BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 8 2014, pages 1081–1086
doi:10.1093/bioinformatics/btt727

Sequence analysis

Advance Access publication December 25, 2013

SigHunt: horizontal gene transfer finder optimized for eukaryotic genomes

Kamil S. Jaron[1,*], Jiří C. Moravec[1] and Natália Martínková[1,2,*]

[1]Institute of Biostatistics and Analyses, Masaryk University and [2]Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno, Czech Republic

Associate Editor: John Hancock

ABSTRACT

Motivation: Genomic islands (GIs) are DNA fragments incorporated into a genome through horizontal gene transfer (also called lateral gene transfer), often with functions novel for a given organism. While methods for their detection are well researched in prokaryotes, the complexity of eukaryotic genomes makes direct utilization of these methods unreliable, and so labour-intensive phylogenetic searches are used instead.

Results: We present a surrogate method that investigates nucleotide base composition of the DNA sequence in a eukaryotic genome and identifies putative GIs. We calculate a genomic signature as a vector of tetranucleotide (*4-mer*) frequencies using a sliding window approach. Extending the neighbourhood of the sliding window, we establish a local kernel density estimate of the *4-mer* frequency. We score the number of *4-mer* frequencies in the sliding window that deviate from the credibility interval of their local genomic density using a newly developed discrete interval accumulative score (DIAS). To further improve the effectiveness of DIAS, we select informative *4-mers* in a range of organisms using the tetranucleotide quality score developed herein. We show that the SigHunt method is computationally efficient and able to detect GIs in eukaryotic genomes that represent non-ameliorated integration. Thus, it is suited to scanning for change in organisms with different DNA composition.

Availability and implementation: Source code and scripts freely available for download at http://www.iba.muni.cz/index-en.php?pg=research–data-analysis-tools–sighunt are implemented in C and R and are platform-independent.

Contact: 376090@mail.muni.cz or martinkova@ivb.cz

Received on August 9, 2013; revised on November 18, 2013; accepted on December 9, 2013

1 INTRODUCTION

Horizontal gene transfer (HGT) occurs when a DNA sequence passes between organisms otherwise than by reproductive descent. It results in a relationship of orthologous sequences that is not tree-like and contains reticulations. Notorious examples include antibiotic resistance plasmids transferred between bacterial strains (Freeman, 1951), pathogenicity islands (Friesen *et al.*, 2006), incorporation of retroviruses (Jern and Coffin, 2008) and artificial HGT in the forms of genetically modified organisms (Wolfenbarger and Phifer, 2000). When a horizontally transferred gene becomes fixed in a population, it is termed a genomic island (GI). Relatively frequent HGT occurs between organisms of similar complexity, such as between prokaryotes, but successful HGT between domains and kingdoms is also known. Incorporation of the alien sequence into a recipient genome must be compatible with survival of the cell; it should not, for example, knock out an essential gene. In eukaryotes, distortion of the open reading frame with HGT is less likely due to the sparseness of coding sequences; yet alien genes face molecular biological limitations relating to metabolism in the recipient organism. When genes are transferred from prokaryotes to eukaryotes, the genetic code difference might hinder correct protein translation. In cases of HGT between eukaryotes, incorrect intron splicing would render the gene product altered and potentially dysfunctional. Nevertheless, successful implementation of HGT in a suitable place within the genome could result in expression of the relevant protein. Proteins encoded in a GI that could be expressed in the recipient organism might provide a novel, highly adaptive function (Casacuberta and Gonzlez, 2013; Schönknecht *et al.*, 2013). To find such a GI is to discover exciting information that is often transformative for the given research field.

