# PreHGT: A scalable workflow that screens for horizontal gene transfer within and between kingdoms

**Horizontal gene transfer (HGT) is the exchange of DNA between species. It can lead to the acquisition of new gene functions, so finding HGT events can reveal genome novelty. preHGT is a pipeline that uses multiple existing methods to quickly screen for transferred genes.**

Version 5, published Jun 7, 2024. Originally published Jul 19, 2023.

Arcadia Science DOI: 10.57844/arcadia-jfbp-7p11

# Purpose

Horizontal gene transfer (HGT) is the exchange of DNA between an organism and another organism that is not its offspring. It can lead to the rapid acquisition of novel functional traits in the recipient species, leaving distinctive genomic patterns behind in the process. While not all HGT events are maintained in a genome or lead to adaptive benefit, looking for patterns of HGT across a diverse array of organisms is one way we can survey for functional novelty. Many tools exist for computational discovery of HGT events from genome sequencing data, targeting different genomic patterns and with varying sensitivity, specificity, speed, and scalability. We designed the preHGT pipeline to be a flexible and rapid tool for pre-screening genomes for HGT events. Our goal was to create a pipeline to screen for putative HGT events in as many genomes as are publicly available, or that become available in the future. We wanted an approach that could successfully screen eukaryotic, bacterial, and archaeal genomes and that could screen for transfer events between closely or distantly related species.

The preHGT pipeline uses multiple existing methods for HGT screening and the elimination of false positives. It quickly produces a candidate list of genes that

researchers can further investigate with more stringent HGT detection methods, different data modalities, or wet lab experimentation.

We hope this pipeline will be useful to researchers interested in exploring HGT in RefSeq or GenBank genomes.

- This pub is part of the **platform effort**, "Software: Useful computing at Arcadia." Visit the platform narrative for more background and context.
- The **preHGT pipeline** is available in this GitHub repository.

# The context

Adaptation and evolutionary innovation often occur through vertical inheritance and gradual evolutionary processes. Lateral transmission of genomic sequences via HGT is a contrasting evolutionary process that occurs between species instead of from parent to offspring. When genes are transferred, HGT can be a source of rapid functional innovation. Not all HGT events lead to adaptation — some may be neutral, detrimental, or may not be maintained by natural selection and are subsequently lost [1]. Nevertheless, HGT has been the underlying mechanism for many functional adaptations [2][3].

HGT occurs across all domains of life with different frequencies and via many different mechanisms [4][5][6]. In bacteria, HGT most frequently occurs via transduction, conjugation, or transformation. As asexual reproducers with dedicated machinery for HGT, horizontal transfer is one of the most prominent mechanisms for quickly generating genetic diversity. This can catalyze rapid evolution and adaptation to different environmental conditions [3]. However, bacteria also combat HGT by degrading foreign DNA with restriction enzymes and CRISPR [7][8]. Although eukaryotes can undergo HGT through transposable elements, hybridization, and viral transfer, the rate of HGT is relatively low compared to bacteria [5]. This is in part due to structural barriers such as the nucleus that impede the transfer of foreign DNA into the recipient's genome. In sexually reproducing eukaryotes, the frequency of successful horizontal transfer is further reduced because foreign genomic material must reach germline cells to be transmitted from parent to offspring [9].

Surprisingly, HGT events leave behind similar signatures in recipient genomes independent of the domain of life in which the transfer event occurred. When a gene is transferred, the gene has a different evolutionary history than that of other genes in the recipient's genome. This manifests in different ways depending on how closely related the donor species is to the recipient species. The transferred gene may conserve the function of the gene in the donor genome, may carry a transfer-associated gene annotation, may be abnormally distributed in the species pangenome, or may deviate from species-specific expectations in GC content or other characteristics [10]. The strength of these signals often depends on how much time has passed since the transfer event. Transferred DNA undergoes a process called amelioration, whereby the sequence accumulates mutations over time and becomes less and less distinguishable from the recipient's genome and more and more different from the donor's genome [10]. Other evolutionary processes can further scramble the strength or clarity of a transfer event signature. For example, if many speciation events occurred since the time of the transfer event, it may be difficult to determine whether a horizontal transfer event occurred or if the incongruent evolutionary history is due to other evolutionary processes such as incomplete lineage sorting [11]. If multiple transfer events of the same gene have occurred, or if there have been gene duplications and losses post-speciation, the evolutionary history of a gene may be even more difficult to disentangle. Lastly, convergent evolution and genome contamination can confound HGT discovery by genome sequence analysis as these processes can leave behind similar genomic signatures as bona fide HGT events [12][13][14].

Given this variation, detecting HGT in genome sequence data can be difficult, or at the very least, may require multiple strategies to find different types of transfer events. Luckily, researchers have developed many computational methods to interrogate the genomic signatures left behind in genome sequence data by HGT in different ways (Table 1). These methods fall into two general categories: parametric and phylogenetic [10].

| Tool | Category | Taxonomic scope | Event scope | Summary |
|---|---|---|---|---|
| **Alien_hunter [15]** | Parametric | Bacteria & archaea | Composition | Interpolated variable order motifs from compositional biases to identify and predict horizontally transferred regions in genomic sequences. |
| **Alienness [16]** | Phylogenetic implicit | All | Kingdom | Measures alien index and HGT score from BLASTp results on a web server. |
| **APP [17]** | Phylogenetic implicit | Bacteria | Pangenome | Alienness by Phyletic Pattern; Phyletic pattern of query gene distribution in closely related genomes. |
| **AnGST [18]** | Phylogenetic explicit | All | All | Analyzer of Gene and Species Trees; Compares gene trees to species trees and identifies discrepancies under a generalized parsimony criterion. |
| **AvP [19]** | Phylogenetic explicit | All | All | Alienness vs Predictor; Finds homologous sequences, produces multiple alignments, and constructs a phylogeny to analyze the topology for HGT. |
| **BLAST2HGT [20]** | Phylogenetic implicit | All | Kingdom | Measures alien index, donor distribution index, and bit score differences from BLASTp results. |
| **DarkHorse [21]** | Phylogenetic implicit | All | Kingdom, sub-kingdom | Measures lineage probability index from BLASTp results. |
| **GeneMates [22]** | Phylogenetic implicit | Bacteria | Pangenome | Network analysis from gene presence-absence and SNP variants. |
| **GIPSy [23]** | Parametric, phylogenetic implicit | Bacteria | Composition | Genomic Island Prediction Software; Predicts genomic islands using features such as abnormal GC content and presence of mobility genes. |
| **HGT-DB [24]** | Parametric | Bacteria & archaea | Composition | A database of potential HGT events detected using deviations in GC content and codon and amino acid usage. |
| **HGT-Finder [25]** | Phylogenetic implicit | All | Sub-kingdom | Measures transfer index from BLASTp results. |

