Characterising nullomers and rare sequences in the pan-genome of *Arabidopsis thaliana*

Jack Morgan Book [1], Amanda Clare [1], Wayne Aubrey [1*]
[1]Department of Computer Science, Aberystwyth University, Aberystwyth, UK

Abstract

Absent or rare sequences in genomes are of interest to genome analysts, for medical and diagnostic purposes or because their absence may indicate otherwise harmful sequence. The extraction and characterisation of rare sequences enables us to better understand their properties and potential. We use a recently published set of 69 *Arabdopsis thaliana* genome assemblies to investigate the nullomers and rare sequences that are present and absent across this model plant species. We find that, while CpG sites are a strong factor influencing core absent and rare sequences, the distribution of accessory nullomers is very variable across the genomes.

Introduction

Nullomers are short sequences of DNA that are absent from the genome of a species. These sequences can be missing from a specific genome, absent across an entire species, or or even absent from multiple species (Hampikian and Andersen, 2006). Sequences whose absence is conserved may have deleterious consequences to the organism or be structurally infeasible due to their chemical and physical properties. Negative selection has been observed when analysing nullomers in eukaryotes (Georgakopoulos-Soares et al., 2021). While some sequences may be absent by chance, studies have shown that there are more sequences absent than would be expected by random chance (Koulouras and Frith, 2021).

Previous research suggested that the hyper mutability of CpG sites led to the absence of some sequences, due to the high number of mammalian nullomers containing CpG sites at the time (Acquisti et al., 2007). However, further studies on dinucleotide content in vertebrates concluded that CpG islands alone were insufficient to explain all absent sequences. (Vergni and Santoni, 2016) computed mean helical rise for nullomer sequences from several mammalian genomes and found this to be significantly greater than for present sequences, leading them to hypothesise that DNA-histone interactions were partly the reason for the absence of these sequences, as higher helical rises allow more room for the phosphate-histone interaction.

Other research has suggested a potential role for nullomers in immune response (Patel et al., 2012). If a sequence is detected that is absent from the host but present in a pathogen, this difference can be utilised by the host in an immune response. Sequences that are consistently present in one organism but absent in another can be useful as genetic signatures (Silva et al., 2015; Pratas and Silva, 2020).

Previous studies have primarily focused on mammalian, bacterial and viral genomes (Vergni and Santoni, 2016; Georgakopoulos-Soares et al., 2021; Montgomery et al., 2024). In this study, we characterise the nullomers and rare sequences in a large set of plant genomes. *Arabidopsis thaliana*, the most well characterised model plant, now has many high-quality annotated genome assemblies available (Lian et al., 2024), allowing us to inspect the variation of nullomers and rare sequences across its pangenome.

We present a summary of the nullomers and rare sequences in the pan-genome of *Arabidopsis thaliana*, including analysis of their GC content, as well as their locations and annotations when present in some of the genomes.

Materials and Methods

Data

The 69 *Arabidopsis thaliana* genomes were downloaded from NCBI GenBank all under the accession number PRJNA1033522. The associated annotaion files are also available under the same accession number. Details of the genomes, assem-

blies and annotation are available in the original paper from Lian et al. (2024).

The *Brachypodium distachyon* genome was downloaded from Genbank under BIOProject ID PRJNA32607. The *Solanum lycopersicum* genome was downloaded from the European Nucleotide Archive under study accession PRJEB44956.

Nullomer Generation

A pipeline (see Figure 2) was developed in Python that when provided with genome(s) and a desired k-mer length (k), produced a set of nullomers for each genome. Jellyfish (Marcais and Kingsford, 2011) was used to identify the set of present k-mers. The set of all possible k-mers for a given length k was then generated. The set of nullomers was then determined by comparing the set of all possible k-mers with the present k-mers. From the resulting 69 sets of nullomers, a subset of nullomers that are unique across all 69 genomes can be found.

