BioTech

MDPI

Article

Retrotransposons and the Evolution of Genome Size in *Pisum*

T. H. Noel Ellis [1,*] and Alexander V. Vershinin [2,3]

1 John Innes Centre, Norwich Research Park, Colney Lane, Norwich NR4 7UH, UK
2 Institute of Molecular and Cellular Biology, Acad. Lavrentiev Ave. 8/2, 630090 Novosibirsk, Russia; avershin@mcb.nsc.ru
3 Department of Natural Sciences, Novosibirsk State University, Pirogova 2, 630090 Novosibirsk, Russia
* Correspondence: Noel.Ellis2@jic.ac.uk; Tel.: +44-1603-450-000

Received: 25 September 2020; Accepted: 16 November 2020; Published: 26 November 2020

Abstract: Here we investigate the plant population genetics of retrotransposon insertion sites in pea to find out whether genetic drift and the neutral theory of molecular evolution can account for their abundance in the pea genome. (1) We asked whether two contrasting types of pea LTR-containing retrotransposons have the frequency and age distributions consistent with the behavior of neutral alleles and whether these parameters can explain the rate of change of genome size in legumes. (2) We used the recently assembled v1a pea genome sequence to obtain data on LTR-LTR divergence from which their age can be estimated. We coupled these data to prior information on the distribution of insertion site alleles. (3) We found that the age and frequency distribution data are consistent with the neutral theory. (4) We concluded that demographic processes are the underlying cause of genome size variation in legumes.

Keywords: genome size; retrotransposons; pea; legumes

1. Introduction

Variation in the size of nuclear genomes among organisms has been a long-standing area of interest [1,2].

Within the legumes (Leguminosae, or Fabaceae), genome sequences are now available for a broad diversity of Papilionoid (Faboid) taxa [3] and these show that legume genomes typically have ca. 37,000 ± 10,000 annotated genes, similar to that for angiosperms more widely [4]. Among diploid legume species, genome size ranges about 40-fold, from ca. 340 Mb in several *Trifolium* species to a little over 14,000 Mb in *Lathyrus vestitus* [5]. Genome size variation among legumes is in contrast to their relatively constant gene number. However, genome size in *Pisum* seems to be stable, despite underlying variation in the presence and absence of retrotransposon insertions [6,7]. The one exception to this stability is the approximately 10% larger genome size noted in *P. abyssinicum* and *P. fulvum* [6], which are notably distinct taxa [7] within the genus.

Much of the variation in diploid legume genome size is attributable to variation in the content of LTR (long terminal repeats) retrotransposons [8–10]. Retrotransposons replicate by a copy and paste mechanism [11] and so they have the potential to accumulate to a great extent in nuclear genomes. It was suggested that this behavior implies that genome size should increase irrevocably [12] unless mechanisms exist by which retrotransposons may be removed [13–18]. This process was discussed recently by Jedlicka et al. [19].

In this study, our aim is to investigate the properties of retrotransposon insertions in the *Pisum* genome in order to constrain population genetical models of their dynamics. This requires a description of the age and frequency distribution of retroelement insertions in order to put limits on population genetic parameters of the neutral theory [20]. The details of these models are described in Section 2

below. We selected two contrasting elements for this analysis. The first, *PDR1*, is a *Ty1/copia* superfamily retrotransposon present in about 200 copies per haploid genome, evenly distributed along all pea chromosomes as was shown genetically [21], by in situ hybridization (Vershinin, unpublished data) and as is clear from the available genome assembly [22]. *PDR1* is about 4 kb in length, and its LTRs, at 156 bp [23], are exceptionally short. The second, *Cyclops*, has the typical pol region of the *Ty3/gypsy* superfamily of retrotransposons and is present in about 5000 copies [24]. *Cyclops* elements are approximately 12 kb long, including very long LTRs of about 1500 bp.

Previous studies in *Pisum* reached two important conclusions about its retroelement content; the first is that allelic variation in the genus *Pisum* is very broadly distributed and "recombination, introgression, and segregation between pea inbred lineages is common, although this may be rare per plant generation" [7]. The second conclusion is that the average age of retrotransposon insertions is one to two Myr [25]. Now that a genome sequence of *Pisum* was been assembled [22], further study of divergence between LTRs of individual elements and a more complete understanding of their genomic location is possible. Here, we are interested in how treating retrotransposition as an analogue of neutral base substitution provides insight into the expected age and frequency distribution of retrotransposon insertions. In other words we are asking whether genetic drift alone can explain the variation in genome size in the *Viceae*.

2. Materials and Methods

2.1. Plant Material and the Selection of Accessions for Analysis

Accessions from the John Innes Pisum Germplasm are designated JI*x*, where *x* is a number [26]. The analysis of this collection was carried out by the SSAP (Sequence Specific Amplification Polymorphism) technique and the data obtained were used to generate a pairwise distance matrix of allelic differences [7]. Principal coordinate analysis was used to order the distance matrix of all pairwise differences, and reduce to one member, pairs or groups of accessions that shared 95% or more of the marker alleles. A selection of 44 accessions was made from these data after excluding those that were closely related. This eliminated one *P. sativum* accession (JI 188), two *P. sativum* ssp *transcaucasicum* accessions (JI 2547 & JI 196), and four *P. abyssinicum* accessions (JI 1556, JI 2385, JI 130 and JI 2); leaving the following accessions: *P. abyssinicum*: JI 225; *P. fulvum*: JI 224, JI 1006, JI 1010, JI1796; *P. elatius*: JI 64, JI 254, JI 261, JI 262, JI 199, JI 1074, JI 1092, JI 1093, JI 1096, JI 2201, JI 2055, and JI 1794 (sometimes called *P. humile*). The *P. sativum* accessions included JI 45 and JI 2546 (designated ssp *transcaucasicum*), JI 156, JI 185, JI 189, JI 281 (African landraces), the Asian landraces JI 85, JI 95, JI 102, JI 109, JI 181, JI 241,JI 804, JI 1346, JI 1428, JI 1854, JI 2545, JI 250 (sometimes called *P. jomardii*), JI 52, JI 201, JI 209, JI 284, JI 399, JI 1030, JI 1089, JI 1846, and JI 2713. All accessions are available from the John Innes *Pisum* germplasm collection [26].

2.2. Population Genetic Considerations

The effective population size is the number of individuals that would be needed to generate any given statistic of population genetics for the population, if it comprised a set of individuals that interbreed freely and at random, i.e., are in Hardy–Weinberg equilibrium.

2.2.1. Allele Frequency Distribution

The presence or absence of retrotransposons at individual locations in the pea genome was observed by the SSAP technique [21]. We treated these data as genetic loci with two allelic states. The ancestral condition, which is the absence of an insertion, is called the unoccupied or empty site, and an evolutionarily derived allele, the occupied site, is defined by the insertion of a retrotransposon at this previously unoccupied site. The derived allele can suffer subsequent loss of the internal region (between the LTRs) by LTR–LTR recombination creating a solo LTR, or the deletion of the genomic region carrying the insertion. These events are not discussed further as they occur in a fraction of the

individuals in the population that carry an occupied site allele, and their subsequent behavior would follow the same trajectory as the initial insertion allele.

Retrotransposon insertion creates a new allele with a frequency (p) for the occupied site and the frequency of the empty site becomes ($1 - p$); initially $p = 1/2N$, where N is the population size, and the factor of 2 is because the species is diploid. These values define the effective heterozygosity; $H_e = 2p(1 - p)$, which is the chance that two alleles chosen at random are different. Effective heterozygosity, for neutral alleles, is related to the population genetic parameters of effective population size (N_e) and mutation rate (μ) [20,27–29]:

$$4N_e\mu = H_e/(1 - H_e)\,; \quad \mu = H_e/4N_e(1 - H_e) \tag{1}$$

When $4N_e\mu$ is estimated from p, the lowest frequency for which an allele can be observed is the reciprocal of the number of individuals that were genotyped. As $p \rightarrow 0$, $4N_e\mu \rightarrow 2p$ determines the resolution of the observable values of $4N_e\mu$.

