Genetics and Molecular Biology, 35, 4 (suppl), 1078-1084 (2012)

www.sbg.org.br

Review Article

Methods for detection of horizontal transfer of transposable elements in complete genomes

Marcos Oliveira de Carvalho[1] and Elgion L.S. Loreto[1,2]

[1]*Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.*
[2]*Departamento de Biologia, Universidade Federal de Santa Maria, Santa Maria, RS, Brazil.*

Abstract

Recent advances in nucleic acid sequencing technology are creating a diverse landscape for the analysis of horizontal transfer in complete genomes. Previously limited to prokaryotes, the availability of complete genomes from close eukaryotic species presents an opportunity to validate hypotheses about the patterns of evolution and mechanisms that drive horizontal transfer. Many of those methods can be transported from methods previously used in prokaryotic genomes, as the assumptions for horizontal transfer can be interpreted as the same. Some methods, however, require a complete adaptation, while others need refinements in sensitivity and specificity to deal with the huge datasets generated from next-generation sequencing technologies. Here we list the types of methods used for horizontal transfer detection, as well as theirs strengths and weakness.

Keywords: transposable element, horizontal transfer, genome, computational analysis, evolution.

Horizontal Gene Transfer and its Detection

Horizontal transfer can be defined as the exchange of genetic material between species without the aid of any form of sexual mechanism (Gilbert *et al.*, 2009). This phenomenon is widely documented in prokaryote species and plays a special role in prokaryotic and eukaryotic evolution and adaptation (Biémont and Vieira, 2006; Silva *et al.*, 2004). Prokaryotes usually perform horizontal transfer of genetic material through Type IV secretion systems (Juhas *et al.*, 2008), conjugation (Weinert *et al.*, 2009), transformation (Fall *et al.*, 2007) or transduction (Zaneveld *et al.*, 2008), all being biological mechanisms that ease the exchange of DNA. There are many cases of horizontal transfer documented for eukaryotic species as well (Keeling and Palmer, 2008), although with a lower frequency than in prokaryotes, due to the lack of a well defined biological process for the exchange of genetic material between eukaryotic lineages without a sexual mechanism.

However, genetic elements like transposable elements (TEs) are capable of encoding enzymes that permit the integration of their DNA sequences into specific regions of the genome (Schaack *et al.*, 2010). This characteristic enables transposable elements to insert themselves into different hosts, without the aid of special biological mechanism for DNA transfer. Several processes have been suggested in horizontal transfer of TEs in eukaryotes. For example, direct transfer of episomes (O'Brochta *et al.*, 2009) or some retrotransposons capable of generating viral particles (Kim *et al.*, 1994). Also, transposable elements can putatively explore events like virus infections (Dupuy *et al.*, 2011) and parasite mediated transfer (Gilbert *et al.*, 2010) from one host to another.

Horizontal acquisition of genes is an important force in evolution, with examples of influence in the evolutionary history of many species (Gurudatta and Corces, 2009; Zhou and Wang, 2008). Of special importance is the acquisition of pathogenicity islands in prokaryotic species (Gal-Mor and Finlay, 2006), as those cases of horizontal transfer can promote the development of new pathogenic bacteria strains. Thus, the accessibility of accurate and precise methods to quickly identify horizontally transferred genes is of crucial significance to the study and broad comprehension of the processes that have shaped genomes. However the first challenge a researcher faces when identifying horizontal transfer events in whole genomes, and specifically in the case of transposable element horizontal transfer analysis, is the annotation of TE sequences (Flutre *et al.*, 2011). Currently many methodologies are applied for TE annotation, however, there are no established ontologies like the Gene Ontology (Ashburner *et al.*, 2000) that greatly help

Send correspondence to Marcos Oliveira de Carvalho. Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil. E-mail: marcos.carvalho@gmail.com.

the automated curation and annotation of large-scale datasets (Plessis *et al.*, 2011).

