# Structural variation across the pangenome of flax (*Linum usitatissimum*)

Esme Padgett

St. Chad's College

Durham University

A thesis submitted for the Degree of *Integrated Masters of Biology (MBiol)*

Easter 2025

Supervisor of Thesis:

Dr. Adrian Brennan

# Abstract

Against mounting agricultural challenges, crop breeders are mining plant genomes for novel genetic variation and trait diversity. To better record genomic diversity, advancements in genome assembly and allelic variant detection are facilitating a transition from single reference genomes to pangenomes, which can compile genotype-phenotype associations and genetic variance across cultivated and wild plant species. Modern agriculture is genetically restricted, but pangenomes can innovate crop breeding by informing multi-trait selection and elucidating the evolutionary processes underpinning crop domestication and diversification.

Like many crops, flax (*Linum usitatissimum*) breeding is hindered by a limited understanding of complex and agronomic trait genetics. Here, we present the first pangenome of *L. usitatissimum*, collating genes and genome structural details from five published haplotype-collapsed assemblies of wild and cultivated flax, aided by high-throughput chromosome conformation capture (HiC) sequencing of two wild flaxes (*Linum bienne*).

Following scaffolding improvements to existing flax assemblies, the 564.8 Mb graph-based flax pangenome was generated without a reference genome to mitigate the reference bias within the study. Despite informational limits of haplotype-collapsed assemblies, 57,729 detected structural variants (SVs) establish a more comprehensive account of flax trait diversity; insertion, deletion, duplication, and inversion variants were well represented, measuring ~20 Mb across the pangenome. 1,252 gene ontology functions were identified within variant regions, wherein traits were associated with crop domestication. Preliminary assessments found wild flax SVs enriched with key agronomic functions; other pangenome SVs bear annotations of architectural, developmental, and stress response changes associated with domestication.

# Table of Contents

**Structural variation across the pangenome of flax (*Linum usitatissimum*) ..... 1**

**Abstract ..... 2**

**Table of contents ..... 3**

Table of Figures ..... 4

Acronyms and Abbreviations ..... 5

**Introduction ..... 6**

**Loss of genetic diversity in ancient and modern agriculture ..... 6**

**Advancing flax (*Linum usitatissimum*) breeding ..... 7**

Existing understanding of flax ..... 7

Surpassing technological limitations to variation detection ..... 8

Pangenomes to capture structural variation ..... 9

**Surveying flax with pangenomics ..... 10**

**Methods ..... 11**

**Genome assembly from public databases ..... 11**

Genome accession download ..... 11

Note on haploid assembly inputs ..... 12

Starting assembly statistics ..... 12

Starting genome assembly and quality control ..... 13

*L. usitatissimum* var CDC Bethune ..... 13

*L. usitatissimum* var Atlant ..... 14

*L. usitatissimum* var Longya, Heiya, and *L. bienne* ..... 15

**Assembly of the *L. bienne* genome ..... 15**

Wild plant material ..... 16

HiC read generation ..... 16

HiC scaffolding and the Dovetail Genomics™ HiRise assembly ..... 16

HiC sequence cleaning ..... 16

HiC scaffolding using HapHiC ..... 17

HiC-scaffolded assembly visualization ..... 17

**Final assembly of *L. usitatissimum* cultivars and *L. bienne* genomes ..... 18**

RagTag assembly ..... 18

**Post-assembly genome quality assessments ..... 19**

**Pangenome assembly ..... 19**

Inputs of PanGenome Graph Builder (PGGB) ..... 19

Algorithms inside PGGB pangenome creation ..... 20

Execution of PGGB ..... 20

**Structural Variation (SV) calling ..... 21**

Genome-level SV calling ..... 21

Pangenome-level SV calling ..... 22

SV cross-dataset validation ..... 23

SV Analyses ..... 23

Pangenome SVs ..... 23

PCA Plot ..... 23

**Functional Annotation ..... 24**

Annotating SVs of the pangenome ..... 24

Defining Gene Ontology Terms ..... 24

SV Set Enrichment Analysis ..... 25
**Results ..... 26**
**L. bienne assembly aided by HapHiC HiC scaffolding ..... 26**
**Improved genome assembly from publicly available accessions ..... 28**
QUAST hinted differences amongst varietal assemblies ..... 29
**Pangenome Construction ..... 30**
**Structural Variation Calling ..... 31**
Validated pangenome SV calls ..... 31
Diversity of pangenome SVs ..... 32
**Gene Ontology ..... 35**
SV Set Enrichment Analysis ..... 35
**Discussion ..... 39**
**The first pangenome draft of flax ..... 39**
**Inter-varietal differences demonstrated through pangenome SVs ..... 40**
Conservative estimation of wild flax diversity ..... 41
**Functional annotation differences in domesticated and wild flax ..... 42**
Pangenome-wide annotations ..... 42
Functional differences in domesticated and wild flax ..... 43
**Final thoughts ..... 44**
**References ..... 45**
**Appendix ..... 51**

## Table of Figures

*Figure 1 ..... 7*
*Figure 2 ..... 9*
*Table 1 ..... 12*
*Table 2 ..... 13*
*Table 3 ..... 15*
*Table 4 ..... 22*
*Table 5 ..... 22*
*Figure 3 ..... 24*
*Figure 4 ..... 26*
*Figure 5 ..... 27*
*Table 6 ..... 28*
*Figure 6 ..... 29*
*Figure 7 ..... 30*
*Figure 8 ..... 32*
*Table 7 ..... 32*
*Figure 9 ..... 33*
*Figure 10 ..... 34*
*Figure 11 ..... 35*
*Table 8 ..... 38*
*Table 9 ..... 39*

# Acronyms and Abbreviations

Acronym: Description

AM: Association Mapping

BUSCO: Benchmarking University Single-Copy Orthologs

bp: base pair(s), or just “b” when following a metric prefix e.g. kb for kilobase pairs

CEGMA: Core Eukaryotic Genes Mapping Approach

CWR: Crop Wild Relative

GFA: Graphical Fragment Assembly

GO: Gene Ontology

GWAS: Genome-wide Association Study

HiC: High-throughput Chromosome Conformation Capture

HPC: High-Performance Computing

MAPQ: MAPping Quality

NW: Read edit distance (based on Needleman-Wunsch algorithm)

PCA: Principal Component Analysis

PGGB: PanGenome Graph Builder

QUAST: QUality ASsessment Tool

RAPD: Random Amplified Polymorphic DNA

SLURM: Simple Linux Utility for Resource Management

SNP: Single-Nucleotide Polymorphism

SSR: Simple Sequence Repeat

SV: Structural Variant/Variation

VCF: Variant Call File

# Introduction

## Loss of genetic diversity in ancient and modern agriculture

Upwards of 350,000 plant species have been described in the World Checklist of Vascular Plants [1]. Yet among 7,000 domesticated, edible species [2], just a handful of species, like maize, rice, and wheat, work to meet global caloric needs. The modern human diet relies on ~255 plants: 26 cereals, 17 roots and tubers, 26 pulses, 44 vegetables, 69 fruits, 14 nuts, 28 oils, 24 herbs and spices, 3 sugars, and 4 stimulant crops [2].

These cultivated varietals, "cultivars", have been gradually domesticated from wild edible species, progressively biasing the genetic diversity in agriculture. Among allelic variants that emerged in wild and semi-domesticated "landrace" populations, favorable trait alleles became fixed, remaining within cultivated lineages. Food digestibility increased as organoleptic (sensorial) and nutritional quality traits were fixed in domesticated lineages. Beyond aligning with consumer preferences, domestication has promoted more regular and high-yielding harvests, aided by reduced seed shattering and adaptations to local climate conditions and pest-pathogen threats. Relative to their wild progenitors or crop wild relatives (CWRs), cultivated lineages heavily favor the growth of target tissues (e.g., fruit in tomatoes, seed stalks in rice).

Critical trait fixation in major crops, such as non-shattering (indehiscent) seeds in wheat, barley, and rice, arose over several millennia of domestication selection pressures. Modern breeding, however, imposes more intensive selection pressures that accelerate the emergence of favorable traits. During the mid-1960s to the mid-1980s, for instance, the fixation of semi-dwarfism in rice and wheat occurred in mere decades, supported by material institutional and governmental investment.

Through high-yield, dwarf varieties, the globe witnessed massive increases in agricultural productivity [3]. In two decades, the food supply of developing countries grew by 12%. During this Green Revolution, crop breeding efforts conservatively reduced arable land requirements in developing nations by 20 million hectares [4]. Consumers globally experienced lowered food prices. Without the productivity gains provided by high-yield cultivar innovation, caloric availability per capita would drop up to 7% in the poorest regions of the world [4].

Careful cultivar generation has introduced superior agronomic functions into the field, feeding millions. However, as an unintended consequence, modern agriculture has become dependent on a narrower set of crops. Already, 25% of animal and plant species face extinction threats [5]. Within modern agriculture's development, land use transitions and the replacement of landrace species with cultivars have shown the clearest negative consequences for genetic diversity [6–9]. Introducing modern cultivars can temporarily boost genetic diversity in agricultural systems, but over generations, there is a net decline in diversity [9].

The gradual domestication of crop progenitors has generated a genetic bottleneck. Selected traits that persisted through this bottleneck have supported the growth of human societies through agriculture. However, we are witnessing the genetic bottleneck of domestication advance dangerously further.

Threats to agricultural productivity and adaptability from reduced genetic diversity are not just a portending anxiety. Reduced dietary diversity worsens human health in devastatingly broad strokes: millions of lives are currently controlled by hidden hunger, obesity, starvation, and food insecurity. The limited knowledge of and access[1] to crop genetic resources obscure the full diversity of domesticated species and have largely led to the omission of wild, landrace, and CWRs from agricultural systems. Currently, the effective diversity within agriculture may be restored using the genetic diversity of major and non-major crops, and their CWRs.

# Advancing flax (*Linum usitatissimum*) breeding

## Existing understanding of flax

Genetic diversity serves as the natural capital of ecosystems to tolerate environmental change. In increasingly homogenous agricultural systems, safeguarding wild and domesticated plant diversity and engineering superior crop lineages largely depend on understanding the mechanisms and genetic bases for prized agronomic traits.

Already, domesticated species have been widely used to probe the mechanisms underpinning genotype-phenotype relationships. Wild progenitors, CWR, landrace varietals, and cultivated relatives have been valuable resources for isolating domestication trait loci, genes, and networks associated with abiotic and biotic stress responses and developmental and architectural changes in edible plant structure [11,12].

However, even in some of the earliest domesticated crops, much of the genetic variance responsible for changes in agricultural performance remains understudied. Flax (*Linum usitatissimum*) was among the first species cultivated in the Fertile Crescent approximately 8,000 years ago, alongside wheat and barley, as humans transitioned to agrarian societies.

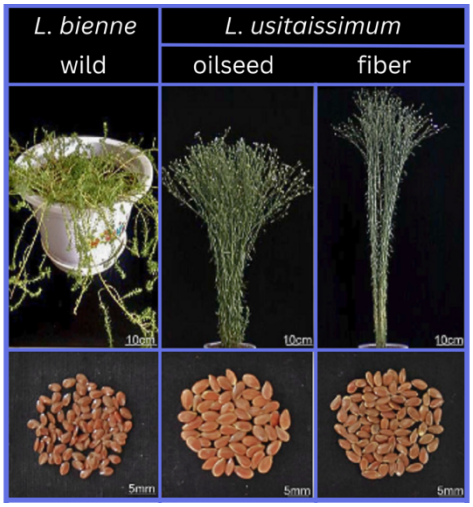



*Figure 1 Two distinct morphotypes can be seen in domesticated flax (Linum usitatissimum), for oilseed and fibrous varietals, relative to the wild progenitor (Linum bienne).*

Distinct morphotypes emerged from domesticating the wild progenitor, pale flax (*Linum bienne*) [13], supporting the historical and modern use of flax as a dual-function crop (**Figure 1**). With divergent fatty acid [14,15] and cell wall compositions [16], modern *L. usitatissimum* cultivars are applied as oilseed and stem fiber crops. Oil varieties have reduced heights and large, indehiscent seeds composed of ~40% oil [17]. Fiber varieties, relatively, have greater branching and shoot heights and yield fewer seeds.

[1] Many commercial crops rely heavily on private-sector breeding programs. Since the 1990s, rising mergers and acquisitions within the agricultural sector has introduced market constraints on agricultural genetic diversity. What used to be a race between the "Big Six" (BASF, Bayer, Dupont, Dow Chemicals, Monsanto, and Syngenta), is now a race of the "Big Three". Between 2015 and 2018, Bayer acquired Monsanto, DuPont Pioneer merged with Dow Agrosciences—creating Corteva—and ChemChina acquired Syngenta, in a series of sales totaling ~$236 billion. Eleven companies now execute 75% of global seed market sales [10], concentrating control over the distribution of seed genetic diversity into a handful of predominantly privately traded companies.

Yet, despite its agricultural importance since antiquity, the domestication of *Linum* cultivars from *L. bienne* remains speculative. Molecular evidence suggests that *L. usitatissimum* extended across Europe and North Africa from a single origin of domestication as an oilseed crop [17]. This conflicts, however, with early archeological evidence that portrays *L. usitatissimum* as a "highly useful" dual-function oilseed and fiber crop [18].

Early explorations of cultivated flax genetic diversity show major geographic divisions between Europe, West Africa, and South Asia [17]. However, the history of flax is confounded by the similar geographic distribution of the progenitor *L. bienne*, atop complications from the involvement of multiple domestication processes in *L. usitatissimum* [16].

During the 2000s and 2010s, short-read sequencing approaches rose in popularity and accessibility. Using small genetic variants, studies worked to decipher plant evolutionary processes, genetic trait mapping, and population structures at relatively low costs [19]. Resolving simple sequence repeats (SSRs)[2] and, later, single-nucleotide polymorphisms (SNPs) through short-read approaches—high and low-throughput—allowed small genetic variants to inform multiple descriptions of the genetic basis of flax's oilseed and fiber phenotypes.

Global assessments of *L. ustatitissimum*, using random amplified polymorphic DNA markers (RAPD) [17] and the association mapping (AM) of genetic markers to phenotypes [20], agreed that there is low genetic variance across cultivated flax. Curiously, intranational accession groups appeared relatively more diverse than international comparisons. RAPD variation also revealed the genetic base of oilseed flax to be broader than that of fiber flax [17,21]. From the core collection of flax germ cells or "germplasm", 407 distinct *L. usitatissimum* accessions [22] were later characterized with genomic regions controlling fiber and oilseed characters using 448 SSRs [20]. A subsequent AM study considered additional *L. usitatissimum* germplasm and SSR markers [23], characterizing additional polymorphism mutations within genes associated with cell wall, fatty acid, and polysaccharide biosynthesis. Aided by short-read sequencing advancements and gene-wide association studies (GWAS) [24,25], larger genomic regions continued to be linked to oil and fiber qualitative traits in flax. Largely, however, the genetic understanding of flax remains limited.

## Surpassing technological limitations to variation detection

Despite the phenotypic importance and numerous downstream applications of small genetic variants, the level of genome coverage in SSR and SNP studies incompletely detects present variance. Inherent restrictions within GWAS can also fail to describe the complete phenotypic variation attributable to genetic factors [26]. A single SNP represents a larger segment of genetic material, so samples can be misassigned as identical variants based on the SNP location. The impact of SNPs thus becomes diluted in GWAS studies when SNP variation is averaged across the larger genome segment and multiple samples [27]. Additionally, as the focus has shifted from SNPs to whole-genome sequencing, polygenic and complex traits have become increasingly associated with changes in larger segments of the genome architecture.

Pressing for more comprehensive accounts of genetic diversity, multi-genome analyses outperform GWAS. Advancements in genomic technologies offer an avenue to resolve missing genomic variants, particularly the resolution of large variants (< 50 base pairs (bp)). Both

[2] Microsatellite markers of one to six nucleotide-long motifs, repeated potentially up to 50 times.

reference genome creation and resequencing efforts in plant species have accelerated the study of large-scale changes in the organization of the genome, searching for structural variants (SVs) [28].

Despite initial assumptions that SVs were rare in plants, emerging evidence has determined that SVs are widely distributed across plant species [14,28–35]. Mounting crop studies associate SVs with adaptation, selection, and diverse agronomic phenotypes. Several classes of these SVs have been reviewed in *L. usitatissimum* through genome-level analyses: compared against a single reference genome, 216,863 gene presence/absence variation and copy number variants were found across the genomes of 100 fiber, dual-purpose, oilseed, and landrace flax accessions [32].

However, genome assembly methods introduce inherent limits to functional conclusions. Genome assembly is generally confounded by repeat sequences and polyploidization, which are common to plant species. Within genomic investigations, there are significant risks of reference biases and underestimations of variance, distorted by sampling and genome assembly methods. "Pan-genomes" are emerging at the forefront of genomes as a computational tool that mitigates estimation and reference genome biases. Pangenomes can also assess more diverse genome samples.

## Pangenomes to capture structural variation

Pangenome assemblies aggregate variation across input genomes to better represent genetic diversity. Within a graph data structure, pangenomes collapse identical genomic regions and expand variants (**Figure 2**). However, during pangenome assembly, researchers must contend with "haplotype-phased" or "haplotype-collapsed" genome inputs.

Groups of variable alleles or polymorphisms are often inherited together as "haplotypes", but their variation may be collapsed in genome assembly for computational efficiency. These haplotype-collapsed assemblies represent poorer resolutions of genetic diversity relative to haplotype-phased assemblies, which retain haplotype variation information. Phased genome assemblies ultimately aim to represent complete DNA sequences associated with each chromosome set, i.e., two haplotypes for both chromosome sets of a diploid individual.

Despite the emergence of the technology to generate haplotype-phased assemblies in the mid-2010s [36], few high-quality haplotype-phased assemblies exist [37]. Most haploid genome assemblies, even reference-quality genomes, are

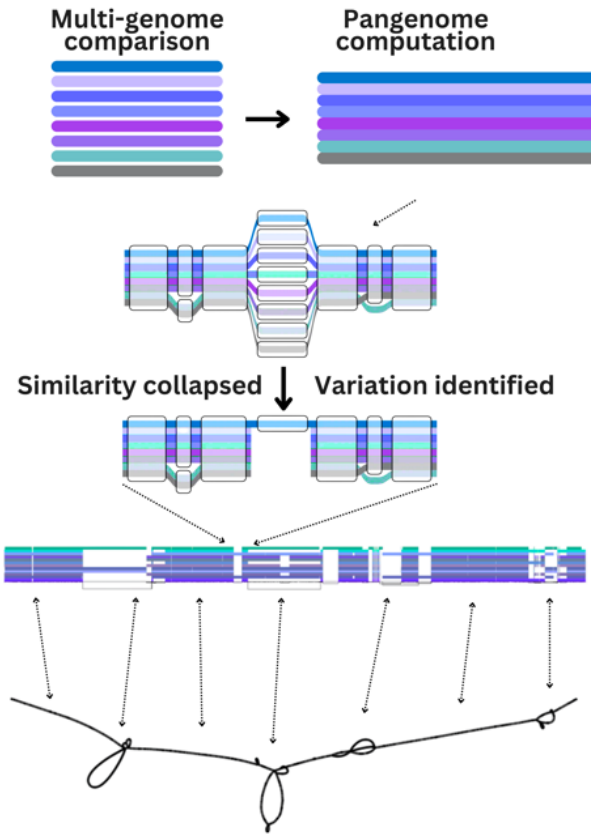



*Figure 2 Simple representation of the scale of genomic diversity resolved through a graph-based pangenome assembly.*

chimerized from multiple genome samples and haplotypes to produce more continuous, complete assemblies.

Just as genome assembly benefits from advancements in computational biology and genomic technology, so does pangenome assembly. As long-read sequencing and chromosome-level assemblies better capture haplotypic variation, pangenomes can efficiently incorporate these haplotypes into directed graphs. In computational biology, directed graphs[3] describe highly efficient data structures that allow tens to hundreds of haplotypes to be compounded together. All haplotype variants can be retained within the pangenome, offering a new lens on intraspecific diversity.

As the field of genomics shifts to favor pangenome references, interpopulation—even interspecific—diversity can be resolved through a computational transition from singular genome reference assemblies to pangenome assemblies [37–39].

A recent series of crop pangenome investigations [14,29,30,38] has captured species' genetic diversity and evolutionary history in greater detail than recoverable by single or incomplete genome assemblies. As a more comprehensive record of genomic variants, pangenomes may be leveraged for functional and evolutionary inferences. In staple cereals [14,29], annotating pangenome fragments with structure and function labels has already identified SVs that shape temperature tolerance, nutrient release, and disease resistance. Pangenomes are still used primarily to catalogue the large-scale genetic diversity of crop cultivars, landraces, and CWRs. However, characterizing new causal variants could innovate crops against future climate pressures and productivity demands.

## Surveying flax with pangenomics

Even with extensive germplasm collections, limited characterizations of genetic variance stunt flax breeding advancements. Few chromosome-level assemblies exist for *Linum* sp., and as a non-major crop, flax pangenomics has yet to be approached. However, the members of genus *Linum* possess diverse qualitative characteristics; the core *L. usitatissimum* germplasm collection alone organizes 79 character-state[4] phenotype combinations [22]. A holistic representation of flax diversity would benefit commercial parties wanting to innovate flax within agricultural, nutraceutical, textile, and alternative energy industries, and support scientific parties interested in conserving the modern and historical position of flax genetic diversity.

With little need for additional sequencing, *L. usitatissimum* stands to benefit from pangenome construction and the annotation of genome structures and gene functions within. As scientific institutions face funding uncertainty in the wake of political action [39], public data reusage in pangenome investigations also offers a low-cost research opportunity for laboratories with the computational resources to run pangenome software[5].

---

[3] Directed graphs are made of sets of vertices (more commonly "nodes" in bioinformatics), connected by a series of unidirectional arcs or edges, describing a one-way relationship between the graph network.

[4] A qualitative character with several states, i.e. the character of seed capsule dehiscence can present either a dehiscent or indehiscent state.

[5] Pangenome analyses may be done through High-Performance Computing (HPC) environments, by direct access often through Simple Linux Utility for Resource Management (SLURM) systems or by cloud-based platforms like Amazon Web Services, Google Cloud Platform.

For this project, public repositories were mined for several high-quality, haplotype-collapsed, haploid assemblies of *L. usitatissimum* and *L. bienne* (n = 15). While haplotype-collapsed assemblies fail to meet benchmarks for "gold-standard" pangenomes [42], due to variant information loss, haplotype-collapsed genomes can still provide advanced representations of diversity across many samples and remain relevant in discussing structural differences amongst crop varieties.

This report describes the improved genome assemblies of four *L. usitatissimum* cultivars and three *L. bienne* varieties, and their assimilation into a single pangenome graph. The constructed pangenome represents wild, oilseed, and fibrous flax populations across the Northern hemisphere, and the graph's structural and functional annotations offer insights into interpopulation diversity possibly missed in genomic investigations. Annotated SVs were associated with fundamental metabolic and architectural traits, including oil production traits that may become pertinent to unravelling the story of flax domestication and informing breeding strategies.

# Methods

## Genome assembly from public databases

### Genome accession download

The increasing availability of genomic data within public repositories provides strategies to reduce the cost of genomic investigations. This project used publicly available genome assemblies, which removed the immediate need to generate new sequence material to investigate flax genetic diversity. The five starting genomes included in the annotated pangenome are directly downloadable from NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/). Accessions were selected based on sequence completeness and/or geographic origin, such that the pangenome varietals (**Table 1**) represent flax populations spanning the Northern Hemisphere. The flax varietals also serve diverse agricultural functions as oilseed, fiberstem, or dual-function crops. For instance, *L. usitatissimum* var Longya and Heiya are respectively bred as oil-use and fiber-use cultivars [16]. For easier assembly, all selected species share chromosome-level organization (2n = 30), but were reported in databases as haploid assemblies (n = 15).