Furthermore, the expected frequency distribution of the abundance of an allele $\Phi(x)$ is determined by the effective population size N_e and the mutation rate μ [20,27,29] as follows:

$$\Phi(x) = 4N_e\mu(1-x)^{4N_e\mu}/x \tag{2}$$

We used the average frequency of occupied sites to determine the expected frequency distribution using Equations (1) and (2). We then determined whether or not the observed data were a good fit to this expectation using χ^2 test. For clarity, we used the term ρ for the retrotransposition rate to avoid confusion with single base substitution, μ.

The SSAP data are available in Tables S1 and S2.

2.2.2. Age Distribution of Occupied Sites

LTR retrotransposons replicate by a copy and paste mechanism [11] where a transcript is initiated in the 5′ LTR and terminated in the 3′ LTR. Reverse transcription of this RNA and second strand synthesis generates a circularly permuted intermediate dsDNA where the LTR of this DNA is derived from one copy of the LTR sequence [30,31]. Upon insertion into the genome, the single LTR of the cDNA is replicated and defines the two ends of the element. Thus, at the time of insertion these two DNA sequences are derived from a single molecule and are therefore expected to be identical. Differences between these LTRs can accumulate due to mutation, and for this reason the comparison of the LTR sequences at an individual insertion site was used as a measure of the time since insertion, based on the assumption that these sequence differences arise by mutation at the same rate as silent substitution [32].

Using the pea v1a genome sequence [22] and prior data [7], we compared the age of an insertion estimated from LTR–LTR sequence divergence to the expected age of a neutral allele in a population, as determined by population genetic parameters. Kimura and Ohta [33] derived formulas for the age of neutral alleles that first achieve a given frequency in a population:

$$\bar{t}_x(0) = 4N_e\left\{\frac{1-x}{x}\ln(1-x) + 1\right\} \tag{3}$$

where x is the frequency of a neutral allele in the population after an average of $\bar{t}$ generations, having started at a very low frequency ($1/2N$), which can be considered to be effectively 0. Note that age is independent of the retrotransposition rate as it describes the fate of an allele once it has been formed. Kimura and Ohta [33] showed that the average or expected age, $E(age)$, of a neutral allele is a function of effective population size N_e and the current frequency of that allele x, such that:

$$E(age) = -4N_e[x/(1-x)]\ln(x) \tag{4}$$

As x ranges between 0 and 1, the term $[x/(1-x)]\ln(x)$ ranges between 0 and −1, which means that $E(age) \leq 4N_e$. We used estimates of N_e obtained from the allele frequency distribution (above) to determine whether the observed (from LTR–LTR divergence) and expected (from Equation (4)) ages of retrotransposon insertion sites were compatible.

3. Results

3.1. Retrotransposons in Legumes

The taxonomic distribution of legume genome size (Figure 1) shows that the largest genomes occur within the Viceae (Fabeae) tribe, which includes *Pisum*, *Lathyrus*, *Vicia*, and *Lens*. The Viceae genomes are not uniformly large, but also contain species with genomes of a size more typical for legumes generally. The distribution of genome sizes within the Viceae is consistent with an evolutionary history of both increase and decrease in genome size (Appendix A, Figure A1).

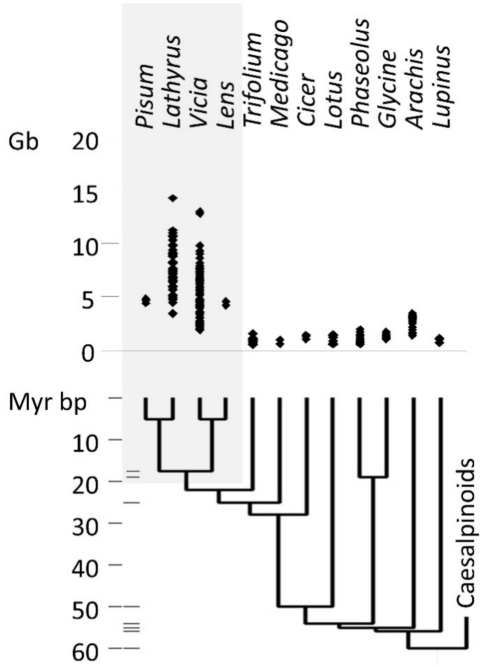

Figure 1. Diploid genome sizes in legumes. Genome sizes are from the plant C-value database [5]. The phylogenetic tree and the dates for divergences are from [34]. Grey shading indicates taxa in the tribe Viceae; horizontal bars to the right of the 10 Myr time scale represent splits supported by fossil evidence [34].

The distribution of genome sizes in the Vicieae suggests that evolutionary change in diploid genome size occurred within 5 My (Figures 1 and A1) and is therefore rapid, which is consistent with the differences being due to differential accumulation of retrotransposons.

3.2. Allele Frequency Distribution

Jing et al. [25] compared the observed and expected frequency distribution of insertion site alleles for the *Ty1/copia* element *PDR1* and found it a good fit to the expectation from the neutral theory. Here, we undertook the same analysis for the more abundant *Ty3/gypsy* element *Cyclops* [24] using the data from [7] in a selection of 44 pea accessions that represent the diversity of *Pisum* and that does not include multiple closely related accessions (see Section 2). The *PDR1* data for this subset of 44 accessions is compared to the frequency distribution of *Cyclops* insertion alleles in Figure 2.

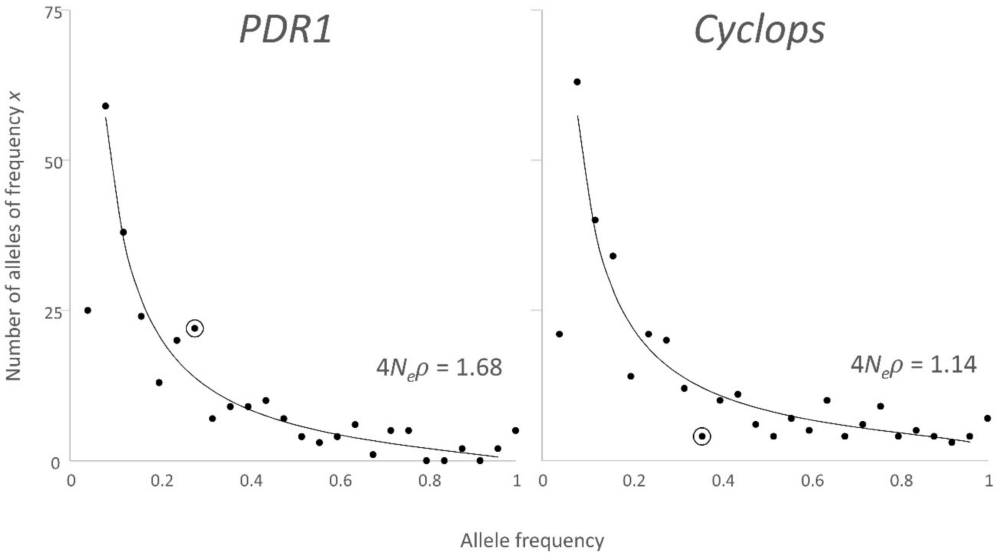

Figure 2. Frequency distribution of insertion alleles. The occurrence of occupied sites was observed for 329 *Cyclops* and 281 *PDR1* SSAP markers within a set of 44 accessions representing *Pisum* diversity [7]. The frequency of these occupied site allele is given on the x axis, binned in groups of 0.04 (0–0.04, 0.04–0.08, … 0.96–1). The y axis is the number of alleles within the frequency class on the x axis. The black line is the fit of $\Phi(x)$ with the minimum total χ^2. The curve for $\Phi(x)$ has $4N_e\rho$ as 1.68 and 1.14 for *PDR1* and *Cyclops*, respectively (Table S3, Appendix B). The χ^2 test showed that the highlighted values (ringed) differ from expectation at 5%, but not 1%, level. Note that in Equation (2) where $x = 0$, $\Phi(x)$ is unbounded.