Before the large availability of genome sequencing, the support for horizontal transfer detection was sparse, relying on standard molecular biology hybridization protocols or a small number of sequences (Clark *et al.*, 2002; Daniels *et al.*, 1990). As genome sequencing technology progressed, the era of complete genomes brought a great opportunity for large-scale analysis of horizontal transfer events, where a big volume of data could be explored to develop new theories and analysis about the nature of this intricate evolutionary process (Ragan, 2009).

Initially, restrained in its ability to compute and process large quantities of data from full genome projects, indirect methods of detection for horizontal transfer were developed (Ragan, 2001). These methods relied on the detection of differential compositional patterns from nucleotide sequences (Putonti *et al*., 2006). They were based on the premise that sequences that have been transferred horizontally from a distinct donor to a new host would harbor a different nucleotide composition (Rocha *et al.*, 2006). This would allow one to infer horizontal transfer in a certain region of the genome if its nucleotide composition statistically differs from the genome's average nucleotide composition. These methods, called surrogate methods, proved to be useful in identifying pathogenicity islands in prokaryotic genomes. However, they cannot be widely applied without a careful analysis of the results, since they tend to produce a large number of false positives and false negatives (Azad and Lawrence, 2011). The availability of comparative methods that use phylogenetic inference and tree incongruence greatly expanded the set of tools for horizontal transfer, adding specificity but also increasing the computational costs for each analysis (Lyubetsky and V'yugin, 2003). With the recently developed next-generation sequencing technologies and the low cost for complete genome sequencing genome sequence data is even more readily available. These data can now be used for a data-driven discovery of horizontal transfer, where the phenomenon can be modeled through the use of massive datasets bonded with exploratory and machine learning techniques.

Next, we will present the methods discussed above in more detail, also including a section on large-scale data visualization for horizontal transfer analysis.

Horizontal Transfer Analysis in Whole Genomes

Surrogate methods for detection of horizontal transfer events can be defined as those that do not employ the construction of phylogenetic trees or other direct phylogenetic analysis (Ragan, 2001). Although surrogate methods present advantages in large-scale analysis of horizontal transfer, they have major drawbacks, especially the high rate of false positives that this kind of method can return. Errors of this type can be traced to the breaking of assumptions of the method. It has been demonstrated that intragenomic variation of codon bias can be large enough to be confounded with true significant variation that otherwise would be attributed to horizontal transfer (Guindon and Perrière, 2001). Moreover, it has been suggested that both codon bias and base compositional indexes, common indexes used in surrogate methods, are poor indicators of horizontal transfer (Koski *et al.*, 2001). Surrogate methods can, however, be used when a quick scan for horizontal transfer is needed, but insufficient genome data are available for comparison. This is the case for a new genome with no closely related species genomes sequenced. In this particular case, comparative methods could not be applied with a high level of confidence, since the lack of paralogous/orthologous data can lead to mistaken results of horizontal transfer (Capy and Gibert, 2004). One remaining weakness of surrogate methods is related to the power of resolution over the time since the occurrence of the horizontal transfer. Genes that were introgressed will slowly acquire the host genome codon usage and compositional values in a phenomenon called amelioration (Marri and Golding, 2008). This will lead to a masking of compositional differences, reducing the sensitivity of surrogate methods (Becq *et al.*, 2010). It also has been suggested that for successful integration and maintenance of genes in the host genome after a horizontal transfer event, both the recipient genome and the transferred genes must have codon bias compatibility (Medrano-Soto *et al.*, 2004). This is a valid assumption for horizontal transfer of bacterial genes; however, its importance for the successful expression of TE genes remains to be accessed.