Aiming to input publicly available sequences into the pangenome, all downloaded genomes are archived under NCBI Bioprojects (**Table 1**). Amongst the five genomes, *L. usitatissimum* CDC Bethune and *L. bienne* isolate 15003 are the current NCBI-recognized reference genomes for their respective species.

*Table 1 NCBI-downloaded genomes are listed with the varietal name, the most recently published assembly with the publication citation, NCBI Bioproject number, and the level of genome assembly reported by NCBI. Within the full varietal name, the abbreviated name is bolded. Genomes were downloaded on Nov 21, 2024.*

| ***Linum species*** | ***Accession [Publication]*** | ***Bioproject*** | ***Origin*** | ***Level of assembly*** |
|---|---|---|---|---|
| *L. usitatissimum* var **CDC Bethune** | GCA_000224295.2 [40] | PRJNA68161 | Canada | **Chromosome** |
| *L. usitatissimum* var **Longya**-10 | GCA_010665275.2 [16] | PRJNA505721 | China | **Chromosome** |
| ***L. bienne*** isolate 15003 | GCA_010665285.1 [16] | PRJNA449140 | S Europe/ NW Africa | **Scaffold** |
| *L. usitatissimum* var **Heiya**-14 | GCA_010665265.1 [16] | PRJNA449140 | China | **Scaffold** |
| *L. usitatissimum* var **Atlant** (line 3896) | GCA_030674075.2 [41] | PRJNA648016 | Russia | **Contig** |

### Note on haploid assembly inputs

Trends in genome and pangenome assembly currently aim to phase haplotypes to better index genomic variation through haplotype and allele-resolved assemblies. However, despite the significance of haplotype information in pangenome construction, all downloaded genomes (**Table 1**) are haplotype-collapsed.

Original input and intermediate files involved in creating haploid assemblies may be used to reconstruct haplotypes. Still, the majority of genome assemblers and pipeline software responsible for the published genome assemblies—**SOAPdenovo** [42], **BioNano** Assembler [43], **Allpaths-LG** [44]—do not produce haplotype-resolved outputs.

Therefore, although some software in our methodology offers haplotype-sensitive options, these features were not incorporated into the improved genome and pangenome assemblies. For instance, **HapHiC** [45] provides the option of haplotype-aware contig assignment. However, all input assemblies (**Table 1**) are haplotype-collapsed, so this study did not exploit this **HapHiC** feature nor the haplotype-aware features of the PanGenome Graph Builder (**PGGB**) [46].

## Starting assembly statistics

The downloaded genomes exhibit variable fragment numbers and sizes, constituting contig-level, scaffold-level, and chromosome-level genome assemblies. From raw sequence inputs, continuous sequences, contigs, can be pieced together from matching overlapping fragments from raw sequence inputs. Scaffolds are often considered at a higher assembly level than contigs, as scaffolds orient and orderly combine contigs based on additional sequence information, such as a reference genome.

The published "starting" assembly statistics are summarized in **Table 2**. While many metrics are easy to understand, sequencing depth, N50, and L50 statistics may require brief explanation. Sequencing depth describes the number of times a nucleotide was read during sequencing, where increased depth generally increases genome sequencing confidence. If the fragments within an assembly were ordered by length, N50 is the contig length at which 50% of the total assembly

length is reached, while L50 is the number of longest contigs needed to reach that threshold. Likewise, N90 and L90 values apply when 90% of the assembly length is reached.

*Table 2 Assembly statistics for the downloaded (starting) genomes. Where there was a conflict between the assembly statistics reported in the source publication and the NCBI database report, the NCBI statistics were recorded below. Bolded numbers correspond to the NCBI-reported level of assembly.*

| ***Name*** | *Genome size (Mb)* | *Ungapped size (Mb)* | *Sequencing depth* | *GC content (%)* |
|---|---|---|---|---|
| **CDC Bethune** | 316.2 | 269.9 | 94x | 39.5 |
| **Longya** | 306.4 | 300.2 | 47.6x | 39.0 |
| ***L. bienne*** | 293.6 | 287.9 | 92.9x | 39.0 |
| **Heiya** | 303.7 | 300.8 | 141x | 39.0 |
| **Atlant** | 434.2 | 434.2 | 92.9x | 38.5 |

*Table 2 Continued*

| ***Name*** | ***Reported assembly statistics*** | | | | |
|---|---|---|---|---|---|
| | *Chromosome number* | *Scaffold number* | *Contig number* | *N50* | *L50* |
| **CDC Bethune** | **15** | 15 | - | 20.5 Mb | 7 |
| **Longya** | **15** | 1608 | - | 18.2 Mb | 8 |
| ***L. bienne*** | - | **2654** | 6369 | 383.8 kb | 201 |
| **Heiya** | - | **2772** | 4581 | 699.9 kb | 130 |
| **Atlant** | - | - | **1516** | 6.2 Mb | 25 |

## Starting genome assembly and quality control

### *L. usitatissimum* var CDC Bethune

The current chromosome-level reference genome for *L. usitatissimum* var CDC Bethune (GCA_000224295.2 [40]) was produced from the previous, scaffold-level assembly of (GCA_000224295.1 [47]).

GCA_000224295.1 scaffold sequences were extracted from flax seeds of the original breeder (Prof. em. Gordon Rowland, University of Saskatchewan, Canada); DNA extractions were confirmed by flow cytometry (Accuri C6) before short shotgun sequencing. In creating the more recent genome assembly, Wang et al. [47] filtered the previous assembly's raw Illumina reads that contained > 2% unknown nucleotide bases or > 10 bp of suspected adapter contaminants. Filtered reads were *de novo* assembled with **SOAPdenovo**.

Published GCA_000224295.1 reads [47] were additionally filtered for contaminants against the **UniVec database** (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/). This cleaning step used the tuning options **(-task blastn -reward 1 -penalty -5 -gapopen 3 -gap extend 3 -dust yes -soft_masking true -value 700 -search 1750000000000)** to remove > 15 kb of sequences.

Filtered scaffolds were then guided and corrected against a **BioNano** genome optical map [40], a high-resolution map of DNA sequences. The final, consensus optical map was generated only with single-molecule map contigs that carried p-values < 5e-8, < 5e-9, and < 5e-9 during the respective pairwise assembly, extension/refinement, and final refinement stages of the **BioNano** Assembler pipeline. The resulting consensus map was then checked and treated for chimeric contigs, and redundant (≥ 90% alignment similarity) and low-confidence alignments were removed.

The corrected scaffolds and 251 **BioNano** optical map contigs (317 Mb in length) were then "superscaffolded" against the physical map of flax produced by Ragupathy et al. [48] That physical map was constructed using large DNA fragments cloned into bacterial artificial chromosomes [48]. The use of physical maps improved contig orientation and continuity to the chromosome-level genetic map of *L. usitatissimum* var CDC Bethune (GCA_000224295.2, **Table 1**).

### *L. usitatissimum* var Atlant

The Russian *L. usitatissimum* var Atlant (GCA_030674075.2) is based on the previous Oxford Nanopore long-read-based assembly of Atlant (GCA_030674075.1 [49]). Dvoriainova et al. [41] improved the existing assembly with additional high-molecular-weight DNA and Oxford Nanopore sequencing, raising the assembly level beyond the previous contig-level assembly [49]. Adapter sequences were removed after sequencing using **Porechop**[6] (https://github.com/rrwick/Porechop).

Filtered Atlant reads (GCA_030674075.2) underwent genome assembly without adapters using **Canu v2.2 (genomeSize = 400 m --nano-raw)** [51] before three rounds of polishing. The first two polishing rounds used **Racon v1.4.10 (-m 8 -x -6 -g -8 -w 500)** [52], which uses partial order alignment graphs to refine assemblies. For comparison, multiple sequence alignment software also use alignment graph structures to represent the possible overlaps and junctions between sequence fragments [53]. A final round of polishing used **Medaka v1.4.10 (-m r941_min_sup_g507)** (https://github.com/nanoporetech/medaka). Using neural networks, **Medaka** identifies consensus and variable sequences to correct mismatches and insertion-deletion errors from Oxford Nanopore reads.

Following **Canu** assembly and **Racon** (x2)-**Medaka** polishing, Dvoriainova et al. [41] assessed the output assembly using **QUality ASsessment Tool (QUAST) v5.0.2** [54] assembly parameters (**Table 2**) and **Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.2** [55] completeness scores (**Table 3**) with the defined lineage **(-l eudicots_odb10 )**.

**QUAST** offers benchmarking standards for genome assembly quality, with or without a reference; **BUSCO** estimates the completeness of an assembly based on near-universal single-copy orthologs from a specified evolutionary lineage and classifies the single-copy orthologs as "single", "duplicated", "fragmented", or "missing". Single and duplicated BUSCO genes may be collectively referred to as "complete" BUSCOs.

[6] **Porechop** is no longer supported software (as of 2018), but **Porechop_ABI** [50] is reported as an extension of **Porechop**.

*Table 3 Reported assembly statistics at the end of published genome assembly pipelines (i.e., statistics for "Atlant" refers to the assembly following Canu, Racon (x2), and Medaka). BUSCO and CEGMA assessment results are reported where applicable.*

| **Assembly** | ***BUSCO*** | | | | ***CEGMA*** | |
|---|---|---|---|---|---|---|
| | *Single* | *Duplicated* | *Fragmented* | *Missing* | *CEGs (% of 458)* | *Highly conserved CEGs (% of 248)* |
| **Longya** | 35.42 | 56.11 | 1.88 | 6.60 | 99.13 | 97.98 |
| ***L. bienne*** | 42.0 | 47.64 | 2.29 | 7.99 | 98.69 | 98.79 |
| **Heiya** | 34.65 | 56.18 | 2.29 | 6.88 | 98.91 | 97.98 |
| **Atlant** | 31.60 | 62.20 | 0.70 | 5.50 | - | - |

Published **QUAST** statistics of *L. usitatissimum* var Atlant [41] used an earlier assembly of *L. usitatissimum* var CDC Bethune (GCA_000224295.1 [47]) as the reference genome. Our own **QUAST v5.3.0** [56] analyses, however, used the more recent *L. usitatissimum* var CDC Bethune assembly (GCA_000224295.2).

### *L. usitatissimum* var Longya, Heiya, and *L. bienne*

The starting assemblies of *L. usitatissimum* var Longya, Heiya, and *L. bienne* were constructed using Illumina HiSeq2500 paired-end sequences with 133x, 142x, and 93x sequencing depths, respectively. Sequences across the three *de novo* assemblies were filtered for low-quality, contaminated, and adaptor reads, but exact quality control methods were not specified [16].

Downstream, sequences were put through an assembly pipeline combining **Allpaths-LG,** with **SSPACE v2.3** [57] and **SOAPdenovo2 GapCloser** [58] software. The genome assembly of Longya was further scaffolded with paired-end high-throughput chromatin conformation capture (HiC) reads using **BWA** [59] and **LACHESIS** [60]. HiC sequencing is explained in *HiC scaffolding and the Dovetail Genomics HiRise Assembly*. No software tuning options (beyond default parameters) or quality control steps were detailed for the genome assembly pipeline or the HiC scaffolding.

Genome assembly quality was assessed by sequencing error rates (or single-nucleotide error rates) in the **BWA** assembly of *L. usitatissimum* var Longya, and **Core Eukaryotic Genes Mapping Approach (CEGMA) v2.5** [61] and **BUSCO v3.0.2** [62] assessments were run for all three assemblies (**Table 3**). Zhang et al. do not specify the BUSCO lineage. However, their analysis assessed 1440 BUSCOs [16], so the lineage is suspected to be the embryophyta_odb9 dataset. The presence of BUSCOs and CEGs in **Table 3** is reported as percentages of the total surveyed genes: 2326 BUSCOs (eudicots_odb10) were surveyed for *L. usitatissimum* var Atlant, and 1440 BUSCOs were surveyed for others (likely embryophyte_odb9).

## Assembly of the *L. bienne* genome

The majority of genomic data in this study is from previously published material (**Table 1**). However, there was a financial opportunity to construct a superior chromosome-level assembly of *L. bienne* using previously unpublished HiC sequences as scaffolding material. In 2020, HiC

sequences for two wild varieties of *L. bienne* were produced by the work of Dr. Adrian Brennan (supervisor) and the commercial services of Dovetail Genomics™ (now under Cantana Bio, LLC). Those sequences were factored into the starting assembly of *L. bienne* [16].

## Wild plant material

Samples of *L. bienne* were collected from the following locations: Spanish wild flax samples (var 1_6) were collected at (36.80044, -5.39258) at 624m altitude in southern Andalusia by Rocio Perez-Barrales in 2016; Israeli[7] wild flax samples (var Isr) were collected at (33.24028, 35.75056) at 1,066m altitude in the northern region of the Golan Heights before 2019 (exact date unknown). Initial stages of Dovetail HiC (officially Omni-C™) sequencing require live nuclei, so live shoots of *L. bienne* vars 1_6 and Isr were preserved for Dovetail HiC sequencing and HiC scaffolded (**HiRise**) assembly.

## HiC read generation

### HiC scaffolding and the Dovetail Genomics™ HiRise assembly

HiC sequencing was initially devised to probe chromatin dynamics surrounding genome conformations [64]. However, improved investigations of chromatin state changes have refined inferences of long-range chromatin interactions to reconstruct genomes' three-dimensional structures. More detailed reviews discuss the principles of HiC library sequencing and assembly, as applied in the proprietary Dovetail Genomics pipeline [65–67].

Presently, HiC sequences can be applied to improve contig orientation and scaffolding to chromosomes, generating more complete genome assemblies [68]. *L. bienne* var 1_6 and Isr samples were initially used to generate proprietary Chicago and HiC library reads for **HiRise** scaffolding, Dovetail Genomics' proprietary pipeline. **HiRise** sequences were scaffolded against the existing *L. bienne* assembly [16].

The output **HiRise** assemblies are classifiable as scaffold-level assemblies, with 1,005 scaffolds for *L. bienne* var 1_6 and 961 scaffolds for *L. bienne* var Isr (**Table A1**). However, due to poor assembly metrics (**Table A1**) and reduced BUSCO quality scores (**Table A2**), the **HiRise** assembly was discarded to pursue a different scaffolding option: **HapHiC** [45].

## HiC sequence cleaning

Before **HapHiC** HiC scaffolding, input *L. bienne* var 1_6 and Isr were cleaned based on Dovetail HiC FastQC reports (not shown). The non-random distribution of nucleotide and GC content at the terminus of forward and reverse read pairs (< 5 bp) suggested minor adaptor contamination during library preparation. Potential contaminants were cleaned from the *L. bienne* var 1_6 and Isr HiC sequences by **Trimmomatic** [69] **(HEADCROP:5)**. The **Trimmomatic** paired-end (PE) read trimming command accounts for the HiC forward and reverse read pairs.

[7] The United States is the only nation to recognize the Israeli annexation of the Golan Heights, in violation of international law and treaties that accept the region as Syrian [63].

After cleaning, 1,290 and 1,131 input HiC reads for *L. bienne* var 1_6 and Isr, respectively, were concatenated into a larger scaffold dataset to improve the existing scaffold-level *L. bienne* genome assembly (GCA_010665285.1, **Table 1**).

## HiC scaffolding using HapHiC

In the absence of a chromosome-level reference genome for *L. bienne*, **HapHiC v1.0.6** [45] was selected as a reference-free HiC scaffolding software. Compared to the popular HiC scaffolding software alternative, **AllHiC** [70], **HapHiC** captures high-throughput HiC information with fewer contig misassignments and higher scaffold continuity, up to chromosome-level scaffold assignments [45]. Chimeric contigs are also tolerated or misjoin-corrected by **HapHiC**, which suits the HiC input of collated, multi-varietal *L. bienne* reads.

As part of its scaffolding pipeline, **HapHiC** indexed the existing *L. bienne* assembly scaffolds (GCA_010665285.1), softmasked variation within the assembly, and aligned the unpublished HiC reads of *L. bienne* var 1_6 and Isr to the softmasked assembly. In a softmasked assembly, genetic variation between samples is not overwritten, like in hardmasked assemblies, but merely flagged using lowercase lettering.

Output **HapHiC** HiC alignments were manually filtered based on assignment confidence, as 25.4% of reads carried MAPping Quality (MAPQ) values of 0. MAPQ values vary between programs (e.g., **BWA-MEM** [59] or **Bowtie 2** [71]), but MAPQ values represent the logarithmic probability of correctly placing a read. While the upper bounds of MAPQ scores are variable, a score of zero strongly suggests incorrect placement.

Following the recommended **HapHiC** pipeline parameters, only reads with MAPQ $\geq 1$ were kept for HiC scaffolding. After filtering 7.2 million reads, over 210 million reads remained as scaffold inputs. However, contrary to **HapHiC** recommendations, reads were not filtered based on read edit distance (NW), a metric for detecting sequence similarity. This decision was led by technical difficulties while scripting sequence filtering. Standard procedure filters NW < 3 to remove similar reads where fewer than three single-nucleotide edits would be required to convert one string into another.

**HapHiC** then scaffolded the MAPQ-filtered HiC reads of *L. bienne* var 1_6 and Isr against the starting, scaffold-level assembly of *L. bienne* var isolate 15003 (GCA_010665285.1 [16]). With the chromosome number **(nchrs)** defined as 15, **HapHiC** generated a chimeric genome assembly for *L. bienne*. The final scaffolded assembly contained 2,763 sequences after undergoing two rounds of corrections **(--correct_nrounds 2)**. The uncorrected assembly contained 2,159 uncorrected scaffold sequences, but the differences between the assemblies were minute. The **HapHiC** assembly reads became the inputs for the final genome assembly of *L. bienne* during further scaffolding with **RagTag** [72].

### HiC-scaffolded assembly visualization

The **HapHiC plot** command with the options **(--origin top_left --border_style outline --outline_width 0.5)** was used to visualize a HiC interaction heatmap. HiC interactions are based on the alignment of post-**HapHiC** scaffolds and smaller reads of *L. bienne*.

500 kb (default) bin sizes were used, and only scaffolds > 1 Mb long were mapped across HiC interaction heatmaps. Nine well-defined chromosomes are identified through **HapHiC plot**, but ten, smaller scaffolds are also demarcated in blue. To improve small scaffold visualization, the **HapHiC plot** option **(--separate_plots)** visualized scaffolds < 1 Mb.

**HapHiC** HiC assembly reads were generated in the **YaSH** [38] format, which also allows for manual HiC heatmap curation using **Juicebox** [39] (**Figure A1**). Mapped reads were not limited to scaffolds > 1 Mb in the **Juicebox** viewer, although **Figure A1** only covers ~200 Mb of HiC interactions to condense the visualization. In our study, **HapHiC plot** outperformed manual curation in **Juicebox**.

## Final assembly of *L. usitatissimum* cultivars and *L. bienne* genomes

The assembly pipelines of *L. usitatissimum* var Atlant, Longya, Heiya, and *L. bienne* during our study, for pangenome creation, are described below. For all assemblies of this study, *L. usitatissimum* var CDC Bethune (GCA_000224295.2 [40]) acts as the chromosome-level reference genome. The reference genome required no further assembly and underwent no additional editing or quality control.

### RagTag assembly

Assemblies of *L. usitatissimum* var Atlant, Longya, Heiya, and *L. bienne* used the **scaffold** command of the reference genome-guided (homology-based) software **RagTag v2.1.0** [72]. **RagTag** has been used previously in multi-genome and pangenome projects, with similar aims to investigate structural variation (SV) and trait landscapes of crop species. While creating haplotype-sensitive pangenomes for the tea plant (*Camellia sinensis*) [75], and grapevine varieties (*Vitis* spp.) [34], **RagTag** respectively handled the *de novo* genome assembly of 22 and 18 accessions from each plant.

The starting assemblies of *L. usitatissimum* var Atlant, Longya, Heiya (**Table 1**), and the **HapHiC**-scaffolded assembly of *L. bienne* were subject to a single round of **RagTag** scaffolding under default parameters against the reference *L. usitatissimum* var CDC Bethune (**Table 1**).

In a separate script, the **(-C)** option was applied to amalgamate less poorly aligned sequences within an arbitrary "Chromosome 0" file **(Table A3)**. The resultant assemblies presented chromosome-level scaffolds (n = 15). However, these were misrepresentative. Chromosome 0s for each assembly contained sizable proportions of the scaffolded genome, thereby removing regions of higher divergence, which are pertinent to variation analysis.

The length of Chromosome 0 strongly correlates with the starting level of assembly (**Table 1**), but in the most extreme case, Chromosome 0 comprised 24% of the total genome length (Atlant, **Table A3**). Reasons for poor alignment could arise from contamination, haplotypic and allelic variation, or sequence redundancy. The latter is both likely and valuable for repetitive, largely unannotated plant genomes. It was concluded that executing **RagTag scaffold** without **(-C)** produced more complete assemblies.

## Post-assembly genome quality assessments

Following **RagTag scaffold**, the assembly quality and completeness of the four final genomes were each assessed by **QUAST** and **BUSCO v5.3.2** [55] software. **QUAST** evaluation of assembly quality applied the **(--eukaryote –large)** options, and generated post-assembly k-mer statistics using the **(--k-mer-stats --k-mer-size 120)** option.

Typical k-mer reports are in the context of distribution analysis: k-mers are spread across raw, often short-read sequences, and k-mer read coverage is used to estimate genome size. In contrast, **QUAST --k-mer-stats** operated on the output assembly scaffolds of **RagTag**, and describes misassembly patterns and relative divergence when compared between assemblies. “Regular” or “k-mer-based” prepositions are included where relevant in the discussions of **QUAST** misassembly metrics.

The reported k-mer-based misjoins account for k-mer-based relocations and translocations. At different contig coordinates (relocations) or potentially on different chromosomes (translocations), contig-level k-mer patterns may be found at separate locations of the reference and sampled genomes.

The final genome assemblies were also subject to **BUSCO** genome assessments for the eudicots lineage **(-l eudicots_odb10 -m geno)**.

## Pangenome assembly

### Inputs of PanGenome Graph Builder (PGGB)

Generating a pangenome from haplotype-collapsed genomes sacrifices greater variation than if input genomes are *de novo* assembled from haplotype-aware raw sequences. Additionally, genome scaffolding by homology-based methods, while beneficial to increasing assembly continuity, may also mask intervarietal genetic diversity. The choice of pangenome software aimed to represent maximum genetic diversity from haplotype-collapsed assemblies while reducing reference-based genetic diversity loss.

**PGGB v0.7.2** [49] was used to produce a “general-purpose” pangenome graph [37] from *L. usitatissimum* var CDC Bethune, Atlant, Heiya, Longya, and *L. bienne* as well as make pangenomic SV calls. SV calls describe locations where an SV was detected. As an “all-to-all” graph constructor, **PGGB** is preferable to the iterative construction methods used in pipelines like **Minigraph-Cactus** [77]. Therefore, **PBBG** could mitigate variation loss from reference bias introduced during the genome or pangenome stages.

The input sequence names of the starting assembly of *L. usitatissimum* var CDC Bethune and the post-**RagTag** output assemblies for *L. usitatissimum* var Atlant, Heiya, Longya, and *L. bienne* were reformatted under **PanSN** conventions (https://github.com/pangenome/PanSN-spec). This was per the recommendations of the **PGGB** pipeline. Renaming sequences also supports a standard pangenome naming convention, which is helpful in downstream data handling across programs. Assembly fragments were renamed as follows: Varietal_name#haplotype_ID#contig_number (e.g. Bethune#1#contig1) (**Script A1**). For all

inputs, the haplotype_ID was set to one; PGGB accepts haplotype-collapsed assemblies. Scaffolds and contigs were named equally as "contig", followed by the appropriate fragment number for that varietal assembly.