Figure 2 shows that, for both plots, there are fewer alleles with a frequency in the range 0–0.04 than is expected from the neutral theory. Presumably, this is because frequencies less than 1/44 cannot be observed. There is also an excess for the 'fixed' class (allele frequency = 44/44), where all accessions carry the occupied site allele. This observed fixed class also includes alleles with a frequency greater than 44/45. Hence, this frequency class is expected to be overrepresented. That is, we cannot distinguish between insertion sites in all individuals in the genus from insertion sites present in just these 44 accessions.

The area under the curve corresponding to $\Phi(x)$ has to be estimated numerically, because the function has an improper integral; the area under the tails of the curve cannot be determined.

The occupied sites, which are present in only one accession, are distributed widely, for *PDR1* there are 25 of these, while for *Cyclops* there are 21. Of these, 4 are in the single *P. abyssinicum* accession and 9 in the 4 *P. fulvum* accessions, consistent with the differentiation of these taxa.

A χ^2 test for the observed vs. expected number of alleles in each frequency class, other than the two extremes, shows which observed values are significantly different from expectation. For this test, all frequency classes with an expectation less than or equal to 5 were combined into a single group. For both retrotransposons, a single class (ringed in Figure 2) had a significant value, $\chi^2 = 4.84$ & 5.43, $p = 0.0278$, and 0.020 for *PDR1* and *Cyclops*, respectively. For *PDR1* occupied site allele frequencies ≥ 0.56, the expected number was equal to or less than 5, so these were treated as a single class, $\chi^2 = 0.001$, $p = 0.98$. For *Cyclops* occupied site allele frequencies ≥0.76, the expected number was equal to or less than 5, so these were treated as a single class, $\chi^2 = 0.86$, $p = 0.35$. These data suggest that with the exception of the fixed alleles and the lowest frequency class, the data are an excellent fit to the prediction of the neutral theory. We know from the discussion above, that the fixed alleles and lowest frequency class do not have a properly defined expectation. If we accept the interpretation that the data are a good fit to the neutral theory, then Equation (2) suggests that only ca. $\frac{1}{4}$ of occupied site alleles expected to be found with a frequency less than 0.04 were detected in this sample of accessions. For both retrotransposons, this is about 1/3 of the total number of occupied sites detected, implying that we have detected about 75% of the number of insertion sites in *Pisum* that could be detected in a

sample of this size. The expected abundance of alleles with a very low frequency is arbitrarily large, implying that a very large number of insertions are extremely rare and are very quickly lost from the population.

With the exception of the extreme values discussed above, the neutral theory appears to give an adequate description of the frequency distribution of occupied site alleles of retrotransposons in *Pisum*. The estimated values of $4N_e\rho$ are remarkably similar for *PDR1* and *Cyclops*, two very different elements. This is not expected and implies that the survival rate of new insertions in the population is similar. We therefore asked whether the neutral theory can also explain the age distribution of retrotransposon insertions.

3.3. The Distribution of Cyclops Elements in the Cameor Genome Assembly

Ty3-gypsy elements are often described as being clustered in pericentric regions, as for example in *Arachis* [35]. Using the theory of runs [36] to examine the location of *Cyclops* elements in the pea cv. Cameor genome [22] provided no evidence for their having a non-random distribution at the scale of 100 kb blocks (Appendix C). The low recombination pericentric regions occupy ca. 720 Mb or roughly 18% of the 3.92 Gb assembly. Given the random distribution of *Cyclops* elements, we expect ca. 18% to lie within these low recombination regions. This should correspond to about 60 of the insertion sites assayed in the genetic diversity study.

3.4. Occupied Site Allele Age Distribution

Cyclops LTR sequences in the Cameor genome that were in the same orientation and separated by 8 to 10 kb were identified as candidates for pairs flanking a single element. This selection was further refined by removing sequences where more than two alignments with the LTR were found (Appendix C). This left 390 LTR sequences with the appropriate spacing and orientation. Neighbor Joining trees of these sequences were generated to test whether the adjacent LTRs were each other's most similar sequences. This further stringent filtering step left a list of 49 LTR pairs that had the expected characteristics from a single insertion event (Appendix C). These paired LTRs were compared to each other using BLASTn, noting the alignment length and the number of mismatches (gap openings were ignored), to determine the number of substitutions between LTR pairs.

Twenty-five LTR pairs of the *PDR1* retroelement present in the Cameor genome were identified and filtered in a similar way, and the number of pairwise differences was determined. These data were compared to the 49 LTR pairs previously described by Jing et al. [25], as presented in Figure 3 and Table 1.

Table 1. LTR pair divergence.

Element	μ ± SD, *n* [1]	Estimated Age [2]	Source
PDR1	0.013 ± 0.011, 49	1.89 ± 0.33	[25]
PDR1	0.013 ± 0.014, 25	1.86 ± 0.55	This work, [22]
PDR1	0.013 ± 0.013, 74	1.88 ± 0.30	combined
Cyclops	0.016 ± 0.009, 49	2.20 ± 0.30	This work, [22]

[1] Fraction of pairwise substitutions, n number of LTR pairs. [2] Myr (μ ± SEM).

These estimates of sequence divergence are not significantly different from one another, or from the data of Jing et al. [25].

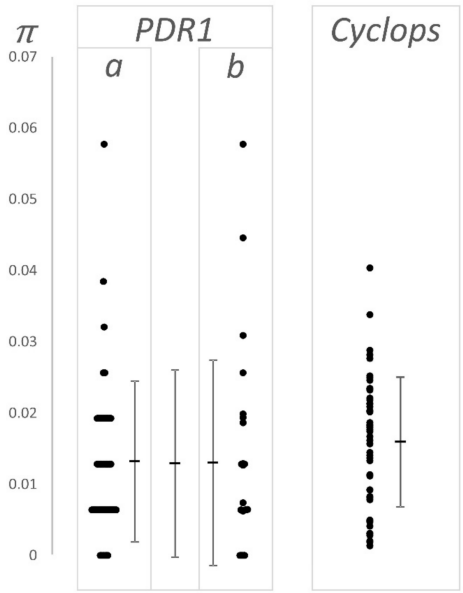

Figure 3. LTR–LTR divergence. The y axis, π (= 4 $N_e\mu$), is the fraction of single nucleotide substitutions observed over the length of the compared sequences. This ignores the variation due to indels, but includes their length. The data for *PDR1* are (**a**) taken from [25], (**b**) derived from the Cameor genome sequence [22] as is the data for *Cyclops*. All individual values are plotted (side by side when they have the same value) and the mean and standard deviation of the values are also plotted (see also Table 1). For *PRD1*, the central mean and standard deviation is for the combined data set. The bunching of values, giving a ladder-like appearance to the *PDR1* data is because this LTR is short; the observable values of π increment by the reciprocal of the LTR length, e.g., the first 'rung' of points above zero for *PDR1* is at 1/156 (ca. 0.006).

4. Discussion

We have investigated the age and location of two retrotransposons in pea genomes. *PDR1* is a *Ty1/copia* class element present in about 200 copies per genome, while *Cyclops* is a *Ty3/gypsy* element present in about 5000 copies [23,24]. The insertion sites of these two contrasting types of retrotransposon have a similar age and frequency distribution in *Pisum*, therefore it seems plausible that common factors have shaped these features of the elements. Inevitably, they have shared a similar population biology of their host plant, which is one obvious factor in common.

4.1. Nucleotide Diversity and Effective Population Size

Jing et al. [37] estimated the nucleotide diversity ($\pi = 4N_e\mu$) among 39 genes in 46 *Pisum* accessions as 0.011 ± 0.007. Estimates of π are also available from Sulima et al. [38] based on three genes among 110 accessions and from the 30 sequences derived from 25 genes among 100 accessions analyzed by Carpenter et al. [39]; these are 0.019 ± 0.003 and 0.006 ± 0.005 (mean ± SD), respectively. Kreplak et al. [22] estimated the nucleotide diversity of *Pisum* as ca. 8×10^{-4}, which is about an order of magnitude lower than in the other three studies. Estimates of nucleotide diversity depend on the range of accessions analyzed, and the first three data sets were designed to capture the diversity of *Pisum* as a whole, while Kreplak et al. [22] were primarily concerned with the sequence of the cultigen Cameor in the context of cultivated pea and its relatives; accordingly, this set was dominated by cultivated forms. These accessions included 16 cultivars, 15 landraces, 2 *P. abyssinicum*, and 10 wild accessions. This difference in the representation of wild accessions, which carry the bulk of the diversity of *Pisum*, is consistent with the lower estimate of π in Kreplak et al. [22]. An estimate of $4N_e\mu$ for *Pisum* as a whole in the range 0.005 to 0.01 is compatible with all these previous data. If we take the mutation rate as ca. 10^{-8}, then the estimate of N_e is ca. $3–4 \times 10^4$.