Specific software implementations exist for different approaches of horizontal transfer identification via surrogate methods. The Alien-Hunter software (Vernikos and Parkhill, 2006) uses Interpolated Variable Order Motifs (IVOMs) to explore compositional biases for detection of horizontal transfer. This method employs variable order motif distributions to capture more reliably the local composition of a sequence compared with fixed-order methods (Vernikos and Parkhill, 2006). It is assumed that the genome has a reasonably constant background sequence composition, derived from uniform mutational pressure over the complete genome. Thus, atypical sequences are inferred as horizontally transferred if the present window of analysis over the genome induces the current HMM to change from the "typical" to "atypical" state. Clustering of proxy variables from sequence composition has also proved to be an effective alternative to discrete modeling through hidden Markov models. The entropic clustering method (Azad and Lawrence, 2007) uses the Jensen-Shannon divergence measure as a variable for posterior clustering of all genes in the genome in analysis, allowing identification of horizontal transferred sequences based on the dissimilarity of the cluster distributions.

However, it is also possible to use more flexible approaches in the identification of horizontal transfer using nucleotide and codon bias data. This can be achieved by the independent calculation of compositional indexes for each gene in the genome (or in a sliding window fashion), be it nucleotide indexes or codon bias values, and run the statistical analysis in a separate framework. The EMBOSS suite of software contains specific software implementations both for nucleotide indexes and codon bias calculations, and can be easily integrated in complex pipelines. The INCA (Supek and Vlahovicek, 2004) system is a user-friendly graphical software program that allows the calculation of many codon bias parameters, also allowing the determination of sequence clusters via the self-organizing maps machine learning method (Wang *et al.*, 2001). The results of the independent codon bias and nucleotide indexes calculations can then be submitted to statistical analysis using the R system or other statistical frameworks.

Comparative methods rely on the existence of evolutionary related sequence data to identify horizontal transfer of genes in a determined species or group of species. This group of methodologies include phylogenetic and tree analysis, phylogenomic approaches and statistical analysis of phylogenetic indexes.

Local alignment similarity searches provide a quick and relatively inexpensive (in computational terms) way to identify related sequences in different databases. This property was explored to develop a method of horizontal transferred genes where simple assumptions are taken into account (Shi *et al.*, 2005). First, for a group of species, all genes are searched against each other in the different species of the dataset. From the results, only hits with an *e*-value of less than 1e-20 and with the following five hits homologous to the searched taxon are retained. Second, all of its homologs (of the hit selected in the previous step) are from a distant taxon, or the *e*-value of the closest homolog from a distant taxon is significantly lower than the *e*-value of the closest homolog not from the distant taxon (Shi *et al.*, 2005). This method uses the *e*-value as a proxy variable to compare putative horizontally transferred DNA from different genomes, assuming the *e*-value as an indicator of similarity.

The DarkHorse method (Podell and Gaasterland, 2007) also employs local alignment similarity searches as its start point for horizontal transfer detections in genomic scale. However, this methodology uses a specific metric to compare different horizontally transferred gene candidates, instead of the default *e*-value. This metric is called "lineage probability index" or LPI and represents the likelihood that the current gene under search was horizontally transferred taking into consideration the similarity of the gene with distant taxa and also the other genes in the dataset in relation to different taxa.

More recently a method based on phylogenetic distances (Distance Method) was introduced to avoid dealing with the intrinsic bias introduced in the process of local alignment similarity search (Wei *et al.*, 2008). In this methodology all phylogenetic distances from a gene family are calculated from different species, forming a dataset of all-against-all distance pairs. The distances are then analyzed to identify horizontal transfer through the comparative analysis of distances values between pairs of species according to a pre-defined species phylogeny. The method assumes that between pairs of species from the same branch of the three, all genes must have smaller distance values between each other than with other species in the tree. By transversing the tree and comparing all distances for the different pairs of species, the Distance Method can identify putatively transferred genes if the distances between taxa from other branches of the tree are smaller than with species within the same branch.

Comparative methods proved to be more sensitive and specific (Poptsova and Gogarten, 2007). However, it is important to note that to work efficiently, comparative phylogenetic methods need a robust phylogenetic tree from the species under analysis as a reference. If this requirement cannot be met, surrogate methods can be applied in conjunction to provide additional support to the results, although a complete phylogenetic validation should be preferentially used if possible (van Passel *et al.*, 2004).

