RESEARCH

Pangenomic QTL analysis

Lily Sakellaridi*

Correspondence:
lygeri.sakellaridi@wur.nl
Department of Bioinformatics, WUR, Droevendaalsesteeg 1, Wageningen, Netherlands
Full list of author information is available at the end of the article
*Equal contributor

Abstract

Many agriculturally important traits are quantitative: examples include grain yield, disease resistance, and flowering time. Quantitative traits are affected by multiple genes, lying in a large genomic region called a quantitative trait locus (QTL). Typically, a quantitative trait is affected by multiple such regions (ie multiple QTL). Due to the large number of genes in a QTL, pinpointing the causal genes for a trait constitutes an important challenge. In an attempt to pinpoint such causal genes, the genomic content present in quantitative trait loci (QTL) may be analyzed manually or through the use of automated gene prioritization methods [1][2][3]. However, both manual and automated analyses are restricted in that they are typically performed in the context of a single reference genome. In the midst of an era where genomic data are so vastly and readily available, reference-centric approaches greatly limit the amount of available genomic knowledge that is actually utilized. As a result, genomics has been shifting from reference-centric to pangenomic approaches.

In this project, we have developed a pangenomic approach for QTL analysis. In this context, QTL analysis refers to the study of the structural and sequence-level variation within a QTL region across different genomes, with the ultimate aim of identifying causal genes for a specific trait. In this project we made several critical steps towards that goal. We developed a fast, accurate and robust workflow to integrate QTL data into pangenomes, extract homologous regions of the QTLs, and analyze the genomic variation between the regions. Our method is applicable to all diploid plants. Computational testing was performed using a mix of artificial and real QTL data from *A.thaliana* and *O.sativa*. From a biological perspective, results from our method were validated against the previously studied use case of strigolactone biosynthesis (SLB) in rice. Moreover, the output of our workflow was used to explore several types of structural variation in rice QTL. Our results show that structural variation is very prominent even within a small, closely related set of genomes; gene presence-absence variation (PAV) was found to the most common type of those examined.

Our workflow immediately provides an overview of the structural variation present in regions of interest across a set of genomes, enabling researchers to manually examine such regions for variants of interest and candidate causal genes. However, its greatest strength lies in its potential combination with different data types and analyses. Gene lists from the detected regions can be functionally annotated, combined with RNA-seq or phenotypic data, and/or used as inputs for gene prioritization methods. Thus, while our method stands on its own as an immediately useful application, it also serves as a strong yet flexible foundation for the development of something even greater.

Keywords: pangenomics; QTL; PanTools

Introduction

Plant breeding is used to facilitate the improvement of crops in terms of traits that affect food production[4]. Examples of such agriculturally important traits include grain yield, disease resistance and flowering time. Many of these traits are complex quantitative traits: they are affected by multiple genes, lying in a large genomic region called a quantitative trait locus (QTL). Typically, a quantitative trait is affected by multiple such regions (ie multiple QTL). Due to the large number of genes in a QTL, pinpointing the causal genes for a trait constitutes an important challenge. Prioritization methods have been developed to address that challenge [1][3][2].

However, such methods are limited by the fact that they are usually applied and interpreted in the context of a single reference genome. Phenotypic variation is naturally tied to genomic variation. Therefore, to pinpoint the genes that result in different phenotypes for a trait, one must examine the subset of genes that vary in some way between individual organisms. It is possible to study certain types of variation in the context of a single reference genome: for example, single nucleotide polymorphisms (SNPs) can be studied by use of resequencing data, or comparing multiple accessions to the reference. However, structural variation is much harder to detect in that way. In an era where genomic data are generated at an unprecedented scale, it becomes almost paradoxical to study traits in the context of a single reference genome, since that misses a tremendous amount of known genomic variation present within a species.

In order to capitalize on this genomic variation, we introduce a pangenomic approach to the study of quantitative traits. Our approach is in line with a recent paradigm shift in comparative genomics from the reference-centric approach to the pangenomic approach[5]. This shift has been due to the aforementioned availability of a vast amount of high-throughput sequencing data, which naturally generates the need of more efficient methods to represent, store and analyze the data. Here, we use the term “pangenome” to refer to a graph-based representation that stores, and enables the comparative study of, multiple genomes and their annotations. This representation allows for capturing the entire genomic variation of the included genomes, including structural variation (SV), single nucleotide polymorphisms (SNPs) and small insertions - deletions (indels).

In this project, we used PanTools [6] to construct pangenomes of two plant species: *Arabidopsis thaliana* and *Oryza sativa* (rice). We developed several new functionalities for PanTools that enable the study of complex quantitative traits in a pangenomic context. The new functionalities can be divided in three categories: 1) integration and summary of QTL data; 2) detection of homologous QTL regions; and 3) variation analysis of the detected regions. We applied the new functionalities to analyze the use case of strigolactone biosynthesis in rice, a trait that has been previously studied[7][8]. We use this case study for validation of our results and further examine the structural variation in the QTL regions involved. We demonstrate that our methods are already useful for the study of quantitative traits. Moreover, we propose that they can serve as a basis for the develoment of future functionalities related to functional annotation, integration of phenotypic data, and gene prioritization.

Methods

Construction of pangenomes

Pangenomes were constructed in PanTools for two diploid plant species: *Arabidopsis thaliana* and *Oryza sativa*. The pangenome of *A.thaliana* was built using genomic and annotation data from 19 accesssions[9]. All *A.thaliana* assemblies are at chromosome level.

The pangenome of *O.sativa* was built with genomes and annotations of three cultivars: Nipponbare, DJ123 and IR64. *O.sativa* is divided in two major subspecies: Indica and Japonica. Nipponbare is a Japonica cultivar, while DJ123 and IR64 are Indica cultivars. For Nipponbare, we used the chromosome-level MSU v7.0 assembly and annotation[10]. For DJ123 and IR64, scaffold level assemblies and their corresponding annotations were used[11].