4.2. Age Distribution of LTR Pairs and Effective Population Size

A critique of the LTR–LTR comparison method for dating the age of retrotransposon insertions was made by Jedlicka et al. [19], who claimed that biases exist, some attributed to conversion events, such that longer LTR pairs were more similar to each other than shorter LTR pairs. The authors also commented that this phenomenon was partly reproduced in data simulation (although the reason for this was not clear). It should be noted that LTR length is not independent of retrotransposon family and different retrotransposons may have different genomic locations that may contribute to differences in recombination and/or gene conversion rate. Potential gene conversion events were identified by comparing the "ratio of solo LTR/FL"—presumably comparing the sequence of solo LTRs with that of the paired LTRs of intact elements. This method assumes that the sequence diversity of solo LTRs and the LTRs of a given intact element is the same, which may not be the case because of subfamily structure within retroelements [40].

Furthermore, Jedlicka et al. [19] noted that for approximately a quarter of nested insertions, the targeted element appeared to be younger than the element that was subsequently inserted. There are several mechanisms by which this may occur, but the observation highlights the need for caution and emphasizes the possibility that recombination-like processes may lead to an underestimate of the divergence between LTR pairs. Nevertheless these authors note that, for a wide range of species, most estimates of the mean age of LTR retroelement insertions are in the range of 1–3 Myr, which is consistent with the estimates obtained here.

From Equation (4), the expected age of any allele with a frequency $x = 0.28$ (the average of the frequencies for both *PDR1* and *Cyclops*) is ca. 60,000 to 120,000 years. This means that the measured age of retrotransposon insertions (1–3 Myr) is very much greater than expected, but can be understood as follows (see also Appendix D). Retrotransposon insertions, which carry a sequence difference between the LTRs are necessarily derived alleles; they must have occurred in a pre-existing insertion. The expected age in Equation (4) corresponds to the length of time until the insertion allele *first* reaches the frequency x, not the average age of an allele of this frequency. An insertion in which there is one difference between the LTRs arose from an insertion allele in which the LTRs were identical. The derived allele was therefore at the frequency 1/2N when the original insertion event occurred, and again 1/2N when the mutation defining the derived allele occurred. The number of times an insertion allele has visited the frequency 1/2N is therefore at least equal to the number of differences between the LTRs. Each time this occurs, the probability that the allele will be lost by chance alone is high, thus we do not expect a large number of SNP variants per insertion site, nor do we expect such variants to exist at an appreciable frequency in *Pisum* as a whole.

With a nucleotide substitution rate of ca. 10^{-8}, for a retroelement with LTR length 0.1 to 1 kb a single base change will on average occur within about 10^5 to 10^6 years. If this variant reaches a moderate frequency in the population, then a further period of about 10^4 to 10^5 years will have elapsed. Thus, the estimated age of retrotransposons of the order of 1–2 million years is consistent with the mutation rate in the LTRs and the population dynamics that permit only a few of these derived alleles to achieve a moderate frequency.

4.3. Gain and Loss

The similar estimates of $4N_e\rho$ (Figure 2, i.e., the equivalent of π for retrotransposon insertions) above suggest that the long-term transposition rate ρ is very similar for *PDR1* and *Cyclops*, and is about 1.5×10^{-7}. The similarity of these two values of ρ may simply reflect a long-term average, with transposition rate varying between the elements from time to time. It is necessarily the case that ρ is the transposition rate for insertion sites that survive in the population, which is not necessarily the same as the rate at which retrotransposition occurs in a given individual.

Our study suggests that there is little remarkable about the age and frequency distribution of retrotransposon insertion site alleles in pea, yet we know that pea and its relatives in the Viceae present a diversity of genome sizes. Many of these species have genome sizes larger than pea, and others

with smaller genomes. The size difference in these genomes seems to be accounted for by differential accumulation of retrotransposons [8,10], and the taxonomic distribution of genome size suggests that both gain and loss has occurred (Appendix A, Figure A1). The overall abundance of retrotransposons could change if their transposition rate changed coordinately. However, it seems more likely that this genome-wide property is a consequence of the period of time for which they remain in the genome, and this is determined solely by effective population size. Thus, we propose that historical (evolutionary) changes in effective population size are the main reason for the diversity of genome size in the Viceae. An increase in effective population size, all other things being equal, would lead to an accumulation of (polymorphic) retrotransposon insertions and hence an increase in average genome size. On the other hand, a reduction in effective population size would reduce allelic diversity by facilitating the loss of alleles. Reduction in effective population size would not necessarily reduce genome size, but would replace the mean genome size with the mean genome size of a sub-population. However, with a reduced effective population size, the number of polymorphic alleles that could accumulate would be reduced, so the effectiveness of retrotransposition to increase average genome size would be reduced.

In general, recombination rate per kb is negatively correlated with genome size [2] and, through hitchhiking effects, recombination rate influences the effective population size [2,41] such that the effective population size is increased in regions of higher recombination rate. With higher effective population size, the average age of alleles is increased (Equation (4) above). So regions of high recombination rate will include polymorphic retrotransposon insertions, which would be fixed (as either the empty or occupied site) more rapidly in regions of low recombination or lower effective population size.

Bertioli et al. [35] showed that the *A. ipaensis* genome is 10%–20% larger than that of *A. duranensis* with more frequent duplications and a higher transposon content. Several of the corresponding chromosomes in these genomes differ by having a large distal inversion so that the telomeric region of one is closer to the centromere in the other. The alignment of the pseudomolecules shows that the physical distance between matched sequences is longer in the species where these are nearer the centromere and shorter where these are nearer the telomere. This effect is continuous and gradual, as revealed by an arc in dot-plots of homeologous chromosomes [35], and is associated with a difference in transposon abundance; the extra transposons accounting for the increased length. These inversions have moved sequences from a region of high recombination (closer to the telomere) to a region of low recombination (closer to the centromere) and the consequence of this change is seen in the repetitive sequence content of these regions of the genome.

The occupied site allele for a retrotransposon is initially rare, so these are usually lost by genetic drift, but, by chance, a few may become relatively abundant in the population. A change in effective population size can have a systematic effect, as in the example from *Arachis* [34]. In pea, the recombination rate per kb is low with respect to its close relatives with smaller genomes. This is simply because chromosome arms, irrespective of size, typically have 1 or 2 crossovers; chromosomes require at least one crossover for proper disjunction. As recombination rate per kb simply describes the situation, it cannot be taken as an explanation for the increased retrotransposon content in pea (and many other members of the Viceae) compared to other legumes. However, if effective population size in a small genome ancestor of pea increased, then a longer time would have to elapse before the loss of insertion alleles and this effect, therefore, may have led to an abundance of polymorphic retrotransposon insertions as is seen in extant pea lineages.

4.4. Comparison with Other Studies

Our general conclusion from these observations is that the age and frequency distribution of *PDR1* and *Cyclops* retrotransposons in pea can be accounted for according to the Neutral Theory. In other words, their age and abundance are dominated by demographic processes. We infer that these processes would act on the genome as a whole, although they would be modulated somewhat by local genomic effects on effective population size. For this reason, we would expect to see coordinated behavior of

retrotransposons and in consequence their age distribution of insertions would be similar, reflecting these events. Note that the age distribution of insertions is entirely distinct from the age distribution of an element; an element may be much older than its individual incarnations, which successively occupy sites that are fleetingly present in the population.

Jedlicka et al. [19] used two methods to estimate the frequency distribution of the ages of insertions of 22 retroelement families in 15 diverse plant taxa. Both methods show broad similarity in the age distribution of insertion sites of different elements within each species, but with clear differences between species. In their 'complex' approach [19], attempting to account for conversion events, there is for example a marked bimodal age distribution for several elements in tomato. This type of pattern would be expected in species which have undergone changes in effective population size within the period of time that these insertion sites have survived.

