GrAnnoT, a tool for efficient and reliable annotation transfer through pangenome graph

Nina Marthe[1], Matthias Zytnicki[2], and Francois Sabot[1]

DOI not yet assigned

Abstract

The increasing availability of genome sequences has highlighted the limitations of using a single reference genome to represent the diversity within a species. Pangenomes, encompassing the genomic information from multiple genomes, thus offer a more comprehensive representation of intraspecific diversity. However, pangenomes in form of a variation graph often lack annotation information and tools for, which limits their utility for downstream analyses. We introduce here GrAnnoT, a tool designed for an efficient and reliable integration of annotation information in such variation graphs. It projects existing annotations from a source genome to the variation graph and subsequently to other embedded genomes. GrAnnoT was benchmarked against state-of-the-art tools on pangenome variation graphs and linear genomes from Asian rice, and tested on human and *E. coli* data. The results demonstrate that GrAnnoT is consensual, conservative, and fast. It provides informative outputs, such as presence-absence matrices for genes, and alignments of transferred features between source and target genomes, helping in the study of genomic variations and evolution. GrAnnoT's robustness and replicability across different species make it a valuable tool for enhancing pangenome analyses. GrAnnoT is available under the GNU GPLv3 licence at https://forge.ird.fr/diade/dynadiv/grannot.

Keywords: Pangenome, graph, annotation transfer

[1]DIADE unit, UM, Cirad, IRD, 911 avenue Agropolis Montpellier F-34391, France, [2]MIAT, INRAE, Auzeville-Tolosane F31320, France

Correspondence
nina.marthe@ird.fr, francois.sabot@ird.fr

Introduction

Recent advances in genome sequencing and assembly methods give access to a massive and increasing number of genome sequences per species for the scientific community. Consequently, while still currently prevalent, the use of a single reference genome has been shown to bias many analyses (Chen et al., 2021; Martiniano et al., 2020; Maurstad et al., 2024), as it favors variant calling toward reference alleles and thus hinders the identification of non-reference sequences. From that rises the idea that a single individual is not enough to represent the diversity of a given species or group. This led to the development and diffusion of the concept of pangenomics across the whole tree of life (Bayer et al., 2020; Liao et al., 2023; Miga and Wang, 2021; Rouli et al., 2015; Shi et al., 2023).

A pangenome aims to represent the complete genomic information from several genomes of the same species or group, in order to better represent the intra-specific/group diversity. While the concept emerged from bacterial studies (Tettelin et al., 2005), it is now applied to larger and more complex eukaryotic genomes. Many studies in pangenomics have been published, and allowed a better understanding of genomic diversity, population dynamics, and evolution (Rice et al., 2023; Secomandi et al., 2025; Tranchant-Dubreuil et al., 2019; Zhou et al., 2022).

This pangenomic information can be stored in different structures depending on the type of organism and study involved. Ranging from gene set to whole pangenome graphs, the methods to build, manipulate and study these structures differ. Some representations of pangenomes have an extensive toolset allowing in depth analysis (e.g. bacterial pangene set with tools such as PPanGGOLiN (Gautreau et al., 2021)), some can handle thousands of genomes (e.g. de Bruijn graphs built by Bifrost (Holley and Melsted, 2020)), and some others still suffer from methodological shortage.

Annotation is an important element for studying genomic sequences and understanding their potential biological functions, if any, to help interpret the variations found in the pangenome in regard of phenotypes. Numerous tools have already been proposed to produce, cluster, visualize, or manipulate pangenome annotation (Durant et al., 2021; Gautreau et al., 2021; Horsfield et al., 2023; Pedersen et al., 2016).

The variation graph (Outten and Warren, 2021) is a promising structure for representing a pangenome, but it still lacks adapted tools to integrate and manipulate annotation. This structure represents the whole sequence information of the embedded genomes, including intergenic regions, small and large variants (SNPs as well as large indels, duplications and translocations - to some extent), and is easier to use for large or complex genomes with high repeat content (Andreace et al., 2023; Secomandi et al., 2025) compared to de Brujin graphs, for instance. In variation graphs, the nodes represent sequences by stretch of DNA, the links (or edges) show the adjacencies between two sequences (nodes) in at least one embedded genome, and the paths reconstruct these embedded genome sequences. Variation graphs are usually built from complete alignments between whole genomes, and the currently most popular tools include PGGB (Garrison et al., 2024) and minigraph-cactus (Hickey et al., 2023). Additionally, tools like VG (Garrison et al., 2018) and ODGI (Guarracino et al., 2022) manipulate these graphs and perform various tasks. These variation graphs are currently used for better alignment of reads, genotyping, and structural variation detection (Hickey et al., 2020; Sirén et al., 2021). However, the variation

graph only encodes sequence information, and does not carry any annotation natively. Such biological information is crucial to give context to any variation in genomic sequences, and is often available for the linear reference genomes. Integrating these existing annotations to the variation graphs would enrich these structures and make them a better tool for studying pangenomes.

While tools exist to visualize annotations on a graph (Jonkheer et al., 2022; Liu et al., 2024; Miao and Yue, 2025; Wick et al., 2015), their use is not adapted for large-scale analyses of thousands of annotations. To answer this issue, VG annotate recently proposed to project genomic annotation on a graph, and offers options to efficiently index and query the resulting graph annotation (Novak et al., 2024). However, it cannot project annotations from the graph to the genomes, which would in turn allow to output genome annotation and to identify the variants between the embedded genomes in the annotated regions, to study their impact. Tools for transferring annotation exist for linear to linear genomes, with the most used of them being Liftoff (Shumate and Salzberg, 2021), but it relies on gene-by-gene sequence alignment and does not provide explicit information about variations between genomes. In this regard, the use of the variation graph, which represents in its essence a whole genome alignment and models the synteny between embedded genomes, is a natural way to transfer annotations from genome to graph, or from graph to genome, and to identify the differences between genomes.

To fill this gap, we developed GrAnnoT, a command line tool that manipulates genomic annotations in a variation graph space under its native GFA format. Starting from projecting on the graph the annotation of a single genome in a GFF format, GrAnnoT then outputs the graph annotation in GAF format (as VG annotate), but it also projects this annotation on the other embedded genomes with a dedicated GFF output for each of them. In addition to this transfer, GrAnnoT compares the annotated regions between the genomes in the graph, and outputs transfer statistics (i.e. transfer rate, or mean sequence identity and coverage), lists and types of variants, alignments, and a presence-absence matrix for gene features. These operations rely on the structure of the graph, speeding up the transfers by harnessing the multiple genome alignment it represents (Hickey et al., 2023; Li et al., 2020). We applied it to variation graphs built from different species, and compared it to existing methods to ensure the annotation transfer is valid. As an graph annotation transfer tool, GrAnnoT is fast and conservative, and it allows to study the annotated regions of a pangenome variation graph and of its embedded genomes in an easier way and at a larger scale than any tool before.

1. Implementation

GrAnnoT is implemented in Python 3.10+, as a Linux command-line tool that can be installed as a standard python package. It only requires the *tqdm* package and the external program bedtools (Quinlan, 2014) (that must be accessible in the user or in the global path). To ease its installation and use, an AppTainer container definition is available in the GrAnnoT repository on the IRD forge (`https://forge.ird.fr/diade/dynadiv/grannot`); all the codes are under the GNU GPLv3 license.

1.1. Code overview

GrAnnoT performs annotation transfer from an annotated genome (the source genome) to a variation graph (Figure 1). It can also transfer the annotation from the graph to one, several or all other genomes embedded in the graph (the target genomes). It takes as input the annotation of

the source genome in GFF3 format, and the variation graph that includes the source and target genome in GFA 1.1 format (without overlap between the nodes).

The annotation transfer only relies on the graph structure, harnessing the multiple alignment and synteny it naturally represents. GrAnnoT projects the coordinates between the graph and the genomes, transferring annotations in a fast, alignment-free manner.

Once the annotation has been loaded, GrAnnoT outputs the graph annotation in GAF format (Li et al., 2020). This tab-delimited text format was originally proposed to represent sequence-to-graph alignment. However, it can also be used for graph annotation (Novak et al., 2024), where, instead of describing the paths of the mapped reads, it describes the paths of the annotated features through the graph. GrAnnoT can then output the annotation in GFF3 format for a chosen set of target linear genomes included in the variation graph. These transfers can be filtered through sequence identity and coverage scores similarly to the BLAST approach (Altschul et al., 1990). For these transfers, the alignment of each feature between the source genome and the target one can be outputted in a Clustal-like format, as well as a list of all the variants recorded in the alignments. These alignments are not computed by GrAnnoT, but directly extracted from the graph structure. Finally, a presence-absence matrix for gene features summarizes the transfer on the target genomes.

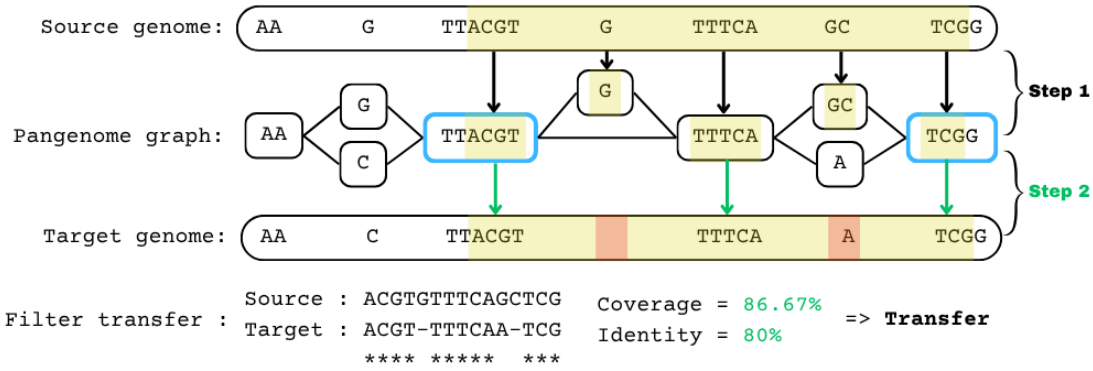
1.2. Implementation details

The first step is to find the start and stop positions of each node from the graph on the embedded genomes (Figure 1, step 1). For that, GrAnnoT follows the paths of these genomes in the graph and computes the start and stop positions of the nodes for each of them; these positions are then stored in BED files, one per contig/chromosome per genome. Then, the BED files representing the source genome are compared to its annotation file using bedtools intersect (Quinlan, 2014). The resulting BED file is processed to compute the paths of the features in the graph and output the graph annotation in the GAF format.

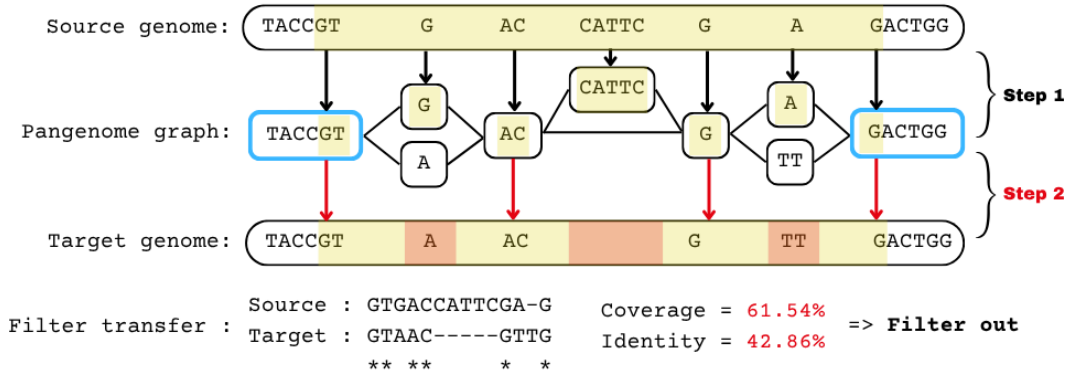
In order to transfer an annotation to a target genome, the sub-path of the genome corresponding to the feature is extracted (Figure 1a, step 2). For that, all the nodes from the original feature path are looked for in the target genome paths. These nodes are then grouped into copies of the feature, since the nodes corresponding to a feature can sometimes be found multiple times in the target genome path (duplication). For each copy, the first and the last nodes are considered as the ends of the feature's copy in the target genome, and all copies are transferred by default. An option allows to only transfer the copy with the highest sequence identity and coverage. All the nodes between the first and last segment in the target genome path are expected to be part of the feature's copy to transfer, including the nodes absent from the original feature path, corresponding to insertions. Nodes from the original feature path that are not found in the target genome correspond to deletions. An insertion and a deletion at the same locus in the variation graph correspond to a substitution.