QTL data integration

Within the rice pangenome, we incorporated the regions of nine (9) QTL relevent to strigolactone biosynthesis, retrieved from a Bala x Azucena recombinant inbred line (RIL) population. In the *A.thaliana* pangenome, we used a subset of QTL from a larger published collection, which were mapped using a Bay-0 x Shahdara RIL population [12]. Artificial QTL (i.e. simple text files defining genomic regions) were added in oth pangenomes for development and testing purposes.

To facilitate the QTL annotation, we implemented a QTL feature node in PanTools. Explaining the context for this decision requires some understanding of how pangenomes are represented in PanTools. A pangenome in PanTools is represented as a compressed generalized de Brujn graph. The graph has separate layers for sequence, annotation and relationships (eg homology groups and clusters). In the annotation layer, genomic features such as genes and mRNAs are represented as nodes. We added an additional type of feature node to represent QTLs. In order to take advantage of pre-existing annotation functionality of PanTools, we reformatted the QTL data as genomic features in generic feature format (gff3); attributes that are specific to QTL are included as tags in the attribute column (S1). We justify this reformatting by the fact that QTL are included in the controlled vocabulary of the Sequence Ontology project, which makes them valid inputs for gff3 files [13]. We further implemented functionality to generate a summary table that gives an overview of all QTL present in the pangenome. The table includes six standard columns: genome, chromosome or scaffold, start position, end position, size, and number of genes in the QTL. Depending on the content of the QTL file, the table can adapt to include columns for logarithm of odds (LOD) score, explained variance, and parental lines.

Homology detection

The idea behind our approach for homology detection is quite simple. First , we retrieve the homologs of every gene in a QTL. Then, on a given genome, we stitch the homologous genes together to define a homologous region, which may be fragmented and consist of multiple sub-regions. Throughout this paper, the term 'original genome' refers to the genome where the QTLs are mapped, and 'queried genome'

refers to the genome(s) we want to detect homologous regions on.

Grouping

To retrieve the homologous genes, we utilized the homology grouping functionality already included in PanTools[14]. This is a fast, k-mer based approach that identifies similar mRNA across the pangenome and places them into groups, such that all members of a group are considered to be homologous.

Grouping in PanTools is influenced by a set of parameters which can be set by the user. There are eight default groups of parameter settings, ranging from D1 (strictest) to D8 (most relaxed). Stricter settings are more suitable for pangenomes where the individual genomes are phylogenetically close. Based on the results of previous work[14], we used D2 settings for the *A.thaliana* pangenome. In order to determine the optimal settings for the *O.sativa* pangenome, we benchmarked all eight settings against the BUSCO Liliopsida dataset [15]. We evaluated the results based on the resulting precision, recall, and F-score measures. We selected the D3 setting for the *O.sativa* pangenome because it results in the highest F-score (S5).

Construction of the homology tables

Homology tables summarize relationships between genes and their homologs. The required input is a QTL-annotated pangenome graph. The method takes two optional arguments: 1) the names of the QTLs to be analyzed, and 2) the genomes tp be queried (i.e. the genomes where homologous regions should be detected.) By default, the method will examine all QTLs and all genomes. The algorithm iterates first over the QTL names, then over the genome numbers. The output is a homology table for each QTL-queried genome combination. Each homology table contains the relationships between genes and their homologs, as well as positional information of the genes.

The construction itself works as follows: the algorithm walks through the QTL and examines each mRNA node, detecting the homology group it belongs to and, by extension, all other members of the same group. Detected members are retained if they belong to either 1) the currently queried genome, or 2) the original genome but are outside the boundaries of the QTL itself. Subsequently, mRNAs are replaced by their parent genes, a step that effectively compresses alternatively spliced isoforms and intron-variants. Finally, the groups and their associated information are returned in tabular format.

In the table, the queried genome is allocated four columns corresponding to: the gene identifier, the chromosome or scaffold the gene is on, the start position of the gene, and the end position of the gene. By sorting by scaffold and position, we can easily detect the boundaries of the homology regions on other genomes in the pangenome. The original genome is allocated eight columns that follow the same scheme. The extra four columns are to detect in-paralogs that fall outside the boundaries of the QTL. These in-paralogs are not used in detecting homology regions per se; however, they provide useful additional context: for example, we may wish to exclude detected fragments that clearly originate from paralogous regions.

Expansion of homology tables

As mentioned, homologous regions may be fragmented into sub-regions. From the rice homology tables, we used the start and end position of each detected scaffold as boundaries for that particular sub-region. Sub-regions were taken into account in subsequent steps, namely walk-through, visualization, and variation analysis. Isolated genes and smaller (less than 4 genes) fragments were kept in the homology table, but not taken into account in subsequent analusis within the scope of this project. This choice stems from a trade-off between the need to capture as much information as possible, and the impracticability of homologous regions being too fragmented.

In the *A.thaliana* pangenome, we applied the same rules but excluded fragments that clearly originated from paralogous regions of the original genome (S10) .

We implemented functionality to walk through the detected homologous regions and add genes that are present in the queried genome, but absent from the original genome. The input of this function is a file that contains the path to the homology table file constructed in the previous step, and the coordinates of the regions to be searched. The output is the extended homology table that also contains the added genes.

Visualizations and variation analysis

In order to provide context to the visualizations and the structural variation analysis, we classified our groups into core, core and single copy orthologs, accessory and unique. Classification was performed with a custom Python script on the output file of homology grouping. The classification terms are defined as follows:

1. core: present in all genomes; may have multiple members per genome.
2. core and single copy orthologous: present in all genomes; exactly one member per genome.
3. accessory: present in more than one genome, but not all.
4. unique: present in exactly one genome.

We visualized QTLs in Python/Matplotlib.

Additional information was added to the homology tables for structural variation analysis: specifically, known prioritititized genes from the rice use case were marked as such. Moreover, when a gene was annotated as a transposable element, it was marked as such in the homology tables.

We quantified structural variation events on the rice QTLs using custom Python scripts. We examined presence-absence variation (PAV), copy number variation (CNV) and complex variation, defined as a combination of PAV and CNV (S6). We compared the frequency of occurrence of various event types with their overall frequency in the rice pangenome.