A notable feature of pea is that it is predominantly self-pollinating. This led us to suspect that it should have a small effective population size as compared to the expectation under outcrossing. In turn, this would lead us to expect a relatively small genome size. However, as Vershinin et al. [7] noted, *Pisum* diversity is marked by recombination, introgression, and segregation. Presumably, this reflects outcrossing between stands of relatively homogeneous and homozygous individuals (Appendix D). The persistence of these stands should be assisted by self-fertility, and their persistence is required for successful outcrossing.

Macas et al. [9] have shown that *Ogre* elements represent by far the greatest bulk of LTR retrotransposons in the Viceae and that variation in their copy number is most strongly correlated with genome size within this tribe. Furthermore, although *Ogre* elements are present in other eudicots, including the Trifoleae, sister to the Viceae, it is only within the Viceae that they have reached such a high fraction of the genome [9]. These authors showed that *Ogre* elements are the main drivers of genome size variation in this tribe, while recognizing that "contrasting population sizes and different ecological and mating strategies" are likely to be significant forces shaping the retroelement composition of plant genomes. Here, we argued that effective population size and transposition rate together define these dynamics. The amplification and diversification of elements is represented by variation in transposition rate, but the dynamics of their accumulation or elimination needs to be understood in terms of population genetical history.

5. Conclusions

We propose that the uniformity of genome size in *Pisum* reflects the randomization of insertion alleles throughout the genus, rather than their fixation. Treating retrotransposon insertions as effectively neutral alleles can explain their age and frequency distribution in *Pisum*. If the elements we analyzed are representative of all pea retrotransposons, we can conclude that genetic drift alone can explain the variation in genome size in the *Viceae*. This further suggests that a large effective population size, which would maintain a high level of insertion site polymorphism, is the underlying cause of the large genome size in pea.

Supplementary Materials: The following are available online at http://www.mdpi.com/2673-6284/9/4/24/s1, Table S1: PDR1 SSAPs, Table S2: Cyclops SSAPs, Table S3: Data for Figure 2, Table S4: Data for Figure A2.

Author Contributions: Conceptualization, T.H.N.E. and A.V.V.; methodology, T.H.N.E. and A.V.V.; formal analysis, T.H.N.E. and A.V.V.; investigation, T.H.N.E. and A.V.V.; writing—original draft preparation, T.H.N.E.; writing—review and editing, T.H.N.E. and A.V.V.; visualization, T.H.N.E.; supervision, T.H.N.E. and A.V.V.; project administration, T.H.N.E.; funding acquisition, T.H.N.E. and A.V.V. All authors have read and agreed to the published version of the manuscript.

Funding: T.H.N.E. gratefully acknowledges the support of an Institute Strategic Fellowship from The John Innes Institute and the BBSRC SASSA UPGRADE project (BB/R020604/1). A.V.V. acknowledges the support of the Russian fundamental scientific research program (project 0310-2019-0003).

Acknowledgments: We thank B. Steuernagel, J. Hofer, C. Domoney, and C. Martin for useful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Taxonomic Distribution of Genome Sizes in the Viceae

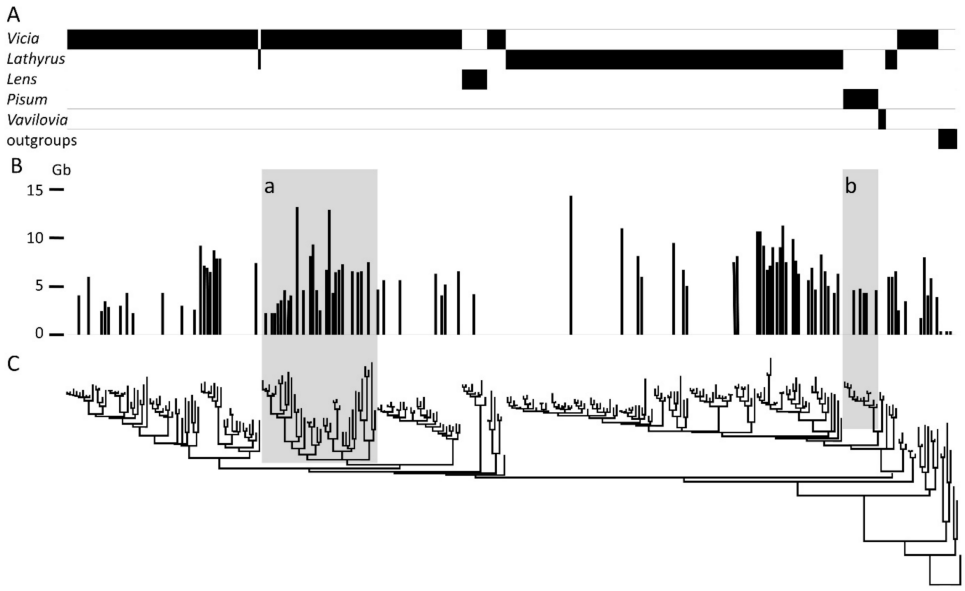

Figure A1. Phylogenetic distribution of diploid genome size in the Viceae. (**A**) The distribution of generic names within the phylogenetic tree of the Viceae. (**B**) Genome size for diploid taxa from [5]. (**C**) Phylogenetic tree, redrawn from Supplementary File 10 of [42]. The shaded regions are 'a' part of *Vicia* section *Vicia* including the small genome size species *V.amphicarpa*, *V. sativa*, and *V. faba* with a large genome size (see bar height); 'b' marks the genus *Pisum*.

Figure A1 shows that genome size in the Viceae (which ranges from ca. 1.8 to 14 Gb) does not have a simple phylogenetic distribution. The clade marked 'a' includes *V. amphicarpa* and *V. sativa*, with the second and third smallest genomes in the Viceae (2.1 and 2.2 Gb) as well as *V. faba* with the second largest of these genomes (13 Gb). In clade 'a' the small genome species are embedded among species with larger genomes. The neighboring taxa all have larger genomes, indicating that the small genome sizes of *V. amphicarpa* and *V. sativa* are derived; i.e., these genomes have reduced in size from that of the common ancestor of this clade. Similarly, the genus *Pisum* is a monophyletic group with a smaller genome than the majority of adjacent taxa in the phylogeny. These taxonomic relationships are consistent with some taxa having reduced, and others increased, genome size with respect to their progenitors.

Appendix B. Estimating $4N_e\mu$ from $\Phi(x)$

Equation (1) shows that $4N_e\mu$ can be determined directly from the allele frequency, but this is problematic because of the inability to count the number of low frequency alleles in a finite sample of accessions and because alleles fixed in the sample are not necessarily fixed in the population as a whole. For these reasons, $4N_e\mu$ is best estimated from the allele frequency distribution, which is described theoretically as $\Phi(x)$. Figure A2 illustrates a family of curves of $\Phi(x)$ for a range of values of $4N_e\mu$ (Equation (2)).

Fitting the curve of $\Phi(x)$ to the observed allele frequency data (Figure 2) was done by evaluating the predicted number of alleles with the frequencies 0.04, 0.08, etc., for a particular value of the parameter $4N_e\mu$ (see Figure 2). Initially, $4N_e\rho$ was in the range 0.5 to 2, incrementing the parameter in steps of 0.1. A χ^2 was then calculated for each frequency class and the value of $4N_e\rho$ with the minimum sum of χ^2 values was found. This process was then repeated in the vicinity of $4N_e\rho$ of interest, but with 10-fold smaller steps. For each estimate of $4N_e\rho$, the number of alleles expected in each frequency class was estimated by scaling the sum of the expected number of alleles to equal their observed number.

The highest and lowest observed frequency classes were not included in this sum for the reasons discussed above and in the main text.