Given a genome $g \epsilon G$, and a desired k-mer length m we denote a present k-mer of length m in genome g by k_{mg} and the set of all present k-mers of length m for a genome g by K_{mg}. We further denote the set of all possible k-mers of length m by K_m.

The set of nullomers of length m for a genome g is:

$$N_{mg} = K_m - K_{mg}$$

The set of *core* nullomers across the collection of 69 genomes is C_{mG} and represents nullomers absent from every genome:

$$C_{mG} = (N_{mg_1} \cap N_{mg_2} \cap ... \cap N_{mg_{69}})$$

The set of *all unique* nullomers across the collection of 69 genomes is A_{mG} and represents nullomers absent from at least one of the genomes:

$$A_{mG} = (N_{mg_1} \cup N_{mg_2} \cup ... \cup N_{mg_{69}})$$

Discovering location information

The set of accessory nullomers now contains sequences that are absent from some genomes but present in others. To identify the locations of these sequences, a pipeline was developed (see Figure 1). A Bowtie 2 index was generated for each of the 69 genomes (Langmead and Salzberg, 2012).

	Nullomers	Present k-mers
10-mer	778	1,047,798
11-mer	82,219	4,112,085
12-mer	2,522,690	14,254,526

Table 1: The numbers of nullomers on average not present per genome and the number of k-mers present on average per genome

Using the set of shared nullomers, we queried the databases, returning locations of where a sequence was present in each genome. Using bedtools intersect (Quinlan and Hall, 2010), the annotations for each genome were intersected with the locations of nullomers. The nullomers were grouped by chromosome, to provide a count of nullomers in each chromosome.

Results

The nullomers

Table 1 shows the average number of nullomers generated from each of the 69 genomes, for k-mer lengths, 10, 11 and 12. A full list of these nullomers can be downloaded from `https://github.com/Somerset-Jack/Nullomers`.

While each genome has on average 778 nullomers of length 10, they are not all unique. A rarefaction graph demonstrates how many new nullomers additional genomes add to the set of unique nullomers (Figure 3). Clearly, the set of 69 genomes is not yet sufficient to demonstrate all nullomers that might be absent from *A. thaliana*.

Only 6 nullomers of length 10 are core (absent from all 69 genomes), as can be seen in Table 2. These core nullomers are: GGGAGCGCGC, GGCGCCCGCG, ACGGGCGCGC, GGGCGCGCCG, CCGCGCGCCC, GCGCCCGCGT. We note that these six are particularly GC-rich k-mers and each contain several CpG sites, which have long been known to be under-represented in genomes due to the mutation rates of CG dinucleotides (Josse et al., 1961). For 11-mers this core set comprises 9381 nullomers. Table 2 also summarises the average number of nullomers per genome and the size of the union of all nullomers generated from all the 69 genomes.

The presence and absence of each nullomer sequence across the 69 genomes of the pan-genome shows huge variability in the rareness of the nullomers. Some nullomers are present in most genomes but absent from just one or a few (for

