Ambiogous regions found in public plant genomes

Lucaciu R.[1], Rattei T.[1]

[1]*Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria*

Abstract

The DNA extracted from plants for plant sequencing projects can contain other eukaryotic, microbial and viral DNA. Although unintended, plant genome sequencing projects are thereby sources for DNA from members of their microbial communities. This phenomenon has so far received relatively little attention in the plant community, although well-known and even utilized for animal genomes. Intrinsic information, such as k-mer frequencies and sequencing coverage, indicates regions of unexpected characteristics, which might consist of foreign DNA, HGT or repeats [1], [2]. Database searches identify regions of unexpected similarity between unrelated genomes, which are candidates for HGT or contamination. However, none of these approaches alone allows reasonably specific detection of contamination in genomes assemblies. To get a first impression of the current status of possible microbial contaminations on plant genomes, we used the database-centric approach. Plant assembly genomes from Genebank were split and screened with kraken [3] software against microbial human and viral contaminations. For many species, but most clear for Arabidopsis thaliana we observed that one or few genome assemblies had many more microbial, viral or human segments than typical, Table1. Such assemblies would be reasonable targets for assembly evaluation, reassembly or even resequencing.

In this study we propose a method to determine possible contaminated regions in the genomes based on raw information obtained from short-read coverage. The method assume that the coverage on every contig (chromosome) follow a general type of distribution. This is calculated from random coverage of segments from the genome. With a sliding window, genome coverage is screened and regions that do not fit to the distribution (base on a selected Z-score) are called ambiguous regions and might indicate possible contamination.

Observation

For a quick first impression about contamination in plant genomes, all latest assembly versions of plant genomes in NCBI Genbank were split into chunks of 10kb, overlapping each other by 5kb. These regions were screened with the Kraken software [3] for microbial and viral contaminations. To avoid false positive by short, unspecific hits of high similarity, only consecutive 25kb regions of 4 overlapping fragments, which all matched the same taxonomic lineage in the Kraken database, were counted. In Table 1 we showed the number of segments found in Arabidopsis genome.

Arabidopsis Genome ID	Length	Potential foreign segments				
		viral	bacterial	archaeal	fungal	human
GCA_000523005.1_Ahal_1.0	102276753	0	6	0	1	15
GCA_900078215.1_Ahal2.2	188585209	0	171	0	9	51
GCA_000004255.1_v.1.0	202142215	0	456	0	12	233
GCA_000524985.1_Alyr_1.0	39278059	0	2	0	0	0
GCA_000001735.1_TAIR10	119146348	0	58	0	8	72
GCA_000211275.1_ASM21127v1	93654490	0	32	0	4	58
GCA_000222325.1_Bur-0_2010-09-30	91404641	0	5	0	0	2
GCA_000222345.1_C24_2010-09-30	94827584	0	13	0	3	7
GCA_000222365.1_Ler-1_2010-09-30	93324667	0	8	0	3	11
GCA_000222385.1_Kro-0_2010-09-30	90729513	0	3	0	1	5
GCA_000835945.1_ASM83594v1	124475305	0	134	0	14	67
GCA_001651475.1_Ler_Assembly	118890721	0	24	0	7	54
GCA_001742845.1_AthNd1_v1.0	109108457	0	16	0	7	48
GCA_001753755.2_Athal_Col0Cvi0F1	243352215	0	97	0	22	121