| Tool | Category | Taxonomic scope | Event scope | Summary |
|---|---|---|---|---|
| **HGTector** [26] | Phylogenetic implicit | All | Sub-kingdom | Measures likelihood of HGT from between self and close & distal groups from BLASTp results. |
| **HGTphyloDetect** [27] | Phylogenetic implicit, phylogenetic explicit | All | All | Measures alien index and out_pct from BLASTp results, followed by phylogenetic inference on initial candidates. |
| **HGTree** [28] | Phylogenetic explicit | Bacteria & archaea | All | A database of potential HGT events inferred using tree reconciliation. |
| **Islander** [29] | Parametric | Bacteria | Bacteria | Targeted identification of tDNAs. |
| **IslandHunter** [30] | Parametric | Bacteria | Composition | Predicts genomic islands using features such as abnormal GC content and presence of mobility genes. |
| **IslandPath-DIMOB** [31] | Parametric | Bacteria | Composition | Predicts genomic islands using dinucleotide composition and presence of mobility genes. |
| **IslandPick** [32] | Phylogenetic implicit | Bacteria | Species, Strain | Predicts genomic islands by comparing closely related genomes. |
| **IslandViewer4** [33] | Parametric, phylogenetic implicit | Bacteria & archaea | See other tools | Integrates IslandPick, IslandPath-DIMOB, SIGI-HMM, and Islander |
| **Near HGT** [34] | Phylogenetic implicit | Bacteria | Species, Strain | Measures synteny index and constant relative mutability from comparisons |
| **PGAP-X** [35] | Phylogenetic implicit | Bacteria | Pangenome | Pan-genome Analysis Pipeline; Pangenome gene presence absence |
| **RANGER-DTL** [36] | Phylogenetic explicit | All | All | Rapid ANalysis of Gene family Evolution using Reconciliation-DTL; Reconciles gene and species trees to detect duplications, transfers, and losses. |
| **RecentHGT** [37] | Phylogenetic implicit | Bacteria & archaea | Species, Strain | Expectation maximization algorithm on global protein sequence alignments |
| **RIATA-HGT** [38] | Phylogenetic explicit | All | All | Identifies incongruencies between gene trees and species trees. |

| Tool | Category | Taxonomic scope | Event scope | Summary |
|---|---|---|---|---|
| **SIB** [39] | Parametric | Bacteria & archaea | Species, Strain | Sequential Information Bottleneck; Signals derived from k-mer co-occurrence to identify transferred regions |
| **ShadowCaster** [40] | Parametric, phylogenetic explicit | Bacteria & archaea | Composition | Uses a support vector machine on compositional features to identify candidates and then filters results by assessing ortholog similarity at increasing taxonomic distances. |
| **SigHunt** [41] | Parametric | Eukaryotes | Composition | Sliding window of 4-mer frequencies. |
| **SIGI-HMM** [42] | Parametric | Bacteria & archaea | Composition | Predicts genomic islands using a combination of codon usage bias and hidden Markov models. |
| **T-REX** [43] | Phylogenetic explicit | All | All | Tree-based search for Reticulate Evolution; Incongruities in phylogenetic trees |
| **TF-IDF** [44] | Parametric | Bacteria & archaea | Species, Strain | Term frequency-inverse document frequency to identify unusual sequence features |

Table 1. **Non-exhaustive list of computational tools for HGT discovery**.

**Composition**: Composition different from acceptor genome.

**Pangenome**: Any set of organisms one can reasonably build a pangenome from (clade, species, genus).

**Kingdom**: Cross-kingdom detection, usually by user-defined definition of ingroup and outgroup.

**Sub-kingdom**: Any taxonomic level lower than kingdom and higher than species or strain, usually with decreasing accuracy at higher taxonomic resolution.

Parametric methods analyze the genome of interest to identify regions that deviate from species-specific expectations in GC content, codon usage, amino acid usage, k-mer frequencies, gene annotations, or other characteristics [10]. These methods are fast, but natural differences in genome uniformity can lead to over-prediction and they are often limited to recent transfer events for which amelioration of transferred DNA is limited [10]. Parametric approaches can also be biased by gene length [45][46], so they may be difficult or impossible to use on

genes, which vary in size, as opposed to sliding windows across the genome, which are a consistent length.

Phylogenetic methods detect inconsistencies between gene and species evolution [10]. This category can be further divided into explicit and implicit methods. Explicit methods test alternative evolutionary scenarios using tree-based analysis, while implicit methods rely on implied phylogenetic relationships derived from comparative genomic approaches. Gene-by-gene explicit phylogenetic methods are the gold standard in horizontal gene transfer detection [10][47]. The most robust of these approaches works by formally reconciling gene family tree topologies (where each tip is a protein sequence belonging to a species) with the species tree topology (each tip is a species) under explicit Maximum Likelihood inference for models of gene family duplication, transfer, and loss [48][49]. These methods identify candidate ancestral HGT events while accounting for the confounding impacts of gene duplication and loss on these inferences. Although powerful, these methods require that gene homology is already known and that gene family trees of these homologous sequences have already been inferred. Consequently, these methods are typically ideally suited for focused application to a set of gene families of special interest and thus are less computationally tractable to apply at scale than other HGT prediction methods.

Without *a priori* knowledge about the donor and recipient genomes for horizontally transferred genetic material, it becomes necessary to sample in a taxonomically broad and unbiased manner. In this respect, implicit phylogenetic methods are particularly well suited to hypothesis-free discovery of HGT events, as they scale more readily to hundreds of genomes than do explicit methods. Implicit methods rely on patterns that correlate with evolutionary history to infer HGT. For example, you can use BLAST to identify homologous genes with different taxonomic labels than the query gene, which can be analyzed to find patterns consistent with HGT [19][25][50][51][52]. Similarly, you can use the pangenome — the full complement of genes shared between a set of closely related organisms — to investigate HGT by determining the presence or absence of genes across all genomes [53][54].

Across the HGT literature and tool space, including both parametric and phylogenetic methods, genome contamination is often underappreciated.

Contaminant sequences in genomes can look like HGT events. This has led to rebuttals [14][55] against high-profile papers [56][57] that claimed detection of high fractions of horizontally transferred genes, and may more generally impact the biological interpretation of HGT predictions. At least 0.54% of genomes in GenBank and 0.34% in RefSeq are contaminated [58]. While some methods incorporate careful contamination checks [19], others rely on filtering heuristics [16] or omit them entirely.

## The problem

We sought a scalable computational approach for predicting HGT candidate genes. We wanted the pipeline to be able to screen for HGT events across the tree of life and across taxonomic scopes (from family- to kingdom-level transfers), and to assess the likelihood that a candidate transfer event was instead the result of genome contamination.