HGT detection is well studied in prokaryotes, having started from measuring variability of oligonucleotide frequency along the genome (Karlin and Burge, 1995). Eukaryotic genomes, however, are comparatively heterogeneous in their composition and much more extensive. That situation complicates the HGT search. Therefore, those methods developed for prokaryotes fail either due to their inability to handle the sequence heterogeneity or because their computational requirements skyrocket. Researchers studying HGT in eukaryotes use two types of methods: surrogate and comparative. Surrogate methods use nucleotide base composition of the DNA sequence. They have been applied in the form of chaos game representation clustering based on tetranucleotide composition (Mallet *et al.*, 2010). Comparative methods are computationally intensive because they compute phylogenetic comparisons between large numbers of identified genes or they use local alignment comparisons against a reference sequence database. They require prior annotation of a genome, an extensive database of comparable orthologues and computationally intensive phylogenetic analyses. Despite these limitations, the results from comparative methods are considered most reliable, as they enable identification of HGT and the donor organism in the form of testable hypotheses. Therefore, methods have been developed to reduce the candidate dataset for GIs while balancing the numbers of false positives and false negatives (Boc and Makarenkov, 2011;

*To whom correspondence should be addressed.

Mallet *et al.*, 2010; Podell and Gaasterland, 2007). To date, the effort in eukaryotic genomes has been demonstrated consistently to fail on these criteria due to the genome heterogeneity in eukaryotes (Mallet *et al.*, 2010), thus leading to the need for an expanding reference database for comparative analyses.

We show here, however, that genome complexity may be overcome by carefully examining the pattern of genomic signature along a sequence. We use both the selection of tetranucleotides with greatest interspecific variability in their frequencies and local composition shifts that take into account natural variation of tetranucleotide frequency changes along a chromosomal arm to locate regions that differ and might thus represent recent acquisitions via HGT.

2 ALGORITHM

2.1 Calculation of *4-mer* sliding density

The principle of the SigHunt (genomic signature hunter) method is to use a sliding window approach both to detect the HGT and to take into account DNA sequence composition change along chromosomes. The genomic signature is calculated as a frequency vector of tetranucleotides (*4-mer*) within the window of a genomic sequence. Within each sliding window, frequency of every *4-mer* is calculated as

$$F_m(S) = \frac{C_m(S)}{N_S - 3}$$

where $F_m(S)$ is the frequency of the *4-mer m* within a sequence in the sliding window (S), $C_m(S)$ is the count of m in S and N_S is the length of the sequence in the window S. Only fully resolved sites are counted towards $F_m(S)$. Tetranucleotides that contain an unresolved base are omitted. Unresolved sites are skipped, and the length of the window, but not the length of the sequence, is increased by the given number of nucleotides until it reaches 20% of the window size. This is the maximum extension of the window. There are two reasons for choosing oligonucleotide length. First, *4-mers* provide a vector with a number of dimensions sufficient for complex comparisons. Second, the *4-mers* frequency vector is representative even for relatively short sliding windows, which is of interest in cases when short alien genes could be expected.

Along a chromosomal sequence, DNA base composition changes with functional regions, such as coding and non-coding regions, repetitive elements, telomeres, centromeres or other structural elements that stabilize a chromosomal arm. Differences between repetitive and coding regions in particular are prone to variable frequency of a few *4-mers* that could either distort or inform a signal from alien fragments. To account for this, we develop a novel scoring system and introduce a sliding-density concept. This takes into account genomic regions D that are directly adjacent to the sliding window S, offset at the 5′ and 3′ ends of the sequence. Within the long sliding window of D, we calculate a kernel density estimate for each *4-mer*. The measured *4-mer* frequency in the short sliding window of sequence S is tested for whether or not it is located outside of the credibility interval (CI) of the *4-mer* density in D

$$F_m(S) \in \left(0, \Phi_D^{-1}\left(\frac{\alpha}{2}\right)\right) \cup \left(\Phi_D^{-1}\left(1 - \frac{\alpha}{2}\right), 1\right)$$

where Φ_D is a cumulative distribution function of $F_m(S)$ in D and α is a confidence level. Values found to be outside of the CI are scored in three intervals for $\alpha \in \{0.05, 0.025, 0.01\}$, adding 1, 2 and 3, respectively, to the discrete interval accumulative score (DIAS). To avoid autocorrelation in measuring $F_m(S)$ on D, D is selected in such a way that S and a specified number of its surrounding sliding windows (x) are excluded from the *4-mer* density calculation (eye-of-the-storm approach). Thus, we compare the S sequence to its context but not to its immediate surroundings. To compare the sliding window approach to previous genome-wide signature studies, we also show global *4-mer* density.

