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, offer thus a more comprehensive representation of intraspecific diversity. However, pangenomes in form of graph often lack annotation information, which limits their utility for forward analyses. We introduce here GrAnnoT, a tool designed for efficient and reliable annotation transfer using such graphs, by projecting existing annotations from a source genome to the graph and subsequently to other embedded genomes. GrAnnoT was benchmarked against state-of-the-art tools on pangenome graphs and linear genomes from rice, human, and *E. coli*. The results demonstrate that GrAnnoT is consensual, conservative, and fast, outperforming alignment-based methods in accuracy or speed or both. It provides informative outputs, such as presence-absence matrices for genes, and alignments of transferred features between source and target genomes, aiding 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 have allowed access to a massive and increasing number of genome sequences per species. This has highlighted the fact that a single individual is not enough to represent the whole diversity of a species. Indeed, while currently prevalent, the use of a single reference genome has been shown to bias some analysis (Chen et al., 2021; Martiniano et al., 2020; Maurstad et al., 2024). This has led to the development 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; Tranchant-Dubreuil et al., 2019). 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. This information can be organized in different ways depending on the type of study involved. The pangenome graph structure has recently emerged as a solution to store and model these pangenomes. This structure has the advantage of containing the whole sequence information (genic and inter-genic regions) and of encoding the relationships between the genomes (which regions are identical, where are the variations) in a compact and comprehensive way. These pangenome graphs can be stored in variation graphs in GFA format, the standardized text format used by many tools. It is human-readable and represents the multiple genome alignment. Multiple forms of graphs that can be stored in the GFA format; the graphs we consider in this paper are bi-directed and acyclic genome graphs (see 2.1 for details on the graphs used).

Pangenome graphs have already proven their usefulness to better understand the structure and dynamics of genomes, for structural variations detection, or for genotyping for instance (Rice et al., 2023; Zhou et al., 2022). However, their practical use still has limitations, as the tools to manipulate them are often still in development. In particular, these graphs usually do not contain any annotation information. Similarly to a genome without annotation, a pangenome graph without embedded biological information is less useful, and the variations present in the graph are harder to interpret. Therefore any variation found in the graph has to be reported back to a single linear annotated genome to see if it overlaps with a region of interest. An annotated graph would then allow to more easily study the structure and evolution of a species pangenome.

Furthermore, despite the recent appearance of AI-based tools (Holst et al., 2023), *de novo* genome annotation is a long and complicated process. However, good quality and manually curated annotations already exist for many species, usually for a single linear individual genome. These existing annotations can be transferred to other non-annotated genomes to add meaningful biological information. This transfer operation is much faster than a *de novo* annotation, requiring far less computation and resources. Annotation transfer between linear genomes is usually performed using a blast-like approach (e.g. Liftoff (Shumate and Salzberg, 2021)), by mapping/aligning the sequence of the annotated elements on the target genome. However, this approach is not currently adapted to pangenome graphs. Indeed, alignment on a graph is more complex than alignment on a linear sequence, and existing graph alignment tools are quite recent and still under maturation. However, the graph itself already represents a global alignment between the annotated genome and the other embedded genomes. Thus, this alignment can be used to project the coordinates of the annotated elements from the linear genome to the graph, and any information concerning a linear genome position can be transferred to the graph. Once the annotation is transferred onto the graph, it can be transferred back to any other genome

embedded in the graph. Therefore, any information added to the graph will benefit all the embedded genomes.

We developed GrAnnoT (Graph Annotation Transfer) to perform these operations. This command-line tool will allow to gather more information in pangenome graphs and better harness the many graphs that have been already produced (Rice et al., 2023; Shi et al., 2023; Zhou et al., 2022). It includes functions to transfer annotations and to study the variations present in the graph in the annotated elements. We applied it to different pangenome graphs to ensure it works with various species, and compared it to existing methods for annotation transfer. Compared to these methods, GrAnnoT is consensual, conservative, and fast. It also offers informative outputs to allow the user to review the transfers performed, such as a presence-absence matrix, alignments between the genes in different genomes or a list of the variations found in annotated regions.

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 uses the package tqdm and the external program bedtools (Quinlan, 2014) (that must be accessible in the user or in the global path). The code is available on the IRD forge (`https://forge.ird.fr/diade/dynadiv/grannot`) under the GNU GPLv3 licence.

1.1. Code overview

GrAnnoT performs annotation transfer from an annotated genome (the source genome) to a pangenome 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 a pangenome graph in GFA 1.1 format (which includes the source and target genomes), and the annotation of the source genome in GFF3 format. For the sake of clarity, in the present paper the graph considered has a unique path for each haplotype, but the method works for assemblies with several chromosomes/contigs.

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

Once the annotation has been loaded, GrAnnoT outputs the graph annotation in GAF format, which describes the paths of the annotated features in the graph. It can then output the annotation in GFF format of a chosen set of target linear genomes included in the graph. These transfers can be filtered through sequence identity and coverage scores. 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 variations recorded in the alignments. 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 source and target 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. Then, the BED file representing the source genome is 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 GAF format.