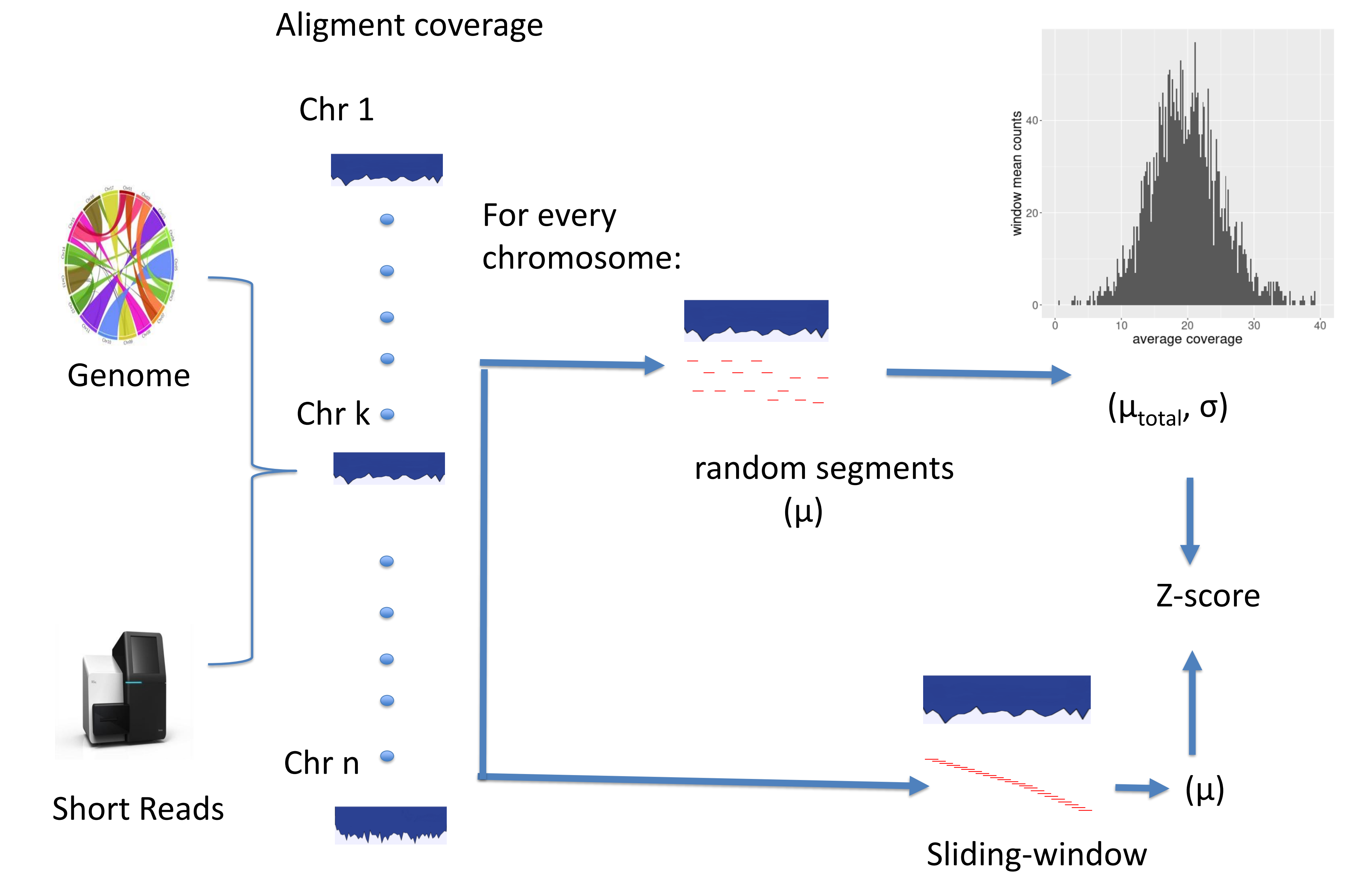
Hypothesis

In Arabidopsis plant genomes we observe that one or few genome assemblies had many more microbial, viral or human segments than typical. We expect that TAIR10 to be well polished and checked for contamination and we expect a close number of bacterial segments to appear in all other genomes in raport with the genome size. This might be an indicative that such regions might contain foreign DNA.

Discussion

The discussed method is still a work in progress that need to be tested more and validated. So far based on results we have and in the given examples we noticed that there are ambiguous regions found in plant genomes that might correspond to microbial, viral or human DNA. There are also chances to determine regions that could represent contaminated DNA but due to the limitation in the database they were not found by kraken. There are also microbial segment that might have the coverage that fit to the distribution. The method will be further tested with different values for parameters (z-score threshold and sliding-window size) for a better understanding of ambiguous regions. We might also include in the method genomic information related to structure composition of the genome.