Parental data

We retrieved resequencing data for Bala and Azucena from the European Nucleotide Archive (ENA). The study accession is SRP011382. These samples were selected because they were used the original study we compared our results with[7]. We used

the read mapping functionality of PanTools[16] to align the resequencing parental data against the rice pangenome. A sequence alignment/map (sam) file was generated for each genome in the pangenome. We used samtools[17] to add read groups and sort the files, and sambamba[18] to mark duplicate reads. Using samtools, we filtered the bam files for reads with a mapping quality of at least 15. Joint variant calling was performed with Freebayes[19] using three samples for Bala and two for Azucena. Filtering for read depth equal to or over 20, and subsetting of the resulting vcf files, were performed using vcftools[20]. The major deletions observed in the case study, as well as Variants of the major QTL for strigolactone biosynthesis in, were visualized in the Integrative Genomics Browser (IGV)[21].

Results

Functionality development

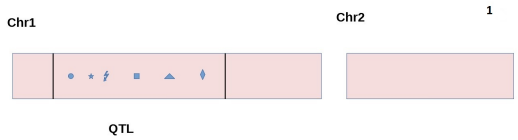
PanTools was expanded with novel functionalities for analyzing QTLs in a pangenomic context. The pangenomic QTL analysis workflow is outlined in Figure 1. In this context, QTL analysis involves annotating the pangenome with known QTLs, extracting homologous regions of the QTLs on other genomes in the pangenome, and constructing an overview of the regions and the links between homologous genes. The required inputs are a pangenome graph and some QTL data to be integrated into the pangenome and analyzed. Optional inputs include a file specifying which QTL should be analyzed, and a second file specifying which genomes should be included in the analysis. If the optional inputs are not given, all QTL will be analyzed for all genome combinations.

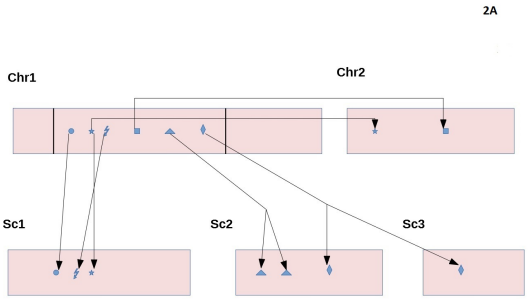
The workflow includes four steps. In the first step the pangenome is annotated with QTL data, which are added to the pangenome as feature nodes. It is assumed that the QTL data are defined in genomic coordinates. It is important to realize that QTL are typically mapped on the reference genome of a species: therefore, at this point in the workflow, QTLs are associated with only one of the genomes in the pangenome. The annotated QTL regions are used as a starting point for subsequent steps.

The next steps involve the detection and analysis of homologous regions of the annotated QTLs. This is the turning point where the 'pangenomic' part of the analysis starts to come into play: in order to analyze the differences between a certain region and its equivalent region on a different genome, we first need to detect that equivalent region. By utilizing a pangenome graph, it is possible to detect homologous regions that for any combination of the included genomes. However, the question remains on how the homology detection, on any given genome, will be performed.

The approach we use for homology detection relies on the previously developed homology grouping functionality of PanTools. Homology grouping detects homologous genes across all genomes, and places them in groups: genes that are considered homologous to each other are placed in the same homology group.

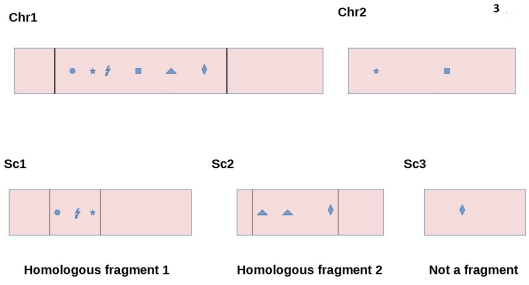
Figure 1: **Project workflow.** Visualization of workflow. (1) QTL data are integrated in the pangenome. (2A), (2B) Homologous genes are detected and organized in a homology table. (3) Homologous regions are extracted from the table. (4A), (4B) Detected homologous regions are searched for genes absent in the original genome. These genes are added to the tables.

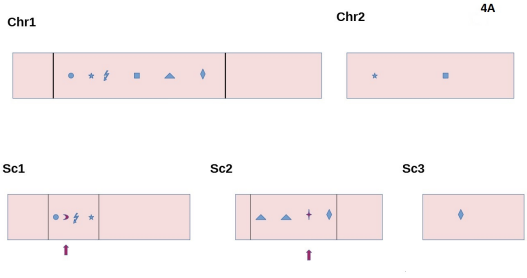




2B

Group	Genome 1	Genome 1 (homologs)	Genome 2
1	Circle	__	Circle'
2	Star	Star'	Star''
3	Bolt	__	Bolt'
4	Square	Square'	__
5	Triangle	__	Triangle' Triangle''
6	Diamond	__	Diamond' Diamond''





4B

Group	Genome 1	Genome 1 (homologs)	Genome 2
1	Circle	__	Circle'
2	Star	Star'	Star''
3	Bolt	__	Bolt'
4	Square	Square'	__
5	Triangle	__	Triangle' Triangle''
6	Diamond	__	Diamond' Diamond''
7	__	__	Moon