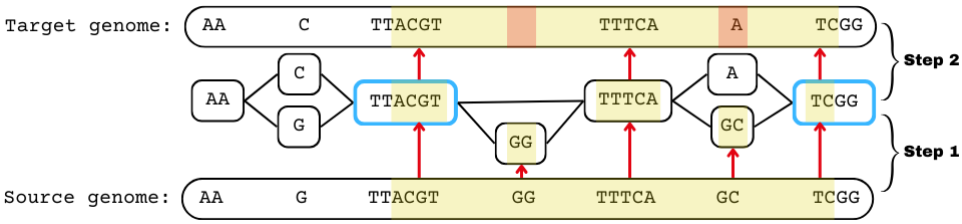


Figure 1 – GrAnnoT overview. Step 1: the position of the feature is projected from the source genome to the graph using the positions of the nodes on the source genome. Step 2: the position of the feature is projected from the graph to the target genome. The first and last nodes from the feature that 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.

In order to transfer an annotation to a target genome, the sub-path of the genome corresponding to the feature is extracted (Figure 1, step 2). For that, all the nodes from the original feature path are looked for in the target genome path. These nodes are then grouped into copies of the feature, and for each copy the first and the last nodes are considered as the ends of the feature's copy in the target genome. All the nodes between them 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 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. These parameters are estimated by computing the cumulated length of the shared and different nodes between the paths of the features in the two genomes. The output is finally printed out in the GFF 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 embedded target genome. For that, it can output the variations details in a text format that describes all the variations 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 a 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 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%).

GrAnnoT was also tested on a graph of the human chromosome 1 with 92 haplotypes (from Liao et al., 2023) and an *E. coli* 12 genomes (Jangir et al., 2022) graph 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.

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 relies primarily on the alignment of the nucleic sequences of the annotated features from the source genome upon the target one. 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 below).

Liftoff approach to transfer annotations can be adapted to a pangenome graph 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.

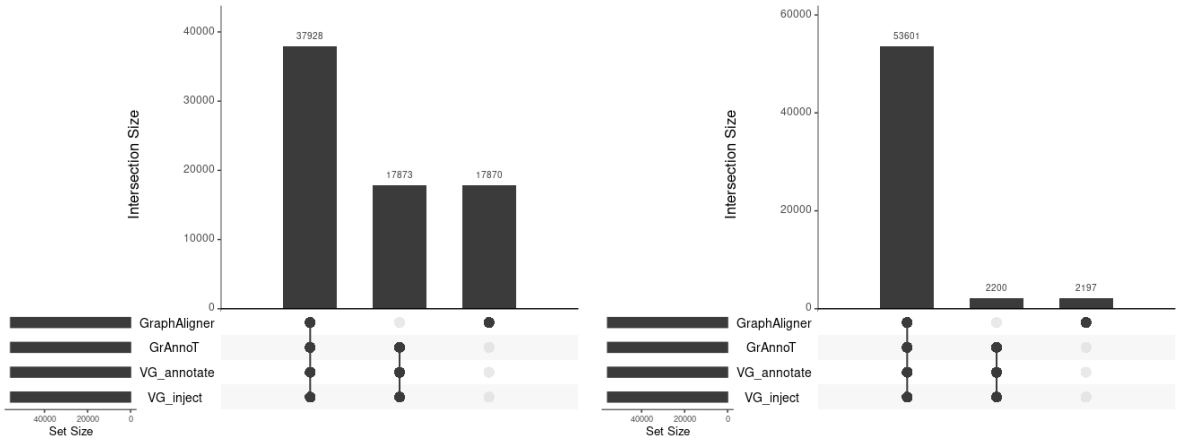
VG and ODGI are state-of-the-art tools for pangenome graph manipulation (Garrison et al., 2018; Guarracino et al., 2022). They do not have options specifically designed to transfer annotations between genomes of the graph, but they do have options to project coordinates between the graph and its embedded genomes, specifically *odgi position* and *vg inject/surject*. These options can be adapted to transfer annotations between genomes, and from a genome to the graph. VG also has an option for annotation transfer on the graph (*vg annotate*). 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 transfer was 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.



(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).

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

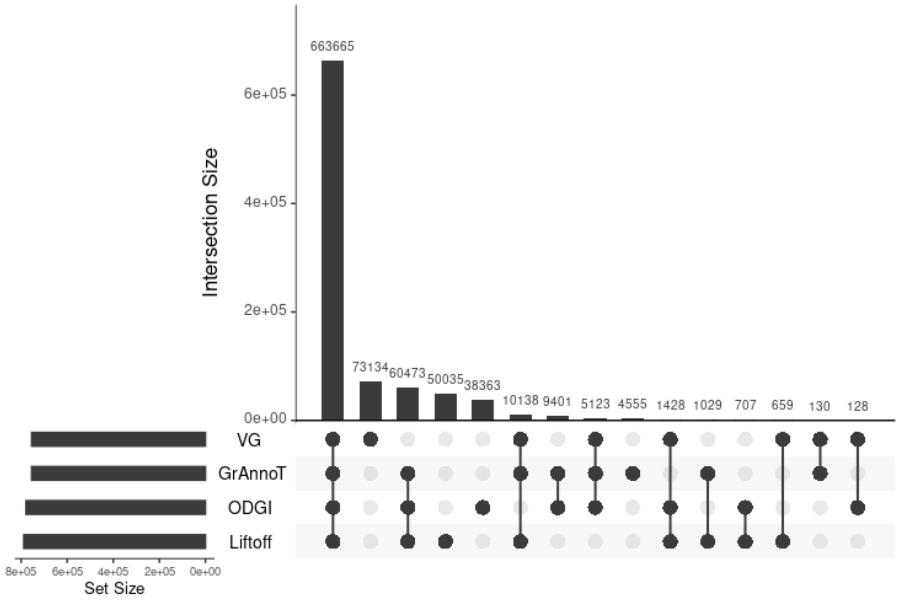
Genome to graph transfer:

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 annotation transfer.