DIAS measures how many *4-mers* deviate in their frequency from the local background of the genomic sequence and by how much. While this is more stable along a chromosome than any other compositional measure known to us, it is nevertheless sensitive to local changes.

2.2 Selection of informative *4-mers*

To further improve computational speed of the SigHunt method, informative *4-mers* can be selected to reduce the signal-to-noise ratio. Multiple genomes are used to train the procedure for *4-mer* selection based on their intra- and inter-genomic variability. $F_m(S)$ values for all consecutive windows are used to calculate the *4-mer* density in a given chromosome. All chromosomes are used for the training genomes. Informative *4-mers* are selected as those where means of F_m in organisms are distinctive from the overall estimates and within-genome F_m variance is small. We score the *4-mers* using the tetranucleotide quality score (TES) for each *4-mer*

$$TES_m = \sum_{k=1}^{n} \left(K_k - \bar{K}\right)^2 - \sum_{k=1}^{n} (A_k + B_k + D_k) + E$$

$$K = \frac{1}{c} \sum_{i=1}^{c} \mu_i$$

$$A = \sum_{i=1}^{c} \frac{1}{c} (\mu_i - K)^2$$

$$B = \sum_{i=1}^{c} \frac{1}{c} (\sigma_i - \overline{\sigma_c})^2$$

$$D = \bar{\sigma}_c^2$$

$$E = \sum_{k=1}^{n} (K_k - K_e)^2 - (A_e + B_e + D_e)$$

where n is the number of all organisms used for training, c is the number of chromosomes in the given organism, μ_i is mean *4-mer* frequency on the given chromosome $\left[\mu_i = \left(\overline{F_m(S)}\right)_i\right]$, $\bar{K}$ is the average of all mean frequencies of the given *4-mer* on all tested chromosomes in all organisms within the dataset, σ denotes respective variances and e is the estimated organism intended for the SigHunt search.

Using TES, we are interested to learn the extent to which *4-mer* frequencies vary between organisms with respect to their variability within a genome. To achieve this, K estimates average frequency of a *4-mer* that is found in the given organism. It is calculated as a mean of means to avoid weighting of the value according to the number and size of chromosomes. A sums

squared differences between each typical *4-mer* frequency on a chromosome compared with the whole genome; *B* similarly penalizes *4-mers* that have deviant frequency variance in a chromosome compared with the background genome. By including the *D* component into the equation, we ensure that the frequency variance in an organism is small to facilitate finding and interpreting *4-mer* frequencies outside of the confidence interval of their density. *E* provides a measure that would stress usefulness of the given *4-mer* for discrimination of the home sequence. TES increases where *4-mer* frequencies differ between organisms, they are stable for a given organism and they exhibit little variation within a genome. We demonstrate below that *4-mers* with high TES scores will be informative in recognizing putative HGT. GIs identified with SigHunt should subsequently be verified using comparative methods (Fig. 1).

