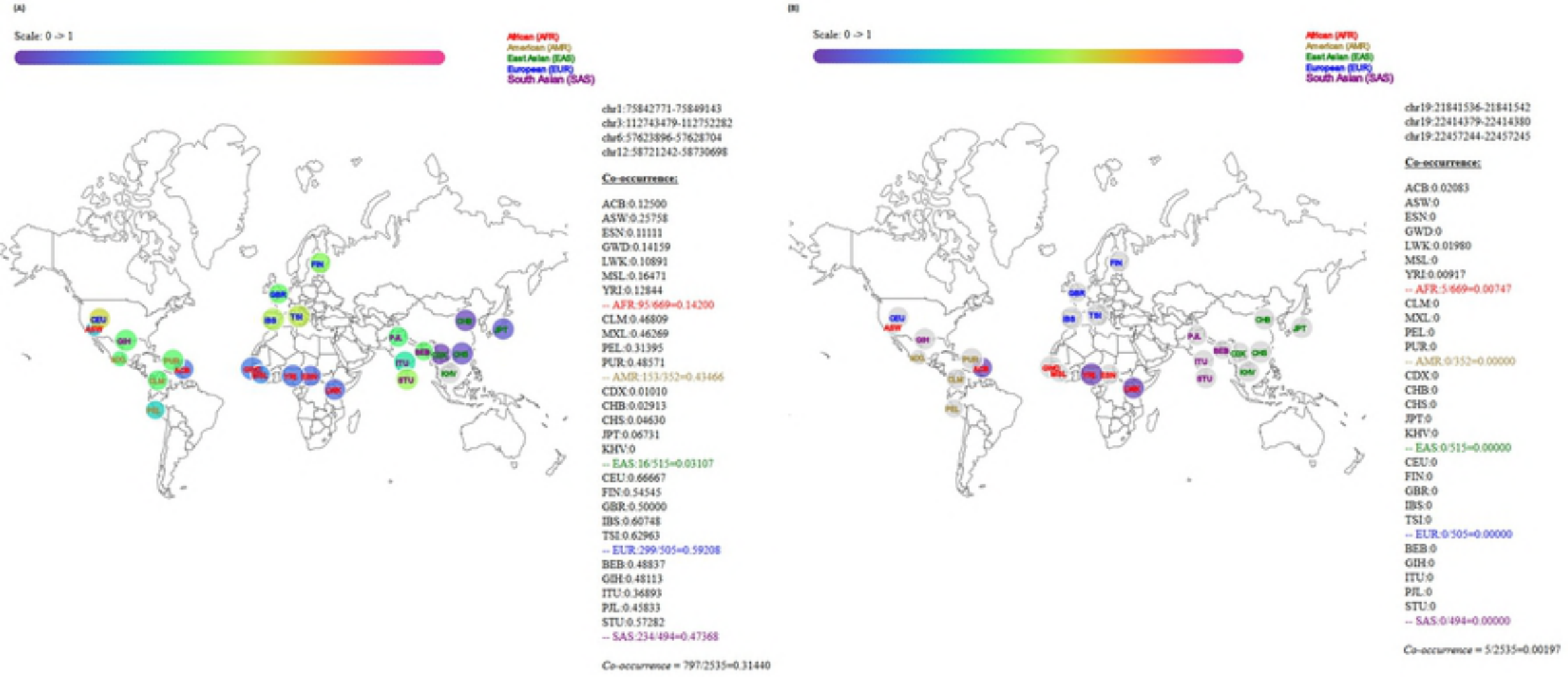
603


(A)
chr12:55727215-55728183 (K=70)
depth
n/T
(B)
chr12:55727215-55728183 (K=50)
depth
n/T
23
4
28
16
14
6
25
12