For the transfer itself, only the two nodes at the ends of the feature path on the target genome are considered (nodes in blue in Figure 1). The BED file previously computed reporting the positions of the nodes on the target genome is used to locate these two nodes on the genome.

Transferred features are then filtered based on the coverage (in base) and the identity level between the source and the target genomes (Figure 1b), both set at 80% by default (but can be defined by the user). These parameters are estimated by computing the cumulated length



(a) Transfer from genome to graph (step 1) and from graph to genome (step 2)



(b) Transfer on target genome rejected due to low coverage and sequence identity.

Figure 1 – GrAnnoT overview. Step 1: the position of the feature is projected from the source genome to the graph using the positions of the nodes on the source genome. Step 2: the position of the feature is projected from the graph to the target genome (a). The first and last nodes from the feature that are on the target genome (the blue ones) are the ends of the feature in this genome, and everything in between is considered as part of the feature. The differences between the two genomes in this region in terms of path in the graph mirror the differences between the two versions of the feature. If these differences are too important and the transfer does not reach the coverage and/or sequence identity thresholds, the annotation is not transferred (b).

of the shared and different nodes between the paths of the features in the two genomes. The genes are excluded if they do not meet the threshold set, and their child features (exons, CDS, UTR) are not transferred. This filtering ensures that the sequence of the annotated feature is conserved between the source and target genome, to remove spurious transfers. The output is finally printed out in the GFF3 format.

If the user is interested in the differences between the source and target annotation, GrAnnoT can provide a detailed comparison between the feature alternative paths in the source and any of the embedded target genomes. For that, GrAnnoT can output the variants details in a human-readable text format that describes all the variants present in the feature (node deletion, insertion, substitution). A Clustal-like alignment file of all the transferred features based on their alternative paths is similarly generated.

2. Benchmark

2.1. Data and tools for benchmarks

The main test data used in this paper is an Asian rice pangenome graph built with 13 genomes (Kawahara et al., 2013a; Zhou et al., 2020) using minigraph-cactus v2.8.2 with default options (Hickey et al., 2023; see supplementary data for the exact commands) and the cv Nipponbare IRGSP1.0 as reference. The rice genome is 380-410Mb long and has 12 chromosomes. The annotation used as source (Kawahara et al., 2013b) includes 57,585 *gene* features for 813,790 total features, and is rich in transposable elements (15,848/57,858≈27% of *gene* features are annotated as transposable elements).

GrAnnoT was also tested on a graph of the human chromosome 1 with 92 haplotypes (from Liao et al., 2023) and an *E. coli* 13 genomes graph (see supplementary data for the genomes used) built using the same protocol as for rice (detailed commands available online, Marthe and Sabot, 2025b).

GrAnnoT was compared to existing and state-of-the-art tools (see below) that can also perform annotation transfer in order to assess its efficiency, and using the different data presented before to test its replicability and robustness. All analyses were ran on a biprocessor Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz with 48 HT CPU computer with 144Gb of RAM, under RockyLinux 9.1 Blue Onyx.

In all transfers made with GrAnnoT, the parameters for sequence identity and coverage based filtering were left as default, so at 80% both. These parameters can be changed by the user, and an evaluation of their effect on the transfers in rice data is available in supplementary data (Table S14).

Multiple tools are available to perform annotation transfer between linear genomes, with different approaches. Tools like CAT (Fiddes et al., 2018), RATT (Otto et al., 2011), FLO (Pracana et al., 2017), or CrossMap (Zhao et al., 2013) use a form of whole genome alignment to convert the positions of the annotated features from one genome to another. Tools like Liftoff (Shumate and Salzberg, 2021), GeMoMa (Keilwagen et al., 2019) and LiftOn (Chao et al., 2024) align the sequence of each annotated feature on the target genome to find its position. The current state-of-the-art annotation transfer tool for linear genome sequences is Liftoff (Shumate and Salzberg, 2021). It is widely used (Alonge et al., 2022; Kim et al., 2021; Wang et al., 2021; Yang et al., 2023), and was chosen here test the validity of GrAnnoT's genome to genome annotation transfers. However, since Liftoff does not use a pangenome graph to transfer annotations, the comparison with GrAnnoT is biased by the graph itself, whose structure partially impacts the results of GrAnnoT transfer (see Discussion).

For annotation transfer on graph through alignment, Liftoff approach can be mimicked by aligning the sequences of the annotated features to the graph. Graph pangenome alignment tools can be thus compared to GrAnnoT for graph annotation transfer: GraphAligner was chosen for this purpose (Rautiainen and Marschall, 2020), as a state-of-the-art tool for aligning long sequences on a graph.

To transfer annotations between genomes through a pangenome graph and use graph properties instead of gene sequence alignment, VG and ODGI were used. They are state-of-the-art tools for pangenome graph manipulation (Garrison et al., 2018; Guarracino et al., 2022), and

while they do not have options specifically designed to transfer annotations between genomes of the graph, they do have options to project coordinates between the graph and its embedded genomes. *odgi position* and *vg inject/surject* project coordinates between the genomes of the graph and can be used to transfer annotations, with the limitation that there is no filtering based on sequence identity of coverage.

Additionally, VG recently implemented an option to annotate the graph by projecting an annotation from a genome to the graph. Although *vg inject* already performed this task, *vg annotate* is specifically designed for this purpose (Novak et al., 2024) and is easier to use. It also offers efficient ways to index an query the graph annotation.

The results and execution time of all these functions were compared to GrAnnoT.

The versions of the tools used are available in the supplementary data. The complete exact commands used for those benchmark are available online (Marthe and Sabot, 2025b). The Jupyter notebooks used for the analysis are available on our Forge (`https://forge.ird.fr/diade/dynadiv/grannot`, Marthe et al., 2025). All the data used for the analysis and the outputs are available online (Marthe and Sabot, 2025a,b).

2.2. Comparison of the transfers

Comparison with other tools

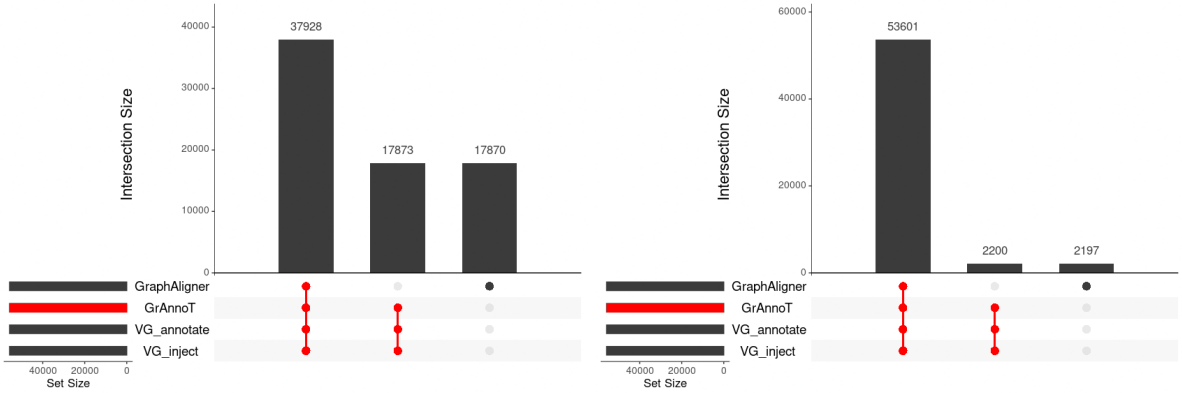
Results were evaluated for the two types of transfers that GrAnnoT can perform: from genome to graph and from genome to genome. In both cases, the transfers were performed with the different tools described before when possible. Then, for each transferred feature, its positions provided by the different tools were compared. Given a feature, we consider two transfers as different if they placed the feature at different positions. A transfer is specific to a tool if it is different from all the other transfers. By definition, a feature transfer is also specific to a tool if the feature is only transferred by this tool.

We tested GrAnnoT, GraphAligner, VG *inject* and VG *annotate* for the transfer of the annotation of the cv Nipponbare (Kawahara et al., 2013b) to the rice pangenome graph. We tested GrAnnoT, Liftoff, VG *inject/surject* and ODGI *position* for the transfer of the annotation of the cv Nipponbare to the cv Azucena.

Genome to graph transfer:

For the genome to graph transfer, only the gene features were used. Indeed, VG *annotate* requires the GFF3 features to have a *ID* and a *Name* attribute to be transferred, while GrAnnot does not. It was the case only for the gene and mRNA features in the IRGSP annotation of the cv Nipponbare. For simplicity, we thus only selected the gene features in this annotation, as the mRNA are always included in a gene, and did not add other type of features to test the transfers.

The three methods that do not perform alignment (GrAnnoT, VG *inject* and VG *annotate*) have the exact same results for all the features. For ~32% of the features transferred by GraphAligner (17,870 features out of 55,798), the output is different from the other tools (Figure 2a). However, when allowing a difference of 1 bp on the position on the path, ~88% of the GraphAligner-specific transfers (15,673 out of 17,870 transfers) are then considered identical to the transfers from the other tools (Figure 2b). Further verification showed that these 1 bp differences from GraphAligner are alignment errors, where 1 bp is missing in 5' or 3' in the transferred feature sequence. Such differences are minor and acceptable for certain applications, but not in the context of annotation. Because of that, the current version of GraphAligner does not seem to be suitable for precise annotation transfer.



(a) Features transferred at the exact same position **(b)** Features transferred at a distance of 1 bp maximum

Figure 2 – Genome to graph transfer comparison, Upset representation. Each vertical bar represents the number of identical transfers between the different tools specified below the bar. Two transfers are considered identical if they placed the feature at the exact same path in the graph and either at the exact same position on the nodes (a) or at a distance of maximum 1 nucleotide (b). The horizontal bars on the left represent the total number of transfers for each tool. GrAnnoT transfers are highlighted in red.

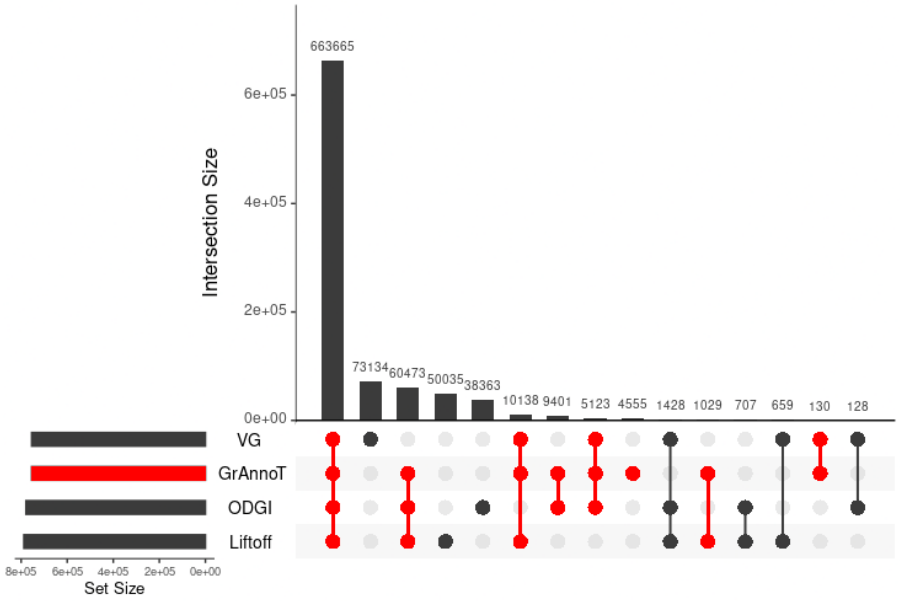
Genome to genome transfer:

Most of the transfers between genomes are identical between the four tools (663,665/918,973≈72%). GrAnnoT seems to be the most consensual tool as it has the least specific transfers (Figure 3a) compared to the other tools. Liftoff has the most transfers, and seems to perform better than other tools (see below).