3 METHODS

The sensitivity and specificity of the SigHunt method was tested using a receiver operating characteristic (ROC) curve on simulated data. We introduced alien sequences into the recipient sequence by randomly selecting and replacing DNA fragments between 10 organisms with complete genomic sequences. Eukaryotes were represented by the fungi *Aspergillus fumigatus* (Nierman *et al*., 2005), *Encephalitozoon cuniculi* (Katinka *et al*., 2001) and *Saccharomyces cerevisiae* (Goffeau *et al*., 1996); the red alga *Cyanidioschyzon merolae* (Matsuzaki *et al*., 2004); the chromalveolates *Cryptosporidium parvum* (Abrahamsen *et al*., 2004), *Plasmodium falciparum* (Hall *et al*., 2002) and *Thalassiosira pseudonana* (Armbrust *et al*., 2004); and an animal, *Drosophila melanogaster* (Adams *et al*., 2000). SigHunt's performance in prokaryotes was demonstrated in *Buchnera* sp. (Shigenobu *et al*., 2000) and *Escherichia coli* (Welch *et al*., 2002). In each of 500 replicates of the procedure, we replaced three genomic fragments with alien DNA from other organisms. The length of each introduced fragment was 2, 5 and 15 kb in each chromosome, where the origin of each fragment was randomly drawn from the pool of analysed organisms and chromosomes. We scored the sequence according to DIAS for 5 kb windows and 1 kb sliding windows. Those regions with known alien sequences were scored as the only GIs present. We calculated the area under the curve (AUC) from the ROC curve in R (R Development Core Team, 2011; Robin *et al*., 2011; Sing *et al*., 2005) and estimated the optimal threshold while maximizing AUC. The same analysis was conducted using INDeGeNIUS, which is a recent surrogate method that uses oligonucleotide frequencies (Shrivastava *et al*., 2010), and with Alien_Hunter, which uses interpolated variable order motifs of DNA composition (Vernikos and Parkhill, 2006).

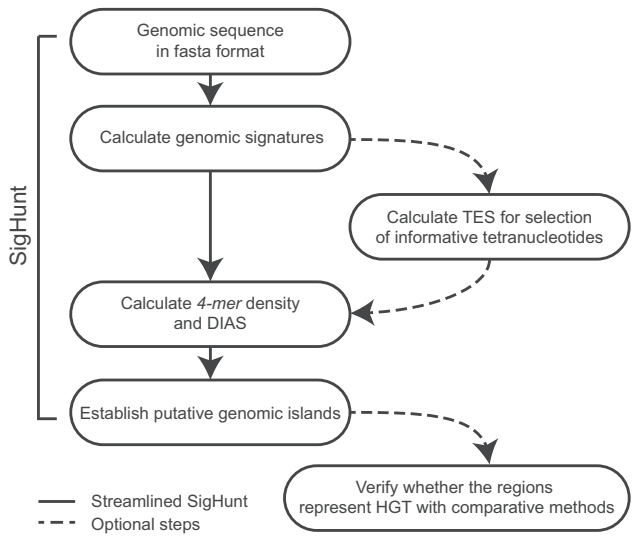

Fig. 1. Flow diagram of GIs analyses using the SigHunt method

3.1 Case studies

To test the SigHunt method on real biological data, we used genomic sequences of *Aspergillus*, *Cryptosporidium* and *Saccharomyces*, as listed above, and added genomic sequences not yet assembled to chromosomes for organisms with known GIs. The latter were the red algae *Galdieria sulphuraria*, wherein the horizontally transferred genes provide multiple environmental adaptations (Schönknecht *et al*., 2013), and the fungus *Pyrenophora tritici-repentis*, which recently acquired a pathogenicity island (Friesen *et al*., 2006) and had additional proteins originating from HGT (Sun *et al*., 2013). We used organisms where GIs had been identified previously and their location was specific in the available genomic sequence (Hall *et al*., 2005; Huang *et al*., 2004; Mallet *et al*., 2010; Schönknecht *et al*., 2013; Sun *et al*., 2013). Contigs at least 200 kb were used from these organisms. This enabled us to cross-check SigHunt against previous studies and thereby to demonstrate its utility. The optimal threshold value for the DIAS as tested with ROC on simulated data was 6.04, and we relaxed this value further to account for sequence amelioration. The cut-off value used here was DIAS $\geq$ 5. Two windows adjacent to the previously identified GI were assessed to compensate for the fact that most GIs were identified as a coding gene sequence and the transferred region could likely include flanking regions (Friesen *et al*., 2006).