No masking operation was performed on the input genomes, as repetitive variants will be subject to smoothing by the final graph sorting stages of the pangenome pipeline. In the competitor pangenome graph construction software, **Minigraph-Cactus**, softmasked inputs are ignored, and masking is explicitly not advised (from **v2.1.0** onwards).

## Algorithms inside PGGB pangenome creation

All-to-all alignment directs input sequences to one another using the pairwise alignment algorithm, **wfmash** [78]; **wfmash** enables any sequence to serve as a reference. **PGGB**'s all-to-all approach outperforms linear reference-based approaches[8] [79] and remains inclusive of SV even during later graph construction and refinement stages.

The **seqwish** [80] algorithm then transforms the collection of pairwise alignments into a tree data structure. **seqwish** creates the pangenome graph by directing tree node and edge information. Unlike the construction of phylogenetic trees, where edges are undirected and represent hierarchical information, the **seqwish** graph instead represents *uni*directional (directed) SV paths [80].

An advantage to **seqwish** is that variants of all lengths are considered in the directed, acyclic **PGGB** variation graph [37]. From the trees of aligned sequences, pangenome variation graphs are constructable without input loss and at scale [80]. Algorithms **GFAfix** (https://github.com/marschall-lab/GFAffix) and **smoothxg** (https://github.com/pangenome/smoothxg) refine the variant graph (graph sorting) to yield a "smoothed", nonredundant graph structure.

Variation graphs are often modeled acyclically, as in **PGGB**, which avoids the creation of cyclic SV "loops" within the graph. Despite some loss of complex structural variation (e.g., translocations or recombination loops) [37], different SV classes and lengths remain well-represented in the acyclic graph, and alternative variant paths are clearly defined [76]. Data structures proposed for pangenome analysis are described further in [81]. Where applicable, these paths may also be haplotype-associated.

## Execution of PGGB

**PGGB** was run with **(--n-haplotypes 1 --map-pct-id 70 --segment-length 5000)**. The pairwise identity threshold **(--map-pct-id)**, which describes the average sequence similarity between aligned sequences as a percent, was set notably lower than the typical 90. As input assemblies are not all at chromosome-levels of assembly, every genome underwent a reference step against *L. usitatissimum* var CDC Bethune during genome assembly. In the face of possible reference

[8] A "linear" reference structure is a simpler coordinate system, describing linear tracks along chromosomes or sequences. Structures of a pangenome, particularly complex structural variants, may be restricted by the linear coordinate system.

bias from that scaffolding step, the pairwise identity threshold was lowered to preserve the remaining variation amongst varietals during alignment.

**PGGB** is currently the only pangenome tool that can generate variation graphs without a reference [37], but SV calling requires the definition of at least one reference sequence. The **(--vcf-spec)** option was set to generate a variant call format file (VCF) for resulting SV calls and requires the definition of FASTA-type reference sequence(s). An entire genome cannot be used as a reference unless the chromosome-level scaffolds are concatenated. Therefore, all 15 chromosome sequences of *L. usitatissimum* var CDC Bethune were set as references for SV calls.

Several notes on **PGGB** execution are: first, by default, **PGGB** generates a graphical fragment assembly (GFA) output containing variable and shared sequence information. GFAs are transferable to other pangenome programs, such as the VG **toolkit** (https://github.com/vgteam/vg) and **ODGI** [82]. Second, while most software programs of this project have been run in a high-performance computing (HPC) environment, all-to-all pangenome assembly is particularly computationally expensive. For **PGGB** to be executed in a reasonable time, 64 cores needed to be reserved for a run generating a single GFA alongside default outputs and **(--vcf-spec)**.

# Structural Variation (SV) calling

## Genome-level SV calling

At the genome level, SVs were called using **SVIM-asm v1.0.3** [83]. **SVIM-asm** is one of the few SV callers tailored for genome assemblies, alongside **Smartie-sv** (https://github.com/zeeev/smartie-sv) and **SyRI** [84]. However, despite the high assembly of our final genomes, **SyRI** strictly operates on chromosome-level genome assemblies and, thus, could not generate any SV calls from our final genome assemblies.

**SVIM-asm** first requires indexed read alignments, so our haploid **RagTag** assemblies were processed by **minimap2 v2.24 (-a -x asm5 --cs -r2k)** [85] and **samtools v1.21 sort** and **index** [86]. Indexing assigns unique IDs to organize fragments within multi-sequence files. **SVIM-asm** was executed by the **(svim-asm haploid)** command, with the default minimum MAPQ value of 20, minimum SV length of 40 bp, and maximum SV length of 100 kb. A custom script using base Linux (not provided) was applied to filter SVs < 50 bp. The **SVIM-asm** VCF outputs are organized by variety in **Table 4**.

*Table 4 Genome-level SV calls ( > 50 bp) by SVIM-asm, organized by SV type (deletions, insertions, inversions, tandem duplications, and breakend points). SVIM-asm reports tandem insertions, but the count across varietals was zero and was thus omitted.*

| ***Genome assembly*** | Longya | Bienne | Heiya | Atlant |
|---|---|---|---|---|
| **Total SV calls** | 20,583 | 24,172 | 20,361 | 21,122 |
| **Deletions** | 9,180 | 9,973 | 8,927 | 9,419 |
| **Insertions** | 11,281 | 11,158 | 11,308 | 11,481 |
| **Inversions** | 11 | 15 | 12 | 10 |
| **Tandem duplications** | 19 | 29 | 12 | 80 |
| **Breakend points** | 92 | 3,002 | 102 | 132 |

Other SV callers designed for long or short-read alignments, such as **Sniffles** [87], **Manta** [88], or **Delly**[56], can be executed on completed genome assemblies. However, these programs were not included in the SV dataset, as most alignment-level SV callers are unfit for assembly-level SV detection.

Generally, alignment-level SV callers' outputs diverge drastically from assembly-level callers. Using an SV caller designed for sequence alignments risks misrepresenting genomic variation. For example, **Sniffles v2.5.2** [87] called a total of 2,044 SVs across the four varietals (**Table A4**). The assembly-based SV caller, **SVIM-asm**, meanwhile, detected 42 times the number of calls under the same > 50 bp length restriction (**Table 4**).

## Pangenome-level SV calling

SV calls were generated at the pangenome level by **PGGB (--vcf-spec)** against every chromosome of *L. usitatissimum* var CDC Bethune. Referencing SVs to each chromosome of the reference genome assembly is akin to the all-to-all assembly used previously during pangenome construction. The **PGGB (--vcf-spec)** operation requires a sequence or multiple sequences to be specified, so a single reference—artificially concatenated from all chromosomes—may also have been suitable if not for the computational resources demanded.

From the pangenome SV call, 15 VCF files containing variants of all lengths were output. Each chromosome contained an average of 6.4 million calls, including single-nucleotide polymorphisms (SNPs). However, to present only large SVs, the VCF outputs of **PGGB** were filtered only to contain SVs > 50 bp in length (**Table 5**).

*Table 5 Pangenome-level SV calls, as referenced to CDC Bethune by PGGB.*

| ***Chromosome*** | SV calls |
|---|---|
| **1** | 13951 |
| **2** | 11723 |
| **3** | 11639 |
| **4** | 9467 |
| **5** | 8366 |
| **6** | 8015 |
| **7** | 8266 |
| **8** | 9743 |
| **9** | 9878 |
| **10** | 8188 |
| **11** | 10645 |
| **12** | 9432 |
| **13** | 9986 |
| **14** | 10418 |
| **15** | 7573 |

## SV cross-dataset validation

**PGGB** was executed with *L. usitatissimum* var CDC Bethune chromosomes as references, and **SVIM-asm** used genome assemblies scaffolded against *L. usitatissimum* var CDC Bethune. Therefore, all detected SVs share coordinate positioning relative to the reference flax genome. To finalize the SV dataset for functional analysis and annotation using reference-based positions, only SV calls detected by both **PGGB** and **SVIM-asm** were retained.

The 147,290 SVs called, in total, by **PGGB**, were cross-referenced against the 86,238 SVs called by the four genome-level **SVIM-asm** inquiries. SVs were considered to be equivalent calls based on two concurrent positions:

1. **Chromosomal alignments** were in agreement. For instance, an SV assigned to *L. usitatissimum* var CDC Bethune chromosome 3 by both programs would pass.
2. **Starting nucleotide position** of the SV are defined within 30 bp of one another. A shift of 30 bp in either direction was tolerated, determined after manually assessing the positions and represented variant sequences for SV calls across both programs.

Of the starting SVs called by **PGGB**, 57,729 variants were validated by the **SVIM-asm** VCFs using a custom R script (**Script A2**). Therefore, every SV in the final dataset is defined within the pangenome graph and contextualizable within one or more varietals: *L. usitatissimum* var Atlant, Heiya, Longya, and/or *L. bienne*.

## SV Analyses

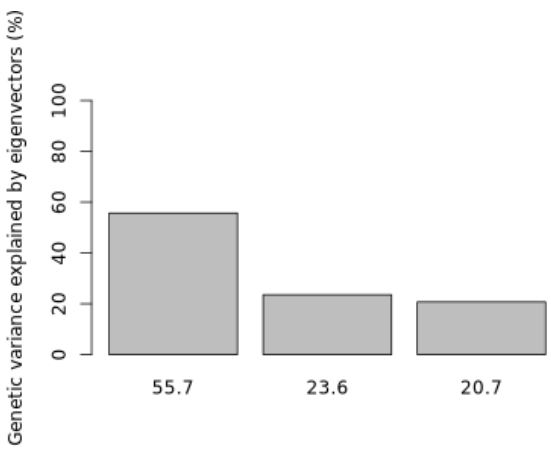
### Pangenome SVs

To represent the additional pangenome length provided by SVs alone, a custom R script was used to estimate the pangenome length from insertions and inversions (**Script A3**). Pangenome length estimations are based on all pairwise genome comparisons (from single genomes, and double, triple, and quadruple/all genome comparisons).

To investigate SV length and distribution along the pangenome, insertions, inversions, and deletions were considered. As breakend points consist of only one positional coordinate, opposed to a start and end coordinate, and represent the placement of an SV (not a unique type of SV, themselves), they were omitted during SV distribution mapping. Undefined SVs were also removed from the final dataset of 57,729 cross-validated SVs. Due to conflicting location coordinate formats between **SVIM-asm** and **PGGB**, tandem duplications' lengths could not be confirmed, and are removed from the discussion until gene ontology analyses. Statistical tests for SV length and distribution patterns were run in base R (**Script A2**).

### PCA Plot

To interpret the intervarietal diversity of SVs between *L. usitatissimum* var Atlant, Heiya, Longya, and *L. bienne*, a covariance-matrix-based Principal Component Analysis (PCA) was conducted using a custom R script (**Script A4**).




A scree plot was produced using **adegenet** [89]. For the filtered SV calls, the first two Eigenvalues represent 79.3% of the variance in our SV dataset (**Figure 3**).

*Figure 3 Scree plot for the variance of SVs across L. usitatissimum var Atlant, Heiya, Longya, and L. bienne within the pangenome.*

# Functional Annotation

## Annotating SVs of the pangenome

Annotation data for *L. usitatissimum* was publicly available from **JGI Phytozome v13** (https://phytozome-next.jgi.doe.gov/). The annotation data (Lusitatissimum_200_v1.0.transcript_primaryTranscriptOnly.fa) was published in 2013 for the transcript of the first genome assembly of *L. usitatissimum* var CDC Bethune (GCA_000224295.1).

The annotated transcript was aligned using **minimap2 (-ax splice)** to our reference assembly GCA_000224295.2. This alignment ensured the older annotations corresponded to the more recent assembly sequences of *L. usitatissimum* var CDC Bethune used during SV calling.

As the pangenome retained chromosome-level coordinate labels corresponding to *L. usitatissimum* var CDC Bethune, annotations from the aligned transcript could be assigned to the locations of SVs in the pangenome. However, due to their physical removal in the pangenome, deletion SVs did not gain any annotations. Therefore, by our methods, only inversion, insertion, and tandem duplication SVs, totaling 12,483 SVs, were functionally annotated from the published transcript data (**Script A5-A6**).

## Defining Gene Ontology Terms

Homology-based annotation applies previous genetic investigations of shared gene (or protein) functions to later studies. Therefore, annotations must be highly organized to be transferable and scalable for genome-wide or multigenome-wide features within or between species.

The published annotation file of *L. usitatissimum* contained multiple systems of homology-based annotations, but Gene Ontology (GO) terms were selected based on their abundance in the data. In genomic-level functional analysis, the most ubiquitous functional annotation system is GO. GO creates three overarching functional domains, describing grouped and serial molecular functions or events, subcellular or macromolecular protein localization, and the molecular impacts of a protein products. For instance, after searching the function for the GO term

GO:0016743, the GO database, **AmiGO** (https://amigo.geneontology.org/amigo), would call these five annotations first: "carboxyltransferase activity", "carbamoyltransferase activity", "carbamoyltransferase", and "ornithine carbamoyltransferase." Thus, the conclusion can be confidently made that the annotated region of interest carries carboxy- or carbamoyltransferase activity.

The functional annotations of all GO terms were extracted first using the **Gene Ontology** database (https://api.geneontology.org/), extracting the "label" functional annotations with a custom script (**Script A5**). Where the GO database connection did not retrieve an annotation label, functional annotations were manually added using **AmiGO**. **AmiGO** is the online repository for accessing data from the **Gene Ontology** consortium database. These GO annotations combined biological processes, cellular components, and molecular functions.

## SV Set Enrichment Analysis

Of the 12,483 pangenomic SVs, 55,079 GO terms were assigned across SV regions. Each GO term labels a unique function; multiple functional annotations were often made to individual SVs. Collectively, the 55,079 GO terms represented 1,252 unique functions from the **Gene Ontology** and **AmiGO** databases (**Scripts A5-A6**).

Functional analyses are frequently conducted through gene set enrichment analysis. However, our investigation considers functions prescribed to larger SV regions, spanning multiple genes. Insertions, inversions, and tandem duplications were functionally annotated from the public transcript of *L. usitatissimum*. Using this methodology, however, deletion SVs could not be annotated (explained in *Annotating SVs of the pangenome)*.

In our investigation, functional variation in SVs between varieties was of particular interest. To distinguish the functional impact of SVs in the cultivated flax varietals (*L. usitatissimum* var Atlant, Heiya, and Longya) and in the wild flax varietal (*L. bienne*), two SV sets were generated to divide SVs. The “Cultivar” set contained SVs present in all the cultivars, but not in *L. bienne*. The “Wild” set contained SVs present only in *L. bienne* and were absent from *L. usitatissimum* var Atlant, Heiya, and Longya. *L. usitatissimum* var CDC Bethune is not listed, as it was the reference assembly for SV calling.

The SV enrichment ratio infers which functions are more strongly represented in the Cultivar and Wild SV sets. The ratio compares the frequency of an SV, associated with a given function, that arises in respective Cultivar or Wild SV sets compared to the entire pangenome (**Script A6**).

# Results

## L. bienne assembly aided by HapHiC HiC scaffolding

Previous attempts to scaffold *L. bienne* var isolate 15003 with HiC reads used the **HiRise** assembly, the proprietary pipeline of Dovetail Genomics. **HiRise** made over 480 joins across the *L. bienne* var 1_6 and Isr assemblies. However, the resultant HiC heat maps (**Figure 4)** suggest that the **HiRise** pipeline failed to inform chromosome-level grouping, ordering, and orientation of input contigs; the **HiRise** assemblies showed L50 values ≤ 3, and only five chromosome structures (**Figure 4, Table A1**), which is an underestimation of the true chromosome number: 15 [90].

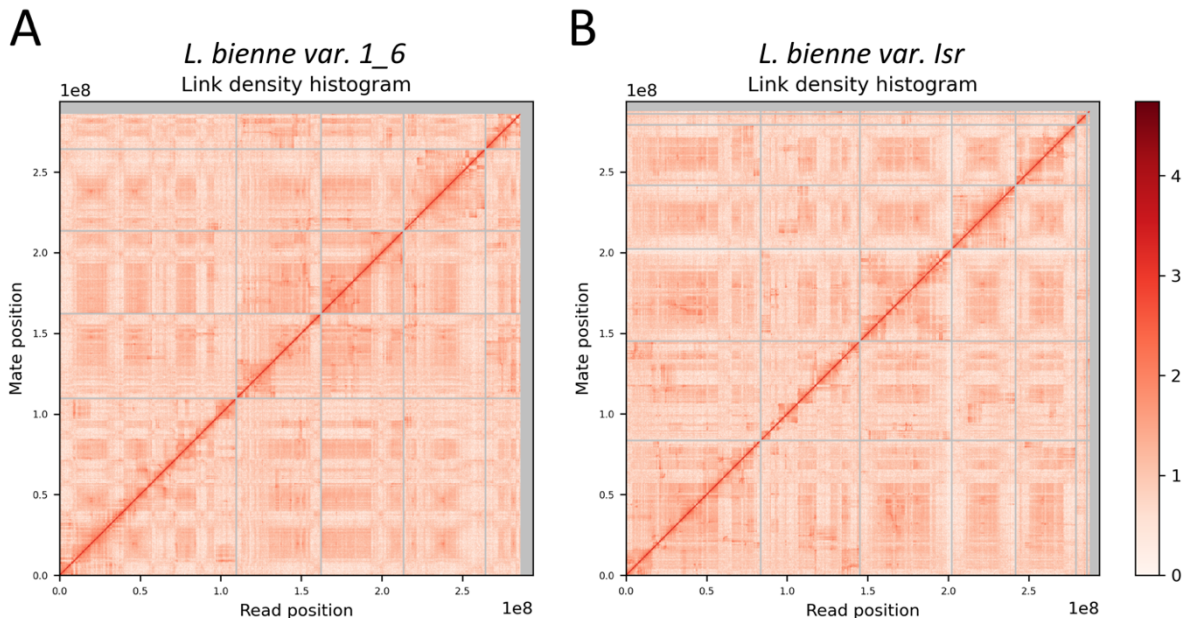



*Figure 4 HiC heat maps for the Dovetail HiC-scaffolded HiRise assemblies of (A) L. bienne var 1_6 and (B) Isr.*

In contrast, **HapHiC** scaffolding achieved a relative improvement in genome assembly. **HapHiC** scaffolded the starting *L. bienne* isolate 15003 assembly (GCA_010665285.1) against over 2,000 HiC reads from *L. bienne* var 1_6 and Isr to create more continuous and complete scaffolds. The chimeric **HapHiC** assembly of *L. bienne* presented nine chromosomes (**Figures 5-6**), progressing beyond **HiRise** towards a chromosome-level genome assembly.

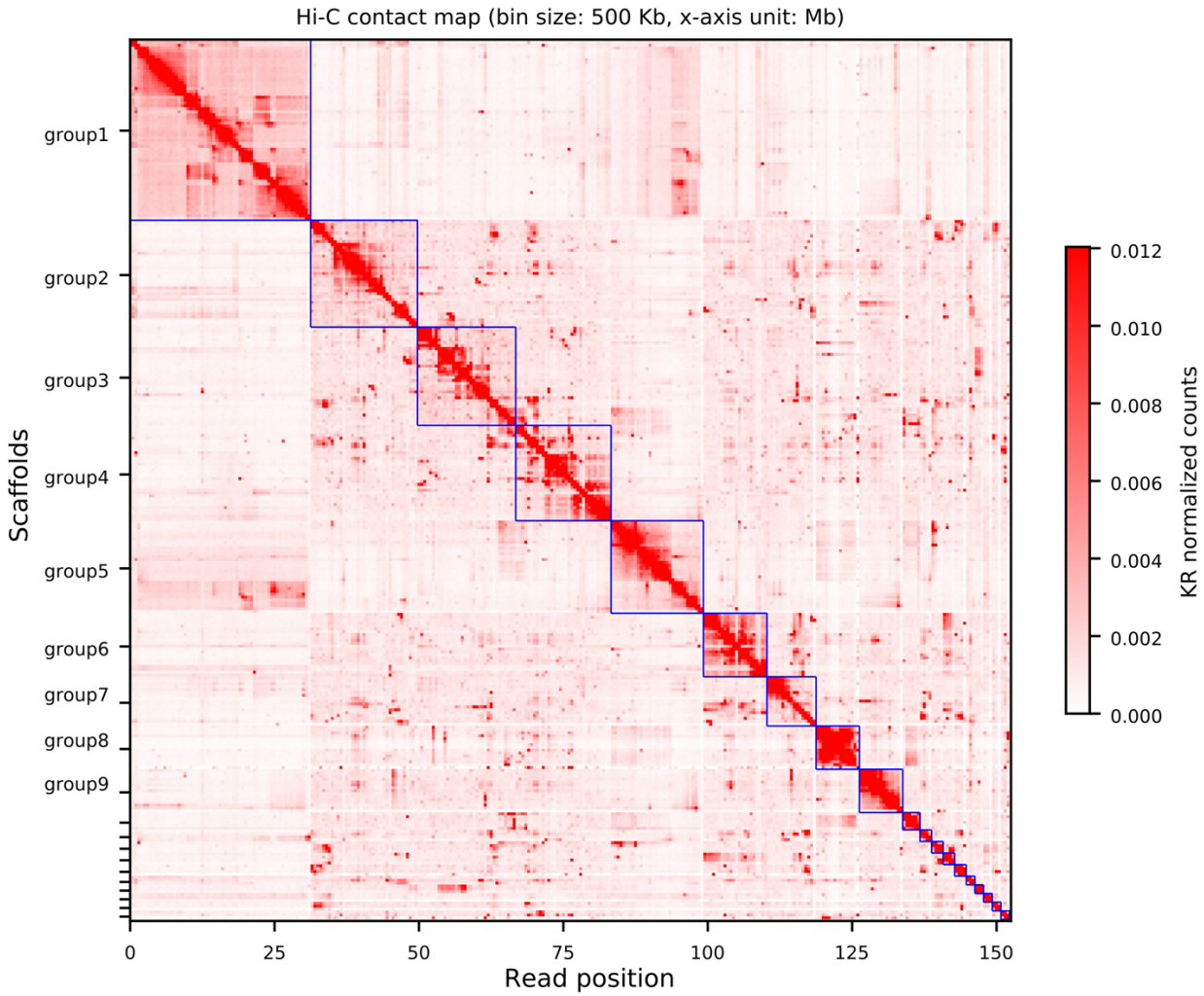



*Figure 5 HiC interaction heatmap of L. bienne HiC scaffolds produced by HapHiC alignment software. Scaffold group assignments, as calculated by HapHiC, are identified in blue. Read and (read) mate positions from the HiC assembly are defined along the x and y axes, respectively.*

To clarify the HiC scaffold groupings from **Figure 5**, **Figure A2** presents the HiC contact maps of individual contact groups, in which groups 1-9 are defined as chromosome-level groups. Smaller scaffolds were left with their HiC alignment read names (e.g., QMEG…, **Figure A2**).

MAPQ-filtered reads from the HiC scaffolded assembly were then used as inputs for **RagTag** scaffolding against *L. usitatissimum* var CDC Bethune. Therefore, the *L. bienne* genome output by **RagTag** represents genome information from the wild varieties isolate 15003, 1_6, and Isr. Sequencing multiple varieties increased the absolute data input to construct a more complete *L. bienne* genome. However, achieving a higher assembly level likely sacrificed some of the innate diversity of *L. bienne* var isolate 15003, 1_6, and Isr.

## Improved genome assembly from publicly available accessions

All non-reference *Linum* genome assemblies (*L. usitatissimum* var Atlant, Longya, Heiya, and *L. bienne*) were improved by homology-based scaffolding. In all four instances, the **scaffold** command of **RagTag** was used, with *L. usitatissimum* var CDC Bethune as a reference.