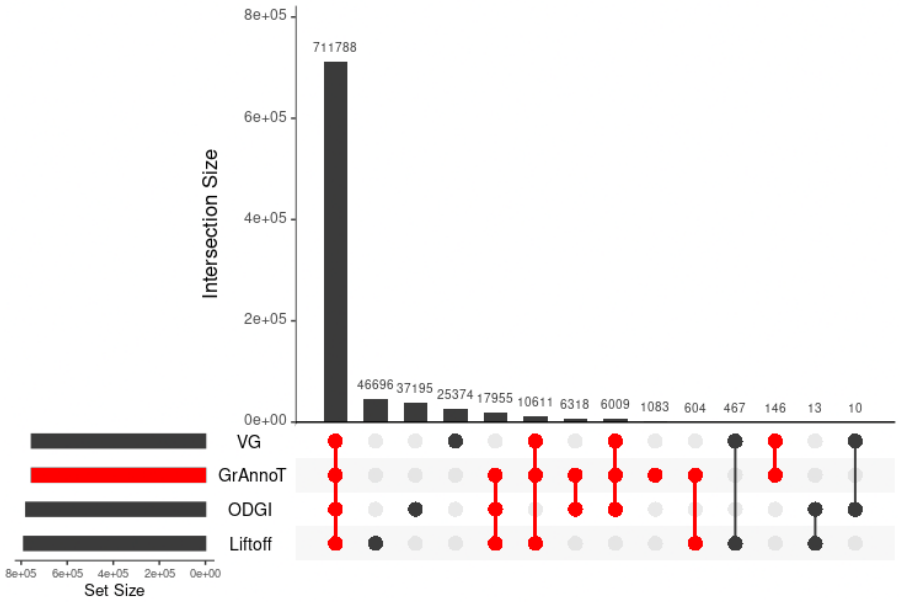
When looking at the tool-specific transfers, VG stands out the most, with 73,134 specific transfers. However, when allowing a difference of 10 bp between the transfers, VG has ~65.3% less specific transfers. Some of these VG specific transfers were manually compared to the transfers from the other tools for the same feature, and were identified as errors from VG (see supplementary Figure S10 for an example). This suggests that VG produces small errors during the *surject* step, since VG *inject* was shown to have the exact same results as GrAnnoT in Figure 2a. The 10bp difference tolerance revealed Liftoff and ODGI as the most divergent tools (with 46,696 and 37,195 specific transfers, respectively; Figure 3b).

Regarding the Liftoff-specific transfers, most are features that only Liftoff can transfer. Indeed, ~56% of them are inter-chromosomal translocations, *i.e.* features that are on a different chromosome between the source and the target genome (Figure 4). These transfers cannot be performed with GrAnnoT, VG or ODGI, as variation graphs are currently built chromosome-per-chromosome to reduce complexity, and therefore cannot represent such events (Andreace et al., 2023; Mergez et al., 2024). Thus, features on different chromosomes between Nipponbare and Azucena cannot be transferred by any of the graph-based approaches, and are found only by Liftoff. This could explain why Liftoff has the most transfers between the four tools.

Furthermore, when the annotations of these Liftoff specific-transferred features were thoroughly looked at, it appeared that they are enriched in transposable elements (TE) (p-value < 0.01; Figure 4). This could explain this discrepancy of chromosomal location between the two varieties, since transposable elements are mobile in the genome and can jump between chromosomes (Hayward and Gilbert, 2022; Wicker et al., 2007). The Liftoff-specific transfers that are on the same chromosome are also enriched in TEs (p-value < 0.01; Figure 5), as their ability to



(a) Features transferred at the exact same position



(b) Features transferred at a distance of 10 bp maximum

Figure 3 – Genome to genome transfer comparison, Upset representation. Each vertical bar represents the number of identical transfers between the different tools specified below the bar. Two transfers are considered identical if they placed the feature either at the exact same positions on the target genome (a) or at a distance of maximum 10 nucleotide (b). The horizontal bars on the left represent the total number of transfers for each tool. GrAnnoT transfers are highlighted in red.

move in the genome makes them often not syntenic: encoding the relationships between such TEs in the graph with the current variation graph tools still seems complex (Eizenga et al., 2020). Indeed, the graph we used was built by minigraph-cactus, that aligns genomes to the graph in construction (Li et al., 2020). During this process, a non-syntenic region is more difficult to align, and its sequence can be represented in the variation graph by two different nodes carrying the same information. Because of that, some duplications, inversions, translocations, and TEs are not

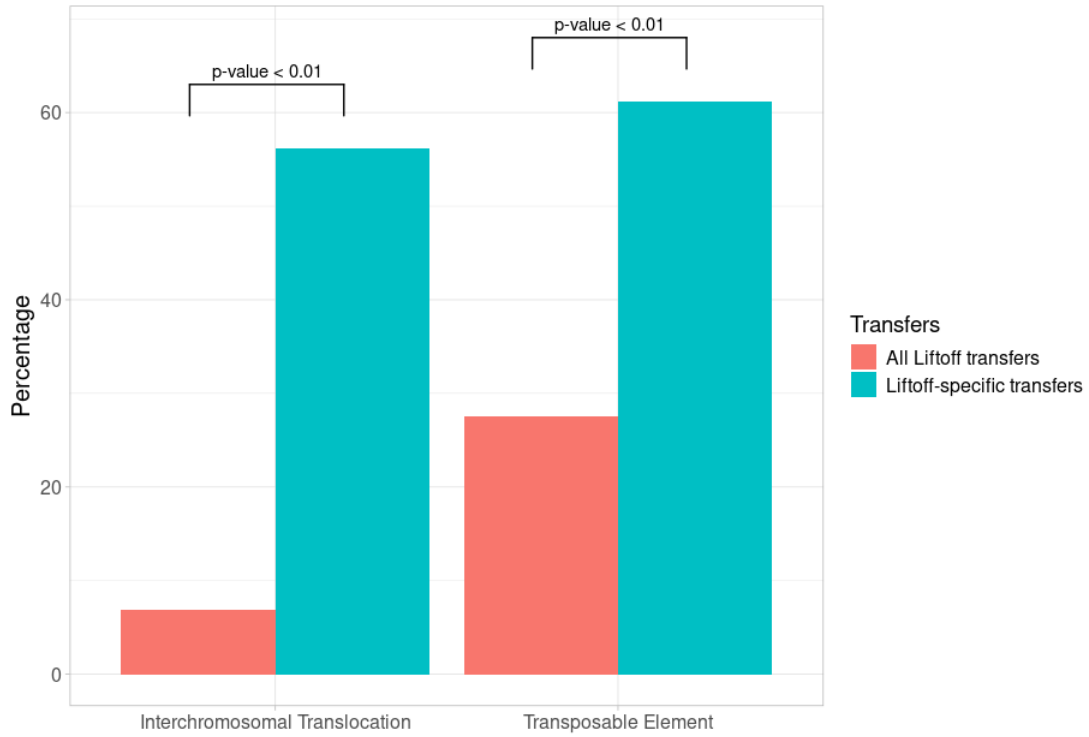


Figure 4 – Transposable element and inter-chromosomal translocation percentages in all Liftoff transfer *vs* Liftoff-specific transfers. The Liftoff-specific transfers are enriched in interchromosomal translocations and transposable elements compared to all the other Liftoff transfers. Detailed data and *p*-value calculation are available in supplementary data (Tables S6 and S7).

detected in the graph (Lemaitre, 2021; Romain et al., 2025), and annotations in these regions are not transferrable using only the variation graph structure. Since the relationships between the TEs are not always correctly encoded by the graph, TE annotation transfers cannot be reliably performed by tools such as GrAnnoT, VG or ODGI, that only use the structure of the graph.

Overall, Liftoff-specific transfers seem valid, and demonstrate a limitation in the variation graph approaches: they completely rely on the variation graph structure, which is not perfect and struggles to connect non-syntenic shared elements. The three graph tools tested are thus not suited for studying mobile sequences, such as interchromosomal translocations or TEs.

Most of the ODGI-specific transfers place a feature on a very small interval on the target genome. For instance, among the 37,195 ODGI-specific transfers, ~65% of the features (24,058) are placed on an interval of a length of zero nucleotide, and ~30% (11,320) on an interval of a length of one nucleotide. These transfers should be discarded, as they are of no biological meaning in terms of genes.

Robustness

Back and forth transfer: Two consecutive transfers (back and forth) with Liftoff and GrAnnoT allowed to compare how conservative these tools are. The first transfer was performed from the cv Nipponbare to the cv Azucena with the two tools. Then, the resulting Azucena GFF3 file was used as source to perform the second transfer, from the cv Azucena back to the cv Nipponbare. The resulting GFF3 for the cv Nipponbare was compared to its first original annotation in order to measure the loss or corruption of information during these transfers.

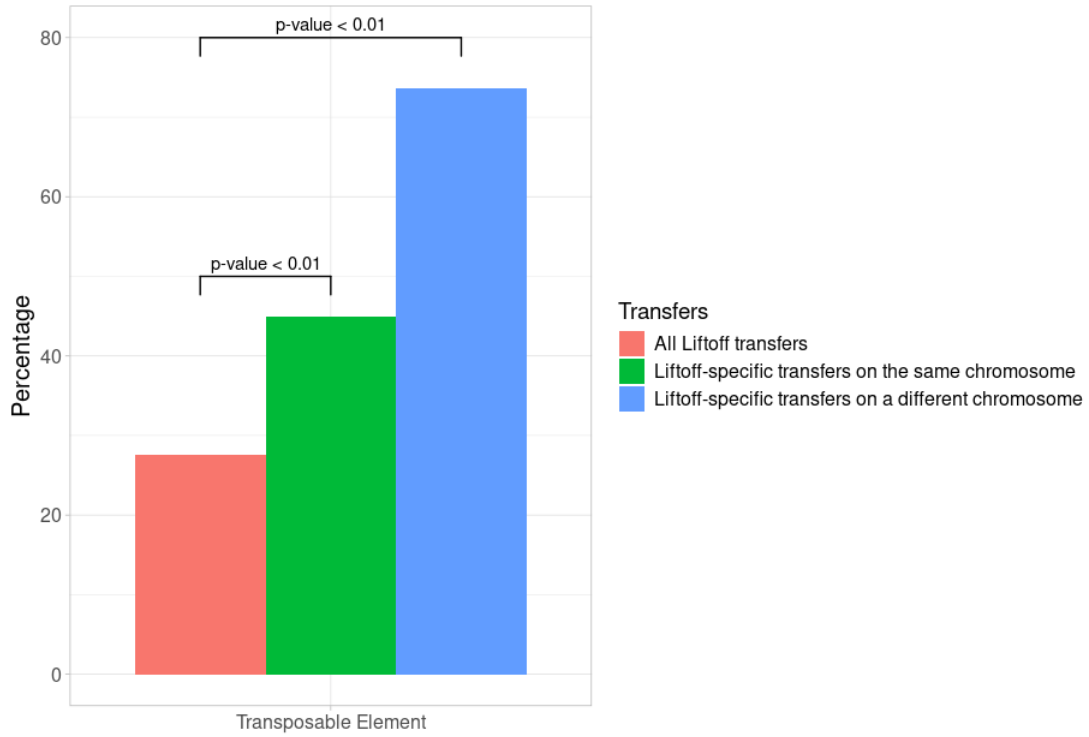


Figure 5 – Transposable element percentages in all Liftoff transfer *vs* Liftoff-specific transfers. Compared to the other Liftoff transfers, Liftoff-specific transfers are enriched in transposable elements, whether or not they are in a translocation. Detailed data and *p*-value calculation are available in supplementary data (Tables S8 and S9).

Liftoff loses less features during the two-round process (Table 1). This can be explained by the fact that Liftoff is better at finding non-syntenic features and handles interchromosomal translocations, as shown previously. However, while GrAnnoT did not lose any more annotation on the way back to the original sequence, Liftoff lost 257 features in the way back. In addition, when comparing the positions of the features before and after the two transfers (original *vs* transferred twice), GrAnnoT shows better results that Liftoff, with only 3.8% of the features being located at a different position compared to the original annotation, *versus* 10.4% of discrepancies for Liftoff. Moreover, after manual verification, it appeared that all the features misplaced by GrAnnoT in the second transfer are features where an extremity was shortened during the first transfer due to a deletion at the beginning or the end of the feature. Thus, while the feature transferred during the second transfer was incomplete regarding the true annotation, the transfer itself occurred correctly.

Finally, some features transferred by Liftoff back to the cv Nipponbare are placed on a different chromosome than the original one, as the transfer is alignment-based only and does not rely on synteny. In this regard GrAnnoT is more conservative than Liftoff. Indeed, orthologous copies are sometimes considered to guarantee a better conservation of gene function compared to paralogous copies, according to the ortholog conjecture (Nevers et al., 2020; Rogozin et al., 2014). As the variation graph conserves the synteny, GrAnnoT is more likely to transfer annotations between orthologous copies than between paralogous copies.

Overall, while GrAnnoT transfers less annotations than Liftoff (as seen in Figure 3a), it is more conservative with the annotations it does transfer.