As other projects at Arcadia are developing explicit phylogenetic methods for the inference of gene family evolution, we sought a solution that we could use upstream of this tool to produce candidate species lists for further validation, and tried to avoid using trees so as not to duplicate efforts.

## Our solution

We built a pipeline that we're calling "preHGT" to quickly find *pre*liminary *HGT* candidates in genomes with gene predictions. Our approach blends parametric and phylogenetic implicit methods to generate a list of candidate genes that may have been horizontally transferred (Figure 1). The preHGT pipeline uses compositional scans, pangenome inference, and BLAST-based searches. It combines information from these approaches, as well as annotation information, to highlight candidate genes that are more likely to be contamination than HGT. By implementing multiple HGT screens in one pipeline, we aimed to combine approaches that target different signatures of HGT, to provide a more comprehensive HGT screening strategy.

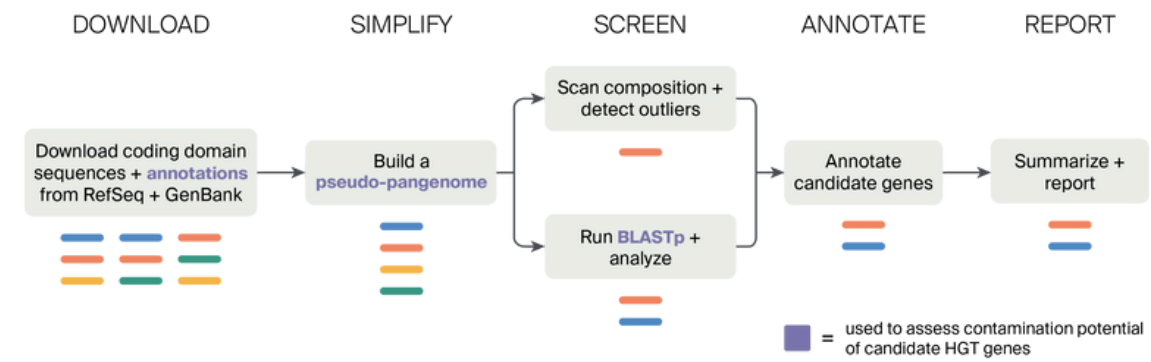


Figure 1. **Conceptual overview of the HGT screening approach implemented in the preHGT pipeline**.

Starting from a genus or genera, preHGT scans GenBank and RefSeq and downloads matching genomes with gene models (coding domain sequences) and annotation files. The coding domain sequences are represented by colored bars, and genes of the same identity are the same color. The pipeline uses the input genes from all genomes of the same genus to build a pseudo-pangenome. These genes are provided as input to two HGT screening methods — compositional scans and BLASTp-based approaches. These steps return HGT candidates that are then annotated to predict function. Information from each of these steps is summarized and returned in a final table.

As we were designing the pipeline, we were concerned about overall run times, especially given that BLAST searches can be computationally expensive. We implemented clustering heuristics at two key places to keep the pipeline fast. First, we clustered the genes in input genomes to reduce the number of genes we investigated for HGT potential. Given our eventual goal of running this pipeline on all publicly available genomes, we wanted to assess the potential for HGT in redundant genes only once. We did this by clustering genes in closely related genomes — those of the same genus — prior to screening for HGT. Second, we clustered the NCBI BLAST non-redundant protein database, reducing its size by over half, to increase the speed of BLAST searches **[59]**.

One of the reasons we were particularly excited to include BLAST in our pipeline was to take advantage of a rich literature of BLAST-based HGT predictor indices (Table 1, Table 2). Many creative and insightful HGT screening methods exist, each with its own strengths. However, these methods are contained in different tools. Since BLAST is the most expensive computational step of our pipeline and none of the methods rely on a clustered BLAST database, we re-implemented them in the preHGT workflow. This consolidation allows HGT screening using a single tool and a single BLAST run (Table 2).

We implemented the pipeline as both a Snakemake [60] and a Nextflow [61] workflow, with software environments controlled by conda or Docker. The modular nature of the workflow will allow us to incorporate additional methods over time.

The preHGT pipeline does not implement any new algorithms for HGT candidate screening. However, the pipeline contributes to this space by:

1. Combining multiple existing HGT screening algorithms in one pipeline.
2. Using pangenome inference on eukaryotic genomes to inform a gene's contamination potential and phyletic distribution, and to reduce compute required to run the pipeline.
3. Reducing the BLAST database size by clustering similar proteins, thereby reducing compute required to run the pipeline and diversifying taxonomic lineages represented in top hits.
4. Providing multiple information sources to help assess an HGT candidate's contamination potential.

# The resource

> The **preHGT Snakemake and Nextflow workflows** are available at this GitHub repository (DOI: 10.5281/zenodo.8169269).

Below we provide an overview of each step in the preHGT pipeline (Figure 2).

1. **Retrieving gene sequences and annotation files**. The pipeline begins with the user providing a genus or genera of interest in a TSV file. The pipeline then scans GenBank [62] and RefSeq [63] for matching genomes and downloads gene models and genome annotation files using ncbi-genome-download. When a genome is available in both GenBank and RefSeq, only the RefSeq version is retained.
2. **Building a pseudo-pangenome**. For each genus, the pipeline then combines genes into a pseudo-pangenome by clustering the nucleotide sequences at 90% length and identity using `mmseqs easy-cluster` [64]. For each

cluster, MMSeq2 selects a single representative sequence by retaining the sequence with the most alignments. The clustered nucleotide sequences are then translated into amino acid sequences using EMBOSS `transeq` [65]. We refer to this as a pseudo-pangenome, and not a pangenome, because we empirically cluster genes based on sequence similarity and not by constructing orthologous groups or by considering the evolutionary history of each sequence [66][67]. We recognize that while this may collapse functionally different paralogs, it is unlikely to obscure patterns of HGT from distant donor genomes; paralogous genes share a common ancestor, so while they may serve different purposes for the organism, at >90% identity only one copy of the gene needs to be screened for HGT potential. Using a pseudo-pangenome is useful in two ways for the pipeline. First, it reduces the number of genes that are investigated which reduces run times. Second, it provides metadata about the gene. Singletons are more likely to be contaminants, and genes that are only present in a subset of genomes may have interesting evolutionary histories (e.g. gene loss).

3. **Screening for HGT candidates**. Using the genes in the pseudo-pangenome, the preHGT pipeline then uses two approaches to screen for HGT candidates.