A comprehensive analysis of multiple eukaryotic genomes can generate many megabytes of data, if not gigabytes. This is also true for the analysis of multiple horizontal transfer events, especially if using comparative methodologies. To tackle the problem of identifying data signals that lead to the identification of horizontal transfer, the use of large scale data analysis is imperative. Data clustering is the assignment of data points into subset classes, where the intraclass similarities are statistically more significant than the interclass similarities. Cluster analysis is thus a form of unsupervised learning, where no prior knowledge is used for the determination of the classes. There are two fundamental types of cluster analysis, one that employs a hierarchical approach and another that employs a partition approach to the data classification. Hierarchical methods find successive clusters by seeding the actual cluster with previous classified data. When no cluster is available, like in the initial round of clustering, each data point is considered an initial cluster. The algorithm then progresses by agglomerating similar data points based on metrics derived from distance measures between the data points. This kind of hierarchical clustering is called bottom-up, where each data point starts as a cluster and is fused with other data points as the cycles of clustering progress. It is also possible to use top-bottom approaches, where the whole dataset is initially considered as a single cluster and is progressively separated into small clusters, where the analysis of each data point identifies its separation based on distance measure metrics.

Hierarchical clustering analysis of large datasets is commonplace in many fields of biology today like expression analysis of microarray and RNA-seq data. However, this technique can be successfully exploited in the analysis of whole genome horizontal transfer datasets if the analysis considers each gene as a data point. In this manner, if a gene has a fixed number of variables associated with it, where such variables are indicators of phylogenetic or evolutionary events, those genes can be grouped through the clustering of one or more of those variables. It is important to consider the variance of the variables used for analysis and also the scale of each variable prior to the use of more than one as a group. If there are significant differences between the variables, a normalization step must be considered before the hierarchical clustering. There are many implementations of clustering analysis in software ranging from complete statistical packages like SPSS or SAS to specific library packages for the R platform. However, much of the analysis in whole genome horizontal transfer analysis needs to be integrated in a pipeline fashion, where flexibility is a positive characteristic in a software implementation of hierarchical clustering. In this form, the most useful software platforms to implement this kind of analysis are programming libraries. The Python programming language is considered one of the most flexible and user-friendly modern languages, with multi-paradigm programming capabilities and also clean syntax. Besides the availability of biology-centered programming libraries like Biopython (Cock *et al.*, 2009), PyCogent (Knight *et al.*, 2007) and Corebio, the Python language has many large-scale data analysis libraries, including for hierarchical clustering. One of the most documented and maintained data analysis library comes from the SciPy project, with the scipy-cluster plug-in. This library has implemented more than twenty clustering methods, including complete linkage, ward clustering and the centroid/UPGMC algorithm. Those can be quickly implemented in a pipeline alongside other statistical methods allowing for great flexibility in the large-scale analysis of genetic horizontal transfer. The R platform also contains a large number of programming libraries with different clustering methods. One advantage of the R platform is the possibility to prototype specific statistical analysis in integrated graphical interfaces like Rkward and R-Studio, before integration in the pipeline. The R libraries can also be conveniently accessed from the Python environment using the RPy library, making the two technologies very suitable for building large pipelines with many integrated complex statistical analyses. To conduct an exploratory analysis using hierarchical clustering it is also useful to have a graphical interface where it is possible to test different kinds of distance measures and normalizations in a sample of the data that is being analyzed. This kind of analysis can be conveniently performed in the HCE-Explorer software. This tool provides an interactive graphical user interface to explore large datasets before and after the hierarchical clustering analysis. Also, there are options to compare different runs of clustering, depending on the comparison of the parameters used for the analysis. This software was initially developed for the analysis of gene expression data; however, any dataset can be loaded if it conforms to the standard data format used by HCE-Explorer. This mainly consists of rows describing the main data point with variable values in the respective columns and can be loaded in the CSV data file format. Although only Windows binaries are available, the HCE-Explorer can run on Unix systems under the Wine platform.