Scaffolding against *L. usitatissimum* var CDC Bethune successfully reduced the number of contigs within all assemblies. **RagTag** assembly statistics describe more contiguous contigs and scaffolds[9] for all assemblies. Averaged across the four assemblies, 97% of contigs were joined to another contig, with default gap lengths of 100 bp (**Table A5**).

**QUAST** assembly statistics (**Table 6**) also demonstrate that the non-reference *L. usitatissimum* cultivars and *L. bienne* assemblies are more complete than their initial assemblies (**Table 2**). The final assembly statistics suggest that near-chromosome levels of assembly were achieved for *L. usitatissimum* var Atlant, Heiya, and *L. bienne*. If the contigs of an assembly were sorted by size, the L50 and N50 statistics, respectively, represent the number of contigs and the length of the contig at the halfway mark. Because *L. usitatissimum* and *L. bienne* are diploid (2n = 30), their haploid assemblies have 15 chromosomes; halfway through a 15 chromosome assembly, between seven and eight chromosomes should be covered. Therefore, the L50 values ≤ 9 and N50 values ≥ 16.8 Mb (**Table 6**) most closely resemble a chromosome-level genome.

*Table 6 QUAST assembly statistics of final assemblies. The starting chromosome-level assembly of CDC Bethune was used as the reference genome for the QUAST assessments.*

| ***Name*** | ***QUAST assembly statistics*** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Contig number* | *Genome size (Mb)* | *Genome fraction* | *N50 (Mb)* | *N90 (Mb)* | *L50* | *L90* | *GC (%)* |
| CDC Bethune | 15 | 316.2 | - | 20.5 | 17.7 | 7 | 14 | 39.49 |
| Longya | 349 | 305.7 | 93.496 | 20.0 | 17.5 | 7 | 14 | 39.06 |
| *L. bienne* | 573 | 292.6 | 78.829 | 16.8 | 9.84 | 8 | 15 | 38.94 |
| Heiya | 713 | 302.0 | 93.633 | 17.7 | 14.2 | 8 | 15 | 38.95 |
| Atlant | 1319 | 434.2 | 95.096 | 19.8 | 8.30 | 9 | 416 | 38.49 |

Among our starting assemblies, *L. usitatissimum* var Longya was reported as a chromosome-level assembly, with a 306.4 Mb genome and 1,608 contigs grouped into 15 chromosomes (**Table 2**), with L50 = 8 and N50 = 18.2 [16]. The similarities between those reported statistics and **Table 6** supports the designation of our final assemblies as near-chromosome level. Relative to the starting assembly of *L. usitatissimum* var Longya (GCA_010665275.2, **Table 1**), our final assembly also showed reduced fragmentation (decrease in L50 value to 7) and increased contig continuity (increase in N50 value to 20.0 Mb). While there was <1 Mb change in the *L.*

[9] **RagTag "scaffold"** [24] equally tolerates the contig and scaffold-level inputs, but only reports "contigs" in output statistics. Contigs are gapless sequence intervals, while scaffolds are constructed of contigs and gaps; a subtle difference such that scaffolds carry additional relative sequence orientation and positioning information. For consistency, we report results with the appropriate software or program-specific terminology.

*usitatissimum* var Longya genome size, the number of scaffolds in the assembly was reduced by 78% (**Tables 2** and **6**).

**BUSCOs** were also used to assess the completeness of the final assemblies. Combining single and duplicated BUSCOs, the final genome assemblies of *L. usitatissimum* var Atlant, Heiya, Longya, and *L. bienne,* respectively, had complete BUSCO scores of 95.27, 95.18, 94.93, and 93.98% (**Figure 6**, **Table A6**). Of the 2326 BUSCOs in the eudicot_odb10 lineage, 94.15% were complete in the reference *L. usitatissimum* var CDC Bethune.

Even the least assembled final genome, *L. usitatissimum* var Atlant (**Table 6**), showed improvement from the publicly available genome. The previous *L. usitatissimum* var Atlant assembly [44] was subject to a BUSCO assessment in 2022 against the same lineage as our analyses (**Table 3**). Comparatively, our final assembly showed a 29.6% reduction in "missing" BUSCOs, and a greater proportion of "duplicated" to "single" complete BUSCOs (~3:1 versus ~2:1, **Figure 6, Table A6**). As more BUSCO genes were successfully detected, the BUSCO assessment indicates that the final *L. usitatissimum* var Atlant assembly is more complete.

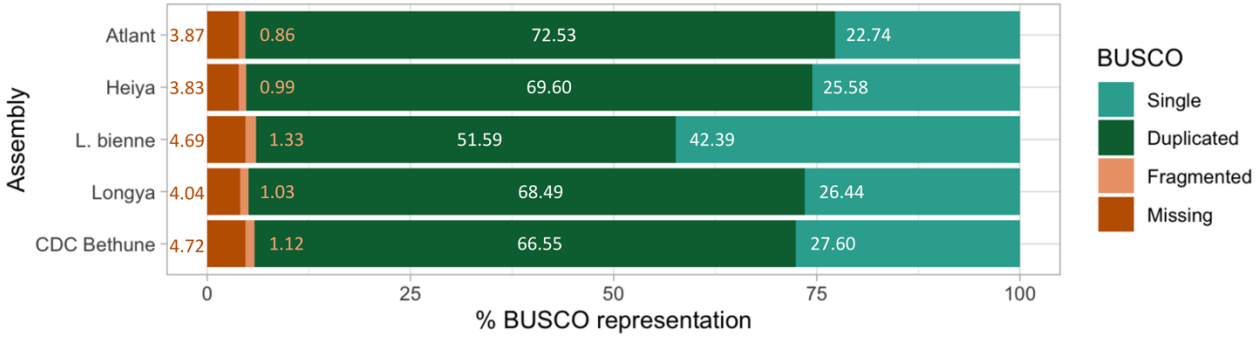



*Figure 6 BUSCO assessments of the final assemblies (Atlant, Heiya, Longya, CDC Bethune, and L. bienne) using the eudicots_odb10 lineage. Present single, duplicated, fragmented, and missing BUSCOs are shown as percentages (%).*

## QUAST hinted differences amongst varietal assemblies

**QUAST** was also used to generate less-commonly reviewed k-mer statistics for **RagTag**'s output assemblies, rather than the typical use of k-mers to estimate genome. In this instance, **QUAST** k-mer statistics represent the respective divergence between *L. bienne* and *L. usitatissimum* var Atlant, Longya, Heiya, and the reference genome of *L. usitatissimum* var CDC Bethune.

Similar k-mer completeness scores of *L. usitatissimum* var Heiya (77.61%) and Longya (77.81%) **RagTag** assemblies, therefore, suggest similar levels of variation from *L. usitatissimum* var CDC Bethune, relative to *L. usitatissimum* var Atlant (79.94%).

While their k-mer completeness scores suggest similar degrees of divergence from *L. usitatissimum* var CDC Bethune, k-mer-based misjoins statistics in *L. usitatissimum* var Heiya (577 misjoins) and *L. usitatissimum* var Longya (1,030 misjoins) suggest that the underlying variation between the *L. usitatissimum* var Heiya and Longya assemblies also diverges from one another, unsurprisingly.

Regular **QUAST** misassembly metrics also vary among the varietals. Most critically, comparative trends in the k-mer-based and regular misjoins across flax varieties imply that the starting genome assembly level influences the number of misassembled contigs in the assemblies.

## Pangenome Construction

The final assemblies (**Table 6**) served as inputs for pangenome construction without a reference genome using **PGGB**. The resultant pangenome graph was composed of 8,897,875 nodes and 12,104,212 edges, and 5,359 paths corresponding to variation within the graph.

Of the detected variation in the graph, 99.62% is represented acyclically (see *Algorithms inside PGGB pangenome creation*). Cyclic variation, and variation that was not definable as an insertion, deletion, inversion, duplication, or breakend, was omitted from the VCF output or removed during the SV validation and cleaning.

The graph length was reported as 564,843,077 bp, from the total 1,650,585,360 bp accumulated across the five input *Linum* sp. genomes. If the pangenome length were considered only from additive SVs (insertions, inversions), the added length would be around 20 Mb (**Figure 7**). Notably, there was little length difference added from the SVs of *L. bienne* relative to the *L. usitatissimum* varieties. At the pangenome scale, SVs are balanced from each incorporated genome.

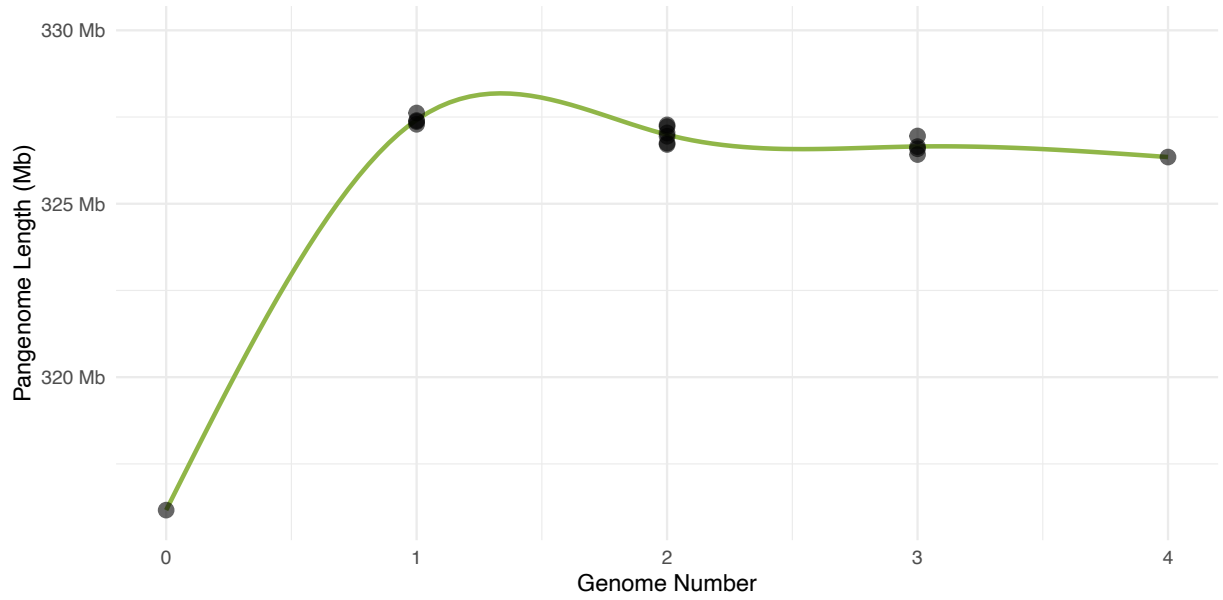



*Figure 7 The length of the pangenome, if estimated by SVs alone from the reference genome length (316.2 Mb). The pangenome length (y-axis) increases by adding a single genome's SVs. Successively adding SVs shared between multiple genomes (x-axis, > 1) leads to a flattening of the pangenome size estimate curve.*

The pangenome effectively captures SVs from a small sample size, and many SVs appear shared amongst flax varieties (**Figure 7**). In more divergent genomes, we could expect to see the length estimate of the pangenome supplemented by each additional genome. However, successive combinations of genomes (**Figure 7**, genome number > 1) do not drastically improve the flax pangenome length. With neither significant improvement nor loss in pangenome length, it can be inferred that SVs present within a single variety are near-ubiquitously detected within relatives.

# Structural Variation Calling

## Validated pangenome SV calls

147,290 pangenome SV calls were generated through **PGGB**, and **SVIM-asm** generated 86,238 genome-level SV calls. Both programs detected variants against the reference sequences of *L. usitatissimum* var CDC Bethune. Of the original **PGGB** calls, 40% of variants were validated by **SVIM-asm** within 30 bp exact position matches (**Figure 8B**). The final SV dataset included 57,729 cross-validated pangenomic SVs.

Our findings agree with the consensus that pangenome-based SV calling outperforms linear genomes [26,34,37,78,93,94]. Calling SVs from pangenome alignments currently offers the best strategy to compile complete genomic variation. Despite haploid genome inputs, variants were detected and well-represented amongst pangenome SVs. Therefore, future pangenome investigations should continue to incorporate collapsed genomes, currently shelved in repositories, to explore genetic diversity at greater scales.

While most **PGGB** calls were not found by genome-level SV calling, **PGGB** comfortably outperformed **SVIM-asm**, even after amalgamating four rounds of genome-level SV calls (**Figure 8**).

Assembly-level calls between non-reference varieties (e.g., *L. usitatissimum* var Atlant against *L. bienne*) would be preferable to validate for additional variants from the pangenome. However, varying levels of assembly between genomes limit the ability to detection comparable variation, compared to the simpler use of a single, shared reference. Using multiple assembly-based callers may also help to validate more pangenome SV calls.

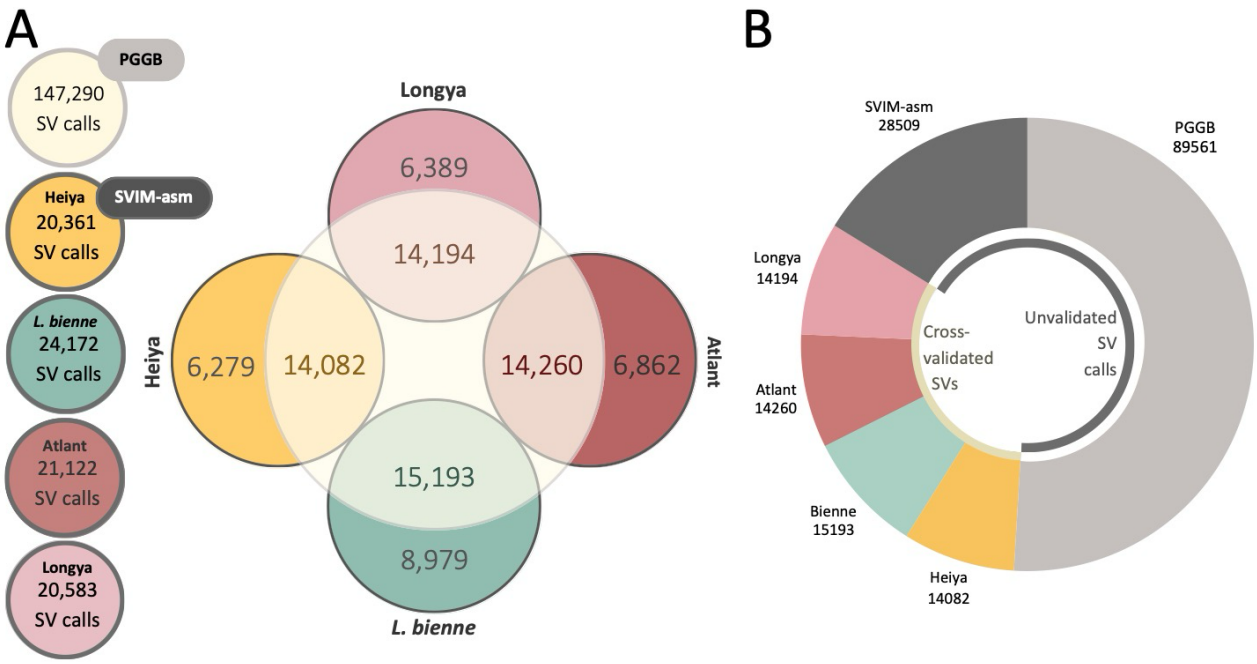



*Figure 8 (A) SV calls > 50 bp were produced independently by SVIM-asm for each non-reference varietal and by PGGB for the flax pangenome, against CDC Bethune. (B) Pangenomic SV calls are colored by varietal, if validated by SVIM-asm, relative to SV calls left unvalidated at the genome level (dark grey) or pangenome level (light grey).*

## Diversity of pangenome SVs

Of the SVs validated from the pangenome and genome collections, there were 35,138 insertions, 21,239 deletions, 22 inversions, and 762 undefined SVs. **PGGB** explicitly presented entire nucleotide sequences for SVs as it is present in the reference, and the (alternate) variant(s) in the non-reference genome(s). In contrast, **SVIM-asm** offered only the change in sequence length for the specified SV (e.g., SVLEN = -1306). While **SVIM-asm** detected tandem duplications with defined coordinates, conflicting data formats with **PGGB** meant that our cross-referencing method (**Script A2**) could not validate the length of duplication SVs. Tandem duplications are still included in the cross-validated SV dataset, but were omitted from **Table 7** and **Figures 9-10** to present length information consistently.

SV counts for deletions, insertions, and inversions are summarized nonredundantly in **Table 7**, as present in the pangenome. That is, an SV detected in *L. usitatissimum* var Atlant and *L. bienne* in the pangenome is counted as one SV. All-inclusive SV counts are summarized in **Table A7**.

*Table 7 Cross-validated SV counts for deletions, insertions, and inversions across the chromosome groups of the pangenome. Chromosome groups were assigned based on L. usitatissimum var CDC Bethune coordinates.*

| Chromosome | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **Total SVs** | 5108 | 4468 | 4829 | 3606 | 3186 | 3239 | 3209 |
| **Deletions** | 1930 | 1820 | 1824 | 1322 | 1237 | 1148 | 1238 |
| **Insertions** | 3177 | 2646 | 3003 | 2284 | 1949 | 2089 | 1971 |
| **Inversions** | 1 | 2 | 2 | 0 | 0 | 2 | 0 |

*Table 7 Continued*

| Chromosome | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|
| **Total SVs** | 3907 | 3946 | 3869 | 3382 | 3797 | 3876 | 3318 | 2659 |
| **Deletions** | 1360 | 1472 | 1586 | 1259 | 1317 | 1516 | 1220 | 990 |
| **Insertions** | 2546 | 2474 | 2282 | 2118 | 2480 | 2355 | 2097 | 1667 |
| **Inversions** | 1 | 0 | 1 | 5 | 0 | 5 | 1 | 2 |

**Figure 10** plots deletion, insertion, and inversion SVs of the pangenome along the chromosomal groups of *L. usitatissimum* var CDC Bethune. Following **SVIM-asm** validation, each varietal presented highly similar SV counts (**Figure 8**). However, dividing **Figure 10** by flax variety (not shown) confirmed that SVs in the pangenome are similarly distributed among *L. usitatissimum* cultivars and *L. bienne*, and arise evenly from the input varietals, as suggested in **Figures 7-8**.

Beyond SV distribution patterns, varietals showed non-significant differences in SV length (Kruskal-Wallis, p-value = 0.3123). However, there are significant length differences (Kruskal-Wallis, p-value < 2.2e-16) within individual classes of SVs (deletions, inversions, insertions), and the frequent appearance of large variant outliers (**Figure 9**).

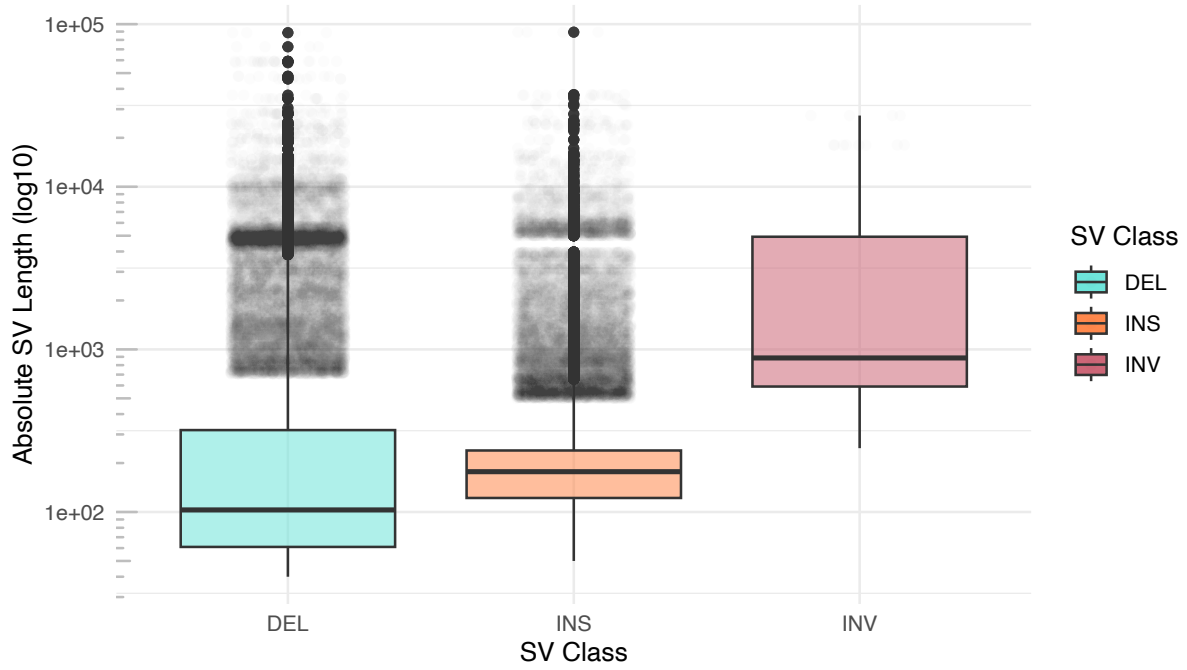



*Figure 9 The length of pangenome SVs, as ordered by deletions (DEL), insertions (INS), and inversions (INV).*

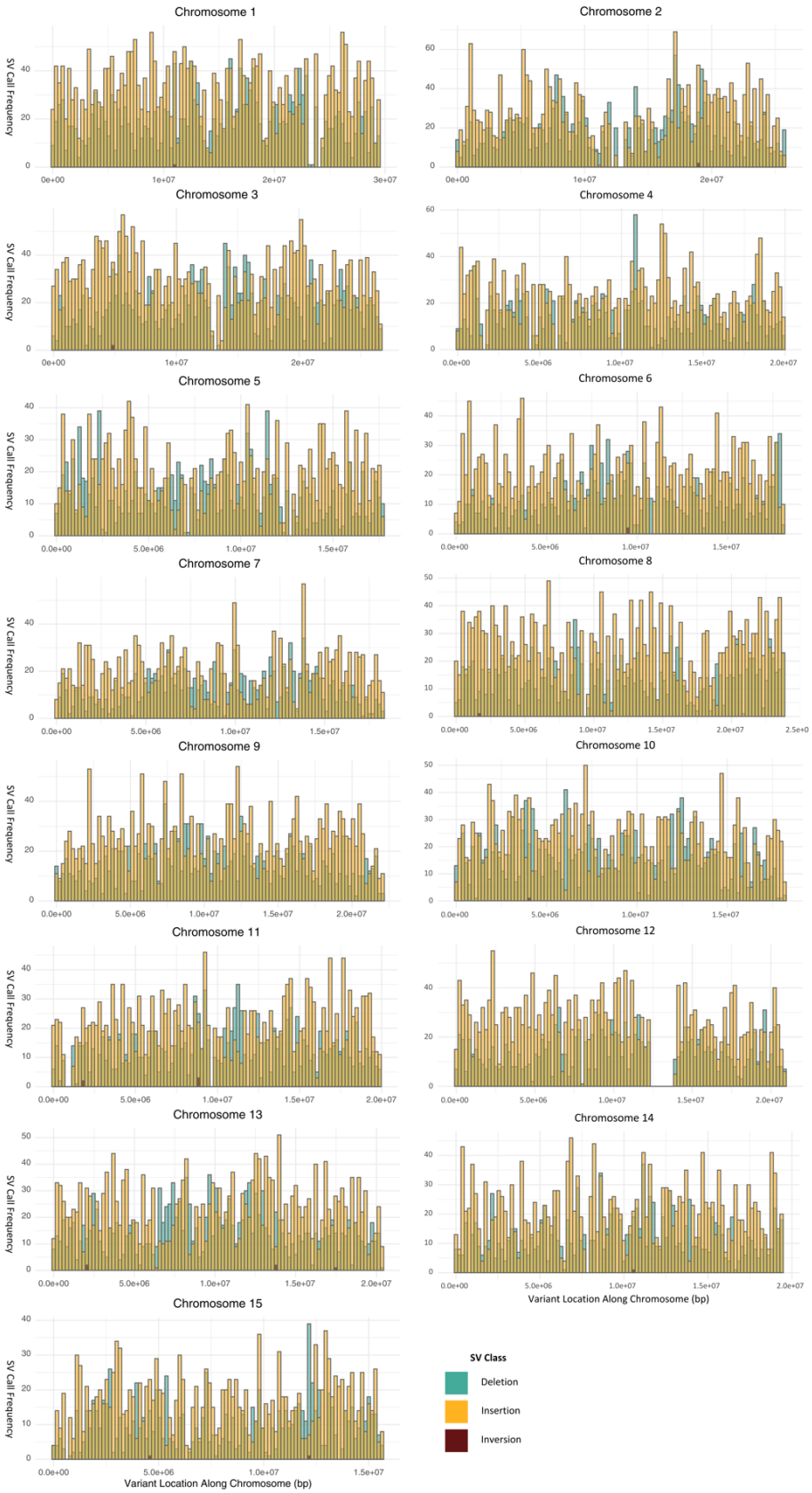



*Figure 10 Deletions (blue), insertions (yellow), and inversions (red) SVs plotted against the pangenome chromosomes.*

A PCA of SVs (n = 57,729) validates the intervarietal diversity previously implied by post-assembly **QUAST** k-mer statistics (**Figure 11**). The most prominent distinction is between the SVs of wild *L. bienne* and the *L. usitatissimum* cultivars, relative to the variance between the fibrous (Heiya, Atlant) and oilseed (Longya) flax cultivars.