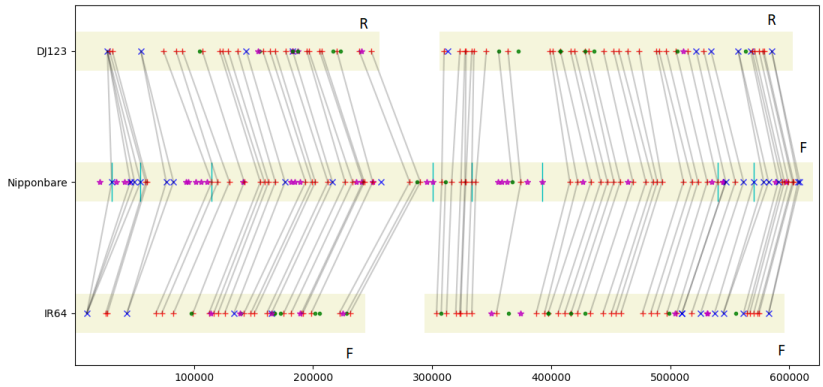
Our novel homology detection method then capitalizes on that functionality, and expands it to regions. The basic idea is as follows: having knowledge of all genes in a region and their order, and being able to detect the homologs of each gene on another genome, we can then stitch the detected homologs together in order to define a homologous region. Homology detection encompasses steps 2, 3 and 4.
In step 2, homology grouping is run under optimal settings . Our algorithm then walks over the QTL region under study and examines each gene, detecting all links to homologous genes. Since homology grouping is performed at a transcript level, this step requires an under the hood simplification of mRNAs to genes . All detected genes and the relationships between them are stored in a homology table. A simplified version of the table is shown in Figure 1; the full table also contains positional information (S2).
In step 3, boundaries (start and end positions) are established for the homologous regions on other genomes. Regions may be discontinuous. Scattered genes are not fragments, and therefore are not used in this step, although they are still included in the table. What constitutes a fragment is explained in more detail in . Finally, in step 4, the algorithm walks through the homologous fragments to detect extra genes that were not present in the original genome. These genes are then also added to the homology table.
To demonstrate what the output looks like, we applied our workflow to the *A.thaliana* pangenome and visualized a QTL mapped on the Col-0 accession (region Chr3:14720735-15245133) (S3). Due to high similarity between the accessions, there are few accessory and no unique groups, and the order of the genes is retained.

Use case: strigolactone biosynthesis in rice

To demonstrate the functionalities and interpret the results, we used a case study regarding strigolactone biosynthesis in rice. Strigolactones are a class of plant hormones that affect plant architecture and are involved in rizosphere signaling [22][23]. The pathway of strigolactone biosynthesis was independently studied by Cardoso et al. [7] and Zhang et al. [8]. Both studies identified the same two causal genes: SLB1 and SLB2. In Cardoso et al.[7], these genes were found to be present in rice cultivar Azucena, but absent in Bala. Azucena belongs to the Japonica subspecies, while Bala belongs to the Indica subspecies. Further analysis showed that the genes were consistently absent across Indica cultivars, and consistently present across Japonica cultivars.
We used data of nine QTL affecting strigolactone biosynthesis [7], as well as a list of genes that were identified as candidates for affecting the trait using a gene prioritization method [1].
The workflow was applied to the rice pangenome. The figure below displays qSLB1.1, which is the major QTL affecting strigolactone biosynthesis in rice as identified by Cardoso et al. [7]. Genomes of the rice pangenome are represented by continuous or fragmented bars. The middle bar represents qSLB1.1 mapped in chromosome 1 on Nipponbare; the collection of fragmented bars at the bottom correspond to IR64 (scaffolds 181 and 203), and the bars at the top correspond to DJ123 (scaffolds 30 and 107). Genes are displayed by several different markers

that correspond to different gene types. Specifically, red crosses are single copy orthologs; blue x's are core and not single copy orthologs; green circles are accessory; and magenta stars are unique. Genes that are homologous to each other are linked with black lines; such links are shown between Nipponbare and DJ123 as well as Nipponbare and IR64, but not between DJ123 and IR64, or within genomes. Finally, prioritized genes are additionally marked by cyan lines.

Figure 2: **Rice qSLB1.1 on Nipponbare and homologous regions on DJ123 and IR64.** QTL1 mapped on Nipponbare (middle, chromosome 1) and its homologous regions on DJ123 (top, scaffolds 30 and 107) and IR64 (bottom, scaffolds 181 and 203). Markers are genes, with black lines representing links between homologous genes. Red crosses are core and single-copy orthologs; blue x's are core and not single-copy orthologs; green circles are accessory; magenta are unique. Prioritized genes are denoted by cyan lines.



Homology grouping placed SLB1 and SLB2 in the same group, along with a third homolog, LOC_Os01g50590 (Table 1). LOC_Os01g50590 is a known homolog of SLB1 and SLB2 that has been demonstrated to be inactive[8]. The Indica cultivars of the rice pangenome have one gene each in the same homology group.

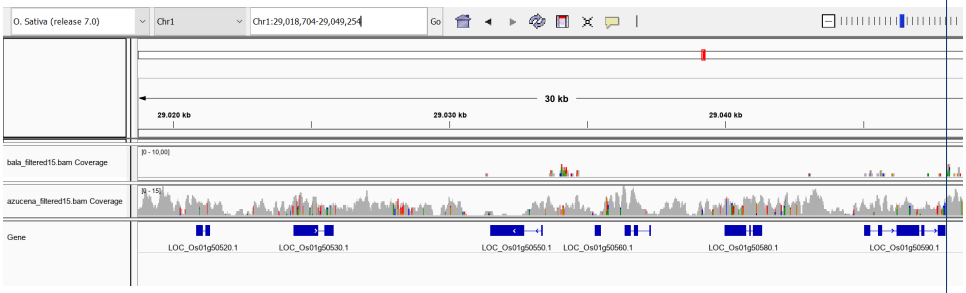
Table 1: The homology group containing the casual genes. Columns correspond to genomes Nipponbare, DJ123 and IR64. Cells contain gene identifiers.

Nipponbare	DJ123	IR64
LOC_Os01g50520 (SLB1)	maker-scaffold_30-snap-gene-2.35	maker-scaffold_203-snap-gene-1.34
LOC_Os01g50580 (SLB2)	maker-scaffold_30-snap-gene-2.35	maker-scaffold_203-snap-gene-1.34
LOC_Os01g50590	maker-scaffold_30-snap-gene-2.35	maker-scaffold_203-snap-gene-1.34

Pairwise alignments of the genes showed that the Indica genes are more similar to the inactive homolog than SLB1 or SLB2 (S4) . This supports the conclusion that SLB1 and SLB2 are absent in Indica.
To further support this, we mapped parental resequencing data on Nipponbare and

visualized the alignment (bam) files in IGV (after filtering for alignments that have a mapping quality of at least 15). The upper tracks show the coverage distributions on Bala (top track) and Azucena (middle track); the lowest track shows the gene models. The deletions of LOC_Os01g50520 (SLB1) and LOC_Os01g50580 (SLB2) on Bala can be inferred by the lack of mapped reads on these loci. A small amount of Bala reads map on the inactive ortholog; this is still a smaller amount than the Azucena reads that map on the same position. This can be explained by the fact that Azucena is taxonomically closer to Nipponbare (both Japonica cultivars) than Bala (an Indica cultivar) is. It is also consistent with the query covers reported by BLAST (S4) for DJ123 and IR64.