    - **Compositional scan**. The first approach uses relative amino acid usage to detect proteins with outlying composition. It measures relative amino acid usage using the EMBOSS `pepstats` function [65], produces a distance matrix with the base R function `dist()`, and hierarchically clusters the distance matrix with fastcluster's `hclust` [68]. It detects outliers by cutting the resultant tree with `height/1.5` and retaining any cluster that contains fewer than 0.1% of the pseudo-pangenome size. Relative amino acid usage is the frequency that each amino acid is used in each gene, normalized by the total number of amino acids in that gene. For example, if alanine is used 27 times in a protein that is 100 amino acids long, the relative usage would be 27%. Relative amino acid usage is generally conserved across a genome and reflects an organism's environment [69]. We tried many compositional metrics such as tetranucleotide frequency, GC content, and codon usage. However, we found that outlying proteins were driven by abnormal length for all metrics other than relative amino acid usage. Given that this is a reference-free approach, genes returned by this screening method do not have accompanying donor species predictions, which makes interpretation more challenging. Aberrant relative amino acid usage can also arise from mechanisms other than HGT and this method does not distinguish between potential sources.

- **BLASTp scan**. The second approach uses BLASTp to identify homologous proteins. All genes in the pseudo-pangenome are BLASTed against a clustered version of NCBI's nr database (90% length, 90% identity) [59] using DIAMOND `blastp` [70]. The pipeline then adds lineage information to the BLASTp search using dplyr, dbplyr, and RSQLite [71]. It scans these results for signatures of transfer events using multiple, previously published algorithms (Table 2) [19][25][50][51][52]. One modification we made throughout is using length-corrected bit scores output by DIAMOND `blastp` instead of raw bit scores. Bit scores are sensitive to gene length, so using corrected bit scores reduces biases associated with gene length in HGT screening [72]. The choice of database will dramatically impact the results produced by this screen. We chose to use a clustered version of the NCBI nr database [59] both to make the BLASTp step faster and to ensure the results contain a variety of taxonomic lineages in cases where many near and distant homologs exist. Using this database, combined with our methods of choice (Table 2), the preHGT pipeline screens for HGT events that occur in seven domains of NCBI's taxonomy: bacteria, archaea, fungi, plants, metazoa, other eukaryotes, and viruses ("kingdom" taxonomic resolution). It will also screen for HGT events between lineages that are in the same domain as the query genus but are different up to the family level from that genus ("sub-kingdom" taxonomic resolution).

| Index | Tool | Taxonomic resolution | Data used | Calculated by |
|---|---|---|---|---|
| Aggregate hit support+ [19] | AvP | Kingdom | All bit scores | Subtracting the sum of normalized bit scores in the donor group from the sum of normalized bit scores in the acceptor group. |
| Alien index [50] | NA | Kingdom | Minimum e-value | Subtracting the transformed e-value of the best donor hit from the transformed e-value of the best non-self acceptor hit. |
| HGT score [51] | NA | Kingdom | Maximum bit score | Subtracting the best non-self acceptor hit bit score from the best donor hit bit score and normalizing this value. |
| Donor distribution index [52] | NA | Kingdom | Number of hits per kingdom | Measuring the dispersion query homologs across groups by determining the number of hits per kingdom against the total number of possible kingdoms. |
| Gini coefficient | NA | Kingdom | Number of hits per kingdom | Measuring inequality among values of a distribution, where values are the number of BLAST hits observed for each kingdom. |
| Entropy | NA | Kingdom | Number of hits per kingdom | Measuring disorder among values, where values are the number of BLAST hits observed for each kingdom. |
| Transfer index [25] | HGT-Finder | Kingdom, Sub-kingdom | All bit scores | Considering taxonomic distances between query and hit, bit score ratios, and rank and total number of BLAST hits. |

**Table 2**. **Algorithms that parse BLASTp results to predict HGT candidates**.

*NA: Not applicable.

+Aggregate hit support is calculated by subtracting the sum of all normalized BLAST bit scores for all hits in an in-group from an out-group. We use a different normalization equation than the original method, which leads to different results.

1. **Annotation**. We then annotate the HGT candidates. For each candidate HGT amino acid sequence, we use two different approaches for ortholog annotation. First, the pipeline uses KofamScan for KEGG ortholog annotation [73]. Next, the pipeline uses HMMER3 `hmmscan` to assign annotations to HGT candidates. `hmmscan` compares each HGT candidate sequence against hidden Markov models (HMMs) of proteins in a database. We built a custom HMM database to target specific annotations of interest.

The HMM database currently contains Virus Orthologous Groups from VOGDB and biosynthetic genes and can be extended in the future to meet user annotation interests.

2. **Reporting**. The last step combines all information that the pipeline has produced and outputs the results in a TSV file. The results include the GenBank protein identifier for the HGT candidate, BLAST and relative amino acid usage scores, pangenome information, gene and ortholog annotations, and contextualizing information about the gene such as position in the contiguous sequence.