Visualization of large-scale datasets can be a powerful tool to help identify patterns of horizontal transfer in genomic data. Although a young subject, many approaches from other fields of science can be applied to the visualization of horizontal transfer, such as graph analysis. Graph visualization is specially suited for the analysis of large-scale horizontal transfer because of the inherently high connection of data points, namely, the genes in a horizontal transfer analysis. These data points are not expected to behave in a tree-like fashion as observed in tree reconstruction phylogenetic analysis, since the variables accounting for genes that were subjected to horizontal transfer should reproduce the characteristics of transfer from one species to another. In a graph, data points that represent genes that have undergone horizontal transfer should connect more distinctly than data points that have not been horizontally transferred. Using variables like phylogenetic distance, one should expect to have connections between data points from distant species as an indicator of putative horizontal transfer. This kind of analysis can provide instant visual information about the patterns of organization of the genes that have undergone horizontal transmission. This kind of visualization was well employed in the analysis of the horizontal transfer events in the genomes of *Mycoplasma synoviae* and *Mycoplasma gallisepticum,* where a cluster of genes that had undergone horizontal transfer was clearly seen in a graph visualization of genes as data points, and the local sequence similarity score as a graph *e.g.* measure (Vasconcelos *et al.*, 2005). Two specific software programs implement useful methods for large-scale graph visualization. The Phylographer software program is a graphical environment that allows for flexible large-scale graph construction, with a simple data file specification where each gene can be considered as a node and any kind of variable can be attributed as an edge. This flexibility allows for the use of either similarity or phylogenetic variables as *e.g.* connection data. A drawback of the Phylographer software is the need for the TK/TCL platform and a less user-friendly interface than most modern packages for graph visualization. However, large graphs (more than 60,000 nodes) can be built in a matter of hours in a high-end workstation, producing a lightweight interactive interface that allows the manipulation of both nodes and edges and identification of dissimilar graph regions. A more modern system for graph visualization is

the Gephi system, used extensively in the graph analysis community and under active development. The Gephi system is an open-source suite built in the Java language and is capable of multiple graph visualization layouts, including the arbitrary mapping of variables over node and *e.g.* properties. This system allows for more information to be added to the graph than the Phylographer software does. One example would be the mapping of gene length over node color, phylogenetic distance over *e.g.* width and GC content over node color. That kind of graph would allow for a complete picture of three different variables and their relationship with putative horizontal transferred genes, providing useful insights in the development of posterior specific analysis. The Python language also provides many libraries that can be effectively used to integrate large-scale graph visualization with analysis pipelines like the NetworkX and the python-graph libraries.

As a complex biological phenomenon, horizontal transfer is modulated and influenced by a number of variables, many unknown and inaccessible with our current set of tools. This lack of knowledge about the specific factors that drive a specific event can be partially overcome by massive datasets derived from analysis of the phenomenon. These datasets can be created by attaching specific variables to each gene in the analysis, in many related genomes and an integrated search for differential patterns of gene evolution carried out over the complete dataset. If evolutionary assumptions for horizontal transfer like small phylogenetic distance between distance species, low dS rates in genes in relation to the core genes in the complete genome and codon bias (as in surrogate methods) are taken into account, a specific model of horizontal transfer can be derived from an initial exploratory clustering analysis. This initial model can be supplied to supervised machine learning methodologies to identify similar patterns of genes in genomes of related species. This approach has the advantage of the use of real evolutionary signals, represented as the variables attached to each gene in the genome, in a large dataset to build a model that most closely represents horizontal transfer.

Figure 1 presents an organogram representation of methods used for inference of horizontal transfer in complete genomes.

Conclusions

Large-scale horizontal transfer analysis is a recent development, being fueled mainly by recent developments in sequencing technology. With the availability of large datasets of genomic sequences, many hypotheses about horizontal transfer of sequences between distant species could be elucidated, as well as the development and testing of new hypotheses regarding the specific evolutionary patterns of sequences that have undergone horizontal transfer.