Figure 3: **Region with deletions of SLB causal genes in Indica.** Region on Nipponbare where the deletions of causal genes SLB1 and SLB2 are observed. Gene models are shown on the lwoer track, while gene coordinates are on the upper track. The two middle tracks represent coverage distributions of mapped reads (mapping quality ¿= 15) for Bala (upper middle) and Azucena (lower middle). SLB1 and SLB2 are deleted on Bala.



Structural variation overview

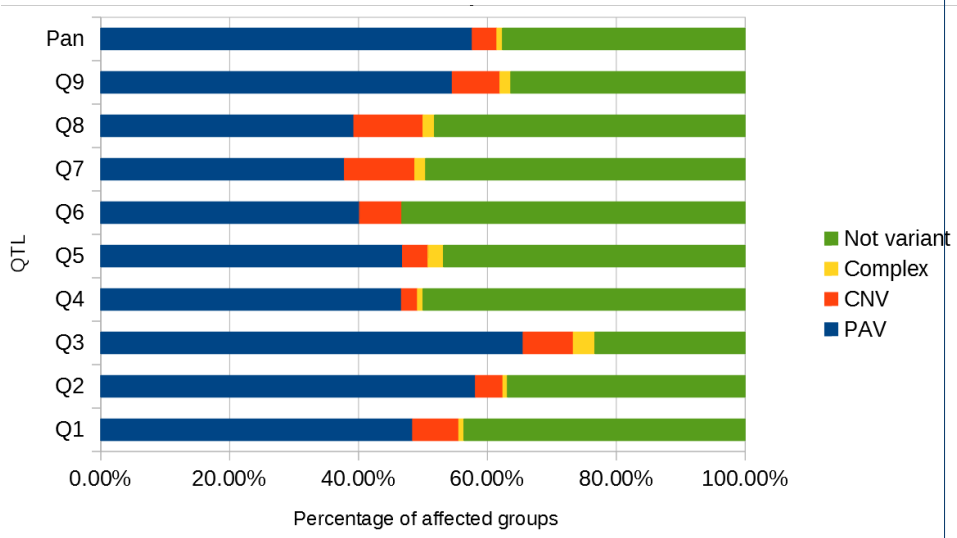
Three types of structural variation were considered: presence absence variation (PAV), copy number variation (CNV) and complex variation, which is a combination of PAV and CNV.

Some definitions of PAV consider it as an extreme version of CNV [24], while others treat the two separately [25]. In this project, PAV is defined as a gene and its homologs being absent from one or more genomes. In constrast, CNV refers to genes being present in all genomes, but with variation in the number of copies. Complex events are when a gene is affected by both PAV and CNV: for example, it has 3 copies on the first genome, zero on the second, and two in the third.

Under these definitions, a correspondence can be drawn between homology groups, and types of structural variation. Unique groups, as well as accessory groups without copy number variation, correspond to PAV events; core groups with copy number variation correspond to CNV events; and accessory groups that also have copy number variation correspond to complex events. The rest of the groups are considered non-variant within the scope of this project, although it should be noted that in reality they may still be affected by another type of structural variation, such as an inversion.

Percentages of affected groups were calculated in each QTL as well as across the pangenome. Results are displayed in the figure below.

Figure 4: **Structural variation across rice QTLs.** Percentage of groups affected by various types of SV across SLB-related rice QTL. Blue represent PAV, red represents CNV, yellow represents complex variation, and green represents non-variant groups. The top bar corresponds to the overall variation in the pangenome; the other nine bars correspond to the nine rice QTLs.



Additionally, PAV variation was broken down to unique and accessory groups for all individual genomes and pairwise genome combinations (S8) . Groups that contain transposable elements (TE) were marked as such.

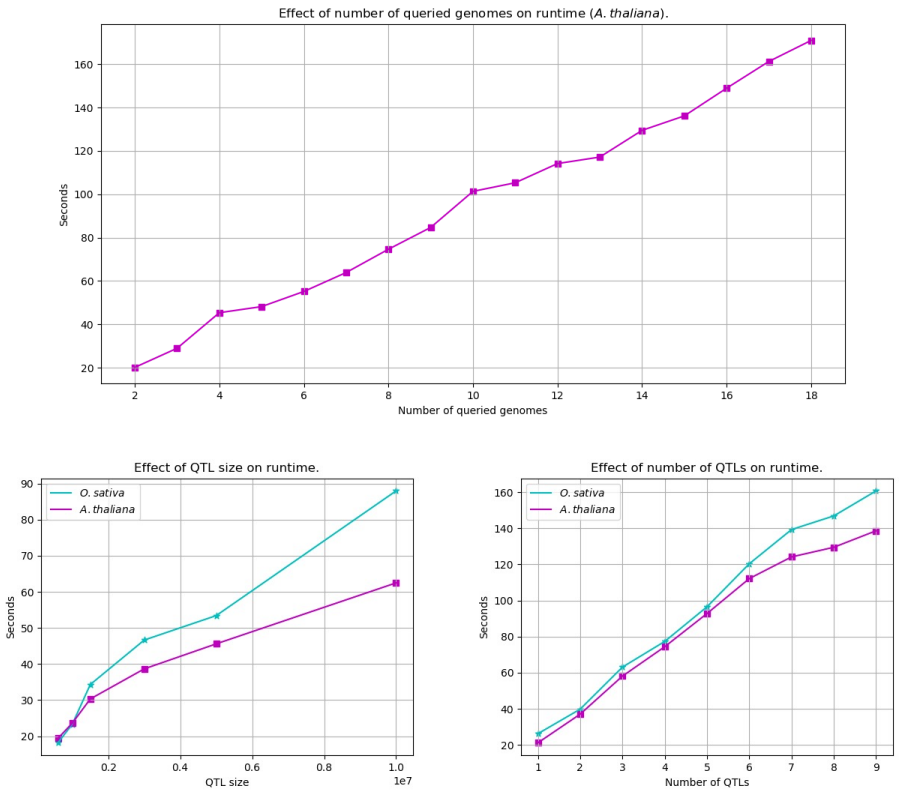
Runtime analysis

The most time-consuming step of homology detection is the construction of homology tables. The runtime of this step was measured against three factors: 1) the size of QTLs analyzed, 2) the number of QTLs analyzed, and 3) the number of queried genomes. For all data points, three measurements were taken and the median was reported.