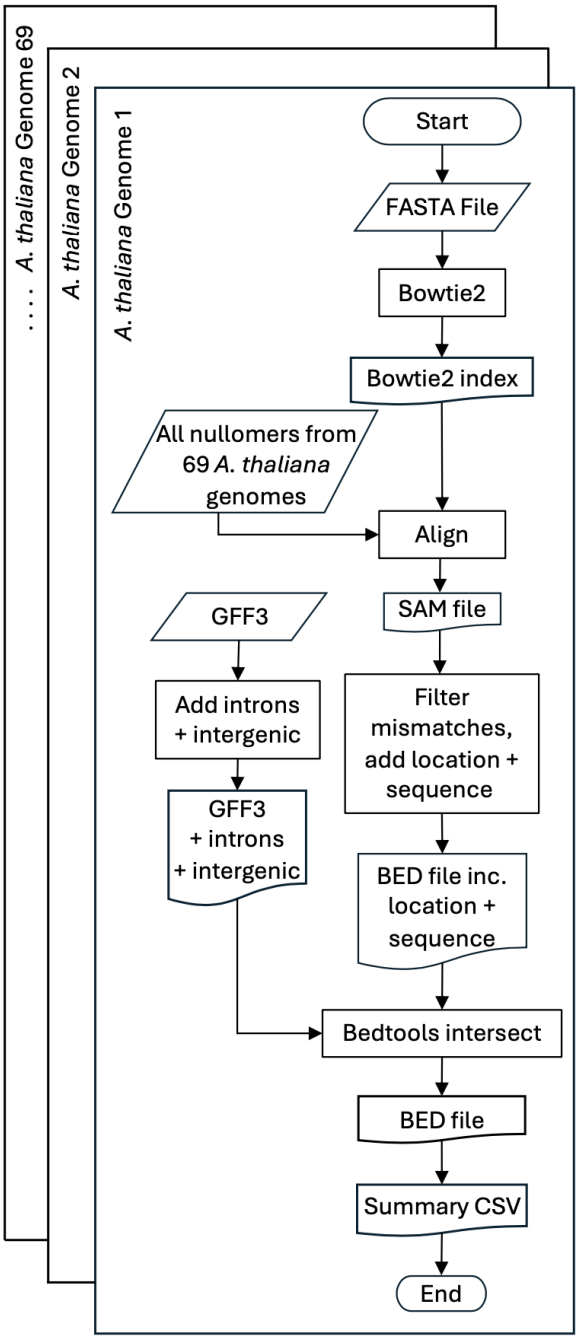

Figure 1: Flowchart illustrating the determination of sequence locations present in one genome but nullomers in another. The Bowtie2 software is used to align nullomer sequences back to all 69 *A. thaliana* genomes. By comparing these locations with the annotations in the GFF3 file, a summary CSV file is produced to record whether a nullomer resides within an annotated region of the genome (CDS, intron, exon, UTR) or in an annotated intergenic region.

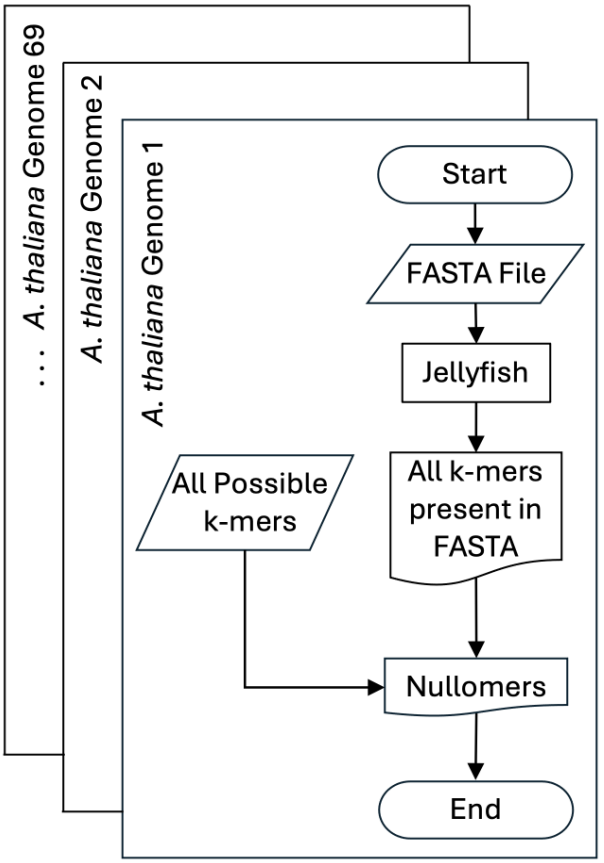

Figure 2: Flowchart illustrating the generation of nullomer sequences for each *A. thaliana* genome. The Jellyfish software indexes all unique k-mer sequences of a specified length present in an input FASTA file. The set of nullomers is then determined by comparing the set of all possible k-mers with the present k-mers, removing the intersection of both sets from the set of all possible k-mers.