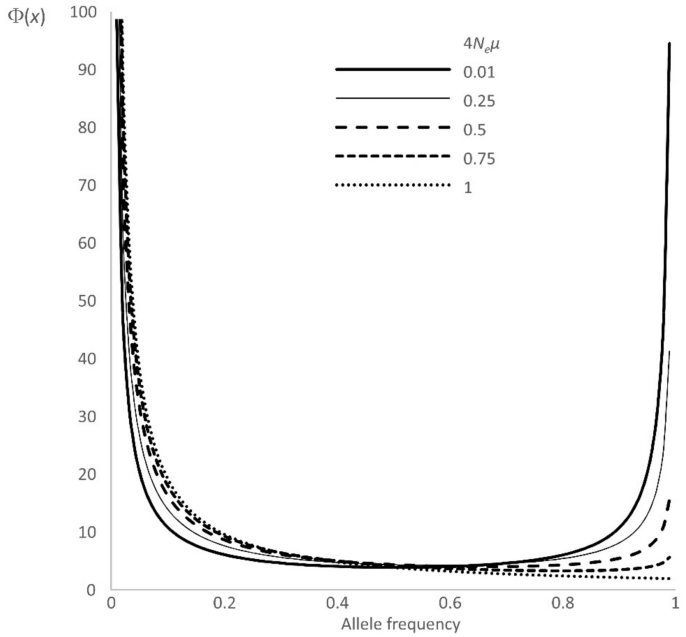

Figure A2. $\Phi(x)$ for different values of $4N_e\mu$. Equation (2) was evaluated and plotted for a range of values of $4N_e\mu$. This is typically a U-shaped curve when the mutation rate is much less than $1/N_e$ as is usual for base substitution rates. For the curves plotted in Figure A2 the y axis is the expected number of alleles of frequency x when sampled from 1000 alleles. See Table S4.

Note that high and low allele frequencies represent stable values, intermediate frequencies are unstable and variation from one generation to the next pushes allele frequency towards elimination or fixation.

Appendix C. Identification of *Cyclops* LTRs

Cyclops LTR1 and LTR2 sequences from the accession AJ000640 were aligned by Muscle to create a consensus sequence, very similar to the longer LTR2 sequence, but with LTR1 sequence replacing the few "-" in the LTR2 alignment. This consensus sequence was then used as the subject in a BLASTn query vs. Cameor v1a genome sequence [22,43] with a threshold e-value of 10^{-150}. This identified 5301 sequences, which were sorted by position in pseudomolecule (and scaffold). Among these, 928 were overlapping, presumably because of internal sequence duplications and/or insertion of one *Cyclops* into another. Of the overlapping sequences, 6 began at exactly the same nucleotide, presumably alternative ways of matching to the LTR sequence. These presumed insertions clustered near position 80 in the LTR (Figure A3).

None of the highly represented positions correspond to the U3/R or RU/5 boundaries, annotated as 1253..1258 "TATA_box" and 1211..1214 "CAAT_signal", respectively.

Next, the length of the sequence between LTRs was determined, and is plotted as their frequency distribution vs. $\log_{10}$ of the distance between adjacent LTRs in Figure A4. The bimodal distribution represents the distance between LTR pairs from individual elements (left peak) and the distance between the LTRs of neighboring *Cyclops* elements in the genome (right peak).

The blue lines are the same analysis, but excluding the four ringed positions.

In Figure A4, there is a clear peak near 10^4 (red), which corresponds to the size of the intact element (12,314 − 1504 = 10,810; complete element −1 LTR; or 9306 +1504 internal sequence plus one LTR). This left-hand peak is noticeably broad and skewed to larger sizes than the intact element, presumably reflecting insertions into the element. The second, right hand, broad peak near 10^6 nt reflects the spacing between elements and/or solo LTRs. Ignoring the spacing classes in red, the mean and standard

deviation of the spacing between LTRs of neighboring *Cyclops* elements is: 1,060,032 ± 1,764,315, $n = 3019$.

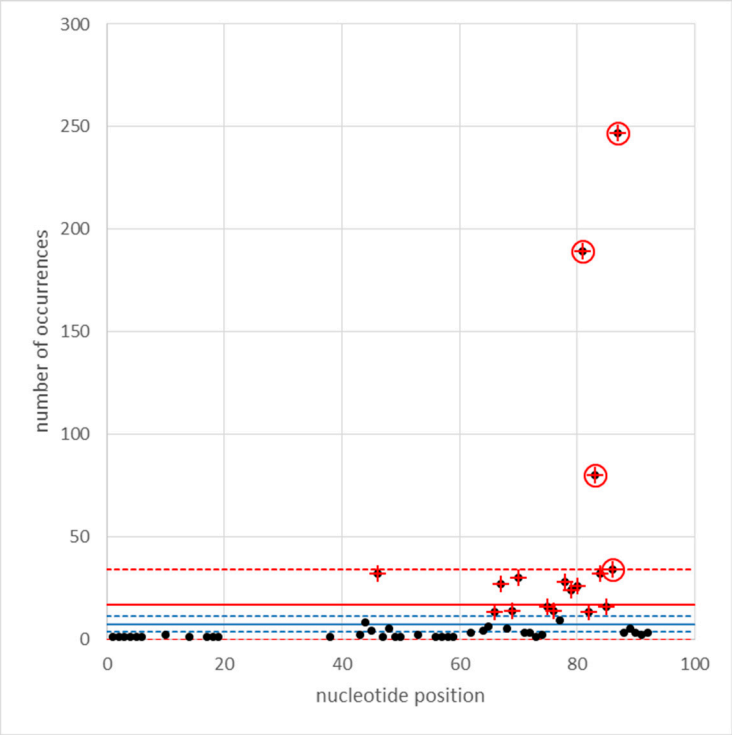

Figure A3. Insertions detected in Cyclops LTRs. The frequency distribution of the positions of disruption to the *Cyclops* LTRs are plotted. The red solid line refers to the mean frequency and the dashed lines to ± 3 standard error units of the number of occurrences per nucleotide (ignoring positions where no insertion occurred). The blue lines are the mean ± 3 standard error units excluding the data points marked with a red cross. The points ringed in red are outside the + 3SEM limit.

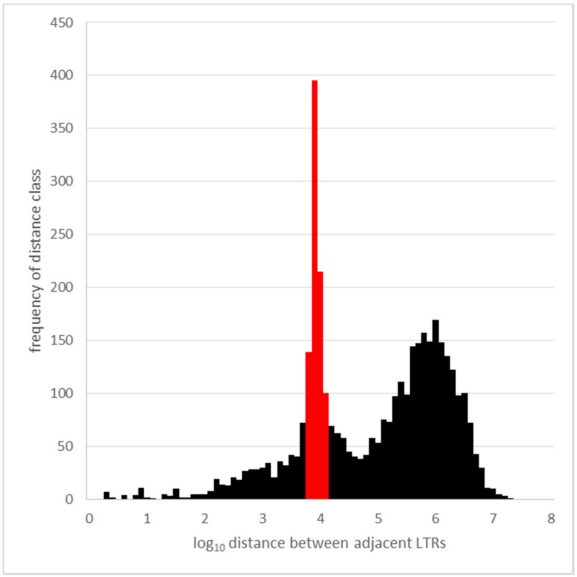

Figure A4. The spacing of Cyclops LTRs. The frequency distribution of distances between adjacent *Cyclops* LTRs in the Cameor genome assembly is plotted where the x axis is $\log_{10}$ of the spacing incremented in steps of 0.1, and the y axis is the count of the number adjacent LTRs with that spacing.

Considering the genome as a sequence of 100 kb blocks, each block may or may not contain LTR sequence(s). The expectation of the frequency distribution of the length of uninterrupted runs of blocks that do not contain LTRs can be derived from the theory of runs [36] as follows:

When two types of blocks are arranged in a row, n_1 of type 1 and n_2 of type 2, where ($n = n_1+n_2$), then the expected number of runs of type 1 blocks of exactly length i is given by Mood [36] in his Equation (3.6):

$$E(r_{1i}) = [(n_2 + 1)^2 n_1{}^i]/n^{(i+1)}$$

If $n_2 + 1 \simeq n_2$ and $p = n_2/n$, then the expected frequency of runs of type 1 elements is:

$$F = p^2(1 - p)^i, \text{ or } \log(F) = 2\log(p) + i\log(1 - p)$$

So, if we plot log(F) vs. i (Figure A5) we expect a straight line with a slope and intercept defined by the single parameter p. For *Cyclops* LTRs in the Cameor genome, the distribution of run length of 100 kb blocks that do not contain an LTR is shown in Figure A5.