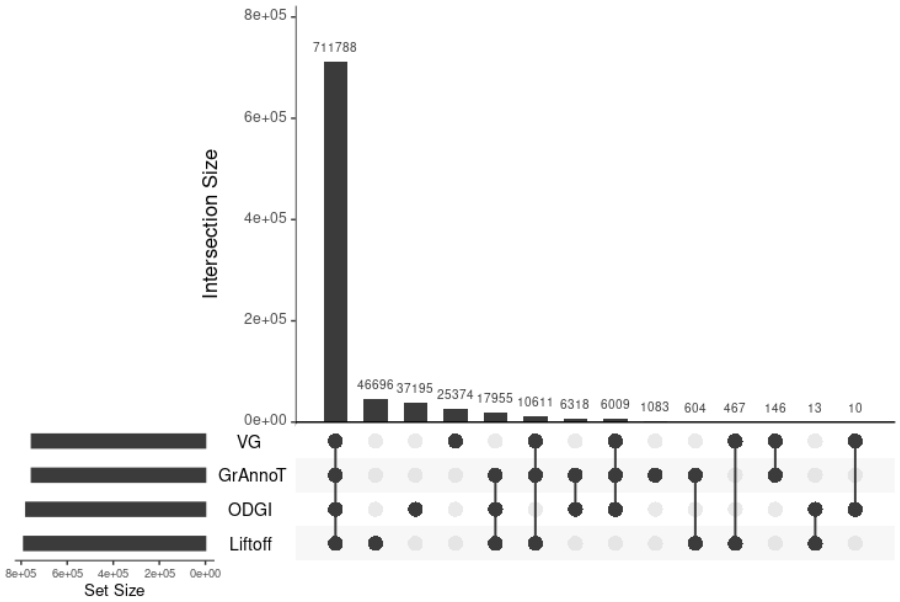
Genome to genome transfer:

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.

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). The 10bp difference tolerance revealed Liftoff and ODGI as the most divergent tools (with 46,696 and 37,195 specific transfers, respectively; Figure 3b).



(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).

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 graphs are currently built chromosome-per-chromosome to reduce complexity, and therefore cannot represent such events. 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.

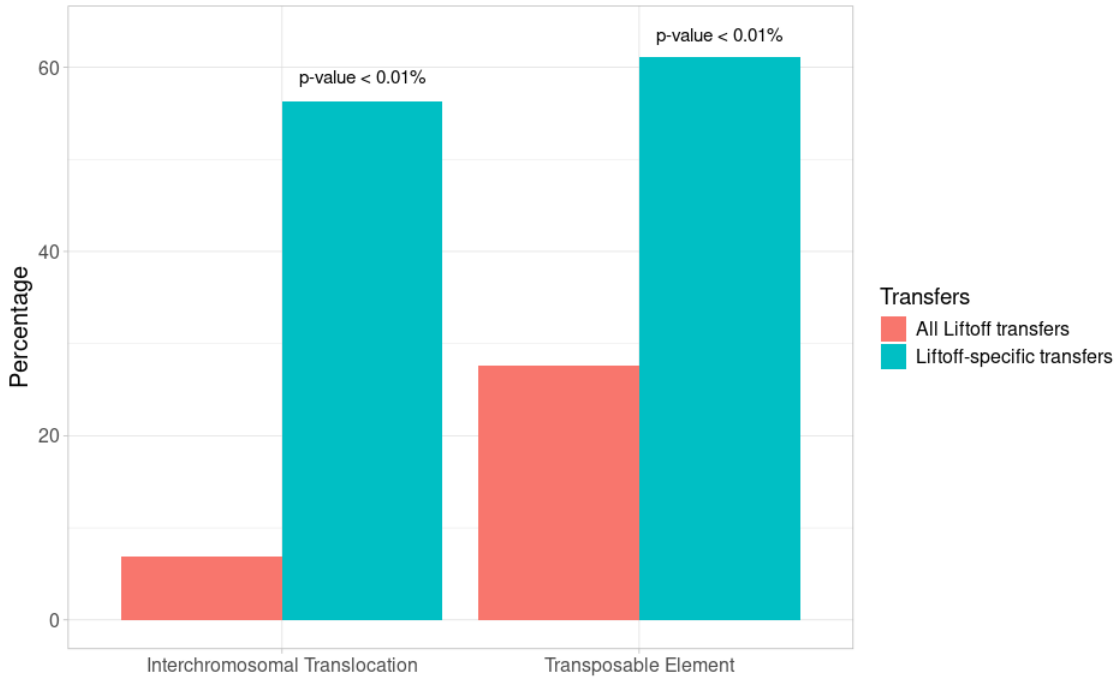


Figure 4 – Transposable element and inter-chromosomal translocation percentages in all Liftoff transfer *vs* Liftoff-specific transfers. The Liftoff-specific transfers are enriched in translocations and in transposable elements compared to all the other Liftoff transfers. Detailed data and *p*-value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

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 why so many elements are on different chromosomes 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 transposable elements (*p*-value < 0.01%; Figure 5), as their ability to move in the genome makes them often not syntenic: encoding the relationships between such elements in the graph with the current pangenome graph tools still seems complex. Since these relationships are not correctly encoded by the graph, the TE annotation transfer cannot be reliably performed by tools such as GrAnnoT, which only uses the structure of the graph.