The effect of each factor is shown in Figure 5. The upper panel shows the effect of the number of queried genomes (2-18): this is only measured in *A.thaliana* because the pangenome of *O.sativa* only contains 3 genomes. The lower left panel shows the effect of QTL size on runtime. These measurements were taken on two (2) queried genomes, so that a comparison between *O.sativa* and *A.thaliana* could be drawn. QTL sizes between 600000 and 10000000 bp were tested. Finally, the lower right panel displays the effect of the number of QTLs on runtime. These measurements were taken under a constant QTL size of 600000 bp and on two (2) queried genomes;

this was again to enable comparisons between the two organisms.

Figure 5: **Runtime analysis.** Effect of various factors on the runtime of homology table construction. The upper graph displays the effect of the number of queried genomes. The lower graphs display the effect of QTL size (left) and QTL number (right).



Discussion

Homology detection based on groups

We opted for a homology detection method that relies on the previously developed homology grouping approach. It would be possible to use whole genome alignments instead. However, global alignments are very slow, and assume homology over the entire length of the region, while local alignments score too low on both sensitivity and specificity when applied at a genome-wide scale[26]. In the end we opted for our approach because its reliance on homology grouping makes it fast, accurate, and applicable to sets of genomes with various degrees of evolutionary distance.

Structural variation analysis

We introduced a method to quantify structural variation based on groups rather than genes . We argue that this approach is more in line with the pangenomic

perspective than a gene-based quantification would be.
Structural variation was shown to be quite widespread, affecting from 46.72% to 76.67% of groups across QTL regions (S7) . Under the metrics used, PAV appears to be the most common type ranging from 37.82% to 65.56%, followed by CNV (2.5-10.92%), and finally complex variation (0-3.33%). It is also important to take into account the low number of genomes (3) in the rice pangenome: structural variation percentages would likely be higher in a larger pangenome, as added individuals introduce new genes and thus increase diversity.
In the course of this study, we made the hypothesis that QTL regions would be more enriched in SV, compared to the entirety of the pangenome. This hypothesis was based on a variety of studies which investigate a potential link between structural variation and quantitative traits [24] [27] [28] [29] [30]. The results do not support this hypothesis: structural variation varies quite a bit between different QTL, which can have a slightly lower or slightly higher percentage than the pangenome average. An explanation for this is that every part of the pangenome is responsible for some kind of functionality.
In the breakdown of PAV and complex variation, we can see that Nipponbare has the highest number of unique groups, while the DJ123-IR64 combination has the highest number of accessory groups. This is consistent with the evolutionary history of *Oryza sativa*: DJ123 and IR64 belong to the Indica subspecies, and as such are expected to be more similar to each other than either is to Nipponbare.
TE groups appear relatively frequently. This observation seems to be in agreement with previous studies that have identified TEs as sources of structural variation and adaptive evolution in plants [31] [32].

Automation and speed

QTL annotation is a fully automated process; the resulting summary table is adaptable to specific attributes of the data, such as LOD score or names of the parental lines. In homology detection, some user input is required between the original construction of the tables and their expansion. While this choice sacrifices some amount of automation, the gain in customization is far higher: the user can chose exactly which fragments to focus on for their specific analysis, making the method very adaptable and generally applicable.
Homology detection is very fast (Figure 5). The most impactful factor appears to be the number of queried genomes, while the least impactful is QTL size. While three factors were explicitly tested, a fourth factor that influences runtime is the size of the organism's genome. The effect of this factor can implicitly observed in the plots: O.sativa (with a genome size of approximately 500 Mb) generally requires a higher running time than A.thaliana (genome size of approximately 135 Mb).

Limitations

The type of visualization used in this project works well for up to three genomes, but does not scale further than that. A potential solution to this problem would be the utilization of circos plots [33]. Circos plots are very flexible and have been

extensively used for the visualization of comparative genomic data. We propose two specific applications: 1) a user might wish to examine structural or sequence level variation from the perspective of a certain genome. Then, the point-of-view genome would be represented as the outer ideogram in the plot. The other genomes would be represented as inner circles, and points of variation would be marked as such. The possibilities for customization are endless: for example, higher bars would represent higher gene copy numbers (for structural variation), or SNPs might be marked with a different color than insertions-deletions (for sequence-level variation). Alternatively, a user might wish to focus on a specific QTL. Then, the examined QTL could be visualized as a part of the circumference of a discontinuous circular ideogram. The homologous regions on other genomes would represent the other parts of the circumference, lined up next to each other. Homologous genes would be linked by curved lines.

Our methods assume a haploid representation of a genome, and do not attempt to deal with genetic phasing and heterozygosity. One way to handle that would be to make PanTools haplotype-aware: namely, by using resolved haplotypes as inputs for pangenome building, and representing each resolved haplotype as a separate 'genome' in the pangenome graph. In the long term, a 'nested pangenome' structure could be developed: the pangenome would consist of separate genomes, and each separate genome would be its own 'mini pangenome', consisting of different haplotypes. However, we are aware that the implementation of these ideas would require availability of genomes with fully resolved haplotypes, which is a significant challenge in itself[34], especially for polyploid data [35]; thus, it might not be feasible in the near future.

We tested our methods on diploid data, as well as on both chromosome- and scaffold-level assemblies. However, both of them pangenomes we used consisted of subspecies that are taxonomically close to each other. In the future, further testing could be performed on polyploid data, as well as on pangenomes consisting of genomes that are more taxonomically distant from each other.

In structural variation analysis, we considered PAV, CNV and complex variation, where 'complex' is defined as consisting of a combination of PAV and CNV variation. Going forward, more types of structural variation could be taken into account, such as inversions or translocations.