4 RESULTS

We estimated *4-mer* density and its variance in the 10 reference genomes. Comparing the *4-mers* using TES, we selected the 16 most informative *4-mers*. These were used for all subsequent analyses.

4.1 Sensitivity and specificity of SigHunt

SigHunt showed average AUC values equal to 0.77 for global density, 0.72 for sliding density and 0.77 for the eye-of-the-storm approach, meaning that sensitivity and specificity of detecting GIs in simulated data were high (Table 1). We assigned a GI only where it had been artificially introduced, and the remaining home sequence still contained its natural GIs, which were disregarded for the purpose of this test (Table 2). This could have lowered the performance indicators. The analyses of individual chromosomes required 8–60 min to calculate DIAS for all three approaches presented here. INDeGeNIUS and Alien_Hunter performed in a similar way with respect to their ability to correctly score the introduced GIs, and no differences between the methods were significant. INDeGeNIUS showed the highest values of AUC from the tested methods. However, those analyses took 20 min–20 h per chromosome, and *Drosophila* could not be analysed due to extensive memory requirements. The speed of Alien_Hunter was 20–120 min per chromosome.

4.2 HGT detection with SigHunt

We estimated that the studied genomes exhibit regions with deviant genomic signature that could be considered alien. These provide a genome-wide assessment of candidate regions

Table 1. Average area under a ROC curve for SigHunt, INDeGeNIUS and Alien_Hunter analyses on 500 random replacements of three GIs into reference chromosomal sequences from the model organisms

Organism	Global density	Sliding density	Eye of the storm	INDeGeNIUS	Alien_Hunter
Fungi					
Aspergillus	0.75 (0.11)	0.72 (0.11)	0.75 (0.12)	0.84 (0.07)	0.65 (0.17)
Encephalitozoon	0.71 (0.09)	0.70 (0.08)	0.74 (0.10)	0.88 (0.07)	0.88 (0.11)
Saccharomyces	0.81 (0.07)	0.72 (0.07)	0.78 (0.06)	0.90 (0.08)	0.83 (0.12)
Red alga					
Cyanidioschyzon	0.83 (0.08)	0.81 (0.07)	0.86 (0.07)	0.91 (0.07)	0.84 (0.15)
Animal					
Drosophila	0.74 (0.11)	0.68 (0.09)	0.72 (0.09)	n/a	0.75 (0.13)
Chromalveolates					
Cryptosporidium	0.77 (0.07)	0.68 (0.08)	0.74 (0.07)	0.86 (0.07)	0.78 (0.17)
Plasmodium	0.67 (0.12)	0.63 (0.09)	0.70 (0.12)	0.87 (0.07)	0.82 (0.17)
Thalassiosira	0.87 (0.09)	0.81 (0.07)	0.85 (0.08)	0.89 (0.06)	0.78 (0.18)
Prokaryotes					
Buchnera	0.83 (0.06)	0.73 (0.07)	0.81 (0.06)	0.91 (0.06)	0.85 (0.16)
Escherichia	0.76 (0.09)	0.68 (0.09)	0.72 (0.09)	0.85 (0.07)	0.82 (0.16)

Note: Standard deviation is given in parentheses.
n/a, not available.