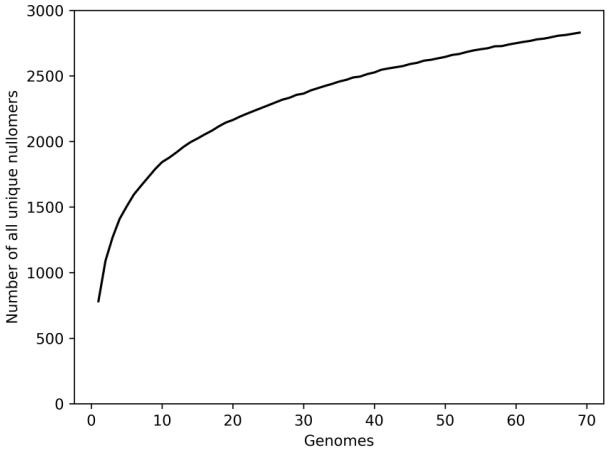

Figure 3: Rarefaction plot showing how the number of all unique length 10 nullomers generated from the genomes increases as a larger number of *A. thaliana* genomes is considered.

Nullomers	Average per genome	Core nullomers	All unique nullomers
10-mer	778	6	2830
11-mer	82,219	9381	194,224
12-mer	2,522,690	787,486	4,329,466

Table 2: The number of 10, 11 and 12-mer nullomers on average not present per genome, and the numbers of core 10, 11, and 12-mer nullomers and all unique 10, 11 and 12-mer nullomers in the pan-genome.

Appearing once in one genome	Appearing once in two genomes
GCGCGCACGC	CGCGCCCCGC
TGCGCGCTCG	GCGCGTCCCC
GGGCGCTGCC	AGGGGCCCCG
AGGCGCGCTC	ATGCGCGCCC
GCGCCCCTAT	CGCGGGGCGC
CGGGGGCCTA	CGCCCGGGGT
GCCGGGCGCG	TGCGGGCGCC
GCGCGCCCCC	CGGGGCCCCG
GCGGACCCCC	GGGGCGTGCG
CCCGGGGGGC	GGGGCCCCCT
GCGCGCGGGG	CAGGGGCGCC
GGGCGCCCGC	GGGTGCCCGC
CGGGCGCGCA	CTGGGCGCAC
GGGGCGCGCG	AGGGCGGGGC
CGTCGCGCCC	
GGCAGCGCCC	
CCCTGGGCGC	

Table 3: List of rare k-mers that appear once in one genome and once in two genomes

example CGCGCTTAGG, which is only absent from GCA_036941985.1 but present in all other genomes). Some nullomers are present multiple times within one genome, and absent from others (for example ACCCAGGGCG, which is present 994 times in genome GCA_036940305.1 but not present at all in 19 other genomes).

Table 3 lists all 10-mers that appear once in genome and once in two genomes, with some of these sequences also absent from other species.

The counts of the rare and absent sequences, where each sequence is compared with each genome is available at `https://github.com/Somerset-Jack/Nullomers`.

Sequence regions of the nullomers

Where nullomers were present in some genomes but not in others we were able to note the annotation for the region in which they were present: intronic, exonic, genic, intergenic, CDS, 3'UTR and 5'UTR.

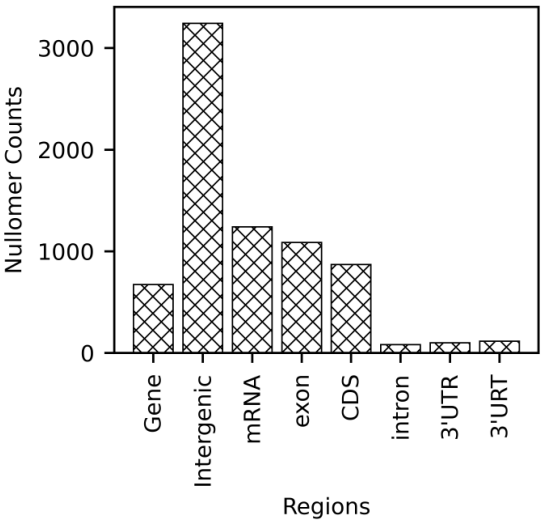

Figure 4: Bar plot showing counts of how many nullomer sequences were found in neighbouring genomes and their location within that genome