We used some simple heuristics to define what constitutes a homologous sub-region within the scope of this project: namely, a minimum threshold of four genes per sub-region, and, for chromosome-level assemblies (i.e. the *A.thaliana* pangenome) the exclusion of sub-regions that correspond to paralogous sub-regions on the original genome. In order to improve this step in the future, distance between sub-regions that lie on the same scaffold should be taken into account as an additional parameter. Within this project, we allowed discontinuity in the sense that sub-regions can lie on separate scaffolds. This worked well for our rice pangenome because it was highly fragmented and the genomes were relatively close taxonomically. Going forward, it would make sense to make it possible for sub-regions on the same chromosome, or scaffold, to be discontinuous: if two fragments on the same chromosome are distant enough from each other, they should be considered as separate fragments.

It is important to distinguish between the case where the QTL region is split on the queried genome, and the case where the QTL region has a paralogous smaller region on the original genome that lies outside the QTL boundaries, resulting in a corresponding smaller region on the queried genome (S10) . In the *A.thaliana* pangenome, we encountered the second scenario, and excluded the smaller regions on the basis that they correspond better to paralogous fragments within the original genome, rather than the QTL itself.

In either case, whether the smaller sub-regions should be considered in subsequent steps (ie table expansion), depends on the specific analysis. For example, if the purpose is to search for gene function, it might make sense to take all subregions into account, regardless of their origin.

A final potential limitation concerns the definition of 'copies' in the analysis of structural variation. For the purposes of the analysis, all members of a homology group were considered as copies of each other. This is not incorrect per se, but it does not distinguish between paralogs and orthologs, which may be important information for some types of studies (eg evolutionary studies). Distinguishing between orthologs and paralogs is an everpresent challenge in genomics and does not have an easy solution.

Future work

Homology regions extracted from our workflow can currently be manually examined for interesting genes, or queried for structural variation, as demonstrated in this report with the example of strigolactone biosynthesis. Going forward, the genes in the detected homologous regions can be used as input lists for gene prioritization methods, thus allowing for efficient discovery of candidate causal genes. While this was possible before for one genome, the fact that different gene lists can now originate from multiple genomes opens possibilities for adding statistical power to the prioritization methods. A possible direction to explore in this regard would be to check whether, using a given method, the same (homologous) genes are consistently prioritized.

Taking it a step further: prioritization methods typically make use of various data types, such as functional annotation terms (ref), evolutionary (ref), or RNA-seq [36] data. These types of data could be integrated into PanTools and used in conjunction with our homology detection approach: for example, instead of just outputting a homologous genomic region, PanTools would output a homologous genomic region and its associated gene ontology (GO) terms. Moreover, integration of various data types would be useful by itself, outside of gene prioritization methods. For example, detecting homologous regions and summarizing the structural variation for a QTL could be integrated with phenotypic data for that QTL: this is quite powerful, as it provides an immediate link between genotypic and phenotypic variation.

Finally, PanTools provides functionality to map resequencing data to any of the genomes in a pangenome[16]. This means we can currently map resequencing data from the parental population used to map the trait: instead of mapping the parental data on the reference genome, each parent can be mapped on the taxonomically closest genome available in the pangenome. Potentially, variants could be called

and visualized within PanTools, which would be a powerful technique for getting an overview of sequence-level variation of a parent in comparison to its closest relative. Variant calling and visualization in PanTools is currently under development. To provide a glimpse of what this functionality might look like in the future, we visualized variants of Bala on cultivar DJ123 using IGV (S9).

Conclusion

In conclusion, we developed novel tools to annotate pangenomes with QTL data and analyze these data in a pangenomic context. Our workflow is immediately useful, as it allows the user to get a detailed overview of the structural variation inside the regions of interest. This immediate application was demonstrated on the example of strigolactone biosynthesis in rice, where we observed high prevalence of structural variation events, with the most common type being PAV.

Moreover, the work performed in this project opens exciting possibilities for integration with diverse data types, and lays the foundation for the development of further pangenomic functionalities. For example, gene lists from homology regions detected by our methods can be enhanced with e.g. functional annotation terms or RNA-seq data, and used as inputs in gene prioritization methods. Further, the combination of a structural variation summary derived from our tools with phenotypic data could provide a clear description of the link between genotypic and phenotypic variation in regions of interest.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

I would like to thank my supervisors Sandra Smit and Aalt-Jan van Dijk for the opportunity to work in this project and their support throughout. I would also like to acknowledge Eef Jonkheer for his help with PanTools, and Raul Wijfjes for his help with FreeBayes.

References

1. Bargsten, J.W., Nap, J.-P., Sanchez-Perez, G.F., van Dijk, A.D.: Prioritization of candidate genes in qtl regions based on associations between traits and biological processes. BMC plant biology **14**(1), 330 (2014)
2. Lin, F., Fan, J., Rhee, S.Y.: Qtg-finder: a machine-learning based algorithm to prioritize causal genes of quantitative trait loci in arabidopsis and rice. G3: Genes, Genomes, Genetics **9**(10), 3129–3138 (2019)
3. Warwick Vesztrocy, A., Dessimoz, C., Redestig, H.: Prioritising candidate genes causing qtl using hierarchical orthologous groups. Bioinformatics **34**(17), 612–619 (2018)
4. Voss-Fels, K.P., Stahl, A., Hickey, L.T.: Q&a: modern crop breeding for future food security. BMC biology **17**(1), 18 (2019)
5. Pantoja, Y., da Costa Pinheiro, K., Araujo, F., da Costa Silva, A.L., Ramos, R.: Bioinformatics approaches applied in pan-genomics and their challenges. In: Pan-genomics: Applications, Challenges, and Future Prospects, pp. 43–64. Elsevier, ??? (2020)
6. Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D., Smit, S.: Pantools: representation, storage and exploration of pan-genomic data. Bioinformatics **32**(17), 487–493 (2016)
7. Cardoso, C., Zhang, Y., Jamil, M., Hepworth, J., Charnikhova, T., Dimkpa, S.O., Meharg, C., Wright, M.H., Liu, J., Meng, X., *et al.*: Natural variation of rice strigolactone biosynthesis is associated with the deletion of two max1 orthologs. Proceedings of the National Academy of Sciences **111**(6), 2379–2384 (2014)
8. Zhang, Y., Van Dijk, A.D., Scaffidi, A., Flematti, G.R., Hofmann, M., Charnikhova, T., Verstappen, F., Hepworth, J., Van Der Krol, S., Leyser, O., *et al.*: Rice cytochrome p450 max1 homologs catalyze distinct steps in strigolactone biosynthesis. Nature chemical biology **10**(12), 1028 (2014)
9. Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., *et al.*: Multiple reference genomes and transcriptomes for arabidopsis thaliana. Nature **477**(7365), 419–423 (2011)
10. Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., *et al.*: Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. Rice **6**(1), 4 (2013)