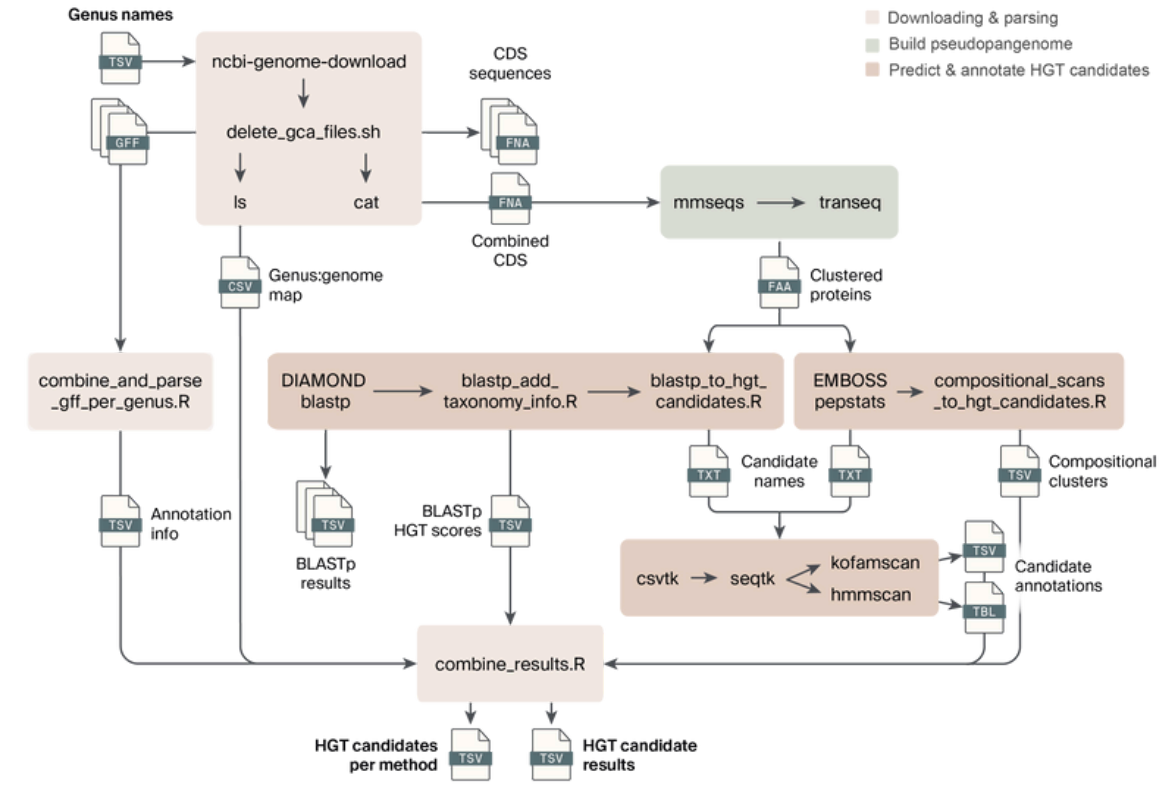


Figure 2. **Overview of the preHGT pipeline steps, inputs, and outputs**.
Users provide a genus or genera of interest in a TSV file as input to the pipeline. The workflow then downloads and parses the available genomes for those genera, builds a pseudo-pangenome, and predicts and annotates horizontally transferred gene candidates.

# Types of HGT events that the pipeline screens for

While we tried to create a fast and generalized pipeline, preHGT is better at detecting some patterns of HGT than others. The preHGT pipeline screens for HGT events where the donor and recipient differ in taxonomy at the family level or above. It is most likely more accurate when the transfer events occur between

more distantly related organisms and where the recipient gene retains homology to the gene in the donor genome. We anticipate the primary use of this approach will be to identify candidate transfer events and donor and recipient groups to which more granular approaches can be applied to better disentangle the evolutionary history of the gene.

The parametric approach we implemented screens for genes with outlying relative amino acid usage compared to the rest of the genes in the genome. This requires that the donor and acceptor species differ in amino acid composition, and that these differences persist in the transferred genes, a scenario that is most typical of recent transfer events among evolutionarily divergent species.

The BLAST-based implicit phylogenetic approaches we implemented screen for genes that exhibit a greater degree of sequence similarity among designated taxonomic outgroups than within ingroups. In the original tools and papers in which these algorithms were generated, the authors implemented or validated their approaches at specific taxonomic levels that the preHGT pipeline adheres to (Table 2). Some are designed to screen for cross-kingdom transfer events, while others can screen for sub-kingdom-level events. However, because the chance of spurious inference of homology increases among more closely related species, results should be more carefully scrutinized at lower taxonomic levels (e.g. order, family). Homology detection also becomes increasingly difficult at larger taxonomic distances, so the pipeline may miss highly diverged homologs.

# Additional considerations and caveats

## How we deal with contamination and other sources of false positives

HGT screens often return many false positives [56][57]. We used contextualizing information about HGT candidates to reduce the number of false positives reported by the pipeline.

Contamination is the biggest source of false positives in BLAST-based HGT screening algorithms. Many genomes in GenBank and RefSeq are contaminated [58]. Contamination arises from impure sampling, contaminated reagents, lab cross-contamination, sequencing artifacts, or reference database

errors [74]. To combat the presence of contamination, we incorporated multiple corroborating lines of evidence to assess whether contamination is more likely than HGT. First, we determine the length of the contiguous sequence within which the candidate gene is found. Short contiguous sequences are more likely to be contaminants [58][75]. Next, we determine how many genes are in the candidate gene's cluster from our pseudo-pangenome approach. Depending on the contamination source, it is unlikely that the same contamination will occur in multiple genomes [76]. Therefore, if a homolog is present in multiple genomes, it is less likely to be a contaminant. Lastly, for BLAST-based results, we assess the percent identity between the donor and acceptor genes. Amelioration deteriorates sequence identity after a transfer event [10], so the more similar two genes are, the more likely similarity is driven by contamination. Many methods use a cutoff of 70%–80% identity for contamination [16][77], but we instead weigh this against other corroborating information.

In the future, we hope to further contextualize contamination potential against the general contamination score for the acceptor genome. The more contamination a genome contains, the more likely a candidate is to be a contaminant itself.

BLAST-based methods may also generate false positives arising from alignment errors or alignment due to sequence similarity that does not arise from shared ancestry, such as from convergent evolution or random chance. Alignment errors from short or low-complexity sequences or from short, highly conserved domains may give the appearance of a horizontal transfer event. To protect against this, we filter corrected bit scores to those greater than 100, or, to rescue true homologs that are very divergent, with a query coverage of greater than 70%. We also provide gene annotations from multiple annotation sources to highlight hits that might be ultra-conserved, such as those from ribosomal proteins. Over time, we hope to curate a list of genes that the preHGT pipeline frequently detects as false positives and to develop a strategy to filter them out.

## Verifying bona fide HGT requires work beyond preHGT

The preHGT pipeline provides a list of candidate HGT events. These candidates need to be carefully scrutinized to determine whether they are biologically interesting and whether they are more or less likely to be false positives. We built preHGT as a generalized precursor to more in-depth HGT analysis (Figure 3). We

envision that preHGT can inform genome selection for comprehensive explicit phylogenetic inference, which can help disentangle alternate evolutionary trajectories, or highlight when not enough information is available to support HGT inference.

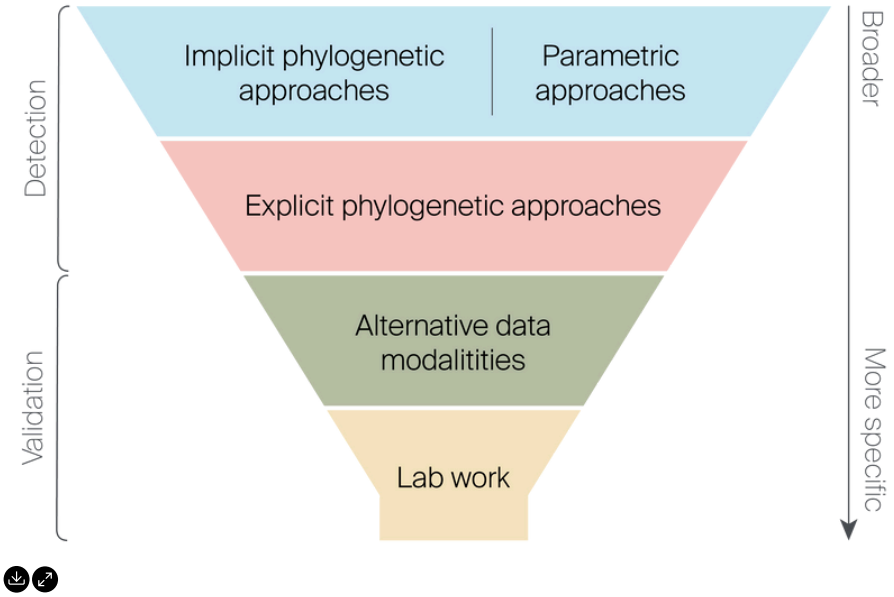


Figure 3. **Funnel of methods for HGT screening and validation**.