In some cases, nullomer sequences were located more than once in each genome, so the total number of annotations for nullomer sequences is larger than that of the original set of nullomers. Across the 69 genomes there was an average of 674 nullomer sequences found within genes, 3243 in intergenic regions, 1241 in mRNA regions, 1086 in exonic regions, 83 in intronic regions, 871 in CDS, 100 in 3'UTR, and 115 in 5'UTR. Figure 4 shows how the counts of nullomer sequences falling within different regions compare.

GC content

The average GC content of the *A. thaliana* genomes was 36.3%. *A. thaliana* is an AT-rich organism. The average GC content of the nullomers of length 10 was 88.1%, far higher than the average GC content of the genomes. However, for those nullomer sequences that were present in some genomes and absent from others, their GC content was still extremely high, being 85.2% on average for mapped nullomers of length 10.

Dinucleotide Content

Chromosome locations

Table 4 shows the chromosome locations of the nullomer sequences, reporting the average count

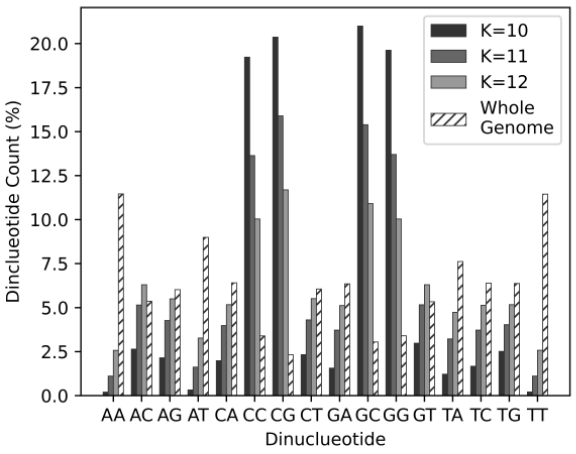

Figure 5: Enter Caption

per chromosome across the genomes. Given the genome lengths, the observed distribution of these counts is significantly different to the expected distribution, using a chi-squared test. There are fewer nullomer sequences than expected found on chromosome 2 and 3 and more sequences than expected found on chromosome 4. The sequence ACCCAGGGCG, which is present 994 times in genome GCA 036940305.1 is found entirely on chromosome 4.

Clustering

The presence/absence counts for each nullomer sequence across the set of genomes form vectors which can be clustered. The nullomer-based clusters formed by a hierarchical clustering (with average linkage distance, Euclidean distance, then flattened on a distance threshold), correspond directly with the geographic origin groups (Africa, Asia, Europe, Madeira, admix) as can be seen in Table 5.

Presence of the nullomers in other plant genomes

Nullomer sequences from *A. thaliana* were checked against the genomes of tomato

The tomato genome had only one 10-mer nullomer sequence. This was not a member of the set of *A. thaliana* nullomer sequences. Brachypodium had no 10-mer nullomer sequences.

Table 6 shows the number of nullomers shared between two different species, comparing the total number generated from one tomato genome, and the set of unique nullomers generated from all 69 Arabidopsis genomes. Over half of the nullomers generated from *Arabidopsis thaliana* were also absent from the *Solanum lycopersicum* genome suggesting that if any sequences did have deleterious properties, these were shared across species.

Discussion

The six core nullomers of length 10 that were absent from all of the genomes, were GC rich and contained several CpG sites. Rare k-mers appearing at most once or twice were also GC rich with CpG sites. The genomes have a very open set of nullomers, with a small nullomer core and a large accessory set. Some of the accessory nullomer sequences appeared very frequently in other genomes.

We characterised nullomer sequences generated from the 69 *Arabidopsis thaliana* genomes into various annotation categories, noting that although the majority of nullomers fell in the intergenic regions of the genome, of the nullomers that were within the genes, the majority of those were within exonic regions.

