1 **plotsr: Visualising structural similarities and rearrangements** 2 **between multiple genomes**

3 Manish Goel^{1,2}, Korbinian Schneeberger^{1,2,*}

4 ¹Faculty of Biology, LMU Munich, Großhaderner Strasse 2, 82152 Planegg-Martinsried, Germany

5 ²Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829
6 Cologne, Germany

7 *Corresponding Author: k.schneeberger@lmu.de

9 **Abstract**

10 **Summary:** Third-generation genomic technologies have led to a sharp increase in the
11 number of high-quality genome assemblies. This allows the comparison of multiple
12 assembled genomes of individual species and demands for new tools for visualising their
13 structural properties. Here we present plotsr, an efficient tool to visualize structural
14 similarities and rearrangements between multiple genomes. It can be used to compare
15 genomes on chromosome level or to zoom in on any selected region. In addition, plotsr can
16 augment the visualisation with regional identifiers (e.g. genes or genomic markers) or
17 histogram tracks for continuous features (e.g. GC content or polymorphism density).

18 **Availability and implementation:** plotsr is implemented as a python package and uses the
19 standard matplotlib library for plotting. It is freely available under the MIT license at GitHub
20 (<https://github.com/schneebergerlab/plotsr>) and bioconda
21 (<https://anaconda.org/bioconda/plotsr>).

22 **Contact:** Manish Goel (manish.goel@lmu.de), Korbinian Schneeberger
23 (k.schneeberger@lmu.de)

24

25 **Introduction**

26 Third-generation sequencing technologies like PacBio high-fidelity sequencing (HiFi) and
27 Oxford-nanopore sequencing (ONT) have revolutionised genome assembly (Koren and
28 Phillippy, 2015). Together with methods like Hi-C, trio-binning or gamete-binning (that
29 perform efficient phasing and scaffolding of the contigs) haplotype-resolved, chromosome-
30 level assemblies can easily be generated (Zhang *et al.*, 2019; Koren *et al.*, 2018; Campoy *et al.*, 2020). This has led to a sharp increase in the number of available high-quality genome
31 assemblies. Chromosome-level assemblies allow for the identification of not only small
32 genomic differences (like SNPs and indels) but also for finding large structural
33 rearrangements (SRs) like inversions and translocations. They are therefore considered as
34 the gold standard for the identification of genomic differences (Simpson and Pop, 2015).

36 An interesting observation can be made when comparing the different genomes of an
37 individual species (intra-genome comparison). The overall structure of these genomes is
38 usually highly conserved as the genomes need to exchange chromosome arms during
39 sexual reproduction. This implies the presence of extensive amounts of syntenic regions
40 between the genomes, where the exchange of chromosome arms can happen without
41 affecting the overall structure of the genomes. During genome assembly comparison these
42 syntenic regions (the “syntenic backbone”) can be identified. All remaining regions in the
43 genomes are structural rearrangements by definition and can then be classified into
44 duplications, translocations or inversions based on their orientation and location in the
45 genomes. Recently, we used these principles to develop SyRI, a tool to identify genomic
46 differences in whole-genome assemblies of the same species (Goel *et al.*, 2019).

In this work, we present **plotsr** (**plot structural rearrangements**) a novel visualisation tool for genomic differences, which is based on the synteny concepts included in SyRI. Using the syntenic backbone as a guide, plotsr can visualize structural differences on a chromosome level, as well as zoom in on specific regions. plotsr is a simple-to-use yet flexible and powerful visualisation tool. It can be used to compare multiple haploid genomes as well as different haplotypes of individual polyploid genomes. In addition, plotsr can mark specific loci as well as plot histogram tracks to show distributions of genomic features along the chromosomes.

Implementation

plotsr is a command-line tool. It requires the assembly sequences (in fasta format) and the synteny and SRs information between the assemblies in a pairwise manner as input. For example, to visualise genomes A, B and C in this order, plotsr requires the comparison of A vs. B and B vs. C. These can be generated using SyRI or can be provided in BEDPE format. Firstly, plotsr validates that the assemblies and structural information are consistent. Then, by using the pairwise synteny between genomes, it groups homologous chromosomes across the genomes and plots the syntenic regions as well as SRs between them. The output plot can be generated in pdf, png or svg format. In addition to the genomic similarities and differences, plotsr can include markers at specific loci (e.g. genes, TEs, or genomic markers) using BED files. plotsr can also plot histogram tracks showing the distribution of genomic features along the chromosomes (e.g. distribution of SNPs, sequencing reads etc). This provides a visual comparison between sequence features and structural properties of the chromosomes. To adjust these plots to specific needs, plotsr includes multiple parameters to control the visual properties (colour, size, spacing etc) of genomes, markers and tracks.

plotsr can also be used to zoom in on specific regions in any of the input genomes. For this, plotsr identifies the corresponding homologous regions in all other genomes. This is a non-trivial task as some regions might include multiple rearrangements that obfuscate the syntenic regions in the other genomes. The identification of all syntenic regions would require whole-genome alignments of all genomes against the genome of interest implying the need for an all-vs-all genome alignment as input. This is computationally prohibitive once tens of genomes are involved. Instead, plotsr overcomes this challenge by using the syntenic backbone between the genomes to zoom in on any given region. For this, plotsr iteratively selects the regions syntenic to the selected region using pairwise genome comparisons until all genomes are covered. It then filters the structural information to only plot information overlapping these homologous regions resulting in a zoomed-in view of the genomes. Markers and feature tracks are also filtered automatically to plot those overlapping with the homologous regions.