Implicit phylogenetic and parametric approaches are fast, generalized methods for screening for HGT in genes or genomes, but these methods are prone to false positives. Explicit phylogenetic methods can help eliminate some false positives or determine when there is not enough evidence to support HGT candidacy.

After these computational approaches, validation requires additional methods. Alternative data modalities like transcriptome sequencing or laboratory experiments like FISH or PCR can provide additional evidence in support of HGT. While powerful, these methods require curated information about the donor and acceptor genomes and the candidate genes and thus can usually only be used after initial exploration.

After phylogenetic analysis, more analysis is still required to reject the null hypothesis that no transfer event occurred. The appropriate experiments for this will depend on the HGT candidate event itself. For example, if a bacterial gene has been transferred into a eukaryotic genome, it may be appropriate to interrogate the candidate gene for the presence of introns, or if transcriptome information is available, for the presence of transcription- and eukaryotic-specific RNA modifications such as 5′ caps or Kozak sequences. In the lab, PCR, FISH, or Southern blots may confirm the presence of the sequence in the genome of interest, while Western blot or mass spectrometry can confirm that the gene is transcribed and translated into a protein.

## Limitations of the preHGT pipeline

Given our approach, we have identified multiple shortcomings. The most conspicuous limitation is our focus on genes. The preHGT pipeline can only scan genomes with gene models. We elected not to implement genome annotation as an early step in the pipeline given that annotation procedures differ for eukaryotic versus bacterial and archaeal genomes, and that eukaryotic genome annotation remains a challenging problem from the genome alone [78][79]. This limits the preHGT pipeline to those genomes with gene models (approximately 21%) and creates blind spots for HGT detection across the tree of life. Of 56 eukaryotic phyla with genomes, only 45 have at least one genome with gene models. Similarly, by treating genes in their entirety as the unit that is horizontally transferred, we are unlikely to detect genes for which only a nested region of the coding sequence was horizontally transferred.

There are also limitations born out of our decision to use composition or BLAST-based HGT screening methods. First, these methods require that the gene has not ameliorated to the composition of the acceptor genome or that it maintains detectable homology to the donor genome. This may limit our detection of ancient HGT events. Second, these methods will be less sensitive to HGT events that occur between closely related organisms. Third, since BLAST-based approaches rely on taxonomies, there are risks since taxonomies may be wrong and since they do not account for branch lengths in the relatedness of species. Lastly, false positives may arise from alignment between short or low-complexity sequences or from natural sequence similarity such as what might arise from convergent evolution or from highly conserved gene sequences. To combat both cases, we have implemented filtering criteria to help eliminate these issues.

Lastly, we did not integrate an explicit phylogenetic approach to better resolve the evolutionary histories of HGT candidates. We elected to forgo this step because another team at Arcadia is developing a tree-based workflow. We are currently experimenting with how to facilitate handoff between the two tools to rapidly enable this next step in validation.

# Additional methods

We used ChatGPT to add comments to our code and suggest wording ideas. We also used ChatGPT to add comments to external code to help us better understand how it worked when trying to implement some existing tools in another language.

# Key takeaways

- preHGT is a scalable pipeline that screens for potential HGT events in genomes with gene models across the tree of life and taxonomic scales.
- The pipeline leverages compositional and BLASTp scans, pangenome inference, annotation, and reporting techniques to provide comprehensive results.
- Multiple checks and filters defend against false positives, including contamination detection and sequence alignment artifact filtering.
- The pipeline is implemented in both Snakemake and Nextflow. Its modular design means it's easily extensible to incorporate more methods in the future.
- preHGT aims to identify HGT events that users further investigate with other approaches such as tree-based ones.

# Next steps

Our follow-up plans include:

1. **Eukaryotic HGT prediction**: We plan to run the pipeline on all eukaryotic genomes in GenBank and RefSeq that have gene models and to make the results available.
2. **Building a user interface for results exploration**: We plan to build a simple user interface to explore results produced by the pipeline. Exploration modes will allow users to dive into gene transfer events by donor or acceptor taxonomy, predicted functions of genes involved, or by strength of result, and to visualize the results in their genomic context.

3. **Adapting the pipeline to take transcriptome assemblies as input**: We plan to extend the pipeline to run on assembled transcriptomes by incorporating upstream gene prediction rules. We will then run the pipeline on the transcriptomes in the NCBI Transcriptome Shotgun Assembly database and make the results publicly available.
4. **Integrating new algorithms for HGT screening**: Other algorithms exist for the interpretation of BLASTp results. We plan to integrate those from other tools into this pipeline in the future.

We welcome feedback on the user experience, the results we include, or additional algorithms or metrics that would be helpful to incorporate.

---

# Contributors (A–Z)

- **Adair L. Borges**: Critical Feedback
- **Feridun Mert Celebi**: Resources, Validation
- **Rachel J. Dutton**: Supervision
- **Jonathan A. Eisen**: Critical Feedback
- **Megan L. Hochstrasser**: Editing, Visualization
- **Elizabeth A. McDaniel**: Validation
- **Austin H. Patton**: Critical Feedback, Editing
- **Taylor Reiter**: Conceptualization, Software, Visualization, Writing
- **Dennis A. Sun**: Validation
- **Emily C.P. Weiss**: Critical Feedback, Validation

# References


1. Soucy SM, Huang J, Gogarten JP. (2015). Horizontal gene transfer: building the web of life. https://doi.org/10.1038/nrg3962
2. Husnik F, McCutcheon JP. (2017). Functional horizontal gene transfer from bacteria to eukaryotes. https://doi.org/10.1038/nrmicro.2017.137
3. Arnold BJ, Huang I-T, Hanage WP. (2021). Horizontal gene transfer and adaptive evolution in bacteria. https://doi.org/10.1038/s41579-021-00650-4
4. Zaneveld JR, Nemergut DR, Knight R. (2008). Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. https://doi.org/10.1099/mic.0.2007/011833-0