	GrAnnoT	Liftoff
Loss in first transfer	7,961	1,622
Loss in second transfer	0	257
Total loss	7,961	1,879
Same position	47,184	48,482
Different position	841	5,625
1-10bp difference	393	360
11-100bp difference	275	648
101-1000bp difference	165	776
>1000bp difference	8	1,210
Different chromosome	0	2,631
Total transfers	48,025	54,107

Table 1 – GrAnnoT and Liftoff comparison on back and forth transfer. The input annotation for first transfer included 55,986 features. The loss corresponds to the number of features not transferred in either transfer (first cv Nipponbare to cv Azucena or second cv Azucena to cv Nipponbare). The other rows show how many features were at the same or at different positions before and after the two transfers.

	Nipponbare reference	Natel Boro reference
Total transfers	48,025	45,946
Loss	7,961	10,040
Specific transfers	2,376	297
	Comparison between the two graphs	
Common transfers	45,256	
Different transfers	393	
1-10bp difference	169	
11-100bp difference	87	
101-1000bp difference	76	
>1000bp difference	61	
Different chromosome	0	

Table 2 – Comparison of GrAnnoT transfers using graphs with different reference genomes. The input annotation for the transfer included 55,986 features. The loss corresponds to the number of features not transferred. The other rows show how many features were placed at the same or at different positions when transferred with the two graphs.

Impact of the reference genome for graph construction: The variation graphs used were built with minigraph-cactus, which requires a reference genome as anchor, that can thus bias the graph structure (Andreace et al., 2023). To test the replicability of the GrAnnoT approach, transfers through two different graphs were compared. The two graphs have the same genomes embedded (13 Asian rice genomes), but were built with a different reference genome to initiate the graph. The reference genomes used for the two graphs are the annotated genome IRGSP-1.0 (cv Nipponbare), and Os127652RS1 (cv Natel Boro) (Zhou et al., 2020). Annotation transfer from cv Nipponbare to cv AzucenaRS1 was performed with these two graphs, and the positions of the common transferred features were compared.

Among the 48,322 features transferred on the cv Azucena, 2,673 (~5.5%) were not transferred by both graphs. Among the 45,649 features transferred by both graphs, only 393 (~0.9%) were not transferred at the same exact location (Table 2).

The amount of features not transferred by both graphs is not negligible, even though they mostly consist of TEs (see before). However, it can be explained by the choice of the reference genome for the second graph construction, the cv Natel Boro. Indeed, among the 11 genomes in the graph that are not involved in the transfer (neither the cv Nipponbare nor the cv Azucena), the cv Natel Boro is among the furthest, genetically speaking, as shown in the phylogenetic tree in the genomes original paper (Zhou et al., 2020). Thus, it makes sense that the graph centered around the cv Nipponbare displays better performance for annotation transfer from the cv Nipponbare. This showcases the importance of the choice of the reference genome for the graph construction, that must be adapted to the use case of the graph.

Comparison with other species: GrAnnoT was compared to Liftoff using two other datasets: a pangenome variation graph of the human chromosome 1 (Liao et al., 2023) and an *E. coli* pangenome variation graph (**data_ecoli**), both made with minigraph-cactus. For the rice graph, the transfer was again made from the cv Nipponbare to the cv Azucena; for the human graph, the transfer was made from the CHM13 to the GrCH38 haplotype; for the *E.coli* graph, the transfer was made from the O157_H7_EC4115_0a2c271 strain to the S88_fa4fe08 one. These comparisons checked if the positions of the features transferred by both approaches are consistent, to assess if the results observed in the rice pangenome graph were replicable with graphs from other type of dataset/organisms/phylum.

It appears that for the rice and human datasets, most of the features are transferred by both tools (~85.7% for rice and ~91.3% for human) (Table 3). For *E. coli*, only ~62.5% of features are transferred by Liftoff and GrAnnoT, and both tools have relatively low transfer rates (below 70%, see Table 3). This suggests that the features not transferred by either tool reflect a difference in gene content between the two strains, rather that a technical error.

Additionally, a large part of the features transferred by both tools are placed at the exact same position by Liftoff and GrAnnoT (~96.8% for rice, ~99.6% for human and ~95.5% for *E. coli*). As expected, some features are transferred only by Liftoff, but for the human graph GrAnnoT-specific transfers appear in negligible quantities. This better transfer capacity for the two tools in human may be due to the lesser diversity of human genomes compared to rice (mean 15.6 millions SNP for 64 human haplotypes *vs* 9.4 millions for only 16 rice ones, respectively; Ebert et al., 2021; Wei et al., 2024), and even more so compared to *E. coli*. In addition, the annotation of human genes is probably better curated than in rice, with less hypothetical genes that may be false positive, also explaining the better transfer for both tools on human reference.
Liftoff-specific transfers for E.Coli were manually inspected and most of them are related to unknown protein domains or related to biotic and abiotic stress responses. Such type of genes in bacteria are generally related to mobile structures such as ICE, e.g. Zheng et al., 2023.

Scalability:

To ensure GrAnnoT can work with larger datasets, it was tested on a variation graph build using minigraph-cactus with 69 *A.thaliana* genomes (Lian et al., 2024; Mergez et al., 2024). GrAnnoT was used to transfer the annotation of the genome *Abd-0* to the graph and then to all 68 other genomes. This operation was performed in ~2h52min, with an average of ~2min30sec per transfer (including the transfer from *Abd-0* to the graph).

	Rice		Human Chr1		E.coli	
Total features to transfer	55,986		282,668		11,460	
Features transferred by GrAnnoT	48,025	85.78%	276,681	97.88%	7,407	64.63%
GrAnnoT-specific transfers	72	0.13%	4,647	1.64%	239	2.09%
Features transferred by Liftoff	54,363	97.10%	275,249	97.38%	7,939	69.28%
Liftoff-specific transfers	6,410	11.45%	3,264	1.12%	767	6.69%
Features transferred by both tools	47,951	85.65%	258,092	91.31%	7,167	62.54%
Same position	46,431	82.93%	256,927	90.89%	6,844	59.72%
Different position	1,520	2.71%	1,165	0.41%	323	2.82%
1-10bp difference	795	1.42%	623	0.22%	184	1.61%
11-100bp difference	284	0.51%	150	0.05%	70	0.61%
101-1000bp difference	155	0.28%	46	0.02%	15	0.13%
>1000bp difference	164	0.29%	346	0.12%	54	0.47%
Different chromosome	122	0.22%	0	0%	0	0%
Runtime Liftoff	00:23:45		00:10:27		00:00:09	
Runtime GrAnnoT	00:08:11		00:14:47		00:00:22	

Table 3 – Comparison between GrAnnoT and Liftoff in several species. Each feature in the input annotation was transferred using GrAnnoT and Liftoff. When the feature has been transferred by both tools, the two positions given were compared to see how different they are.

	GrAnnoT	Liftoff	VG *inject/surject*	ODGI *position*
Graph construction	04:23:22	-	04:23:22	04:23:22
Graph conversion/index	-	-	00:14:34	00:05:37
Annotation transfer	**00:08:11**	**00:23:45**	**07:12:17**	**70:18:04**
Total time	04:31:33	00:23:45	11:49:73	74:46:63

Table 4 – Run time comparison for genome to genome transfer in rice. GrAnnoT, VG *inject/surject* and ODGI *position* transfer through the graph, and VG inject/surject and ODGI convert/index the graph before the transfer. Liftoff directly uses the genome fasta files.

2.3. Run time comparison

The execution time for the transfer from genome to genome with the different tools was measured on Asian rice data between the cv Nipponbare and the cv Azucena, using the command */usr/bin/time* (Table 4). The graph tools use as input the annotation file in GFF3 and the variation graph in the adapted format. VG *inject/surject* (as we deal here with genome to genome transfer) and ODGI *position* require the variation graph to be converted/indexed in their format before use (.xg and .og, respectively), and GrAnnoT uses directly the variation graph in native GFA format. Liftoff uses as input the fasta files for the genomes and the annotation file in GFF3 format. The results show that GrAnnoT has the best run time, and that ODGI and VG *inject/surject* are substantially slower that GrAnnoT and Liftoff.

GrAnnoT was further compared to Liftoff in terms of run time.

Several transfers were performed with both tools to compare the run times, because GrAnnoT is designed to facilitate the transfer toward multiple target genomes: it starts by pre-processing the variation graph and loading the graph annotation, which only needs to be done once, no matter how many target genomes are included in the graph.

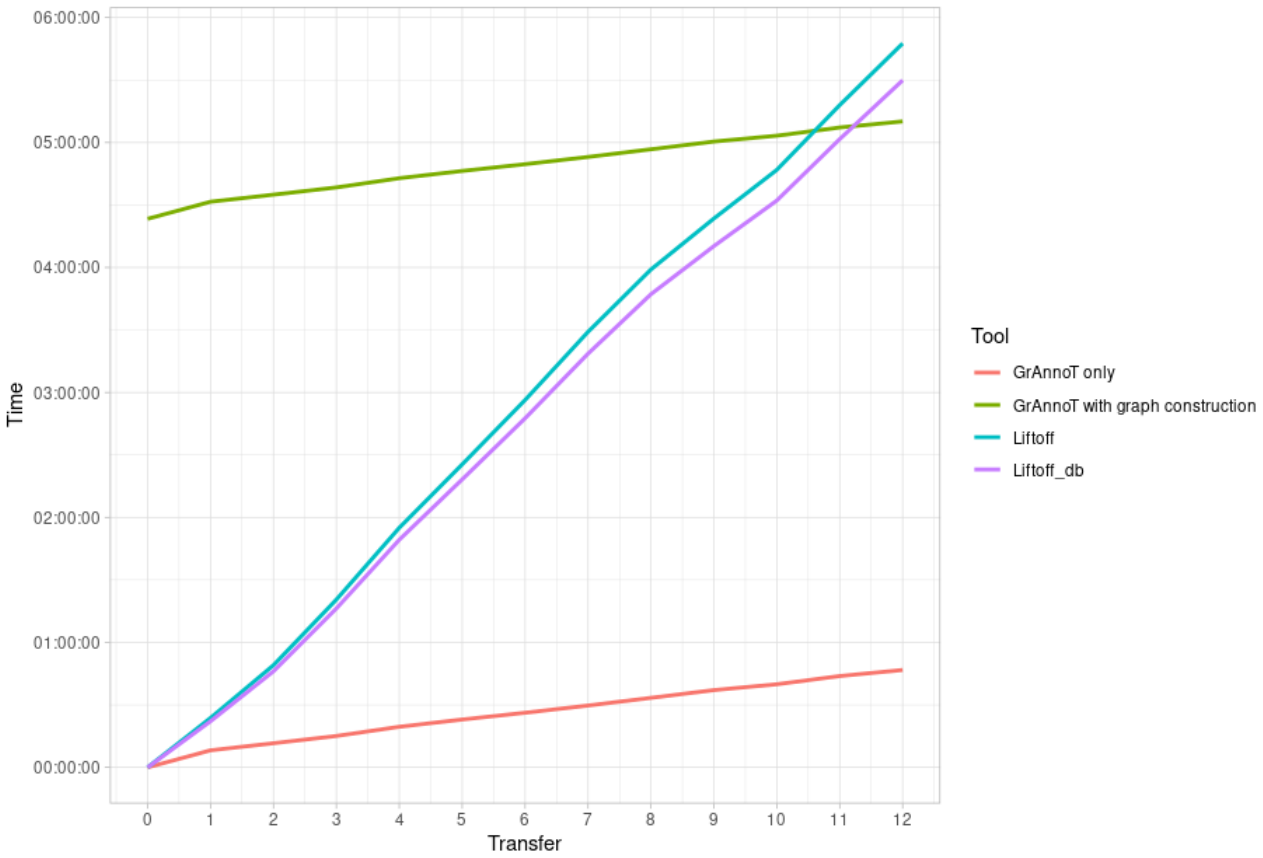


Figure 6 – Genome to genome transfer comparison. GrAnnoT and Liftoff run time for 1-12 transfers were measured using the command */usr/bin/time*. Liftoff was run both in GFF and DB mode. GrAnnoT values are presented including or not the graph building time. Detailed time points are available in supplementary data in Table S10.

Liftoff can be run in GFF mode or in database mode; the database mode needs less time since the GFF annotation file has already been processed. Both of these mode were compared to GrAnnoT.

The commands timed are transfers from the cv Nipponbare to all the other genomes in the rice pangenome graph, with 12 transfers in total.