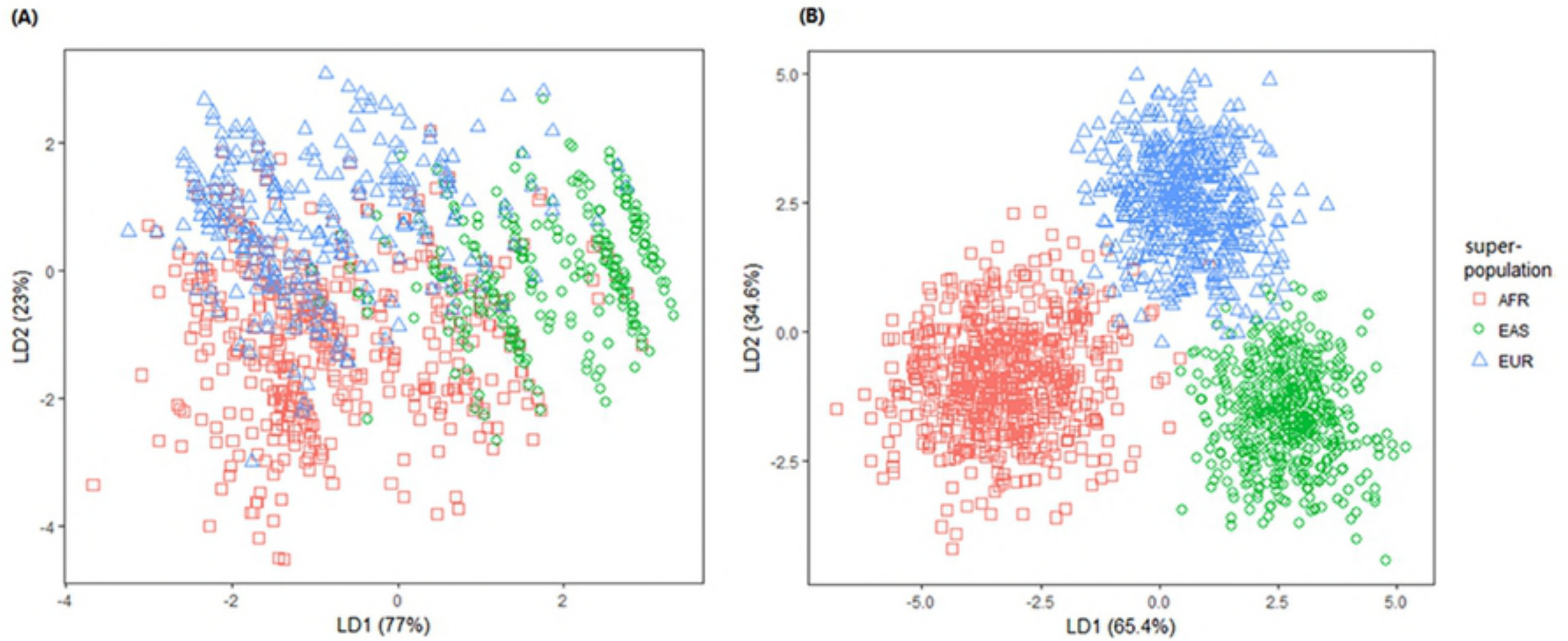
Figure


30
20
10
0
Percentage (%) of super-population
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Number of proviruses per individual
super_
population
AFR
AMR
EAS
EUR
SAS