5. Van Etten J, Bhattacharya D. (2020). Horizontal Gene Transfer in Eukaryotes: Not if, but How Much?. https://doi.org/10.1016/j.tig.2020.08.006
6. Wagner A, Whitaker RJ, Krause DJ, Heilers J-H, van Wolferen M, van der Does C, Albers S-V. (2017). Mechanisms of gene flow in archaea. https://doi.org/10.1038/nrmicro.2017.41
7. Corvaglia AR, François P, Hernandez D, Perron K, Linder P, Schrenzel J. (2010). A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. https://doi.org/10.1073/pnas.1000489107
8. Marraffini LA, Sontheimer EJ. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. https://doi.org/10.1126/science.1165771
9. Gabaldón T. (2020). Patterns and impacts of nonvertical evolution in eukaryotes: a paradigm shift. https://doi.org/10.1111/nyas.14471
10. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. (2015). Inferring Horizontal Gene Transfer. https://doi.org/10.1371/journal.pcbi.1004095
11. Knowles LL, Huang H, Sukumaran J, Smith SA. (2018). A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. https://doi.org/10.1002/ajb2.1064
12. Copley SD, Dhillon JK. (2002). Lateral gene transfer and parallel evolution in the history of glutathione biosynthesis genes. https://doi.org/10.1186/gb-2002-3-5-research0025
13. Cote-L'Heureux A, Maurer-Alcalá XX, Katz LA. (2022). Old genes in new places: A taxon-rich analysis of interdomain lateral gene transfer events. https://doi.org/10.1371/journal.pgen.1010239
14. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. https://doi.org/10.1073/pnas.1600338113
15. Vernikos GS, Parkhill J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. https://doi.org/10.1093/bioinformatics/btl369
16. Rancurel C, Legrand L, Danchin E. (2017). Alienness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. https://doi.org/10.3390/genes8100248

17. Sengupta S, Azad RK. (2023). Leveraging comparative genomics to uncover alien genes in bacterial genomes. https://doi.org/10.1099/mgen.0.000939
18. David LA, Alm EJ. (2010). Rapid evolutionary innovation during an Archaean genetic expansion. https://doi.org/10.1038/nature09649
19. Koutsovoulos GD, Granjeon Noriot S, Bailly-Bechet M, Danchin EGJ, Rancurel C. (2022). AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers. https://doi.org/10.1371/journal.pcbi.1010686
20. Li M, Zhao J, Tang N, Sun H, Huang J. (2018). Horizontal Gene Transfer From Bacteria and Plants to the Arbuscular Mycorrhizal Fungus Rhizophagus irregularis. https://doi.org/10.3389/fpls.2018.00701
21. Podell S, Gaasterland T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. https://doi.org/10.1186/gb-2007-8-2-r16
22. Wan Y, Wick RR, Zobel J, Ingle DJ, Inouye M, Holt KE. (2020). GeneMates: an R package for detecting horizontal gene co-transfer between bacteria using gene-gene associations controlled for population structure. https://doi.org/10.1186/s12864-020-07019-6
23. Soares SC, Geyik H, Ramos RT, de Sá PH, Barbosa EG, Baumbach J, Figueiredo HC, Miyoshi A, Tauch A, Silva A, Azevedo V. (2016). GIPSy: Genomic island prediction software. https://doi.org/10.1016/j.jbiotec.2015.09.008
24. Garcia-Vallve S. (2003). HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. https://doi.org/10.1093/nar/gkg004
25. Nguyen M, Ekstrom A, Li X, Yin Y. (2015). HGT-Finder: A New Tool for Horizontal Gene Transfer Finding and Application to Aspergillus genomes. https://doi.org/10.3390/toxins7104035
26. Zhu Q, Kosoy M, Dittmar K. (2014). HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. https://doi.org/10.1186/1471-2164-15-717
27. Yuan L, Lu H, Li F, Nielsen J, Kerkhoven EJ. (2023). HGTphyloDetect: facilitating the identification and phylogenetic analysis of horizontal gene transfer. https://doi.org/10.1093/bib/bbad035
28. Choi Y, Ahn S, Park M, Lee S, Cho S, Kim H. (2022). HGTree v2.0: a comprehensive database update for horizontal gene transfer (HGT) events detected by the tree-reconciliation method. https://doi.org/10.1093/nar/gkac929

29. Hudson CM, Lau BY, Williams KP. (2014). Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. https://doi.org/10.1093/nar/gku1072

30. Baichoo S, Goodur H, Ramtohul V. (2014). IslandHunter – A Java-based GI detection software. https://doi.org/10.7287/peerj.preprints.716v1

31. Bertelli C, Brinkman FSL. (2018). Improved genomic island predictions with IslandPath-DIMOB. https://doi.org/10.1093/bioinformatics/bty095

32. Langille MG, Hsiao WW, Brinkman FS. (2008). Evaluation of genomic island predictors using a comparative genomics approach. https://doi.org/10.1186/1471-2105-9-329

33. Bertelli C, Laird MR, Williams KP, Group SFURC, Lau BY, Hoad G, Winsor GL, Brinkman FS. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. https://doi.org/10.1093/nar/gkx343

34. Adato O, Ninyo N, Gophna U, Snir S. (2015). Detecting Horizontal Gene Transfer between Closely Related Taxa. https://doi.org/10.1371/journal.pcbi.1004408

35. Zhao Y, Sun C, Zhao D, Zhang Y, You Y, Jia X, Yang J, Wang L, Wang J, Fu H, Kang Y, Chen F, Yu J, Wu J, Xiao J. (2018). PGAP-X: extension on pan-genome analysis pipeline. https://doi.org/10.1186/s12864-017-4337-7

36. Bansal MS, Kellis M, Kordi M, Kundu S. (2018). RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. https://doi.org/10.1093/bioinformatics/bty314

37. Li X, Tong W, Wang L, Rahman SU, Wei G, Tao S. (2018). A Novel Strategy for Detecting Recent Horizontal Gene Transfer and Its Application to Rhizobium Strains. https://doi.org/10.3389/fmicb.2018.00973

38. Nakhleh L, Ruths D, Wang L-S. (2005). RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer. https://doi.org/10.1007/11533719_11

39. Narechania A, Bobo D, DeSalle R, Mathema B, Kreiswirth B, Planet PJ. (2021). What do we gain when tolerating loss? The information bottleneck wrings out recombination. https://doi.org/10.1101/2021.08.27.457981

40. Sánchez-Soto D, Agüero-Chapin G, Armijos-Jaramillo V, Perez-Castillo Y, Tejera E, Antunes A, Sánchez-Rodríguez A. (2020). ShadowCaster: Compositional Methods under the Shadow of Phylogenetic Models to Detect Horizontal Gene Transfers in Prokaryotes. https://doi.org/10.3390/genes11070756

41. Jaron KS, Moravec JC, Martínková N. (2013). SigHunt: horizontal gene transfer finder optimized for eukaryotic genomes.