11. Schatz, M.C., Maron, L.G., Stein, J.C., Wences, A.H., Gurtowski, J., Biggers, E., Lee, H., Kramer, M., Antoniou, E., Ghiban, E., *et al.*: Whole genome de novo assemblies of three divergent strains of rice, oryza sativa, document novel gene space of aus and indica. Genome biology **15**(11), 506 (2014)
12. Serin, E.A., Snoek, L.B., Nijveen, H., Willems, L.A., Jiménez-Gómez, J.M., Hilhorst, H.W., Ligterink, W.: Construction of a high-density genetic map from rna-seq data for an arabidopsis bay-0× shahdara ril population. Frontiers in genetics **8**, 201 (2017)
13. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M.: The sequence ontology: a tool for the unification of genome annotations. Genome biology **6**(5), 44 (2005)
14. Anari, S.S., de Ridder, D., Schranz, M.E., Smit, S.: Efficient inference of homologs in large eukaryotic pan-proteomes. BMC bioinformatics **19**(1), 340 (2018)
15. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.: Busco: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19), 3210–3212 (2015)
16. Anari, S.S., de Ridder, D., Schranz, M.E., Smit, S.: Pangenomic read mapping. bioRxiv, 813634 (2019)
17. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The sequence alignment/map format and samtools. Bioinformatics **25**(16), 2078–2079 (2009)
18. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., Prins, P.: Sambamba: fast processing of ngs alignment formats. Bioinformatics **31**(12), 2032–2034 (2015)
19. Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 (2012)
20. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., *et al.*: The variant call format and vcftools. Bioinformatics **27**(15), 2156–2158 (2011)
21. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in bioinformatics **14**(2), 178–192 (2013)
22. Ruyter-Spira, C., Al-Babili, S., Van Der Krol, S., Bouwmeester, H.: The biology of strigolactones. Trends in plant science **18**(2), 72–83 (2013)
23. Al-Babili, S., Bouwmeester, H.J.: Strigolactones, a novel carotenoid-derived plant hormone. Annual review of plant biology **66**, 161–186 (2015)
24. Gabur, I., Chawla, H.S., Lopisso, D.T., von Tiedemann, A., Snowdon, R.J., Obermeier, C.: Gene presence-absence variation associates with quantitative verticillium longisporum disease resistance in brassica napus. Scientific reports **10**(1), 1–11 (2020)
25. Springer, N.M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., Rosenbaum, H., et al.: Maize inbreds exhibit high levels of copy number variation (cnv) and presence/absence variation (pav) in genome content. PLoS genetics **5**(11) (2009)
26. Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., Hardison, R.C.: Cross-species sequence comparisons: a review of methods and available resources. Genome Research **13**(1), 1–12 (2003)
27. Gabur, I., Chawla, H.S., Snowdon, R.J., Parkin, I.A.: Connecting genome structural variation with complex traits in crop plants. Theoretical and applied genetics **132**(3), 733–750 (2019)
28. Alix, K., Gérard, P.R., Schwarzacher, T., Heslop-Harrison, J.: Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Annals of Botany **120**(2), 183–194 (2017)
29. Neik, T.X., Barbetti, M.J., Batley, J.: Current status and challenges in identifying disease resistance genes in brassica napus. Frontiers in plant science **8**, 1788 (2017)
30. Żmieńko, A., Samelak, A., Kozłowski, P., Figlerowicz, M.: Copy number polymorphism in plant genomes. Theoretical and applied genetics **127**(1), 1–18 (2014)
31. Akakpo, R., Carpentier, M.-c., Hsing, Y.I., Panaud, O.: The impact of transposable elements on the structure, evolution and function of the rice genome. New Phytologist (2020)
32. Lisch, D.: How important are transposons for plant evolution? Nature Reviews Genetics **14**(1), 49–61 (2013)
33. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A.: Circos: an information aesthetic for comparative genomics. Genome research **19**(9), 1639–1645 (2009)
34. Browning, S.R., Browning, B.L.: Haplotype phasing: existing methods and new developments. Nature Reviews Genetics **12**(10), 703–714 (2011)
35. Tang, H.: Disentangling a polyploid genome. Nature plants **3**(9), 688–689 (2017)
36. Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., Myers, C.L.: Integrating coexpression networks with gwas to prioritize causal genes in maize. The Plant Cell **30**(12), 2922–2942 (2018)

Additional Files

S1

The nine QTLs involved in strigolactone biosynthesis before and after reformatting for integration in PanTools.

S2

Full homology tables of qSLB1.1.

S3

A.thaliana QTL mapped on Col0 and visualized on Col0, Zu0 and Ler0.

S4

Pairwise alignments of DJ123 (lower panel) and IR64 (upper panel) genes against Nipponbare genes in the SL-related homology group.

S5
Table outlining recall, precision and F-score measurements calculated to find the optimal grouping settings for the rice pangenome.

S6
Figure that shows how structural variation percentages were calculated and the reasoning for basing the calculations on affected groups instead of affected genes.

S7
Structural variation overview on the nine rice QTL involved in strigolacone biosynthesis as well as on the pangenome.

S8
A more detailed presence-absence variation overview that notes accessory and unique groups as well as transposons.

S9
Variants of Bala with DJ123 (scaffold 30) as a reference.

S10
Visualization of the two distinct scenarios that may result in separate sub-regions on the same chromosome.