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 length 0 nucleotide, and ~30% (11,320) on an interval of length 1 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 with Liftoff and GrAnnoT allowed to compare how conservative these tools are. The first transfer was performed from Nipponbare to Azucena with the two tools. Then, the resulting Azucena GFF was used as source to perform the second

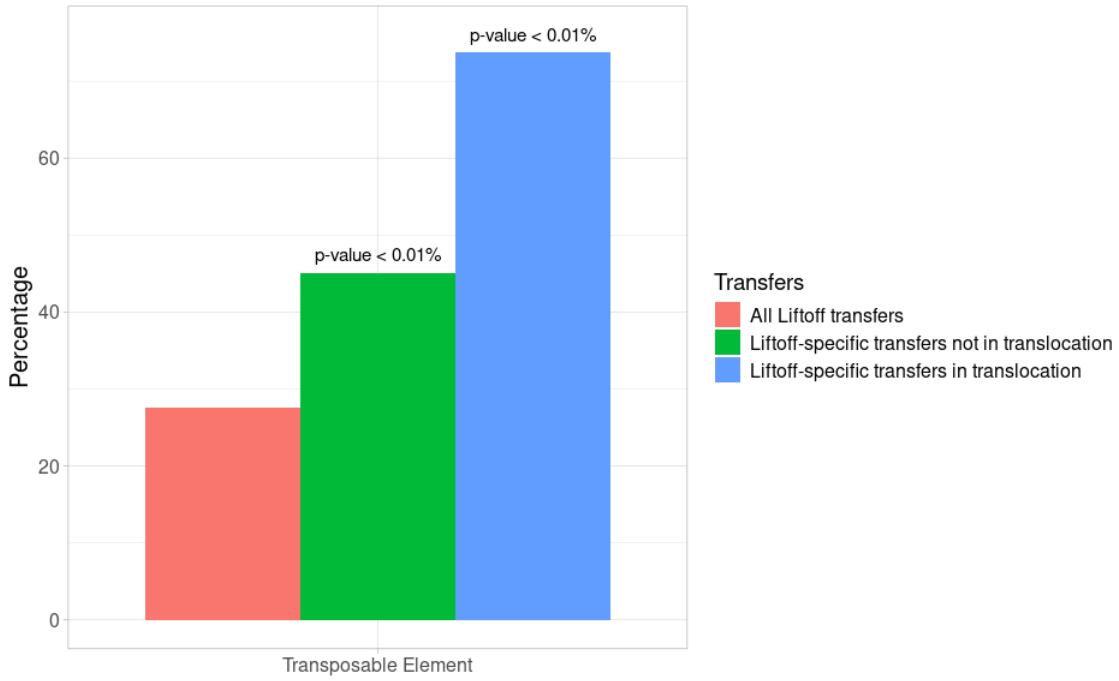


Figure 5 – Transposable element percentages in all Liftoff transfer *vs* Liftoff-specific transfers. Compared to the other Liftoff transfers, Liftoff-specific transfers are enriched in transposable elements, whether or not they are in a translocation. Detailed data and *p*-value calculation are available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

transfer, from Azucena back to Nipponbare. The resulting GFF for Nipponbare was compared to its first original annotation to measure the loss of information during these transfers.

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 loose any more annotation on the way back to the original sequence, Liftoff lost an additional 257 of them. In addition, when comparing the positions of the features before and after the two transfers, 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. In addition, after manual verification, it appeared that the features misplaced by GrAnnoT in the second transfer are features where an extremity was shortened during the first transfer due to a deletion. Thus, the feature transferred during the second transfer was incomplete regarding the true annotation, but the transfer itself occurred correctly.

Finally, some features found by Liftoff are placed on a different chromosome than the original, 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 graph conserves the synteny, GrAnnoT is more likely to transfer annotations between orthologous copies than between paralogous copies.

Impact of the reference genome for graph construction: The 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

	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 (Nipponbare to Azucena or Azucena to 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.

different graphs were compared. The two graphs have the same genomes embedded, but a different reference genome to initiate the graph. The reference genomes used for the two graphs are the annotated genome IRGSP-1.0 (Nipponbare), and Os127652RS1 (Natel Boro) (Zhou et al., 2020). Annotation transfer from genome to genome was performed with these two graphs, and the positions of the common transferred features were compared.

Among the 48,322 features transferred on 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 location (table 2).