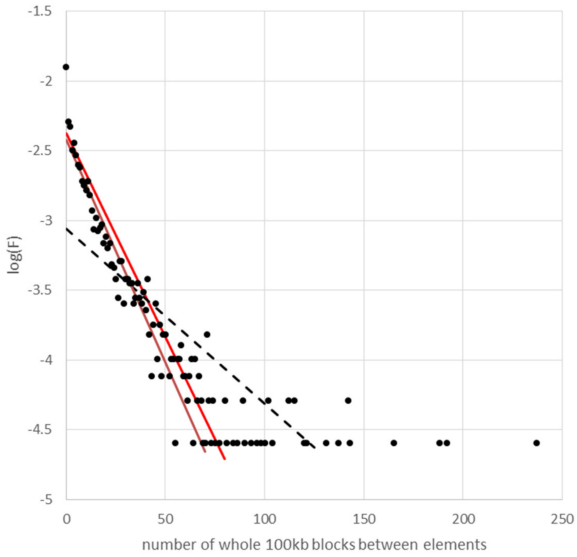

Figure A5. The frequency distribution of spacing between *Cyclops* elements. For the relationship log(F) = 2log(p) + $i$log(1 − p), the y axis is log(F) and the x axis is i the number of successive 100 kb blocks that do not contain an LTR.

The diagonal black dotted line is the regression line for all the points. As the y values get lower, the points spread out on the x axis; the bottom row of points are length classes that occur once. This contributes greatly to the regression, there being many of these points. The brown line is the regression for all values of i greater than the highest value of i, which was not observed; in this case the range is i = 0 to i =50. The red line is the expected distribution given the number of LTRs identified and the size of the assembled genome.

The observed slope and intercept of the brown line in Figure S4 are −0.032 ± 0.001 and −2.425 ± 0.043 (regression coefficient ± standard error) vs. The expected values of −0.029 and −2.375 (red line).

From the slope and intercept of the regression line we can estimate p as 0.066 ± 0.006 vs. The expected value of 0.065. On this basis, there is no evidence for a non-random distribution of *Cyclops* LTRs at the scale of 100 kb blocks.

Appendix D. Population Structure and Taxonomy of Natural and Domesticated *Pisum*

Pea is a crop species grown in many countries and is used in food as a dry seed, as a vegetable, or as a processed product. The dry seed, and the haulm are also used as animal feed or fodder. Global

pea production in 2018 was ca. 13.5 Mt dry seed (ca. 8.9 Mha in 96 countries) and 21 Mt of the fresh crop (ca. 4.4 Mha in 84 countries) [44]. Dry pea seeds are typically ca. 200 mg, so global production represents of the order of 10^{14} seeds. The species is currently dominated by agricultural production, nevertheless there remains a wild population distributed mainly around the Mediterranean basin, but extending from the Atlantic coast to the Indian subcontinent [45].

Wild pea has two major subgroups *P. elatius* and *P. fulvum* [7]. There were two independent domestications from *P. elatius*, one in the Fertile Crescent and the other probably in Ethiopia or Yemen (Ellis et al. 1998, Trněný et al. 2018), the resultant domesticated types are usually referred to as *P. sativum* and *P. abyssinicum*, respectively. *P. elatius* contains the bulk of the diversity of *Pisum*; some authors give additional subspecies names, but these are not well supported as monophyletic clades, nor are they remarkably distinct [7].

The phylogenetic relationships within *Pisum* given by Schaefer et al. [42] (Figure A1) describes the relationships between accessions. We do not consider it likely that all individuals assigned the same taxon names would lie on the same branch [7]. The different *Pisum* genome sizes in the Kew database [5] associated with (sub-)specific taxon names must not be taken to indicate the genome size of all individuals with the same name. Furthermore, the genome size of pea has many estimates, but the study of Baranyi et al. [6] is notable because it used two different methods for each accession and shows the contribution of experimental variation and also identified consistent differences. A major conclusion of their study is that the genome size of *P. fulvum* and *P. abyssinicum* is similar and about 10% greater than for other *Pisum*.

In the wild, *Pisum* grows in small groups of 10 or so individuals in maquis or disturbed ground [46]. The species usually has a low frequency of outcrossing; Blixt [47] collated information on outcrossing rates and concluded this was less than 1 in 30,000 plants. Nevertheless, the diversity of *Pisum* is clearly marked by exchange among lineages and a wide degree of allele sharing [7]; presumably this reflects gene-flow between such stands.

The breeding system for *Pisum*, of small stands of highly inbred plants yet with significant gene flow between lineages, suggests that the population genetics of random mating populations needs to be interpreted with caution when applied to this species. While meiosis is essentially annual, representing the plant generations, crossing between lineages is relatively rarer. This means that equating rates that occur per plant generation and per sexual cycle is probably incorrect. The measure of time in Equation (3) may need adjustment to match estimates of mutation rate of Equation (1).

Mutational differences accumulate at the rate of meioses, that is, per plant generation; however, the population genetic effects accumulate at the rate of outcrosses between lineages, and these two rates are quite different in an inbreeding species like pea. If there are multiple selfing generations (m) per outcross, then the mutation rate may appear m-fold higher than expected from standard population genetics. From the perspective of this analysis, our estimate of K_s, upon which age estimates are based, may be over estimated by the factor m, and if $E(age)$ is over estimated in Equation (4), then N_e is overestimated in proportion. If $m \simeq 10$, then $N_e \simeq 10^5$ and the transposition rate per outcrossing generation is 10-fold higher, but per meiotic generation, it would be unchanged. This would increase the estimates of 60,000–120,000 years (Section 4.2 above) to 0.6 to 1.2 Myr, closer to the estimate of insertion site age as estimated from LTR–LTR divergence, permitting few visits to the frequency 1/2N, as is consistent with observation.

References

1. Eddy, S.R. The C-value paradox, junk DNA and ENCODE. *Curr. Biol.* **2012**, *22*, R898-9. [CrossRef]
2. Lynch, M. *The Origins of Genome Architecture*; Sinauer Associates, Inc.: Sunderland, MA, USA, 2007.
3. LIS—Legume Information System. Available online: https://legumeinfo.org/species (accessed on 23 September 2020).
4. Cannon, S.; Iowa State University, Ames, IA, USA. Personal communication, 2016.
5. Plant DNA C-values Database. Available online: https://cvalues.science.kew.org/ (accessed on 23 September 2020).