https://doi.org/10.1093/bioinformatics/btt727

42. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. https://doi.org/10.1186/1471-2105-7-142
43. Boc A, Diallo AB, Makarenkov V. (2012). T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. https://doi.org/10.1093/nar/gks485
44. Cong Y, Chan Y-b, Ragan MA. (2016). A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. https://doi.org/10.1038/srep30308
45. Moriyama E. (1998). Gene length and codon usage bias in Drosophila melanogaster, Saccharomyces cerevisiae and Escherichia coli. https://doi.org/10.1093/nar/26.13.3188
46. Urrutia AO, Hurst LD. (2003). The Signature of Selection Mediated by Expression on Human Genes. https://doi.org/10.1101/gr.641103
47. Keeling PJ, Palmer JD. (2008). Horizontal gene transfer in eukaryotic evolution. https://doi.org/10.1038/nrg2386
48. Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. (2020). GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. https://doi.org/10.1093/molbev/msaa141
49. Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. https://doi.org/10.1093/sysbio/syt054
50. Gladyshev EA, Meselson M, Arkhipova IR. (2008). Massive Horizontal Gene Transfer in Bdelloid Rotifers. https://doi.org/10.1126/science.1156407
51. Boschetti C, Carr A, Crisp A, Eyres I, Wang-Koh Y, Lubzens E, Barraclough TG, Micklem G, Tunnacliffe A. (2012). Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. https://doi.org/10.1371/journal.pgen.1003035
52. Ma J, Wang S, Zhu X, Sun G, Chang G, Li L, Hu X, Zhang S, Zhou Y, Song C-P, Huang J. (2022). Major episodes of horizontal gene transfer drove the evolution of land plants. https://doi.org/10.1016/j.molp.2022.02.001
53. Freschi L, Vincent AT, Jeukens J, Emond-Rheault J-G, Kukavica-Ibrulj I, Dupont M-J, Charette SJ, Boyle B, Levesque RC. (2018). ThePseudomonas aeruginosaPan-Genome Provides New Insights on Its Population

Structure, Horizontal Gene Transfer, and Pathogenicity. https://doi.org/10.1093/gbe/evy259

54. Fan X, Qiu H, Han W, Wang Y, Xu D, Zhang X, Bhattacharya D, Ye N. (2020). Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. https://doi.org/10.1126/sciadv.aba0111

55. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR. (2001). Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. https://doi.org/10.1038/35082058

56. Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Osborne Nishimura E, Tintori SC, Li Q, Jones CD, Yandell M, Messina DN, Glasscock J, Goldstein B. (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. https://doi.org/10.1073/pnas.1510461112

57. International Human Genome Sequencing Consortium, Research: WIfBRCfG, Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, The Sanger Centre:, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Center WUGS, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Institute: UDJG, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J-F, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Center: BCoMHGS, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Center: RGS, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, UMR-8030: GaC, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Biotechnology: DoGAloM, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Center: GS, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Center: BGIG, Yang H, Yu J, Wang J, Huang G, Gu J, Biology: MSCTIfS, Hood L, Rowen L, Madan A, Qin S, Center: SGT, Davis RW, Federspiel NA, Abola AP, Proctor MJ,

Technology: UoOACfG, Roe BA, Chen F, Pan H, Genetics: MPIfM, Ramser J, Lehrach H, Reinhardt R, Center: CSHLLAHG, McCombie WR, de la Bastide M, Dedhia N, Biotechnology: GRCf, Blöcker H, Hornischer K, Nordsiek G, headings): *AG(iaoaiiluo, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen H-C, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang S-P, Yeh R-F, Health: SmNHGRIUNIo, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Center: SHG, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Center: UoWG, Olson MV, Kaul R, Raymond C, Medicine: DoMBKUSo, Shimizu N, Kawasaki K, Minoshima S, Dallas: UoTSMCa, Evans GA, Athanasiou M, Schultz R, Energy: OoSUDo, Patrinos A, The Wellcome Trust:, Morgan MJ. (2001). Initial sequencing and analysis of the human genome. https://doi.org/10.1038/35057062

58. Steinegger M, Salzberg SL. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. https://doi.org/10.1186/s13059-020-02023-1

59. Reiter T. (2023). Clustering the NCBI nr database to reduce database size and enable faster BLAST searches. https://doi.org/10.57844/arcadia-w8xt-pc81

60. Köster J, Rahmann S. (2012). Snakemake—a scalable bioinformatics workflow engine. https://doi.org/10.1093/bioinformatics/bts480

61. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. (2017). Nextflow enables reproducible computational workflows. https://doi.org/10.1038/nbt.3820

62. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry S, Karsch-Mizrachi I. (2021). GenBank. https://doi.org/10.1093/nar/gkab1135

63. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D,

Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. https://doi.org/10.1093/nar/gkv1189

64. Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988
65. Glick L, Mayrose I. (2021). Panoramic: A package for constructing eukaryotic pan genomes. https://doi.org/10.1111/1755-0998.13344
66. Hu Z, Wei C, Li Z. (2020). Computational Strategies for Eukaryotic Pangenome Analyses. https://doi.org/10.1007/978-3-030-38281-0_13
67. Müllner D. (2013). **fastcluster**: Fast Hierarchical, Agglomerative Clustering Routines forRandPython. https://doi.org/10.18637/jss.v053.i09
68. Moura A, Savageau MA, Alves R. (2013). Relative Amino Acid Composition Signatures of Organisms and Environments. https://doi.org/10.1371/journal.pone.0077319
69. Buchfink B, Reuter K, Drost H-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. https://doi.org/10.1038/s41592-021-01101-x
70. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. (2019). Welcome to the Tidyverse. https://doi.org/10.21105/joss.01686
71. Emms DM, Kelly S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. https://doi.org/10.1186/s13059-015-0721-2
72. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. (2019). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. https://doi.org/10.1093/bioinformatics/btz859
73. Chou S, Reiter T. (2023). Speeding up the quality control of raw sequencing data using seqqc, a Nextflow-based solution. https://doi.org/10.57844/arcadia-cxn6-ch62
74. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. https://doi.org/10.1101/gr.245373.118

75. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. (2015). Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. https://doi.org/10.1186/s13059-015-0607-3

76. Haimlich S, Fridman Y, Khandal H, Savaldi-Goldstein S, Levy A. (2022). Widespread horizontal gene transfer between plants and their microbiota. https://doi.org/10.1101/2022.08.25.505314

77. Salzberg SL. (2019). Next-generation genome annotation: we still struggle to get it right. https://doi.org/10.1186/s13059-019-1715-2

78. Blaxter M, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, Durbin R, Edwards SV, Graves JAM, Hackett KJ, Hall N, Jarvis ED, Johnson RN, Karlsson EK, Kress WJ, Kuraku S, Lawniczak MKN, Lindblad-Toh K, Lopez JV, Moran NA, Robinson GE, Ryder OA, Shapiro B, Soltis PS, Warnow T, Zhang G, Lewin HA. (2022). Why sequence all eukaryotes?. https://doi.org/10.1073/pnas.2115636118