The methods employed for such analysis have evolved at the same rate as data has accumulated. However, many of these lack implementation or have implementations that are not user friendly. This means more expense

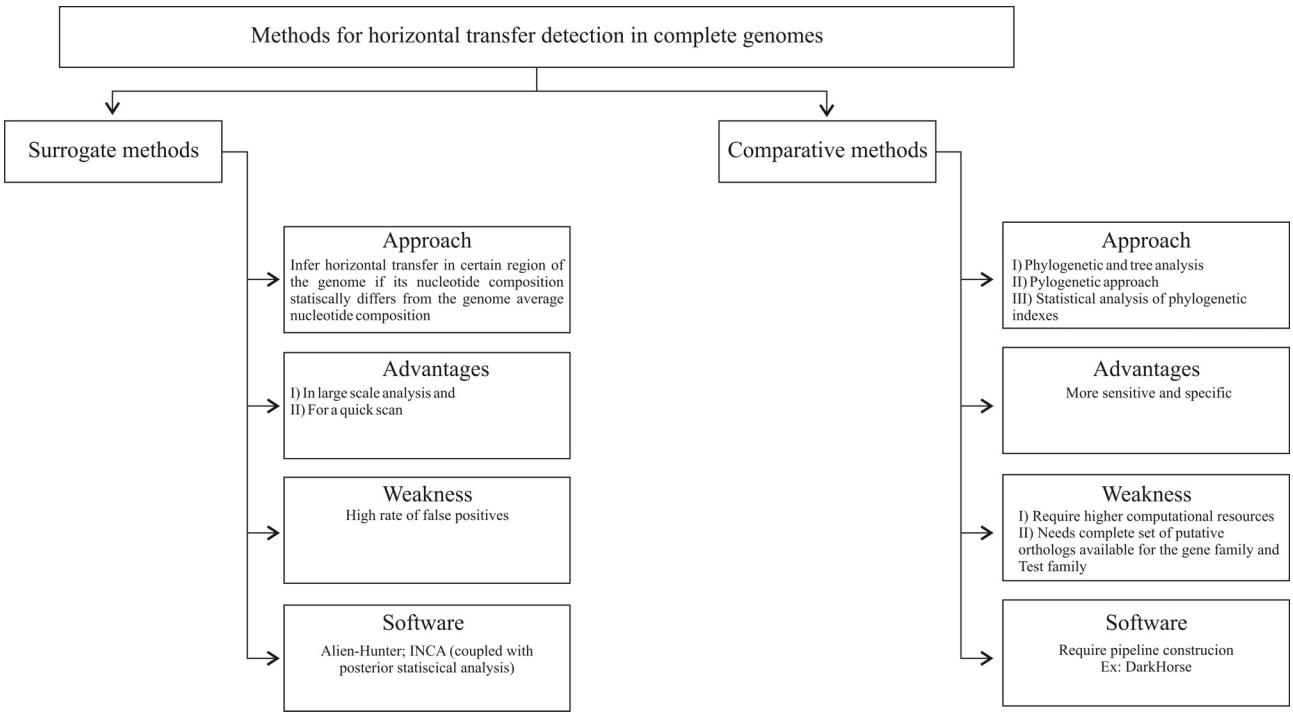

Figure 1 - Organogram representation of methods used for inference of horizontal transfer in complete genomes.

and time spent in the process of analyzing large-scale datasets, a laborious and time consuming endeavor *per se*. Additionally, the lack of a strong community built around the data and methods for analysis of horizontal transfer of transposable elements hinder the development of more advanced tools like common ontologies, essential for an efficient system of data communication. These issues need to be addressed in order to allow effective use of the available data and empower the development of more efficient large-scale analysis of horizontal transfer events.