While there is no clear varietal grouping by agricultural function (oilseed and fiber flax), the PCA shows the distinct grouping of *L. usitatissimum* var Longya and Atlant from *L. usitatissimum* var Heiya. Post-assembly k-mer statistics suggest that if *L. usitatissimum* var CDC Bethune were plotted, the Canadian oilseed variety would be placed equidistant from Longya and Heiya, and remain closest to Atlant.

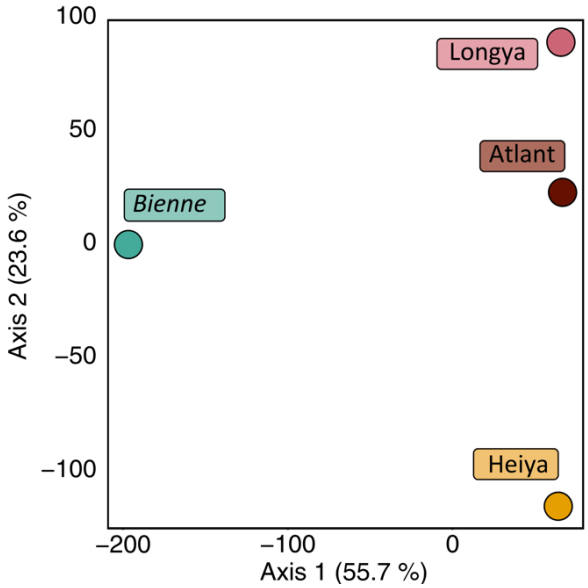



*Figure 11 PCA plot with two principal components (79.3% of variance explained) for the SV of L. usitatissimum var Atlant, Heiya, Longya, and L. bienne.*

## Gene Ontology

Genome transcript data from the 2012 assembly of *L. usitatissimum* var CDC Bethune [47] was used to functionally annotate SV regions of the pangenome. After scripted and manual labelling of functional annotations, 12,483 SVs were assigned 55,079 GO terms. Often, multiple GO terms were assigned to a single SV. The GO terms corresponded to 1,252 unique functions.

A custom script (**Script A5**) associated 97.8% of the unique GO terms to a single functional label from the **Gene Ontology** database. 28 GO IDs returned "no matches", so the **AmiGO** database was then used to label functions manually. In all instances, the first **AmiGO** hit for the ambiguous annotations was "alkaloid biosynthetic process," suggesting there may have been an error in fetching the URL for that GO ID.

Alkaloid biosynthetic processes describe reactions and pathways forming an alkaloid; the alkaloids in these cases are classified separately from "nonprotein amino acids, amines, peptides, cyanogenic glycosides, glucosinolates, cofactors, phytohormones, or primary metabolites" [93]. The second potential annotation for ambiguous annotations was "monooxygenase activity," reflecting the involvement of monooxygenases in the oxidation reactions during alkaloid biosynthesis [96]. "No matches" GO terms were labeled simply as "alkaloid biosynthetic process" (**Appendix File A1**).

### SV Set Enrichment Analysis

GO functional analysis selected from the pangenome SVs, only SVs that were present in *L. bienne* alone (Wild SV set) or present in all three flax cultivars: *L. usitatissimum* var Atlant, Heiya, and Longya (Cultivar SV set). After functionally annotating both SV sets, 208 unique functions were labeled across 1,415 SVs in the Wild SV set, and 380 unique functions were

labeled across 2,890 SVs in the Cultivar SV set. **Appendix File A1** contains all SV annotations and the function's presence in each SV set.

The functional annotations, grouped by SV set, aim to distinguish functional differences between cultivated and wild flax, relative to the pangenome. **Figure 12** lists the 25 most uniquely represented functions in the SV sets by measuring an SV enrichment ratio—modified from the concept of "gene ratio" [94].

The SV enrichment ratio ("Ratio", **Figure 12**) reflects the frequency with which an SV was labeled a particular function within an SV dataset relative to the entire pangenome. Higher SV enrichment ratios, therefore, represent functional annotations that are more unique to the appropriate SV set and the corresponding flax varietals. In other words, the SV enrichment ratio identifies functions within SV sets that are more distinct from the functions present across the pangenome's SV.

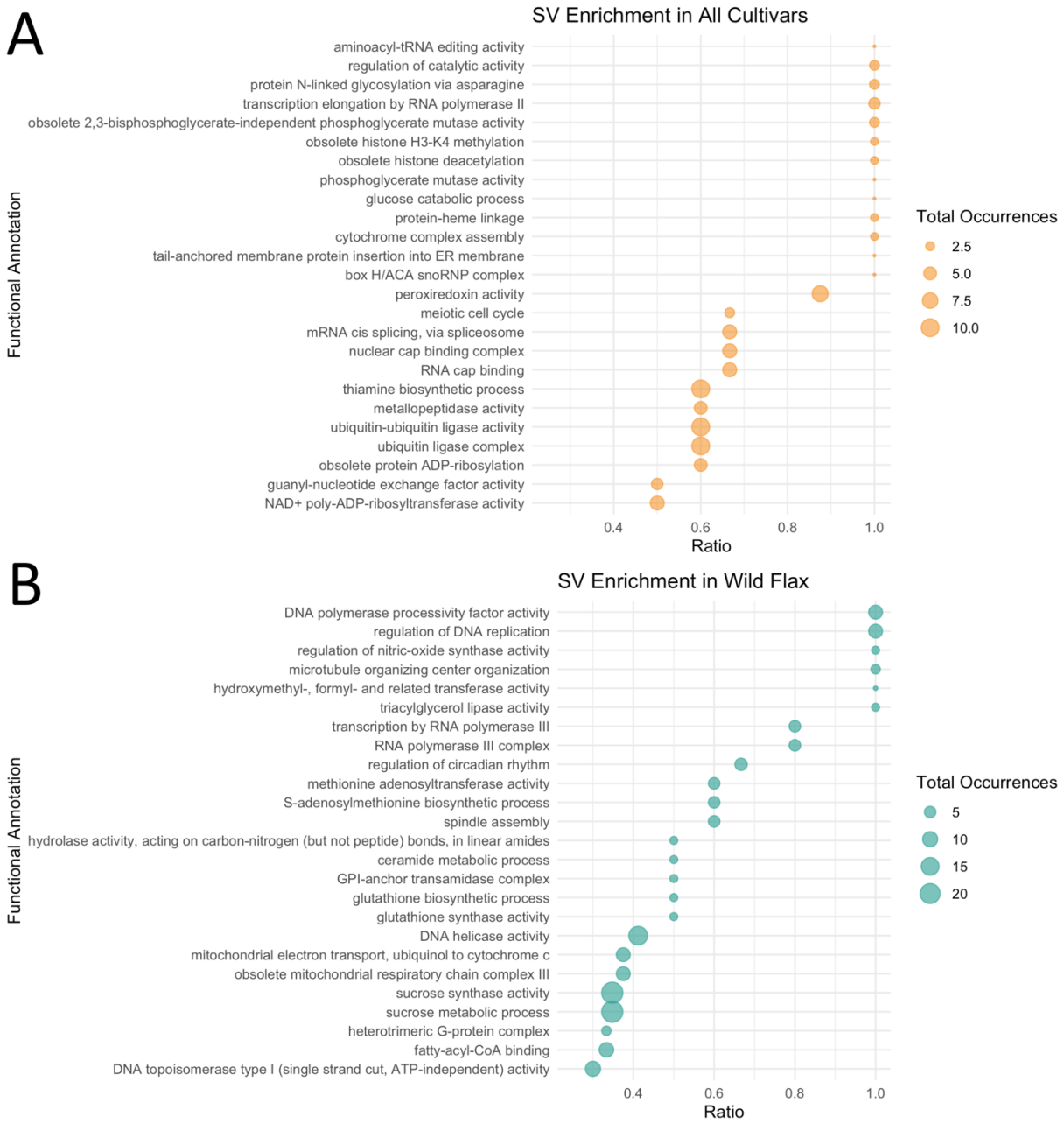



*Figure 12* GO enrichment analysis for SVs in Atlant, Heiya, and Longya (A, All Cultivars) or *L. bienne* (B, Wild Flax) relative to all annotations in the pangenome (Total Occurrences). The 25 most significant GO terms (Functional Annotation) are provided for the Cultivar SV and Wild SV sets based on their SV enrichment ratio (Ratio). The SV enrichment ratio calculates the frequency of a specific function when assigned to SVs in a given SV set relative to the assignment of that function across all SVs in the pangenome. Total Occurrences refers to the frequency at which the function was reported in the pangenome.

The functions with the highest enrichment ratios (**Figure 12**) do not equate to the most abundantly annotated functions across the SV sets or the pangenome itself. Due to their generalizability, the most frequently annotated functions are less relevant to discussing intervarietal functional differences. Nonetheless, **Table 8** reports the 20 functions assigned most frequently to either SV set. Functional annotations found in the "top 20" of both SV sets are

bolded (**Table 8**) to emphasize which functions were greatly affected by SV in both cultivar and wild flax varietals.

Table 8 *The 20 most commonly assigned functions in SVs present, respectively, only in cultivated flax, Atlant, Heiya, Longya, (Cultivar SV Set) and wild flax, L. bienne (Wild SV Set).*

| ***Cultivar SV Set*** | | | ***Wild SV Set*** | | |
|---|---|---|---|---|---|
| ***Function*** | *Annotation frequency* | *Ratio* | ***Function*** | *Annotation frequency* | *Ratio* |
| **Protein binding** | 171 | 0.061 | **ATP binding** | 89 | 0.039 |
| **ATP binding** | 122 | 0.054 | **Membrane** | 60 | 0.045 |
| **Obsolete oxidation-reduction process** | 108 | 0.058 | **Alkaloid biosynthetic process** | 60 | 0.029 |
| **Alkaloid biosynthetic process** | 101 | 0.049 | **Protein binding** | 59 | 0.021 |
| **Protein kinase activity** | 93 | 0.056 | **Protein kinase activity** | 56 | 0.034 |
| **Protein phosphorylation** | 93 | 0.056 | **Protein phosphorylation** | 56 | 0.042 |
| **Membrane** | 72 | 0.054 | **Regulation of DNA-templated transcription** | 50 | 0.042 |
| **Regulation of DNA-templated transcription** | 67 | 0.057 | DNA binding | 38 | 0.034 |
| **Metabolic process** | 64 | 0.068 | DNA-binding transcription factor activity | 33 | 0.055 |
| Zinc ion binding | 61 | 0.065 | **Obsolete oxidation-reduction process** | 33 | 0.018 |
| Oxidoreductase activity | 58 | 0.067 | Nucleus | 32 | 0.043 |
| **DNA binding** | 52 | 0.047 | GTP binding | 20 | 0.056 |
| **Transmembrane transport** | 46 | 0.051 | **Transmembrane transport** | 20 | 0.022 |
| Catalytic activity | 42 | 0.059 | Ribosome | 19 | 0.034 |
| Nucleic acid binding | 40 | 0.055 | **Structural constituent of ribosome** | 19 | 0.032 |
| Proteolysis | 37 | 0.050 | **Translation** | 19 | 0.032 |
| **Structural constituent of ribosome** | 37 | 0.063 | **Intracellular anatomical structure** | 17 | 0.031 |
| **Translation** | 37 | 0.062 | **Metabolic process** | 17 | 0.018 |
| Carbohydrate metabolic process | 36 | 0.051 | Acetylglucosaminyltransferase activity | 15 | 0.163 |
| **Intracellular anatomical structure** | 36 | 0.066 | Heme binding | 15 | 0.029 |

**Appendix File A1** contains the full annotation results, but **Table 9** selects annotation results with domestication-relevant functions. While non-exhaustive, **Table 9** organizes GOs that may relate to the overarching character changes associated with plant domestication.

*Table 9 A conservative list of the functional annotations made to flax SVs that can be associated with crop domestication (Domestication trait). Annotation frequency records the number of times the GO term was assigned to pangenome SVs.*

| Domestication trait | | GO terms | Annotation frequency |
|---|---|---|---|
| **Plant architecture** | **Cell wall** | Cell wall biogenesis; cell wall modification; cell wall structure; cell wall organization | 310 |
| | **Architectural signaling** | Cell morphogenesis; auxin transport; response to auxin; brassinosteroid mediated signaling; cytokinin metabolic process | 129 |
| | **Cytoskeleton** | Actin cytoskeleton organization; microtubule cytoskeleton organization | 89 |
| | **Leaf** | Leaf development; leaf formation | 6 |
| **Growth** | **Photoperiod-associated development and photosynthesis** | Photoperiodism, flowering; circadian rhythm regulation, response to light stimulus; photosynthesis, light harvesting; photosynthesis, light reaction; photosystem assembly and stability; chlorophyll biosynthetic process; electron transporter, electron transfer activity, electron transport chain | 165 |
| | **Seed development** | Nutrient reservoir activity; starch binding; sucrose biosynthetic process; sucrose metabolic process; triglyceride biosynthesis; sphingolipid biosynthesis; glycolipid biosynthesis | 92 |
| | **Flowering** | Regulation of flower development; reproductive structure development | 19 |
| **Stress response** | **Abiotic stress** | Response to oxidative stress; response to salt stress; response to freezing; response to heat; response to water | 179 |
| | **Genetic stability** | DNA repair; DNA mismatch repair; DNA recombination; chromatin remodeling; chromosome segregation; | 158 |
| | **Biotic stress** | Response to biotic stimulus; plant-type hypersensitive response (immune response) | 45 |

# Discussion

## The first pangenome draft of flax

Following continuity improvements to four of the five published genome assemblies in this study (**Figures 5-6, Table 6**, and **Figure A2**), the final pangenome graph included sequences from four varietals of *L. usitatissimum* (vars CDC Bethune, Atlant, Heiya, Longya) and *L. bienne* (isolate 1003). *L. bienne* var 1_6 and Isr guided sequence placement within the *L. bienne* genome, but their sequences were not incorporated into the pangenome. This graph can be considered the first draft pangenome of flax, expanding the variation across the five input genomes (average genome length = 331.14 Mb, **Figure 4)** into a 564.8 Mb acyclic, directed graph.

The flax pangenome's draft status, despite being a full pangenome graph, is based on the input of haplotype-insensitive genome assemblies. Public genomic data slightly impairs the construction of pangenomes, as most genomes reported to NCBI and equivalent databases are haplotype-collapsed genomes. Such haploid assemblies combine haplotypes from different chromosomes

(e.g., paternal and maternal). That collapse minimizes the full variation of diploid organisms and inherently limits the possible investigation of genetic diversity [37]. As the starting assemblies of this project were haploid, the flax pangenome—despite representing large quantities of undocumented genetic variants—failed to take advantage of haplotype-sensitive program options and existing technologies capable of representing haplotype information.

In later processing, reference-based scaffolding helped achieve the finalized, near-chromosome assemblies. However, these scaffolding stages could have biased genome organization [98] and, in turn, reduced the quality of input genomes for pangenome construction. For pangenome SV detection, reference biases carry over as a concern from genome-level investigations, manifesting possibly as reduced or falsely generated pangenomic variation. Future investigations should consider additional sequencing and reference-quality genome generation to define the structural consequence of pangenomic reference bias.

In the flax pangenome, efforts to minimize reference bias were concentrated in the software selection of **PGGB**, which avoids reference-based graph alignment. Pangenomes may be constructed through all-to-all [49] or successive pairwise alignments [77]. The former reduces reference bias by considering all genome segments equivalently, whereas successive pairwise alignments require a starting genome upon which to lay successive comparisons. For successive alignments, non-aligning regions against the reference offer variant information. However, the order of iteratively added genomes may affect the resulting pangenome variations. These consequences have yet to be explored. Of course, in the greater context of genetic studies, pangenomes themselves still counter possible reference biases, as they replace inherently biased single-genome references.

## Inter-varietal differences demonstrated through pangenome SVs

SVs validated at the pangenome and genome level are present across the entire pangenome and comprise ~20 Mb of the graph's total length (**Figure 7, 8A,** and **10**). The 57,729 cross-validated SVs measured between 50 and 100,000 bp to capture realistic causal mutations. These size limitations prevent the false detection of small genetic variants (e.g., SSRs) or the misclassification of chromosomal groups as large SVs. Lower length limits avoid detecting mutations that may have been explored previously in genetic and genomic investigations; upper length limits avoid flagging variants that would be too large to engineer into flax in future breeding applications of the pangenome, as the flax genome is relatively small.

Only 1,330 (2.3%) SV lengths could not be confirmed across software, due to conflicting data formats between the **SVIM-asm** and **PGGB**. Both programs present length information, but their presentation uses different formats (described in *Diversity of pangenome SVs*). For researchers wishing to assign length labels, though, either data format should be accurate.

Summary statistics of the pangenome SVs show little variance: SVs were evenly distributed across the pangenome chromosome groups and within each input varietal (**Table 7**, **Figure 10**). While fewer in number, inversion SVs consistently have long lengths, relative to the broader size distribution of insertion and deletion SVs (**Figures 9-10**). There is also an abundance of large insertion and deletion outliers (**Figure 9**). To inform multi-trait selection, SVs comprising bigger genomic regions may be more useful candidates for isolating causal variants. The reduced

recombination of large inversions and inversions' implication in adaptation and speciation present them as particularly optimistic candidate variants [99].

Subsequent analyses of SVs describe genetic differences in flax between wild and cultivated varieties. PCA analysis (**Figure 11**) indicates a general divergence between *L. bienne* and *L. usitatissimum* var Longya, Atlant, Heiya SVs relative to *L. usitatisssimum* var CDC Bethune. Despite the grouping amongst the cultivars, though, there is no functional correlation between the fibrous (Atlant and Heiya) and oilseed (Longya) flax. The SV PCA likely reflects the dominant, industrial practice of crossbreeding between fibrous and oilseed flax. However, there is an additional possibility that the cross-functional grouping of *L. usitatissimum* var Longya and Atlant (**Figure 11**) may be due to SV identification against the CDC Bethune reference.

Other pangenome studies, in barley and grapevine [34,40], have described logarithmic additions to the length of the pangenome through single-copy gene lengths and additional haplotypes, as more samples are incorporated. Comparatively, the length addition pattern from SVs in the flax pangenome was less pronounced (Figure 7), potentially from fewer genome inputs or smaller inter-sample genetic diversity.

Notwithstanding some distinction between wild and cultivated samples, the pangenome structure chiefly validates previous genetic results: *L. usitatissimum* demonstrates low intraspecific diversity [17]. While the genetic likeness of the cultivar varieties to the reference (*L. usitatissimum* var CDC Bethune) should advantage *L. bienne*, the single phylogenetic outgroup, the representation of cultivar genetic diversity is disadvantaged. Haploid assemblies will organically involve variation loss. Furthered by reference-based scaffolding, inherent SVs may have been lost from our best-assembled haploid genomes for cultivars *L. usitatissimum* var Longya and Heiya.

As we've described, pangenomic SV calling begins at a deficit where available genomes exist at lower assembly levels. Future work should confirm and measure both reference and assembly-level biases by reassembling published haploid assemblies from their initial sequencing data. To maximize the re-usability of genomic data, it is critical to understand how starting assembly levels introduce reference bias and distort downstream variation detection.

## Conservative estimation of wild flax diversity

Validated pangenome SV calls were evenly detected across varietals (**Figure 8A**) despite *L. bienne* being the outgroup and referenced against a heterospecific variety. *L. bienne* pulled marginally greater SV counts (24,172 before cross-validation, 15,193 after), likely representing a greater natural genetic diversity amongst the wild population. However, overall, there appears to be low intervarietal diversity (**Figure 11**).

To clarify, this low intervarietal diversity is not an underestimation of *L. bienne* SV diversity due to a reference bias. If anything, the presence of a reference bias should exaggerate the genetic difference of *L. bienne* relative to the diversity amongst cultivated flax. The low genetic difference amongst cultivated flax has been found previously [17,20]. However, the newfound low genetic difference between *L. bienne* and the cultivars may be a technical underestimation arising from the assembly of *L. bienne*; the true genomes of *L. bienne* var isolate 3005, 1_6, or Isr, if assembled completely and independently, would likely yield more SV calls than the *L.*

*bienne* input into the flax pangenome. The possible loss of variation from each incorporated varietal and the reference bias from homology-based scaffolding should be considered when evaluating *L. bienne*'s genetic diversity, and in future efforts to leverage SVs from its assemblies.

Prioritizing the addition of wild accessions to pangenomes demarcates the difference between pangenomes and "super pangenomes", so the continued study of flax should explore deeper sequencing of *L. bienne* and incorporate other wild *Linum* sp. into the pangenome draft. Including more wild flax variants may produce useful information regarding flax's climatic adaptation across Europe, such as the high-altitude adaptations of *L. perenne* (2n = 36) or grassland adaptations of *L. lewisii* (2n = 18). Unfortunately, other wild flaxes may be challenging to incorporate due to chromosome number variation.

# Functional annotation differences in domesticated and wild flax

## Pangenome-wide annotations

Functions that may be linked to agronomic traits integrate SV in the story of flax domestication and diversification. Of over 1,200 assigned functions, many GO terms can be associated with crop domestication (**Table 9**). Moreover, that selection is a conservative representation, omitting over 1,000 annotations for gene transcription and expression regulation. These regulatory functions are undoubtedly intercalated with the biological pathways associated with the traits listed in **Table 9**.

Significant overlap existed in the most frequently annotated functions of the Cultivated and Wild SV sets (**Table 8**). These shared annotations include fundamental cellular processes (protein binding, alkaloid biosynthetic process, translation, **Table 8**) and broad subcellular localizations (membrane, intracellular anatomical structure, **Table 8**). While highly relevant functions to domestication are prevalent across the functional annotations, **Table 9**, like **Table 8**, also describes very generalizable functions across the pangenome. Therefore, it remains unclear whether the functional annotations linked to domestication traits above are truly involved in flax domestication or if these functions are fundamental to the viability of flax at any stage of domestication-cultivation.