The results show that GrAnnoT is faster than Liftoff to perform one annotation transfer (~8 minutes *vs* ~22 minutes), and even more to perform twelve (~47 minutes *vs* ~5 hours and 30 minutes, see Figure 6). However, this comparison does not take into account the time needed to build the graph. When adding the graph construction time (~4h23mn on our infrastructure) to the GrAnnoT 12 transfers time, we still get a duration (~5h10mn) slightly shorter to Liftoff transfers (~5h47min or ~5h29min). Additionally, GrAnnoT can give supplementary informative output that describe the transfers performed such as a presence-absence matrix or alignment files of the transferred features, as well as the graph annotation.

For the human graph, GrAnnoT is not faster than Liftoff for one transfer (see the last lines of Table 3). However, as shown on Figure 6, for several transfers GrAnnoT is more advantageous. We tested the runtime of GrAnnoT for the annotation transfer on 10 haplotypes, and got ~45 minutes in total. This is significantly lower than the time for one transfer multiplied by 10 (~1h44min), which is what we can expect of 10 Liftoff transfers from the results in Figure 6.

3. Applications

To assess the use of GrAnnoT annotation transfer, in particular the informative outputs complementary to the GFF3 itself, we analyzed a few characteristics of the annotation transfers between the Nipponbare and Azucena cultivars. More precisely, we verified that the variations in the graph reported by GrAnnoT are distributed as biologically expected, in a way that does not disrupt the proteins coded by the gene features.

3.1. Indel rate in different feature types

We looked at the positions of the indel variants (insertion or deletion) in the different feature types that correspond to different parts of the genes. These variants are expected to be less present in the CDS compared to the rest of the gene due to selection pressure, because the resulting changes in the coded protein are more important.

The feature types that were compared are:

- the whole gene feature itself
- the mRNA
- the 5'UTR
- the exons
- the CDS
- the introns
- the 3'UTR

These feature types have different average lengths, inducing a bias in the number of indel found by feature type; if the indels are randomly distributed, we expect more indels in the feature type that has the longest cumulated length. To counter this bias, for each feature type we reported the total number of indels found to its cumulated length, obtaining the average number of indel per position.

The results displayed in Figure 7 show that, as expected, the CDS have the fewest indels and the non-coding regions (UTR and introns) have the most. This confirms that the variations in the graph reported by GrAnnoT are consistent with the current understanding of genome variation selection.

3.2. Frameshift mutations in different feature types

Indels can modify the protein coded by a gene, but indels in CDS are particularly impactful when they change the reading frame. We calculated the rate of frameshift mutations (indel whose length is not a multiple of 3) among the indels, for all feature types. We expect to have a lower ratio of frameshift mutations in the CDS compared to the non-coding regions, because of the selection pressure.

The results displayed in Figure 8 show that the CDS have the lowest percentage of frameshift variation from their indels, and that the introns have the highest.

3.3. Substitutions position in different feature types

Substitutions are usually smaller variants than indels, so they are expected to have a smaller impact. However their distribution in CDS is not expected to be uniform. Indeed, substitutions on the third position of a codon is more likely to be silent than a substitution on the two other

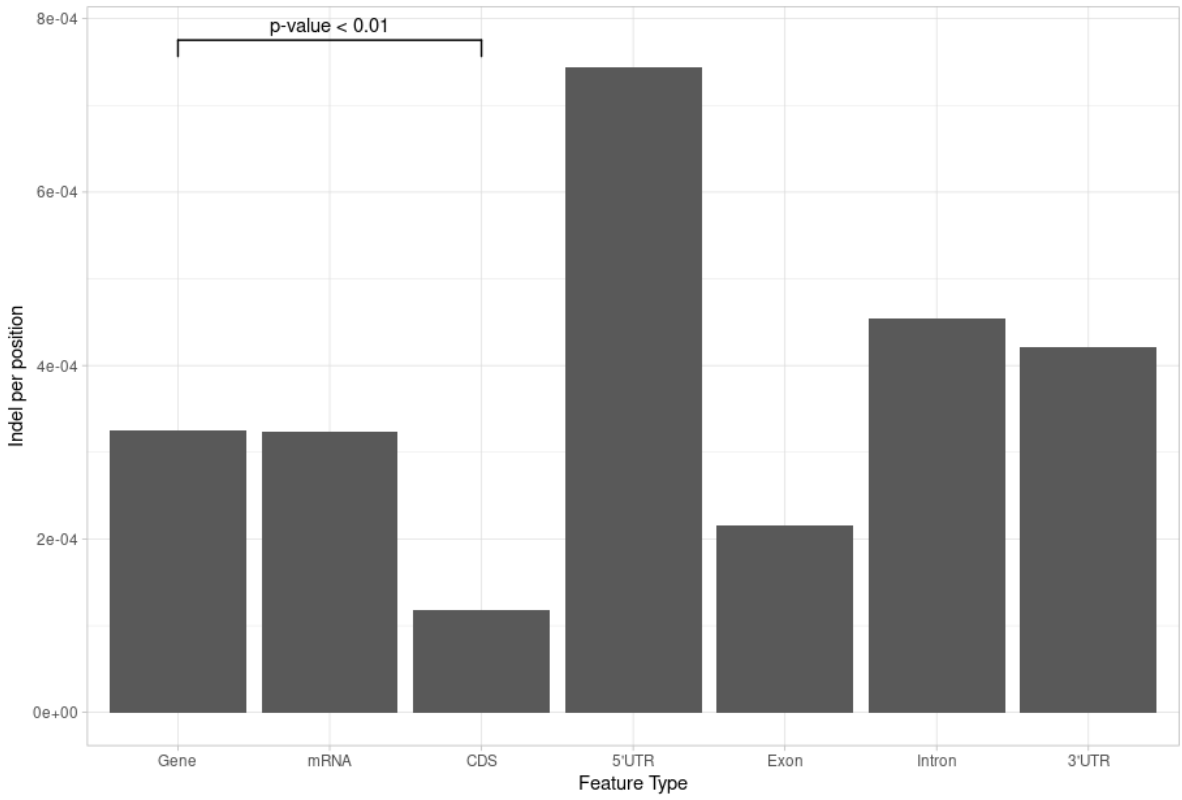


Figure 7 – Indel count. Each bar represents the number of indels (insertion or deletion) per position in the corresponding feature type. As expected, the CDS are the most conserved and thus have the least indels, and the non-coding regions (UTR and introns) are the least conserved and have the most indels. Detailed data and *p*-value calculation are available in supplementary data (table S11).

positions. Because of that, in CDS the third codon position usually has more substitutions than the two other positions (Sanchez et al., 2005).

On Figure 9, we show that the CDS indeed has more substitutions on the third codon position than the other two positions, while the other gene elements have more homogeneous substitution distributions. Details about how we computed the substitution position in the CDS are available in supplementary data (3).

3.4. Pangene set analysis

The PAV matrix output was computed for all the genomes in the Asian rice graph (minus the source genome cv Nipponbare), and was used to compute the core, dispensable and shell gene set from the cv Nipponbare in this pangenome.

We found ~58% of core genes and ~33% of dispensable genes (table 5) in our variation graph, which is similar to what is seen in the literature when accounting for the different threshold chosen in each study (with ~53-62% of core and ~38% of dispensable gene families for instance; Wang et al., 2018).

Discussion and conclusion

Annotation of pangenomes is a broad and crucial topic, that has been explored in various ways depending on the type of pangenome considered. Some pangenomes consist in a gene set, where several genomes are separately annotated and the resulting genes are clustered into

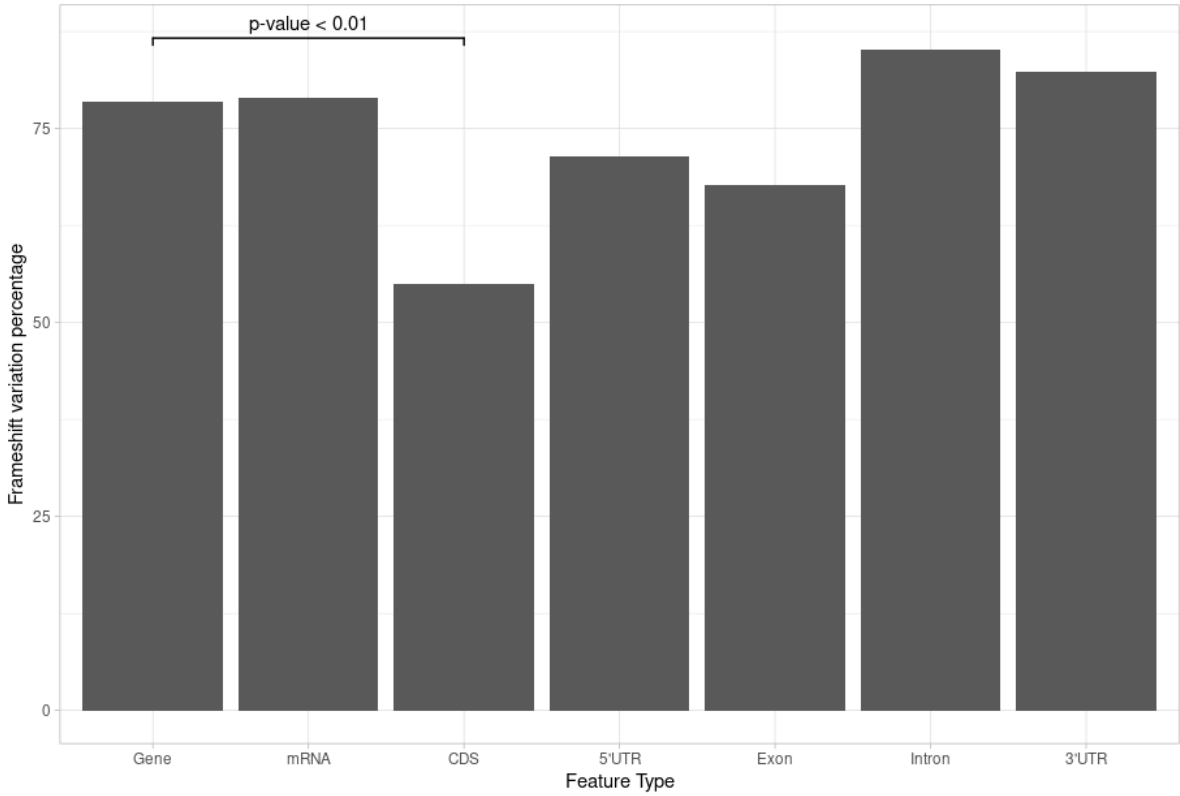


Figure 8 – Frameshift indel count. Each bar represents the percentage of frameshift variant (length not multiple of 3) among all the indels in each feature type. As expected, the CDS have the fewest frameshift variant, since these variants impact significantly the protein coded. Detailed data and *p*-value calculation are available in supplementary data (table S12).

	Core genes	Dispensable genes	Shell genes
Presence percentage	100% - 95%	95% - 10%	10% - 0%
Number of genes	32,537	18,403	5 046
Percentages of genes	58.1%	32.9%	9%

Table 5 – Core, dispensable and shell gene set. The population size is 12, and there are 55,986 genes in total.

orthologous groups (Gautreau et al., 2021; Gordon et al., 2017). Alternatively, pangenomes can be stored in de Bruijn graphs, that ggCaller (Horsfield et al., 2023) can annotate *de novo* (for bacteria), or that Pantools (Jonkheer et al., 2022) can visualize and project annotation on this visualization. However the variation graph structure has less manipulation options, and most of them rely on the visualization on the graph (Liu et al., 2024; Miao and Yue, 2025). Visualizing whole variation graphs can be challenging, as these structures are usually complex and non-linear for large genomes, resulting in a hairball-like structure which is difficult to interpret (Durant, 2022). It is easier to project and visualize only a few annotations on a targeted region of the graph to study the variations, but this approach is less scalable, poorly reproducible, and usually requires to know in advance which regions to study.