Results

We demonstrated the applicability of plotsr by visualising structural rearrangements between eight divergent strains of the model plant *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000; Jiao and Schneeberger, 2020) (Figure 1). For this, we performed seven pairwise whole-genome alignments using minimap2 and then identified synteny and structural rearrangements using SyRI (Li, 2018; Goel *et al.*, 2019). In addition, we obtained some general genome features for further labelling (including gene annotation, population-wide polymorphism data and centromere coordinates (Lamesch *et al.*, 2012; Alonso-Blanco *et al.*, 2016; Giraut *et al.*, 2011)). The visualization by plotsr showed that the genomes were predominantly syntenic (grey alignments) except for some large inversions including a 2.5MB inversion on Chr3 in the genome of Sha and a 1.1MB inversion on Chr4 in Col-0, which we labelled using plotsr as “Inversion 1” and “Inversion 2”. Using the annotation tracks, the

increased density of SRs within the centromere regions became readily apparent in the visualization. Moreover, we also plotted the genes and frequency of SNPs along the chromosomes, which revealed the known inverse relationship between gene and SNP density (The Arabidopsis Genome Initiative, 2000) as well as the depletion of genes in the centromeres.

We also used the zoom-in option of plotsr to highlight the genomic differences at Chr3:6,600,000-10,000,000 in the genome of “col-0” (Figure 2). The region was provided as a command-line parameter to plotsr which then automatically filtered and plotted the syntenic regions in all of the remaining genomes. In this example, we observed two large regions without any alignments (labelled as “Not aligned 1” and “Not aligned 2”) that signify divergence between the two strains as well as two more inversions that become visible during zoom in.

Conclusion

The advent of high-quality long-read sequencing technologies simplified the generation of high-quality genome assemblies. To support the visual analysis of such assemblies, we presented plotsr, a computational tool for visualising structural similarities and rearrangements between genomes. plotsr is highly adjustable and allows additional visualisation of genomic features as well as zoom-in views on specific regions. It generates publication-quality visualisations (Zamyatin *et al.*, 2021; Zhang *et al.*, 2021; Li *et al.*, 2021) and will help towards a better understanding of differences within genomes. Given the great importance of genomic analysis in many research fields, we are continuously developing plotsr to add more useful parameters allowing for more control and customisation.

Funding

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2048/1– 390686111.

References

- Alonso-Blanco, C. *et al.* (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*, **166**, 481–491.
- Campoy, J.A. *et al.* (2020) Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol.*, **21**, 1–20.
- Giraut, L. *et al.* (2011) Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet.*, **7**, e1002354.
- Goel, M. *et al.* (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.*, **20**, 277.
- Jiao, W.-B. and Schneeberger, K. (2020) Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.*, **11**, 989.
- Koren, S. *et al.* (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*, **36**, 1174–1182.
- Koren, S. and Phillippy, A.M. (2015) One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.*, **23**, 110–120.

140 Lamesch,P. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene
141 annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.

142 Li,A. *et al.* (2021) Two Reference-Quality Sea Snake Genomes Reveal Their Divergent
143 Evolution of Adaptive Traits and Venom Systems. *Mol. Biol. Evol.*, **38**, 4867–4883.

144 Li,H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**,
145 3094–3100.

146 Simpson,J.T. and Pop,M. (2015) The Theory and Practice of Genome Sequence Assembly.
147 *Annu. Rev. Genomics Hum. Genet.*, **16**, 153–172.

148 The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering
149 plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

150 Zamyatin,A. *et al.* (2021) Chromosome-level genome assemblies of the malaria vectors
151 *Anopheles coluzzii* and *Anopheles arabiensis*. *Gigascience*, **10**, 1–16.

152 Zhang,W.-P. *et al.* (2021) Dead-End Hybridization in Walnut Trees Revealed by Large-Scale
153 Genomic Sequence Data. *Mol. Biol. Evol.*, **39**.

154 Zhang,X. *et al.* (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes
155 based on Hi-C data. *Nat. Plants* 2019 58, **5**, 833–845.

156

157

158 **Figures**

159

160 **Figure 1: Example for a genome-wide visualization using plotsr.** We used plotsr to
 161 visualize syntenic regions and structural rearrangements between eight strains of *A. thaliana*.
 162 The visualisation is created using plotsr without further modifications. It shows synteny and

SRs between the eight genomes. Tracks for three genomic features: genes, number of SNPs and centromeric regions were included using optional parameters. In the genes track, small blue lines correspond to transcribed regions and long blue lines represent coding-sequences (CDS). We also marked two large inversions, one each on chromosome 3 and chromosome 4.

Figure 2: Using plotsr to zoom in on Chr3:6,600,000-10,000,000 in the “col-0” genome. The visualisation is created using plotsr without any further modifications. Zooming in on a specific location allows for resolved visualisation of the local genomic differences. Here, we observe two large not aligned regions and two smaller inversions, which have been labelled with plotsr.