6. Baranyi, M.; Greilhuberv, J.; Święcicki, W.W. Genome size in wild *Pisum* species. *Theor. Appl. Genet.* **1996**, *93*, 717–721. [CrossRef]
7. Vershinin, A.V.; Allnutt, T.R.; Knox, M.R.; Ambrose, M.J.; Ellis, T.H.N. Transposable elements reveal the impact of introgression, rather than transposition, in *Pisum* diversity, evolution and domestication. *Mol. Biol. Evol.* **2003**, *20*, 2067–2075. [CrossRef] [PubMed]
8. Hill, P.; Burford, D.; Martin, D.M.; Flavell, A.J. Retrotransposon populations of Vicia species with varying genome size. *Mol. Genet. Genom.* **2005**, *273*, 371–381. [CrossRef] [PubMed]
9. Macas, J.; Novák, P.; Pellicer, J.; Čížková, J.; Koblížková, A.; Neumann, P.; Fuková, I.; Doležel, J.; Kelly, L.J.; Leitch, I.J. In Depth Characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. *PLoS ONE* **2015**, *10*, e0143424. [CrossRef] [PubMed]
10. Vondrak, T.; Robledillo, L.Á.; Novák, P.; Koblížková, A.; Neumann, P.; Macas, J. Characterization of repeat arrays in ultra-long Nanopore reads reveals frequent origin of satellite DNA from retrotransposon-derived tandem repeats. *Plant J.* **2020**, *101*, 484–500. [CrossRef] [PubMed]
11. Boeke, J.D.; Chapman, K.B. Retrotransposition mechanisms. *Curr. Opin. Cell Biol.* **1991**, *3*, 502–507. [CrossRef]
12. Bennetzen, J.L.; Kellogg, E.A. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **1997**, *9*, 1509–1514. [CrossRef] [PubMed]
13. Ma, J.; Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 12404–12410. [CrossRef] [PubMed]
14. Bennetzen, J.L.; Ma, J.X.; Devos, K. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* **2005**, *95*, 127–132. [CrossRef]
15. Vitte, C.; Panaud, O. LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model. *Cytogenet. Genome Res.* **2005**, *110*, 91–107. [CrossRef] [PubMed]
16. Vitte, C.; Panaud, O.; Quesneville, H. LTR retrotransposons in rice (*Oryza sativa*, L.): Recent burst amplifications followed by rapid DNA loss. *BMC Genom.* **2007**, *8*, 218. [CrossRef] [PubMed]
17. Hawkins, J.S.; Proulx, S.R.; Rapp, R.A.; Wendel, J.F. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 17811–17816. [CrossRef] [PubMed]
18. Tian, Z.X.; Rizzon, C.; Du, J.C.; Zhu, L.C.; Bennetzen, J.L.; Jackson, S.A.; Gaut, B.S.; Ma, J.X. Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* **2009**, *19*, 2221–2230. [CrossRef]
19. Jedlicka, P.; Lexa, M.; Keinovsky, E. What can Long Terminal Repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Front. Plant Sci.* **2020**, *11*, 644. [CrossRef]
20. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1983.
21. Ellis, T.H.N.; Poyser, S.J.; Knox, M.R.; Vershinin, A.V.; Ambrose, M.J. Polymorphism of insertion sites of Ty1-copia class retrotransposons and its use for linkage and diversity analysis in pea. *Mol. Gen. Genet.* **1998**, *260*, 9–19. [CrossRef]
22. Kreplak, J.; Madoui, M.-A.; Cápal, P.; Novák, P.; Labadie, K.; Aubert, G.; Bayer, P.E.; Gali, K.K.; Syme, R.A.; Main, D.; et al. A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* **2019**, *51*, 1411–1422. [CrossRef]
23. Lee, D.; Ellis, T.H.N.; Turner, L.; Hellens, R.P.; Cleary, W.G. A *copia*-like element in *Pisum* demonstrates the uses of dispersed repeated sequences in genetic analysis. *Plant Molec. Biol.* **1990**, *15*, 707–722. [CrossRef]
24. Chavanne, F.; Zhang, D.-X.; Liaud, M.-F.; Cerff, R. Structure and evolution of *Cyclops*: A novel giant retrotransposon of the *Ty3/Gypsy* family highly amplified in pea and other legume species. *Plant Mol. Biol.* **1998**, *37*, 363–375. [CrossRef]
25. Jing, R.C.; Knox, M.R.; Lee, J.M.; Vershinin, A.V.; Ambrose, M.; Ellis, T.H.N.; Flavell, A.J. Insertional polymorphism and antiquity of *PDR1* retrotransposon insertions in *Pisum* species. *Genetics* **2005**, *171*, 741–752. [CrossRef]
26. SeedStor Homepage. Available online: https://www.seedstor.ac.uk/ (accessed on 23 September 2020).
27. Kimura, M.; Crow, J.F. The number of alleles that can be maintained in a finite population. *Genetics* **1964**, *49*, 725–738. [PubMed]
28. Kimura, M.; Ohta, T. *Theoretical Aspects of Population Genetics*; Monographs in Population Biology. 4; Princeton University Press: Princeton, NJ, USA, 1971.
29. Kimura, M. Rare variant alleles in the light of the neutral theory. *Mol. Biol. Evol.* **1983**, *1*, 84–93. [PubMed]

30. Garfinkel, D.J.; Stefanisko, K.M.; Nyswaner, K.M.; Moore, S.P.; Oh, J.; Stephen, H.; Hughes, S.H. Retrotransposon Suicide: Formation of Ty1 Circles and Autointegration via a Central DNA Flap. *J. Virol.* **2006**, *80*, 11920–11934. [CrossRef] [PubMed]
31. Schulman, A.H. Retrotransposon replication in plants. *Curr. Opin. Virol.* **2013**, *3*, 604–614. [CrossRef]
32. SanMiguel, P.; Gaut, B.S.; Tikhonov, A.; Nakijama, Y.; Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **1998**, *20*, 43–45. [CrossRef]
33. Kimura, M.; Ohta, T. The age of a neutral mutant persisting in a finite population. *Genetics* **1973**, *75*, 199–5312.
34. Lavin, M.; Herendeen, P.S.; Wojciechowski, M.F. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst. Biol.* **2005**, *54*, 575–594. [CrossRef]
35. Bertioli, D.J.; Cannon, S.B.; Froenicke, L.; Huang, G.; Farmer, A.D.; Cannon, E.K.S.; Liu, X.; Gao, D.; Clevenger, J.; Dash, S.; et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **2016**, *48*, 438–446. [CrossRef]
36. Mood, A.M. The Distribution Theory of Runs. *Ann. Mat. Stat.* **1940**, *11*, 367–392. [CrossRef]
37. Jing, R.; Johnson, R.; Seres, A.; Kiss, G.; Ambrose, M.J.; Knox, M.R.; Ellis, T.H.N.; Flavell, A.J. Gene-based sequence diversity analysis of field pea (*Pisum*). *Genetics* **2007**, *177*, 2263–2275. [CrossRef]
38. Sulima, A.S.; Zhukov, V.A.; Afonin, A.A.; Zhernakov, A.I.; Tikhonovich, I.A.; Lutova, L.A. Selection signatures in the first exon of paralogous receptor kinase genes from the *Sym2* region of the *Pisumsativum*, L. genome. *Front. Plant Sci.* **2017**, *8*, 1957. [CrossRef] [PubMed]
39. Carpenter, M.A.; Shaw, M.; Cooper, R.D.; Frew, T.J.; Butler, R.C.; Murray, S.R.; Moya, L.; Coyne, C.J.; Timmerman-Vaughan, G.M. Association mapping of starch chain length distribution and amylose content in pea (*Pisum sativum* L.) using carbohydrate metabolism candidate genes. *BMC Plant Biol.* **2017**, *17*, 132. [CrossRef] [PubMed]
40. Casacuberta, J.M.; Vernhettes, S.; Audeon, C.; Grandbastien, M.-A. Quasispecies in retrotransposons: A role for sequence variability in *Tnt1* evolution. *Genetica* **1997**, *100*, 109–117. [CrossRef]
41. Hahn, M.W. Toward a selection theory of molecular evolution. *Evolution* **2008**, *62*, 255–265. [CrossRef] [PubMed]
42. Schaefer, H.; Hechenleitner, P.; Santos-Guerra, A.; Menezes de Sequeira, M.; Pennington, R.T.; Kenicer, G.; Carine, M.A. Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evol. Biol.* **2012**, *12*, 250. [CrossRef]
43. Pea Genome Project. Available online: https://urgi.versailles.inra.fr/Species/Pisum/Pea-Genome-project (accessed on 25 September 2020).
44. FAOSTAT. Available online: http://www.fao.org/faostat/en/#data/QC (accessed on 25 September 2020).
45. Smýkal, P.; Hradilová, I.; Trněný, O.; Brus, J.; Rathore, A.; Bariotakis, M.; Das, R.R.; Bhattacharyya, D.; Richards, C.; Coyne, C.J.; et al. Genomic diversity and macroecology of the crop wild relatives of domesticated pea. *Sci. Rep.* **2017**, *7*, 17384. [CrossRef]
46. Mumtaz, A.S.; Shehadeh, A.; Ellis, T.H.N.; Ambrose, M.J.; Maxted, N. The collection and ecogeography of non-cultivated peas (*Pisum*, L.) from Syria. *PGR Newslet.* **2006**, *146*, 3–8.
47. Blixt, S. Mutation genetics in *Pisum*. *Agri. Hort. Genet.* **1972**, *30*, 1–293.