Table 2. Number of correctly assigned GIs in model organisms from those identified previously as retrieved by SigHunt eye-of-the-storm variant, INDeGeNIUS and Alien_Hunter

Organism	Number of chromosomes or contigs[a]	Sequence length (Mb)	Previously established GIs	SigHunt	INDeGeNIUS	Alien_ Hunter	References
Fungi							
Aspergillus	8	29.4	189	150	189	54	(Mallet *et al*., 2010)
Pyrenophora	19[a]	33.9	17[b]	6	11	0	(Friesen *et al*., 2006; Sun *et al*., 2013)
Saccharomyces	16	12	10[b]	5	2	5	(Hall *et al*., 2005)
Red algae							
Galdieria	17[a]	4.4	79[b,c]	48	15	33	(Schönknecht *et al*., 2013)
Chromalveolates							
Cryptosporidium	8	9.1	30[b]	9	12	11	(Huang *et al*., 2004)

Note: In SigHunt, a selection of 16 *4-mers* identified by TES was used. DIAS ≥ 5 was used as a cut-off value and two windows adjacent to the island borders were considered. Sequence length = size of the analysed genomic sequence.
[a]Number of assessed contigs.
[b]In annotated genes.
[c]GIs found in 13.7 Mb of the genomic sequence, but only the longest contigs were analysed here.

for HGT. We searched for consistency of specific sequences in *Aspergillus*, *Saccharomyces*, *Pyrenophora*, *Galdieria* and *Cryptosporidium* between SigHunt and published results. Compared with the extent of HGT identified in previous studies that used predominantly comparative methods on annotated genes, SigHunt found from 30% (*Cryptosporidium*) to 80% (*Aspergillus*) of previously identified GIs (Table 2). In *Aspergillus*, we were able to find the majority of published GIs as per Mallet *et al*. (2010), which can be expected given that those authors used a surrogate method verified with phylogenetic comparison to localize GIs. We identified 5 of 10 putative GIs recognized in *Saccharomyces* by Hall *et al*. (2005). The missed GIs were protein-coding genes <1.2 kb, for which our chosen window size might not be optimal. In *Pyrenophora*, we searched for 17 genes and 6 were retrieved by SigHunt. HGT in *Galdieria* can be attributed to ~9% of the genomic sequence in the longest scaffolds, which we analysed (Schönknecht *et al*., 2013). SigHunt was able to recognize the genomic signatures of the published GIs as being alien in 61% of the cases. In *Cryptosporidium*, we found 9 of 30 previously identified GIs (Huang *et al*., 2004). As in previous cases with low success, the GIs in *Cryptosporidium* that were missed consisted predominantly of short genes. In these cases, experimentation with window size would be beneficial. INDeGeNIUS found more known GIs in *Aspergillus*, *Pyrenophora* and *Cryptosporidium*, but the analyses took an order of magnitude longer than in SigHunt. SigHunt outperformed Alien_Hunter in recognizing the previously established HGT events in most tested organisms. The exception

was *Cryptosporidium*, where Alien_Hunter found 11 of 30 GIs (Table 2).

5 DISCUSSION

We present here a tool for identifying genomic regions as candidates for HGT assessment in eukaryotes. To our knowledge, this is the first surrogate method primarily optimized for eukaryotic genomes. It detects non-ameliorated HGT in large genomic sequences, it is computationally efficient and its implementation provides step-wise user access to results that enables data exploration and analytical optimization.

We demonstrated good success in using SigHunt to find introduced GIs across kingdoms and GIs in real genomic sequences (particularly in some fungi). Considering from a biological perspective reproduction within this group, HGT might be more common in fungi than in other eukaryotes (Rosewich and Kistler, 2000). With HGT events being relatively common within the group, one could speculate that some of these will be non-ameliorated and thus easily detectable using surrogate methods. The further example of *Galdieria*, within which GIs were plentiful across the genome, seems to corroborate this.

5.1 SigHunt's advantages

The advantage of SigHunt lies in its utilization of informative *4-mers*. Selecting only parts of the genomic signature that are most informative according to TES reduces noise in the data and computational requirements, thereby speeding up the analysis. Computational demands are not negligible in eukaryotic genomics. For example, INDeGeNIUS analysis of the largest chromosome in *Drosophila* would require ≈70 TB of memory, which is beyond the capacity available to many researchers, including our group, and the current version of the program does not allow for changes in memory use. By contrast, SigHunt analysed the same problem using 900 MB of maximum allocated memory. At a given time, SigHunt stores in memory only that chromosome sequence needed to calculate the signatures, or, once the corresponding signatures are calculated and the chromosome sequence deleted from memory, the signatures and densities themselves. Such orderly memory utilization reduces the memory requirements and thus allows computations of even large datasets on regular office computers.