GrAnnoT fills this gap by efficiently projecting an annotation on a variation graph and its embedded genomes, and providing useful information on the variants in the annotated regions between the different genomes embedded in this graph. It is easy to install and to use, and the

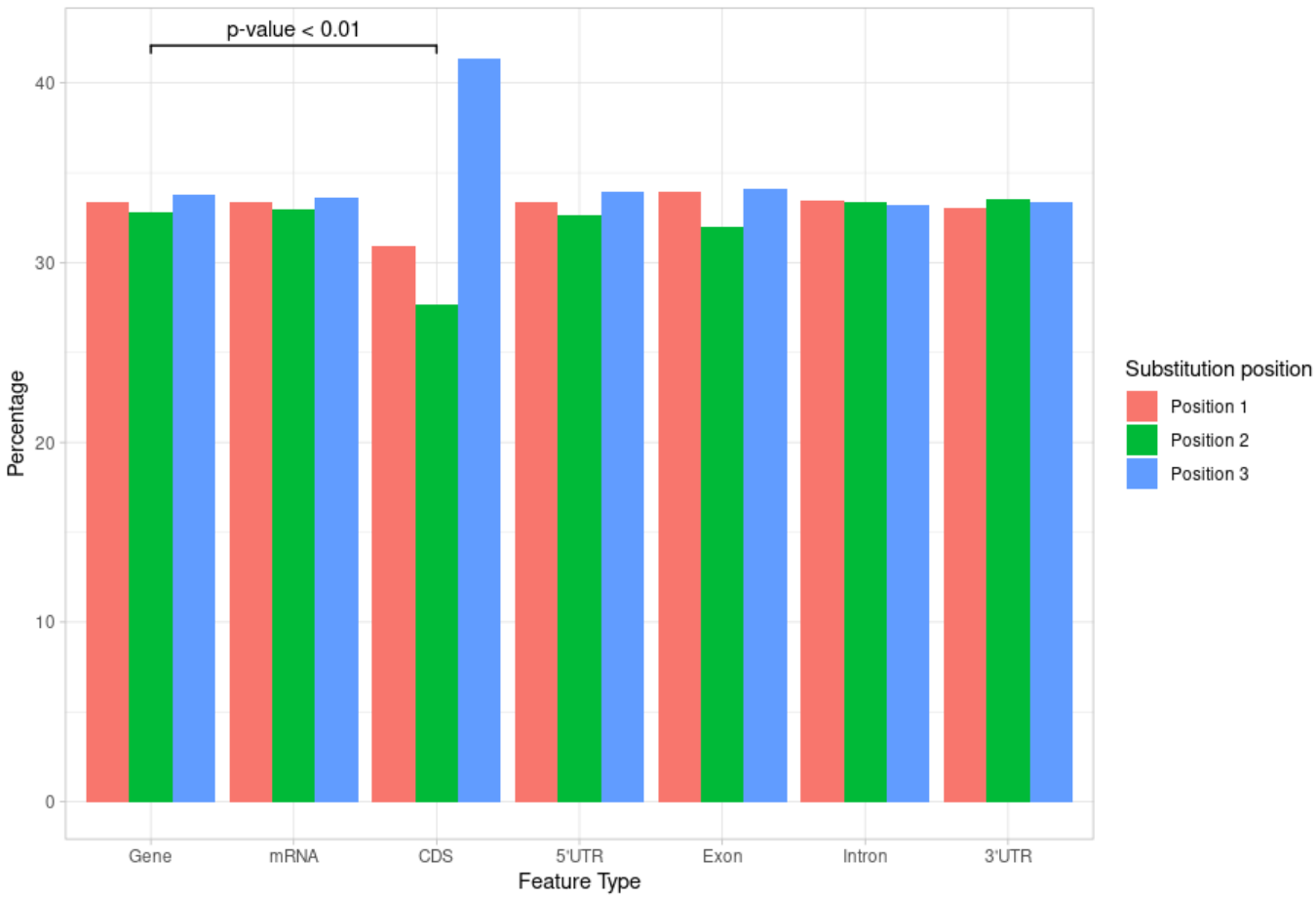


Figure 9 – Substitution positions. For each feature type, the percentage of substitutions that are on each of the three codon positions is displayed. In the CDS, the third position has more substitutions than the two other positions. For the other feature types, we don't see that the positions multiple of 3 have more substitutions than the others. Detailed data and *p*-value calculation are available in supplementary data (table S13).

operations it proposes are fast, reliable, reproducible, and it gives informative and user friendly outputs. Once the annotation has been transferred in the variation graph and its variations between the genomes have been reported, subgraphs of the regions or genes of interest can be extracted with tools like VG or ODGI, or GrAnnoT in the future. A visualization of these subgraph can help study in more details the variations found in the regions targeted with GrAnnoT.

The main limitation of GrAnnoT, as displayed in this paper, comes from the variation graph itself : GrAnnoT solely relies on this graph structure and does not perform any alignment (in order to conserve speed); thus, any flaw in the variation graph will hinder GrAnnoT in its transfers. As a first example, graphs built by minigraph-cactus (such as the ones we used) separate the chromosomes to build independent graphs, and thus cannot represent interchromosomal events (such as translocation or transposition). In addition, the example of the transposable elements in the Asian rice dataset highlights that minigraph-cactus currently struggles to align some non syntenic elements (transposed copies), even on the same chromosome. Therefore, those annotations cannot be transferred by GrAnnot. GrAnnoT can handle gene duplication and inversion when the information is present in the variation graph; however, the detection of these variants (by minigraph-cactus) can be inconsistent (Lemaitre, 2021; Romain et al., 2025). When the graph

builders will be able to inform on interchromosomal relationships and duplications, GrAnnot will be able to immediately transfer the annotations corresponding to these regions.

Indeed, applying GrAnnoT to a graph built with another tool than minigraph-cactus would be an interesting option, to see if other graph builders are better at representing non-syntenic relationships.

To test that, GrAnnoT has been applied to PGGB graphs, as this graph builder has the benefit of being completely reference-free. The annotation transfer on the PGGB graph was performed correctly, but the transfer on a target genome was very slow and erroneous in repeated regions, for reasons not yet identified (data not shown). One possible explanation could be the tendency of PGGB graphs to have many cycles, at the opposite of minigraph-cactus graphs (Andreace et al., 2023). This resulting complexity could be the reason for the high runtime and errors of graph-to-genome transfers, as GrAnnoT algorithms are not suited for such complex graph topologies. However, in the future, we plan to improve the compatibility between PGGB and GrAnnoT, in order to be adaptable to more type of variation graphs.

Minigraph is another tool that builds pangenome graphs, but the output format (rGFA format) is incompatible with GrAnnoT as its graphs do not include paths or walks, and thus cannot inform to which path belong the current annotation.

GrAnnoT was compared to existing tools that perform some similar operations, but none correspond exactly to its scope or its full capacities.

For annotation transfer to the graph, GrAnnoT was compared with VG annotate, and the two tools gave the exact same results. Both are straightforward to use, and are good options. VG inject has the same behavior (outside of its complexity in use) as VG annotate, and thus is also efficient. On the opposite, GraphAligner is too slow and showed some errors in placing gene borders. The runtimes are in the same order of magnitude between GrAnnoT and VG annotate, with the VG strategy consisting in investing time to index the graph for rapid individual transfers later on, while GrAnnoT uses directly the GFA in its native form.

For annotation transfer between genomes, GrAnnoT has showed good results and is comparable to Liftoff in terms of performance for syntenic elements. As previously mentioned, Liftoff is better at transferring non syntenic annotations, and should thus be preferred by users interested in such elements. In terms of CPU time, once the graph is built GrAnnoT is faster than Liftoff, and the graph construction time can be compensated when performing several transfers. While GrAnnoT can transfer annotations between genomes as accurately as Liftoff, its primary objective is to integrate annotations to a pangenome variation graph, and therefore to inform the user of the variability of the population in terms of genes, and of the variation within the genes sequences. In comparison, Liftoff only provides a percentage of similarity, and not any alignment or information on the variants.

For performing both tasks, VG offers the commands *inject/surject*, and ODGI the command *position*. However the runtimes for these tools are prohibitive, and their transfers are not reliable, probably because they were aimed to transfer a few set of coordinates in non-complex, human genomic regions.

Another addition in GrAnnoT compared to VG and ODGI is the filtering based on sequence identity and coverage, which ensures that a substantial part of the annotated feature is present

in the target genome. The threshold used can be easily modified by the user, to adapt to the species, pangenome diversity, graph quality, type of feature annotated, etc.

In the future, we plan to improve GrAnnoT capacity to annotate a pangenome variation graph by including more than one annotation. Indeed, GrAnnoT current approach for transferring annotation comes with the downside of only informing on regions that are shared with the originally annotated genome. However, one advantage of the pangenome is to study dispensable regions, that are not shared by every genome and thus that are not always present in the annotated genome. When available, integrating annotations from several genomes in the variation graph would give a more complete picture of the gene set in the population, and help study the whole pangenome.

Another future development of GrAnnoT could be to aim to reduce the loss of non-syntenic annotations. For that, we would first need to detect them, by selecting gene annotations poorly transferred for example. Then, using a graph alignment tool, we could find all the different sets of nodes representing their sequence, *i.e.* all the different locations of the annotated sequence in the graph. This approach would help transfer annotations lost due to transposition or chromosomal rearrangement, but requires accurate detection of elements not transferred because of the graph structure. Indeed, the alignment of gene sequences on the graph is time consuming, and naively aligning all the genes would be too long and redundant, as the graph correctly aligns most of the genomes regions.

In conclusion, the present study introduced GrAnnoT, the first tool able to efficiently transfer annotation on a pangenome variation graph from one of its embedded genomes and reverse. It relies on the already performed alignment that created the graph to identify syntenic segments. We benchmarked GrAnnoT on Asian rice, bacteria and human pangenomes, and showed that it is fast, scalable, reliable and efficient, and performs adequately compared to state-of-the-art tools for linear genomes. It is a robust, replicable tool working on any type of species for which a variation graph is available. In addition, GrAnnoT can provide useful outputs, such as the alignments of the gene sequence between source and target, or a presence/absence matrix.

Acknowledgements

The authors want to thank Sebastien Ravel from PHIM/CIRAD for his valuable discussions and help on packaging and containerization.

The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/- http://www.southgreen.fr

Fundings

This work is part of project AgroDiv of the Agroecology and Digital Technologies research program and received government funding managed by the Agence Nationale de la Recherche under the France 2030 program, reference ANR-22-PEAE-0005.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: FS is a PCI Genomics recommender. The LLM agent LeChat was used to prepare the initial abstract section before manual editing once the manuscript was finished.

Data, script, code, and supplementary information availability

Data and results are available online (data: `https://doi.org/10.23708/DO1RTF`,Marthe and Sabot, 2025a, results: `https://doi.org/10.23708/RRSKRA`, Marthe and Sabot, 2025b). Scripts and codes are available online (`https://doi.org/10.23708/TW3KYV`, Marthe et al., 2025).

Supplementary data

Data and tools used :

- Data
  - *E.coli* graph (built with data obtained using the same protocol as Heumos et al., 2024, see below)
  - Human graph (from Liao et al., 2023)
  - Rice graph (built with data from Kawahara et al., 2013a; Zhou et al., 2020)
- Tools
  - minigraph-cactus v2.8.2 (Hickey et al., 2023)
  - Liftoff v1.6.3 (Shumate and Salzberg, 2021)
  - ODGI v0.8.6-11-ga1f169cc (Guarracino et al., 2022)
  - VG v1.58.0 (Garrison et al., 2018)
  - GraphAligner Branch master commit daec67f67a2f50d648a6aa30cbbe5a2949583061 (Rautiainen and Marschall, 2020)

NCBI ID of *E.coli* genomes used to build the graph:

- NC_000913.3
- NC_002655.2
- NC_004431.1
- NC_007779.1
- NC_008253.1
- NC_008563.1
- NC_009800.1
- NC_010468.1
- NC_010473.1
- NC_011353.1
- NC_011601.1
- NC_011741.1
- NC_011742.1

Substitution positions: To find the codon positions for the subsection 3.3, we had to take into account the splicing of the mRNA. Indeed, the CDS elements in the annotation only correspond to a fraction of the real CDS in the mRNA. Thus the substitution positions are not relative to the real CDS, and finding the third position of the codon required to add the context of the preceding

```
LOC_Os01g01050                  ACAAGTCACAGGGAGGAGTC    20
GrAnnoT_Lifotff_ODGI_transfer   ACAAGTCACAGGGAGGAGTC    20
VG_transfer                     ACAAGTCACAGGGAGGAGTC    20
                                ********************
                            ...
                            ...
                            ...
LOC_Os01g01050                  TCTAT--------CTATCTA    512
GrAnnoT_Lifotff_ODGI_transfer   tctatctatctatctatcta    520
VG_transfer                     tctatctatctatctatcta    520
                                *****        *******
                            ...
                            ...
                            ...
LOC_Os01g01050                  TATACATGACGATATGATCC    4131
GrAnnoT_Lifotff_ODGI_transfer   TATACATGACGATATGATCC    4139
VG_transfer                     TATACATGACGA--------    4131
                                ************
```

Figure S10 – Extraction of the alignment of LOC_Os01g01050 gene and its transfers on cv Azucena by different tools. VG *inject/surject* transfer appears to have an error as the positions it gives miss the last 8 bases of the gene. The gene total length is conserved in VG transfer because there is an insertion in cv Azucena in the middle of the gene.