Methods



Aligment coverage

Genome

Short Reads

For every chromosome:

random segments (μ)

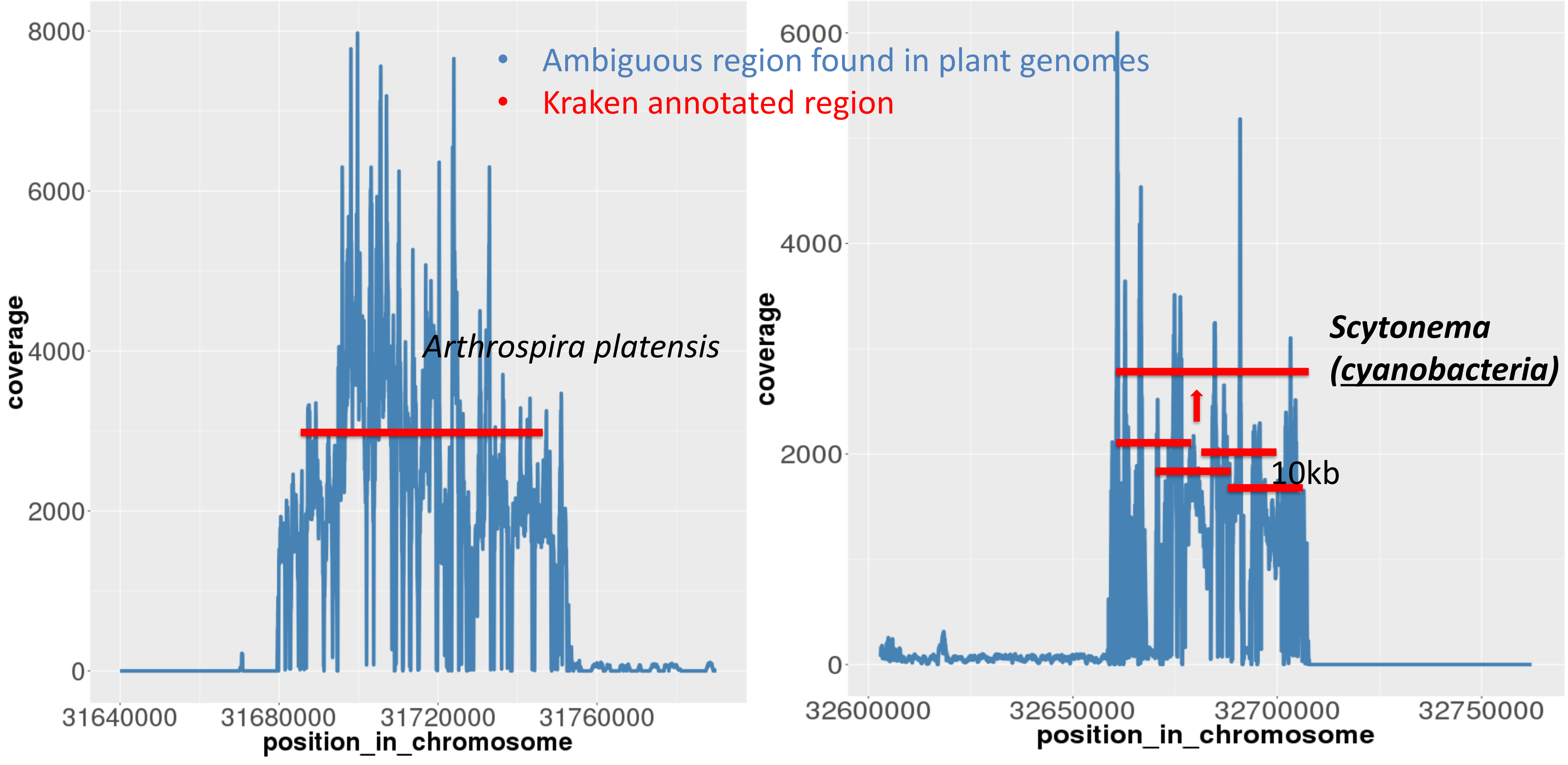
(μ_{total}, σ)

Z-score

Sliding-window

(μ)

Examples of ambiguous region determined with the method



Examples of ambiguous region found in one of the Arabidopsis genome (GCA_001753755.2) for which consecutive 25kb ,or greater, regions of at least 4 overlapping fragments (of 10kb) were annotated with kraken.

References

1 Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Berger B, ed. Bioinformatics. 2017;33(4):574-576. doi:10.1093/bioinformatics/btw663.

2 Delmont, Tom O., and A. Murat Eren. "Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies." PeerJ 4 (2016): e1839.

3 Wood, Derrick E., and Steven L. Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments." Genome biology 15.3 (2014): R46..

Contact information:
Thomas Rattei:thomas.rattei@univie.ac.at

We gratefully acknowledge funding received from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 675657.

Horizon 2020
European Union Funding
for Research & Innovation

MARIE CURIE ACTIONS

Cube
Division of Computational
Systems Biology