Contrary to other recent methods that use a genomic signature for HGT detection (Elhai *et al*., 2012; Shrivastava *et al*., 2010; but see Mallet *et al*., 2010), SigHunt does not assume a point estimate of the genomic signature. Instead, it uses a density distribution and thus acknowledges the natural variability of DNA composition and distinguishes only those regions that deviate from the broad 'norm' for an organism. We recognize that the density distributions must contain tails even for home signatures. Due to this, DIAS measures accumulations of deviant *4-mer* frequencies rather than their mere occurrence. With sliding density and its eye-of-the-storm variant that avoids autocorrelation, SigHunt is able to find putative GIs in any region of the chromosome. This includes regions rich in repetitive DNA because SigHunt assumes a differing genomic signature typical for a genomic region rather than for the whole chromosome.

SigHunt makes it unnecessary to have knowledge as to the exact position on a chromosome of the examined sequence (Podell and Gaasterland, 2007). It can successfully analyse unassembled genomes, provided that the supercontigs are sufficiently long. It also does not require information about gene locations. By filtering nucleotide positions in a sequence that are not fully resolved, we limit the amount of information while increasing the accuracy. In case of a long eukaryotic genome, a trade-off in favour of accuracy is paramount for SigHunt.

5.2 SigHunt's disadvantages

Unfortunately, SigHunt is not a universal black box solution for all HGT problems. Its very principle denies universality, as it rises and falls on the assumption that there are differences in oligonucleotide frequencies between organisms (Karlin and Burge, 1995). This is not always sufficiently true, as shown by our analyses on manipulated and real data. Some random islands were undifferentiated from the home signature. For others, the GI size in real datasets might have been too small to accurately estimate the *4-mer* frequency density for the DIAS calculation. In addition to reasons of there being similar genomic signatures among the organisms involved, SigHunt is prone to false negatives due to amelioration over time of the compositional bias in the horizontally transferred region compared with the host genome. The false-positive rate might also be increased. In regions with strong selection bias and functional restrictions, home signature might vary locally. The extent to which this is the case remains to be tested.

SigHunt expands the search for GIs across the genome without annotation limitations, yet it is able to guide the comparative search more effectively than do other similar methods. We have shown that SigHunt provides a fair basis of target regions for comparative assessment that consists of true GIs as confirmed by phylogenetic analyses in the published data (Friesen *et al*., 2006; Hall *et al*., 2005; Huang *et al*., 2004; Mallet *et al*., 2010; Schönknecht *et al*., 2013; Sun *et al*., 2013).

5.3 Global density paradox

We claim that the advantage of SigHunt lies in its ability to account for variation of the genomic signature along a chromosomal sequence. Yet, in Table 1, the sensitivity and specificity test shows the highest (albeit not statistically significant) AUC values to be for the global density estimate in six cases. This is probably caused by the fact that we introduced HGT directly between organisms that spanned kingdoms and our islands were thus devoid of any amelioration. In other words, while one would rarely encounter such an event in practice, it is one that is relatively easy to capture by means of genomic signatures. Analysing real biological data would require a more subtle approach. Therefore, both the sliding density and its eye-of-the-storm variant provide room for fine-tuning the method. These are parameterized for window size, sliding window size, sliding-density window size, and eye-of-the-storm size. All these parameters might be optimized to further improve SigHunt for any specific target organism. On the other hand, the global density reached the height of its performance in this study. We nevertheless consider global-density DIAS calculation a useful approach in view

of the fact that it is computationally effective and has low memory demands.