From the methods presented to identify horizontal transfer, it is clear that surrogate methods should be employed only if no data are suitable for the use of comparative methods. Surrogate methods lag behind comparative methods in terms of resolution and specificity. Comparative methods, however, although more specific and with more power to identify old horizontal transfer events, require higher computational resources for their application on a complete genome scale and increase the complexity of large-scale analysis. Additionally, comparative methods should be applied only when a complete set of putative orthologs are available for the gene family or TE family under analysis.

As a whole, the field of methodologies for large-scale horizontal transfer analysis is in its infancy with many exciting developments under way. Many improvements are needed especially in the automation of TE annotation and in the sensitivity and specificity of the current methods to identify horizontal transfer events. Some of the drawbacks can be addressed with the development of specific machine learning implementations for horizontal transfer detections, as those methodologies can make use of hidden features of evolutionary variables from complete genome datasets to identify the subtle differences between genes inherited vertically or horizontally more accurately.

However, with the growing availability of complete genomes and the growing importance of the understanding of horizontal transfer as a force in the evolution of many species, the methodologies under development should quickly develop into fully established standards. Together with the growing datasets of genomic information, mature methods for horizontal transfer identification could help establish a new way of thinking about the long-term evolution of species.

Acknowledgments

The authors thank CAPES and CNPQ for fellowships. This work was supported by CNPq (grant 473375/2009-5) and PRONEX- FAPERGS (grant 10/0028-7).

References

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-29.

Azad RK and Lawrence JG (2007) Detecting laterally transferred genes: Use of entropic clustering methods and genome position. Nucleic Acids Res 35:4629-4639.

Azad RK and Lawrence JG (2011) Towards more robust methods of alien gene detection. Nucleic Acids Res 39:e56-e56.

Becq J, Churlaud C and Deschavanne P (2010) A benchmark of parametric methods for horizontal transfers detection. PLoS One 5:e9989.

Biémont C and Vieira C (2006) Junk DNA as an evolutionary force. Nature 443:521-524.

Capy P and Gibert P (2004) *Drosophila melanogaster*, *Drosophila simulans*: So similar yet so different. Genetica 120:5-16.

Clark JB, Silva JC and Kidwell MG (2002) Evidence of horizontal transfer of P transposable elements. In: Syvanen M and Kado CI (eds) Horizontal Gene Transfer. Academic Press, San Diego, pp 161-171.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, *et al.* (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422-1423.

Daniels S, Peterson K, Strausbaugh L, Kidwell M and Chovnick A(1990) Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics 124:339-355.

Dupuy C, Periquet G, Serbielle C, Bézier A, Louis F and Drezen J-MM (2011) Transfer of a chromosomal Maverick to endogenous bracovirus in a parasitoid wasp. Genetica 139:489-496.

Fall S, Mercier A, Bertolla F, Calteau A, Gueguen L, Perrière G, Vogel TM and Simonet P (2007) Horizontal gene transfer regulation in Bacteria as a "spandrel" of DNA repair mechanisms. PLoS One 2:e1055.

Flutre T, Duprat E, Feuillet C and Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS ONE 6:e16526.

Gal-Mor O and Finlay BB (2006) Pathogenicity islands: A molecular toolbox for bacterial virulence. Cell Microbiol 8:1707-1719.

Gilbert C, Pace JK and Feschotte C (2009) Horizontal SPINning of transposons. Commun Integr Biol 2:117-119.

Gilbert C, Schaack S, Pace II JK, Brindley PJ and Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature 464:1347-1350.

Guindon S and Perrière G (2001) Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. Mol Biol Evol 18:1838-1840.

Gurudatta B and Corces VG (2009) Chromatin insulators: Lessons from the fly. Brief Funct Genomics Proteomics 8:276-282.

Juhas M, Crook DW and Hood DW (2008) Type IV secretion systems: Tools of bacterial horizontal gene transfer and virulence. Cell Microbiol 10:2377-2386.

Keeling PJ and Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605-618.