CDS fragments. This adjustment was only done for the CDS elements in the annotation, since they are the only splited elements. This explains why the exons do not follow the CDS tendency in Figure 9, contrary to Figures 7 and 8.

	Different chromosome	Same chromosome	*P*-value
Liftoff-specific transfers	3,604	2,806	< 0.01
Other Liftoff transfers	122	47,831	

Table S6 – Interchromosomal translocation rates in Liftoff transfers. The *p*-value measures the enrichment in interchromosomal translocations in the Liftoff-specific transfers, and was computed with Pearson's Chi-squared test.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers	3,919	2,491	< 0.01
Other Liftoff transfers	11,053	36,900	

Table S7 – Transposable elements rates in Liftoff transfers. The *p*-value measures the enrichment in transposable elements in the Liftoff-specific transfers, and was computed with Pearson's Chi-squared test.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers on the same chromosome	1,263	1,543	< 0.01
Other Liftoff transfers	13,709	37,848	

Table S8 – Transposable elements rates in Liftoff transfers. The *p*-value measures the enrichment in transposable elements in the Liftoff-specific transfers on the same chromosome, and was computed with Pearson's Chi-squared test.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers on a different chromosome	2,656	948	< 0.01
Other Liftoff transfers	12,316	38,443	

Table S9 – Transposable elements rate in Liftoff transfers. The *p*-value measures the enrichment in transposable elements in the Liftoff-specific transfers in interchromosomal translocations, and was computed with Pearson's Chi-squared test.

	GrAnnoT	Liftoff GFF	Liftoff DB
1 transfer	00:08:11.64	00:23:45	00:22:08
2 transfers	00:11:36.00	00:49:03	00:46:07
3 transfers	00:15:04.47	01:20:34	01:16:18
4 transfers	00:19:29.69	01:55:02	01:49:17
5 transfers	00:22:59.03	02:25:29	02:18:16
6 transfers	00:26:13.64	02:56:24	02:47:37
7 transfers	00:29:41.71	03:29:06	03:18:38
8 transfers	00:33:23.48	03:59:03	03:47:11
9 transfers	00:37:06.03	04:23:31	04:10:16
10 transfers	00:39:54.23	04:46:52	04:32:09
11 transfers	00:43:50.98	05:17:58	05:01:38
12 transfers	00:46:45.57	05:47:34	05:29:53

Table S10 – GrAnnoT and Liftoff CPU time comparison for 1 to 12 transfers

	Positions without indel	Positions with indel	*P*-value
Gene	141,980,198	46,148	< 0.01
CDS	74,201,787	8,762	

Table S11 – Indel rates in genes and CDS. Annotations were transferred between cv Nipponbare and cv Azucena, and the number of insertions and deletions was analyzed. The *p*-value measures the enrichment in indel in gene features, and was computed with Pearson's Chi-squared test.

	Non-frameshift indel	Frameshift indel	*P*-value
Gene	9,959	36,189	< 0.01
CDS	3,950	4,812	

Table S12 – Frameshift indel rates in genes and CDS. Annotations were transferred between cv Nipponbare and cv Azucena, and the insertions and deletions lengths were analyzed. The *p*-value measures the enrichment in indel causing a frameshift in gene features, and was computed with Pearson's Chi-squared test.

	Substitutions on position 1 or 2	Substitutions on position 3	*P*-value
Gene	171,763	87,638	< 0.01
CDS	73,005	51,562	

Table S13 – Substitution positions in nucleotide triplets in genes and CDS. Annotations were transferred between cv Nipponbare and cv Azucena, and the substitution positions were analyzed. The *p*-value measures the enrichment in substitutions on position 3 of the nucleotide triplets in CDS features, and was computed with Pearson's Chi-squared test.

	Filter 50%	Filter 60%	Filter 70%	Filter 80%	Filter 90%	Filter 95%
Transferred genes	48,847	48,600	48,363	48,011	47,428	46,556
Not transferred genes	7,139	7,386	7,623	7,975	8,558	9,430
Transferred TE	11,443	11,330	11,217	11,088	10,926	10,749
Transfer rate	87.2%	86.8%	86.4%	85.8%	84.7%	83.2%
TE rate in transfers	23.4%	23.3%	23.2%	23.1%	23.0%	23.1%

Table S14 – Impact of sequence identity and coverage based filtering on annotation transfers between cv Nipponbare and cv Azucena using GrAnnoT. The first row gives the value of both parameters. The filtering has a limited effect on gene transfer (transfer rate), and does not impact the transfer of transposable elements (TE).

References

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S (2022). *Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing*. *Genome Biology* **23**. https://doi.org/10.1186/s13059-022-02823-7. URL: http://dx.doi.org/10.1186/s13059-022-02823-7.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). *Basic local alignment search tool*. *Journal of Molecular Biology* **215**, 403–410. https://doi.org/10.1016/s0022-2836(05)80360-2. URL: http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Andreace F, Lechat P, Dufresne Y, Chikhi R (2023). *Comparing methods for constructing and representing human pangenome graphs*. *Genome Biology* **24**. https://doi.org/10.1186/s13059-023-03098-2. URL: http://dx.doi.org/10.1186/s13059-023-03098-2.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020). *Plant pan-genomes are the new reference*. *Nature Plants* **6**, 914–920. https://doi.org/10.1038/s41477-020-0733-0. URL: http://dx.doi.org/10.1038/s41477-020-0733-0.

Chao KH, Heinz JM, Hoh C, Mao A, Shumate A, Pertea M, Salzberg SL (2024). *Combining DNA and protein alignments to improve genome annotation with LiftOn*. *Genome Research*. https://doi.org/10.1101/gr.279620.124. URL: http://dx.doi.org/10.1101/gr.279620.124.

Chen NC, Solomon B, Mun T, Iyer S, Langmead B (2021). *Reference flow: reducing reference bias using multiple population genomes*. *Genome Biology* **22**. https://doi.org/10.1186/s13059-020-02229-3. URL: http://dx.doi.org/10.1186/s13059-020-02229-3.

Durant É (2022). *Design of novel visual representations and tools applied to plant pangenome visualization*. Theses. Université de Montpellier. URL: https://theses.hal.science/tel-04135739.

Durant É, Sabot F, Conte M, Rouard M (2021). *Panache: a web browser-based viewer for linearized pangenomes*. *Bioinformatics* **37**. Ed. by Tobias Marschall, 4556–4558. https://doi.org/10.1093/bioinformatics/btab688. URL: http://dx.doi.org/10.1093/bioinformatics/btab688.

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y, Ren J, Santamarina M, et al. (2021). *Haplotype-resolved diverse human genomes and integrated analysis of structural variation*. *Science* **372**, eabf7117. https://doi.org/10.1126/science.abf7117. eprint: https://www.science.org/doi/pdf/10.1126/science.abf7117. URL: https://www.science.org/doi/abs/10.1126/science.abf7117.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E (2020). *Pangenome Graphs*. *Annual Review of Genomics and Human Genetics* **21**, 139–162. https://doi.org/10.1146/annurev-genom-120219-080406. URL: http://dx.doi.org/10.1146/annurev-genom-120219-080406.

Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, Gordon D, Earl D, Keane T, Eichler EE, Haussler D, Stanke M, Paten B (2018). *Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation*. *Genome Research* **28**, 1029–1038. https://doi.org/10.1101/gr.233460.117. URL: http://dx.doi.org/10.1101/gr.233460.117.

Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, Ashbrook DG, Thorell K, Rusholme-Pilcher RL, Liti G, Rudbeck E, Golicz AA, Nahnsen S, Yang Z, Mwaniki MN, Nobrega FL, et al. (2024). *Building pangenome graphs*. *Nature Methods* **21**, 2008–2012. https://doi.org/10.1038/s41592-024-02430-3. URL: http://dx.doi.org/10.1038/s41592-024-02430-3.

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R (2018). *Variation graph toolkit improves read mapping by representing genetic variation in the reference*. *Nature Biotechnology* **36**, 875–879. https://doi.org/10.1038/nbt.4227. URL: http://dx.doi.org/10.1038/nbt.4227.

Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, Perrin A, Médigue C, Calteau A, Cruveiller S, Matias C, Ambroise C, Rocha EPC, Vallenet D (2021). *PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph*. *PLOS Computational Biology* **17**, e1009687. https://doi.org/10.1371/journal.pcbi.1009687. URL: http://dx.doi.org/10.1371/journal.pcbi.1009687.

Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J, Lipzen A, Dochy N, Phillips J, Barry K, Geuten K, Budak H, Juenger TE, Amasino R, Caicedo AL, et al. (2017). *Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure*. *Nature Communications* **8**. https://doi.org/10.1038/s41467-017-02292-8. URL: http://dx.doi.org/10.1038/s41467-017-02292-8.

Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E (2022). *ODGI: understanding pangenome graphs*. *Bioinformatics* **38**. Ed. by Peter Robinson, 3319–3326. https://doi.org/10.1093/bioinformatics/btac308. URL: http://dx.doi.org/10.1093/bioinformatics/btac308.

Hayward A, Gilbert C (2022). *Transposable elements*. *Current Biology* **32**, R904–R909.

Heumos S, Heuer ML, Hanssen F, Heumos L, Guarracino A, Heringer P, Ehmele P, Prins P, Garrison E, Nahnsen S (2024). *Cluster-efficient pangenome graph construction with nf-core/pangenome*. *Bioinformatics* **40**. Ed. by Can Alkan. https://doi.org/10.1093/bioinformatics/btae609. URL: http://dx.doi.org/10.1093/bioinformatics/btae609.

Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B (2020). *Genotyping structural variants in pangenome graphs using the vg toolkit*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-1941-7. URL: http://dx.doi.org/10.1186/s13059-020-1941-7.

Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, Abel HJ, Antonacci-Fulton LL, Asri M, Baid G, Baker CA, Belyaeva A, Billis K, Bourque G, Buonaiuto S, Carroll A, Chaisson MJP, Chang PC, Chang XH, Cheng H, et al. (2023). *Pangenome graph construction from genome alignments with Minigraph-Cactus*. *Nature Biotechnology* **42**, 663–673. https://doi.org/10.1038/s41587-023-01793-w. URL: http://dx.doi.org/10.1038/s41587-023-01793-w.

Holley G, Melsted P (2020). *Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-02135-8. URL: http://dx.doi.org/10.1186/s13059-020-02135-8.

Horsfield ST, Tonkin-Hill G, Croucher NJ, Lees JA (2023). *Accurate and fast graph-based pangenome annotation and clustering with ggCaller*. *Genome Research* **33**, 1622–1637. https://doi.org/10.1101/gr.277733.123. URL: http://dx.doi.org/10.1101/gr.277733.123.

Jonkheer EM, Workum DJM, Sheikhizadeh Anari S, Brankovics B, Haan JR, Berke L, Lee TAJ, Ridder D, Smit S (2022). *PanTools v3: functional annotation, classification and phylogenomics*. *Bioinformatics* **38**. Ed. by Tobias Marschall, 4403–4405. https://doi.org/10.1093/bioinformatics/btac506. URL: http://dx.doi.org/10.1093/bioinformatics/btac506.

Kawahara Y, Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, et al. (2013a). *Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data*. *Rice* **6**. https://doi.org/10.1186/1939-8433-6-4. URL: https://doi.org/10.1186/1939-8433-6-4.

Kawahara Y, Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, et al. (2013b). *Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data*. *Rice* **6**. https://doi.org/10.1186/1939-8433-6-4. URL: http://dx.doi.org/10.1186/1939-8433-6-4.

Keilwagen J, Hartung F, Grau J (2019). *GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data*. In: *Gene Prediction*. Springer New York, pp. 161–177. https://doi.org/10.1007/978-1-4939-9173-0_9. URL: http://dx.doi.org/10.1007/978-1-4939-9173-0_9.

Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ER, Pelaez J, Aguilar JM, Haji D, Matsunaga T, Armstrong EE, Zych M, Ogawa Y, Stamenković-Radak M, Jelić M, Veselinović MS, Tanasković M, et al. (2021). *Highly contiguous assemblies of 101 drosophilid genomes*. *eLife* **10**. https://doi.org/10.7554/elife.66405. URL: http://dx.doi.org/10.7554/eLife.66405.

Lemaitre C (2021). *Méthodes bioinformatiques pour l'étude des Variants de Structure avec des données de séquençages génomiques*. Accreditation to supervise research. Université Rennes 1. URL: https://theses.hal.science/tel-03497793.

Li H, Feng X, Chu C (2020). *The design and construction of reference pangenome graphs with minigraph*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-02168-z. URL: http://dx.doi.org/10.1186/s13059-020-02168-z.

Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, Mercier R (2024). *A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range*. *Nature Genetics* **56**, 982–991. https://doi.org/10.1038/s41588-024-01715-9. URL: http://dx.doi.org/10.1038/s41588-024-01715-9.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, Buonaiuto S, Chang XH, Cheng H, Chu J, Colonna V, Eizenga JM, Feng X, Fischer C, Fulton RS, Garg S, et al. (2023). *A draft human pangenome reference*. *Nature* **617**, 312–324. https://doi.org/10.1038/s41586-023-05896-x. URL: http://dx.doi.org/10.1038/s41586-023-05896-x.

Liu M, Zhang F, Lu H, Xue H, Dong X, Li Z, Xu J, Wang W, Wei C (2024). *PPanG: a precision pangenome browser enabling nucleotide-level analysis of genomic variations in individual genomes and their graph-based pangenome*. *BMC Genomics* **25**. https://doi.org/10.1186/s12864-024-10302-5. URL: http://dx.doi.org/10.1186/s12864-024-10302-5.

Marthe N, Sabot F (2025a). *Data for GrAnnoT*. Version V1. https://doi.org/10.23708/DO1RTF. URL: https://doi.org/10.23708/DO1RTF.

Marthe N, Sabot F (2025b). *Output for Grannot*. Version V1. https://doi.org/10.23708/RRSKRA. URL: https://doi.org/10.23708/RRSKRA.

Marthe N, Sabot F, Zytnicki M (2025). *GrAnnoT source code and scripts*. Version V1. https://doi.org/10.23708/TW3KYV. URL: https://doi.org/10.23708/TW3KYV.

Martiniano R, Garrison E, Jones ER, Manica A, Durbin R (2020). *Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-02160-7. URL: http://dx.doi.org/10.1186/s13059-020-02160-7.

Maurstad MF, Hoff SNK, Cerca J, Ravinet M, Bradbury I, Jakobsen KS, Præbel K, Jentoft S (2024). *Reference genome bias in light of species-specific chromosomal reorganization and translocations*. https://doi.org/10.1101/2024.06.28.599671. URL: http://dx.doi.org/10.1101/2024.06.28.599671.

Mergez A, Racoupeau M, Klopp C, Gaspin C, Legeai F (2024). *Pan1c (Pangenome at chromosome level)*. URL: https://hal.science/hal-05034842v1.

Miao Z, Yue JX (2025). *Interactive visualization and interpretation of pangenome graphs by linear reference-based coordinate projection and annotation integration*. *Genome Research*. https://doi.org/10.1101/gr.279461.124. URL: http://dx.doi.org/10.1101/gr.279461.124.

Miga KH, Wang T (2021). *The Need for a Human Pangenome Reference Sequence*. *Annual Review of Genomics and Human Genetics* **22**, 81–102. https://doi.org/10.1146/annurev-genom-120120-081921. URL: http://dx.doi.org/10.1146/annurev-genom-120120-081921.

Nevers Y, Defosset A, Lecompte O (2020). *Orthology: Promises and Challenges*. In: *Evolutionary Biology—A Transdisciplinary Approach*. Springer International Publishing, pp. 203–228. https://doi.org/10.1007/978-3-030-57246-4_9. URL: http://dx.doi.org/10.1007/978-3-030-57246-4_9.

Novak AM, Chung D, Hickey G, Djebali S, Yokoyama TT, Garrison E, Narzisi G, Paten B, Monlong J (2024). *Efficient indexing and querying of annotations in a pangenome graph*. https://doi.org/10.1101/2024.10.12.618009. URL: http://dx.doi.org/10.1101/2024.10.12.618009.

Otto TD, Dillon GP, Degrave WS, Berriman M (2011). *RATT: Rapid Annotation Transfer Tool*. *Nucleic Acids Research* **39**, e57–e57. https://doi.org/10.1093/nar/gkq1268. URL: http://dx.doi.org/10.1093/nar/gkq1268.

Outten J, Warren A (2021). *Methods and Developments in Graphical Pangenomics*. *Journal of the Indian Institute of Science* **101**, 485–498. https://doi.org/10.1007/s41745-021-00255-z. URL: http://dx.doi.org/10.1007/s41745-021-00255-z.

Pedersen TL, Nookaew I, Wayne Ussery D, Månsson M (2016). *PanViz: interactive visualization of the structure of functionally annotated pangenomes*. *Bioinformatics* **33**. Ed. by John Hancock, 1081–1082. https://doi.org/10.1093/bioinformatics/btw761. URL: http://dx.doi.org/10.1093/bioinformatics/btw761.

Pracana R, Priyam A, Levantis I, Nichols RA, Wurm Y (2017). *The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB*. *Molecular Ecology* **26**, 2864–2879. https://doi.org/10.1111/mec.14054. URL: http://dx.doi.org/10.1111/mec.14054.

Quinlan AR (2014). *BEDTools: The Swiss-Army Tool for Genome Feature Analysis*. *Current Protocols in Bioinformatics* **47**. https://doi.org/10.1002/0471250953.bi1112s47. URL: http://dx.doi.org/10.1002/0471250953.bi1112s47.

Rautiainen M, Marschall T (2020). *GraphAligner: rapid and versatile sequence-to-graph alignment*. *Genome Biology* **21**. https://doi.org/10.1186/s13059-020-02157-2. URL: http://dx.doi.org/10.1186/s13059-020-02157-2.

Rice ES, Alberdi A, Alfieri J, Athrey G, Balacco JR, Bardou P, Blackmon H, Charles M, Cheng HH, Fedrigo O, Fiddaman SR, Formenti G, Frantz LAF, Gilbert MTP, Hearn CJ, Jarvis ED, Klopp C, Marcos S, Mason AS, Velez-Irizarry D, et al. (2023). *A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants*. *BMC Biology* **21**. https://doi.org/10.1186/s12915-023-01758-0. URL: http://dx.doi.org/10.1186/s12915-023-01758-0.

Rogozin IB, Managadze D, Shabalina SA, Koonin EV (2014). *Gene Family Level Comparative Analysis of Gene Expression in Mammals Validates the Ortholog Conjecture*. *Genome Biology and Evolution* **6**, 754–762. https://doi.org/10.1093/gbe/evu051. URL: http://dx.doi.org/10.1093/gbe/evu051.

Romain S, Dubois S, Legeai F, Lemaitre C (2025). *Investigating the topological motifs of inversions in pangenome graphs*. https://doi.org/10.1101/2025.03.14.643331. URL: http://dx.doi.org/10.1101/2025.03.14.643331.

Rouli L, Merhej V, Fournier PE, Raoult D (2015). *The bacterial pangenome as a new tool for analysing pathogenic bacteria*. *New Microbes and New Infections* **7**, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005. URL: http://dx.doi.org/10.1016/j.nmni.2015.06.005.

Sanchez R, Morgado E, Grau R (2005). *Gene algebra from a genetic code algebraic structure*. *Journal of Mathematical Biology* **51**, 431–457. https://doi.org/10.1007/s00285-005-0332-8. URL: http://dx.doi.org/10.1007/s00285-005-0332-8.

Secomandi S, Gallo GR, Rossi R, Rodríguez Fernandes C, Jarvis ED, Bonisoli-Alquati A, Gianfranceschi L, Formenti G (2025). *Pangenome graphs and their applications in biodiversity genomics*. *Nature Genetics* **57**, 13–26. https://doi.org/10.1038/s41588-024-02029-6. URL: http://dx.doi.org/10.1038/s41588-024-02029-6.

Shi J, Tian Z, Lai J, Huang X (2023). *Plant pan-genomics and its applications*. *Molecular Plant* **16**, 168–186. https://doi.org/10.1016/j.molp.2022.12.009. URL: http://dx.doi.org/10.1016/j.molp.2022.12.009.

Shumate A, Salzberg SL (2021). *Liftoff: accurate mapping of gene annotations*. *Bioinformatics* **37**. Ed. by Alfonso Valencia, 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016. URL: http://dx.doi.org/10.1093/bioinformatics/btaa1016.

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang PC, Carroll A, Gupta N, Gabriel S, Blackwell TW, Ratan A, Taylor KD, Rich SS, Rotter JI, Haussler D, Garrison E, Paten B (2021). *Pangenomics enables genotyping of known structural variants in 5202 diverse genomes*. *Science* **374**. https://doi.org/10.1126/science.abg8871. URL: http://dx.doi.org/10.1126/science.abg8871.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, et al. (2005). *Genome analysis*

of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* **102**, 13950–13955. https://doi.org/10.1073/pnas.0506758102. URL: http://dx.doi.org/10.1073/pnas.0506758102.

Tranchant-Dubreuil C, Rouard M, Sabot F (2019). *Plant Pangenome: Impacts on Phenotypes and Evolution*. https://doi.org/10.1002/9781119312994.apr0664. URL: http://dx.doi.org/10.1002/9781119312994.apr0664.

Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, Dong Q, Ye K (2021). *High-Quality Arabidopsis Thaliana Genome Assembly with Nanopore and HiFi Long Reads*. *Genomics, Proteomics & Bioinformatics* **20**, 4–13. https://doi.org/10.1016/j.gpb.2021.08.003. URL: http://dx.doi.org/10.1016/j.gpb.2021.08.003.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, et al. (2018). *Genomic variation in 3, 010 diverse accessions of Asian cultivated rice*. *Nature* **557**, 43–49. https://doi.org/10.1038/s41586-018-0063-9. URL: http://dx.doi.org/10.1038/s41586-018-0063-9.

Wei X, Chen M, Zhang Q, Gong J, Liu J, Yong K, Wang Q, Fan J, Chen S, Hua H, Luo Z, Zhao X, Wang X, Li W, Cong J, Yu X, Wang Z, Huang R, Chen J, Zhou X, et al. (2024). *Genomic investigation of 18,421 lines reveals the genetic architecture of rice*. *Science* **385**, eadm8762. https://doi.org/10.1126/science.adm8762. eprint: https://www.science.org/doi/pdf/10.1126/science.adm8762. URL: https://www.science.org/doi/abs/10.1126/science.adm8762.

Wick RR, Schultz MB, Zobel J, Holt KE (2015). *Bandage: interactive visualization of de novo genome assemblies*. *Bioinformatics* **31**, 3350–3352. https://doi.org/10.1093/bioinformatics/btv383. URL: http://dx.doi.org/10.1093/bioinformatics/btv383.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. (2007). *A unified classification system for eukaryotic transposable elements*. *Nature reviews genetics* **8**, 973–982.

Yang C, Zhou Y, Song Y, Wu D, Zeng Y, Nie L, Liu P, Zhang S, Chen G, Xu J, Zhou H, Zhou L, Qian X, Liu C, Tan S, Zhou C, Dai W, Xu M, Qi Y, Wang X, et al. (2023). *The complete and fully-phased diploid genome of a male Han Chinese*. *Cell Research* **33**, 745–761. https://doi.org/10.1038/s41422-023-00849-5. URL: http://dx.doi.org/10.1038/s41422-023-00849-5.

Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L (2013). *CrossMap: a versatile tool for coordinate conversion between genome assemblies*. *Bioinformatics* **30**, 1006–1007. https://doi.org/10.1093/bioinformatics/btt730. URL: http://dx.doi.org/10.1093/bioinformatics/btt730.

Zheng Q, Li L, Yin X, Che Y, Zhang T (2023). *Is ICE hot? A genomic comparative study reveals integrative and conjugative elements as "hot" vectors for the dissemination of antibiotic resistance genes*. *mSystems* **8**. Ed. by Li Cui. https://doi.org/10.1128/msystems.00178-23. URL: http://dx.doi.org/10.1128/msystems.00178-23.

Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K, Zhang J, Lyu H, Lin T, Gao Q, Saha S, Mueller L, Fei Z, Städler T, Xu S, Zhang Z, et al. (2022). *Graph pangenome captures missing heritability and empowers tomato breeding*. *Nature* **606**, 527–534. https://

doi.org/10.1038/s41586-022-04808-9. URL: http://dx.doi.org/10.1038/s41586-022-04808-9.

Zhou Y, Chebotarov D, Kudrna D, Llaca V, Lee S, Rajasekar S, Mohammed N, Al-Bader N, Sobel-Sorenson C, Parakkal P, Arbelaez LJ, Franco N, Alexandrov N, Hamilton NRS, Leung H, Mauleon R, Lorieux M, Zuccolo A, McNally K, Zhang J, et al. (2020). *A platinum standard pan-genome resource that represents the population structure of Asian rice*. *Scientific Data* **7**. https://doi.org/10.1038/s41597-020-0438-2. URL: http://dx.doi.org/10.1038/s41597-020-0438-2.