The amount of features not transferred by both graphs is not negligible, but it can be explained by the choice of the reference genome for the graph construction, Natel Boro. Indeed, among the 11 genomes in the graph that are not involved in the transfer (not Nipponbare or Azucena), 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

	Rice		Human		E.coli	
Total features to transfer	55,986		282,668		9,467	
Features transferred by GrAnnoT	48,025	85.78%	276,681	97.88%	7,230	76.37%
GrAnnoT-specific transfers	72	0.13%	4,647	1.64%	81	0.86%
Features transferred by Liftoff	54,363	97.10%	275,249	97.38%	7,940	83.87%
Liftoff-specific transfers	6,410	11.45%	3,264	1.12%	790	8.34%
Features transferred by both tools	47,951	85.65%	258,092	91.31%	7,146	75.48%
Same position	46,431	82.93%	256,927	90.89%	6,789	71.71%
Different position	1,520	2.71%	1,165	0.41%	357	3.77%
1-10bp difference	795	1.42%	623	0.22%	216	2.28%
11-100bp difference	284	0.51%	150	0.05%	60	0.63%
101-1000bp difference	155	0.28%	46	0.02%	18	0.19%
>1000bp difference	164	0.29%	346	0.12%	63	0.67%
Different chromosome	122	0.22%	0	0%	0	0%
Runtime Liftoff	00:23:45		00:10:27		00:00:08	
Runtime GrAnnoT	00:08:11		00:14:47		00:00:20	

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.

around Nipponbare displays better performance for annotation transfer from 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. In the case of annotation transfer, using the annotated source genome is a good solution.

Comparison with other species: GrAnnoT was compared to Liftoff using two other datasets: a pangenome graph of the human chromosome 1 (Liao et al., 2023) and an *E. coli* pangenome graph (Jangir et al., 2022). Both of these graphs were made with Minigraph-Cactus. For the rice graph, the transfer was again made from Nipponbare to Azucena; for the human graph, the transfer was made from CHM13 to GrCH38; for the *E.coli* graph, the transfer was made from K_12_MG1655_09949b0 to O127_H6_E2348_69_193637c. These comparisons checked if the positions of the features transferred by both approach are consistent, to see 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 three species, most of the features are transferred by both tools (~85.7% for rice, ~91.3% for human and ~75.5% for *E. coli*) (table 3). Additionally, a large part of these features are placed at the exact same position by Liftoff and GrAnnoT (~96.8% for rice, ~99.6% for human and ~95% 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 tool 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 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.

	GrAnnoT	Liftoff	VG	ODGI
Run time	00:08:11	00:23:45	07:26:54	70:23:41

Table 4 – Run time comparison for genome to genome transfer in rice.