Kim AC, Terzian P, Santamaria A, Pélisson N, Prud'homme and Bucheton A (1994) Retroviruses in invertebrates: The gypsy retrotransposon is apparently an infectious retrovirus of

Drosophila melanogaster. Proc Natl Acad Sci USA 91:1285-1289.

Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton B, Eaton M, Hamady M, Lindsay H, Liu Z, *et al*. (2007) PyCogent: A toolkit for making sense from sequence. Genome Biol 8:R171.

Koski LB, Morton RA and Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol 18:404-412.

Lyubetsky VA and V'yugin VV (2003) Methods of horizontal gene transfer determination using phylogenetic data. In silico Biol 3:17-31.

Marri PR and Golding GB (2008) Gene amelioration demonstrated: The journey of nascent genes in bacteria. Genome 51:164-168.

Medrano-Soto A, Moreno-Hagelsieb G, Vinuesa P, Christen AA and Collado-Vides J (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol Biol Evol 21:1884-1894.

O'Brochta DA, Stosic CD, Pilitt K, Subramanian RA, Hice RH and Atkinson PW (2009) Transpositionally active episomal hAT elements. BMC Mol Biol 10:108-120.

Passel M, van Bart A, Pannekoek Y and Ende A (2004) Phylogenetic validation of horizontal gene transfer? Nat Genet 36:1028.

Plessis L, du Skunca N and Dessimoz C (2011) The what, where, how and why of gene ontology - A primer for bioinformaticians. Brief Bioinform 12:723-735.

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

Poptsova MS and Gogarten JP (2007) The power of phylogenetic approaches to detect horizontally transferred genes. BMC Evol Biol 7:e45.

Putonti C, Luo Y, Katili C, Chumakov S, Fox GE, Graur D and Fofanov Y (2006) A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences. Mol Biol Evol 23:1863-1868.

Ragan MA (2009) Thinking laterally about genomes. Genome informatics. Int Conf Genome Informat 23:221-222.

Ragan MA (2001) On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett 201:187-191.

Rocha E, Touchon M and Feil E (2006) Similar compositional biases are caused by very different mutational effects. Genome Res 16:1537-1547.

Schaack S, Gilbert C and Feschotte C (2010) Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537-546.

Shi S-Y, Cai X-H and Ding D (2005) Identification and categorization of horizontally transferred genes in prokaryotic genomes. Acta Biochim Biophys Sinica 37:561-566.

Silva JC, Loreto EL and Clark JB (2004) Factors that affect the horizontal transfer of transposable elements. Curr Issues Mol Biol 6:57-71.

Supek F and Vlahovicek K (2004) INCA: Synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20:2329-2330.

Vasconcelos ATR, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, Almeida DF, Almeida LG, Almeida R, Alves-Filho L, *et al* (2005) Swine and poultry pathogens: The complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*. J Bacteriol 187:5568-5577.

Vernikos GS and Parkhill J (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the Salmonella pathogenicity islands. Bioinformatics 22:2196-2203.

Wang H, Badger J, Kearney P and Li M (2001) Analysis of codon usage patterns of bacterial genomes using the self-organizing map. Mol Biol Evol 18:792-800.

Wei X, Cowen L, Brodley C, Brady A, Sculley D and Slonim D (2008) A distance-based method for detecting horizontal gene transfer in whole genomes. In: Mandoiu I, Sunderraman R and Zelikovsky A (eds) Bioinformatics Research and Applications. Springer, Berlin, pp 26-37.

Weinert LA, Welch JJ and Jiggins FM (2009) Conjugation genes are common throughout the genus Rickettsia and are transmitted horizontally. Proc Biol Sci R Soc Lond 276:3619-3627.

Zaneveld JR, Nemergut DR and Knight R (2008) Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. Microbiology 154:1-15.

Zhou Q and Wang W (2008) On the origin and evolution of new genes - a genomic and experimental perspective. J Genet Genomics 35:639-648.