For instance, SVs associated with triacylglycerol (triglyceride) lipase activity (**Table 9**) may suggest that these regions diverged during flax evolution and promoted oilseed morphotype in *L. usitatissimum* var CDC Bethune or var Longya. However, outside of agronomic functions, triglycerides are critical energy stores—particularly during seed germination—and necessary components for membrane integrity. The signaling of lipophilic molecules also operates downstream of triglyceride biosynthesis; those molecules include hormones present in many biological pathways, like jasmonic acid (JA), abscisic acid (ABA), and gibberellins (GA). The functional involvement of lipids in plant biology cannot be overstated. Therefore, a triglyceride lipase could relate to a myriad of biological functions and pathways and could be related to non-domestication selection pressures.

In its current state, the flax pangenome offers only a guide towards genomics regions of interest for agronomic trait design and domestication trait identification. Pangenomic SV also fails to

suggest any bias by *L. bienne* towards cultivars of either agricultural function (**Figure 11**). Certainly, the draft pangenome has the potential to extract causal variants relevant for the stem fiber and oilseed morphotypes of flax and to extrapolate the evolutionary processes involved in the functional diversification of *L. usitatissimum*. At present, however, causal variants amongst flax SVs have yet to be determined, and it cannot be ascertained here if the annotations of domestication-associated functions reflect human-driven selection pressures.

The field of pangenomics has not advanced to the point of breeding causal variants in crop lines, only identifying candidate variants [26,34,40,81,94]. As causal functions emerge from flax SVs, the potential of dual-functionality within *L. bienne* should be pursued.

## Functional differences in domesticated and wild flax

The generalizability of functional annotations impairs their usefulness in understanding domestication and informing genomics-assisted breeding. However, using the pangenome as a background to distinguish the SVs in domesticated and wild flax varieties (**Figure 12**) resolves some of the functional annotation ambiguity discussed above. Among the gene ontologies, the SVs in *L. bienne* were enriched with annotations of ceramide metabolic processes, regulation of circadian rhythm, sucrose metabolism, and triacylglycerol lipase activity (Ratios > 0.35, **Figure 12**, **File A1**).

High-ratio annotations regarding oil biosynthesis and metabolism regulation in *L. bienne* relative to *L. usitassimum* cultivars optimistically suggest that SVs have directed changes in flax morphology and development. While there are concerns against overgeneralizing these functional annotations, SVs have likely been involved in the selection pressures upon *L. bienne* and *L. usitatissimum* during domestication.

The draft pangenome currently identifies regions of variability between cultivated and wild flax, but present analyses do not carry decisive conclusions regarding the domestication or agricultural transformation of flax. Commentary on SV functions is also restricted to insertion, tandem duplication, and inversion information. The functional information of the 21,239 deletion SVs is absent, presenting a significant limitation of our annotation methodology. These preliminary conclusions are also contingent upon transcript information from 2012; annotations were derived from short shotgun sequencing data and the 2011 version of the Pfam database [50].

In gene ontology analyses, the robustness of any functional conclusions depends on the annotations' origins; annotation is rooted in protein and gene homology. Generalizable functional annotations at the pangenome level (**Tables 9-10**) and the fundamental biological processes arising in SV set-exclusive annotations (**Figure 12**) suggest that the annotated transcript [50] has limited our functional analyses. Function specificity for flax genomes—and identifying core gene families and single-copy gene variants—may benefit from more extensive and recent annotations, possibly from an alternate crop species or *Arabidopsis thaliana*. Also, annotating deletion SVs would compensate for existing annotation weaknesses to better inform fiber and oilseed trait breeding. As the functional variance between wild and domesticated flax becomes clearer, the enriched SV annotations will become more relevant to the domestication and possible agricultural engineering of flax.

## Final thoughts

Homogeneous agricultural systems exacerbate the risks and inequalities of food insecurity and the loss of diversity among crops. Pushing against that, pangenome construction helps to unveil the evolutionary processes underlying past crop domestication and diversification. Across domesticated and wild species, SV characterization is uncovering allelic variants responsible for multiple agronomic traits and deleterious variants.

With sights turned towards SV-based trait prediction and selection, pangenomes look to accelerate multiple-trait breeding in major and non-major crops. Pangenomes also share the research benefits from publicly updated genome and transcript information, generating highly comprehensive and complete accounts of agronomic trait diversity with reduced need to generate additional sequencing data.

The draft pangenome of *L. usitatissimum* described here contains graph-wide SVs annotated with functions relevant to the domestication and agronomic improvement of the dual-use crop. The structural and allelic diversity from pangenomes constitute critical tools to explore the evolutionary processes responsible for historic and recent crop improvement.

Meeting the agricultural and biodiversity needs of the coming generation hinges upon the conservation and understanding of genetic diversity. Though the field is just emerging, crop pangenomes are aligning to those demands. Accounts of crop genetics are mounting in number and functional detail. A growing consortium of pangenomes, now including flax, offers insights into agronomic phenotype diversity and evolution, with hopes leading towards crops for the future. In the age of pangenomes, trait prediction has never been so accessible and so capable of empowering a resilient and diverse global food system.

# References


1. Antonelli A, Smith R, Fry C, Simmonds MS, Kersey PJ, Pritchard H, et al. State of the World's Plants and Fungi. 2020;
2. Khoury CK, Sotelo SH, Hawtin G, Wibisono J, Amariles D, Guarino L, et al. The Plants That Feed the World: baseline data and metrics to inform strategies for the conservation and use of plant genetic resources for food and agriculture. 2022;
3. Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci. 2014;111(17):6139–46.
4. Pingali PL. Green revolution: impacts, limits, and the path ahead. Proc Natl Acad Sci. 2012;109(31):12302–8.
5. Watson R, Baste I, Larigauderie A, Leadley P, Pascual U, Baptiste B, et al. Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES Secr Bonn Ger. 2019;22–47.
6. Fu YB. The Vulnerability of Plant Genetic Resources Conserved Ex Situ. Crop Sci. 2017;57(5):2314–28.
7. Van de Wouw M, Kik C, van Hintum T, van Treuren R, Visser B. Genetic erosion in crops: concept, research results and challenges. Plant Genet Resour. 2010;8(1):1–15.
8. Crawford CL, Wiebe RA, Yin H, Radeloff VC, Wilcove DS. Biodiversity consequences of cropland abandonment. Nat Sustain. 2024;1–12.
9. Shaw RE, Farquharson KA, Bruford MW, Coates DJ, Elliott CP, Mergeay J, et al. Global meta-analysis shows action is needed to halt genetic diversity loss. Nature. 2025;1–7.
10. Shoham J. Revisiting Seed Company Sales and Profit [Internet]. Food and Agriculture Organization of the United Nations (FAO); 2024 [cited 2025 Apr 17]. Available from: https://openknowledge.fao.org/server/api/core/bitstreams/0535a5cd-2373-414c-8758-2349227dd52e/content
11. Abbo S, van-Oss RP, Gopher A, Saranga Y, Ofner I, Peleg Z. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. Trends Plant Sci. 2014 Jun 1;19(6):351–60.
12. Kumar R, Sharma V, Suresh S, Ramrao DP, Veershetty A, Kumar S, et al. Understanding omics driven plant improvement and de novo crop domestication: some examples. Front Genet. 2021;12:637141.
13. Fu YB. Genetic evidence for early flax domestication with capsular dehiscence. Genet Resour Crop Evol. 2011;58:1119–28.
14. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet. 2018;50(2):278–84.
15. Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E, Walichnowski A, et al. Integrated consensus genetic and physical maps of flax (Linum usitatissimum L.). Theor Appl Genet. 2012 Dec 1;125(8):1783–95.
16. Zhang J, Qi Y, Wang L, Wang L, Yan X, Dang Z, et al. Genomic comparison and population diversity analysis provide insights into the domestication and improvement of flax. Iscience. 2020;23(4).
17. Allaby RG, Peterson GW, Merriwether DA, Fu YB. Evidence of the domestication history of flax (Linum usitatissimum L.) from genetic diversity of the sad2 locus. Theor Appl Genet. 2005;112:58–65.

18. Helback H. Domestication of Food Plants in the Old World: Joint efforts by botanists and archeologists illuminate the obscure history of plant domestication. Science. 1959;130(3372):365–72.
19. Vieira MLC, Santini L, Diniz AL, Munhoz C de F. Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol. 2016;39:312–28.
20. Soto-Cerda BJ, Diederichsen A, Ragupathy R, Cloutier S. Genetic characterization of a core collection of flax (Linum usitatissimum L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. BMC Plant Biol. 2013;13:1–15.
21. Fu YB. Geographic patterns of RAPD variation in cultivated flax. Crop Sci. 2005;45(3):1084–91.
22. Diederichsen A, Kusters PM, Kessler D, Bainas Z, Gugel RK. Assembling a core collection from the flax world collection maintained by Plant Gene Resources of Canada. Genet Resour Crop Evol. 2013;60(4):1479–85.
23. Chandrawati, Singh N, Kumar R, Kumar S, Singh P, Yadav V, et al. Genetic diversity, population structure and association analysis in linseed (Linum usitatissimum L.). Physiol Mol Biol Plants. 2017;23:207–19.
24. Guo D, Jiang H, Yan W, Yang L, Ye J, Wang Y, et al. Resequencing 200 flax cultivated accessions identifies candidate genes related to seed size and weight and reveals signatures of artificial selection. Front Plant Sci. 2020;10:1682.
25. You FM, Xiao J, Li P, Yao Z, Jia G, He L, et al. Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. Int J Mol Sci. 2018;19(8):2303.
26. Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, et al. Graph pangenome captures missing heritability and empowers tomato breeding. Nature. 2022;606(7914):527–34.
27. Maher B. Personal genomes: The case of the missing heritability. 2008;
28. Gabur I, Chawla HS, Snowdon RJ, Parkin IAP. Connecting genome structural variation with complex traits in crop plants. Theor Appl Genet. 2019 Mar 1;132(3):733–50.
29. Yan H, Sun M, Zhang Z, Jin Y, Zhang A, Lin C, et al. Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. Nat Genet. 2023;55(3):507–18.
30. Li N, He Q, Wang J, Wang B, Zhao J, Huang S, et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet. 2023;55(5):852–60.
31. Jin S, Han Z, Hu Y, Si Z, Dai F, He L, et al. Structural variation (SV)-based pan-genome and GWAS reveal the impacts of SVs on the speciation and diversification of allotetraploid cottons. Mol Plant. 2023;16(4):678–93.
32. Duk M, Kanapin A, Samsonova A, Rozhmina T, Samsonova M. Analysis of Structural Variation in Flax (Linum usitatissimum L.) Genomes. Biophysics. 2022;67(2):175–9.
33. Guan J, Xu Y, Yu Y, Fu J, Ren F, Guo J, et al. Genome structure variation analyses of peach reveal population dynamics and a 1.67 Mb causal inversion for fruit shape. Genome Biol. 2021;22:1–25.
34. Liu Z, Wang N, Su Y, Long Q, Peng Y, Shangguan L, et al. Grapevine pangenome facilitates trait genetics and genomic breeding. Nat Genet. 2024;1–11.
35. Goldberg JK, Olcerst A, McKibben M, Hare JD, Barker MS, Bronstein JL. A de novo long-read genome assembly of the sacred datura plant (Datura wrightii) reveals a role of tandem

gene duplications in the evolution of herbivore-defense response. BMC Genomics. 2024;25(1):15.
36. Duitama J. Phased genome assemblies. In: Haplotyping: Methods and Protocols. Springer; 2022. p. 273–86.
37. Andreace F, Lechat P, Dufresne Y, Chikhi R. Comparing methods for constructing and representing human pangenome graphs. Genome Biol. 2023;24(1):274.
38. Benoit M, Jenike KM, Satterlee JW, Ramakrishnan S, Gentile I, Hendelman A, et al. Solanum pan-genetics reveals paralogues as contingencies in crop engineering. Nature. 2025;1–11.
39. Cheng L, Wang N, Bao Z, Zhou Q, Guarracino A, Yang Y, et al. Leveraging a phased pangenome for haplotype design of hybrid potato. Nature. 2025;1–10.
40. Jayakodi M, Lu Q, Pidon H, Rabanus-Wallace MT, Bayer M, Lux T, et al. Structural variation in the pangenome of wild and domesticated barley. Nature. 2024;1–9.
41. Garisto D, Tollefson J. "Totally broken": how Trump 2.0 has paralysed work at US science agencies. Nature.
42. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312–24.
43. You FM, Xiao J, Li P, Yao Z, Jia G, He L, et al. Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. Plant J. 2018;95(2):371–84.
44. Dvorianinova EM, Bolsheva NL, Pushkova EN, Rozhmina TA, Zhuchenko AA, Novakovskiy RO, et al. Isolating Linum usitatissimum L. nuclear DNA enabled assembling high-quality genome. Int J Mol Sci. 2022;23(21):13244.
45. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966–7.
46. Shelton JM, Coleman MC, Herndon N, Lu N, Lam ET, Anantharaman T, et al. Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. BMC Genomics. 2015;16:1–16.
47. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci. 2011;108(4):1513–8.
48. Zeng X, Yi Z, Zhang X, Du Y, Li Y, Zhou Z, et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. Nat Plants. 2024;10(8):1184–200.
49. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. Nat Methods. 2024;1–5.
50. Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, et al. The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. Plant J. 2012;72(3):461–73.
51. Ragupathy R, Rathinavelu R, Cloutier S. Physical mapping and BAC-end sequence analysis provide initial insights into the flax (Linum usitatissimum L.) genome. BMC Genomics. 2011 May 9;12(1):217.
52. Dmitriev AA, Pushkova EN, Novakovskiy RO, Beniaminov AD, Rozhmina TA, Zhuchenko AA, et al. Genome sequencing of fiber flax cultivar atlant using oxford nanopore and illumina platforms. Front Genet. 2021;11:590282.

53. Bonenfant Q, Noé L, Touzet H. Porechop_ABI: discovering unknown adapters in Oxford Nanopore Technology sequencing reads for downstream trimming. Bioinforma Adv. 2023;3(1):vbac085.
54. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
55. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.
56. Rausch T. Dissecting multiple sequence alignment methods: the analysis, design and development of generic multiple sequence alignment components in SeqAn. 2010;
57. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
58. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–54.
59. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142–50.
60. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578–9.
61. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1(1):2047-217X.
62. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv Prepr ArXiv13033997. 2013;
63. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31(12):1119–25.
64. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7.
65. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
66. DARBAJ AK. The US Recognition of Israeli Sovereignty over the Golan Heights. Insight Turk. 2024;26(2):117–38.
67. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. science. 2009;326(5950):289–93.
68. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75.
69. Salter JF, Johnson O, Stafford III NJ, Herrin Jr WF, Schilling D, Cedotal C, et al. A highly contiguous reference genome for Northern Bobwhite (Colinus virginianus). G3 Genes Genomes Genet. 2019;9(12):3929–32.
70. Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26(3):342–50.

71. Yamaguchi K, Kadota M, Nishimura O, Ohishi Y, Naito Y, Kuraku S. Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. Mol Ecol. 2021;30(23):5923–34.
72. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
73. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019;5(8):833–45.
74. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
75. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23(1):258.
76. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39(1):btac808.
77. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3(1):99–101.
78. Chen S, Wang P, Kong W, Chai K, Zhang S, Yu J, et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant Camellia sinensis. Nat Plants. 2023;9(12):1986–99.
79. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol. 2024;42(4):663–73.
80. Marco-Sola S, Eizenga JM, Guarracino A, Paten B, Garrison E, Moreto M. Optimal gap-affine alignment in O (s) space. Bioinformatics. 2023;39(2):btad074.
81. Wang J, Yang W, Zhang S, Hu H, Yuan Y, Dong J, et al. A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biol. 2023;24(1):19.
82. Garrison E, Guarracino A. Unbiased pangenome graphs. Bioinformatics. 2023;39(1):btac743.
83. Computational pan-genomics: status, promises and challenges. Brief Bioinform. 2018;19(1):118–35.
84. Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome graphs. Bioinformatics. 2022;38(13):3319–26.
85. Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. Bioinformatics. 2020;36(22–23):5519–21.
86. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:1–13.
87. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
88. Petr Danecek, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
89. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42(10):1571–80.
90. Chen K, Wang Y, Zhang R, Zhang H, Gao C. CRISPR/Cas Genome Editing and Precision Plant Breeding in Agriculture. Annu Rev Plant Biol. 2019;70(1):667–97.

91. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070–1.
92. Valdés-Florido A, Tan L, Maguilla E, Simón-Porcar VI, Zhou YH, Arroyo J, et al. Drivers of diversification in Linum (Linaceae) by means of chromosome evolution: correlations with biogeography, breeding system and habit. Ann Bot. 2023;132(5):949–62.
93. Abondio P, Cilli E, Luiselli D. Human Pangenomics: Promises and Challenges of a Distributed Genomic Reference. Life. 2023;13(6).
94. Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, et al. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. Nat Commun. 2024;15(1):657.
95. Dey PM, Harborne JB, Bonner J. Plant biochemistry. San Diego: Academic Press; 1997.
96. Mai Z, Kim K, Richardson MB, Deschênes DAR, Garza-Garcia JJO, Shahsavarani M, et al. Oxidation of four monoterpenoid indole alkaloid classes by three cytochrome P450 monooxygenases from Tabernaemontana litoralis. Plant J. 2024;120(6):2770–83.
97. Kulkarni O, Sugier PE, Guibon J, Boland-Augé A, Lonjou C, Bacq-Daian D, et al. Gene network and biological pathways associated with susceptibility to differentiated thyroid carcinoma. Sci Rep. 2021;11(1):8932.
98. Secomandi S, Gallo GR, Sozzoni M, Iannucci A, Galati E, Abueg L, et al. A chromosome-level reference genome and pangenome for barn swallow population genomics. Cell Rep. 2023;42(1).
99. Berdan EL, Barton NH, Butlin R, Charlesworth B, Faria R, Fragata I, et al. How chromosomal inversions reorient the evolutionary process. J Evol Biol. 2023;36(12):1761–82.

# Appendix

*Table A1* HiRise scaffolding reports from Dovetail Genomics for *L. bienne* var. 1_6 and Isr assemblies, using the reference *L. bienne* genome from Zhang et al. [16]

| ***L. Bienne* variety** | | ***1_6*** | | ***ISR*** |
|---|---|---|---|---|
| **Assembly** | Input | **Dovetail HiRise** | Input | **Dovetail HiRise** |
| **Summary** | | | | |
| ***Total length (bp)*** | 293,719,246 | **293,736,946** | 293,718,681 | **293,749,081** |
| ***N50*** | 2,936,960 | **57,144,684** | 1,744,805 | **52,558,052** |
| ***L50*** | 30 | **3** | 49 | **2** |
| ***N90*** | 733,753 | **37,586,277** | 429,599 | **22,308,750** |
| ***L90*** | 99 | **5** | 169 | **5** |
| **Continuity** | | | | |
| ***Largest scaffold*** | 10,579,187 | **83,557,733** | 6,357,957 | **109,628,437** |
| ***Scaffold number*** | 1,131 | **961** | 1,290 | **1,005** |
| ***Scaffold (> 1kb) number*** | 1,086 | **916** | 1,242 | **957** |
| ***Number of gaps*** | 5,547 | **5,724** | 5,560 | **5,864** |
| ***Uncalled bases per 100 kb*** | 1,980.74 | **1,986.65** | 1,981.05 | **1,991.20** |

*Table A2* BUSCO statistics for the HiRise assemblies of *L. bienne var. 1 6* and *Isr* samples, scaffolded with Dovetail HiC and reference sequences from the *de novo L. bienne* genome of Zhang et al. [16] In the eukaryota_obd10 lineage dataset, 255 BUSCOs were searched.

| ***L. bienne* variety** | **Assembly** | ***BUSCO*** | | | |
|---|---|---|---|---|---|
| | | *Single* | *Duplicated* | *Fragmented* | *Missing* |
| **1_6** | **Input** | 63.14 | 32.16 | 3.14 | 1.57 |
| | **Dovetail HiRise** | 78.04 | 17.65 | 3.14 | 1.18 |
| **Isr** | **Input** | 57.65 | 38.04 | 2.35 | 1.96 |
| | **Dovetail HiRise** | 75.69 | 19.61 | 3.53 | 1.18 |

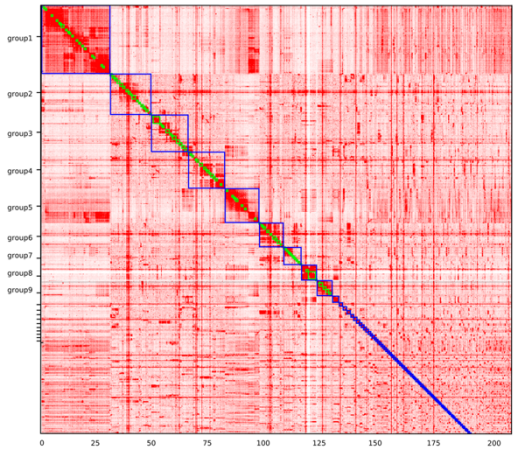



*Figure A1* Based on **HapHiC** HiC alignments, Juicebox manual curation produced a poorer quality interaction heatmap than the **HapHiC plot** function. Scaffold name labels were added based on Figure 6. Although scaffolds < 1 Mb are mapped, as scaffolds become progressively smaller approaching the bottom-right corner of the figure, no scaffold smaller than *group9* is labeled. Chromosome annotations, defined in the **HapHiC** assembly output, are made in blue; scaffold annotations, associated with the initial assembly scaffolds, are made in green.

