Supporting Text

Supporting Materials and Methods

The number of hits from *Brassica oleracea* could not be directly converted into transposable element (TE) copy number because the Institute for Genomic Research (TIGR) *B. oleracea* database consists of short reads (on average ~650 bp), and two hits from two different reads could represent different regions of the same element. Thus, the hits from each TBLASTN search can be divided into two types: those covering the entire query (full-length hits) and those that are truncated due to cloning and only contain part of the query (partial hits). Each full-length hit represents one element, whereas statistically each partial hit represents half an element. Let the probability that a particular hit is full length be P_f, the probability that a hit is partial should be 1 - P_f. The copy number of a certain type of element in the current *B. oleracea* database (N_{cp}) can be estimated based on the number of hits (N_{hits}) as:

$$N_{cp} = N_{hits}\left[P_f + \frac{1}{2}\left(1 - P_f\right)\right] \quad \textbf{[1]}$$

or

$$N_{cp} = N_{hits}\frac{1}{2}\left(1 + P_f\right) \quad \textbf{[2]}$$

P_f can be estimated based on the effective query length (L_{eq}) and the length of reads in the database (L_{dr}) as:

$$P_f = \left(L_{dr} - L_{eq}\right)/\left(L_{dr} + L_{eq}\right) \quad \textbf{[3]}$$

Combining Eqs. **1** and **2** and considering the current *B. oleracea* database covers approximately one third of the genome, the total copy (N) in the entire genome should be:

$$N = \frac{3}{2} N_{hits} \left[1 + \left(L_{dr} - L_{eq} \right) / \left(L_{dr} + L_{eq} \right) \right] \quad \textbf{[4]}$$

Thus, the copy number of a certain type of elements in the *B. oleracea* genome can be estimated based on the number of hits and query length by using Eq. **4**.

Eq. **4** contains two correlated variables, effective query length (L_{eq}) and number of hits (N_{hits}). A longer query would yield more hits. However, if Eq. **4** is correct, the total copy number should not change (as it is a constant). This was tested on *B. oleracea* long interspersed nuclear elements (LINEs) as an example. Nine queries, ranging from 70 to 150 aa in length, were derived from the most conserved RT region in *Arabidopsis* LINEs and used in TBLASTN searches against the TIGR *B. oleracea* database. As shown in Fig. 6, when query length increased from 70 to 150 aa, N_{hits} increased from 1,329 to 1,729. However, there was very little variation in N_{cp} values (on average ~1,100). Thus, the total copy number of LINEs in *B. oleracea* was estimated to be ~3,300.

A second test of equation E was performed on a simulated database containing a known number of LINEs, constructed as following. One hundred *Arabidopsis* genomic fragments, each 10 kb long and containing a LINE element, were combined, and the resulting 1 Mb sequence was randomly divided into 1,534 segments. Each segment was 600-700 bp long and the average length was 652 bp, thus resembling the TIGR *B. oleracea* database but containing a known number (100) of LINEs (database available upon request). The same set of queries used in test 1 were then used in TBLASTN searches against the simulated database, and the results are shown in Fig. 6*b*. Although there appeared to be a slight underestimate, the average value of N_{cp} (97) is very close to the actual copy number of LINEs in the database. Taken together, the results from these two tests indicated that it is feasible to use Eq. **4** to estimate TE copy numbers in *B. oleracea* based on the number of TBLASTN hits from the TIGR database.