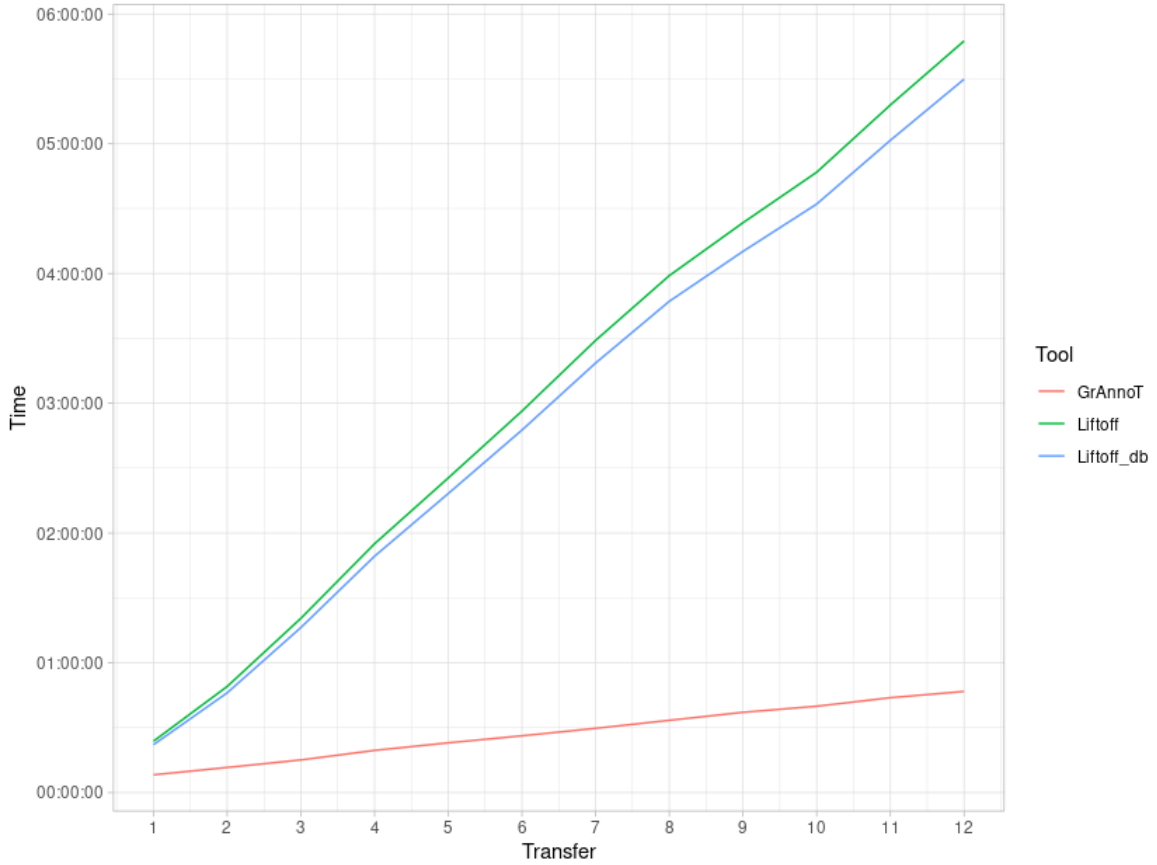


Figure 6 – Genome to genome transfer comparison. GrAnnoT and Liftoff run time for 1-12 transfers were measured using the command */usr/bin/time*. Liftoff was run both in GFF and DB mode. Detailed time points are available in supplementary data in table S10.

2.3. Time and memory usage results

The execution time for the transfer from genome to genome with the different tools was measured using the command */usr/bin/time* (table 4). The results show that GrAnnoT has the best run time, and that ODGI and VG are substantially slower that GrAnnoT and Liftoff.

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

Several transfers were performed with both tools to compare the run times, because GrAnnoT is advantaged when multiple transfers are requested, as GrAnnoT starts by pre-processing the graph and loading the graph annotation. These steps only need to be done once, no matter how many annotation transfers to target genomes are performed.

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 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 doesn't take into account the time needed to build the graph. When adding the graph construction time (~4h55mn on our infrastructure) to the GrAnnoT 12 transfers time, we still get a duration (~5h41mn) equivalent 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.

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 GFF 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 variations (insertion or deletion) in the different feature types that correspond to different parts of the genes. These variations 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, with the gene being the longest element (since it contains all the others), and the CDS being shorter than the exon summed size, for instance. This induces 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 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.

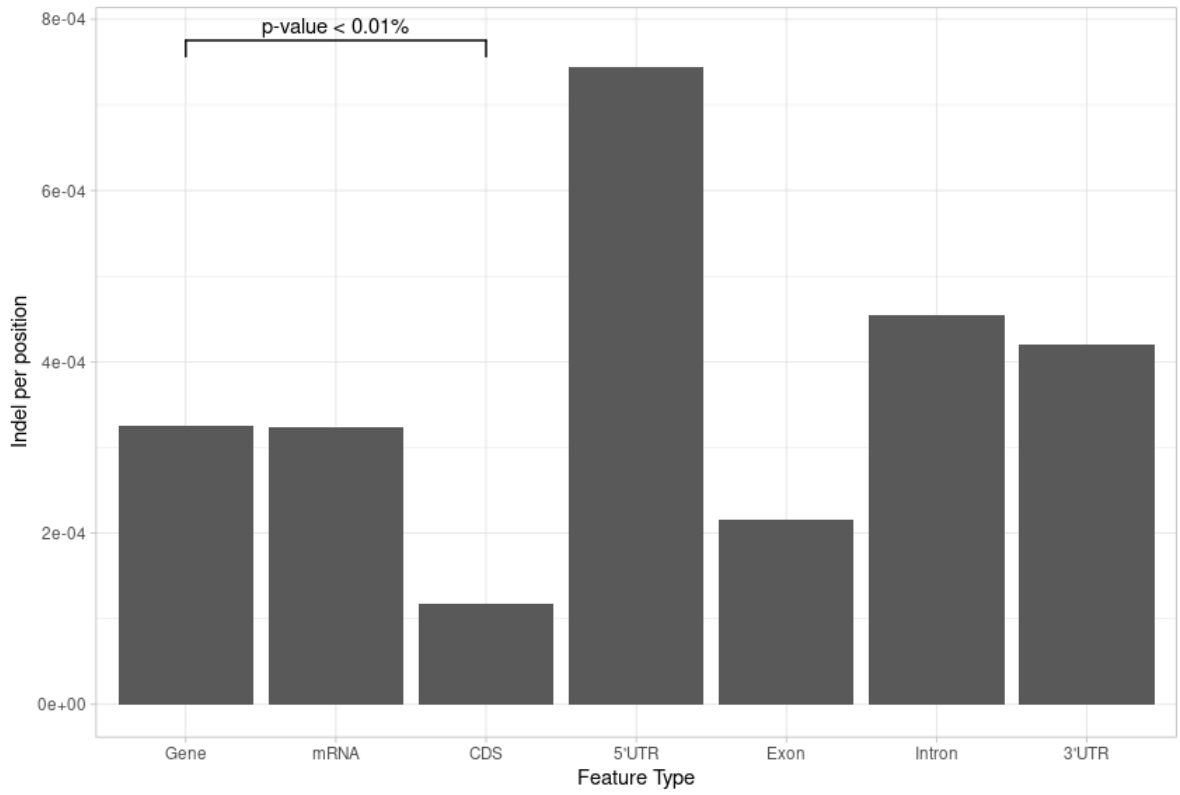


Figure 7 – Indel distribution. Each bar represents the number of indels (insertion or deletion) per position in the corresponding feature type. As expected, the CDS are the most conserved and thus have the least indels, and the non-coding regions (UTR and introns) are the least conserved and have the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

3.2. Frameshit 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

The substitutions are smaller variations than the 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 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.