*Table A3* The number of poorly aligned sequences (contigs or scaffolds) concatenated into Chromosome 0, by **RagTag (-C)**, is reported for each assembly. The total Chromosome 0 length (Mb) and proportion (%) of Chromosome 0-inclusive genome length occupied by Chromosome 0 are also reported.

| *Assembly* | *Contained sequences* | *Length (Mb)* | *Proportion of assembly (%)* |
|---|---|---|---|
| **Longya** | 1687 | 5.8 | 1.89 |
| ***L. bienne*** | 2553 | 30.1 | 10.28 |
| **Heiya** | 3703 | 18.7 | 6.15 |
| **Atlant** | 2621 | 104.5 | 24.07 |

*Script A1* A sample Linux (Bash) script used to rename PGGB and SVIM-asm VCF outputs in the PanSN format.

```
# This is a Bash script. For those running the script in a SLURM-based
HPC environment, these were the job resource requests. Those not
executing a SLURM job should skip to 'Renaming SVIM-asm VCF'.

#!/bin/bash

#SBATCH --job-name={}
#SBATCH -p shared
#SBATCH --mem=10G
#SBATCH -c 4
#SBATCH -t 1:59:59
#SBATCH --mail-type=ALL

# Renaming SVIM-asm VCF so chromosome names match PanSN format
## Note that {$Haplotype_ID} should be a number, 1 is acceptable for
haploid assemblies. Repeat "-e ..." for as many chromosomes as you have
```

```
sed -e
's/{$ORIGINAL_CHROMOSOME_NAME_STRING}/{$VARIETAL}#{$Haplotype_ID}#{$con
tig1/g}'
'{$input_svim-asm_file.vcf}' > '{$renamed_input_svim-asm_file.vcf'
## Repeat for as many SVIM-asm VCFs as you need to process

# Renaming PGGB VCF so chromosome names match PanSN format
## Define PGGB VCF file inputs (a single VCF file is produced for every
defined reference sequence in the PGGB --vcf-spec command; defined all
the VCFs output)
vcf_files=("{$pggb_vcf1.vcf}" "{$pggb_vcf2.vcf}")
# Define output file
output_file="{$renamed_pggb_concatenated.vcf}"
# Print the first 10 lines of the first VCF file with ##info
head -n 10 "${vcf_files [0]}" >> "$output_file"
# Print the contig ids from each VCF file (chromosome names)
for vcf in "${vcf_files [@]}"; do
  sed -n '11p' "$vcf" >> "$output_file"
done
# Print the column header line from the first VCF file
sed -n '12p' "${vcf_files [0]}" >> "$output_file"
# Print all lines from the 13th line onwards for each VCF file
for vcf in "${vcf_files [@]}"; do
  tail -n +13 "$vcf" >> "$output_file"
done

# All PanSN renaming is done. And we've concatenated the (many) PGGB
VCF outputs into 1 file.
```

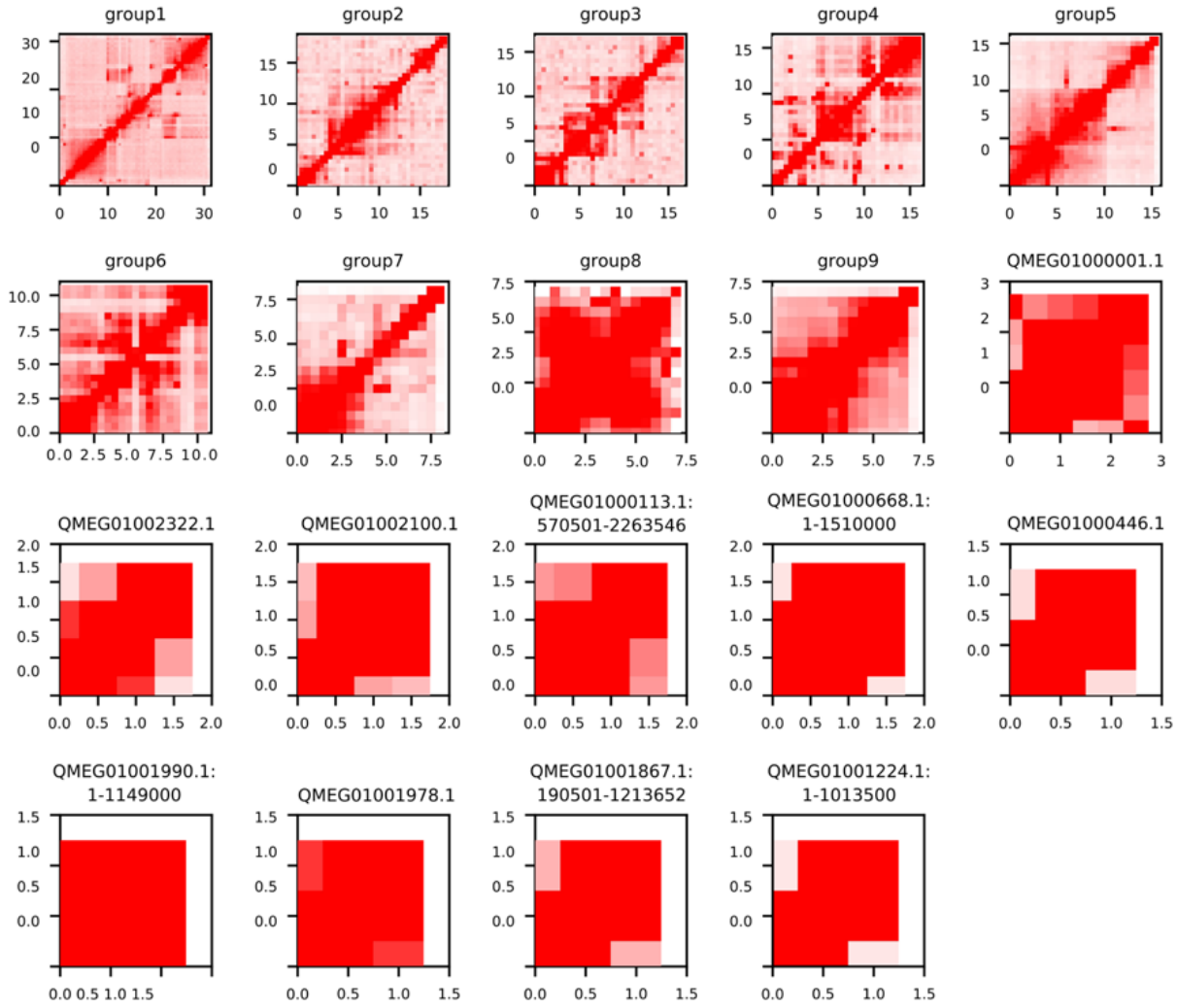
*Table A4* The SV calls (> 50 bp) by the long-read SV caller, **Sniffles** is reported by varietal.

| ***Genome assembly*** | **Longya** | **Bienne** | **Heiya** | **Atlant** |
|---|---|---|---|---|
| **Total SV calls** | 484 | 712 | 410 | 408 |
| **Deletions** | 30 | 87 | 61 | 45 |
| **Insertions** | 484 | 625 | 348 | 363 |
| **Inversions** | 0 | 0 | 1 | 0 |
| **Duplication** | 0 | 0 | 0 | 0 |
| **Breakend** | 0 | 0 | 0 | 0 |

*Table A5* Selected RagTag assembly statistics for the final genome assemblies of *L. usitatissimum* Atlant, Heiya, Longya, and *L. bienne*

| Listed output file | Assembly statistic | CDC Bethune | Atlant | Bienne | Heiya | Longya |
|---|---|---|---|---|---|---|
| **genome_stats/ genome_info.txt** | **Genome fraction** | NA | 95.096 | 78.829 | 93.633 | 93.496 |
| | **Duplication ratio** | NA | 1.060 | 1.029 | 1.017 | 1.017 |
| | **Gaps number** | NA | | | | |
| **report.txt** | **Contigs** | | | | | |
| | ≥ *0 bp* | 15 | 1326 | 1292 | 1867 | 859 |
| | ≥ *1000 bp* | 15 | 1326 | 1260 | 1839 | 842 |
| | ≥ *5000 bp* | 15 | 1326 | 401 | 534 | 259 |
| | ≥ *10000 bp* | 15 | 1326 | 279 | 282 | 136 |
| | ≥ *25000 bp* | 15 | 1326 | 214 | 118 | 55 |
| | ≥ *50000 bp* | 15 | 1326 | 169 | 76 | 35 |
| | **Length** | | | | | |
| | ≥ *0 bp* | 316167074 | 434178430 | 293725949 | 303775269 | 306453633 |
| | ≥ *1000 bp* | 316167074 | 434178430 | 293695390 | 303748136 | 306437269 |
| | ≥ *5000 bp* | 316167074 | 434178430 | 291904607 | 301277600 | 305310171 |
| | ≥ *10000 bp* | 316167074 | 434178430 | 291039931 | 299545186 | 304380547 |
| | ≥ *25000 bp* | 316167074 | 434178430 | 289973971 | 297129833 | 303159979 |
| | ≥ *50000 bp* | 316167074 | 434178430 | 288314398 | 295588744 | 302447177 |
| | **Number of contigs** | 15 | 1319 | 573 | 713 | 349 |
| | **Largest contig** | 29425369 | 34435688 | 23294354 | 25361847 | 24566470 |
| | **Total length** | | 434178430 | 292571030 | 301995403 | 305673423 |
| | **GC** | 39.49 | 38.28 | 38.94 | 38.95 | 39.06 |
| | **Reference GC** | NA | 39.49 | 39.49 | 39.49 | 39.49 |
| | **N50** | 20483506 | 19813820 | 16810307 | 17737391 | 20038188 |
| | **N90** | 17699753 | 82992 | 9841106 | 14173484 | 17504909 |
| | **L50** | 7 | 9 | 8 | 8 | 7 |
| | **L90** | 14 | 416 | 15 | 15 | 14 |
| | **Number of misassemblies** | NA | 2412 | 3739 | 1532 | 2139 |
| | **Misassembled contigs** | NA | 85 | 201 | 55 | 27 |
| | **Misassembled contig length** | NA | 336223124 | 285041767 | 289608710 | 301259227 |

*Figure A2* Hi-C interactions heatmaps for the individual scaffolds (> 1Mb) of L. bienne produced by HapHiC




alignment software. Scaffold names are identified by "groupX" when chromosome grouping was possible, and by the alignment read name (e.g. QMEG…) for smaller scaffolds. The scaffolds have been reoriented to begin at the bottom left corner, where they corresponding scaffold in Figure R2 begins at the top left corner

*Table A6* BUSCO assessments of the final assemblies (Atlant, Heiya, L. bienne, Longya, and the reference assembly CDC Bethune), using the eudicots_odb10 lineage (2326 BUSCOs assessed). The present complete (single and duplicated BUSCOs), single, duplicated, fragmented, and missing BUSCOs are listed as %.

| **Assembly** | ***BUSCO*** | | | | |
|---|---|---|---|---|---|
| | *Complete (S/D)* | *Single* | *Duplicated* | *Fragmented* | *Missing* |
| **CDC Bethune** | 94.15 | 27.60 | 66.55 | 1.12 | 4.72 |
| **Longya** | 94.93 | 26.44 | 68.49 | 1.03 | 4.04 |
| ***L. bienne*** | 93.98 | 42.39 | 51.59 | 1.33 | 4.69 |
| **Heiya** | 95.18 | 25.58 | 69.60 | 0.99 | 3.83 |
| **Atlant** | 95.27 | 22.74 | 72.53 | 0.86 | 3.87 |

*Script A2* An R script used to validate the SV calls of a PGGB VCF using the calls of SVIM-asm VCFs.

```
# This is an R script. If you're running a SLURM job, you must set up
your bash script appropriately for your system to run RStudio-
compatible packages and libaries

# Packages and Libraries
install.packages("dplyr")
install.packages("vcfR")
install.packages("stringr")
install.packages("tidyr")
install.packages("ggplot2")
library(ggplot2)
library(dplyr)
library(vcfR)
library(stringr)
library(tidyr)

##  Analysis: SV call overlap between the PGGB VCF and the SVIM-asm
VCFs for each varietal
# VCF data loading
pggb.vcf <- read.vcfR('{$renamed_pggb_concatenated.vcf') #I recommend
you define full filepaths
svim1.vcf <- read.vcfR('renamed_input_svim-asm_file.vcf') #repeat for
as many SVIM-asm VCFs you need to validate against

#  VCF data formatting
pggb_gt_matrix <- extract.gt(pggb.vcf, return.alleles = FALSE)
pggb.df <- as.data.frame(pggb.vcf@fix)
pggb.df <- cbind(pggb.df, pggb_gt_matrix)
pggb.df$POS <- as.numeric(as.character(pggb.df$POS))

svim1.df <- as.data.frame(svim1.vcf@fix)
svim1.df$POS <- as.numeric(as.character(svim1.df$POS))
svim1.df <- svim1.df %>%
  rename(POS2 = POS)

# Function to detect near-equivalent POS values in PGGB calls (POS) and
SVIM calls (POS2). Tolerated difference in position is current set to
30 (base pairs) for my dataset, you may need to adjust.
find_matching_svim1pos2 <- function(chrom, pos, svim1.df) {
  svim1pos2_values <- svim1.df %>%
    filter(abs(POS2 - pos) <= 30 & CHROM == chrom) %>%
    pull(POS2)
  return(ifelse(length(svim1pos2_values) > 0, paste(svim1pos2_values,
collapse = ","), NA))
}
svim1_matching_svcalls <- pggb.df %>%
  rowwise() %>%
  filter(any(abs(svim1.df$POS2 - POS) <= 30 & svim1.df$CHROM ==
CHROM)) %>%
  mutate(POS2 = find_matching_svim1pos2(CHROM, POS, svim1.df))

## Results: SV call overlap between PGGB and SVIM for each varietal
print(c("SVIM-asm1 Matching SV Calls:",nrow(svim1_matching_svcalls)))

## Analysis: Kruskal-Wallis Statistics
# data formatting
```

```
svim1.df$POS2 <- as.numeric(as.character(svim1.df$POS2))
svim1_matching_svcalls$POS2 <-
as.numeric(as.character(svim1_matching_svcalls$POS2))
svim1_highconfidence_svcalls.df <- svim1_matching_svcalls %>%
  left_join(svim1.df, by = c("CHROM", "POS2"))
#colnames(svim1_highconfidence_svcalls.df)
svim1_highconfidence_svcalls.df <- svim1_highconfidence_svcalls.df %>%
  select("CHROM", "POS", "ID.y", "REF.x", "ALT.x", "QUAL.x",
"FILTER.y", "INFO.y", "SVIM1") #Add additional columns for other SVIM-
asm VCF names
colnames(svim1_highconfidence_svcalls.df) <- c("CHROM", "POS", "ID",
"REF", "ALT", "QUAL", "FILTER", "INFO", "SVIM1") #Add additional
columns for other SVIM-asm VCF names
svim1_highconfidence_svcalls.df$SVTYPE <-
str_extract(svim1_highconfidence_svcalls.df$ID, "(?<=\\.) [A-
Z]{3}(?=\\.)")
svim1_highconfidence_svcalls.df$SVLEN <-
str_extract(svim1_highconfidence_svcalls.df$INFO, "(?<=SVLEN=) [-0-
9]+")
svim1_highconfidence_svcalls.df$SVLEN <-
as.numeric(as.character(svim1_highconfidence_svcalls.df$SVLEN))

#combine genome-level dataframes w PGGB-validated validated
highconfidence_svcalls.df <- rbind(SVIM1_highconfidence_svcalls.df,
{$all_other_SVIM-asm_VCF_highconfidence_svcalls.df})
highconfidence_svcalls.df_long <-
pivot_longer(highconfidence_svcalls.df, cols = c("SVIM1", {$other_SVIM-
asm VCF names}), names_to = "Location", values_to = "Presence")
highconfidence_svcalls.df_long <- highconfidence_svcalls.df_long
[!is.na(highconfidence_svcalls.df_long$Presence), ]
highconfidence_svcalls.df_long$Presence <-
as.numeric(as.character(highconfidence_svcalls.df_long$Presence))

## Mini-Results: General SV Call Stats
SV_presence_count <- highconfidence_svcalls.df_long %>%
  group_by(Location) %>%
  summarize(SV_presence = sum(Presence))
print(SV_presence_count)
SV_type_count <- highconfidence_svcalls.df_long %>%
  count(SVTYPE)
print(SV_type_count)


## Results: Kruskal-Wallis tests
print("KW Results by SVTYPE")
svtype_kruskal_results <- kruskal.test(SVLEN ~ SVTYPE, data =
highconfidence_svcalls.df_long)
print(svtype_kruskal_results)

print("KW Results by Location")
kruskal_results <- kruskal.test(SVLEN ~ Location, data =
highconfidence_svcalls.df_long)
print(kruskal_results)

vcf_INSdf_long <- highconfidence_svcalls.df_long %>%
  filter(SVTYPE == "INS")
print("KW Results by Location, for Insertions")
```

```
insertion_kruskal_results <- kruskal.test(SVLEN ~ Location, data =
vcf_INSdf_long)
print(insertion_kruskal_results)

vcf_DELdf_long <- highconfidence_svcalls.df_long %>%
  filter(SVTYPE == "DEL")
print("KW Results by Location, for Deletions")
deletion_kruskal_results <- kruskal.test(SVLEN ~ Location, data =
vcf_DELdf_long)
print(deletion_kruskal_results)

#the Lengths for Inversions, Tandem Duplications, Tandem Insertions,
and Breakend points are not directly provided by PGGB or SVIM-asm
outputs. Something like the script below, however, may work to get
Inversions, Tandem Duplications, and Tandem Insertions. Breakend points
have no length to extract.
#highconfidence_svcalls.df$POS <-
as.numeric(as.character(highconfidence_svcalls.df$POS))
#highconfidence_svcalls.df $SVLEN <-
str_extract(highconfidence_svcalls.df $INFO, "(?<=SVLEN=) [-0-9]+")
#highconfidence_svcalls.df $SVLEN <-
as.numeric(as.character(highconfidence_svcalls.df $SVLEN))
#highconfidence_svcalls.df $SVLEN <- abs(highconfidence_svcalls.df
$SVLEN)
#highconfidence_svcalls.df <- highconfidence_svcalls.df %>%
#  mutate(END = as.numeric(str_extract(INFO, "(?<=END=)\\d+"))) %>%
#  mutate(ABS_LEN = abs(END - POS))
#lengthsupplemented_highconfidence_svcalls.df <-
highconfidence_svcalls.df %>%
#  mutate(SVLEN = ifelse(is.na(SVLEN), ABS_LEN, SVLEN))
```

*Table A7* Redundant, all-inclusive count of SVs in the pangenome. For example, should a SV be detected in both *L. usitatissimum* var Atlant and Heiya, the SV is counted twice.

| ***Chromosome*** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **Total SVs** | 20592 | 17996 | 19624 | 14340 | 8100 | 13152 |
| **Deletions** | 7720 | 7280 | 7296 | 4948 | 152 | 4592 |
| **Insertions** | 12708 | 10584 | 12012 | 9136 | 7796 | 8356 |
| **Inversions** | 4 | 8 | 8 | 0 | 0 | 8 |
| **Breakend points** | 160 | 124 | 308 | 256 | 152 | 196 |

*Table A7 Continued*

| 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|
| **12944** | 15752 | 15848 | 15584 | 13796 | 15340 | 15676 | 13512 | 10796 |
| **4952** | 5440 | 5888 | 6344 | 5036 | 5268 | 6064 | 4880 | 3960 |
| **7884** | 10184 | 9896 | 9128 | 8472 | 9920 | 9420 | 8388 | 6668 |
| **0** | 4 | 0 | 4 | 20 | 0 | 20 | 4 | 8 |
| **108** | 124 | 64 | 108 | 268 | 152 | 172 | 240 | 160 |