5.4 Usefulness of SigHunt

As a method for investigating genomic signature in complex eukaryotic genomes, SigHunt can provide a rapid analysis tool for ongoing sequencing projects. In particular, it will be sensitive to recent HGT, such as in cases of emergent pathogens that have acquired novel genes. The choice of informative *4-mers* could increase resolution and success for binning of metagenomic DNA fragments (Saeed and Halgamuge, 2009). With the recent finding of bacterial horizontally transferred genes in human tumour cells (Riley *et al*., 2013), SigHunt shows promise to be used in rapid screening of such events in genomic assemblies of specific cell lines. We assume that further research will reveal more fields within which sliding density of DNA composition in eukaryotic genomes and the selection of informative oligonucleotides will prove advantageous.

ACKNOWLEDGEMENT

The bioinformatic analyses were conducted at the MetaCentrum computing facility of Masaryk University.

Funding: Czech Science Foundation (grant number P506/12/1064). The access to the MetaCentrum was funded by the Ministry of Education, Youth, and Sports of the Czech Republic (grant number LM2010005).

Conflict of Interest: none declared.

REFERENCES

Abrahamsen,M.S. *et al*. (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**, 441–445.

Adams,M.D. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Armbrust,E.V. *et al*. (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science*, **306**, 79–86.

Boc,A. and Makarenkov,V. (2011) Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Res*., **39**, e144.

Casacuberta,E. and Gonzlez,J. (2013) The impact of transposable elements in environmental adaptation. *Mol. Ecol*., **22**, 1503–1517.

Elhai,J. *et al*. (2012) Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. *BMC Genomics*, **13**, 245.

Freeman,V.J. (1951) Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J. Bacteriol*., **61**, 675.

Friesen,T.L. *et al*. (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat. Genet*., **38**, 953–956.

Goffeau,A. *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546–567.

Hall,C. *et al*. (2005) Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot. Cell*, **4**, 1102–1115.

Hall,N. *et al*. (2002) Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature*, **419**, 527–531.

Huang,J. *et al*. (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biol*., **5**, R88.

Jern,P. and Coffin,J.M. (2008) Effects of retroviruses on host genome function. *Annu. Rev. Genet*., **42**, 709–732.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*., **11**, 283–290.

Katinka,M.D. *et al*. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*, **414**, 450–453.

Mallet,L. *et al*. (2010) Whole genome evaluation of horizontal transfers in the pathogenic fungus *Aspergillus fumigatus*. *BMC Genomics*, **11**, 171.

Matsuzaki,M. *et al*. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10d. *Nature*, **428**, 653–657.

Nierman,W.C. *et al*. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, **438**, 1151–1156.

Podell,S. and Gaasterland,T. (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol*., **8**, R16.

R Development Core Team. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riley,D.R. *et al*. (2013) Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. *PLoS Comput. Biol*., **9**, e1003107.

Robin,X. *et al*. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.

Rosewich,U.L. and Kistler,H.C. (2000) Role of horizontal gene transfer in the evolution of fungi. *Annu. Rev. Phytopathol*., **38**, 325–363.

Saeed,I. and Halgamuge,S. (2009) The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics*, **10**, S10.

Schönknecht,G. *et al*. (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, **339**, 1207–1210.

Shigenobu,S. *et al*. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. aps. *Nature*, **407**, 81–86.

Shrivastava,S. *et al*. (2010) INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *J. Biosci*., **35**, 351–364.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Sun,B.F. *et al*. (2013) Multiple interkingdom horizontal gene transfers in *Pyrenophora* and closely related species and their contributions to phytopathogenic lifestyles. *PLoS One*, **8**, e60029.

Vernikos,G.S. and Parkhill,J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, **22**, 2196–2203.

Welch,R.A. *et al*. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.

Wolfenbarger,L.L. and Phifer,P.R. (2000) The ecological risks and benefits of genetically engineered plants. *Science*, **290**, 2088–2093.