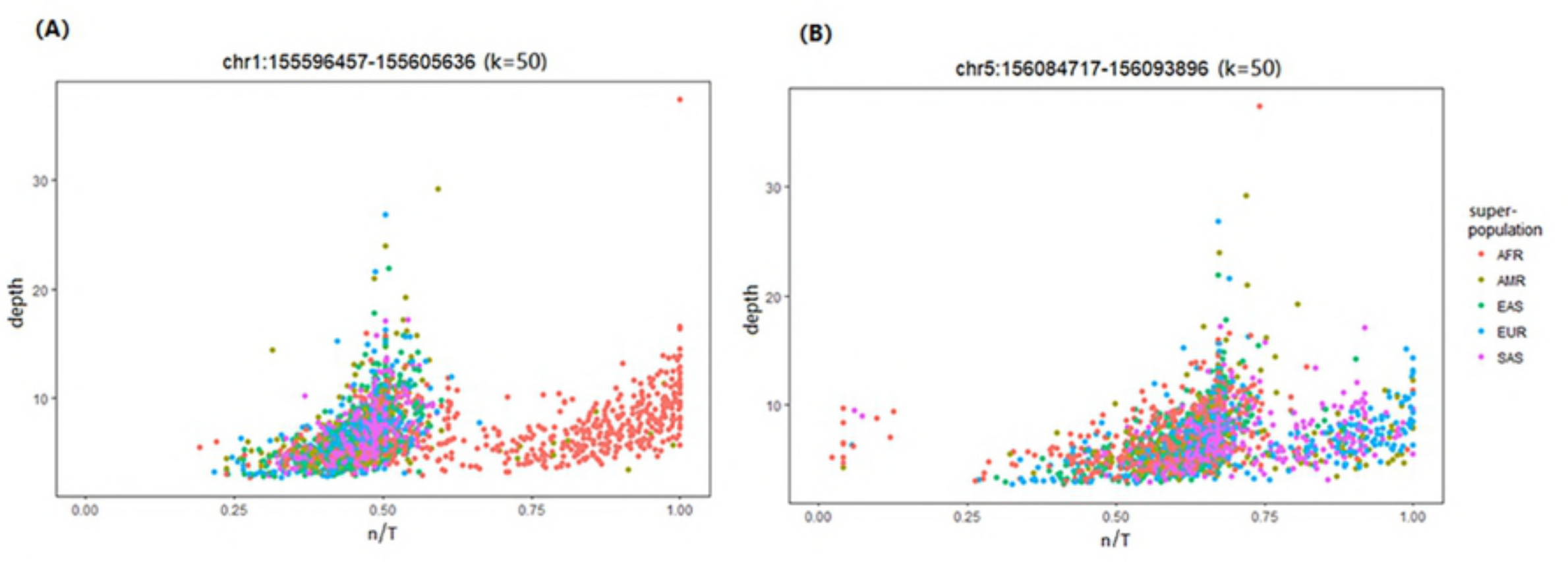
Figure


(A)
Scale: 0 -> 1
African (AFR)
American (AMR)
East Asian (EAS)
European (EUR)
South Asian (SAS)
chr1:75842771-75849143
chr3:112743479-112752282
chr6:57623896-57628704
chr12:58721242-58730698
Co-occurrence:
ACB:0.12500
ASW:0.25758
ESN:0.11111
GWD:0.14159
LWK:0.10891
MSL:0.16471
YRI:0.12844
-- AFR:95/669=0.14200
CLM:0.46809
MXL:0.46269
PEL:0.31395
PUR:0.48571
-- AMR:153/352=0.43466
CDX:0.01010
CHB:0.02913
CHS:0.04630
JPT:0.06731
KHV:0
-- EAS:16/515=0.03107
CEU:0.66667
FIN:0.54545
GBR:0.50000
IBS:0.60748
TSI:0.62963
-- EUR:299/505=0.59208
BEB:0.48837
GIH:0.48113
ITU:0.36893
PJL:0.45833
STU:0.57282
-- SAS:234/494=0.47368
Co-occurrence = 797/2535=0.31440
(B)
Scale: 0 -> 1
African (AFR)
American (AMR)
East Asian (EAS)
European (EUR)
South Asian (SAS)
chr19:21841536-21841542
chr19:22414379-22414380
chr19:22457244-22457245
Co-occurrence:
ACB:0.02083
ASW:0
ESN:0
GWD:0
LWK:0.01980
MSL:0
YRI:0.00917
-- AFR:5/669=0.00747
CLM:0
MXL:0
PEL:0
PUR:0
-- AMR:0/352=0.00000
CDX:0
CHB:0
CHS:0
JPT:0
KHV:0
-- EAS:0/515=0.00000
CEU:0
FIN:0
GBR:0
IBS:0
TSI:0
-- EUR:0/505=0.00000
BEB:0
GIH:0
ITU:0
PJL:0
STU:0
-- SAS:0/494=0.00000
Co-occurrence = 5/2535=0.00197

Figure



Figure


(A)
chr1:155596457-155605636 (k=50)
depth
n/T
0.00
0.25
0.50
0.75
1.00
10
20
30
(B)
chr5:156084717-156093896 (k=50)
depth
n/T
super-
population
AFR
AMR
EAS
EUR
SAS

Figure