The geographical origin of the genomes had a direct connection with the sets of nullomers. This matches previous reports that genomes originating in Africa are more genetically diverse as they have the most ancient lineage, and the clustering of Asian origins due to the more modern and rapid spread of *Arabidopsis thaliana* in that region (Alonso-Blanco et al., 2016; Lian et al., 2024).

Other k-mer lengths

Nullomers generated for k-mer lengths of 10, 11 and 12 for each genome are available online: `https://github.com/Somerset-Jack/Nullomers`

We are mainly interested in k-mers with a length less than 13 as these are the shortest sequences absent from the genome. They become less significant as length increases due to the genomes total length not being able to contain them. Many nullomers at longer lengths are just shorter nullomers with more nucleotides attached to either end of the nullomer. At k-mer lengths of 10, 11 and 12, nullomers make up 0.07%, 1.96%, and 15.04% of all possible k-mers respectively, at k-mer lengths of 13 this jumps to 54.61% of all possible k-mers. The potential biological impact of the absent sequences is likely greatly reduced when they make up half of all possible sequences.

	chr1	chr2	chr3	chr4	chr5
Average nullomer count	1030	481	674	864	1024
Average chromosome length	33347840	22254544	26325396	22479997	30995619
Ratio of count/length	0.0000309	0.0000216	0.0000256	0.0000384	0.0000330

Table 4: The chromosomes on which nullomer sequences are found. The distribution is significantly different to that which would be expected by chance (chi-squared test). Fewer nullomer sequences are found on chromosomes 2 and 3, and more on chromosome 4.

Genome cluster	Geographic origin
GCA_036940605.1	Africa
GCA_036936955.1	Africa
GCA_036940995.1	Africa
GCA_036941165.1, GCA_036942965.1	Africa
GCA_036927435.1, GCA_036927455.1, GCA_036927475.1, GCA_036940645.1, GCA_036942405.1	Madeira
GCA_036927285.1	admix
GCA_036942575.1	Africa
GCA_036941465.1	Africa
GCA_036937985.1	Africa
GCA_036927355.1	Africa
GCA_036936895.1, GCA_036940335.1	Africa
GCA_036939995.1, GCA_036942075.1	Africa
GCA_036927025.1, GCA_036927045.1, GCA_036927505.1, GCA_036936845.1, GCA_036936905.1, GCA_036937005.1, GCA_036937185.1, GCA_036937335.1, GCA_036937385.1, GCA_036937475.1, GCA_036937505.1, GCA_036937805.1, GCA_036937815.1, GCA_036938785.1, GCA_036940405.1, GCA_036942975.1	Asia
GCA_036926975.1, GCA_036927245.1, GCA_036927265.1, GCA_036927275.1, GCA_036927385.1, GCA_036940715.1, GCA_036940825.1, GCA_036942435.1, GCA_036942445.1	Europe
GCA_036926925.1, GCA_036926965.1, GCA_036927085.1, GCA_036927255.1, GCA_036927365.1, GCA_036927375.1, GCA_036936705.1, GCA_036936715.1, GCA_036936885.1, GCA_036936915.1, GCA_036936945.1, GCA_036937455.1, GCA_036937465.1, GCA_036937515.1, GCA_036940305.1, GCA_036940415.1, GCA_036940625.1, GCA_036940635.1, GCA_036941295.1, GCA_036941635.1, GCA_036941785.1, GCA_036941915.1, GCA_036941985.1	Europe and admix
GCA_036940615.1	Europe
GCA_036927485.1	Europe

Table 5: Genomes clustered by nullomer sequence counts reveal a structure that is closely aligned with the geographic origins determined by the authors of the 69 genomes. Genomes GCA_036941915.1 and GCA_036941985.1 are labelled "admix" but cluster with many genomes labelled "Europe". These two are both from Turkey.