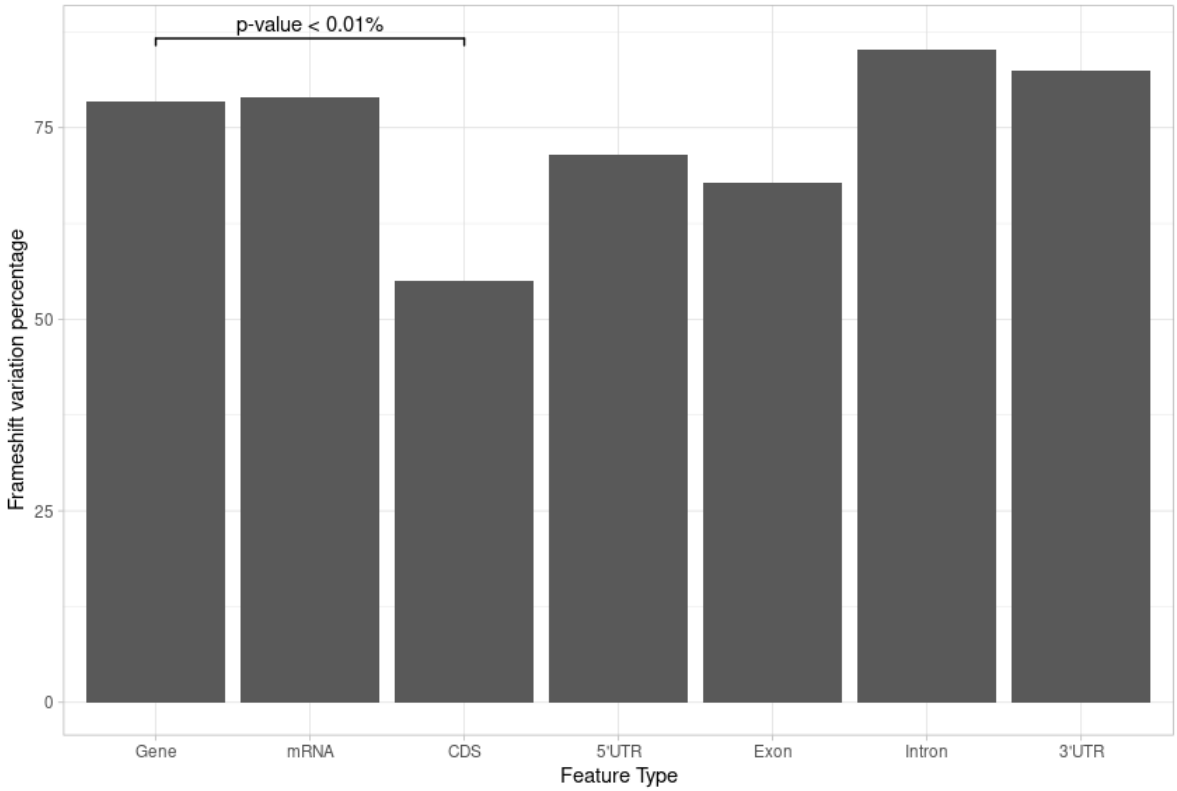


Figure 8 – Frameshift indel distribution. Each bar represent the percentage of frameshift variation (length not multiple of 3) among all the indels in each feature type. As expected, the CDS have the least frameshift variation, since these variations impact significantly the protein coded. Detailed data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

To find the codon positions we had to take into account the splicing of the mRNA. 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 CDS fragments. Thus adjustment was only done for the CDS elements in the annotation, since they were the only element of interest. This explains why the exons do not follow the CDS tendency in Figure 9, contrary to Figures 7 and 8.

3.4. Pangene set analysis

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

We found ~58% of core genes and ~33% of dispensable genes (table 5) in our pangenome 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).

Conclusion

In the present study, we presented the first tool able to efficiently transfer annotation on a pangenome graph from one of its embedded genomes and reverse, GrAnnoT. It relies on the

