Methods

We downloaded whole genome sequencing (WGS) assemblies for Homo sapiens (human), Pan troglodytes (chimpanzee), Gorilla gorilla (gorilla), and Chlorocebus sabeus (green monkey) (Table 1). We searched these genomes for hits using the HSAT\_hitFinder program: chromosomal locations with at minimum 75% sequence identity to the HSATII subfamily-specific 24-mers HSATIIA1, HSATIIA2, and HSATIIB (Table 2). We excluded indels from our search criteria to prevent inclusion of hits referencing sequence containing other HSAT subfamilies. The base pair distance between adjacent hits was calculated for each subfamily on every chromosome. Numerous hits from our human WGS data within a small genomic area were grouped together into preliminary loci, and these loci were aggregated into a training dataset.

We generated bootstrap data from the training dataset, and fit a mixed distribution to the distances between hits in the bootstrap dataset. For data less than three standard deviations from the mean, we used a beta distribution fit to the entire dataset. For data greater or equal to three standard deviations above the mean (our model base point) we used a uniform distribution equal to the probability of our model base point. This mixed distribution distribution constituted our hit-distance model (HDM) (Figure 1). We believe that our HDM is a useful way to describe the distribution of the distances between locus hits, since our right-skewed model captures both the high density of low-distances, and the more uncommon but still perhaps biological pertinent instances of larger distances between HSATII hits. The training dataset that is applied to the HDM gives it additional versatility; different models can be generated based upon the profiles of different organisms.

We compared the distances between hits generated from our WGS searches to our human-derived HGM in a kernel-based iterative process. If the cumulative probability of every hit within a kernel exceeded 0.05, or 5%, then we considered the hit at the origin of the kernel to be a candidate for inclusion in a locus. A locus was labeled if enough candidate hits occurred adjacent to one another (Figure 2). We found that using a kernel size of 3, or +/- 3 hits from the origin of each kernel, allowed us to define the edges of biologically relevant loci with enough smoothing to compensate for larger distances within the loci.

**Figure 2**. Loci Identification Pseudo-code

FOR each HSATII hit

probability = beta(distance to previous hit, alpha, beta)

FOR each neighboring hit +/- kernel size

probability = probability + beta(distance to previous hit, alpha, beta)

probability = probability / ((2\*kernel size)+1)

IF probability < 0.05

Include origin as candidate hit

FOR each candidate hit

IF directly adjacent neighbor candidate hit

Include to potential locus list

ELSE

IF potential loci length > minimum number of hits

Define as a locus, add locus to loci list, clear potential loci list

ELSE

Clear potential locus list

RETURN loci list

**Table 1**. WGS Assembly Sources

|  |  |  |
| --- | --- | --- |
| Organism | GeneBank Assembly Accession | WGS Project Number |
| Homo Sapiens  (Human) | GCA\_000002115.2 | [AADD01](http://www.ncbi.nlm.nih.gov/nuccore/AADD00000000.1/) |
| Pan troglodytes  (Chimpanzee) | GCA\_000001515.5 | [AACZ04](http://www.ncbi.nlm.nih.gov/nuccore/AACZ00000000.4/) |
| Gorilla gorilla  (Gorilla) | GCA\_000151905.3 | [CABD03](http://www.ncbi.nlm.nih.gov/nuccore/CABD000000000.3/) |
| Chlorocebus sabeus (Green monkey) | GCA\_000409795.2 | [AQIB01](http://www.ncbi.nlm.nih.gov/nuccore/AQIB00000000.1/) |

**Table 2**. HSATII Subfamily 24-mers (Altemose et al, May 2014)

|  |  |
| --- | --- |
| HSATIIA1 | TTGATTCCATTAGTTTCCATTGGA |
| HSATIIA2 | CATTCGATTCCATTCGATGATAAT |
| HSATIIB | TTCGATTCCATTTGATGATTCCAT |

**Figure 2**. HSATII Loci Identification Protocol

A WGS genome (A) is searched for HSATII hits (B). Hits are curated into a biologically relevant training dataset (C). Bootstrap data is generated from the training dataset (D), which is used to develop the mixed HDM (E). The HDM is used to identify loci of HSATII hits (F) from the original HSATII data.

**D**

**C**

**A**

**B**

**E**

**F**