	Nullomers	Absent from both
Solanum lycopersicum	194,224	3,419
Arabidopsis thaliana	6,054	

Table 6: The total number of nullomers generated by the *Solanum lycopersicum* genome, all unique nullomers generated by all 69 *Arabidopsis thaliana* genomes and the count of nullomers that are absent from both sets, for a nullomer length of 11.

Conclusion

We have presented an initial summary of the nullomers and rare k-mers in a set of 69 assembled genomes of *Arabidopsis thaliana*.

References

C. Acquisti, G. Poste, D. Curtiss, and S. Kumar. Nullomers: Really a matter of natural selection? *PLoS ONE*, 2(10):e1022, 2007. doi: 10.1371/journal.pone.0001022.

C. Alonso-Blanco, J. Andrade, C. Becker, F. Bemm, J. Bergelson, K. M. Borgwardt, J. Cao, E. Chae, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166(2):481–491, 2016. doi: 10.1016/j.cell.2016.05.063.

I. Georgakopoulos-Soares, O. Yizhar-Barnea, I. Mouratidis, and N. Ahituv. Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biology*, 22:245, 2021. doi: https://doi.org/10.1186/s13059-021-02459-z.

G. Hampikian and T. Andersen. Absent sequences: Nullomers and primes. In *Biocomputing 2007*, pages 355–366. World Scientific, 2006. doi: 10.1142/9789812772435_0034.

J. Josse, A. Kaiser, and A. Kornberg. Enzymatic synthesis of Deoxyribonucleic acid: VIII. frequencies of nearest neighbor base sequences in Deoxyribonucleic acid. *Journal of Biological Chemistry*, 236(3):864–875, 1961. doi: https://doi.org/10.1016/S0021-9258(18)64321-2.

G. Koulouras and M. C. Frith. Significant non-existence of sequences in genomes and proteomes. *Nucleic Acids Research*, 49(6): 3139–3155, 2021. doi: https://doi.org/10.1093/nar/gkab139.

B. Langmead and S. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9: 357–359, 2012.

Q. Lian, B. Huettel, B. Walkemeier, B. Mayjonade, C. Lopez-Roques, L. Gil, F. Roux, K. Schneeberger, and R. Mercier. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nature Genetics*, 56(5):982–991, 2024. doi: 10.1038/s41588-024-01715-9.

G. Marcais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. doi: 10.1093/bioinformatics/btr011.

A. Montgomery, G. C. Tsiatsianis, I. Mouratidis, C. S. Y. Chan, M. Athanasiou, A. D. Papanastasiou, V. Kantere, N. Syrigos, I. Vathiotis, K. Syrigos, N. S. Yee, and I. Georgakopoulos-Soares. Utilizing nullomers in cell-free RNA for early cancer detection. *Cancer Gene Therapy*, 31(6):861–870, June 2024. doi: 10.1038/s41417-024-00741-3. URL `https://www.nature.com/articles/s41417-024-00741-3`.

A. Patel, J. C. Dong, B. Trost, J. S. Richardson, S. Tohme, S. Babiuk, A. Kusalik, S. K. P. Kung, and G. P. Kobinger. Pentamers not found in the universal proteome can enhance antigen specific immune responses and adjuvant vaccines. *PLoS ONE*, page e43802, 2012. doi: https://doi.org/10.1371/journal.pone.0043802.

D. Pratas and J. M. Silva. Persistent minimal sequences of SARS-CoV-2. *Bioinformatics*, 36(21):5129–5132, 2020. doi: https://doi.org/10.1093/bioinformatics/btaa686.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

R. M. Silva, D. Pratas, L. Castro, A. J. Pinho, and P. J. S. G. Ferreira. Three minimal sequences found in Ebola virus genomes and absent from human DNA. *Bioinformatics*, 31(15): 2421–2425, 2015. doi: https://doi.org/10.1093/bioinformatics/btv189.

D. Vergni and D. Santoni. Nullomers and high order nullomers in genomic sequences. *PLoS*

ONE, 11(12):e0164540, 2016. doi: 10.1371/journal.pone.0164540.