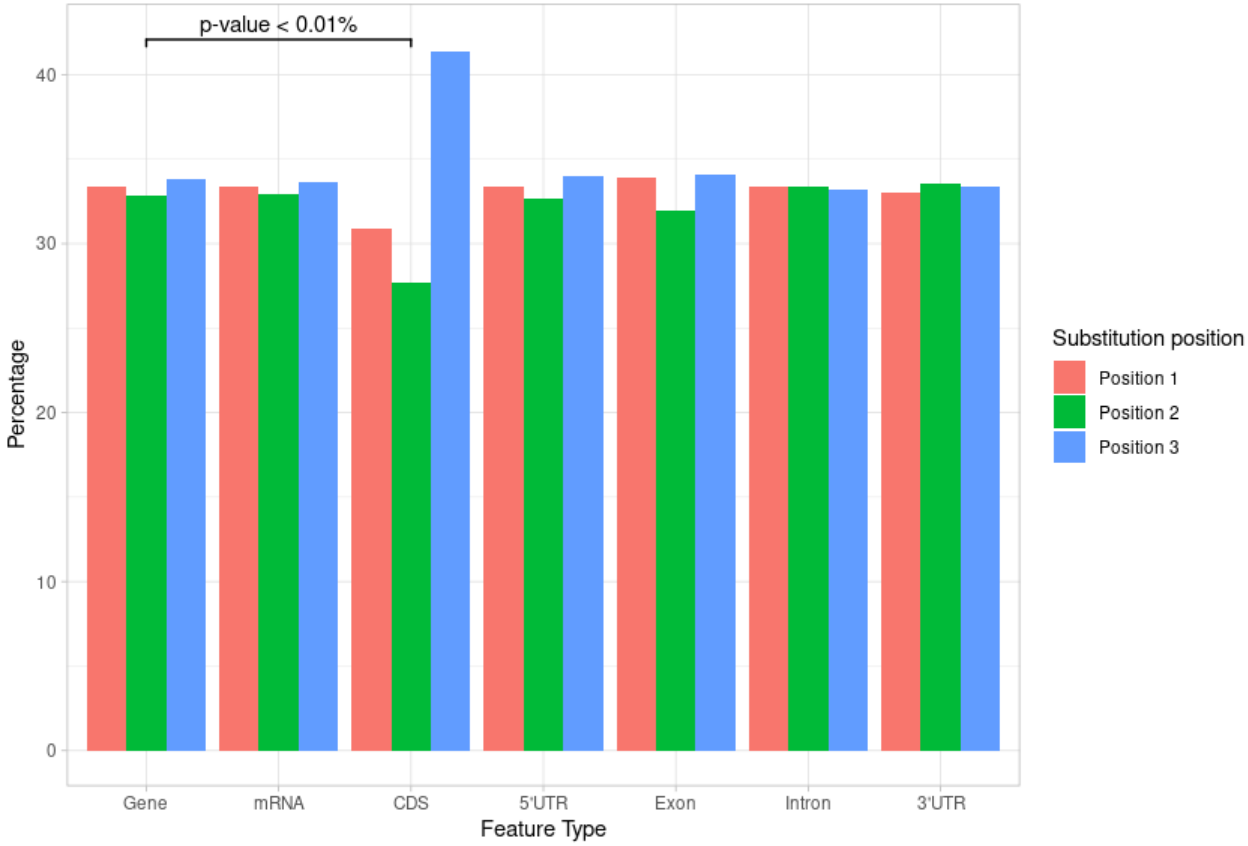


Figure 9 – Substitution positions. For each feature type, the percentage of substitutions that are on each of the three codon positions is displayed. In the CDS, the third position has more substitutions than the two other positions. For the other feature types, we don't see that the positions multiple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

	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.

already performed alignment that created the graph. We benchmarked GrAnnoT on rice and human pangenomes, and showed that it is fast, reliable and efficient, 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 pangenome 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. In the near future, we plan to optimize the transfer time through parallelization, and to implement the inference of the impact of mRNA and CDS mutations on the resulting proteins.

Acknowledgements

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

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

Nina Marthe is supported by a PhD funding from the Agence Nationale de la Recherche as part of the France2030 program under the reference "ANR-22-PEAE-4".

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 recommender. The LLM agent LeChat was used to prepare the abstract section once the manuscript 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). Script 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 from Jangir et al., 2022)
  - 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)

	Different chromosome	Same chromosome	*P*-value
Liftoff-specific transfers	3604	2806	< 0.01%
Other Liftoff transfers	122	47831	

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.

```
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 – Extract of the alignment of gene LOC_Os01g01050 and its transfers in Azucena by different tools. VG 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 Azucena in the middle of the gene.

	Transposable elements	Other features	*P*-value
Liftoff-specific transfers	3919	2491	< 0.01%
Other Liftoff transfers	11053	36900	

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	1263	1543	< 0.01%
All Liftoff transfers	13709	46410	

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	2656	948	< 0.01%
All Liftoff transfers	12316	38443	

Table S9 – Transposable elements rates 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 time comparison for 1-12 transfers

	Positions without indel	Positions with indel	*P*-value
Gene	141980198	46148	< 0.01%
CDS	74201787	8762	

Table S11 – Indel rates in genes and CDS. Annotations were transferred between cv Nipponbare and 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	9959	36189	< 0.01%
CDS	3950	4812	

Table S12 – Frameshift indel rates in genes and CDS. Annotations were transferred between cv Nipponbare and 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	171763	87638	< 0.01%
CDS	73005	51562	

Table S13 – Substitution positions in nucleotide triplets in genes and CDS. Annotations were transferred between cv Nipponbare and 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.

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.

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.

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.

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.

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.

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.

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.

Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöh O, Usadel B, Schwacke R, Bolger M, Weber AP, Denton AK (2023). *Helixer–de novoPrediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model*. https://doi.org/10.1101/2023.02.06.527280. URL: http://dx.doi.org/10.1101/2023.02.06.527280.

Jangir PK, Yang Q, Shaw LP, Caballero JD, Ogunlana L, Wheatley R, Walsh T, MacLean RC (2022). *Pre-existing chromosomal polymorphisms in pathogenic E. coli potentiate the evolution of resistance to a last-resort antibiotic*. *eLife* **11**. https://doi.org/10.7554/elife.78834. URL: http://dx.doi.org/10.7554/eLife.78834.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.