*Script A3* An R script to estimate the pangenome length contributions from Insertions and Inversions, building upon the reference *L. usitatissimum* var CDC Bethune. Pangenome size estimates are based on all pairwise genome comparisons and combinations (e.g. single, double, triple genome comparisons).

```
library(readr)
library(tidyr)
library(dplyr)
library(vcfR)
library(stringr)
library(ggplot2)
library(reshape2)
library(ggridges)
library(forcats)
library(splines)
library(scales)

# data loading
vcfsample <- read.vcfR('~/Users/svdataset.vcf')
vcf_pgest_df <- vcfsample@fix %>%
  as.data.frame()

# formatting into dataframe
generate_unique_ids <- function(ids) {
  unique_ids <- ids
  id_count <- table(ids)
  for (id in names(id_count [id_count > 1])) {
    idx <- which(ids == id)
    unique_ids [idx] <- paste(id, seq_along(idx), sep = "_")
  }
  return(unique_ids)
}
vcf_pgest_df$ID <- generate_unique_ids(vcf_pgest_df$ID)
vcfsample@fix <- as.matrix(vcf_pgest_df) # update the VCF file with
modified IDs
gt_matrix <- extract.gt(vcfsample, return.alleles = FALSE)

#extracting SV information
vcf_pgest_df <- cbind(vcf_pgest_df, gt_matrix)
vcf_pgest_df$SVTYPE <- str_extract(vcf_pgest_df$ID, "(?<=\\.) [A-
Z]{3}(?=\\.)")
vcf_pgest_df$POS <- as.numeric(as.character(vcf_pgest_df$POS))
vcf_pgest_df$SVLEN <- str_extract(vcf_pgest_df$INFO, "(?<=SVLEN=) [-0-
9]+")
```

```
vcf_pgest_df$SVLEN <- as.numeric(as.character(vcf_pgest_df$SVLEN))
View(vcf_pgest_df)
vcf_pgest_df <- vcf_pgest_df %>%
  mutate(END = as.numeric(str_extract(INFO, "(?<=END=)\\d+"))) %>%
  mutate(ABS_LEN = abs(END - POS))
supplemented_vcf_pgest_df <- vcf_pgest_df %>%
  mutate(SVLEN = ifelse(is.na(SVLEN), ABS_LEN, SVLEN))
#View(supplemented_vcf_pgest_df)

#pivot out different sample varietals (in our case, named Atlant,
Bienne, Heiya, and Longya)
vcf_pgest_df_long <- pivot_longer(supplemented_vcf_pgest_df, cols =
c("Atlant", "Bienne", "Heiya", "Longya"), names_to = "Location",
values_to = "Presence")
vcf_pgest_df_long <- vcf_pgest_df_long
[!is.na(vcf_pgest_df_long$Presence), ]
vcf_pgest_df_long <- vcf_pgest_df_long %>%
  filter(SVTYPE %in% c("DEL", "INS", "INV"))
#View(vcf_pgest_df_long)

# Get SVLEN totals for single genome additions
# 1 Genome: Atlant
g1_atlant_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  summarise(total_g1_atlant_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g1_atlant_SVLEN = sum(!is.na(SVLEN)))  # Summarise
SVLEN, ignoring NA
g1_atlant_svlen_sum <- g1_atlant_svlen$total_g1_atlant_SVLEN
g1_atlant_svlen_count <- g1_atlant_svlen$count_g1_atlant_SVLEN
print(c("Atlant:", g1_atlant_svlen_count, "SVs", g1_atlant_svlen_sum,
"bp"))
# 1 Genome: Bienne
g1_b_svlen <- supplemented_vcf_pgest_df %>%
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g1_b_SVLEN = sum(SVLEN, na.rm = TRUE),
            count_g1_b_SVLEN = sum(!is.na(SVLEN)))
g1_b_svlen_sum <- g1_b_svlen$total_g1_b_SVLEN
g1_b_svlen_count <- g1_b_svlen$count_g1_b_SVLEN
print(c("Bienne:", g1_b_svlen_count, "SVs", g1_b_svlen_sum, "bp"))
# 1 Genome: Heiya
g1_h_svlen <- supplemented_vcf_pgest_df %>%
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g1_h_SVLEN = sum(SVLEN, na.rm = TRUE),
            count_g1_h_SVLEN = sum(!is.na(SVLEN)))
g1_h_svlen_sum <- g1_h_svlen$total_g1_h_SVLEN
g1_h_svlen_count <- g1_h_svlen$count_g1_h_SVLEN
print(c("Heiya:", g1_h_svlen_count, "SVs", g1_h_svlen_sum, "bp"))
# 1 Genome: Longya
g1_l_svlen <- supplemented_vcf_pgest_df %>%
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g1_l_SVLEN = sum(SVLEN, na.rm = TRUE),
            count_g1_l_SVLEN = sum(!is.na(SVLEN)))
g1_l_svlen_sum <- g1_l_svlen$total_g1_l_SVLEN
g1_l_svlen_count <- g1_l_svlen$count_g1_l_SVLEN
print(c("Longya:", g1_l_svlen_count, "SVs", g1_l_svlen_sum, "bp"))
```

```
# pairwise genome length comparisons
# 2 Genomes: Atlant-Bienne
g2_ab_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g2_ab_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g2_ab_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g2_ab_svlen_sum <- g2_ab_svlen$total_g2_ab_SVLEN
g2_ab_svlen_count <- g2_ab_svlen$count_g2_ab_SVLEN
print(c("Atlant-Bienne:", g2_ab_svlen_count, "SVs", g2_ab_svlen_sum,
"bp"))
# 2 Genomes: Atlant-Heiya
g2_ah_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g2_ah_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g2_ah_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g2_ah_svlen_sum <- g2_ah_svlen$total_g2_ah_SVLEN
g2_ah_svlen_count <- g2_ah_svlen$count_g2_ah_SVLEN
print(c("Atlant-Heiya:", g2_ah_svlen_count, "SVs", g2_ah_svlen_sum,
"bp"))
# 2 Genomes: Atlant-Longya
g2_al_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g2_al_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g2_al_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g2_al_svlen_sum <- g2_al_svlen$total_g2_al_SVLEN
g2_al_svlen_count <- g2_al_svlen$count_g2_al_SVLEN
print(c("Atlant-Longya:", g2_al_svlen_count, "SVs", g2_al_svlen_sum,
"bp"))
# 2 Genomes: Bienne-Heiya
g2_bh_svlen <- supplemented_vcf_pgest_df %>%
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g2_bh_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g2_bh_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g2_bh_svlen_sum <- g2_bh_svlen$total_g2_bh_SVLEN
g2_bh_svlen_count <- g2_bh_svlen$count_g2_bh_SVLEN
print(c("Bienne-Heiya:", g2_bh_svlen_count, "SVs", g2_bh_svlen_sum,
"bp"))
# 2 Genomes: Bienne-Longya
g2_bl_svlen <- supplemented_vcf_pgest_df %>%
```

```
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g2_bl_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g2_bl_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g2_bl_svlen_sum <- g2_bl_svlen$total_g2_bl_SVLEN
g2_bl_svlen_count <- g2_bl_svlen$count_g2_bl_SVLEN
print(c("Bienne-Longya:", g2_bl_svlen_count, "SVs", g2_bl_svlen_sum,
"bp"))
# 2 Genomes: Heiya-Longya
g2_hl_svlen <- supplemented_vcf_pgest_df %>%
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g2_hl_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g2_hl_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g2_hl_svlen_sum <- g2_hl_svlen$total_g2_hl_SVLEN
g2_hl_svlen_count <- g2_hl_svlen$count_g2_hl_SVLEN
print(c("Heiya-Longya:", g2_hl_svlen_count, "SVs", g2_hl_svlen_sum,
"bp"))

# three-way genome length comparisons
# 3 Genomes: Atlant-Bienne-Heiya
g3_abh_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g3_abh_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g3_abh_SVLEN = sum(!is.na(SVLEN)))  # Summarise
SVLEN, ignoring NA
g3_abh_svlen_sum <- g3_abh_svlen$total_g3_abh_SVLEN
g3_abh_svlen_count <- g3_abh_svlen$count_g3_abh_SVLEN
print(c("Atlant-Bienne-Heiya:", g3_abh_svlen_count, "SVs",
g3_abh_svlen_sum, "bp"))
# 3 Genomes: Bienne-Heiya-Longya
g3_bhl_svlen <- supplemented_vcf_pgest_df %>%
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g3_bhl_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g3_bhl_SVLEN = sum(!is.na(SVLEN)))  # Summarise
SVLEN, ignoring NA
g3_bhl_svlen_sum <- g3_bhl_svlen$total_g3_bhl_SVLEN
g3_bhl_svlen_count <- g3_bhl_svlen$count_g3_bhl_SVLEN
print(c("Bienne-Heiya-Longya:", g3_bhl_svlen_count, "SVs",
g3_bhl_svlen_sum, "bp"))
# 3 Genomes: Atlant-Heiya-Longya
g3_ahl_svlen <- supplemented_vcf_pgest_df %>%
```

```
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g3_ahl_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g3_ahl_SVLEN = sum(!is.na(SVLEN)))  # Summarise
SVLEN, ignoring NA
g3_ahl_svlen_sum <- g3_ahl_svlen$total_g3_ahl_SVLEN
g3_ahl_svlen_count <- g3_ahl_svlen$count_g3_ahl_SVLEN
print(c("Atlant-Heiya-Longya:", g3_ahl_svlen_count, "SVs",
g3_ahl_svlen_sum, "bp"))
# 3 Genomes: Atlant-Bienne-Longya
g3_abl_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g3_abl_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of
SVLEN, ignoring NA
            count_g3_abl_SVLEN = sum(!is.na(SVLEN)))  # Summarise
SVLEN, ignoring NA
g3_abl_svlen_sum <- g3_abl_svlen$total_g3_abl_SVLEN
g3_abl_svlen_count <- g3_abl_svlen$count_g3_abl_SVLEN
print(c("Atlant-Bienne-Longya:", g3_abl_svlen_count, "SVs",
g3_abl_svlen_sum, "bp"))

# SV lengths for SVs shared across all genome
# 4 Genomes
g4_svlen <- supplemented_vcf_pgest_df %>%
  filter(Atlant %in% c("0", "1", "2", "3", "4")) %>%   # Filter rows
where Atlant is not NA
  filter(Bienne %in% c("0", "1", "2", "3", "4")) %>%
  filter(Heiya %in% c("0", "1", "2", "3", "4")) %>%
  filter(Longya %in% c("0", "1", "2", "3", "4")) %>%
  summarise(total_g4_SVLEN = sum(SVLEN, na.rm = TRUE),  # Sum of SVLEN,
ignoring NA
            count_g4_SVLEN = sum(!is.na(SVLEN)))  # Summarise SVLEN,
ignoring NA
g4_svlen_sum <- g4_svlen$total_g4_SVLEN
g4_svlen_count <- g4_svlen$count_g4_SVLEN
print(c("Atlant-Bienne-Heiya-Longya:", g4_svlen_count, "SVs",
g4_svlen_sum, "bp"))

# Create pangenome length dataframe
pangenome_genome_number <- c(0,1,1,1,1,2,2,2,2,2,2,3,3,3,3,4) # number
of comparisions for 4 samples
sv_len <- c(316167074, g1_atlant_svlen_sum, g1_b_svlen_sum,
g1_h_svlen_sum, g1_l_svlen_sum,
            g2_ab_svlen_sum, g2_ah_svlen_sum, g2_al_svlen_sum,
g2_bh_svlen_sum, g2_bl_svlen_sum, g2_hl_svlen_sum,
            g3_abh_svlen_sum, g3_bhl_svlen_sum, g3_ahl_svlen_sum,
g3_abl_svlen_sum,
            g4_svlen_sum)
pangenome_genome_idcode <- c('bethune', 'a','b','h','l',
                             'ab','ah','al','bh','bl','hl',
                             'abh','bhl','ahl','abl',
```

```
                            'abhl')
pangenome_length <- c(316167074, (316167074+g1_atlant_svlen_sum),
(316167074+g1_b_svlen_sum), (316167074+g1_h_svlen_sum),
(316167074+g1_l_svlen_sum),
                      (316167074+g2_ab_svlen_sum),
(316167074+g2_ah_svlen_sum), (316167074+g2_al_svlen_sum),
(316167074+g2_bh_svlen_sum), (316167074+g2_bl_svlen_sum),
(316167074+g2_hl_svlen_sum),
                      (316167074+g3_abh_svlen_sum),
(316167074+g3_bhl_svlen_sum), (316167074+g3_ahl_svlen_sum),
(316167074+g3_abl_svlen_sum),
                      (316167074+g4_svlen_sum))

pangenome_svlen_df <- cbind(pangenome_genome_number, sv_len,
pangenome_genome_idcode, pangenome_length)
#View(pangenome_svlen_df)
pangenome_svlen_df <- as.data.frame(pangenome_svlen_df)
pangenome_svlen_df$pangenome_genome_number <-
as.numeric(pangenome_svlen_df$pangenome_genome_number)
pangenome_svlen_df$sv_len <- as.numeric(pangenome_svlen_df$sv_len)
pangenome_svlen_df$pangenome_length <-
as.numeric(pangenome_svlen_df$pangenome_length)


# Fit a natural spline model using `lm` and `ns()`
spline_model <- lm(pangenome_length ~ ns(pangenome_genome_number, df =
4), data = pangenome_svlen_df)
print(summary(spline_model))

# Generate predictions for plotting the spline curve
predicted <- data.frame(
  pangenome_genome_number =
seq(min(pangenome_svlen_df$pangenome_genome_number),
max(pangenome_svlen_df$pangenome_genome_number), length.out = 100)
)
predicted$pangenome_length <- predict(spline_model, newdata =
predicted)

# Plot the data points and the spline curve
svlen_point <- ggplot(pangenome_svlen_df, aes(x =
pangenome_genome_number, y = pangenome_length / 1e6)) +
  geom_line(data = predicted, aes(x = pangenome_genome_number, y =
pangenome_length / 1e6), color = "#90b648", size = 1)+
  geom_point(size = 3, alpha = 0.6) +
  labs(
    x = "Genome Number") +
  scale_y_continuous(
    name = "Pangenome Length (Mb)",  # Rename the y-axis
    limits = c(316167000 / 1e6, 316210000 / 1e6),
    labels = label_number(scale = 1, suffix = " Mb")  # Add "Mb" to the
labels
  ) +
  theme_minimal()
svlen_point

ggsave(pangenome_len_curve,
```

```
       filename =
"/Users/esmepadgett/mbiolathome/pangenome_len_curve.pdf",
       device = "pdf",
       height = 4, width = 8, units = "in")
```

*Script A4* An R script to generate the Eigenvalues, scree plot, and PCA for the SV calls from a VCF. File

```
library(adegenet)
library(poppr)
library(dplyr)
library(reshape2)
library(ggplot2)
library(RColorBrewer)
library(scales)
library(vcfR)

# Import vcf file
myvcf_file = '{yourfile.vcf}'
mysample.vcf = read.vcfR(myvcf_file, verbose = FALSE)
vcf_df <- as.data.frame(mysample.vcf@fix, stringsAsFactors = FALSE) #
replace "." with "_" in columns 1 (CHROM) and 3 (ID)

# Replace any abnormal text string characters
vcf_df$CHROM <- gsub("\\#", "_", vcf_df$CHROM)
vcf_df$ID <- gsub("\\.", "_", vcf_df$ID)
#make column names unique
extracted_part <- sapply(strsplit(vcf_df$CHROM, "_"), function(x) tail(x,
1))
vcf_df$ID <- paste(extracted_part, vcf_df$ID, sep = "_")
vcf_df$ID <- make.unique(vcf_df$ID)
vcf_df$ID <- gsub("\\.", "_", vcf_df$ID)

# Format file for R to read
mysample.vcf@fix <- as.matrix(vcf_df)
mysample.light <- vcfR2genind(mysample.vcf) #Error in extract.gt(x,
return.alleles = return.alleles) : ID column contains non-unique names
x = tab(mysample.light, NA.method = NULL)
x [is.na(x)] <- 0

# Perform PCA
pca1 = dudi.pca(x, scannf = FALSE, scale = FALSE, nf = 3)
percent = pca1$eig/sum(pca1$eig)*100

png("/Users/esmepadgett/mbiolathome/Eigenplot.png", width = 400, height =
300)
Eigenplot <- barplot(percent, ylab = "Genetic variance explained by
eigenvectors (%)", ylim = c(0,100),
        names.arg = round(percent, 1))
dev.off()

## Visualize PCA results
# Create a data.frame containing individual coordinates
ind_coords = as.data.frame(pca1$li)
# Rename columns of dataframe
colnames(ind_coords) = c("Axis1","Axis2","Axis3")
# Add a column containing individuals
ind_coords$Ind = indNames(mysample.light)
```

```
# Define colour palette, in accordance with how many samples you have
   cols = c("#332288", "#88CCEE", "#44AA99", "#117733", "#999933", "#DDCC77",
   "#661100", "#CC6677", "#882255",
   "#AA4499","#DDDDDD","#555555")#tol10qualitative + two

   #cols = brewer.pal(nPop(data.light), "Set3")
   # Custom x and y labels
   xlab = paste("Axis 1 (", format(round(percent [1], 1), nsmall=1)," %)",
   sep="")
   ylab = paste("Axis 2 (", format(round(percent [2], 1), nsmall=1)," %)",
   sep="")
   # Custom theme for ggplot2
   ggtheme = theme(axis.text.y = element_text(colour="black", size=12),
                   axis.text.x = element_text(colour="black", size=12),
                   axis.title = element_text(colour="black", size=12),
                   panel.border = element_rect(colour="black", fill=NA,
   linewidth=1),
                   panel.background = element_blank(),
                   plot.title = element_text(hjust=0.5, size=15))

   PCAplot <- ggplot(data = ind_coords, aes(x = Axis1, y = Axis2)) +
     geom_point(aes(fill = Ind), shape = 21, size = 6, show.legend = FALSE) +
     geom_label(data = ind_coords,
                aes(label = Ind, fill = Ind, alpha = 0.5, size = 4),
   show.legend = FALSE) +
     scale_fill_manual(values = cols) +
     scale_colour_manual(values = cols) +
     labs(x = xlab, y = ylab) +
     ggtheme

   PCAplot


   ggsave(PCAplot,
          filename = "{output_directory_path}",
          device = "pdf",
          height = 4, width = 4, units = "in")
```

*Script A5* An R script to assign GO term with functions.

```
   library(httr)
   library(jsonlite)
   library(tidytext)
   library(dplyr)
   library(stringr)

   # Read GO terms from the input file
   input_file <- "{/path/to/input_list_of_go_terms.txt}"
   output_file <- "/path/to/output/goterms_functions.csv"

   go_terms <- readLines(input_file)
   print("Initializing an empty data frame for storing results")

   remove(results) #if this isn't your first time runnning this code
   results <- data.frame(GO_Term = character(),
                         Function_Name = character(),
                         stringsAsFactors = FALSE)
```

```
print("Starting iteration over each GO term and processing data")

# Function to fetch the label (Function_Name) for a GO term
get_go_label <- function(go_term) {
  base_url <- "https://api.geneontology.org/api/ontology/term"
  url <- paste0(base_url, "/", go_term)

  response <- GET(url)

  if (status_code(response) == 200) {
    data <- fromJSON(content(response, "text", encoding = "UTF-8"))

    # Ensure the 'label' field exists
    if (!is.null(data$label)) {
      return(data$label)
    } else {
      return(NA)  # Ensure function always returns something
    }
  } else {
    return(NA)
  }
}

# Process each GO term and populate the data frame
for (term in go_terms) {
  label <- get_go_label(term)
  results <- rbind(results, data.frame(GO_Term = term, Function_Name =
label, stringsAsFactors = FALSE))
}

print("Finishing iterations")

#print("First 10 results")
#head(results)

write.csv(results, output_file, row.names = FALSE)
```

*Script A6* An R script to assign GO term functions to SVs. GO terms are associated with a pacid; pacids—from the annotated transcript—are assigned to the appropriate pangenome SV. This allows for SVs to be functionally annotated. Later in the script (`# Creating bubble plots`), SV sets can also be made from within the pangenome, to determine functions assigned within sample groups.

```
library(readr)
library(tidyr)
library(dplyr)
library(vcfR)
library(stringr)
library(ggplot2)
library(reshape2)
library(ggridges)
library(forcats)
library(readxl)

## FUNCTIONAL ANNOTATIONS OF SVS
# GO term and Functional Annotation
```

```
go_functions <- read_csv("{~/goterms_with_functions.csv}") #input list of
go terms and associated functions
colnames(go_functions) <- c("GO", "GO_Function")
norepeat_go_functions <- distinct(go_functions)

# associating pacids (from transcript annotation) present in pangenome
SVs) to their GO term
vcf_filtered_annotation <-
read.delim("{~/annotation_information_with_pacid_and_go.txt}",
header=FALSE, comment.char="#")
vcf_filtered_annotation <- vcf_filtered_annotation %>%
  select(V1, V10)#for pacid and corresponding GO column lists
colnames(vcf_filtered_annotation) <- c("pacid", "GO")
vcf_annotation_split <- vcf_filtered_annotation %>%
  separate_rows(GO, sep = ",") # make sure grouped GO terms (e.g.
GO:0001242, GO:1243456 are separated). This allows multiple functions to
be appropriately extracted from a single pacid

# find the matching functional annotation for each pacid
pacid_gofunction <- merge(norepeat_go_functions, vcf_annotation_split, by
= "GO")

# load pangenome SVs information with associated pacids
pacid_data <- read_excel("{~/list_of_pacids_found_within_SVs}")

filter_rows <- function(pacid_data, pacid_gofunction) {
  # Ensure 'pacid' columns are treated as character
  pacid_data$pacid <- as.character(pacid_data$pacid)
  pacid_gofunction$pacid <- as.character(pacid_gofunction$pacid)

  # Identify rows where any values in 'pacid' match the values in
pacid_gofunction
  pacid_data$Matches <- sapply(pacid_data$pacid, function(text) {
    # Split the text in 'pacid' into individual elements
    values <- unlist(strsplit(text, ","))
    # Check if any of the values are in pacid_gofunction$pacid
    any(values %in% pacid_gofunction$pacid)
  })

  # Filter rows where Matches is TRUE
  filtered_df <- pacid_data [pacid_data$Matches, ]

  # Create a new column with the matching pacid_gofunction$pacid values
  filtered_df$Matched_pacid <- sapply(filtered_df$pacid, function(text) {
    values <- unlist(strsplit(text, ","))
    paste(intersect(values, pacid_gofunction$pacid), collapse = ", ")
  })
  # Drop the temporary 'Matches' column
  filtered_df$Matches <- NULL

  return(filtered_df)
}

pacid_filtered_df <- filter_rows(pacid_data, pacid_gofunction)
View(pacid_filtered_df)

pacid_filtered_df <- pacid_filtered_df %>%
```

```
  select({"CHROM","POS","ID","REF","QUAL","FILTER","INFO”,{Samples
list},"Matched_pacid"}) #select columns for SV information, SV presence in
different varities, and pacid

pacid_expanded_df <- pacid_filtered_df %>% separate_rows(Matched_pacid,
sep = ", ")
colnames(pacid_expanded_df) <-
c("CHROM","POS","ID","REF","QUAL","FILTER","INFO”,{Samples list},"pacid")
#give every pacid it's own row with the appropriate SV information

sv_functionalannotation_df <- merge(pacid_expanded_df, pacid_gofunction,
by = "pacid") #combination of pacid and GO are unique

#to output just a CSV of all your SVs with the functional annotations
write.csv(sv_functionalannotation_df, file =
"{~/directory_path/SV_functional_annotation.csv}", row.names = TRUE)

## MAKING FIGURES FROM FUNCTIONAL ANNOTATION OF SVS
library(ggplot2)

#annotating SVs and functions based on which varietals they occur in (e.g.
I have four varietals (Atlant (cultivar), Bienne (wild), Heiya (cultivar),
and Longya (cultivar))); replace and customize SV sets from “All
Varietals, Multiple Varietals, Wild, Cultivar-only”
sv_functionalannotation_df1 <- sv_functionalannotation_df %>%
  mutate(Varietal = case_when(
    ((Atlant != "NA") & (Bienne != "NA") & (Heiya != "NA") & (Longya !=
"NA")) ~ "All Varietals",

    ((Bienne != "NA") & ((Atlant != "NA") | (Heiya != "NA") | (Longya !=
"NA"))) ~ "Multiple Varietals",

    ((Bienne %in% c(0, 1, 2, 3, 4))) ~ "Wild",

    ((Atlant %in% c(0, 1, 2, 3, 4)) & (Heiya %in% c(0, 1, 2, 3, 4)) &
(Longya %in% c(0, 1, 2, 3, 4))) ~ "Cultivar-Only"
  ))
sv_functionalannotation_df1 <-
subset(sv_functionalannotation_df1, !is.na(Varietal))

go_function_list <- sv_functionalannotation_df1 %>%
  count(GO_Function, Varietal)

go_function_list1 <- go_function_list %>%
  group_by(GO_Function) %>%
  pivot_wider(names_from = Varietal, values_from = n, values_fill = list(n
= NA))
go_function_list1$`Total Occurrences` <- rowSums(go_function_list1 [,
c("All Varietals", "Multiple Varietals", "Cultivar-Only", "Wild")], na.rm
= TRUE)

#calculate the ratio of SV enrichment against the pangenome SV annotation
background
go_function_list2 <- go_function_list1 %>%
  mutate(
    MultipleVarietalRatio = `Multiple Varietals`/`Total Occurrences`,
    CultivarOnlyRatio = `Cultivar-Only`/`Total Occurrences`,
```

```
    WildRatio = Wild/`Total Occurrences`
  )

#if you want the function list by SV set
#allvar_function <- go_function_list2 %>%
#  select("GO_Function", "All Varietals")
#View(allvar_function)
#cultivar_function <- go_function_list2 %>%
#  select("GO_Function", "Cultivar-Only")
#cultivar_function <- na.omit(cultivar_function)
#View(cultivar_function)
#multiple_function <- go_function_list2 %>%
#  select("GO_Function", "Multiple Varietals")
#multiple_function <- na.omit(multiple_function)
#View(multiple_function)
#wild_function <- go_function_list2 %>%
#  select("GO_Function", "Wild")
#wild_function <- na.omit(wild_function)
#View(wild_function)

sv_functionalannotation_df2 <- sv_functionalannotation_df1 %>%
  left_join(go_function_list2, by = "GO_Function") #%>%

# Creating bubble plots
# Plot for All-Cultivars
cultivar_comparison <- sv_functionalannotation_df2 %>%
  subset(Varietal == "Cultivar-Only") %>%
  arrange(desc("CultivarOnlyRatio")) %>%
  select("GO_Function","Total Occurrences","Cultivar-
Only","CultivarOnlyRatio") %>%
  distinct()

sum(cultivar_comparison$`Cultivar-Only`)

cultivar_comparison$GO_Function <- factor(cultivar_comparison$GO_Function,
levels = cultivar_comparison$GO_Function
[order(cultivar_comparison$CultivarOnlyRatio, decreasing = FALSE)])
View(cultivar_comparison)
cultivarplot<-  cultivar_comparison %>%
  arrange(desc(CultivarOnlyRatio)) %>%
  head(25) %>%

  ggplot(aes(x = GO_Function, y = CultivarOnlyRatio, size = `Total
Occurrences`)) +
  geom_point(alpha = 0.6, color = "#F5930C") +
  labs(y = "Ratio", x = "Functional Annotation", title = "SV Enrichment in
All Cultivars") +
  coord_flip() +
  ylim(0.25, 1) + #x axis originally until 0.18
  scale_size_continuous(range = c(1, 10)) + #from 1 to 10
  theme_minimal()
cultivarplot

# Plot for Wild Flax
wild_comparison <- sv_functionalannotation_df2 %>%
  subset(Varietal == "Wild") %>%
  arrange(desc(WildRatio)) %>%
```

```
  select("GO_Function","Total Occurrences","Wild","WildRatio") %>%
  distinct()

wild_comparison$GO_Function <- factor(wild_comparison$GO_Function, levels
= wild_comparison$GO_Function [order(wild_comparison$WildRatio, decreasing
= FALSE)])

wildplot <-  wild_comparison %>%
  arrange(desc(WildRatio)) %>%
  head(25) %>%

  ggplot(aes(x = GO_Function, y = WildRatio, size = `Total Occurrences`))
+
  geom_point(alpha = 0.7, color = "#32A89B") +
  labs(y = "Ratio", x = "Functional Annotation", title = "SV Enrichment in
Wild Flax") +
  coord_flip() +
  ylim(0.25, 1) + #x axis originally from 0.4 to 1
  #    scale_size_continuous(range = c(0.5, 11.5)) + #ranges from 1 to 23
  theme_minimal()
wildplot

#save bubble plots to directory path as png
ggsave("Cultivarplot.png", plot = cultivarplot, path =
"{~/directory_path/}", width = 9.5, height = 5, units = "in", bg =
"white")
ggsave("Wildplot.png", plot = wildplot, path = "{~/directory_path/}",
width = 9.5, height = 5, units = "in", bg = "white")
```

*File A1* Excel file for all the SV functional annotations from the GO database. Details of the SV location in the pangenome are defined. Each SV sample's annotation is assigned to a SV set group label. The "Cultivar" SV set refers to SVs which have variants present in all three domesticated *L. usitatissimum* variatals (Atlant, Heiya, Longya); the "Wild" SV set refers to SVs only detected in *L. bienne* (Bienne); the "Multiple Cultivar" SV set refers to SVs detected in Bienne and at least one domesticated varietal. For trouble accessing file, email esme.padgett@durham.ac.uk or a.c.brennan@durham.ac.uk.

File A1 (SV_functional_annotation2.xlsx)